Journal of Econometrics 186 (2015) 1–18

Contents lists available at ScienceDirect

Journal of Econometrics

journal homepage: www.elsevier.com/locate/jeconom

A spatial autoregressive model with a nonlinear transformation of the dependent variable

Xingbai Xu, Lung-fei Lee *

Department of Economics, The Ohio State University, Columbus, OH, United States

ARTICLE INFO

Article history:
Received 7 August 2013
Received in revised form
12 December 2014
Accepted 12 December 2014
Available online 9 January 2015

JEL classification:
C13
C21
C24
C63

Keywords:
Nonlinear spatial autoregressive model
Near-epoch dependence
Maximum likelihood
Instrumental variables
Asymptotic distribution of estimators

ABSTRACT

This paper develops a nonlinear spatial autoregressive model. Of particular interest is a structural interaction model for share data. We consider possible instrumental variable (IV) and maximum likelihood estimation (MLE) for this model, and analyze asymptotic properties of the IV and MLE based on the notion of spatial near-epoch dependence. We also design a statistical test to compare the nonlinear transformation against alternatives. Monte Carlo experiments are designed to investigate finite sample performance of the proposed estimates and the sizes and powers of the test.

1. Introduction

The linear spatial autoregressive (SAR) model $Y_n = \lambda W_n Y_n + X_n\beta + \epsilon_n$ has been widely studied. Many of the early studies of the model have been summarized in Anselin (1988), Anselin and Bera (1998) and LeSage and Pace (2009). Kelejian and Prucha (1999) and Lee (2007) study the generalized method of moments (GMM) applied to the SAR model. Lee (2004) studies asymptotic properties of the quasi-maximum likelihood estimator of the SAR model.

To obtain asymptotic properties of estimators in nonlinear spatial models, laws of large numbers (LLN) and central limit theorems (CLT) are necessary. Jenish and Prucha (2009) establish the CLT, the uniform and pointwise LLN for spatial mixing processes. Jenish and Prucha (2012) study asymptotic properties of near-epoch dependent (NED) random fields. Subsequently, Jenish (2012) considers the estimation of a nonparametric regression function of NED processes. Even though the previously mentioned studies provide general asymptotic theories of large samples, we found that there are few studies for specific parametric nonlinear spatial models. In this paper, we explore the usefulness of the spatial NED theories for the estimation of a nonlinear SAR model that involves a nonlinear transformation.

Some types of spatial models are designed to deal with share data or positive data. In this study, "share data" refers to samples with observed dependent variables whose values are between zero and one. In this paper, we study share data with values in the open interval (0, 1). An earlier example is in Lin and Lee (2010), which studies a model of share data pertaining to county teenage pregnancy rates. However, they adopted the conventional linear SAR model for their study. As a county's teenage pregnancy rate must be between zero and one, a linear model at best could only approximate the true model. This paper proposes a nonlinear model with interactions, which takes into account the limited range of the share variable. More specifically, because share data take values in (0, 1), we formulate the model as $s_{i,n} = F(\lambda_0 w_{i\cdot,n} S_n + x_{i,n}\beta_0 + \epsilon_{i,n})$, where $F(\cdot)$ is a strictly increasing cumulative probability function on the real line R, and $s_{i,n}$ represents the share variable of unit i while the sample size is n. While the interest of this model is motivated by share variables, we consider a more general setting of such a model with $F(\cdot)$ being a smooth monotonic function and not necessarily a distribution function so that the setting can be also used to study other types of variables, such as positive dependent variables.

* Corresponding author. Tel.: +1 614 292 5508.
E-mail address: lee.1777@osu.edu (L.-f. Lee).

http://dx.doi.org/10.1016/j.jeconom.2014.12.005

This paper suggests estimation methods, namely, the maximum likelihood (ML) method and the two-stage least squares (2SLS) estimation, for the unknown parameters λ_0 and β_0 while maintaining the setting that $F(\cdot)$ is a known function. We first show that the outcome $s_{i,n}$ generated from this model is a spatial NED random field. Then, we provide asymptotic analysis for parameter estimates of this nonlinear spatial model based on the newly developed LLN and CLT in Jenish and Prucha (2012) for spatial NED random fields. Our analysis goes beyond that of the popular SAR model in the spatial literature.

This paper is organized as follows. We introduce the nonlinear SAR model and derive the spatial NED property of the dependent variable generated by this model in Section 2. We consider the estimation of this model by the ML method and prove the consistency and asymptotic normality of the MLE in Section 3. In addition to the ML approach, Section 4 considers the IV estimation, which includes the 2SLS approach, and a procedure to test a nonlinear functional form against some alternatives based on 2SLS estimation. Finally, Monte Carlo experiments are conducted in Section 5 to investigate the finite sample performance of the estimates and sizes and powers of the test. All proofs for propositions and theorems are collected in Appendices.[1]

2. The model and near-epoch dependence

As described in the introduction, we consider the model

$$s_{i,n} = F(\lambda_0 w_{i\cdot,n} S_n + x_{i,n}\beta_0 + \epsilon_{i,n}), \tag{1}$$

for $i = 1, \ldots, n$, where $F(\cdot)$ is a strictly increasing and continuous function on the real line $\mathbb{R}$ and $x_{i,n} = (x_{i1,n}, \ldots, x_{iK,n}) \in \mathbb{R}^K$ is the vector of exogenous variables. In this paper, we consider a parametric model in which the functional form of F is known and does not involve any unknown parameters. For example, $F(\cdot)$ can be the distribution function of the standard normal distribution $\Phi(\cdot)$, the logistic distribution, $F(x) = 1/(1+e^{-x})$, or the function $F(x) = (x + \sqrt{x^2+4})/2$ with range $(0, \infty)$. $S_n = (s_{1,n}, \ldots, s_{n,n})'$ is the n-dimensional column vector of outcomes.

This model covers and goes beyond linear spatial interaction models. Thus it can possibly enable broader application of spatial and network interaction models. Here are some possible applications: (1) Share data and percentage data that satisfy $s_{i,n} \in (0, 1)$. An example is violent crime rates for all US states in a year. Another example is the test pass rates of different schools or school districts, e.g., in Papke and Wooldridge (2008). (2) Many data in economics, such as the GDP of different regions and stock prices belonging to the same industry, are strictly positive and might be spatially correlated. One way to model such data is to choose a strictly increasing and positive $F(\cdot)$. In this paper, we shall consider the estimation of the model (1) by the methods of ML and IV estimation.

There may be some possible concerns about our model and estimation methods.[2] (1) In data sets with a non-negative dependent variable and a significant numbers of observations taking on the value 0, the above model is not suitable, and we should consider using a Tobit model instead (see Xu and Lee, 2014). (2) The strictly increasing assumption of $F(\cdot)$ might be too strong. For this concern, however, we note that, in many economics studies, one may prefer that the marginal effects of exogenous variables maintain the same signs and having a monotonic property. In those situations, the strictly increasing assumption of $F(\cdot)$ is preferred. The strictly monotonic assumption is widely used in the transformation model literature (see, e.g. Horowitz, 1996; Chen, 2002). (3) Since $F(\cdot)$ is strictly increasing, its inverse exists and thus the model can be written as $F^{-1}(s_{i,n}) = \lambda_0 w_{i\cdot,n} S_n + x_{i,n}\beta_0 + \epsilon_{i,n}$ and IV estimation can be applied to estimate the model. This assertion is correct and we will discuss the IV and 2SLS estimations in Section 4. However, properties such as consistency and the asymptotic distribution of an IV estimator do not follow from existing literature on typical IV estimation with cross section or time series data. Also, they do not follow from existing IV estimation for the linear SAR model. Thus, rigorous study of those properties of an IV estimator still needs to be conducted. We also study the MLE as it can be more efficient than IV estimators. (4) It might be a strong assumption that the functional form is known. Without a known function for $F(\cdot)$, the model will be a semi-parametric one. We will explore such a model in future research. This paper will focus on a parametric model, as such a study can be a good starting point to understand the properties of popular estimation methods.

As S_n is endogenous, Eq. (1) is a well-defined model if the system determines a unique vector S_n of outcomes given ϵ_n and X_n, where X_n is an $n \times K$ matrix of exogenous variables $x_{i,n}$'s and ϵ_n is the vector of disturbances. This is possible if there are proper restrictions on the interaction effect λ and the spatial weights matrix W_n, whose ith row is $w_{i\cdot,n}$. The implied system of the specified equations in (1) for all n units is

$$S_n = \begin{pmatrix} F(\lambda w_{1\cdot,n} S_n + x_{1,n}\beta + \epsilon_{1,n}) \\ F(\lambda w_{2\cdot,n} S_n + x_{2,n}\beta + \epsilon_{2,n}) \\ \vdots \\ F(\lambda w_{n\cdot,n} S_n + x_{n,n}\beta + \epsilon_{n,n}) \end{pmatrix}. \tag{2}$$

Before further discussion, we list some of our formal assumptions. The first set of assumptions concerns the geographical setting of spatial units:

Assumption 1. Individual units in an economy are located or living in a region $D_n \subset D \subset \mathbb{R}^d$, where $\lim_{n\to\infty} |D_n| = \infty$ and $\mathbb{R}^d$ is the finite dimensional Euclidean space of dimension d. The distance between every two individuals is larger than or equal to a specific positive constant, say, 1.

The distance, as referred to in Assumption 1, can be defined from the norm $\|(x_1, \ldots, x_d)\|_\infty \equiv \max_i |x_i|$ or other norms. The above assumption is similar to that in Jenish and Prucha (2012). It means, in a bounded space, there are at most a finite number of units even if the population is infinite.

Assumption 2. Only individuals whose distances are less than or equal to some specific constant may affect each other. Without loss of generality, we set it as d_0, which is greater than 1.[3]

The elements of the spatial weights matrix are defined in terms of the strength of neighbors' direct interactions with each other. Under Assumptions 1 and 2, it follows immediately that in every row i and column j in W_n, the total number of non-zero elements is less than or equal to some finite constant uniformly in i, j and n.

For the spatial weights matrix W_n, as n tends to infinity, we have a sequence of square matrices $\{W_n\}$ increasing in dimension. It is valuable to summarize some of the regularity for $\{W_n\}$ in terms of relevant matrix norms. As shown in Kelejian and Prucha (2001), the matrix norms $\|.\|_\infty$ and $\|.\|_1$ induced, respectively, by the vector norms $\|.\|_\infty$ and $\|.\|_1$ are of particular interest. Explicitly, $\|W_n\|_1 = \max_{j=1,\ldots,n} \sum_{i=1}^n |w_{ij,n}|$ is known as the column sum norm, and $\|W_n\|_\infty = \max_{i=1,\ldots,n} \sum_{j=1}^n |w_{ij,n}|$ is the row sum norm.

1 A supplement file which provides additional analysis and results are also available upon request (see Appendix C).

2 We appreciate having these comments from referees.

3 This allows individuals to have interactions with others, as individuals live at least one unit of distance apart in Assumption 2.

In the linear SAR model, it is required that $\sup_{\lambda,n} \|\lambda W_n\|_\infty < 1$. In our paper, we have a similar assumption. As $F(x)$ is strictly increasing, its derivative exists almost everywhere. The next assumption concerns the derivative function F' of F.

Assumption 3. The function $f(x) = F'(x) > 0$ for all $x \in \mathbb{R}$, and the following condition holds: $\zeta \equiv \lambda_m b_f \sup_n \|W_n\|_\infty < 1$, where $b_f = \sup_x f(x)$, $\lambda_m = \sup_{\lambda\in\Lambda} |\lambda|$ with Λ being the compact parameter space of λ on the real line.

Assumption 3 implies that elements in W_n for all n are uniformly bounded. Because the number of nonzero elements in each column is uniformly bounded, $\{W_n\}$ is uniformly bounded in both row and column sum norms. In many studies on linear SAR models in the spatial econometric literature, the uniform boundedness in both row and column sum norms for W_n is a stated assumption. In those cases, the uniform boundedness of elements of W_n is an implied necessary condition. The uniform boundedness of W_n in both row and column sum norms for a linear SAR model is important in order to make the SAR system stable as n tends to infinity. Assumption 2, on the geographical setting, is a stronger than usual assumption for a linear SAR model. However, in many empirical applications, such a specification is used. We find it to be analytically tractable and simpler to adopt this assumption for our asymptotic analysis of estimators for the nonlinear SAR model (1).

As ζ is assumed to be finite, Assumption 3 has implicitly assumed that $f(x)$ is bounded. The logistic, normal, extreme value, Laplace and t distributions satisfy this assumption. The function $F(x) = \frac{1}{2}(x+\sqrt{x^2+4})$ also satisfies this assumption. This assumption is useful to establish the NED property of S_n that will be discussed later. If W_n is row normalized, then $\|W_n\|_\infty = 1$ and $\zeta = \lambda_m b_f$; hence, the condition in Assumption 3 for B will be satisfied if $\lambda_m b_f < 1$. This condition will, in turn, restrict what the parameter space Λ of λ can be. For example, if F is the standard normal distribution, f will be the standard normal density and $b_f = 1/\sqrt{2\pi}$. For W_n being row-normalized, Λ can be taken as a compact subset of $(-\sqrt{2\pi}, \sqrt{2\pi})$. If $F(x) = 1/(1+e^{-x})$ is the logit distribution, then $b_f = 0.25$. For the logit transformation, the possible range of parameter values of λ will be a compact subset of $(-4, 4)$ when W_n is row-normalized. Under Assumption 3, the right hand side of Eq. (2) is a contraction mapping with respect to S_n, so Eq. (2) will surely have a unique solution as in the following proposition:

Proposition 1. *Under Assumption 3, there is exactly one solution S_n for Eq. (2).*

When Assumption 3 fails to hold, it is possible that Eq. (2) has multiple solutions and we do not study such cases in this paper. For example, when $F(x) = \exp(x)$, the system

$$\begin{pmatrix} \ln s_{i,n} \\ \vdots \\ \ln s_{n,n} \end{pmatrix} = \begin{pmatrix} \lambda w_{1\cdot,n} S_n + x_{1,n}\beta + \epsilon_{1,n} \\ \vdots \\ \lambda w_{n\cdot,n} S_n + x_{n,n}\beta + \epsilon_{n,n} \end{pmatrix} \tag{3}$$

might have several solutions. As a specific case, the system $(\ln s_1, \ln s_2) = (0.1 s_2, 0.1 s_1)$ has two solutions: $(s_1, s_2) = (1.1183, 1.1183)$ and $(35.7715, 35.7715)$.

Since our model is a nonlinear one with spatial correlation, in order to show the large sample properties of an estimator, we explore a type of weak dependence on the sample observations generated by the model. We consider NED random fields in this paper due to the intrinsic spatial autoregressive feature of the model. As in Jenish and Prucha (2012), for any random vector Y, $\|Y\|_p \equiv [\mathrm{E}|Y|^p]^{1/p}$, where $|Y|$ is the Euclidean norm of Y. $D_n \subset D$ is a finite set and $|D_n|$ is its cardinality.

Definition 1 (*NED*). Let $Z = \{Z_{i,n}, i \in D_n, n \geq 1\}$ be a random field with $\|Z_{i,n}\|_p < \infty$, $p \geqslant 1$, let $\epsilon = \{\epsilon_{i,n}, i \in D_n, n \geqslant 1\}$ be a random field, where $|D_n| \to \infty$ as $n \to \infty$, and let $d = \{d_{i,n}, i \in D_n, n \geqslant 1\}$ be an array of finite positive constants. Then the random field Z is said to be L_p-near-epoch dependent on the random field ϵ if $\|Z_{i,n} - E(Z_{i,n}|\mathcal{F}_{i,n}(s))\|_p \leqslant d_{i,n}\psi(s)$ for some function $\psi(s) \geqslant 0$ with $\lim_{s\to\infty} \psi(s) = 0$, where σ-field $\mathcal{F}_{i,n}(s) = \sigma(\{\epsilon_{j,n} : d(j, i) \leqslant s\})$. The $\psi(s)$, which is, without loss of generality, assumed to be non-increasing, is called the NED coefficient, and the $d_{i,n}$'s are called NED scaling factors. Z is said to be L_p-NED on ϵ of size $-\lambda$ if $\psi(s) = O(s^{-\mu})$ for some $\mu > \lambda > 0$. Furthermore, if $\sup_n \sup_{i\in D_n} d_{i,n} < \infty$, then Z is said to be uniformly L_p-NED on ϵ. If $\psi(s) = O(\rho^s)$, where $0 < \rho < 1$, then Z is called geometrically L_p-NED on ϵ.

The term of geometrically L_p-NED random fields can be found, for example, in Hill (2010). Obviously, geometrically L_p-NED random fields are also L_p-NED of size $-\lambda$ for any $\lambda > 0$.

Another assumption is needed regarding the disturbances in Eq. (1).

Assumption 4. For each n, $\epsilon_{i,n}$'s are i.i.d. $(0, \sigma_0^2)$ double arrays.

The regressors $x_{i,n}$'s may be treated as deterministic or random variables. For generality, they are treated as random variables with spatial correlation. For the following propositions on NED, the explicit spatial structure on $x_{i,n}$'s is unnecessary, but it will be needed later on.

Lemma 1. *Under Assumptions 3 and 4, if $\sup_n \mathrm{E}|\epsilon_{i,n}|^p < \infty$ and $\sup_{i,k,n} \|x_{ik,n}\|_p < \infty$ for some $0 < p \in \mathbb{Z}$, then $s_{i,n}$ is uniformly L_p bounded, i.e., $\sup_{i,n} \mathrm{E}|s_{i,n}|^p < \infty$.*

Proposition 2. *Under Assumptions 1–4, if $\sup_{i,k,n} \|x_{ik,n}\|_2 < \infty$, then $\|s_{i,n} - \mathrm{E}(s_{i,n}|\mathcal{F}_{i,n}(md_0))\|_2 \leqslant b_f(\sigma_0 + \|\beta_0\|_1 \sup_{i,n} \|x_{ik,n}\|_2)\zeta^{m+1} /(1-\zeta)$, where $\mathcal{F}_{i,n}(s) \equiv \sigma(\{\epsilon_{j,n}, x_{j,n} : d(j, i) \leqslant s\})$, i.e., $\{s_{i,n}\}_{i=1}^n$ is a geometrically L_2-NED random field on $\{\epsilon_{i,n}, x_{i,n}\}_{i=1}^n$ uniformly in i and n.*

Proposition 1 in Jenish and Prucha (2012) discusses the conditions under which a nonlinear system is L_2-NED and we apply their conclusion to obtain the above proposition for our system. Proposition 2 has a useful corollary.

Corollary 1. *Under Assumptions 1–4, if $\sup_{i,k,n} \|x_{ik,n}\|_2 < \infty$, then $\{w_{i\cdot,n}S_n\}_{i=1}^n$ is uniformly and geometrically L_2-NED: $\|w_{i\cdot,n}S_n - \mathrm{E}(w_{i\cdot,n}S_n|\mathcal{F}_{i,n}(md_0))\|_2 \leqslant \sigma_0\zeta^{m+1}/[\lambda_m b_f(1-\zeta)]$; if, in addition, $\sup_n \mathrm{E}|\epsilon_{i,n}|^p < \infty$ and $\sup_{i,k,n} \|x_{ik,n}\|_p < \infty$ for some $0 < p \in \mathbb{Z}$, then $\{w_{i\cdot,n}S_n\}_{i=1}^n$ is uniformly L_p bounded in i and n.*

Another interesting variable is $t_{i,n} := F^{-1}(s_{i,n})$, which is a transformed dependent variable. The model (1) has the following equivalent representation: $t_{i,n} = \lambda_0 w_{i\cdot,n}S_n + x_{i,n}\beta_0 + \epsilon_{i,n}$. Corollary 1 implies immediately that $\{t_{i,n}\}$ is a geometrically L_2-NED random field on $\{\epsilon_{i,n}, x_{i,n}\}_{i=1}^n$ uniformly in i and n.

3. The MLE and its large sample properties

In this section, we would like to consider the MLE method for the model (1). For the MLE approach, Assumption 4 needs to be strengthened such that $\epsilon_{i,n}$'s are normally distributed and we require that $\{x_{i,n}\}_{i=1}^n$ is an α-mixing random field with α-mixing coefficient $\alpha(u, v, r) \leqslant (u+v)^\tau \hat{\alpha}(r)$ for some $\tau \geqslant 0$ and $\lim_{r\to\infty} \hat{\alpha}(r) = 0$. The definition and some discussion of α-mixing random fields can be found in Jenish and Prucha (2009, 2012).

Assumption 5. $f(x) = F'(x)$ is a bounded Lipschitz function.

Assumption 6. $\epsilon_{i,n}$'s are i.i.d. $N(0,\sigma^2)$ double arrays; X_n and ϵ_n are independent.

Assumption 7. (i) $\{x_{i,n}\}_{i=1}^n$ is an α-mixing random field with α-mixing coefficient $\alpha(u,v,r) \leqslant (u+v)^\tau \hat{\alpha}(r)$ for some $\tau \geqslant 0$, where $\hat{\alpha}(r)$ satisfies $\sum_{r=1}^\infty r^{d-1}\hat{\alpha}(r) < \infty$. (ii) $\sup_{i,k,n} \|x_{ik,n}\|_5 < \infty$.[4]

Assumption 8. The parameter space Θ of $\theta = (\lambda, \beta', \sigma^2)'$ is a compact subset of R^{K+2}.

Recall $f_{D_n} = \mathrm{diag}\{f(t_{1,n}), \ldots, f(t_{n,n})\}$ is the diagonal matrix with $f(t_{1,n}), \ldots, f(t_{n,n})$ as its diagonal elements. Then, under normal disturbances, the conditional log-likelihood function of S_n from (1) is

$$\ln L_n(\theta) = -\frac{n}{2}\ln(2\pi\sigma^2) - \frac{1}{2\sigma^2}[F^{-1}(S_n) - \lambda W_n S_n - X_n\beta]' \times [F^{-1}(S_n) - \lambda W_n S_n - X_n\beta] + \ln|f_{D_n}^{-1} - \lambda W_n|. \quad (4)$$

Define $Q_n(\theta) \equiv \mathrm{E}[\ln L_n(\theta)]$. Now we will discuss identification. We shall present some sufficient conditions for identification with a finite sample. As the sample size tends to infinity, we assume that the identification remains valid.[5]

The following lemmas provide some regularity conditions in order to show that, when the sample size is finite, the true parameter vector can be identified as the unique maximizer of $Q_n(\theta)$.

Lemma 2. *Under Assumptions 3 and 6, when $W_n \neq 0$, X_n has full column rank, the characteristic values of $f_{D_n}W_n$ are all real, and $\lim_{x\to+\infty} F(x)/x = 0$, then $Q_n(\theta)$ is uniquely maximized at θ_0.*

The characteristic values of $f_{D_n}W_n$ are all real when W_n is symmetric. It holds also for Ord's case where W_n is constructed from row-normalization of a symmetric spatial matrix (Ord, 1975). To illustrate this point, suppose $W_n = R_n W_n^*$ where W_n^* is a symmetric matrix and R_n is a diagonal matrix with a strictly positive diagonal. As $f_{D_n}R_n$ is positive definite, it has a decomposition $f_{D_n}R_n = B_n B_n'$ where B_n is invertible. Hence, $f_{D_n}W_n = B_n B_n' W_n^* = B_n(B_n' W_n^* B_n)B_n^{-1}$. As $B_n' W_n^* B_n$ is symmetric, there exists an orthonormal matrix Q_n and real eigenvalue matrix Λ_n such that $B_n' W_n^* B_n = Q_n \Lambda_n Q_n'$. In consequence, $f_{D_n}W_n = B_n Q_n \Lambda_n Q_n' B_n^{-1} = P_n \Lambda_n P_n^{-1}$, where $P_n = B_n Q_n$, is diagonalizable and Λ_n is the diagonal matrix of eigenvalues of $f_{D_n}W_n$.

There are also other sufficient conditions that guarantee identification. The following is one of them:

Lemma 3. *Under Assumptions 3 and 7, if $W_n'W_n$ is not a diagonal matrix, elements of $W_n'W_n$ are not all the same, $w_{ii,n} = 0$ for all i, X_n has full column rank, $f(\cdot)$ is differentiable, and there is at least an $x \in \mathbb{R} \cup \{+\infty, -\infty\}$ such that $f'(x) = 0$ while $f(x) \neq 0$, then $Q_n(\theta)$ is uniquely maximized at θ_0.*

All of the technical conditions in Lemma 3 are easy to satisfy, and this lemma includes the linear case: $F(x) = x$. If $F(\cdot)$ is a distribution function, then $f(\cdot)$ is its density function. The condition $f'(x) = 0$ will be satisfied if $f(.)$ has some modes. The sufficient condition $f'(x) = 0$ rules out a strictly convex or concave $F(x)$ if we only consider $x \in \mathbb{R}$. The strictly increasing and strictly convex function $F = (x + \sqrt{x^2+4})/2$, which is considered in the Monte Carlo simulation, does not satisfy the condition $f'(x) = 0$ for some $x \in \mathbb{R}$, but we have $\lim_{x\to\infty} f'(x) = 0$. The preceding sufficient conditions guarantee the $Q_n(\theta)$ is uniquely maximized at θ_0 via the information inequality. In the limit as n tends to infinity, we assume the identification in terms of limiting information inequality remains valid.

Assumption 9. $\liminf_{n\to\infty} \frac{1}{n}[Q_n(\theta_0) - Q_n(\theta)] > 0$ for any $\theta \neq \theta_0$.

Having the identification, we still need to show the uniform convergence: $\frac{1}{n}\sup_{\theta\in\Theta} |\ln L_n(\theta) - Q_n(\theta)| \xrightarrow{p} 0$ and the equicontinuity of $\frac{1}{n}Q_n(\theta)$ in order to establish the consistency of the MLE. In proving the uniform convergence of the log-likelihood function, one of the key points is to show the uniform convergence of the component $[\ln|I_n - \lambda f_{D_n}W_n| - \mathrm{E}\ln|I_n - \lambda f_{D_n}W_n|]/n$, whose form is not similar to the usual form of LLN. To show its uniform convergence, the formula of the Taylor series of $\ln|I_n - \lambda W_n|$ in Qu and Lee (2013) is useful. For $\|\lambda f_{D_n}W_n\|_\infty \leqslant \zeta$, i.e., $|\lambda| \leqslant \zeta/\|f_{D_n}W_n\|_\infty$, which holds under Assumption 3, $\lim_{l\to\infty}\|(\lambda f_{D_n}W_n)^l\|_\infty \leqslant \lim_{l\to\infty}\|\lambda f_{D_n}W_n\|_\infty^l \leqslant \lim_{l\to\infty}\zeta^l = 0$. Because any two norms on a finite dimensional linear space are equivalent (Theorem 4, p. 260 Royden and Fitzpatrick, 2010) and the convergence for all elements in a sequence of matrices with the same dimension is equivalent to the convergence in matrix norm (Theorem 18.2.20, p. 431 Harville, 1997), $\lim_{l\to\infty}(\lambda f_{D_n}W_n)^l = 0$. Then by Theorem 18.2.16 (p. 429 Harville, 1997), $(I_n - \lambda f_{D_n}W_n)^{-1} = \sum_{l=0}^\infty \lambda^l (f_{D_n}W_n)^l$ for $|\lambda| \leqslant \zeta/\|f_{D_n}W_n\|_\infty$. Thus, by Theorem 21(ii) in Amemiya (1985, p. 461), $d\ln|I_n - \lambda f_{D_n}W_n|/d\lambda = -\mathrm{tr}[(I_n - \lambda f_{D_n}W_n)^{-1} f_{D_n}W_n] = -\sum_{l=0}^\infty \lambda^l \mathrm{tr}((f_{D_n}W_n)^{l+1})$.

When $\lambda = 0$, $\ln|I_n - \lambda f_{D_n}W_n| = 0$. When $\lambda \in (0, \zeta/\|f_{D_n}W_n\|_\infty]$, because

$$\left|\sum_{l=0}^L \lambda^l \mathrm{tr}((f_{D_n}W_n)^{l+1})\right| = \left|\sum_{l=0}^L \sum_{i=1}^n \lambda^l ((f_{D_n}W_n)^{l+1})_{ii}\right| \leqslant n\|f_{D_n}W_n\|_\infty \sum_{l=0}^\infty \zeta^l = \frac{n\|f_{D_n}W_n\|_\infty}{1-\zeta}, \quad (5)$$

the dominated convergence theorem is applicable:

$$\begin{aligned}\ln|I_n - \lambda f_{D_n}W_n| &= \int_0^\lambda \frac{d\ln|I_n - v f_{D_n}W_n|}{dv}dv \\ &= -\int_0^\lambda \sum_{l=0}^\infty v^l \mathrm{tr}((f_{D_n}W_n)^{l+1})dv \\ &= -\sum_{l=0}^\infty \int_0^\lambda v^l \mathrm{tr}((f_{D_n}W_n)^{l+1})dv \\ &= -\sum_{l=1}^\infty \frac{\lambda^l}{l}\mathrm{tr}((f_{D_n}W_n)^l) \\ &= -\sum_{l=1}^\infty \frac{\lambda^l}{l}\sum_{i=1}^n ((f_{D_n}W_n)^l)_{ii}. \end{aligned} \quad (6)$$

Similarly, when $\lambda \in [-\zeta/\|f_{D_n}W_n\|_\infty, 0)$, the series expansion also holds. Hence,

[4] Here, we consider the L_5 norm because in Lemma 1 the order of moments is an integer and in the proof, we require the order of moments to be greater than four.

[5] It was pointed out in Wooldridge (1994, p. 2653–2654) that, for M-estimation, "Verifying that θ_0 is the unique minimizer of $\bar{q}$ in either the stationary or heterogeneous case often requires knowing something about the distribution of conditioning variables, and so identification is often taken on faith unless there are reasons to believe it might fail. Newey and McFadden (Section 2.2) give three examples of how to verify identification in examples with identically distributed data". The $\bar{q}$ in Wooldridge (1994) is $\lim_{n\to\infty} n^{-1}Q_n(\theta)$ in this paper. Hence, even in simpler models with dependence and heterogeneity, it is usually hard to establish the identification in the limiting sense.

$$\frac{1}{n}(\ln|I_n - \lambda f_{D_n} W_n| - \mathrm{E}\ln|I_n - \lambda f_{D_n} W_n|)$$
$$= -\frac{1}{n}\sum_{l=1}^{\infty}\frac{\lambda^l}{l}\sum_{i=1}^{n}\sum_{j_1}\sum_{j_2}\cdots\sum_{j_{l-1}} w_{ij_1,n}w_{j_1j_2,n}\cdots$$
$$\times w_{j_{l-1}i,n}(f_i f_{j_1}\cdots f_{j_{l-1}} - \mathrm{E} f_i f_{j_1}\cdots f_{j_{l-1}}). \quad (7)$$

The next proposition is about the NED property of $f_i f_{j_1}\cdots f_{j_{l-1}}$ and the uniform convergence:

Proposition 3. (i) *Let f_i be the ith diagonal element of the diagonal matrix f_{D_n}. Under Assumptions 1–5, for every positive integer l and every point i, pick an arbitrary chain $f_i, f_{i_1}, f_{i_2}, \ldots, f_{i_l}$ such that $d(i, i_1) \leqslant d_0$ and $d(i_p, i_{p+1}) \leqslant d_0$ for all $1 \leqslant p \leqslant l-1$, then $\{f_i f_{i_1}\cdots f_{i_l}\}$ is geometrically L_2-NED uniformly in i and n.*

(ii) $\sup_{\lambda\in\Lambda}(\ln|I_n - \lambda f_{D_n}W_n| - \mathrm{E}\ln|I_n - \lambda f_{D_n}W_n|)/n \xrightarrow{p} 0$.

To show the uniform convergence, we adopt a strategy from Qu and Lee (2013). For any given small positive number $\epsilon > 0$, we can divide the summation in Eq. (7) into two parts ($l \leqslant K_0$ & $l > K_0$) for some constant K_0 that does not depend on n. We show the uniform convergence of the first part by properties of NED random fields and that the second part can be bounded by $\epsilon/2$, and thus we establish the uniform convergence. Details of the proof can be found in Appendices.

Theorem 1. *Under Assumptions 1–9, the MLE $\hat\theta$ is a consistent estimator of θ_0.*

With consistency of the estimator, we next discuss the asymptotic distribution of MLE. The partial derivatives of the log-likelihood function in Eq. (4) are $\frac{\partial \ln L_n(\theta)}{\partial\lambda} = \frac{1}{\sigma^2}(W_nS_n)'[F^{-1}(S_n) - \lambda W_nS_n - X_n\beta] - \mathrm{tr}[(I_n - \lambda f_{D_n}W_n)^{-1}f_{D_n}W_n]$, $\frac{\partial \ln L_n(\theta)}{\partial\beta} = \frac{1}{\sigma^2}X_n'[F^{-1}(S_n) - \lambda W_nS_n - X_n\beta]$ and $\frac{\partial \ln L_n(\theta)}{\partial\sigma^2} = -\frac{n}{2\sigma^2} + \frac{1}{2\sigma^4}[F^{-1}(S_n) - \lambda W_nS_n - X_n\beta]'[F^{-1}(S_n) - \lambda W_nS_n - X_n\beta]$. To deduce the CLT, we write the score as a summation. Denote $z_{i,n} = \sum_{j=1}^n w_{ij,n}s_{j,n}$ and $r_{ii,n} = \sum_{l=0}^\infty \lambda_0^l((f_{D_n}W_n)^{l+1})_{ii}$. From the first order condition, we have

$$\frac{1}{\sqrt{n}}\frac{\partial \ln L_n(\theta_0)}{\partial\theta} = \frac{1}{\sqrt{n}}\sum_{i=1}^n\begin{pmatrix} z_{i,n}\epsilon_{i,n}/\sigma_0^2 - r_{ii,n} - \mathrm{E}[z_{i,n}\epsilon_{i,n}/\sigma_0^2 - r_{ii,n}] \\ x_{i,n}'\epsilon_{i,n}/\sigma_0^2 \\ (\epsilon_{i,n}^2 - \sigma_0^2)/(2\sigma_0^4)\end{pmatrix}. \quad (8)$$

To prove the asymptotic normality of the estimator, a key step is to show that the above sequence of scores would obey a CLT. For that purpose, we need additional regularity conditions:

Assumption 10. θ_0 is in the interior of the parameter space Θ.

Assumption 11. (i) For some $\delta > 0$, the α-mixing coefficient of $\{x_{i,n}\}_{i=1}^n$ in Assumption 7 satisfies

$$\sum_{r=1}^\infty r^{d(\tau_*+1)}\hat\alpha^{\frac{\delta}{4+2\delta}}(r) < \infty,$$

where $\tau_* = \delta\tau/(2+\delta)$. (ii) $\Sigma_X \equiv \mathrm{p}\lim_{n\to\infty} n^{-1}\sum_{i=1}^n x_{i,n}'x_{i,n}$ is a positive definite matrix.

Assumption 12. $\Sigma_0 = \lim_{n\to\infty}\Sigma_n$ exists and is nonsingular, where $\Sigma_n = \frac{1}{n}\mathrm{Var}(\sum_{i=1}^n(\frac{z_{i,n}\epsilon_{i,n}}{\sigma_0^2} - r_{ii,n}, \frac{x_{i,n}\epsilon_i}{\sigma_0^2}, \frac{\epsilon_{i,n}^2 - \sigma_0^2}{2\sigma_0^4})')$.

By our assumptions, we know that $\frac{1}{\sqrt{n}}\sum_{i=1}^n(x_{i,n}\epsilon_{i,n}/\sigma_0^2, \epsilon_{i,n}^2 - \sigma_0^2)' \xrightarrow{d} N(0, \mathrm{diag}(\Sigma_X, 2\sigma_0^4))$, where the asymptotic variance is nonsingular. Therefore, the nonsingularity of Σ_0 may be mainly captured by the asymptotic variance of $\frac{1}{\sqrt{n}}\sum_{i=1}^n(z_{i,n}\epsilon_{i,n}/\sigma_0^2 - r_{ii,n})$ via the inverse form of a partitioned matrix. Alternatively, one may investigate the concentrated log likelihood function $\ln L_{cn}(\lambda)$ of λ with β and σ^2 concentrated out. The corresponding asymptotic variance of the normalized score of λ is $\lim_{n\to\infty}\frac{1}{n}\mathrm{Var}\sum_{i=1}^n(\frac{z_{i,n}\epsilon_{i,n}}{\sigma_0^2} - r_{ii,n}) - B[\mathrm{diag}(\lim_{n\to\infty}\frac{1}{n}\sum_{i=1}^n x_{i,n}'x_{i,n}, 2\sigma_0^4)]^{-1}B = -\lim_{n\to\infty}\frac{1}{n}\mathrm{E}[\partial^2\ln L_{cn}(\lambda_0)/\partial\lambda^2]$, where $B = \lim_{n\to\infty}\frac{1}{n}\mathrm{cov}(\sum_{i=1}^n(\frac{z_{i,n}\epsilon_{i,n}}{\sigma_0^2} - r_{ii,n}), (\frac{x_{i,n}\epsilon_{i,n}}{\sigma_0^2}, \epsilon_{i,n}^2 - \sigma_0^2))$. Thus, Assumption 12 preserves the local identification in the limit.

To establish the asymptotic normality, we apply the CLT from Jenish and Prucha (2012). To do so, we show that $\{[(\frac{z_{i,n}\epsilon_{i,n}}{\sigma_0^2} - r_{ii,n})^2 + (\frac{x_{i,n}\epsilon_i}{\sigma_0^2})^2 + (\frac{\epsilon_{i,n}^2 - \sigma_0^2}{2\sigma_0^4})^2]^{1/2}\}_{i=1}^n$ is uniformly and geometrically NED. Then with Assumption 12, we have the following result:

Proposition 4. *Under Assumptions 1–12, $\frac{1}{\sqrt{n}}\sum_{i=1}^n(\frac{z_{i,n}\epsilon_{i,n}}{\sigma_0^2} - r_{ii,n}, \frac{x_{i,n}\epsilon_{i,n}}{\sigma_0^2}, \frac{\epsilon_{i,n}^2 - \sigma_0^2}{2\sigma_0^4})' \xrightarrow{d} N(0, \Sigma_0)$.*

In order to derive the asymptotic distribution of an extremum estimator, as usual, one may investigate the linearization of the first order condition which characterizes the extremum estimator, by the mean value theorem (see, e.g Amemiya (1985)). For the ML estimation, this linearization will involve the product of the score and the Hessian matrix of the log likelihood. With Proposition 4, the score vector is asymptotically normal. The Hessian matrix can be shown to converge uniformly in probability to a non-singular matrix. Thus, the asymptotic distribution can be derived as in the following theorem:

Theorem 2. *Under Assumptions 1–12, $\sqrt{n}(\hat\theta_n - \theta_0) \xrightarrow{d} N(0, \Sigma_0^{-1})$.*

4. IV and two stage least square estimation

In this section, we consider IV estimation of our model. We keep Assumptions 1–4. Because IV estimation is distributionally free, the independence of disturbances in Assumption 4 will be sufficient and there is no need for the use of the normality in Assumption 6. With this independence assumption, $\{s_{i,n}\}$ remains a uniformly and geometrically L_2-NED random field on $\{\epsilon_n\}$ in Proposition 2.

IV estimation can be applied to the model expressed as $T_n = \lambda W_nS_n + X_n\beta + \epsilon_n = Z_n\delta + \epsilon_n$, where $Z_n = (W_nS_n, X_n)$ and $\delta = (\lambda, \beta')'$. For general 2SLS estimation, let Q_n be an IV matrix. In practice, possible IV variables can be X_n and $W_nX_{2,n}$, where $X_{2,n}$ is the submatrix of X_n with the exclusion of the intercept term ι_n when W_n is row normalized such that $W_n\iota_n = \iota_n$. But if W_n is not row normalized, W_nX_n can be used because $W_n\iota_n$ will not be equal to ι_n and may not be perfectly collinear with X_n. In addition to W_nX_n, $W_n^2X_n$ may also be used. With the IV matrix $Q_n = (q_{1,n}', \ldots, q_{n,n}')'$, the corresponding IV estimator is

$$\widehat{\delta_n} = [Z_n'Q_n(Q_n'Q_n)^{-1}Q_n'Z_n]^{-1}Z_n'Q_n(Q_n'Q_n)^{-1}Q_n'T_n. \quad (9)$$

Assumption 13. (i) The instrumental variable $\{q_{i,n}\}_{i=1}^n$ is a geometric L_2 NED random field on $\{x_{i,n}\}_{i=1}^n$ uniformly in i and n.[6]

[6] We can relax $\{q_{i,n}\}_{i=1}^n$ to be an NED random field with NED coefficient s^{-r} for some constant $r > 0$, but then we need to add a constraint on the ξ in condition (ii): $r > d(2\xi - 4)/(\xi - 4)$. To simplify the statement, we just assume geometric L_2 NED. We have an older version of this paper where we assume that $\{x_{i,n}\}_{i=1}^n$ and $\{q_{i,n}\}_{i=1}^n$ are exogenous deterministic variables, which are uniformly bounded. In that setting, the spatial process properties would not be needed.

(ii) $\sup_{i,n} \|q_{i,n}\|_{\xi} < \infty$ for some $\xi > 4$. (iii) Q_n and ϵ_n are independent for all n. (iv) $\Sigma_{QQ} \equiv \text{p}\lim_{n\to\infty} Q_n'Q_n/n$ exists and is positive definite. (v) $\Sigma_{ZQ} \equiv \text{p}\lim_{n\to\infty}(\text{E}Z_n)'Q_n/n$ exists and has full row rank $K+1$.

It is not difficult to verify that X_n and $W_nX_{2,n}$ satisfy Assumption 13. With Assumption 13, $\{(w_{i\cdot,n}S_n, x_{i,n}) \otimes q_{i,n}\}_{i=1}^n$ and $\{\epsilon_{i,n}q_{i,n}\}_{i=1}^n$ are uniformly $L_{\min(\xi/2,2.5)}$ bounded and geometric NED uniformly in i and n. Therefore, we have $\frac{1}{n}Q_n'\epsilon_n = o_p(1)$, $(Z_n - \text{E}Z_n)'Q_n/n \xrightarrow{p} 0$ and

Corollary 2. $\text{p}\lim_{n\to\infty} Z_n'Q_n/n = \Sigma_{ZQ}$.

Then, the consistency of the IV estimator $\widehat{\delta}_n$ follows because

$$\widehat{\delta}_n - \delta = \left[\frac{1}{n}Z_n'Q_n\left(\frac{1}{n}Q_n'Q_n\right)^{-1}\frac{1}{n}Q_n'Z_n\right]^{-1} \times \frac{1}{n}Z_n'Q_n\left(\frac{1}{n}Q_n'Q_n\right)^{-1}\frac{1}{n}Q_n'\epsilon_n = o_p(1). \quad (10)$$

As usual, σ^2 can be estimated by the sample average of the estimated residuals,

$$\widehat{\sigma_n^2} = \frac{1}{n}(T_n - \widehat{\lambda}_n W_nS_n - X_n\widehat{\beta}_n)'(T_n - \widehat{\lambda}_n W_nS_n - X_n\widehat{\beta}_n). \quad (11)$$

With Corollary 1, Assumptions 6 and 7, $\frac{1}{n}(W_nS_n)'W_nS_n$, $\frac{1}{n}\epsilon_n'W_nS_n$ and $\frac{2}{n}(W_nS_n)'X_n$ are all $O_p(1)$. Because

$$\begin{aligned}\widehat{\sigma_n^2} &= \frac{1}{n}[\epsilon_n - (\widehat{\lambda_n} - \lambda_0)W_nS_n - X_n(\widehat{\beta_n} - \beta_0)]' \\ &\quad \times [\epsilon_n - (\widehat{\lambda_n} - \lambda_0)W_nS_n - X_n(\widehat{\beta_n} - \beta_0)] \\ &= \frac{1}{n}\epsilon_n'\epsilon_n + \frac{1}{n}(\widehat{\lambda_n} - \lambda_0)^2(W_nS_n)'W_nS_n \\ &\quad + (\widehat{\beta_n} - \beta_0)'\frac{1}{n}X_n'X_n(\widehat{\beta_n} - \beta_0) - \frac{2}{n}(\widehat{\lambda_n} - \lambda_0)\epsilon_n'W_nS_n \\ &\quad - \frac{2}{n}\epsilon_n'X_n(\widehat{\beta_n} - \beta_0) + \frac{2}{n}(\widehat{\lambda_n} - \lambda_0) \\ &\quad \times (W_nS_n)'X_n(\widehat{\beta_n} - \beta_0), \end{aligned} \quad (12)$$

the consistency of $\widehat{\lambda_n}$ and $\widehat{\beta_n}$ implies the consistency of $\widehat{\sigma_n^2}$.

We can apply the CLT in Jenish and Prucha (2012) to $Q_n'\epsilon_n/\sqrt{n}$ and obtain the asymptotic normality for the 2SLS estimator:

$$\sqrt{n}(\widehat{\delta_n} - \delta) \xrightarrow{d} N(0, \sigma_0^2(\Sigma_{ZQ}\Sigma_{QQ}^{-1}\Sigma_{ZQ}')^{-1}). \quad (13)$$

If $\text{E}Z_n$ is taken as an IV matrix, then the asymptotic variance becomes $\lim_{n\to\infty} \text{E}Z_n'\text{E}Z_n/n$. Since $(\text{E}Z_n'Q_n)(Q_n'Q_n)^{-1}(Q_n\text{E}Z_n) \leqslant \text{E}Z_n'\text{E}Z_n$ for any IV matrix Q_n, $\text{E}Z_n$ is the optimal IV matrix. In sum, we have

Theorem 3. *Under Assumptions 1–4, 7, 11 and 13,* $\sqrt{n}(\widehat{\delta_n} - \delta) \xrightarrow{d} N(0, \sigma_0^2(\Sigma_{ZQ}\Sigma_{QQ}^{-1}\Sigma_{QZ})^{-1})$. *Furthermore,* $\text{E}Z_n$ *is the optimal IV matrix, with which the asymptotic variance of the estimator is* $\sigma_0^2(\text{E}Z_n'\text{E}Z_n)^{-1}$.

As the distribution of $\epsilon_{i,n}$ is unknown, the optimal IV estimation would not have a closed form expression for convenient use.[7] Intuitively, we propose a feasible simulated optimal IV estimation:

(1) Use a general 2SLS estimator $\widetilde{\delta_n}$ derived from using some IVs such as (X_n, W_nX_n), and get the residuals $\widehat{\epsilon_{i,n}}$'s.

(2) Use the empirical distribution of $\widehat{\epsilon_{i,n}}$'s to generate R number of $\epsilon_{rn} = (\epsilon_{1,rn}, \ldots, \epsilon_{n,rn})'$, and use these to generate RS_{rn}'s, and evaluate their empirical mean as $\widehat{\text{E}S_n}$.

(3) Use $(W\widehat{\text{E}S_n}, X_n)$ as IV to obtain $\widehat{\delta_n} = [(W\widehat{\text{E}S_n}, X_n)'Z_n]^{-1}(W\widehat{\text{E}S_n}, X_n)'T_n$.

The Monte Carlo experiments in Section 5 show that the simulated optimal IV estimator is more efficient than the 2SLS estimator in most cases.

The 2SLS estimation also provides a method to test a specified functional form $F(\cdot)$ against an alternative: H_0: the true functional form is F_1; H_1: the alternative functional form is F_2. Denote $t_{i,n} = F_1^{-1}(s_{i,n})$ and $\tilde{t}_{i,n} = F_2^{-1}(s_{i,n})$. Let $\gamma \neq 0$ be any constant. We consider the following model: $(1-a)t_{i,n} + a\gamma\tilde{t}_{i,n} = \lambda w_{i\cdot,n}S_n + x_{i,n}\beta + \epsilon_{i,n}$, i.e.,

$$t_{i,n} = a(t_{i,n} - \gamma\tilde{t}_{i,n}) + \lambda w_{i\cdot,n}S_n + x_{i,n}\beta + \epsilon_{i,n}. \quad (14)$$

Then H_0 is equivalent to $a = 0$ and H_1 is equivalent to $a = 1$. We can show that $\{\tilde{t}_{i,n}\}$ is also an NED random field for several widely used distributional families when the true $F_1(\cdot)$ is a logit (or normal) transformation.[8] Because $t_{i,n}$ is strictly increasing with respect to $\tilde{t}_{i,n}$, usually there is serious collinearity between $t_{i,n}$ and $\tilde{t}_{i,n}$. Thus we should choose a γ to eliminate some of the possible collinearity. Let $f_1(\cdot)$ and $f_2(\cdot)$ be respectively the derivatives of $F_1(\cdot)$ and $F_2(\cdot)$. Since $\tilde{t}_{i,n} = F_2^{-1}(s_{i,n}) = F_2^{-1}(F_1(t_{i,n}))$,we have $d\tilde{t}_{i,n}/dt_{i,n} = f_1(t_{i,n})/f_2(\tilde{t}_{i,n})$. That is $dt_{i,n}/d\tilde{t}_{i,n} = f_2(\tilde{t}_{i,n})/f_1(t_{i,n})$. Thus we choose γ be the mean of $f_2(\tilde{t}_{i,n})/f_1(t_{i,n})$. Experiments show that this can significantly reduce the multicollinearity. For example, when $F_1(\cdot)$ is the logit and $F_2(x)$ is the standard normal distribution function, the R^2 of regressing $t_{i,n}$ on $\tilde{t}_{i,n}$ is about 0.99 while the R^2 of regressing $t_{i,n}$ on $t_{i,n} - \gamma\tilde{t}_{i,n}$ is only about 0.05. Bootstrapping is utilized to obtain a more precise critical value to test H_0. We can do the test in the following steps:

(1) Estimate $t_{i,n} = \lambda w_{i\cdot,n}S_n + x_{i,n}\beta + \epsilon_{i,n}$ by 2SLS and obtain the residuals $\widehat{\epsilon_n}$, whose empirical distribution is $F_{\widehat{\epsilon_n}}$;

(2) Generate n random draws $\epsilon_{i,n}^{(r)}$'s from the distribution $F_{\widehat{\epsilon_n}}$, and then generate $S_n^{(r)}$ by contraction mapping, calculate $\gamma^{(r)}$ and estimate the equation $t_{i,n}^{(r)} = a(t_{i,n}^{(r)} - \gamma^{(r)}\tilde{t}_{i,n}^{(r)}) + \lambda w_{i\cdot,n}S_n^{(r)} + x_{i,n}\beta + \epsilon_{i,n}^{(r)}$ with 2SLS to obtain $\hat{a}^{(r)}$: we can adopt X_n, $W_nX_{2,n}$ and $W_n^2X_{2,n}$ as the IV's, where $X_{2,n}$ is the exogenous variable matrix without the constant;

(2) Repeat Step (2) R times and obtain the bootstrap critical value for 5% level of significance for a one-sided test $H_0 : a = 0$ against $H_1 : a = 1$.

5. Monte Carlo experiments

5.1. Estimation

In this section, we conduct some Monte Carlo experiments to study the finite sample properties and the robustness of our estimators. Specifically, we would like to investigate the following four issues in the experiments: (1) comparing the marginal effects of nonlinear and linear models; (2) the precision of predictions from nonlinear and linear models if the true model is nonlinear; (3) the finite sample performance of our estimators; and (4) the robustness of the QMLE if ϵ_i is not normally distributed.

In our experiments, $s_{i,n} = F(\lambda w_{i\cdot,n}S_n + \beta_1 + \beta_2 x_{i,n} + \epsilon_{i,n})$, where the true values of coefficients are $(\beta_{1,0}, \beta_{2,0}) = (-1, 1)'$ and $\epsilon_{i,n}$'s are i.i.d. $N(0, \sigma_0^2)$. The $x_{i,n}$'s are designed to allow spatial correlation: $(x_{1,n}, \ldots, x_{n,n})' \sim 1.5(I_n - 0.2W_n)^{-1}N(0, I_n)$. The generation

[7] Even if $\epsilon_{i,n}$ is known to be normally distributed, a closed form expression is still hard to get due to the nonlinearity of the model.

[8] See the supplement material for the proof (see Appendix C).

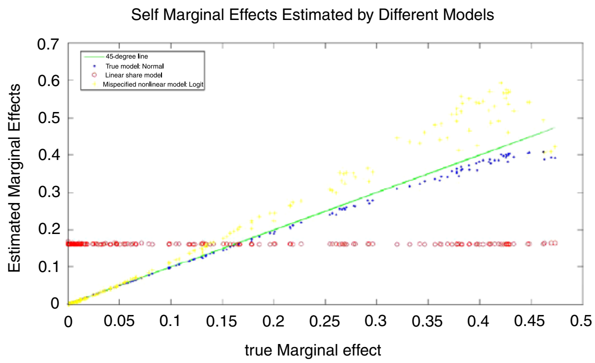

Fig. 1. Self marginal effects when true $F(x) = \Phi(x)$.

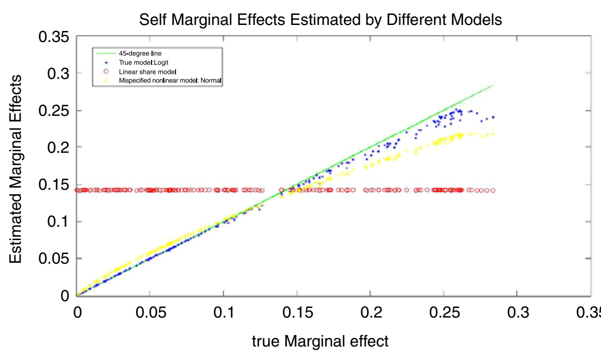

Fig. 2. Self marginal effects when true $F(x) = (1 + e^{-x})^{-1}$.

of W_n will be discussed in the next paragraph. In the experiments, three different nonlinear functions, namely, $F(x) = 1/(1 + e^{-x})$, $F(x) = \Phi(x)$ and $F(x) = 0.5(x + \sqrt{x^2 + 4})$, are considered. The third function is a strictly increasing convex function with two asymptotes, $y = x$ and $y = 0$. When $F(x) = 1/(1+e^{-x})$ and $F(x) = \Phi(x)$, σ_0's are respectively 1.5 and 1, since the normal distribution has thinner tails while the logit distribution has relatively thicker tails. The true λ_0 is designed to be 1 or 1.5 so that the contraction mapping holds for each of these two models. When $F(x) = 0.5(x+\sqrt{x^2 + 4})$, λ_0 is 0.4 or 0.7. We consider various sample sizes of 100, 200, 500 and 1000. Detailed parameters with corresponding designs on $x_{i,n}$ and $F(\cdot)$ are noted in each of the tables in Appendices.

The weights matrix W_n is generated from county data in the US. When the sample sizes n are 100, 200 and 500, W_n is generated from 761 counties in 10 states as in Lin and Lee (2010). First, we construct W_{0n} as follows: $W_{ij,0n}$ equals 1 if county i and county j are contiguous, zero otherwise. In our Monte Carlo experiments, we generate W_n randomly from W_{0n} as follows: we generate a natural number m uniformly distributed between 1 and $(761 - n)$, and then use the entries of W_{0n} that are between the mth row and the $(m + n - 1)$th row and between the mth column and the $(m + n - 1)$th column to form an n by n matrix $\widetilde{W}_n$. Then we row-normalize $\widetilde{W}_n$ to get the weights matrix W_n. When the sample size is 1000, we do it in a similar way, except that W_n is generated from all 3142 counties in the US.

As the conditions of the contraction mapping theorem hold, we can generate S_n using contraction mapping. We start by letting $s^{(0)}_{i,n} = F(\beta_1+\beta_2 x_{i,n}+\epsilon_{i,n})$, then $s^{(j+1)}_{i,n} = F(\lambda W_{i\cdot,n} S^{(j)}_n + \beta_1 + \beta_2 x_{i,n} + \epsilon_{i,n})$. The iteration stops when $\max_i |s^{(j+1)}_{i,n} - s^{(j)}_{i,n}| < 10^{-8}$.

Besides MLE, we also do IV and 2SLS estimation. For IV estimation, we use $W_n X_{2,n} = W_n(x_{1,n}, \ldots, x_{n,n})'$ as the IV for $W_n S_n$, where $X_{2,n}$ is the second column of X_n.[9] For 2SLS, we use $W_n X_{2,n}$ and $W_n^2 X_{2,n}$, as the IVs for $W_n S_n$.

In the last experiment, we investigate the performance of the estimators when the normality of the error terms does not hold. We try four different distributions: uniform, $t(5)$, mixed normal and $\beta(0.5, 0.5)$ distributions. To make our results here comparable to those in the normal distribution case, we normalize and scale these distributions such that their expectations are all zero and their standard deviations are all 1.5. Explicitly, we generate random numbers from the following four distributions: mixed normal (with half probability $N(6/\sqrt{17}, 9/68)$ and half probability $N(-6/\sqrt{17}, 9/68)$); $\sqrt{1.35}$ times $t(5)$, where $t(5)$ is the Student t-distribution with five degrees of freedom; uniform distribution $U(-1.5\sqrt{3}, 1.5\sqrt{3})$; and $\sqrt{18}(\beta(0.5, 0.5) - 0.5)$, where $\beta(a, b)$ is the two-parameter beta distribution with parameters a and b. Notice that the density of the mixed normal has double peaks and that $\beta(0.5, 0.5)$ has a U shape on $(0, 1)$.

To get the empirical means, standard deviations and root mean squared errors (RMSE) of the estimates, we do 1000 repetitions for each design.

Marginal effects of exogenous variables are often considered in empirical studies. Hence we first consider the marginal effects in the Monte Carlo experiments. For illustrative purposes, we focus on the self marginal effect, $\partial s_{i,n}/\partial x_{i,n} = \beta_2[(I_n - \lambda f_{D_n} W_n)^{-1} f_{D_n}]_{ii}$. With a sample size $n = 200$ and $F(x) = \Phi(x)$, we show the self marginal effects in Fig. 1; when $F(x) = 1/(1 + e^{-x})$, the result is shown in Fig. 2. We have the true self marginal effects on the horizontal axis. Thus, points on the 45-degree line are equal to the true self marginal effects. We can see from the graph that the estimated self marginal effects are much more accurate than those estimated by linear models. If we use a linear SAR model, the estimated self marginal effects will be nearly the same for all individuals. Different sample sizes and parameters have been tried and their figures are similar to Figures 1 and 2.[10]

Second, we examine the predictions of different models. Here let us recall binary choice models, which are usually estimated by a probit or logit model, though the linear probability model is easier and usually gives the same signs for estimators of coefficients. One of the drawbacks of the linear probability model for binary choice models is that its predicted probability can be greater than 1 or less than zero. When the range of dependent variables is not $\mathbb{R}$, similar phenomena appear. As can be seen from Fig. 5, 11% of predicted values of the dependent variable from the linear SAR model are out of the interval $(0, 1)$ when the true model is $s_{i,n} = F(\lambda w_{i.,n} S_n + \beta_1 + \beta_2 x_{i,n} + \epsilon_{i,n})$, with $F(x) = 1/(1 + e^{-x})$. Besides, we compare the distance between S_n and its estimated value $\hat{S}_n$ by 1-norm and 2-norm: $\|S_n - \hat{S}^{(true)}_n\|_1 = 16.3704 < 19.2970 = \|S_n - \hat{S}^{(linear)}_n\|_1$ and $\|S_n - \hat{S}^{(true)}_n\|_2 = 2.1037 < 2.3008 = \|S_n - \hat{S}^{(linear)}_n\|_2$. These results show that the true nonlinear model has better prediction. To check the robustness of our conclusion, we also try various sample sizes, parameters and functional forms and we obtain similar figures and conclusions.

From Tables 1–3, we have several observations:

(1) As the sample size increases, both biases and variances of estimators decrease. This verifies the consistency of the estimators.

(2) For most experiments, the biases of IV, 2SLS and simulated optimal IV estimates are less than the bias of MLE.

(3) When we compare the variance of estimators, the simulated optimal IV estimation is more efficient than the 2SLS (especially when the sample size $n \geqslant 200$), and the 2SLS is a little bit more efficient than the IV estimation. The variance of MLE is obviously less than those of IV/2SLS/optimal IV estimators. For instance, from

[9] The first column of X_n is the constant intercept term.

[10] Those figures can be found in a supplement file (see Appendix C).

Table 1
Estimation results when $F(x) = 1/(1+\exp(-x))$.

λ_0	n		IV			2SLS			Optimal IV			MLE		
			mean	sd	RMSE	mean	sd	RMSE	mean	sd	RMSE	mean	sd	RMSE
1	100	λ	0.9111	1.3373	1.3402	0.9784	1.2916	1.2918	0.9028	1.3840	1.3875	0.8669	0.8282	0.8389
		β_1	−0.9598	0.6123	0.6136	−0.9900	0.5919	0.5919	−0.9558	0.6339	0.6354	−0.9409	0.4006	0.4050
		β_2	0.9980	0.1033	0.1033	0.9955	0.1029	0.1030	0.9983	0.1034	0.1034	1.0045	0.0944	0.0945
	200	λ	1.0216	0.9388	0.9390	1.0604	0.9266	0.9285	1.0105	0.9147	0.9147	0.9522	0.5506	0.5526
		β_1	−1.0084	0.4007	0.4008	−1.0243	0.3960	0.3967	−1.0037	0.3916	0.3916	−0.9802	0.2499	0.2506
		β_2	0.9989	0.0889	0.0889	0.9963	0.0884	0.0884	0.9995	0.0879	0.0879	1.0056	0.0735	0.0737
	500	λ	0.9300	0.7243	0.7276	0.9541	0.7138	0.7153	0.9237	0.7069	0.7110	0.9505	0.4115	0.4144
		β_1	−0.9677	0.3101	0.3117	−0.9777	0.3057	0.3065	−0.9650	0.3025	0.3045	−0.9763	0.1850	0.1865
		β_2	1.0009	0.0499	0.0499	1.0000	0.0496	0.0496	1.0011	0.0493	0.0493	1.0013	0.0443	0.0444
	1000	λ	0.9899	0.2563	0.2565	0.9921	0.2540	0.2542	0.9912	0.2470	0.2471	0.9901	0.1576	0.1579
		β_1	−0.9961	0.0962	0.0963	−0.9969	0.0955	0.0955	−0.9966	0.0933	0.0934	−0.9963	0.0647	0.0648
		β_2	0.9983	0.0327	0.0327	0.9982	0.0326	0.0327	0.9983	0.0324	0.0325	0.9988	0.0312	0.0312
1.5	100	λ	1.4176	1.2879	1.2906	1.4796	1.2473	1.2475	1.4101	1.2901	1.2932	1.3715	0.7902	0.8006
		β_1	−0.9594	0.6393	0.6406	−0.9896	0.6197	0.6198	−0.9557	0.6415	0.6430	−0.9381	0.4130	0.4176
		β_2	0.9982	0.1037	0.1037	0.9957	0.1033	0.1034	0.9985	0.1036	0.1036	1.0048	0.0945	0.0946
	200	λ	1.5219	0.8984	0.8987	1.5575	0.8879	0.8897	1.5021	0.8694	0.8694	1.4501	0.5201	0.5225
		β_1	−1.0093	0.4158	0.4159	−1.0251	0.4112	0.4120	−1.0006	0.4033	0.4033	−0.9776	0.2554	0.2564
		β_2	0.9989	0.0890	0.0890	0.9964	0.0885	0.0885	1.0001	0.0878	0.0878	1.0060	0.0733	0.0736
	500	λ	1.4341	0.6870	0.6902	1.4547	0.6777	0.6792	1.4257	0.6735	0.6776	1.4527	0.3899	0.3928
		β_1	−0.9669	0.3200	0.3217	−0.9762	0.3157	0.3166	−0.9631	0.3135	0.3156	−0.9755	0.1903	0.1918
		β_2	1.0009	0.0501	0.0501	1.0002	0.0498	0.0498	1.0013	0.0496	0.0496	1.0013	0.0444	0.0444
	1000	λ	1.4848	0.3872	0.3875	1.4881	0.3842	0.3844	1.4854	0.3727	0.3730	1.4823	0.2374	0.2380
		β_1	−0.9928	0.1807	0.1808	−0.9942	0.1794	0.1795	−0.9931	0.1746	0.1748	−0.9917	0.1175	0.1178
		β_2	0.9983	0.0327	0.0328	0.9982	0.0327	0.0327	0.9983	0.0325	0.0325	0.9989	0.0312	0.0312

$F(x) = 1/(1+\exp(-x))$, $X_{2,n} = (x_{1,n}, \ldots, x_{n,n})' \sim 1.5(I_n - 0.2W_n)^{-1}N(0, I_n)$, ϵ_i *iid* $\sim N(0, 1.5)$, $\beta_0 = (-1, 1)'$.
IV: use $W_n X_{2,n}$ as the IVs of $W_n S_n$. 2SLS: use $W_n X_{2,n}$ and $W_n^2 X_{2,n}$ as the IV of $W_n S_n$. Repetition: 1000.

Table 2
Estimation results when $F(x) = \Phi(x)$.

λ_0	n		IV			2SLS			Optimal IV			MLE		
			mean	sd	RMSE	mean	sd	RMSE	mean	sd	RMSE	mean	sd	RMSE
1	100	λ	0.9440	0.8297	0.8316	0.9837	0.8005	0.8006	0.9396	0.8084	0.8107	0.9185	0.5033	0.5098
		β_1	−0.9777	0.3403	0.3410	−0.9934	0.3288	0.3289	−0.9756	0.3321	0.3330	−0.9681	0.2233	0.2255
		β_2	0.9982	0.1041	0.1041	0.9957	0.1037	0.1038	0.9987	0.1040	0.1040	1.0048	0.0946	0.0947
	200	λ	1.0145	0.5857	0.5859	1.0367	0.5786	0.5798	1.0058	0.5678	0.5678	0.9717	0.3352	0.3364
		β_1	−1.0050	0.2208	0.2208	−1.0129	0.2184	0.2188	−1.0016	0.2154	0.2154	−0.9898	0.1389	0.1392
		β_2	0.9989	0.0890	0.0890	0.9965	0.0885	0.0886	0.9998	0.0879	0.0879	1.0056	0.0733	0.0735
	500	λ	0.9558	0.4547	0.4569	0.9687	0.4477	0.4488	0.9514	0.4375	0.4402	0.9686	0.2524	0.2544
		β_1	−0.9820	0.1705	0.1715	−0.9867	0.1680	0.1685	−0.9804	0.1642	0.1654	−0.9866	0.1020	0.1029
		β_2	1.0009	0.0500	0.0500	1.0002	0.0496	0.0496	1.0012	0.0491	0.0491	1.0013	0.0443	0.0444
	1000	λ	0.9899	0.2563	0.2565	0.9921	0.2540	0.2542	0.9912	0.2470	0.2471	0.9901	0.1576	0.1579
		β_1	−0.9961	0.0962	0.0963	−0.9969	0.0955	0.0955	−0.9966	0.0933	0.0934	−0.9963	0.0647	0.0648
		β_2	0.9983	0.0327	0.0327	0.9982	0.0326	0.0327	0.9983	0.0324	0.0325	0.9988	0.0312	0.0312
1.5	100	λ	1.4519	0.7725	0.7740	1.4862	0.7482	0.7483	1.4446	0.7556	0.7576	1.4258	0.4592	0.4651
		β_1	−0.9777	0.3641	0.3648	−0.9934	0.3532	0.3533	−0.9742	0.3563	0.3573	−0.9666	0.2326	0.2350
		β_2	0.9984	0.1047	0.1047	0.9961	0.1043	0.1043	0.9990	0.1045	0.1045	1.0051	0.0946	0.0947
	200	λ	1.5147	0.5392	0.5394	1.5340	0.5335	0.5346	1.4975	0.5176	0.5176	1.4698	0.3010	0.3025
		β_1	−1.0059	0.2334	0.2335	−1.0138	0.2311	0.2315	−0.9988	0.2253	0.2253	−0.9876	0.1430	0.1436
		β_2	0.9989	0.0891	0.0891	0.9967	0.0886	0.0886	1.0008	0.0877	0.0877	1.0062	0.0729	0.0731
	500	λ	1.4608	0.4113	0.4131	1.4703	0.4052	0.4063	1.4546	0.3994	0.4020	1.4716	0.2272	0.2290
		β_1	−0.9816	0.1785	0.1794	−0.9855	0.1759	0.1765	−0.9789	0.1734	0.1747	−0.9861	0.1059	0.1068
		β_2	1.0010	0.0503	0.0503	1.0004	0.0499	0.0499	1.0014	0.0495	0.0495	1.0014	0.0444	0.0444
	1000	λ	1.4906	0.2355	0.2357	1.4921	0.2322	0.2323	1.4904	0.2257	0.2259	1.4893	0.1433	0.1437
		β_1	−0.9959	0.1021	0.1022	−0.9965	0.1009	0.1009	−0.9958	0.0985	0.0986	−0.9954	0.0675	0.0677
		β_2	0.9984	0.0333	0.0333	0.9983	0.0331	0.0332	0.9984	0.0329	0.0329	0.9989	0.0314	0.0314

$F(x) = \Phi(x)$, $X_{2,n} = (x_{1,n}, \ldots, x_{n,n})' \sim (I_n - 0.2W_n)^{-1}N(0, I_n)$, ϵ_i *iid* $\sim N(0, 1)$, $\beta_0 = (-1, 1)'$.
IV: use $W_n X_{2,n}$ as the IVs of $W_n S_n$. 2SLS: use $W_n X_{2,n}$ and $W_n^2 X_{2,n}$ as the IV of $W_n S_n$. Repetition: 1000.

Table 1, when $\lambda_0 = 1$, we see that the standard errors of the simulated optimal IV estimators are greater than those of the MLE by 56% ∼ 71%.

(4) The RMSE of MLE is obviously less than those of IV/2SLS/optimal IV estimators. The reason is that the standard errors dominate biases. We can see that RMSE ≈ s.d.

We summarize the results when the error terms are not normally distributed in Tables 4 and 5. We can see that the biases of IV/2SLS/optimal IV estimators, but not MLE, decrease when the sample size n increases. This verifies that the normal distribution of the error terms needs to be correctly specified for MLE but needs not be so for the other three estimators.

Table 3
Estimation results when $F(x) = 0.5(x + \sqrt{x^2 + 4})$.

λ_0	n		IV			2SLS			Optimal IV			MLE		
			mean	sd	RMSE	mean	sd	RMSE	mean	sd	RMSE	mean	sd	RMSE
0.4	100	λ	0.3796	0.2418	0.2427	0.3910	0.2341	0.2343	0.3710	0.2294	0.2312	0.3616	0.1519	0.1567
		β_1	−0.9644	0.3828	0.3845	−0.9812	0.3713	0.3718	−0.9518	0.3645	0.3677	−0.9426	0.2668	0.2729
		β_2	0.9984	0.1047	0.1047	0.9959	0.1042	0.1043	1.0007	0.1036	0.1036	1.0074	0.0953	0.0956
	200	λ	0.3995	0.1637	0.1637	0.4044	0.1601	0.1602	0.3904	0.1531	0.1534	0.3862	0.1018	0.1027
		β_1	−0.9994	0.2253	0.2253	−1.0053	0.2209	0.2210	−0.9878	0.2164	0.2167	−0.9851	0.1596	0.1603
		β_2	0.9993	0.0883	0.0883	0.9975	0.0873	0.0874	1.0027	0.0846	0.0846	1.0069	0.0743	0.0747
	500	λ	0.3848	0.1434	0.1442	0.3884	0.1404	0.1409	0.3813	0.1358	0.1371	0.3876	0.0830	0.0839
		β_1	−0.9752	0.2090	0.2105	−0.9803	0.2052	0.2061	−0.9703	0.1995	0.2017	−0.9802	0.1303	0.1318
		β_2	1.0010	0.0498	0.0498	1.0003	0.0493	0.0493	1.0016	0.0486	0.0486	1.0016	0.0447	0.0447
	1000	λ	0.3964	0.0817	0.0818	0.3969	0.0806	0.0806	0.3966	0.0787	0.0787	0.3955	0.0528	0.0530
		β_1	−0.9926	0.1216	0.1219	−0.9933	0.1203	0.1205	−0.9929	0.1182	0.1185	−0.9918	0.0851	0.0855
		β_2	0.9984	0.0328	0.0328	0.9983	0.0327	0.0327	0.9984	0.0325	0.0325	0.9989	0.0313	0.0313
0.7	100	λ	0.6874	0.1374	0.1380	0.6920	0.1329	0.1332	0.6755	0.1322	0.1344	0.6741	0.0856	0.0894
		β_1	−0.9698	0.3187	0.3201	−0.9793	0.3102	0.3109	−0.9449	0.3096	0.3145	−0.9461	0.2294	0.2356
		β_2	0.9992	0.1066	0.1066	0.9974	0.1060	0.1061	1.0047	0.1050	0.1051	1.0099	0.0964	0.0969
	200	λ	0.6994	0.0866	0.0866	0.6998	0.0837	0.0837	0.6917	0.0783	0.0788	0.6904	0.0540	0.0549
		β_1	−0.9997	0.1759	0.1759	−1.0004	0.1714	0.1714	−0.9863	0.1675	0.1681	−0.9864	0.1333	0.1339
		β_2	0.9996	0.0873	0.0873	0.9993	0.0855	0.0855	1.0049	0.0820	0.0821	1.0083	0.0738	0.0743
	500	λ	0.6906	0.0835	0.0841	0.6909	0.0803	0.0808	0.6864	0.0776	0.0788	0.6908	0.0491	0.0500
		β_1	−0.9785	0.1731	0.1744	−0.9791	0.1679	0.1692	−0.9703	0.1632	0.1659	−0.9798	0.1126	0.1145
		β_2	1.0012	0.0500	0.0500	1.0010	0.0490	0.0490	1.0024	0.0482	0.0482	1.0021	0.0449	0.0449
	1000	λ	0.6973	0.0521	0.0522	0.6974	0.0498	0.0499	0.6963	0.0487	0.0489	0.6958	0.0334	0.0337
		β_1	−0.9927	0.1077	0.1080	−0.9928	0.1040	0.1043	−0.9907	0.1025	0.1029	−0.9901	0.0765	0.0771
		β_2	0.9986	0.0341	0.0342	0.9986	0.0337	0.0338	0.9988	0.0334	0.0335	0.9994	0.0319	0.0319

$F(x) = 0.5(x + \sqrt{x^2 + 4}), X_{2,n} = (x_{1,n}, \ldots, x_{n,n})' \sim (I_n - 0.2W_n)^{-1}N(0, I_n), \epsilon_i \ iid \sim N(0, 1), \beta_0 = (-1, 1)'$.
IV: use $W_nX_{2,n}$ as the IVs of W_nS_n. 2SLS: use $W_nX_{2,n}$ and $W_n^2X_{2,n}$ as the IV of W_nS_n. Repetition: 1000.

Table 4
Estimation results without normality (I).

ϵ_n	n		IV			2SLS			Optimal IV			MLE		
			mean	sd	RMSE	mean	sd	RMSE	mean	sd	RMSE	mean	sd	RMSE
MN	100	λ	0.8899	1.3429	1.3474	0.9520	1.3082	1.3090	0.8885	1.3088	1.3135	1.0533	0.7791	0.7809
		β_1	−0.9414	0.6312	0.6339	−0.9696	0.6166	0.6173	−0.9405	0.6170	0.6199	−1.0168	0.3906	0.3909
		β_2	0.9957	0.1052	0.1053	0.9933	0.1044	0.1047	0.9959	0.1047	0.1048	0.9938	0.0927	0.0929
	200	λ	0.9798	1.0044	1.0046	1.0213	0.9954	0.9956	0.9676	0.9915	0.9920	1.1200	0.5496	0.5625
		β_1	−0.9892	0.4353	0.4354	−1.0066	0.4305	0.4305	−0.9843	0.4293	0.4296	−1.0487	0.2525	0.2571
		β_2	1.0001	0.0933	0.0933	0.9975	0.0930	0.0930	1.0008	0.0924	0.0924	0.9931	0.0700	0.0704
	500	λ	0.9771	0.7424	0.7427	0.9981	0.7349	0.7349	0.9720	0.7245	0.7251	1.2491	0.4170	0.4857
		β_1	−0.9905	0.3175	0.3177	−0.9993	0.3146	0.3146	−0.9882	0.3103	0.3105	−1.1047	0.1875	0.2147
		β_2	0.9991	0.0494	0.0494	0.9984	0.0493	0.0493	0.9993	0.0488	0.0488	0.9908	0.0431	0.0441
	1000	λ	0.9757	0.4956	0.4962	0.9907	0.4931	0.4932	0.9700	0.4912	0.4921	1.2656	0.2745	0.3820
		β_1	−0.9884	0.2124	0.2127	−0.9946	0.2115	0.2116	−0.9860	0.2106	0.2111	−1.1093	0.1238	0.1652
		β_2	1.0002	0.0352	0.0352	0.9996	0.0353	0.0353	1.0004	0.0352	0.0352	0.9898	0.0309	0.0325
t(5)	100	λ	0.8338	1.3304	1.3407	0.8887	1.2960	1.3008	0.8301	1.3020	1.3130	0.6718	0.8859	0.9448
		β_1	−0.9243	0.6099	0.6146	−0.9488	0.5944	0.5966	−0.9227	0.5958	0.6008	−0.8530	0.4233	0.4481
		β_2	1.0013	0.1032	0.1032	0.9991	0.1028	0.1028	1.0015	0.1023	0.1023	1.0126	0.0955	0.0964
	200	λ	1.0068	0.9336	0.9336	1.0362	0.9273	0.9280	0.9890	0.9000	0.9001	0.7447	0.6286	0.6785
		β_1	−1.0005	0.3966	0.3966	−1.0126	0.3935	0.3937	−0.9934	0.3825	0.3825	−0.8935	0.2763	0.2961
		β_2	0.9991	0.0902	0.0902	0.9971	0.0903	0.0903	1.0003	0.0887	0.0887	1.0193	0.0798	0.0821
	500	λ	0.9422	0.7286	0.7308	0.9624	0.7195	0.7205	0.9353	0.7058	0.7088	0.7344	0.4885	0.5561
		β_1	−0.9723	0.3122	0.3135	−0.9807	0.3083	0.3089	−0.9693	0.3035	0.3050	−0.8861	0.2123	0.2409
		β_2	0.9996	0.0487	0.0487	0.9989	0.0486	0.0486	0.9999	0.0484	0.0484	1.0084	0.0465	0.0473
	1000	λ	0.9984	0.4910	0.4911	1.0084	0.4884	0.4885	0.9952	0.4755	0.4755	0.7302	0.3187	0.4176
		β_1	−1.0016	0.2066	0.2066	−1.0057	0.2055	0.2055	−1.0003	0.2006	0.2006	−0.8917	0.1401	0.1771
		β_2	1.0003	0.0366	0.0366	0.9999	0.0365	0.0365	1.0004	0.0363	0.0363	1.0114	0.0333	0.0352

MN: mixed normal distribution: half probability $N(6/\sqrt{17}, 9/68)$, half probability $N(-6/\sqrt{17}, 9/68)$.
t(5): $\sqrt{1.35}t(5)$.
$F(x) = 1/(1 + \exp(-x)), X_{2,n} = (x_{1,n}, \ldots, x_{n,n})' \sim 1.5(I_n - 0.2W_n)^{-1}N(0, I_n), \lambda_0 = 1, \beta_0 = (-1, 1)'$.
IV: use $W_nX_{2,n}$ as the IVs of W_nS_n. 2SLS: use $W_nX_{2,n}$ and $W_n^2X_{2,n}$ as the IV of W_nS_n. Repetition: 1000.

5.2. Estimation with misspecified functional forms

Next, we consider consequences when estimating the model with a wrong nonlinear $F(\cdot)$. The results are summarized in Table 6. The model $s_{i,n} = F(\lambda w_{i.,n}S_n + \beta_1 + \beta_2 x_{i,n} + \epsilon_{i,n})$ has true parameters $(\lambda_0, \beta_{10}, \beta_{20}, \sigma_0) = (1, -1, 1, 1)$. We presume that in empirical studies $F(x) = (1 + e^{-x})^{-1}$ and $F(x) = \Phi(x)$ are most frequently used, thus we focus on the estimation with these two functional forms. But the true functional forms can be one of four different distribution functions: logit $F(x) = (1 + e^{-x})^{-1}$, normal $F(x) = \Phi(x)$,

Table 5
Estimation results without normality (II).

ϵ_n	n		IV mean	IV sd	IV RMSE	2SLS mean	2SLS sd	2SLS RMSE	Optimal IV mean	Optimal IV sd	Optimal IV RMSE	MLE mean	MLE sd	MLE RMSE
		λ	0.8725	1.3414	1.3474	0.9438	1.3077	1.3089	0.8676	1.3219	1.3285	0.9950	0.7751	0.7751
	100	β_1	−0.9375	0.6288	0.6319	−0.9698	0.6122	0.6130	−0.9349	0.6219	0.6253	−0.9950	0.3859	0.3860
		β_2	0.9962	0.1036	0.1037	0.9935	0.1033	0.1035	0.9965	0.1030	0.1031	0.9961	0.0930	0.0931
		λ	0.9563	0.9895	0.9904	0.9985	0.9810	0.9810	0.9382	0.9776	0.9796	1.0367	0.5673	0.5684
	200	β_1	−0.9811	0.4298	0.4302	−0.9986	0.4258	0.4258	−0.9738	0.4238	0.4246	−1.0157	0.2609	0.2614
U		β_2	1.0032	0.0916	0.0917	1.0004	0.0914	0.0914	1.0042	0.0906	0.0907	0.9999	0.0710	0.0710
		λ	0.9915	0.7217	0.7218	1.0130	0.7152	0.7153	0.9901	0.7019	0.7020	1.1477	0.4170	0.4424
	500	β_1	−0.9947	0.3102	0.3102	−1.0037	0.3076	0.3076	−0.9940	0.3019	0.3020	−1.0601	0.1891	0.1984
		β_2	0.9994	0.0481	0.0481	0.9987	0.0481	0.0481	0.9995	0.0476	0.0476	0.9950	0.0435	0.0438
		λ	0.9778	0.5006	0.5011	0.9920	0.4991	0.4992	0.9733	0.4919	0.4926	1.1602	0.2943	0.3350
	1000	β_1	−0.9890	0.2127	0.2130	−0.9949	0.2121	0.2121	−0.9871	0.2093	0.2097	−1.0648	0.1301	0.1454
		β_2	1.0000	0.0360	0.0360	0.9994	0.0360	0.0360	1.0002	0.0359	0.0359	0.9935	0.0314	0.0321
		λ	0.8712	1.3419	1.3481	0.9355	1.3121	1.3137	0.8627	1.3392	1.3462	1.0036	0.8063	0.8063
	100	β_1	−0.9275	0.6325	0.6367	−0.9568	0.6181	0.6196	−0.9234	0.6320	0.6366	−0.9904	0.3983	0.3984
		β_2	0.9989	0.1041	0.1041	0.9965	0.1039	0.1039	0.9993	0.1040	0.1040	0.9979	0.0925	0.0926
		λ	0.8904	1.0185	1.0244	0.9257	1.0029	1.0056	0.9200	0.9809	0.9842	1.0881	0.5741	0.5808
	200	β_1	−0.9846	0.4338	0.4341	−0.9991	0.4261	0.4261	−0.9969	0.4181	0.4181	−1.0660	0.2625	0.2706
B		β_2	1.0111	0.0906	0.0912	1.0089	0.0907	0.0911	1.0091	0.0891	0.0896	1.0005	0.0739	0.0740
		λ	0.9809	0.7071	0.7073	1.0063	0.6992	0.6992	0.9761	0.6901	0.6905	1.2218	0.4036	0.4605
	500	β_1	−0.9917	0.3027	0.3028	−1.0023	0.2992	0.2992	−0.9894	0.2963	0.2965	−1.0928	0.1823	0.2045
		β_2	1.0003	0.0495	0.0495	0.9995	0.0494	0.0494	1.0005	0.0493	0.0493	0.9929	0.0449	0.0454
		λ	0.9944	0.5013	0.5013	1.0080	0.4962	0.4962	0.9930	0.4942	0.4943	1.2086	0.2785	0.3479
	1000	β_1	−0.9981	0.2113	0.2113	−1.0038	0.2093	0.2093	−0.9975	0.2087	0.2087	−1.0870	0.1223	0.1501
		β_2	0.9995	0.0342	0.0342	0.9990	0.0341	0.0341	0.9995	0.0340	0.0340	0.9920	0.0297	0.0308

U: $1.5U(-\sqrt{3}, \sqrt{3})$; B: $\sqrt{18}(B(\frac{1}{2}, \frac{1}{2}) - 0.5)$.
$F(x) = 1/(1+\exp(-x))$, $X_{2,n} = (x_{1,n}, \ldots, x_{n,n})' \sim 1.5(I_n - 0.2W_n)^{-1}N(0, I_n)$, $\lambda_0 = 1$, $\beta_0 = (-1, 1)'$.
IV: use $W_nX_{2,n}$ as the IVs of W_nS_n. 2SLS: use $W_nX_{2,n}$ and $W_n^2X_{2,n}$ as the IV of W_nS_n. Repetition: 1000.

Table 6
Compare F is logistic and standard normal distributions.

True F	Logit				Normal			
Estimate with	Logit	Normal	Logit	Normal	Logit	Normal	Logit	Normal
	2SLS		MLE		2SLS		MLE	
λ	0.9536 (0.5818)	0.5391 (0.3305)	0.9546 (0.3339)	0.6167 (0.1936)	2.0612 (1.0499)	0.9619 (0.4654)	0.6985 (0.5493)	0.9675 (0.2553)
β_1	−0.9821 (0.2339)	−0.5603 (0.1327)	−0.9820 (0.1407)	−0.5906 (0.0814)	−2.0640 (0.3971)	−0.9867 (0.1747)	−1.5701 (0.2177)	−0.9882 (0.1040)
β_2	1.0012 (0.0497)	0.5706 (0.0275)	1.0023 (0.0452)	0.5677 (0.0247)	2.1124 (0.1255)	1.0013 (0.0498)	2.1941 (0.1248)	1.0022 (0.0450)
True F	Laplace				Cauchy			
λ	1.2194 (0.6455)	0.6691 (0.3617)	1.5876 (0.3678)	0.9544 (0.2096)	0.7528 (0.5239)	0.4456 (0.3144)	1.0785 (0.3098)	0.6241 (0.1859)
β_1	−1.2668 (0.2476)	−0.7017 (0.1384)	−1.4035 (0.1499)	−0.8078 (0.0850)	−0.7998 (0.2126)	−0.4750 (0.1275)	−0.9298 (0.1305)	−0.5462 (0.0783)
β_2	1.2835 (0.0618)	0.7091 (0.0340)	1.2654 (0.0552)	0.6947 (0.0302)	0.8136 (0.0391)	0.4830 (0.0234)	0.8007 (0.0348)	0.4760 (0.0209)

$X_{2,n} = (x_{1,n}, \ldots, x_{n,n})' \sim (I_n - 0.2W_n)^{-1}N(0, I_n)$, ϵ_i *iid* $\sim N(0, 1)$, $(\lambda_0, \beta_{10}, \beta_{20}) = (1, -1, 1)$.
2SLS: use $W_nX_{2,n}$ and $W_n^2X_{2,n}$ as the IV of W_nS_n. Sample size: 500. Repetition: 1000.

Laplace $F(x) = 1(x < 0)e^x/2 + 1(x \geqslant 0)(1 - e^{-x}/2)$, and Cauchy $F(x) = \frac{1}{2} + \frac{1}{\pi}\arctan x$. All these transformations are nonlinear and thus the estimates of coefficients in Table 6 would be different from the true ones. Instead of comparing estimated coefficients across model specifications with various transformations, it may be more appropriate to compare implied marginal effects.

Figs. 1–4 illustrate differences in the implied marginal effects based on estimated models with those derived from the exact ones (with true coefficients). When the true $F(\cdot)$ is either the logit or normal distribution function, the marginal effects are not far away from each other regardless of whether the specified transformation used is $F(x) = (1+e^{-x})^{-1}$ or $F(x) = \Phi(x)$. But if the true $F(\cdot)$ is the Laplace distribution, then $\Phi(x)$ gives much worse marginal effects than those from the logit $(1+e^{-x})^{-1}$. This can be explained

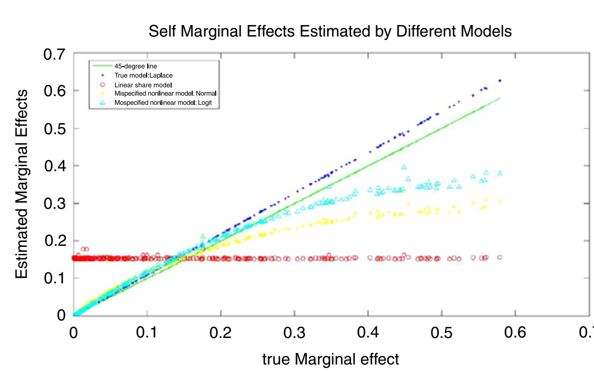

Fig. 3. Self marginal effects when true $F(\cdot)$ is the Laplace distribution function.

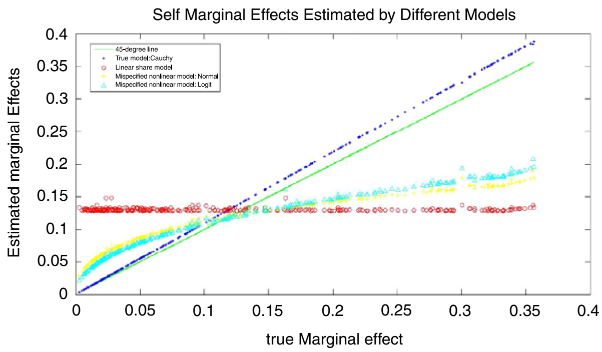

Fig. 4. Self marginal effects when true $F(\cdot)$ is the Cauchy distribution function.

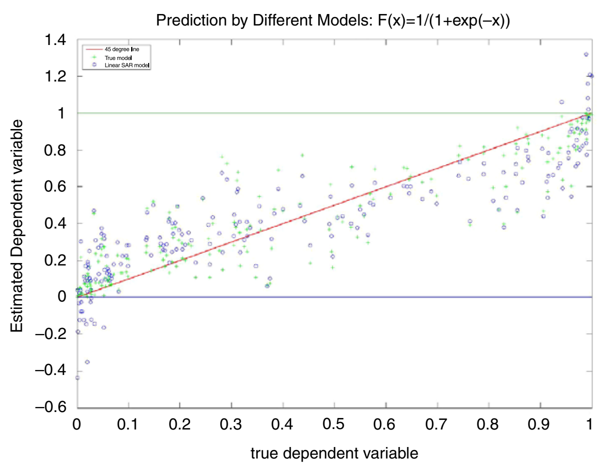

Fig. 5. Prediction when $F(x) = 1/(1 + e^{-x})$.

by the tail behavior of these distributions: the tails of Laplace distribution and logit are similar while $\Phi(x)$ has much thinner tails. When $F(\cdot)$ is the Cauchy distribution, the marginal effects from both $F(x) = (1 + e^{-x})^{-1}$ and $F(x) = \Phi(x)$ are imprecise, even though $F(x) = (1+e^{-x})^{-1}$ gives slightly better estimation. Perhaps that is because Cauchy has much fatter tails than the logit distribution, while the tails of the normal distribution are the thinnest. Figs. 1–4 are generated from MLE. 2SLS estimation, different sample sizes and parameters have been tried in the experiment, and the corresponding figures are very similar.[11]

5.3. Testing functional forms

In this section, we conduct some Monte Carlo experiments on the finite sample performance of testing the specified $F(\cdot)$ transformation as suggested by the end of Section 4. In the experiment, $s_{i,n} = F(\lambda w_{i.,n}S_n + \beta_1 + \beta_2 x_{2i,n} + \beta_3 x_{3i,n} + \epsilon_{i,n})$ has true parameters $(\lambda_0, \beta_{10}, \beta_{20}, \beta_{30}) = (1, -1, 0.5, 0.5)$ and $\epsilon_{i,n}$ is i.i.d. $N(0, 0.7^2)$. The designs of the regressors $x_{2i,n}$ and $x_{3i,n}$ are described under Table 7. Both the sample size and the number of bootstrapping repetitions are 500, and the Monte Carlo repetition is 1000. The true parameters are chosen such that there are few computational problems such as ill-conditioned matrices. We obtain the critical value (one-sided test) of the 5% level of significance. From Table 7, we see that the frequencies of Type I errors are between 5.2% and 6.8%, which are close to the 5% errors. However, from Table 8, the powers for most tests are not large. This is especially true for distributions that have certain similarities, e.g., the power of testing

[11] Those additional results are presented in the supplement file (see Appendix C).

Table 7
Size of test between different transformations.

		H_1				
		Cauchy	Laplace	Logit	Normal	Extreme
H_0	Logit	6.3%	6%	–	5.2%	5.6%
	Normal	5.2%	5.9%	6.8%	–	5.3%

$X_{2,n} = (x_{21,n}, \ldots, x_{2n,n})' \sim 1.5(I_n - 0.2W_n)^{-1}N(0, I_n), X_{3,n} = (x_{31,n}, \ldots, x_{3n,n})' \sim N(0, I_n)$,
ϵ_i *iid* $\sim N(0, 0.7^2)$, $(\lambda_0, \beta_{10}, \beta_{20}, \beta_{30}) = (1, -1, 0.5, 0.5)$.
Use $W_nX_{2,n}, W_nX_{3,n}, W_n^2X_{2,n}$ and $W_n^2X_{3,n}$ as IV.
Bootstrap: 500 times. Sample size: 500. Repetition: 1000.

Table 8
Power of test between different transformations.

		H_1				
		Cauchy	Laplace	Logit	Normal	Extreme
H_0	Logit	48.5%	12.1%	–	5.5%	63.3%
	Normal	57.6%	23.8%	9.7%	–	41.6%

$X_{2,n} = (x_{21,n}, \ldots, x_{2n,n})' \sim 1.5(I_n - 0.2W_n)^{-1}N(0, I_n), X_{3,n} = (x_{31,n}, \ldots, x_{3n,n})' \sim N(0, I_n)$,
ϵ_i *iid* $\sim N(0, 0.7^2)$, $(\lambda_0, \beta_{10}, \beta_{20}, \beta_{30}) = (1, -1, 0.5, 0.5)$.
Use $W_nX_{2,n}, W_nX_{3,n}, W_n^2X_{2,n}$ and $W_n^2X_{3,n}$ as IV.
Bootstrap: 500 times. Sample size: 500. Repetition: 1000.

logit $F(\cdot)$ vs normal $F(\cdot)$ is 5.5% and the power of testing normal vs logit is 11.7%. Laplace $F(\cdot)$ and logit $F(\cdot)$ also have similar behaviors, and the power of testing logit vs Laplace transformation is 12.1%. If two distributions are quite different, then the powers are large. For example, Cauchy $F(\cdot)$ has fat tails but normal $F(\cdot)$ has thin tails, and the power of testing logit vs the Cauchy transformation is 57.6%. We have also examined the relationship between powers and variances of $x_{i,n}\beta_0 + \epsilon_{i,n}$. From Table 9, we see that as we raise the variance of $x_{i,n}\beta_0 + \epsilon_{i,n}$, powers increase for all test except testing logit against Laplace. These phenomena can be explained by the tail behaviors of these distributions: the tails of logit and Laplace distributions are the same, except the scaling factor 2; but the tails of other pairs are of different thickness. When the variance of $x_{i,n}\beta_0 + \epsilon_{i,n}$ increases, more data are located at the tails of these distributions, then it will be easier to differentiate two $F(\cdot)$'s if their tails are more different and it is harder to differentiate them if their tails are similar.

6. Conclusion

In this paper, we consider a generalization of the linear SAR model to a nonlinear one with a strictly increasing nonlinear transformation function. After establishing the NED property of the dependent variable and relevant functions, we show the consistency and asymptotic normality of the ML estimators with normally distributed errors. To consider the case where the distribution of errors is unknown, we also consider IV and 2SLS estimation. Monte Carlo experiments verify our theoretical results in finite samples. The experiments also show that MLE is more efficient relative to the 2SLS estimation.

Our models can be extended in several ways. First, we have not considered heteroskedasticity in our model. As the MLE is generally not consistent for the estimation of the linear SAR model with unknown heteroskedasticity (see Lin and Lee, 2010), we expect that the MLE for a nonlinear SAR would also be inconsistent, if unknown heteroskedasticity were ignored. Thus, it would be of interest to study the nonlinear SAR model with heteroskedasticity. Second, it would also be interesting to generalize our model to panel data. Many results have been obtained for the estimation of linear spatial panel data models (see, e.g. Lee and Yu, 2010), but the research on nonlinear spatial panel models needs to be developed. Third, our model depends crucially on the Lipschitz property of $F(\cdot)$, which gives NED property of the dependent variable and other

Table 9
Power of test.

β_{20}	β_{30}	σ_0	H_0	H_1	H_0	H_1	H_0	H_1	H_0	H_1	H_0	H_1
			Normal	Logit	Normal	Laplace	Logit	Cauchy	Normal	Cauchy	Logit	Laplace
0.5	0.5	0.7	9.7%		23.8%		48.5%		57.6%		12.1%	
0.7	0.7	0.9	11.7%		28.1%		56.1%		64.9%		12%	
1	1	1	16.7%		33.2%		66.9%		74.5%		11.4%	
1.5	1.5	1.5	21.5%		38.2%		75.3%		81.6%		10.4%	

The first row of values duplicates some results in Table 8 for easier comparison.
$X_{2,n} = (x_{21,n}, \dots, x_{2n,n})' \sim 1.5(I_n - 0.2W_n)^{-1}N(0, I_n)$, $X_{3,n} = (x_{31,n}, \dots, x_{3n,n})' \sim N(0, I_n)$,
ϵ_i *iid* $\sim N(0, \sigma_0^2)$, $(\lambda_0, \beta_{10}) = (1, -1)$.
Use $W_nX_{2,n}$, $W_nX_{3,n}$, $W_n^2X_{2,n}$ and $W_n^2X_{3,n}$ as IV.
Bootstrap: 500 times. Sample size: 500. Repetition: 1000.

variables. However, some nonlinear transformation functions in certain models, such as step functions for binary choice models, do not satisfy the Lipschitz property. More work needs to be done in this area. Finally, in empirical applications, we may not know the functional form of $F(\cdot)$. Thus it would be useful to generalize the model to a semiparametric one.

Acknowledgments

We are grateful to two referees, Professor Nazgul Jenish, and the Co-editor, Professor Peter Robinson, for their comments and suggestions to improve this paper.

Appendix A. Some useful lemmas

Lemma A.1 (*A Direct Generalization of Corollary 4.3(b), Gallant and White, 1988*). *If for all i and n, $\|Y_{i,n}\|_{2r} \leqslant \Delta < \infty$ and $\|Z_{i,n}\|_{2r} \leqslant \Delta < \infty$ for some $r > 2$, $\|Y_{i,n} - \mathrm{E}[Y_{i,n}|\mathcal{F}_{i,n}(s)]\|_2 \leqslant d_{i,Yn}\rho^s$ and $\|Z_{i,n} - \mathrm{E}[Z_{i,n}|\mathcal{F}_{i,n}(s)]\|_2 \leqslant d_{i,Zn}\rho^s$, then $\|Y_{i,n}Z_{i,n} - \mathrm{E}[Y_{i,n}Z_{i,n}|\mathcal{F}_{i,n}(s)]\|_2 \leqslant d_{i,n}\tilde{\rho}^s$ where $d_{i,n} = 2^{(3r-2)/(r-1)}(d_{i,Zn} + d_{i,Yn})^{(r-2)/(2r-2)}\Delta^{(3r-2)/(2r-2)}$ and $\tilde{\rho} = \rho^{(r-2)/(2r-2)}$. Specifically, if $\{Y_{i,n}\}$ and $\{Z_{i,n}\}$ are both uniformly L_{2r} bounded and uniformly and geometrically L_2-NED, then $\{Y_{i,n}Z_{i,n}\}$ is still uniformly and geometrically L_2-NED.*

The proof of the above lemma is almost the same as that of Corollary 4.3(b) in Gallant and White (1988), thus we omit it here.

Lemma A.2. *If $\{X_{i,1n}\}, \dots, \{X_{i,kn}\}$ are kL_p-NED random fields on $\{\epsilon_{i,n}\}_{i=1}^n$, for each i, define $Z_{i,n}$ arbitrarily as one among $\{X_{i,1n}, \dots, X_{i,kn}\}$, then $\{Z_{i,n}\}_{i=1}^n$ is also L_p-NED.*

Proof. Because $\|X_{i,jn} - \mathrm{E}(X_{i,jn}|\mathcal{F}_{i,n}(m))\|_p \leqslant d_{i,jn}\psi_j(m)$, we have $\|Z_{i,n} - \mathrm{E}(Z_{i,n}|\mathcal{F}_{i,n}(m))\|_p \leqslant \max_j d_{i,jn} \max_j \psi_j(m)$. □

Lemma A.3. *If $\{X_{i,1n}\}, \dots, \{X_{i,kn}\}$ are $k\,L_2$-NED random fields on $\{\epsilon_{i,n}\}$ such that $\|X_{i,jn} - \mathrm{E}[X_{i,jn}|\mathcal{F}_{i,n}(m)]\| \leqslant d_{i,jn}\psi(m)$, then $\{Z_{i,n} \equiv \sqrt{X_{i,1n}^2 + \cdots + X_{i,kn}^2}\}$ is L_2-NED such that $\|Z_{i,n} - \mathrm{E}[Z_{i,n}|\mathcal{F}_{i,n}(m)]\| \leqslant (\sum_j d_{i,jn})\psi(m)$. If $\{X_{i,1n}\}, \dots, \{X_{i,kn}\}$ are k uniformly and geometrically L_2-NED random fields, then $\{Z_{i,n}\}$ is also a uniformly and geometrically L_2-NED random field.*

Proof. The Euclidean distance function $\|(x_1, \dots, x_k)\| = \sqrt{x_1^2 + \cdots + x_k^2}$ is a Lipschitz function because: $\left|\frac{\partial\|(x_1,\dots,x_k)\|}{\partial x_i}\right| = \left|\frac{x_i}{\sqrt{x_1^2+\cdots+x_K^2}}\right| \leqslant 1$. Then the conclusion comes from Theorem 17.12 in Davidson (1994). □

Appendix B. Proofs

B.1. The proof for Section 2

Proof of Proposition 1. Denote the right hand side of Eq. (2) as $H_n(S_n)$. First we will show that $H_n : \mathbb{R}^n \to \mathbb{R}^n$ is a contraction mapping. Because

$$\frac{\partial H_n(S_n)}{\partial S'} = \begin{pmatrix} \lambda f(\lambda w_{1\cdot,n}S_n + x_{1,n}\beta + \epsilon_1)w_{1\cdot,n} \\ \vdots \\ \lambda f(\lambda w_{n\cdot,n}S_n + x_{n,n}\beta + \epsilon_n)w_{n\cdot,n} \end{pmatrix},$$

it follows that $\|\partial H_n(S)/\partial S'\|_\infty \leqslant |\lambda| \sup_{i=1,\dots,n} f(\lambda w_{i\cdot,n}S_n + x_{i,n}\beta + \epsilon_{i,n})\|W_n\|_\infty \leqslant |\lambda| b_f \|W_n\|_\infty \leqslant \zeta < 1$, where $\|\cdot\|_\infty$ represents the infinite vector norm. By the mean value theorem, we have $H_{j,n}(S_1) - H_{j,n}(S_2) = \frac{\partial H_{j,n}(\bar{S}_j)}{\partial S'}(S_1 - S_2)$ where $\bar{S}_j$ lies between S_1 and S_2. Therefore, $\|H_n(S_1) - H_n(S_2)\|_\infty \leqslant \zeta\|S_1 - S_2\|_\infty$, i.e., H_n is a contraction mapping. Since $\mathbb{R}^n$ is a complete metric space, there is exactly one fixed point for the contraction mapping H_n. □

Proof of Lemma 1. First, we consider the solution of the system of equations: $S_n^0 = F(\lambda_0 W_n S_n^0)$. By mean value theorem, $S_n^0 = F(0)\iota_n + \lambda_0 \overline{f_{D_n}} W_n S_n^0$, where $\iota_n = (1, \dots, 1)'$ and $\overline{f_{D_n}}$ is a diagonal matrix with its jth diagonal element $f(\overline{t_j})$ for some $\overline{t_j}$ between 0 and $\lambda_0 w_{j\cdot,n}S_n^0$. Then $S_n^0 = F(0)(I_n - \lambda_0\overline{f_{D_n}}W_n)^{-1}\iota_n$. Because $\|(I_n - \lambda_0 b_f W_n)^{-1}\|_\infty = \|\sum_{i=0}^\infty(\lambda_0 b_f W_n)^i\|_\infty \leqslant 1/(1-\zeta)$, we have $|s_{i,n}^0| \leqslant |F(0)|/(1-\zeta)$.

Second, we consider the equation $S_n = F(\lambda_0 W_n S_n + \eta_n)$. Then $dS_n = (I_n - \lambda_0 f_{D_n} W_n)^{-1} f_{D_n} d\eta_n$. As elements of W_n and f_{D_n} are non-negative, $(I_n - \lambda_0\overline{f_{D_n}}W_n)^{-1}f_{D_n} \leqslant^* M_n = (m_{ij,n}) \equiv b_f(I_n - |\lambda_0| b_f W_n)^{-1}$, where $\leqslant^*$ means the inequality applied to the absolute value of pointwise entries of the two matrices. Thus S_n is a Lipschitz function of η_n. Apply this conclusion to $S_n = F(\lambda_0 W_n S_n + X_n\beta_0 + \epsilon_n)$ and denote its solution as $S_n(\epsilon_n)$. Then $S_n(0)$ is the solution of $S_n = F(\lambda_0 W_n S_n + X_n\beta_0)$. Because $\|M_n\|_\infty \leqslant b_f/(1-\zeta)$, we have $|s_{i,n}(0)| \leqslant |s_{i,n}^0| + \|M_n\|_\infty |x_{i,n}\beta_0| \leqslant (|F(0)| + b_f|x_{i,n}\beta_0|)/(1-\zeta)$.

Third, $\prod_{j=1}^n \mathrm{E}|\epsilon_{j,n}|^{l_j} = \prod_{j=1}^n \|\epsilon_{j,n}\|_{l_j}^{l_j} \leqslant \prod_{j=1}^n \|\epsilon_{j,n}\|_{l_1+\cdots+l_n}^{l_j} = \mathrm{E}|\epsilon_{j,n}|^{l_1+\cdots+l_n}$ by Lyapunov's inequality. Then, with the multinomial theorem (Sheldon, 2009), we have

$$\begin{aligned}
&\mathrm{E}\left[\left(\sum_{j=1}^n |m_{ij,n}\epsilon_{j,n}|\right)^p\right] \\
&= \mathrm{E}\sum_{l_1+\cdots+l_n=p} \frac{p!}{l_1!\cdots l_n!}\prod_{j=1}^n |m_{ij,n}\epsilon_{j,n}|^{l_j} \\
&= \sum_{l_1+\cdots+l_n=p} \frac{p!}{l_1!\cdots l_n!}\prod_{j=1}^n |m_{ij,n}|^{l_j}\mathrm{E}|\epsilon_{j,n}|^{l_j} \\
&\leqslant \sum_{l_1+\cdots+l_n=p} \frac{p!}{l_1!\cdots l_n!}\prod_{j=1}^n |m_{ij,n}|^{l_j}\mathrm{E}|\epsilon_{1,n}|^{l_1+\cdots+l_n} \\
&= \mathrm{E}|\epsilon_{1,n}|^p\left(\sum_{j=1}^n |m_{ij,n}|\right)^p \leqslant b_f^p\mathrm{E}|\epsilon_{1,n}|^p/(1-\zeta)^p.
\end{aligned}$$

Finally, because $|s_{i,n}(\epsilon_n) - s_{i,n}(0)| \leqslant \sum_{j=1}^n |m_{ij,n}\epsilon_{j,n}|$, it follows by the C_r-inequality (Shorack, 2000, p. 47) that

$$\begin{aligned}
&\mathrm{E}[|s_{i,n}(\epsilon_n)|^p|X_n] \\
&\leqslant \mathrm{E}\left[\left(|s_{i,n}(0)| + \sum_{j=1}^n |m_{ij,n}\epsilon_{j,n}|\right)^p |X_n\right] \\
&\leqslant 2^{p-1}\left[\mathrm{E}(|s_{i,n}(0)|^p|X_n) + \mathrm{E}\left(\sum_{j=1}^n |m_{ij,n}\epsilon_{j,n}|\right)^p\right] \\
&\leqslant 2^{p-1}[(|F(0)| + b_f|x_{i,n}\beta_0|)^p + b_f^p\mathrm{E}|\epsilon_{1,n}|^p]/(1-\zeta)^p.
\end{aligned}$$

Then it is clear that $\sup_{i,n}\mathrm{E}[|s_{i,n}(\epsilon_n)|^p] = \sup_{i,n} E\{\mathrm{E}[|s_{i,n}(\epsilon_n)|^p|X_n]\} < \infty$ since $\sup_{i,k,n}\|x_{ik,n}\|_p < \infty$. □

Proof of Proposition 2. Denote $S_n^{(1)} = F(\lambda_0 W_n S_n^{(1)} + X_n^{(1)}\beta_0 + \epsilon_n^{(1)})$ and $S_n^{(2)} = F(\lambda_0 W_n S_n^{(2)} + X_n^{(1)}\beta_0 + \epsilon_n^{(2)})$. From the proof of Lemma 1, we have $|s_{i,n}^{(1)} - s_{i,n}^{(2)}| \leqslant \sum_{j=1}^n m_{ij,n}|(x_{j,n}^{(1)} - x_{j,n}^{(2)})\beta_0 + (\epsilon_{j,n}^{(1)} - \epsilon_{j,n}^{(2)})|$, where $(m_{ij,n}) \equiv b_f(I_n - |\lambda|b_f W_n)^{-1}$. Then, $\|s_{i,n} - \mathrm{E}(s_{i,n}|\mathcal{F}_{i,n}(md_0))\|_2 \leqslant \|s_{i,n} - \mathrm{E}(s_{i,n}|x_{j,n}\beta_0 + \epsilon_j, d(j,i) \leqslant md_0)\|_2 \leqslant (\sigma_0 + \|\beta_0\|_1 \sup_{i,k,n}\|x_{ik,n}\|_2)\sum_{j:d(j,i)>md_0} m_{ij,n}$, where the second inequality comes from Proposition 1 in Jenish and Prucha (2012) and Minkowski's inequality. Under Assumption 2, we know $(W_n^l)_{ij} = 0$ if $d(i,j) > md_0$ while $l \leqslant m$. Hence, the conclusion follows from

$$\begin{aligned}
\sum_{j:d(j,i)>md_0} m_{ij,n} &= b_f \sum_{j:d(j,i)>md_0} (I_n - |\lambda|b_f W_n)_{ij}^{-1} \\
&= b_f \sum_{j:d(j,i)>md_0}\sum_{l=0}^{\infty}(|\lambda|b_f W_n)_{ij}^l \\
&= b_f \sum_{j:d(j,i)>md_0}\sum_{l=m+1}^{\infty}(|\lambda|b_f W_n)_{ij}^l \\
&= b_f \sum_{l=m+1}^{\infty}\sum_{j:d(j,i)>md_0}(|\lambda|b_f W_n)_{ij}^l \\
&\leqslant b_f \sum_{l=m+1}^{\infty}\|\lambda b_f W_n\|_\infty^l \leqslant b_f\zeta^{m+1}/(1-\zeta). \quad \square
\end{aligned}$$

Proof of Corollary 1. Because $s_{i,n}$ is uniformly L_p bounded and W_n is uniformly bounded in row sums, $\{w_{i\cdot,n}S_n\}_{i=1}^n$ is uniformly L_p bounded. Notice $w_{ij,n} \neq 0$ only if $d(i,j) \leqslant d_0$. Then

$$\begin{aligned}
&\|w_{i\cdot,n}S_n - \mathrm{E}(w_{i\cdot,n}S_n|\mathcal{F}_{i,n}(md_0))\|_2 \\
&= \left\|\sum_{j=1}^n w_{ij,n}[s_{j,n} - \mathrm{E}(s_{j,n}|F_{i,n}(md_0))]\right\|_2 \\
&\leqslant \sum_{j=1}^n w_{ij,n}\|s_{j,n} - \mathrm{E}[s_{j,n}|\mathcal{F}_{j,n}((m-1)d_0)]\|_2 \\
&\leqslant (\sigma_0 + \|\beta_0\|_1 \sup_{i,k,n}\|x_{ik,n}\|_2)\frac{\sigma_0\zeta^{m+1}}{\lambda_m(1-\zeta)},
\end{aligned}$$

where the second inequality comes from Proposition 2. □

B.2. Proofs for Section 3

Proof of Lemma 2. We know that $\ln x \leqslant x - 1$ for any $x \geqslant 0$, which means $\ln\sqrt{x} \leqslant \sqrt{x} - 1$. Therefore, $\ln x \leqslant 2(\sqrt{x} - 1)$ for any $x \geqslant 0$. So we have

$$\begin{aligned}
\mathrm{E}\ln[L_n(\theta)/L_n(\theta_0)] &\leqslant 2\mathrm{E}\left(\sqrt{L_n(\theta)/L_n(\theta_0)} - 1\right) \\
&= 2\int\left(\sqrt{L_n(\theta)/L_n(\theta_0)} - 1\right)L_n(\theta_0)dS_n \\
&= 2\left(\int\sqrt{L_n(\theta)L_n(\theta_0)}dS_n - 1\right) \\
&= -\int\left[\sqrt{L_n(\theta)} - \sqrt{L_n(\theta_0)}\right]^2 dS_n \leqslant 0. \qquad (15)
\end{aligned}$$

This implies in particular the information inequality that $\mathrm{E}\ln L_n(\theta) \leqslant \mathrm{E}\ln L_n(\theta_0)$ for all θ. Thus θ_0 is a maximizer. Eq. (15) also implies that if $\mathrm{E}\ln L_n(\theta) = \mathrm{E}\ln L_n(\theta_0)$, $L_n(\theta) = L_n(\theta_0)$ almost surely (see, e.g., Van der Vaart, 1998) We claim that θ_0 is the unique maximizer as follows. Because $\mathrm{E}\ln L_n(\theta) = \mathrm{E}\ln L_n(\theta_0)$ implies $L_n(\theta) = L_n(\theta_0)$ almost surely, we analyze the equation $\ln L_n(\theta) - \ln L_n(\theta_0) = 0$ with variable T_n while X_n and parameters are fixed. For any square matrix A, denote $\rho(A)$ the spectral radius of A. From spectral radius theorem, we have $\rho(W_n'W_n) \leqslant \|W_n'W_n\|_\infty \leqslant \|W_n\|_\infty\|W_n'\|_\infty \leqslant C^2$ for some $C > 0$. Thus $C^2 I_n - W_n'W_n$ is positive semi-definite. Hence by Cauchy's inequality,

$$\begin{aligned}
&\lim_{\inf_i |t_{i,n}|\to\infty}\frac{|T_n'W_nF(T_n)|}{T_n'T_n} \\
&\leqslant \lim_{\inf_i |t_{i,n}|\to\infty}\frac{(T_n'T_n)^{1/2}[F(T_n)'W_n'W_nF(T_n)]^{1/2}}{T_n'T_n} \\
&\leqslant C\lim_{\inf_i |t_{i,n}|\to\infty}(T_n'T_n)^{-1/2}[F(T_n)'F(T_n)]^{1/2} \\
&= C\lim_{\inf_i |t_{i,n}|\to\infty}\left[\sum_{i=1}^n F^2(t_{i,n})/\sum_{i=1}^n t_{i,n}^2\right]^{1/2} \\
&\leqslant C\lim_{\inf_i |t_{i,n}|\to\infty}[\max_{i=1,\ldots,n} F^2(t_{i,n})/t_{i,n}^2]^{1/2} = 0,
\end{aligned}$$

where the last equation follows from $\lim_{x\to+\infty}F(x)/x = 0$. Applying Cauchy's inequality again, we have $\lim\sup_{T_n'T_n\to\infty}|T_n'X_n\beta_0|/(T_n'T_n) \leqslant \lim_{T_n'T_n\to\infty}(T_n'T_n)^{-1/2}(\beta_0 X_n'X_n\beta_0)^{1/2} = 0$. For any $\lambda \in \Lambda$, $\rho(\lambda f_{D_n}W_n) \leqslant \|\lambda f_{D_n}W_n\|_\infty = \zeta$. Denote the characteristic values of $f_{D_n}W_n$ as λ_i's. Because $\lambda_i \in \mathbb{R}$, we obtain $1 - \zeta \leqslant 1 - \lambda\lambda_i \leqslant 1 + \zeta$ and $\ln|I_n - \lambda f_{D_n}W_n| = \ln\prod_{i=1}^n(1 - \lambda\lambda_i) \in [n\ln(1-\zeta), n\ln(1+\zeta)]$. By $\lim_{\inf_i |t_{i,n}|\to\infty} T_n'W_nF(T_n)/(T_n'T_n) = 0$, $\lim_{T_n'T_n\to\infty}|T_n'X_n\beta_0|/(T_n'T_n) = 0$ and $\ln|I_n - \lambda f_{D_n}W_n| \in [n\ln(1-\zeta), n\ln(1+\zeta)]$, we have $\sigma_0 = \sigma$ because

$$\begin{aligned}
0 &= \lim_{\inf_i |t_{i,n}|\to\infty}[\ln L_n(\theta) - \ln L_n(\theta_0)]/(T_n'T_n) \\
&= \lim_{\inf_i |t_{i,n}|\to\infty} \\
&\quad\times\left\{-\frac{[T_n - \lambda W_nF(T_n) - X_n\beta]'[T_n - \lambda W_nF(T_n) - X_n\beta]}{2\sigma^2 T_n'T_n}\right. \\
&\quad + \frac{[T_n - \lambda_0 W_nF(T_n) - X_n\beta_0]'[T_n - \lambda_0 W_nF(T_n) - X_n\beta_0]}{2\sigma_0^2 T_n'T_n} \\
&\quad\left. + \frac{\ln|I_n - \lambda f_{D_n}W_n| - \ln|I_n - \lambda_0 f_{D_n}W_n|}{T_n'T_n}\right\} \\
&= (\sigma_0^{-2} - \sigma^{-2})/2.
\end{aligned}$$

Because $F^{-1}(S_n) - \lambda W_nS_n - X_n\beta = \epsilon_n + (\lambda_0 - \lambda)W_nS_n + X_n(\beta_0 - \beta) = \epsilon_n + (\lambda_0 - \lambda)W_n(S_n - \mathrm{E}S_n) + [(\lambda_0 - \lambda)W_n\mathrm{E}S_n + X_n(\beta_0 - \beta)]$, we have

$$\begin{aligned}
&\mathrm{E}[(F^{-1}(S_n) - \lambda W_nS_n - X_n\beta)' \\
&\quad\times(F^{-1}(S_n) - \lambda W_nS_n - X_n\beta)] \\
&= n\sigma_0^2 + (\lambda_0 - \lambda)^2\mathrm{E}[(W_n(S_n - \mathrm{E}S_n))'
\end{aligned}$$

$$
\begin{aligned}
&\times (W_n(S_n - \mathrm{E}S_n))] + 2(\lambda_0 - \lambda)\mathrm{E}(\epsilon' W_n S_n) \\
&+ \mathrm{E}\{[(\lambda_0 - \lambda)W_n \mathrm{E}S_n + X_n(\beta_0 - \beta)]' \\
&\times [(\lambda_0 - \lambda)W_n \mathrm{E}S_n + X_n(\beta_0 - \beta)]\} \\
= {}& (\lambda_0 - \lambda)^2 \mathrm{E}[(W_n(S_n - \mathrm{E}S_n))'(W_n(S_n - \mathrm{E}S_n))] \\
&+ 2\sigma_0^2 \mathrm{E}\mathrm{tr}[(\lambda_0 - \lambda)(f_{D_n} - \lambda_0 W_n)^{-1} W_n] \\
&+ n\sigma_0^2 + \mathrm{E}\{[(\lambda_0 - \lambda)W_n \mathrm{E}S_n + X_n(\beta_0 - \beta)]' \\
&\times [(\lambda_0 - \lambda)W_n \mathrm{E}S_n + X_n(\beta_0 - \beta)]\},
\end{aligned}
$$

where the last step is from the first order condition $\mathrm{E}(\epsilon_n' W_n S_n) = \sigma_0^2 \mathrm{E}[(f_{D_n}^{-1} - \lambda_0 W_n)^{-1} W_n]$. Because $(f_{D_n}^{-1} - \lambda_0 W_n)^{-1}(f_{D_n}^{-1} - \lambda W_n) = I_n + (\lambda_0 - \lambda)(f_{D_n}^{-1} - \lambda_0 W_n)^{-1} W_n$, we have

$$
\begin{aligned}
\mathrm{E}\ln L_n(\theta) - \mathrm{E}\ln L_n(\theta_0) = {}& \left(\frac{n}{2}\ln\frac{\sigma_0^2}{\sigma^2} - \frac{n\sigma_0^2}{2\sigma^2} + \frac{n}{2}\right) \\
&+ \mathrm{E}\ln|I_n + (\lambda_0 - \lambda)(f_{Dn}^{-1} - \lambda_0 W_n)^{-1} W_n| \\
&- \frac{\sigma_0^2}{\sigma^2}\mathrm{E}\mathrm{tr}[(\lambda_0 - \lambda)(f_{D_n}^{-1} - \lambda_0 W_n)^{-1} W_n] \\
&- \frac{(\lambda_0 - \lambda)^2}{2\sigma^2}\mathrm{E}[(W_n(S_n - ES_n))'(W_n(S_n - ES_n))] \\
&- \frac{1}{2\sigma^2}\mathrm{E}\{[(\lambda_0 - \lambda)W_n \mathrm{E}S_n + X_n(\beta_0 - \beta)]' \\
&\times [(\lambda_0 - \lambda)W_n \mathrm{E}S_n + X_n(\beta_0 - \beta)]\} \\
= {}& \frac{n}{2}\left(\frac{\sigma_0^2}{\sigma^2} - \ln\frac{\sigma_0^2}{\sigma^2} - 1\right) \\
&+ \mathrm{E}\left(\ln\left|\frac{\sigma_0^2}{\sigma^2}\left[I_n + (\lambda_0 - \lambda)(f_{D_n}^{-1} - \lambda_0 W_n)^{-1} W_n\right]\right|\right. \\
&\left. - \mathrm{tr}\frac{\sigma_0^2}{\sigma^2}[I_n + (\lambda_0 - \lambda)(f_{D_n}^{-1} - \lambda_0 W_n)^{-1} W_n] + n\right) \\
&- \frac{(\lambda_0 - \lambda)^2}{2\sigma^2}\mathrm{E}[(W_n(S_n - \mathrm{E}S_n))'(W_n(S_n - \mathrm{E}S_n))] \\
&- \frac{1}{2\sigma^2}\mathrm{E}\{[(\lambda_0 - \lambda)W_n \mathrm{E}S_n + X_n(\beta_0 - \beta)]' \\
&\times [(\lambda_0 - \lambda)W_n \mathrm{E}S_n + X_n(\beta_0 - \beta)]\}. \qquad (16)
\end{aligned}
$$

Because $I_n + (\lambda_0 - \lambda)(f_{D_n}^{-1} - \lambda_0 W_n)^{-1} W_n = (I_n - \lambda_0 f_{D_n} W_n)^{-1}(I_n - \lambda f_{D_n} W_n)$, the characteristic values of $I_n + (\lambda_0 - \lambda)(f_{D_n}^{-1} - \lambda_0 W_n)^{-1} W_n$ is $\frac{1-\lambda\lambda_i}{1-\lambda_0\lambda_i}$. For any λ, $|\lambda\lambda_i| \leqslant |\lambda|b_f\|W_n\|_\infty < 1$. Therefore, $1 - |\lambda|b_f\|W_n\|_\infty \leqslant 1 - \lambda\lambda_i \leqslant 1 + |\lambda|b_f\|W_n\|_\infty$. Thus, the ratio $\frac{1-\lambda\lambda_i}{1-\lambda_0\lambda_i}$ is bounded from above and bounded away from zero as

$$
\begin{aligned}
0 &< \frac{1 - |\lambda|b_f\|W_n\|_\infty}{1 + |\lambda_0|b_f\|W_n\|_\infty} \leqslant \frac{1 - \lambda\lambda_i}{1 - \lambda_0\lambda_i} \\
&\leqslant \frac{1 + |\lambda|b_f\|W_n\|_\infty}{1 - |\lambda_0|b_f\|W_n\|_\infty} < \infty. \qquad (17)
\end{aligned}
$$

When the characteristic values of an $n \times n$ matrix A are all positive, then $\ln|A| \leqslant \mathrm{tr}(A) - n$ with equality only when all characteristic values are 1. As all the characteristic values of $f_{D_n} W_n$ are real, Eq. (17) implies that all characteristic values of $\frac{\sigma_0^2}{\sigma^2}[I_n + (\lambda_0 - \lambda)(f_{D_n}^{-1} - \lambda_0 W_n)^{-1} W_n] = \frac{\sigma_0^2}{\sigma^2}(f_{D_n}^{-1} - \lambda_0 W_n)^{-1}(f_{D_n}^{-1} - \lambda W_n)$ are positive. Then as $\sigma^2 = \sigma_0^2$, $\mathrm{E}\ln L_n(\theta) = \mathrm{E}\ln L_n(\theta_0)$ must imply that $\lambda = \lambda_0$ and $\beta = \beta_0$. □

Proof of Lemma 3. From the proof of Lemma 2, we know that if $\mathrm{E}\ln L_n(\theta) = \mathrm{E}\ln L_n(\theta_0)$, we have $L_n(\theta) = L_n(\theta_0)$ almost surely, i.e.,

$$
\begin{aligned}
&-\frac{n}{2}\ln\sigma^2 \\
&\quad - \frac{[T_n - \lambda W_n F(T_n) - X_n\beta]'[T_n - \lambda W_n F(T_n) - X_n\beta]}{2\sigma^2} \\
&\quad + \ln|I_n - \lambda f_{D_n} W_n| \\
&= -\frac{n}{2}\ln\sigma_0^2 \\
&\quad - \frac{[T_n - \lambda_0 W_n F(T_n) - X_n\beta_0]'[T_n - \lambda_0 W_n F(T_n) - X_n\beta_0]}{2\sigma_0^2} \\
&\quad + \ln|I_n - \lambda_0 f_{D_n} W_n|
\end{aligned} \qquad (18)
$$

holds for T_n almost surely.

Differentiate Eq. (18) with respect to $t_{k,n}$, we have

$$
\begin{aligned}
&\sigma^{-2}\left[t_{k,n} - \lambda w_{k\cdot,n}F(T_n) - x_{k,n}\beta - \lambda f(t_{k,n})\right. \\
&\quad \left.\times \sum_{i=1}^{n}(t_{i,n} - \lambda w_{i\cdot,n}F(T_n) - x_{i,n}\beta)w_{ik,n}\right] \\
&\quad - \lambda f'(t_{k,n})\mathrm{tr}[(I_n - \lambda f_{D_n} w_n)^{-1}\overline{w_{k\cdot,n}}] \\
&= \sigma_0^{-2}\left[t_{k,n} - \lambda_0 w_{k\cdot,n}F(T_n) - x_{k,n}\beta_0 - \lambda_0 f(t_{k,n})\right. \\
&\quad \left.\times \sum_{i=1}^{n}(t_{i,n} - \lambda_0 w_{i\cdot,n}F(T_n) - x_{i,n}\beta_0)w_{ik,n}\right] \\
&\quad - \lambda_0 f'(t_{k,n})\mathrm{tr}[(I_n - \lambda_0 f_{D_n} W_n)^{-1}\overline{w_{k\cdot,n}}], \qquad (19)
\end{aligned}
$$

where $\overline{w_{k\cdot,n}}$ is an $n \times n$ matrix whose entries are zero except that its kth row is identical to the kth row of W_n. Differentiating the above equation with respect to $t_{j,n}, j \neq k$, we get

$$
\begin{aligned}
&\sigma^{-2}\left[-\lambda w_{kj,n}f(t_{j,n}) - \lambda w_{jk}f(t_{k,n})\right. \\
&\quad \left. + \lambda^2 f(t_{k,n})f(t_{j,n})\sum_{i=1}^{n} w_{ij,n}w_{ik,n}\right] \\
&\quad - \lambda^2 f'(t_{k,n})f'(t_{j,n})\mathrm{tr}[(I_n - \lambda f_{D_n} W_n)^{-1} \\
&\quad \times \overline{w_{j\cdot,n}}(I_n - \lambda f_{D_n} W_n)^{-1}\overline{w_{k\cdot,n}}] \\
&= \sigma_0^{-2}\left[-\lambda_0 w_{kj,n}f(t_j) - \lambda_0 w_{jk,n}f(t_{k,n})\right. \\
&\quad \left. + \lambda_0^2 f(t_{k,n})f(t_{j,n})\sum_{i=1}^{n} w_{ij,n}w_{ik,n}\right] \\
&\quad - \lambda_0^2 f'(t_{k,n})f'(t_{j,n})\mathrm{tr}[(I_n - \lambda_0 f_{D_n} W_n)^{-1} \\
&\quad \times \overline{w_{j\cdot,n}}(I_n - \lambda_0 f_{D_n} W_n)^{-1}\overline{w_{k\cdot,n}}].
\end{aligned}
$$

Let $t_{j,n}$ be such that $f'(t_{j,n}) = 0$ and $f(t_{j,n}) \neq 0$ ($t_{j,n}$ may be $+\infty$ or $-\infty$). Then the above equation implies

$$
\begin{aligned}
&\sigma^{-2}\left[-\lambda w_{kj,n}f(t_{j,n}) - \lambda w_{jk,n}f(t_{k,n})\right. \\
&\quad \left. + \lambda^2 f(t_{k,n})f(t_{j,n})\sum_{i=1}^{n} w_{ij,n}w_{ik,n}\right]
\end{aligned}
$$

$$= \sigma_0^{-2}\Big[-\lambda_0 w_{kj,n}f(t_j) - \lambda_0 w_{jk,n}f(t_{k,n}) + \lambda_0^2 f(t_{k,n})f(t_{j,n})\sum_{i=1}^n w_{ij,n}w_{ik,n}\Big]. \quad (20)$$

First, consider the case $F(\cdot)$ is not a linear function, i.e., $f(\cdot)$ is not a constant. Notice that both sides of the above equation are linear equations of $f(t_{k,n})$, so their constant terms are the same: $\lambda w_{kj,n}f(t_{j,n})/\sigma^2 = \lambda_0 w_{kj,n}f(t_{j,n})/\sigma_0^2$. Because $W_n \neq 0$ while its diagonal elements are all 0, there exist k and j such that $w_{kj,n} \neq 0$. As $f(t_j) \neq 0$, thus $\lambda/\sigma^2 = \lambda_0/\sigma_0^2$. Then Eq. (20) implies

$$\lambda^2 f(t_{k,n})f(t_{j,n})\sum_{i=1}^n w_{ij,n}w_{ik,n}/\sigma^2 = \lambda_0^2 f(t_{k,n})f(t_{j,n})\sum_{i=1}^n w_{ij,n}w_{ik,n}/\sigma_0^2.$$

Therefore, $\lambda^2\sum_{i=1}^n w_{ij,n}w_{ik,n}/\sigma^2 = \lambda_0^2\sum_{i=1}^n w_{ij,n}w_{ik,n}/\sigma_0^2$. Summation over k and j, we have $\lambda^2\sum_{j\neq k}(W_n'W_n)_{jk}/\sigma^2 = \lambda_0^2\sum_{j\neq k}(W_n'W_n)_{jk}/\sigma_0^2$. As $W_n'W_n$ is not a diagonal matrix, $\lambda^2/\sigma^2 = \lambda_0^2/\sigma_0^2$. Combining $\lambda/\sigma^2 = \lambda_0/\sigma_0^2$, we obtain $\lambda = \lambda_0$ and $\sigma = \sigma_0$.

Second, consider the case that $F(\cdot)$ is a linear function. Without loss of generality, assume $F(x) \equiv x$. Then $f(x) \equiv 1$ and Eq. (19) can be written as

$$\sigma^{-2}[T_n - \lambda W_nT_n - X_n\beta]'(I_n - \lambda W_n) = \sigma_0^{-2}[T_n - \lambda_0 W_nT_n - X_n\beta]'(I_n - \lambda_0 W_n).$$

Notice that both sides are linear equations of T_n, thus their "slopes" are the same: $\sigma^{-2}(I_n - \lambda W_n)'(I_n - \lambda W_n) = \sigma_0^{-2}(I_n - \lambda_0 W_n)'(I_n - \lambda_0 W_n)$. Therefore,

$$(\lambda^2\sigma^{-2} - \lambda_0^2\sigma_0^{-2})W_n'W_n - (\lambda\sigma^{-2} - \lambda_0\sigma_0^{-2}) \times (W_n' + W_n) + (\sigma^{-2} - \sigma_0^{-2})I_n = 0. \quad (21)$$

Consider the diagonal elements: $(\lambda^2\sigma^{-2} - \lambda_0^2\sigma_0^{-2})(W_n'W_n)_{ii} + (\sigma^{-2} - \sigma_0^{-2}) = 0$ for all i. Since $(W_n'W_n)_{ii}$'s are not all the same, we have $\lambda^2\sigma^{-2} = \lambda_0^2\sigma_0^{-2}$ and $\sigma^{-2} = \sigma_0^{-2}$. Now consider the off-diagonal elements of $(\lambda\sigma^{-2} - \lambda_0\sigma_0^{-2})(W_n' + W_n) = 0$. Because $W_n \neq 0$ and its elements are non-negative, we obtain $\lambda\sigma^{-2} = \lambda_0\sigma_0^{-2}$. Hence, we can identify λ_0 and σ_0 when $F(x) \equiv x$.

Hence, Eq. (16) implies $\mathrm{E}\ln L_n(\theta) - \mathrm{E}\ln L_n(\theta_0) = -\frac{1}{2\sigma_0^2}\mathrm{E}[(\beta_0 - \beta)'X_n'X_n(\beta_0 - \beta)] = 0$, which can hold only if $\beta_0 = \beta$. □

Proof of Proposition 3. (i) From the discussion after Corollary 1, $\{t_{i,n}\}$ is uniformly and geometrically L_2-NED with $\psi(md_0) = \zeta^m$. Because $f(x)$ is a Lipschitz function, we have that $f_i = f(t_i)$ is also uniformly and geometrically L_2-NED: $\|f_i - \mathrm{E}(f_i|\mathcal{F}_{i,n}(md_0))\|_2 \leqslant C\zeta^m$ for some constant $C > 0$. Denote $i_0 = i$. Then with the inequality $|x_0x_1\cdots x_l - y_0y_1\cdots y_l| \leqslant b_f^l\sum_{i=1}^l|x_i - y_i|$ when all x_i's and y_i's are in $[-b_f, b_f]$, we have

$$\begin{aligned}&\|f_if_{i_1}f_{i_2}\cdots f_{i_l} - \mathrm{E}[f_if_{i_1}f_{i_2}\cdots f_{i_l}|\mathcal{F}_{i,n}(md_0)]\|_2\\ &\leqslant \Big\|\prod_{j=0}^l f_{i_j} - \prod_{j=0}^l \mathrm{E}[f_{i_j}|\mathcal{F}_{i,n}(md_0)]\Big\|_2\\ &\leqslant b_f^l\sum_{j=0}^l\|f_{i_j} - \mathrm{E}[f_{i_j}|\mathcal{F}_{i,n}(md_0)]\|_2\\ &\leqslant b_f^l\sum_{j=0}^l\|f_{i_j} - \mathrm{E}[f_{i_j}|\mathcal{F}_{i_j,n}((m-j)d_0)]\|_2\\ &\leqslant b_f^lC(\zeta^m + \zeta^{m-1} + \cdots + \zeta^{m-l})\\ &= b_f^l\frac{C(\zeta^{-1-l} - 1)}{\zeta^{-1} - 1}\zeta^m. \end{aligned} \quad (22)$$

(ii) For any given small positive number $\epsilon > 0$, we can divide the summation in Eq. (7) into two parts ($l \leqslant K_0$ & $l > K_0$), where the fixed natural number K_0 will be determined later. We will show that the first part converges to zero uniformly and the second part can be bounded by $\epsilon/2$.

To show the convergence of the first part, we only need to calculate its variance. By Lemma A.2 and Eq. (22), we know that for any location i, arbitrarily pick a natural number $l \leqslant K_0$ and locations $j_1, j_2, \ldots, j_{l-1}$ such that $d(i, j_1) \leqslant d_0$ and $d(j_h, j_{h-1}) \leqslant d_0$ for all $2 \leqslant h \leqslant l$, then $\{f_if_{j_1}\cdots f_{j_{l-1}}\}$ are L_2-NED: $\|f_if_{j_1}\cdots f_{j_{l-1}} - \mathrm{E}[f_if_{j_1}\cdots f_{j_{l-1}}|\mathcal{F}_{i,n}(md_0)]\|_2 \leqslant b_f^{K_0}\frac{C(\zeta^{-1-K_0}-1)}{\zeta^{-1}-1}\zeta^m$ for some constant $C > 0$.[12] So by Lemma A.3 in Jenish and Prucha (2012), if locations $i' \to j_1' \to j_2' \to \cdots j_{l-1}'$ also satisfy that $d(i', j_1') \leqslant d_0$ and $d(j_h', j_{h-1}') \leqslant d_0$ for all $2 \leqslant h \leqslant l$, then there exists a constant $C_2 > 0$ $s.t.$ $|\mathrm{cov}(f_if_{j_1}\cdots f_{j_{l-1}}, f_{i'}f_{j_1'}\cdots f_{j_{l-1}'})| \leqslant C_2\zeta^{d(i,i')/3}$.

Denote $g_{nl} = \sum_{i=1}^n\sum_{j_1}\cdots\sum_{j_{l-1}} w_{ij_1,n}w_{j_1j_2,n}\cdots w_{j_{l-1}i,n}(f_if_{j_1}\cdots f_{j_{l-1}} - \mathrm{E}f_if_{j_1}\cdots f_{j_{l-1}})$. Then we have

$$\begin{aligned}&\mathrm{Var}\left(\frac{1}{n}g_{nl}\right)\\ &\leqslant \frac{1}{n^2}\sum_{i=1}^n\sum_{i'=1}^n\sum_{j_1}\sum_{j_2}\cdots\sum_{j_{l-1}}\sum_{j_1'}\sum_{j_2'}\cdots\sum_{j_{l-1}'} w_{ij_1,n}w_{j_1j_2,n}\cdots\\ &\quad\times w_{j_{l-1}i,n}w_{i'j_1',n}w_{j_1'j_2',n}\cdots\\ &\quad\times w_{j_{l-1}'i',n}|\mathrm{cov}(f_if_{j_1}\cdots f_{j_{l-1}},\ f_{i'}f_{j_1'}\cdots f_{j_{l-1}'})|\\ &\leqslant \frac{1}{n^2}\sum_{i=1}^n\sum_{i'=1}^n C_2\zeta^{d(i,i')/3}\sum_{j_1}\cdots\sum_{j_{l-1}}\sum_{j_1'}\cdots\\ &\quad\times\sum_{j_{l-1}'} w_{ij_1,n}\cdots w_{j_{l-1}i,n}w_{i'j_1',n}\cdots w_{j_{l-1}'i',n}\\ &\leqslant \frac{1}{n^2}\|W_n\|_\infty^{2l}\sum_{i=1}^n\sum_{i'=1}^n C_2\zeta^{d(i,i')/3}.\end{aligned}$$

Define $N_i(1,1,m) = \{j : (m-1)d_0 \leqslant d(i,j) \leqslant md_0\}$. Because all the positions are in $\mathbb{R}^d$, there exists a constant C_3 such that $|N_i(1,1,m)| \leqslant C_3m^{d-1}$ from Jenish and Prucha (2009). Then $\sum_{i=1}^n\sum_{i'=1}^n C_2\zeta^{d(i,i')/3} \leqslant \sum_{i=1}^n\sum_{m=1}^\infty C_3m^{d-1}C_2\zeta^{(m-1)d_0/3} = O(n)$ as $\sum_{m=1}^\infty C_3m^{d-1}C_2\zeta^{(m-1)/3} < \infty$. This shows that $\frac{1}{n}g_{nl}(\lambda) = o_p(1)$. The uniform convergence $\sup_{\lambda\in\Lambda}|\frac{1}{n}\sum_{l=1}^{K_0}g_{nl}\lambda^l/l| = o_p(1)$ holds because λ appears as a polynomial.

Now we consider the proof of the remaining part where $l > K_0$.

$$\begin{aligned}&\left|\frac{1}{n}\sum_{l=K_0+1}^\infty g_{nl}(\lambda)\right|\\ &\leqslant \frac{1}{n}\sum_{l=K_0+1}^\infty\frac{\lambda^l}{l}\sum_{i=1}^n\sum_{j_1}\cdots\sum_{j_{l-1}} w_{ij_1,n}w_{j_1j_2,n}\cdots\\ &\quad\times w_{j_{l-1}i,n}|f_if_{j_1}\cdots f_{j_{l-1}} - \mathrm{E}f_if_{j_1}\cdots f_{j_{l-1}}|\\ &\leqslant \frac{2}{n}\sum_{l=K_0+1}^\infty\frac{\|\lambda W_n\|_\infty^l}{l}b_f^l \leqslant 2\sum_{l=K_0+1}^\infty\frac{\zeta^l}{l}\end{aligned}$$

12 Here without loss of generality, we assume $b_f \geq 1$.

$$\leqslant \frac{2}{K_0}\sum_{l=K_0+1}^{\infty}\zeta^l = \frac{2}{K_0}\frac{\zeta^{K_0+1}}{1-\zeta} < \frac{\epsilon}{2}$$

so long as $K_0 > K_\epsilon$ for some positive integer K_ϵ with $2\zeta^{K_\epsilon+1}K_\epsilon^{-1}/(1-\zeta) < \epsilon/2$. Notice that K_ϵ does not depend on the sample size n. Then

$$P\left(\sup_{\lambda\in\Lambda}\left|\frac{1}{n}\ln|I_n-\lambda f_{D_n}W_n| - \mathrm{E}\ln|I_n-\lambda f_{Dn}W_n|\right| > \epsilon\right)$$
$$= P\left(\sup_{\lambda\in\Lambda}\frac{1}{n}\left|\sum_{l=1}^{K_0} g_{nl}(\lambda) + \sum_{l=K_0+1}^{\infty} g_{nl}(\lambda)\right| > \epsilon\right)$$
$$\leqslant P\left(\sup_{\lambda\in\Lambda}\frac{1}{n}\left|\sum_{l=1}^{K_0} g_{nl}(\lambda)\right| + \sup_{\lambda\in\Lambda}\frac{1}{n}\left|\sum_{l=K_0+1}^{\infty} g_{nl}(\lambda)\right| > \epsilon\right)$$
$$\leqslant P\left(\sup_{\lambda\in\Lambda}\frac{1}{n}\left|\sum_{l=1}^{K_0} g_{nl}(\lambda)\right| > \epsilon/2\right) \to 0,$$

as $n\to\infty$. □

Proof of Theorem 1. Because $\ln|f_{D_n}^{-1}-\lambda W_n| = \ln|I_n-\lambda f_{D_n}W_n| - \ln|f_{D_n}|$, it causes no harm to drop the term $\ln|f_{D_n}|$, which does not involve parameters, in the analysis of consistency of an extremum estimator. Then $\ln L_n(\theta) = -\frac{n}{2}\ln(2\pi\sigma^2) - \frac{1}{2\sigma^2}[F^{-1}(S_n) - \lambda W_nS_n - X_n\beta]'[F^{-1}(S_n) - \lambda W_nS_n - X_n\beta] + \ln|I_n-\lambda f_{D_n}W_n|$. In order to establish the consistency of the ML estimator, with the identification condition in Assumption 9, it remains to show the uniform convergence $\frac{1}{n}[\sup_{\theta\in\Theta}|\ln L_n(\theta) - Q_n(\theta)|] \xrightarrow{p} 0$, and the equicontinuity of $\frac{1}{n}Q_n(\theta)$.

Proof of the uniform convergence

Denote $v_{i,n}(\lambda,\beta) = F^{-1}(s_{i,n}) - \lambda w_{i\cdot,n}S_n - x_{i,n}\beta = (\lambda_0-\lambda)w_{i\cdot,n}S_n + x_{i,n}(\beta_0-\beta)+\epsilon_{i,n}$. With Proposition 3, it remains to show that $\mathrm{p}\lim_{n\to\infty}\sup_{\theta\in\Theta}\frac{1}{n}|\sum_{i=1}^n v_{i,n}(\lambda,\beta)^2 - \mathrm{E}v_{i,n}(\lambda,\beta)^2| = 0$. To do so, it is sufficient for us to show the pointwise convergence $\mathrm{p}\lim_{n\to\infty}\frac{1}{n}[\sum_{i=1}^n v_{i,n}(\lambda,\beta)^2 - \mathrm{E}v_{i,n}(\lambda,\beta)^2] = 0$ for each (λ,β), and the stochastic equicontinuity of $v_{i,n}(\lambda,\beta)^2$.

Under Assumptions 6 and 7, Corollary 1 implies that $v_{i,n}(\lambda,\beta)$ is L_5 bounded uniformly in i and n, and geometrically L_2-NED uniformly in i and n. Thus, $v_{i,n}(\lambda,\beta)^2$ is $L_{2.5}$ bounded uniformly in i and n, and geometrically L_2-NED uniformly in i and n by Lemma A.1. Thus, the pointwise convergence holds by the LLN in Jenish and Prucha (2012). By Lemma 1 in Andrews (1992), the stochastic equicontinuity originates in uniform L_2 boundedness of $w_{i\cdot,n}S_n$ and $x_{i,n}$, and

$$\left|\frac{1}{n}\sum_{i=1}^n v_{i,n}(\lambda_1,\beta_1)^2 - \frac{1}{n}\sum_{i=1}^n v_{i,n}(\lambda_2,\beta_2)^2\right|$$
$$= \left|\frac{1}{n}\sum_{i=1}^n [v_{i,n}(\lambda_1,\beta_1) + v_{i,n}(\lambda_2,\beta_2)] \cdot [(\lambda_2-\lambda_1)w_{i\cdot,n}S_n + x_{i,n}(\beta_2-\beta_1)]\right|$$
$$\leqslant \frac{1}{n}\sum_{i=1}^n [4\lambda_m|w_{i\cdot,n}S_n| + 4\sum_{k=1}^K |x_{ik,n}|\cdot\sup_{\beta_k}|\beta_k| + 2|\epsilon_{i,n}|]\cdot[|w_{i\cdot,n}S_n|\cdot|\lambda_2-\lambda_1| + \sum_{k=1}^K|x_{ik,n}|\cdot|\beta_{2k}-\beta_{1k}|].$$

Proof of the equicontinuity

With stochastic equicontinuity and the boundedness of the parameter space, the equicontinuity of $\sigma^{-2}\mathrm{E}[F^{-1}(S_n) - \lambda W_nS_n - X_n\beta]'[F^{-1}(S_n) - \lambda W_nS_n - X_n\beta]$ is a result of Corollary 3.1 in Newey (1991).

Because $\frac{1}{n}\mathrm{E}\ln|I_n-\lambda_1 f_{D_n}W_n| - \frac{1}{n}\mathrm{E}\ln|I_n-\lambda_2 f_{D_n}W_n| = (\lambda_1-\lambda_2)\frac{1}{n}\mathrm{Etr}[(I_n-\bar{\lambda}f_{D_n}W_n)^{-1}f_{D_n}W_n]$, and

$$\|(I_n-\bar{\lambda}f_{D_n}W_n)^{-1}f_{D_n}W_n\|_\infty = \left\|\sum_{l=0}^{\infty}(\bar{\lambda}f_{D_n}W_n)^l f_{D_n}W_n\right\|_\infty$$
$$= \frac{1}{\lambda_m}\left\|\sum_{l=0}^{\infty}(\bar{\lambda}f_{D_n}W_n)^l\lambda_m f_{D_n}W_n\right\|_\infty \leqslant \frac{1}{\lambda_m}\sum_{l=0}^{\infty}\zeta^l = \frac{\zeta}{\lambda_m(1-\zeta)},$$

we have $\left|\frac{1}{n}\mathrm{E}\ln|I_n-\lambda_1 f_{D_n}W_n| - \frac{1}{n}\mathrm{E}\ln|I_n-\lambda_2 f_{D_n}W_n|\right| \leqslant |\lambda_1-\lambda_2|\frac{\zeta}{\lambda_m(1-\zeta)}$. □

Before proving asymptotic normality of the ML estimator, we first prove the uniformly and geometrically L_2-NED property of $\left\{\frac{z_{i,n}\epsilon_i}{\sigma_0^2} - r_{ii,n} - \mathrm{E}(\frac{z_{i,n}\epsilon_i}{\sigma_0^2} - r_{ii,n})\right\}$, where $z_{i,n} \equiv \sum_j w_{ij,n}s_{j,n}$. Denote $x_{i,n} = (x_{i1,n},\ldots,x_{iK,n})$.

Lemma B.4. *$\{z_{i,n}\epsilon_i/\sigma_0^2 - r_{ii,n} - \mathrm{E}[z_{i,n}\epsilon_i/\sigma_0^2 - r_{ii,n}]\}_{i=1}^n$ is uniformly $L_{2.5}$ bounded, and geometrically L_2-NED uniformly in i and n. $\{q_{i,n} \equiv [\sum_{j=1}^K(\frac{x_{ij,n}\epsilon_{i,n}}{\sigma_0^2})^2 + (z_{i,n}\epsilon_i/\sigma_0^2 - r_{ii,n} - \mathrm{E}(z_{i,n}\epsilon_i/\sigma_0^2 - r_{ii,n}))^2 + (\frac{\epsilon_{i,n}^2-\sigma_0^2}{2\sigma_0^4})^2]^{1/2}\}$ is also geometrically L_2-NED uniformly in i and n.*

Proof. By Corollary 1, $\{z_{i,n}\epsilon_i/\sigma_0^2\}$ is uniformly $L_{2.5}$ bounded, and geometrically L_2-NED uniformly in i and n. Because $\sup_{i,n}|\mathrm{E}z_{i,n}\epsilon_i| \leqslant \sup_{i,n}\mathrm{E}|z_{i,n}\epsilon_i| \leqslant \sup_{i,n}\|z_{i,n}\epsilon_i\|_p < \infty$, the $L_{2.5}$ boundedness in the first claim follows from

$$|r_{ii,n} - \mathrm{E}r_{ii,n}|$$
$$\leqslant \sum_{l=0}^{\infty}|\lambda_m^l|\sum_{j_1}\sum_{j_2}\cdots\sum_{j_l} w_{ij_1,n}w_{j_1j_2,n}\cdots \times w_{j_{l-1}j_l,n}w_{j_li,n}|f_if_{j_1}\cdots f_{j_l} - \mathrm{E}f_if_{j_1}\cdots f_{j_l}|$$
$$\leqslant 2\sum_{l=0}^{\infty}\lambda_m^l\|W_n\|_\infty^{l+1}b_f^{l+1} \leqslant \frac{2\zeta}{\lambda_m(1-\zeta)}.$$

Next, we establish the uniformly and geometrically L_2-NED property of $\{r_{ii,n}\}$. For $\{f_{i,n} = f(t_{i,n})\}$, from the proof of Proposition 3, $\|f_{i,n} - \mathrm{E}(f_{i,n}|\mathcal{F}_{i,n}(md_0))\|_2 \leqslant A_1\zeta^m$ for some constant A_1. Since the chain $i\to j_1\to\cdots\to j_l\to i$ is closed, we have $d(j_1,i)\leqslant d_0$, $d(j_2,i)\leqslant 2d_0,\ldots,d(j_{[(l+1)/2]},i)\leqslant[\frac{l+1}{2}]d_0,\ldots,d(j_{l-1},i)\leqslant 2d_0$, $d(j_l,i)\leqslant d_0$. So, with the inequality: $|x_1\cdots x_l - y_1\cdots y_l| \leqslant C^{l-1}\sum_{i=1}^l|x_i-y_i|$ if $|x_i|\leqslant C$ and $|y_i|\leqslant C$ for all i's, when $l<m$, we have

$$\|f_if_{j_1}\cdots f_{j_l} - \mathrm{E}[f_if_{j_1}\cdots f_{j_l}|\mathcal{F}_{i,n}(md_0)]\|_2$$
$$\leqslant \|f_if_{j_1}\cdots f_{j_l} - \mathrm{E}[f_i|\mathcal{F}_{i,n}(md_0)] \times \mathrm{E}[f_{j_1}|\mathcal{F}_{i,n}(md_0)]\cdots\mathrm{E}[f_{j_l}|\mathcal{F}_{i,n}(md_0)]\|_2$$
$$\leqslant b_f^l\left(\sum_{k=1}^l\|f_{j_k} - \mathrm{E}[f_{j_k}|\mathcal{F}_{i,n}(md_0)]\|_2 + \|f_i - \mathrm{E}[f_i|\mathcal{F}_{i,n}(md_0)]\|_2\right)$$
$$\leqslant 2A_1b_f^l\sum_{k=0}^{[(l+1)/2]}\zeta^{m-k} = 2A_1b_f^l\frac{\zeta^{-[(l+1)/2]-1}-1}{\zeta^{-1}-1}\zeta^m. \tag{23}$$

Hence,

$$\left\|\sum_{l=0}^{\infty}\lambda_0^l((f_{Dn}W_n)^{l+1})_{ii} - \mathrm{E}\left[\sum_{l=0}^{\infty}\lambda_0^l((f_{D_n}W_n)^{l+1})_{ii}|\mathcal{F}_{i,n}(md_0)\right]\right\|_2$$

$$\leqslant \sum_{l=0}^{m-1} |\lambda_0^l| \sum_{j_1} \cdots \sum_{j_l} w_{ij_1,n} w_{j_1 j_2,n} \cdots$$
$$\times w_{j_{l-1}j_l,n} w_{j_l i,n} \| f_i f_{j_1} \cdots f_{j_l} - \mathrm{E}[f_i f_{j_1} \cdots f_{j_l} | \mathcal{F}_{i,n}(md_0)] \|_2$$
$$+ \sum_{l=m}^{\infty} |\lambda_0^l| \sum_{j_1} \cdots \sum_{j_l} w_{ij_1,n} w_{j_1 j_2,n} \cdots$$
$$\times w_{j_{l-1}j_l,n} w_{j_l i,n} \| f_i \cdots f_{j_l} - \mathrm{E}[f_i \cdots f_{j_l} | \mathcal{F}_{i,n}(md_0)] \|_2$$
$$\leqslant \sum_{l=0}^{m-1} |\lambda_0^l| \cdot \|W_n\|_\infty^{l+1} 2A_1 b_f^l \frac{\zeta^{-[(l+1)/2]-1} - 1}{\zeta^{-1} - 1} \zeta^m$$
$$+ 2 \sum_{l=m}^{\infty} |\lambda_0^l| \cdot \|W_n\|_\infty^{l+1} b_f^{l+1} \leqslant A_2 \zeta^m$$

for some constant A_2 that does not depend on n, which means that $\{r_{ii,n} - \mathrm{E}r_{ii,n}\}$ is uniformly and geometrically L_2-NED.

The uniformly and geometrically L_2-NED property of $\{q_{i,n}\}$ is a result of Lemma A.3. □

Proof of Proposition 4. We need to check the conditions of the CLT of the L_2-NED sequence, i.e., Assumptions 3 and 4 in Jenish and Prucha (2012) hold for $\{q_{i,n}\}$ defined in Lemma B.4. Assumption 3 in Jenish and Prucha (2012) is satisfied because the error terms are i.i.d. and $\{x_{i,n}\}$ satisfies Assumption 11. With Lemma B.4, conditions (c) and (d) of Assumption 4 in Jenish and Prucha (2012) hold. Under Assumption 12, the condition (b) in Assumption 4 in Jenish and Prucha (2012) is satisfied. So it remains to check the uniform $L_{2+\delta_1}$ integrability for some $\delta_1 > 0$. One sufficient condition (Shorack, 2000, p. 54) is to show $\sup_{i,n} \mathrm{E}q_{i,n}^{2+\delta_2} < \infty$ for some $\delta_2 > 0$. Because $\sup_{i,n} \mathrm{E}|\frac{z_{i,n}\epsilon_{i,n}}{\sigma_0^2} - r_{ii,n} - \mathrm{E}(z_{i,n}\epsilon_i/\sigma_0^2 - r_{ii,n})|^{2.5} < \infty$ from Lemma B.4, $\{\epsilon_{i,n}\}$ is normally distributed and $\{x_{ij,n}\}$ are uniformly L_5 bounded, we have

$$\sup_{i,n} \mathrm{E}q_{i,n}^{2.5} \leqslant \sup_{i,n} (K+2)^{1.5} \left[\sum_{j=1}^K \mathrm{E} \left| \frac{x_{ij,n}\epsilon_{i,n}}{\sigma_0^2} \right|^{2.5} + \mathrm{E} \left| \frac{z_{i,n}\epsilon_{i,n}}{\sigma_0^2} - r_{ii,n} - \mathrm{E}\left(\frac{z_{i,n}\epsilon_{i,n}}{\sigma_0^2} - r_{ii,n} \right) \right|^{2.5} + \mathrm{E} \left| \frac{\epsilon_{i,n}^2 - \sigma_0^2}{2\sigma_0^4} \right|^{2.5} \right] < \infty. \quad \square$$

Proof of Theorem 2. We will show that $\frac{1}{n}|\frac{\partial^2 \ln L_n(\theta_0)}{\partial\theta\partial\theta'} - \mathrm{E}\frac{\partial^2 \ln L_n(\theta_0)}{\partial\theta\partial\theta'}| \xrightarrow{p} 0$ and $\frac{1}{n}|\frac{\partial^2 \ln L_n(\hat\theta_n)}{\partial\theta\partial\theta'} - \frac{\partial^2 \ln L_n(\theta_0)}{\partial\theta\partial\theta'}| \xrightarrow{p} 0$, then $\frac{1}{n}|\frac{\partial^2 \ln L_n(\hat\theta_n)}{\partial\theta\partial\theta'} - \mathrm{E}\frac{\partial^2 \ln L_n(\theta_0)}{\partial\theta\partial\theta'}| \xrightarrow{p} 0$. The second order derivatives of the log likelihood are

$\frac{\partial^2 \ln L_n(\theta)}{\partial\beta\partial\beta'} = -\frac{X_n'X_n}{\sigma^2}$,

$\frac{\partial^2 \ln L_n(\theta)}{\partial\beta\partial\lambda} = -\frac{X_n'W_nS_n}{\sigma^2}$,

$\frac{\partial^2 \ln L_n(\theta)}{\partial\beta\partial\sigma^2} = -\frac{X_n'(F^{-1}(S_n) - \lambda W_nS_n - X_n\beta)}{\sigma^4}$,

$\frac{\partial^2 \ln L_n(\theta)}{\partial\lambda\partial\sigma^2} = -\frac{(W_nS_n)'(F^{-1}(S_n) - \lambda W_nS_n - X_n\beta)}{\sigma^4}$,

$\frac{\partial^2 \ln L_n(\theta)}{\partial\lambda^2} = -\mathrm{tr}[(I_n - \lambda f_{D_n}W_n)^{-1} f_{D_n}W_n (I_n - \lambda f_{D_n}W_n)^{-1} f_{D_n}W_n] - \frac{(W_nS_n)'(W_nS_n)}{\sigma^2}$,

$\frac{\partial^2 \ln L_n(\theta)}{\partial\sigma^2\partial\sigma^2} = \frac{n}{2\sigma^4} - \frac{(F^{-1}(S_n) - \lambda W_nS_n - X_n\beta)'(F^{-1}(S_n) - \lambda W_nS_n - X_n\beta)}{\sigma^6}$.

Similarly to the proof of Theorem 1, with the L_5 boundedness of $\{x_{i,n}\}$, $\{w_{i\cdot,n}S_n\}$ and $\{v_{i,n}(\lambda, \beta) \equiv F^{-1}(s_{i,n}) - \lambda w_{i\cdot,n}S_n - x_{i,n}\beta\}$ uniformly in i and n, and their geometric L_2-NED properties, their products obey the weak LLN in Jenish and Prucha (2012). Thus, in order to prove $\frac{1}{n}|\frac{\partial^2 \ln L_n(\theta_0)}{\partial\theta\partial\theta'} - \mathrm{E}\frac{\partial^2 \ln L_n(\theta_0)}{\partial\theta\partial\theta'}| \xrightarrow{p} 0$, it suffices to show that $\frac{1}{n}\{\mathrm{tr}[(I - \lambda_0 f_{D_n}W_n)^{-1} f_{D_n}W_n]^2 - \mathrm{E}\mathrm{tr}[(I - \lambda_0 f_{D_n}W_n)^{-1} f_{D_n}W_n]^2\} \xrightarrow{p} 0$. To do so, we show that $\{([(I_n - \lambda f_{D_n}W_n)^{-1} f_{D_n}W_n]^2)_{ii}\}$ is uniformly bounded and L_2-NED uniformly in i and n. Because

$$\{[(I_n - \lambda f_{D_n}W_n)^{-1} f_{D_n}W_n]^2\}_{ii}$$
$$= \left(\sum_{l=0}^{\infty} \lambda^l (f_{D_n}W_n)^{l+1} \sum_{l'=0}^{\infty} \lambda^{l'} (f_{D_n}W_n)^{l'+1} \right)_{ii}$$
$$= \sum_{k=0}^{\infty} \sum_{l=0}^{k} \lambda^k \sum_{j_1} \cdots \sum_{j_{k+1}} (f_{D_n}W_n)_{ij_1} \cdots (f_{D_n}W_n)_{j_k j_{k+1}} (f_{D_n}W_n)_{j_{k+1} j_i}$$
$$= \sum_{k=0}^{\infty} (1+k) \lambda^k \sum_{j_1} \cdots \sum_{j_{k+1}} w_{ij_1,n} \cdots w_{j_{k+1}i,n} f_i f_{j_1} \cdots f_{j_{k+1}},$$

the uniform boundedness comes from

$$|\{[(I_n - \lambda f_{D_n}W_n)^{-1} f_{D_n}W_n]^2\}_{ii}| \leqslant \sum_{k=0}^{\infty} (1+k) \lambda_m^k \|W_n\|_\infty^{k+2} b_f^{k+2} \leqslant \lambda_m^{-2} \sum_{k=0}^{\infty} (1+k) \zeta^{k+2} < \infty.$$

When $k \leqslant m$, inequality (23) implies

$$|\lambda_0^k| \sum_{j_1} \cdots \sum_{j_{k+1}} w_{ij_1,n} \cdots w_{j_{k+1}i,n} \| f_i f_{j_1} \cdots f_{j_{k+1}} - \mathrm{E}(f_i f_{j_1} \cdots f_{j_{k+1}} | \mathcal{F}_{i,n}(md_0)) \|_2$$
$$\leqslant \lambda_m^k \|w_n\|_\infty^{k+2} 2A_1 b_f^{k+1} \frac{\zeta^{-[(k+2)/2]-1} - 1}{\zeta^{-1} - 1} \zeta^m \leqslant A_3 \zeta^{m+\frac{k}{2}}$$

for some constant $A_2 > 0$. When $k > m$,

$$|\lambda_0^k| \sum_{j_1} \cdots \sum_{j_{k+1}} w_{ij_1,n} \cdots w_{j_{k+1}i,n} \| f_i f_{j_1} \cdots f_{j_{k+1}} - \mathrm{E}(f_i f_{j_1} \cdots f_{j_{k+1}} | \mathcal{F}_{i,n}(md_0)) \|_2$$
$$\leqslant 2\lambda_m^k \|W_n\|_\infty^{k+2} b_f^{k+2} \leqslant A_3 \zeta^k$$

for $A_3 = 2\lambda_m^{-2}$. So,

$$\| \{[(I_n - \lambda f_{D_n}W_n)^{-1} f_{D_n}W_n]^2\}_{ii} - \mathrm{E}[\{[(I_n - \lambda f_{D_n}W_n)^{-1} f_{D_n}W_n]^2\}_{ii} | \mathcal{F}_{i,n}(md_0)] \|_2$$
$$\leqslant \sum_{k=0}^{\infty} (1+k) |\lambda_0^k| \sum_{j_1} \cdots \sum_{j_{k+1}} w_{ij_1,n} \cdots \times w_{j_{k+1}i,n} \| f_i f_{j_1} \cdots f_{j_{k+1}} - \mathrm{E}(f_i f_{j_1} \cdots f_{j_{k+1}} | \mathcal{F}_{i,n}(md_0)) \|_2$$
$$\leqslant \sum_{k=0}^{m} (1+k) A_2 \zeta^{m+\frac{k}{2}} + \sum_{k=m+1}^{\infty} (1+k) A_3 \zeta^k = O(\zeta^m).$$

Therefore, $\{[((I_n - \lambda f_{D_n}W_n)^{-1} f_{D_n}W_n)^2]_{ii}\}$ is geometrically uniformly L_2-NED.

Thus, we have shown $\frac{1}{n}|\frac{\partial^2 \ln L_n(\theta_0)}{\partial\theta\partial\theta'} - \mathrm{E}\frac{\partial^2 \ln L_n(\theta_0)}{\partial\theta\partial\theta'}| \xrightarrow{p} 0$. Next, we will prove $\frac{1}{n}|\frac{\partial^2 \ln L_n(\hat\theta_n)}{\partial\theta\partial\theta'} - \frac{\partial^2 \ln L_n(\theta_0)}{\partial\theta\partial\theta'}| \xrightarrow{p} 0$. Because $\hat\theta_n - \theta_0 \xrightarrow{p} 0$, it is easy to check the other terms except $\frac{\partial^2 \ln L_n(\theta)}{\partial\lambda^2}$. To do so, we only need to check that $\frac{d}{d\lambda}\frac{1}{n}\mathrm{tr}[((I_n - \lambda f_{D_n}W_n)^{-1} f_{D_n}W_n)^2] = \frac{2}{n}\mathrm{tr}[(I_n - \lambda f_{D_n}W_n)^{-1} f_{D_n}W_n]^3$ is bounded. A sufficient condition is that $([(I_n - \lambda f_{D_n}W_n)^{-1} f_{D_n}W_n]^3)_{ii}$ is uniformly bounded:

$$|([(I_n - \lambda f_{D_n}W_n)^{-1} f_{D_n}W_n]^3)_{ii}|$$
$$= \left(\sum_{l=0}^{\infty} \lambda^l (f_{D_n}W_n)^{l+1} \sum_{l'=0}^{\infty} \lambda^{l'} (f_{D_n}W_n)^{l'+1} \sum_{l''=0}^{\infty} \lambda^{l''} (f_{D_n}W_n)^{l''+1} \right)_{ii}$$
$$= \left| \sum_{k=0}^{\infty} \sum_{l+l'+l''=k} \lambda_0^k \sum_{j_1} \cdots \sum_{j_{k+2}} w_{ij_1,n} \cdots w_{j_{k+2}i,n} f_i f_{j_1} \cdots f_{j_{k+2}} \right|$$

$$\leqslant \sum_{k=0}^{\infty} \sum_{l+l'+l''=k} |\lambda_m^{-3}| \zeta^{k+3}$$

$$= |\lambda_m^{-3}| \sum_{k=0}^{\infty} 0.5(k+1)(k+2)\zeta^{k+2} < \infty.$$

Therefore, from $\frac{1}{\sqrt{n}} \frac{\partial \ln L_n(\hat{\theta}_n)}{\partial \theta} = 0 = \frac{1}{\sqrt{n}} \frac{\partial \ln L_n(\theta_0)}{\partial \theta} + \frac{1}{n} \frac{\partial^2 \ln L_n(\bar{\theta})}{\partial \theta \partial \theta'} \sqrt{n} (\hat{\theta}_n - \theta_0)$, we have $\sqrt{n}(\hat{\theta}_n - \theta_0) = (-\frac{1}{n} \frac{\partial^2 \ln L_n(\bar{\theta})}{\partial \theta \partial \theta'})^{-1} \frac{1}{\sqrt{n}} \frac{\partial \ln L_n(\theta_0)}{\partial \theta} \xrightarrow{d} N(0, \Sigma_0^{-1})$. □

Appendix C. Supplementary data

Supplementary material related to this article can be found online at http://dx.doi.org/10.1016/j.jeconom.2014.12.005.

References

Amemiya, T., 1985. Advanced Econometrics. Harvard University Press, Cambridge, MA.
Andrews, D.W., 1992. Generic uniform convergence. Econometric theory 8, 241–257.
Anselin, L., 1988. Spatial Econometrics: Methods and Models. Kluwer Academic Publisher, The Netherlands.
Anselin, L., Bera, A.K., 1998. Spatial dependence in linear regression models with an introduction to spatial econometrics. In: Ullah, A. (Ed.), Handbook of Applied Economic Statistics. CRC Press, New York: Marcel Dekker, pp. 237–290.
Chen, S., 2002. Rank estimation of transformation models. Econometrica 70, 1683–1697.
Davidson, J., 1994. Stochastic Limit Theory. Oxford University Press, Oxford.
Gallant, A.R., White, H., 1988. A Unified Theory of Estimation and Inference for Nonlinear Dynamic Models. Basil Blackwell, New York.
Harville, D., 1997. Matrix Algebra from a Statistician's Perspective. Springer-Verlag, New York.
Hill, J.B., 2010. Lecture Notes for Time Series: Dependence and Stochastic Limit Theory. Dept. of Economics, University of North Carolina - Chapel Hill.
Horowitz, J.L., 1996. Semiparametric estimation of a regression model with an unknown transformation of the dependent variable. Econometrica 64, 103–137.
Jenish, N., 2012. Nonparametric spatial regression under near-epoch dependence. Journal of Econometrics 167, 224–239.
Jenish, N., Prucha, I.R., 2009. Central limit theorems and uniform laws of large numbers for arrays of random fields. J. Econometrics 150, 86–98.
Jenish, N., Prucha, I.R., 2012. On spatial processes and asymptotic inference under near-epoch dependence. J. Econometrics 170, 178–190.
Kelejian, H.H., Prucha, I.R., 1999. A generalized moments estimator for the autoregressive parameter in a spatial model. Internat. Econom. Rev. 40, 509–533.
Kelejian, H.H., Prucha, I.R., 2001. On the asymptotic distribution of the Moran I test statistic with applications. J. Econometrics 104, 219–257.
Lee, L., 2004. Asymptotic distributions of quasi-maximum likelihood estimators for spatial autoregressive models. Econometrica 72, 1899–1925.
Lee, L., 2007. GMM and 2SLS estimation of mixed regressive, spatial autoregressive models. J. Econometrics 137, 489–514.
Lee, L., Yu, J., 2010. Some recent developments in spatial panel data models. Reg. Sci. Urban Econ. 40, 255–271.
LeSage, J., Pace, R., 2009. Introduction to Spatial Econometrics. CRC Press, Boca Raton.
Lin, X., Lee, L., 2010. GMM estimation of spatial autoregressive models with unknown heteroskedasticity. J. Econometrics 157, 34–52.
Newey, W.K., 1991. Uniform convergence in probability and stochastic equicontinuity. Econometrica 59, 1161–1167.
Ord, J.K., 1975. Estimation methods for models of spatial interaction. J. Amer. Statist. Assoc. 70, 120–126.
Papke, L.E., Wooldridge, J.M., 2008. Panel data methods for fractional response variables with an application to test pass rates. J. Econometrics 145, 121–133.
Qu, X., Lee, L., 2013. Locally most powerful tests for spatial interactions in the simultaneous SAR Tobit model. Reg. Sci. Urban Econ. 43, 307–321.
Royden, H., Fitzpatrick, P., 2010. Real Analysis. China Machine Press, Beijing.
Sheldon, R., 2009. A First Course in Probability, 8/E. Pearson Education India. Upper Saddle River, New Jersey.
Shorack, G., 2000. Probability for Statisticians. Springer-Verlag, New York.
Van der Vaart, A.W., 1998. Asymptotic Statistics. Cambridge university press, New York.
Wooldridge, J.M., 1994. Estimation and Inference for Dependent Processes. In: Engle, R.F., McFadden, D.L. (Eds.), Handbook of Econometrics, vol IV. North-Holland Pub. Co., New York, pp. 2639–2738.
Xu, X., Lee, L., 2014. Maximum likelihood estimation of a spatial autoregressive Tobit model. Working paper, Department of Economics, The Ohio State University.