Smoothed spatial maximum score estimation of spatial autoregressive binary choice panel models

Jinghua Lei †

March 25, 2013

Preliminary, comments welcome

Abstract

This paper considers spatial autoregressive (SAR) binary choice models in the context of panel data with fixed effects, where the latent dependent variables are spatially correlated. Without imposing any parametric structure of the error terms, this paper proposes a smoothed spatial maximum score (SSMS) estimator which consistently estimates the model parameters up to scale. The identification of parameters is obtained, when the disturbances are time-stationary and the explanatory variables vary enough over time along with an exogenous and time-invariant spatial weight matrix. Consistency and asymptotic distribution of the proposed estimator are also derived in the paper. Finally, a Monte Carlo study indicates that the SSMS estimator performs quite well in finite samples.

JEL classification: C14 C21 C23 C25 R15

Keywords: Spatial Autoregressive Models, Binary Choice, Fixed Effects, Maximum Score Estimation.

*I am very grateful to my advisor Pavel Čížek for his enormously valuable advice, guidance and support. I also benefited from conversations with Zhuojiong Gan, Yufeng Huang, Tobias J. Klein, Hong Li, Geng Niu, Ruixin Wang, Wendun Wang, Yifang Yu. All errors are my responsibility.

†CentER and Department of Econometrics and Operation Research, Tilburg University, P.O. Box 90153, 5000 LE Tilburg, The Netherlands. E-mail: J.Lei@uvt.nl
1 Introduction

The spatial binary choice model has been increasing used in the spatial econometrics literature, especially for the spatial probit model. There are several different specifications of spatial binary choice models, the most popular specification for probit model is the spatial lag probit model: a linear regression model with endogenous interaction effects among the unobserved dependent variable

\[ Y_n^* = \lambda_0 W_n Y_n^* + X_n \beta_0 + \epsilon_n, \]  

(1)

where \( Y_n^* \) is an \( n \times 1 \) vector of latent dependent variable, \( X_n \) is an \( n \times q \) matrix of independent variables and \( \lambda_0 \) represents the spatial autoregressive coefficient. Endogenous interaction effects are typically considered as the formal specification for the equilibrium outcome of a spatial or social interaction process, in which the value of the dependent variable for one agent is jointly determined with that of neighboring agents. Many studies have considered this model from a methodological viewpoint: McMillen (1992), LeSage (2000), Pace and LeSage (2011) among others. Moreover, Klier and McMillen (2008) replace the probit by the logit specification. Most recently, Qu and Lee (2011) and Paul et al. (2012) conduct an important variant of the spatial lag probit model in the following form:

\[ Y_n^* = \lambda_0 W_n Y_n + X_n \beta_0 + \epsilon_n, \]

where the latent dependent variable \( Y_n^* \) depends on observed choices represented by \( W_n Y_n \) rather than unobserved ones. However, one of the basic problems of this interaction model is that the equilibrium may not be unique, so inference is only possible by assuming that one particular equilibrium occurs with probability one over the total number of equilibria.

Another specification is a linear regression model with spatially correlated errors:

\[ Y_n^* = X_n \beta_0 + v_n, \quad v_n = \rho_0 W_n v_n + \epsilon_n, \]

(2)

where \( v_n \) reflects the spatially correlated errors with coefficient \( \rho_0 \) and \( \epsilon_n \) follows a multivariate normal distribution with mean \( 0 \) and variance \( I_n \). In this model, the variance of
errors is usually normalized to one as it cannot be separately identified with the parameter $\beta_0$. This spatial error probit model has been studied by Beron and Vijverberg (2004), Fleming (2004), Klier and McMillen (2008) among others. Moreover, Bolduc, Fortin, and Gordon (1997) consider the logit specification in their empirical application such that the probability of $\Pr(y = 1)$ has an analytical solution.

The main assumption of model (1) and (2) is that the distribution of $\epsilon_n$ conditional on $X$ is known up to a finite set of parameters, for example, it is often assumed that $\epsilon_n$ has either the normal or logistic distribution. However, when the distribution of $\epsilon_n$ is misspecified, estimation methods that require specifying the distribution of $\epsilon_n$ yield inconsistent estimators. Furthermore, even if the model is correctly specified, likelihood based estimation methods may suffer from the multidimensional integration problem as the individual error terms are dependent on each other. Many attempts have been proposed to solve this problem, see Paul et al. (2012) for a carefully review.

Moreover, estimation would become much more difficult if a context of panel data with fixed effects is considered. Even if the distribution of the errors is correctly specified and there is no spatial dependence, consistently estimating parameters in binary panel models with fixed effect requires clever estimators, such as conditional logit estimation (Chamberlain, 1984) and maximum score estimator (Manski, 1987; Charlier, Melenberg, and van Soest, 1995). These methods could either generate a likelihood function without fixed effects or eliminate the fixed effects by some rank conditions. However, up to my best knowledge, whether these methods still work or not when there is spatial dependence is still unknown. Therefore, the purpose of this paper is to modify the maximum score estimator of Manski (1987) such that it can consistently estimate the parameters in spatial lag binary panel models with fixed effects.

In this paper, I consider a fixed effects spatial autoregressive (SAR) binary choice model which is a panel version of model (1), where the only assumption that imposed on the errors is time stationary rather than any parametric assumption. Based on this assumption and the exogeneity of time-invariant spatial weight matrix, a similar condition of Lemma 1 in Manski (1987) is derived in this paper. Therefore, a spatial maximum score estimator is defined analogously to that of Manski (1987) and can be smoothed by replacing the sign function with a continuous function as in Horowitz (1992). Although the
smoothed spatial maximum score (SSMS) estimator can not be extended to cross-sectional SAR binary choice models, it is applicable to fixed effects SAR binary choice models with arbitrarily spatial correlation in the errors, when such spatial correlation is time-invariant and satisfies some "fading memory" property as described in section 3.2.

The rest of the paper is organized as follows. Section 2 provides the model specification and the suggested smoothed spatial maximum score estimator. Section 3 proves identification, consistency and the asymptotic normality of the proposed estimator. Section 4 presents the results of a Monte Carlo investigation of the finite-sample properties of the estimator. All the proofs are provided in the appendices.

2 Spatial Autoregressive Binary Choice Models and SSMS

The SAR binary choice model is

\[ Y_{it}^* = \lambda_0 \sum_{j=1}^{n} w_{ij}Y_{jt}^* + X_{it}\beta_0 + \alpha_i + \epsilon_{it}, \quad i = 1, \ldots, n, t = 1, \ldots, T, \]  

(3)

where \( Y_{it}^* \) is the latent dependent variable that links to the observed binary outcome \( Y_{it} \) such that \( Y_{it} = 1 \) if \( Y_{it}^* > 0 \) and \( Y_{it} = 0 \) otherwise, \( X_{it} \) are the regressors for the individual \( i \) in the \( t \)-th period, \( W_n \) is a spatial weight matrix with \( ij \)-th element \( w_{ij} \) and is assumed to be constant across time, \( \lambda_0 \) is a parameter to capture the spatial effect, \( \alpha_i \) is the individual fixed effect which is unobserved and allowed to be correlated with the regressors in an arbitrary way, and \( \epsilon_{it} \) is the idiosyncratic individual error term with a common conditional distribution function, \( F_{\epsilon_n}(\cdot, \alpha, X) \) given \( (\alpha, X) \) with \( X = (X_{n1}, \ldots, X_{nT}) \). This spatial model is an equilibrium model.

For simplicity and without loss of generality, I consider the case when there are only two time periods (\( t = 1, 2 \)). Suppose that the inverse of matrix \( S_n(\lambda_0) = (I_n - \lambda_0 W_n) \) exists and denote \( S_n^{-1} = S_n^{-1}(\lambda_0) \), rearrange equation (3) and rewrite it in matrix notation, the equilibrium vector \( Y_{nt}^* \) is then

\[ Y_{nt}^* = (I_n - \lambda_0 W_n)^{-1}(X_{nt}\beta_0 + \alpha_n + \epsilon_{nt}) = S_n^{-1}(X_{nt}\beta_0 + \alpha_n + \epsilon_{nt}), \quad t = 1, 2. \]  

(4)

\footnote{Note that \( X_{it} \) consists of time-varying covariates as any time-invariant covariates would be absorbed into the fixed effect \( \alpha_i \).}
Under the assumption that inverse of matrix $S_n(\lambda_0)$ exists, $S_n^{-1}e_{nt}$ is a vector of linear combination of the error terms for all individuals. Let $e_i$ denote an $n \times 1$ vector with the $i$-th element equal to one and all other elements equal to zero, then $e_i^T S_n^{-1}$ is the $i$-th row of the matrix $(I_n - \lambda_0 W_n)^{-1}$. Denote $\tilde{e}_{nt} = e_i^T S_n^{-1} e_{nt}$, under the conditional stationarity assumption that $\epsilon_{n1}$ and $\epsilon_{n2}$ are identically distributed conditional on $(\alpha_n, X)$, we know that $\tilde{\epsilon}_{n1}$ and $\tilde{\epsilon}_{n2}$ also have the same distribution. Therefore, we obtain the following relationship for each individual $i$ as in Lemma 1 of Manski (1987) $^2$

\[
E[Y_{i1} - Y_{i2} | \alpha_n, X] > 0 \quad \text{if and only if} \quad e_i^T S_n^{-1} X_{n1} \beta_0 > e_i^T S_n^{-1} X_{n2} \beta_0,
\]

\[
E[Y_{i1} - Y_{i2} | \alpha_n, X] = 0 \quad \text{if and only if} \quad e_i^T S_n^{-1} X_{n1} \beta_0 = e_i^T S_n^{-1} X_{n2} \beta_0,
\]

\[
E[Y_{i1} - Y_{i2} | \alpha_n, X] < 0 \quad \text{if and only if} \quad e_i^T S_n^{-1} X_{n1} \beta_0 < e_i^T S_n^{-1} X_{n2} \beta_0.
\]

Mansi (1987) showed that under some regularity conditions, conditions (5) implies that the true parameter $\theta_0 = (\lambda_0, \beta_0^T)^T$ is the unique maximizer of

\[
G_i(\theta) = E[\Delta Y_i \operatorname{sgn}\{e_i^T S_n^{-1}(\lambda) \Delta X_n \beta\}], \quad i = 1, \ldots, n,
\]

where $\theta = (\lambda, \beta^T)^T$, $\Delta Y_i = (Y_{i1} - Y_{i2})$, $\Delta X_n = X_{n1} - X_{n2}$, $\operatorname{sgn}(x) = 1$ if $x \geq 0$ and $-1$ otherwise. Apparently, $\theta_0$ is also the unique maximizer of the average of equation (6), that is

\[
\theta_0 = \arg \max_\theta \frac{1}{n} \sum_{i=1}^n G_i(\theta) = \arg \max_\theta \frac{1}{n} \sum_{i=1}^n E[\Delta Y_i \operatorname{sgn}\{e_i^T S_n^{-1}(\lambda) \Delta X_n \beta\}]\cdot
\]

Consistently estimating parameter $\theta_0$ can be obtained by maximizing the following objective function (7):

\[
G_n(\theta) = \frac{1}{n} \sum_{i=1}^n \Delta Y_i \operatorname{sgn}\left(e_i^T S_n^{-1}(\lambda) \Delta X_n \beta\right).
\]

Observe that the behavior of $G_n(\cdot)$ is unaffected by removing observations having $Y_{i1} = Y_{i2}$,
thus, the estimator maximizing $G^*_n(\cdot)$ is a conditional maximum score estimator. However, it is difficult to derive its asymptotic distribution as the score function is a step function. Chamberlain (1986) has shown that there is no $n^{-1/2}$-consistent estimator of $\beta_0$ under Manski’s assumptions. Horowitz (1992) then modifies Manski’s maximum score estimator (Manski, 1985) by smoothing the score function to be continuous and differentiable, and shows that the convergence rate of the smoothed maximum score estimator is at least as fast as $n^{-2/5}$ and, depending on how smooth the distribution of $\epsilon_n$ and $X_n\beta_0$ are, can be arbitrarily close to $n^{-1/2}$. In the context of panel data models with fixed effect, Charlier, Melenberg, and van Soest (1995) investigate the smoothed version of Manski (1987)’s estimator and indicate that maximizing $G^*_n(\theta)$ boils down to maximizing

$$
\frac{1}{n} \sum_{i=1}^{n} \frac{1}{2} \Delta Y_i \left[ \text{sgn} \left( e_i^T S_n^{-1}(\lambda) \Delta X_n \beta \right) + 1 \right] = \frac{1}{n} \sum_{i=1}^{n} \Delta Y_i \mathbb{1} \{ e_i^T S_n^{-1}(\lambda) \Delta X_n \beta \geq 0 \} \quad (8)
$$

This objective function can then be smoothed by

$$
G_n(\theta; \sigma_n) = \frac{1}{n} \sum_{i=1}^{n} \Delta Y_i K \left( \frac{e_i^T S_n^{-1}(\lambda) \Delta X_n \beta}{\sigma_n} \right). \quad (9)
$$

where $K(u)$ is some smooth function that converges to the indicator function as $n \to \infty$, and $\sigma_n$ is a sequence of strictly positive real numbers satisfying $\lim_{n \to \infty} \sigma_n = 0$. Note that $K(u)$ could be a cumulative distribution function such as the cumulative standard normal distribution function $\Phi(u)$ as in Horowitz (1992), and $\sigma_n$ can be viewed as the bandwidth.

**Remark 1.** Apparently, when there is no spatial effect ($\lambda_0 = 0$), then the SSMS estimator that maximizes equation (9) degenerates to the standard smoothed maximum score estimator for panel models as in Charlier, Melenberg, and van Soest (1995).

Moreover, the identification and estimation strategy described above also works for models with arbitrarily spatially correlated errors if the spatial correlation is time stationary and satisfies the "fading memory" property as stated in the next section. For example, the mixed spatial lag and spatial error binary choices models with fixed effects:

$$
Y^*_{nt} = \lambda_0 W_{n,1} Y^*_{nt} + X_{nt}\beta_0 + \alpha_n + v_n, \quad v_n = \rho_0 W_{n,2} v_n + \epsilon_n.
$$

Suppose that the inverse of matrices $S_n(\lambda_0) = (I - \lambda_0 W_{n,1})$, $(I - \rho_0 W_{n,2})$ exists and
the spatial weighting matrices $W_{n,1}$ and $W_{n,2}$ are time-invariant, rearranging the above equation and rewrite it in matrix notation, we have

$$Y^*_{nt} = (I_n - \lambda_0 W_{n,1})^{-1} [X_{nt} \beta_0 + \alpha_n + (I_n - \rho_0 W_{n,2})^{-1} \epsilon_{nt}]$$

$$= S_n^{-1}(X_{nt} \beta_0 + \alpha_n) + S_n^{-1}(I_n - \rho_0 W_{n,2})^{-1} \epsilon_{nt}.$$ 

As in equation (5), when there are only two time periods ($t=1, 2$) and $\epsilon_{n1}$ and $\epsilon_{n2}$ are identically distributed conditional on $(\alpha_n, X)$, then $\epsilon_{n1}$ and $\epsilon_{n2}$ also have the same distribution, where $\epsilon_{nt} = [S_n^{-1}(I_n - \rho_0 W_{n,2})^{-1}] \epsilon_{nt}$ and $e_i S_n^{-1}(I_n - \rho_0 W_{n,2})^{-1}$ denotes the $i$-th row of the matrix $(I_n - \lambda_0 W_{n,1})^{-1}(I_n - \rho_0 W_{n,2})^{-1}$. Therefore, we could also obtain the same relationship for each individual $i$ as in equation (5). The identification and estimation strategy will be exactly the same as I discussed previously, however, the parameter $\rho_0$ can not be estimated in this case.

In addition, when $\beta_0 = 0$, the model degenerates to a spatial binary choice models without covariates. In this case, the spatial effect $\lambda_0$ is not identified without imposing additional assumptions on the error terms. Another point that we should notice is that the identification strategy described in this paper cannot be applied to the cross sectional spatial binary choice models directly.

Finally, when the time periods are more than two but finite, the SSMS estimator can be defined analogously by

$$\hat{\theta}_{nT} = \arg \max_{\theta} \frac{1}{nT(T-1)} \sum_{i=1}^{n} \sum_{s<t} (Y_{it} - Y_{ts}) K \left( \frac{e_i S_n^{-1}(\lambda)(X_{nt} - X_{ns})^\beta}{\sigma_n} \right).$$

**Remark 2.** First we note that the objective function of (9) is identical to the absolute loss objective function

$$\hat{\theta} = \arg \max_{\theta} \frac{1}{n} \sum_{i=1}^{n} \left| \Delta Y_i - K \left( \frac{e_i S_n^{-1}(\lambda)\Delta X_n^\beta}{\sigma_n} \right) \right|,$$

and the squared loss objective function

$$\hat{\theta} = \arg \max_{\theta} \frac{1}{n} \sum_{i=1}^{n} \left[ \Delta Y_i - K \left( \frac{e_i S_n^{-1}(\lambda)\Delta X_n^\beta}{\sigma_n} \right)^2 \right].$$
Khan (2012) proposes that when a standard normal distribution \( \Phi(\cdot) \) is applied for the kernel function, then we can define the spatial nonlinear least square (SNLLS) probit estimator as

\[
\hat{\theta}_{\text{SNLLS}} = \arg \max_\theta \frac{1}{n} \sum_{i=1}^{n} \left[ \Delta Y_i - \Phi \left( \frac{e_i^T S_n^{-1} \Delta X_n \beta}{\sigma_n} \right) \right]^2.
\]

The main advantage of this procedure is that the standard NLLS objective function can be extended to the case with spatial correlation, and the standard software packages, such as Stata, can be easily adjusted to compute the SNLLS estimator.

### 3 Identification and asymptotic properties of SSMS

#### 3.1 Identification

In this subsection, identification of parameters in model (3) is provided, and the definition of identification is similar as in Manski (1987). Consider \( (\lambda, \beta^T) \in \Lambda \times \mathbb{R}^q, (\lambda, \beta^T) \neq (\lambda_0, \beta_0^T) \), conditions (5) distinguishes \( (\lambda, \beta^T) \) from \( (\lambda_0, \beta_0^T) \) if there exists a set of \( \Delta X \) values having positive \( F_{\Delta X} \) probability such that condition (5) does not hold when \( (\lambda, \beta^T) \) is substituted for \( (\lambda_0, \beta_0^T) \). In this case, let

\[
V_{(\lambda, \beta)} = \left\{ \Delta X \in \mathbb{R}^q : \text{sgn}(e_i^T S_n^{-1}(\lambda) \Delta X_n \beta) \neq \text{sgn}(e_i^T S_n^{-1} \Delta X_n \beta_0) \right\}
\]

we say that \( (\lambda_0, \beta_0^T) \) is identified relative to \( (\lambda, \beta^T) \) if

\[
R(\lambda, \beta) \equiv \int_{V_{(\lambda, \beta)}} dF_{\Delta X} > 0
\]

**Assumption 1.**

i). \( F_{\epsilon_{i1}|X, \alpha_n} = F_{\epsilon_{i2}|X, \alpha_n} \) for all \( i \) and \( (X, \alpha_n) \).

ii). The support of \( F_{\epsilon_{i1}|X, \alpha_n} \) is \( \mathbb{R}^1 \) for all \( i \) and \( (X, \alpha_n) \).

**Assumption 2.**

i). The support of \( F_{\Delta X} \) is not contained in any proper linear subspace of \( \mathbb{R}^q \).

ii). There exists at least one \( q' \in \{1, 2, \ldots, q\} \) such that \( \beta_{0,q'} \neq 0 \), and for almost every value of \( \Delta \tilde{X}_i = (\Delta X_{i,1}, \Delta X_{i,2}, \ldots, \Delta X_{i,q'-1}, \Delta X_{i,q'+1}, \ldots, \Delta X_{i,q})^T \), the scalar random variable \( \Delta X_{i,q'} \) has everywhere positive Lebesgue density conditional on \( \Delta \tilde{X}_i \) for all \( i = 1, 2, \ldots, n \).
and conditional on $\Delta X_{j,t}$ for all $j \neq i$.

**Assumption 3.** The matrix $S_n(\lambda_0) = I_n - \lambda_0 W_n$ is nonsingular;

Assumptions 1 and 2 have the same forms as Assumptions 1 and 2 in Manski (1987), except that we have different conditionings. As in assumption 1 i), $\epsilon_{it}$ is stationary not only conditional on its own characteristics, but also conditional on the characteristics of other members. Such conditioning also appears in Assumption 2, and is necessary because there is spatial correlation between individuals, which is obvious if we assume the process $\{X_{it}, \alpha_i, \epsilon_{it}\}$ is strong mixing as in section 3.2. Assumption 1 ii) guarantees that the event $Y_{i1} \neq Y_{i2}$ occurs with positive probability for all $\alpha_n$. Assumption 2 i) is the familiar full-rank condition that prevents a global failure of identification, and part ii) is a substantive restriction, which implies that $\Delta X_n \beta$ has everywhere positive density for all $\beta$ such that $\beta_{i0} \neq 0$. Assumption 3 guarantees that the system (3) has an equilibrium and matrix $S_n(\lambda_0)$ is invertible.

Clearly, the scale of $\beta_0$ is not identified. To see this, we can simply set $\lambda = \lambda_0$, then the identification problem degenerates to that of Manski (1987). Identification of $\theta_0$ requires that there is a positive probability such that $\epsilon_i^T S_n^{-1}(\lambda) \Delta X_n \beta$ has a different sign with $\epsilon_i^T S_n^{-1} \Delta X_n \beta_0$. As Assumption 2 imposes no condition on the parameter vector $\theta_0$ except that $\beta_{i0} \neq 0$, it is possible for $\epsilon_i^T S_n^{-1}(\lambda) \Delta X_n \beta$ to have bounded support for all $\theta$, given sharper bounds on $\theta_0$. Therefore, $\theta_0$ is identified, which is stated in the following Lemma.

**Lemma 1.** Under Assumptions 1-3, $(\lambda_0, \beta_0^T)^T$ is identified relative to $(\lambda, \beta^T)^T \in \Lambda \times R^q$, where $\beta/||\beta|| \neq \beta_0/||\beta_0||$.

### 3.2 Consistency

In this subsection, consistency of estimators that maximize the objective function (9) is established. The main difficulty to prove the consistency is that the objective function (9) is based on a dependent and heterogenous process. Therefore, some appropriate "fading memory" property must be guaranteed to support laws of large numbers and uniform laws of large numbers, and the "fading memory" property that I will prove is near epoch dependence which is defined in Definition 1.

To proceed, we need to first define the space and metric (which are not restricted
to physical space and distance) for the convenience of analyzing the spatial correlation structure. Following Jenish and Prucha (2009, 2012), I consider spatial processes located on a (possibly) unevenly spaced lattice that satisfies the following assumption.

**Assumption 4.** The lattice $D \subseteq \mathbb{R}^d, d \geq 1$ is infinite countable. All elements in $D$ are located at distances of at least $d_0 > 0$ from each other, i.e., for all $l_i, l_j \in D : d(l_i, l_j) \geq d_0$, where $l_i$ denotes some location of corresponding unit $i$; without loss of generality, we assume that $d_0 = 1$.

The assumption of a minimum distance ensures the growth of the sample size as the sample regions $D_n = \{l_1, \ldots, l_n\} \subset D$ expand, which means the asymptotic methods that I employ in this paper are increasing domain asymptotics.

The models considered in this paper are actually the Cliff and Ord (1981) type models, which is one of the common approaches to model cross-sectional dependence in the econometrics literature. In the Cliff-Ord type models, the spatial weights $w_{ij,n}$ depend on some measure of distance $d_{ij}$ and decline as the distance increases. Under Assumption 3, model (4) is then $Y_{it} = 1\{e_i^TS_n^{-1}[X_{nt}\beta_0 + \alpha_n + \epsilon_{nt}] > 0\}$, where $e_i^TS_n^{-1}$ could be denoted by a vector $(a_{i1}, \ldots, a_{in})$. Although for fixed $n$, the output process $Y_{it}$ only depends on a finite number of elements of the input process $\eta_{it} = (X_{it}, \alpha_i, \epsilon_{it})^\top$, the mixing property of $\eta_{it}$ may not carry over to $Y_{it}$. The reason is that the number of elements composing the spatial lags grows unboundedly with the sample size so that the mixing property can break down in the limit. This is especially important when analyzing the asymptotic properties of Cliff-Ord type processes. Therefore, towards establishing that $\{Y_{it}, l_i \in D_n\}$ is near epoch dependent (NED) on $\{\eta_{it}, l_i \in D_n\}$, we maintain the following assumptions:

$$\lim_{d \to \infty} \sup_n \sup_{1 \leq i \leq n} \sum_{1 \leq j \leq n : d(l_i, l_j) > d} |a_{ij}| = 0$$ (12)

and

$$\sup_n \sup_{1 \leq i \leq n, t} \|\eta_{it}\|_p < \infty \quad \text{for some} \quad p \geq 1.$$ (13)
Jenish and Prucha (2012) show that a sufficient condition for (12) is that for some $\gamma > 0$,

$$\sup_n \sup_{1 \leq i \leq n} \sum_{j=1}^n |a_{ij}|d(l_i, l_j)^\gamma < \infty.$$ 

A similar condition has been used recently by Kelejian and Prucha (2007), and should be satisfied in a wide range of applications. It is slightly stronger than the typical assumption in the Cliff-Ord literature which imposes that the row and column sums of the absolute elements of the matrix $S_n^{-1}$ are uniformly bounded as in Assumption 6 ii).

Now I define the near epoch denpendence (NED) of random variables $Y_{it}$ based on a process of random variables $\eta_{it}$ as follows:

**Definition 1.** Random variables $Y_{it}, i = 1, \ldots, n, t = 1, 2$ are called near epoch dependent on $\eta_{it}$ if

$$\sup_i ||Y_{it} - E(Y_{it} | \mathcal{F}_{i,n}(m))||_2 = d_i \nu(m) \to 0, \quad \text{as} \quad m \to \infty$$

where $d_i$ is a sequence of positive constant and $\mathcal{F}_{i,n}(m) = \sigma(\eta_{jt,n} : d(l_i, l_j) \leq m)$ is the $\sigma$ field generated by the random variables $\eta_{jt,n}$ located in the $m$-neighborhood of location $i$.

The idea behind the near epoch dependence condition is that given the $m$-neighborhood of imput variables $\eta_{it}, Y_{it}$ should be predictable up to arbitrary accuracy. That is, the approximation error declines "sufficiently fast" as the conditioning set of input variables expands. The base process $\eta_{it}$ needs to satisfy a condition such as strong or uniform mixing or independence.

**Assumption 5.** $\{\eta_{it}\}, i = 1, \ldots, n, t = 1, 2$, is a strict stationary strong mixing process with $\alpha$-mixing coefficient $\alpha(m)$.

**Proposition 1.** Under Assumptions 1, 3-5 and conditions (12)-(13), the process $\{Y_{it}\}$, $\{\Delta Y_t\}$, and $\{\text{sgn}(e_i^T S_n^{-1}(\lambda) \Delta X_n \beta)\}$ are uniformly NED on the process $\{\eta_{it}\}$.

**Remark 3.** Proposition 1 shows that $Y_{it}$ is a sequence of $0/1$ valued random variables that is near epoch dependent on $\eta_{it}$. Then $(Y_{it}, \eta_{it})$ is strong mixing by Theorem 2 of de Jong and Woutersen (2011), and the mixing property of $(Y_{it}, \eta_{it})$ will be used in the proofs for consistency and asymptotic normality of the smoothed spatial maximum score estimator.
Although \( \{Y_{it}\} \) is strong mixing, it is not stationary as the inverse spatial weights \( e_i^T S_n^{-1}(\lambda) \) are different for each individual \( i \) in general. One example for \( \{Y_{it}\} \) to be stationary is where the spatial correlation only exists within groups of the same size, and equal weights are assigned for individuals in the same group.

**Assumption 6.** i). \( \beta_0 = (\beta_{0,1}, \ldots, \beta_{0,q-1})^T \) is contained in a compact subset \( B \) of \( \mathbb{R}^{q-1} \);  

ii). The sequence \( \{W_n\} \) and \( \{S_n^{-1}\} \) are uniformly bounded in both row and column sums; \(^3\) 

iii). \( \{S_n^{-1}(\lambda)\} \) are uniformly bounded in either row or column sums, uniformly in \( \lambda \) in a compact parameter space \( \Lambda \). The true parameter \( \lambda_0 \) is in the interior of \( \Lambda \).

The uniform boundedness condition of \( S_n^{-1}(\lambda) \) in Assumption 6 ii) implies that \( S_n^{-1}(\lambda) \) are uniformly bounded in both row and column sums uniformly in a neighborhood of \( \lambda_0 \) (Lee, 2004). Assumption 6 i) and iii) are needed to deal with the nonlinearity of \( K (e_i^T S_n^{-1}(\lambda) \Delta X_n \beta / \sigma_n) \) as a function of \( \lambda \) and \( \beta \). The parameter space \( \Lambda \times \mathbb{R}^q \) is usually assumed to be a compact convex subset of \( \mathbb{R}^{q+1} \) for a nonlinear extremum estimation. This assumption is required for the uniform convergence of the sample average objective function in the proof of consistency Amemiya (1985). However, Wang and Lee (2012) mention that relaxation of this assumption would be an important issue of future research as it does not cover leading specification for the parameter space of \( \lambda \), which is often taken to be an open set, e.g., \((-1, 1)\).

Under Assumptions 1-6, the following theorem shows the consistency of the smoothed spatial maximum score estimator.

**Theorem 1.** Let Assumptions 1-6 hold. Let \( \theta_n \) be a solution to

\[
\max_{\theta} G_n(\theta; \sigma_n),
\]

\(^3\) The notions of uniform boundedness can be defined in terms of some matrix norms: the maximum column matrix norm \( \| \cdot \|_1 \) of a \( n \times n \) matrix \( A = (a_{ij}) \) is defined as \( \|A_n\|_1 = \max_{1 \leq j \leq n} \sum_{i=1}^{n} |a_{ij}| \), and the maximum row sum matrix norm \( \| \cdot \|_\infty \) is \( \|A_n\|_\infty = \max_{1 \leq i \leq n} \sum_{j=1}^{n} |a_{ij}| \) (see Horn and Johnson (1985), pp.294-295). The uniformly boundedness of \( \{A_n\} \) in column (resp. row) sums is equivalent to the sequence \( \{\|A_n\|_1\} \) (resp. \( \{\|A_n\|_\infty\} \) ) being bounded.

Lemma A.2 of Lee (2004) shows that, for any weights matrix, \( \|\lambda_0 W_n\|_1 < 1 \) and \( \|\lambda_0 W_n\|_\infty < 1 \) for all \( n \), are sufficient conditions for \( S_n^{-1} \) to be uniformly bounded in both row and column sums.

Because a matrix norm \( \| \cdot \| \) has the submultiplicative property that \( \|A_n B_n\| \leq \|A_n\| \cdot \|B_n\| \), Assumption 6 guarantees that products of matrices in our analysis such as \( S_n^{-1} W_n S_n^{-1} \) and \( S_n^{-1} W_n S_n^{-1} W_n S_n^{-1} \), etc., will be uniformly bounded in row and column sums.
where \( G_n(\theta; \sigma_n) = \frac{1}{n} \sum_{i=1}^{n} \Delta Y_i K(\frac{e_i^T S_n^{-1}(\lambda) \Delta X_n \beta}{\sigma_n}) \) and \( \theta = (\lambda, \beta^T)^T \), then \( \theta_n \to^p \theta_0 \). If in addition the strong mixing coefficients satisfy \( \alpha(m) \leq Cm^{-r} \) for positive constants \( C \) and \( r \), then \( \theta_n \to^{a.s.} \theta_0 \).

### 3.3 Asymptotic normality

In this subsection, the asymptotic normality of the smoothed spatial maximum score estimator is established, and the approach is analogous to that of Horowitz (1992) and de Jong and Woutersen (2011) except that the asymptotic properties are built on a dependent and heterogeneous process while the process in Horowitz (1992) is i.i.d and the process in de Jong and Woutersen (2011) is dependent but stationary.

Let Assumptions 1–3 hold, and suppose \( K(\cdot) \) is twice differentiable everywhere. Then \( G_n(\theta; \sigma_n) \) is twice differentiable with respect to \( \tilde{\theta} = (\lambda, \tilde{\beta}^T)^T \), where \( \tilde{\beta} = (\beta_1, \ldots, \beta_{q-1})^T \). Assumption 6 ensures that \( \tilde{\theta}_0 \) is an interior point of \( \bar{\Theta} \). Define \( T_n(\theta; \sigma_n) = \partial G_n(\theta; \sigma_n)/\partial \tilde{\theta} \), and \( Q_n(\theta; \sigma_n) = \partial^2 G_n(\theta; \sigma_n)/\partial \tilde{\theta} \partial \tilde{\theta}^T \). Let \( \theta_n \equiv (\tilde{\theta}_n^T, \beta_{n,q})^T \) denote a solution to problem \( (15) \), then with probability approaching 1 as \( n \to \infty \), \( \tilde{\theta}_n \) is an interior point of \( \bar{\Theta} \), \( \beta_{n,q} = \beta_{0,q} = \pm 1 \), and \( T_n(\theta_n; \sigma_n) = 0 \). A Taylor series expansion of \( T_n(\theta_n; \sigma_n) \) yields

\[
T_n(\theta_n; \sigma_n) = T_n(\theta_0; \sigma_n) + Q_n(\theta_n^*; \sigma_n)(\tilde{\theta}_n - \tilde{\theta}_0) = 0,
\]

where \( \theta_n^* \) is between \( \theta_n \) and \( \theta_0 \). Similar to Horowitz (1992), if there is a real function \( \rho(n) \) such that \( \rho(n)T_n(\theta_0; \sigma_n) \) converges in distribution as \( n \to \infty \), and suppose \( Q_n(\theta_n^*; \sigma_n) \) converges in probability to a nonsingular and nonstochastic matrix \( Q \). Then

\[
\rho(n)(\tilde{\theta}_n - \tilde{\theta}_0) = -Q^{-1}\rho(n)T_n(\theta_0; \sigma_n) + o_p(1).
\]

Thus, we know that \( \tilde{\theta}_n - \tilde{\theta}_0 \) converges to 0 at the rate of \( \rho(n)^{-1} \) and \( \rho(n)(\tilde{\theta}_n - \tilde{\theta}_0) \) is distributed asymptotically as \( -Q^{-1}\rho(n)T_n(\theta_0; \sigma_n) \).

Let \( z_i = e_i^T S_n^{-1} \Delta X_n \beta_0 = e_i^T S_n^{-1} \Delta \hat{X}_n \hat{\beta}_0 + e_i^T S_n^{-1} \Delta X_{n,q} \), then there is a one-to-one relation between \( (\Delta \hat{X}_n, Z_n) \) and \( \Delta X_n \) for any fixed \( \theta_0 \), where \( Z_n = (z_1, \ldots, z_n)^T \). Denote \( Z_{-i} = (z_1, \ldots, z_{i-1}, z_{i+1}, \ldots, z_n)^T \) and \( Z_i = \{ \Delta \hat{X}_n, Z_{-i} \} \). By Assumption 2, the distribution of \( z_i \) conditional on \( Z_i \) has everywhere positive density with respect to Lebesgue
measure for almost every \( \tilde{Z}_i \) and \( i = 1, \ldots, n \). Let \( p_i(z_i | \tilde{Z}_i) \) denote this density. For each positive integer \( j \), define \( p_i^{(j)}(z_i | \tilde{Z}_i) = \partial^j p_i(z_i | \tilde{Z}_i) / \partial z_i^j \) whenever the derivative exists, and define \( p_i^{(0)}(z_i | \tilde{Z}_i) = p_i(z_i | \tilde{Z}_i) \). Let \( P_i \) denote the cumulative distribution function of \( \tilde{Z}_i \), and let \( F_i(\cdot | z_i, \tilde{Z}_i) \) denote the cumulative distribution of \( \tilde{\varepsilon}_i = e_i^\top S_n^{-1} (c_{i1} - c_{i2}) \) conditional on \( z_i \) and \( \tilde{Z}_i \). For each positive integer \( j \), define \( F_i^{(j)}(\cdot | z_i, \tilde{Z}_i) = \partial^j F_i^{(j)}(\cdot | z_i, \tilde{Z}_i) / \partial z_i^j \) whenever the derivative exists. Define the scalar constants \( \alpha_A \) and \( \alpha_D \) by \( \alpha_A = \int_{-\infty}^{\infty} v^h K'(v)dv \) and \( \alpha_D = \int_{-\infty}^{\infty} [K'(v)]^2 dv \) whenever these quantities exist. For each integer \( h \geq 2 \), define the \( q \times 1 \) vector \( A \) and the \( q \times q \) matrices \( D \) and \( Q \) by

\[
A = -2\alpha_A \lim_{n \to \infty} \frac{1}{n} \sum_{i=1}^{n} \sum_{k=1}^{h} \left( \frac{1}{k!(h-k)!} E \left[ F_i^{(k)}(0|0, \tilde{Z}_i)p_i^{(k)}(z_i | \tilde{Z}_i) \tilde{B}_{1,i} \right] \Pr(Y_{i1} \neq Y_{i2}) \right),
\]

\[
D = \alpha_D \lim_{n \to \infty} \frac{1}{n} \sum_{i=1}^{n} E \left[ p_i(0|\tilde{Z}_i) \tilde{B}_{1,i} \tilde{B}_{1,i}^\top \right] \Pr(Y_{i1} \neq Y_{i2}),
\]

\[
Q = \lim_{n \to \infty} \frac{2}{n} \sum_{i=1}^{n} E \left[ F_i^{(1)}(0|0, \tilde{Z}_i)p_i(0|\tilde{Z}_i) \tilde{B}_{2,i} \right] \Pr(Y_{i1} \neq Y_{i2}),
\]

where \( \tilde{B}_{1,i} = \left( e_i^\top S_n^{-1} W_n S_n^{-1} \Delta X_n \beta_0, e_i^\top S_n^{-1} \Delta \tilde{X}_n \right)^\top \) and

\[
\tilde{B}_{2,i} = \begin{pmatrix}
(e_i^\top S_n^{-1} W_n S_n^{-1} \Delta X_n \beta_0)^2 & e_i^\top S_n^{-1} W_n S_n^{-1} \Delta X_n \beta_0 e_i^\top S_n^{-1} \Delta \tilde{X}_n \\
& \Delta \tilde{X}_n [S_n^{-1}]^\top e_i e_i^\top S_n^{-1} \Delta \tilde{X}_n
\end{pmatrix}.
\]

**Assumption 7.** i) The \( \alpha \)-mixing coefficient satisfies \( \alpha(m) \leq C m^{-(2s-2)/(s-2)-\gamma} \) for some \( \gamma > 0 \).

ii) For some sequence \( m_n \geq 1 \) and when \( n \to \infty \),

\[
\sigma_n^{-3(p+q-1)} \sigma_n^{-2(n^{1/s} \alpha(m_n) + \sigma_n^{-2(p+q-1)/2} n^{2/s} \alpha(m_n) + |\log(n m_n)| \left( n^{1-4/s} \sigma_n^{-4} m_n^{-2} \right)^{-1} \to 0.
\]

**Assumption 8.** For all vectors \( \xi \) such that \( |\xi| = 1 \), \( E|\xi^\top \tilde{B}_{1,i}|^s < \infty \) for some \( s > 4 \) and all \( i \).

These two assumptions are identical to Assumptions 6 and 7 of de Jong and Woutersen (2011). They strengthen the fading memory conditions of Assumption 5 in order to establish asymptotic normality.
The following assumptions are analogous to Assumptions 7-11 of Horowitz (1992):

**Assumption 9.**

i) $K(\cdot)$ is twice differentiable everywhere, $|K'(\cdot)|$ and $|K''(\cdot)|$ are uniformly bounded, and each of the following integrals over $(-\infty, \infty)$ is finite: $\int |K'(v)|^2 dv$, $\int |K''(v)|^2 dv$, $\int |v^2 K''(v)| dv$.

ii) For some integer $h \geq 2$ and each integer $k (1 \leq k \leq h)$, $\int |v^k K'(v)| dv < \infty$, and

\[
\int_{-\infty}^{\infty} v^k K'(v) dv = \begin{cases} 
0 & \text{if } k < h, \\
(\text{nonzero}) & \text{if } k = h.
\end{cases}
\]

iii) For any integer $k$ between 0 and $h$, any $\gamma > 0$, and any sequence $\{\sigma_n\}$ converging to 0,

\[
\lim_{n \to \infty} \sigma_n^{k-h} \int_{|\sigma_n v| > \gamma} |v^k K'(v)| dv = 0, \quad \lim_{n \to \infty} \sigma_n^{-1} \int_{|\sigma_n v| > \gamma} |K''(v)| dv = 0.
\]

**Assumption 10.** For all $i$ and each integer $k$ such that $1 \leq k \leq h - 1$, all $z_i$ in a neighborhood of 0, almost every $\tilde{Z}_i$, and some $M < \infty$, $p_i^{(k)}(z_i|\tilde{Z}_i)$ exists and is a continuous function of $z_i$ satisfying $p_i^{(k)}(z_i|\tilde{Z}_i) < M$. In addition, $|p_i(z_i|\tilde{Z}_i)| < M$ for all $z_i$ and almost every $\tilde{Z}_i$.

**Assumption 11.** For all $i$ and each integer $k$ such that $1 \leq k \leq h$, all $z_i$ in a neighborhood of 0, almost every $\tilde{Z}_i$, and some $M < \infty$, $F_i^{(k)}(-z_i|z_i, \tilde{Z}_i)$ exists and is a continuous function of $z_i$ satisfying $F_i^{(k)}(-z_i|z_i, \tilde{Z}_i) < M$.

**Assumption 12.** The true parameter $\tilde{\theta}_0$ is an interior point of $\tilde{\Theta}$.

**Assumption 13.** The matrix $Q$ is negative definite.

In addition to the above assumptions, we still need the following two assumptions that similar to Assumptions 13 and 14 in de Jong and Woutersen (2011). The first assumption is needed to ensure proper behavior of covariance terms, and the second assumption on $K''(\cdot)$ is needed to formally show a uniform law of large numbers for the second derivative of the objective function.

**Assumption 14.** The conditional joint density $p(z_i, z_j|\tilde{Z}_i, \tilde{Z}_j)$ exists and is continuous at $(z_i, z_j) = (0, 0)$ for all $i \neq j$. 

15
Assumption 15. $K''(\cdot)$ satisfies, for some $\mu \in (0, 1]$ and $L \in [0, \infty)$ and all $x, y \in \mathbb{R}$,

$$|K''(x) - K''(y)| \leq L|x - y|^\mu.$$ 

The main results concerning the asymptotic distribution of the smoothed spatial maximum score estimator are given by the following theorem.

**Theorem 2.** Let Assumptions 1-15 hold for some $h \geq 2$, then

(a) If $n\sigma_n^{2h+1} \to \infty$ as $n \to \infty$, $\sigma_n^{-h}(\hat{\theta}_n - \tilde{\theta}_0) \to^p -Q^{-1}A$.

(b) If $n\sigma_n^{2h+1} \to \infty$ has a finite limit $\kappa$ as $n \to \infty$, then

$$\sigma_n^{-h}(\hat{\theta}_n - \tilde{\theta}_0) \to^d N(-\kappa^{1/2}Q^{-1}A, Q^{-1}DQ^{-1}).$$

In order to make the results of Theorem 2 useful in applications, the next theorem shows how $A, D$ and $Q$ could be consistently estimated from observations of $(Y_{nt}, X_{nt}, W_n)$.

**Theorem 3.** Let $\hat{\theta}_n$ be a consistent smoothed spatial maximum score estimator based on $\sigma_n$ such that $\sigma_n \propto n^{-(2h+1)}$. For $\theta \in \{-1, 1\} \times \hat{\Theta}$, define

$$t_i(\theta, \sigma) = \mathbb{1}\{Y_{i1} \neq Y_{i2}\}(2 \cdot \mathbb{1}\{Y_{i1} = 1, Y_{i2} = 0\} - 1)B_i^{(1)}(\theta, \sigma),$$

where $B_i^{(1)}(\theta, \sigma)$ is defined in Appendix A. Let $\sigma_n^*$ be such that $\sigma_n^* \propto n^{-\delta/(2h+1)}$, where $0 < \delta < 1$. Then: (a) $\hat{A} = (\sigma_n^*)^{-h}T_n(\theta_n, \sigma_n^*)$ converges in probability to $A$; (b) the matrix

$$\hat{D}_n = \frac{\sigma_n}{n} \sum_{i=1}^n t_i(\theta_n, \sigma_n)t_i(\theta_n, \sigma_n)^\top$$

converges in probability to $D$; (c) $Q_n(\theta_n, \sigma_n)$ converges in probability to $Q$.

4 Simulation results: NOT COMPLETE

To investigate the finite sample properties of our estimator by a Monte Carlo study, I focus on the spatial scenario in Case (1991) and Lee (2004) with an $R$ number of districts, $m$ members in each district, and with each neighbor of a member in a district given equal weight, i.e., $W = I_R \otimes B_m$, where $B_m = (l_ml_m' - I_m)/(m - 1)$ and $l_m$ is an $m$-dimensional
Table 1: simulation results for equation (9)

<table>
<thead>
<tr>
<th></th>
<th>$m = 10$</th>
<th>$m = 20$</th>
<th>$m = 40$</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Bias</td>
<td>MSE</td>
<td>Bias</td>
</tr>
<tr>
<td>$\beta_{SSMS}, R = 30$</td>
<td>0.09</td>
<td>0.30</td>
<td>0.11</td>
</tr>
<tr>
<td>$\lambda$</td>
<td>-0.44</td>
<td>0.69</td>
<td>-0.51</td>
</tr>
<tr>
<td>$\beta_{SMS}$</td>
<td>0.11</td>
<td>0.25</td>
<td>0.14</td>
</tr>
<tr>
<td>$\beta_{SSMS}, R = 60$</td>
<td>0.14</td>
<td>0.24</td>
<td>0.13</td>
</tr>
<tr>
<td>$\lambda$</td>
<td>-0.35</td>
<td>0.43</td>
<td>-0.44</td>
</tr>
<tr>
<td>$\beta_{SMS}$</td>
<td>0.14</td>
<td>0.20</td>
<td>0.14</td>
</tr>
<tr>
<td>$\beta_{SSMS}, R = 90$</td>
<td>0.10</td>
<td>0.22</td>
<td>0.10</td>
</tr>
<tr>
<td>$\lambda$</td>
<td>-0.27</td>
<td>0.29</td>
<td>-0.39</td>
</tr>
<tr>
<td>$\beta_{SMS}$</td>
<td>0.09</td>
<td>0.14</td>
<td>0.10</td>
</tr>
<tr>
<td>$\beta_{SSMS}, R = 120$</td>
<td>0.13</td>
<td>0.20</td>
<td>0.13</td>
</tr>
<tr>
<td>$\lambda$</td>
<td>-0.26</td>
<td>0.25</td>
<td>-0.33</td>
</tr>
<tr>
<td>$\beta_{SMS}$</td>
<td>0.10</td>
<td>0.13</td>
<td>0.08</td>
</tr>
</tbody>
</table>

The model in the study is a binary spatial process

$$Y_{it}^* = \lambda_0 \sum_{j=1}^{n} w_{ij} Y_{jt}^* + X_{1it} + X_{2it} \beta_0 + \alpha_i + \epsilon_{it}$$  \hspace{1cm} (18)

for $t = 1$ and 2, where $X_{1it}$ is drawn from the standard normal distribution $N(0,1)$, and $X_{2it}$ from the chi-square distribution with one degree of freedom, normalized to have zero mean and unit variance, independent of each other, $\alpha_i = \frac{1}{2}(X_{2i1} + X_{2i2}) + \eta_i$ with $\eta_i$ is from $N(0,1)$ independent of other variables, and $\epsilon_{it}$ is drawn from $N(0,1)$ independent of $(X_{1it}, X_{2it})$, the observed dependent variable $Y_{it}$ is generated by $Y_{it} = 1$ if $Y_{it}^* > 0$ and $Y_{it} = 0$ otherwise. Here the coefficient of $X_{1it}$ is set to 1, and we are considering the estimation of true parameters $\lambda_0 = 0.5$ and $\beta_0 = 1$.

I have experimented with different values of $R$ from 30 to 120 and $m$ from 10 to 40. For each case, there are 500 repetitions.\(^5\) The standard smoothed maximum score estimator $\beta_{SMS}$ is also considered for comparison.

\(^4\)In this example, $h_n = (m - 1)$ and $h_n/n = (m - 1)/(mR) = O(1/R)$. If sample size $n$ increases by increasing both $R$ and $m$, then $h_n$ goes to infinity and $h_n/n$ goes to zero as $n$ tends to infinity. Whether $\{h_n\}$ is a bounded or divergent sequence has interesting implications on the least square approach. The least squares estimators of $\beta$ and $\rho$ are inconsistent when $\{h_n\}$ is bounded, but they can be consistent when $\{h_n\}$ is divergent.

\(^5\)R takes values: 30, 60, 90, 120, and $m$ takes values: 10, 20, 40.
Table 2: (9) with \((I_n - \lambda_0 W_n)^{-1}\) replaced by \(I_n + \lambda_0 W_n + \lambda_0^2 W_n^2 + \lambda_0^3 W_n^3\)

<table>
<thead>
<tr>
<th></th>
<th>m = 10</th>
<th>m = 20</th>
<th>m = 40</th>
</tr>
</thead>
<tbody>
<tr>
<td>(\beta_{SSMS}, R = 30)</td>
<td>Bias</td>
<td>MSE</td>
<td>Bias</td>
</tr>
<tr>
<td>(\beta_{SSMS}, R = 60)</td>
<td>0.14</td>
<td>0.21</td>
<td>0.10</td>
</tr>
<tr>
<td>(\beta_{SSMS}, R = 90)</td>
<td>-0.03</td>
<td>0.15</td>
<td>-0.11</td>
</tr>
<tr>
<td>(\beta_{SSMS}, R = 120)</td>
<td>0.14</td>
<td>0.16</td>
<td>0.10</td>
</tr>
</tbody>
</table>

4.1 matlab package devec3.m is used for optimization

Table 1 reports the simulation results for the estimators that maximizing equation 9. Although the estimates of smoothed spatial maximum score estimator \(\beta_{SSMS}\) are more or less the same as the standard smoothed maximum score estimator \(\beta_{SMS}\), the bias of the \(\lambda_0\) is really large and unnormal. That is might because it is difficult to estimate \(\lambda_0\) because it is in \((I_n - \lambda_0 W_n)^{-1}\). Therefore, another experiment is considered such that equation (9) with \((I_n - \lambda_0 W_n)^{-1}\) replaced by \(I_n + \lambda_0 W_n + \lambda_0^2 W_n^2 + \lambda_0^3 W_n^3\), and the simulation results are reported in Table 2. As we can see, the estimates for \(\lambda_0\) performs much better than the results in Table 1, especially for the case when \(m = 10\) where the bias for \(\lambda_0\) is really small. However, for other cases, the results are still undesirable.

I think the bandwidth would be different from the optimal bandwidth in standard maximum score estimation and the estimates could also be influenced by the range of the covariates because the identification requires there is at least one covariate with everywhere positive probability, so I have tried several ad-hoc bandwidths and let the standard deviation of \(X1 = 5\). Here I only consider the experiment when \(m = 40\) and \(R = 60\), the experiment is repeated for 500 times. The results are reported in Table 3. We can see that the smoothed spatial maximum score estimator has a smaller bias but larger mean square

\[\text{Table 3: Simulation results for the estimators that maximizing equation 9.}\]

\[\beta_{SSMS}, R = 30\] | Bias | MSE | Bias | MSE | Bias | MSE |
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>0.08</td>
<td>0.28</td>
<td>0.13</td>
<td>0.25</td>
<td>0.13</td>
<td>0.19</td>
<td></td>
</tr>
<tr>
<td>-0.01</td>
<td>0.40</td>
<td>-0.20</td>
<td>0.38</td>
<td>-0.29</td>
<td>0.45</td>
<td></td>
</tr>
<tr>
<td>0.12</td>
<td>0.27</td>
<td>0.13</td>
<td>0.21</td>
<td>0.13</td>
<td>0.17</td>
<td></td>
</tr>
<tr>
<td>0.11</td>
<td>0.23</td>
<td>0.11</td>
<td>0.19</td>
<td>0.13</td>
<td>0.16</td>
<td></td>
</tr>
<tr>
<td>0.01</td>
<td>0.26</td>
<td>-0.20</td>
<td>0.27</td>
<td>-0.26</td>
<td>0.33</td>
<td></td>
</tr>
<tr>
<td>0.13</td>
<td>0.21</td>
<td>0.11</td>
<td>0.17</td>
<td>0.12</td>
<td>0.13</td>
<td></td>
</tr>
<tr>
<td>0.14</td>
<td>0.21</td>
<td>0.10</td>
<td>0.15</td>
<td>0.15</td>
<td>0.15</td>
<td></td>
</tr>
<tr>
<td>-0.03</td>
<td>0.15</td>
<td>-0.11</td>
<td>0.18</td>
<td>-0.20</td>
<td>0.22</td>
<td></td>
</tr>
<tr>
<td>0.14</td>
<td>0.16</td>
<td>0.10</td>
<td>0.10</td>
<td>0.09</td>
<td>0.09</td>
<td></td>
</tr>
<tr>
<td>0.11</td>
<td>0.18</td>
<td>0.11</td>
<td>0.17</td>
<td>0.07</td>
<td>0.11</td>
<td></td>
</tr>
<tr>
<td>-0.05</td>
<td>0.12</td>
<td>-0.16</td>
<td>0.20</td>
<td>-0.22</td>
<td>0.23</td>
<td></td>
</tr>
<tr>
<td>0.10</td>
<td>0.12</td>
<td>0.09</td>
<td>0.11</td>
<td>0.06</td>
<td>0.09</td>
<td></td>
</tr>
</tbody>
</table>

Here we implicitly make some assumption about \(\lambda_0 W_n\), as we need \((I_n - \lambda_0 W_n)^{-1} = (I_n + \lambda_0 W_n + \lambda_0^2 W_n^2 + \cdots + \lambda_0^n W_n^n)\) when \(n \to \infty\). This condition holds if the right side converges, which is true if and only if all of the eigenvalues of \(\lambda_0 W_n\) have absolute value smaller than 1.
Table 3: Results for different bandwidths and with St.D of $X_1 = 5$

<table>
<thead>
<tr>
<th>$h$</th>
<th>$\beta_{SSMS}$ Bias</th>
<th>$\lambda$ MSE</th>
<th>$\beta_{SMS}$ Bias</th>
<th>$\lambda$ MSE</th>
</tr>
</thead>
<tbody>
<tr>
<td>$h_1 = N^{-1/5}$</td>
<td>0.0018 0.1768</td>
<td>-0.1091 0.0492</td>
<td>0.0106 0.0929</td>
<td></td>
</tr>
<tr>
<td>$h_2 = N^{-1/5}/3$</td>
<td>0.0066 0.2087</td>
<td>-0.1165 0.0584</td>
<td></td>
<td></td>
</tr>
<tr>
<td>$h_3 = N^{-1/5}/5$</td>
<td>-0.0263 0.1983</td>
<td>-0.1121 0.0578</td>
<td></td>
<td></td>
</tr>
<tr>
<td>$h_4 = N^{-1/5}/7$</td>
<td>0.0195 0.2024</td>
<td>-0.1020 0.0516</td>
<td></td>
<td></td>
</tr>
<tr>
<td>$h_5 = N^{-1/5}/9$</td>
<td>0.0042 0.1799</td>
<td>-0.1159 0.0584</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table 4: Results for three cases

<table>
<thead>
<tr>
<th>Case 1</th>
<th>Case 2</th>
<th>Case 3</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\lambda$ Bias</td>
<td>MSE</td>
<td>Bias</td>
</tr>
<tr>
<td>$\beta_{SSMS}$</td>
<td>-0.02 0.02</td>
<td>0.00 0.01</td>
</tr>
</tbody>
</table>

error than the smoothed maximum score estimator.

4.2 Grid search

The aim of using grid search is to get a first insight of the performance of our spatial maximum score estimator. As it is a simulation study rather than the empirical estimation, we know the true parameter in advance, so we can always set suitable intervals for grid search such that they do contain the true parameters.

The advantage of grid search is that the maximum of objective function is reached or approximately reached, although it is really time consuming. Therefore, I only consider the experiment when $m = 40$ and $R = 60$, but with three different objective functions:

Case 1: original objective function with $\text{sgn}$,

Case 2: $\text{sgn}$ replaced by normal kernel,

Case 3: $\text{sgn}$ replaced by normal kernel and $(I_n - \lambda_0 W_n)^{-1}$ replaced by $I_n + \lambda_0 W_n + \lambda_0^3 W_n^2 + \lambda_0^3 W_n^3$.

The interval of grid search for parameters $\lambda$ and $\beta$ are $[0.05, 0.95]$ and $[0.55, 1.45]$, respectively. There are 90 grids and the experiment is repeated for 100 times. The results are reported in Table 4.
Appendix A: Notations

Let $S_n^{-1} = S_n^{-1}(\lambda_0)$.

As $\frac{\partial S_n^{-1}}{\partial \lambda} = -S_n^{-1} \frac{\partial S}{\partial \lambda} S_n^{-1}$, we know that

$$\frac{\partial e_i^T S_n^{-1}(\lambda)}{\partial \lambda} = e_i^T S_n^{-1}(\lambda) W_n S_n^{-1}(\lambda);$$

Denote $B_i = K\left(\frac{e_i^T S_n^{-1}(\lambda) \Delta X_n}{\sigma_n}\right) = K\left(\frac{z_i}{\sigma_n}\right)$ where $z_i = e_i^T S_n^{-1}(\lambda) \Delta X_n$, then

$$\frac{\partial B_i}{\partial \lambda} = K'\left(\frac{z_i}{\sigma_n}\right) \frac{e_i^T S_n^{-1}(\lambda) W_n S_n^{-1}(\lambda) \Delta X_n}{\sigma_n};$$

$$\frac{\partial B_i}{\partial \beta^T} = K'\left(\frac{z_i}{\sigma_n}\right) \frac{e_i^T S_n^{-1}(\lambda) \Delta \tilde{X}_n}{\sigma_n};$$

$$B_i^{(1)}(\theta, \sigma_n) = (\partial B_i / \partial \lambda, \partial B_i / \partial \beta^T)^T.$$

$$\frac{\partial^2 B_i}{\partial \lambda^2} = K''\left(\frac{z_i}{\sigma_n}\right) \left[\frac{e_i^T S_n^{-1}(\lambda) W_n S_n^{-1}(\lambda) \Delta X_n}{\sigma_n}\right]^2 + 2K'\left(\frac{z_i}{\sigma_n}\right) \frac{e_i^T S_n^{-1}(\lambda) W_n S_n^{-1}(\lambda) W_n S_n^{-1}(\lambda) \Delta X_n}{\sigma_n};$$

$$\frac{\partial^2 B_i}{\partial \beta^T \partial \lambda} = K''\left(\frac{z_i}{\sigma_n}\right) \frac{e_i^T S_n^{-1}(\lambda) W_n S_n^{-1}(\lambda) \Delta X_n}{\sigma_n} + K'\left(\frac{z_i}{\sigma_n}\right) \frac{e_i^T S_n^{-1}(\lambda) W_n S_n^{-1}(\lambda) \Delta \tilde{X}_n}{\sigma_n};$$

$$\frac{\partial^2 B_i}{\partial \beta^T \partial \beta} = K''\left(\frac{z_i}{\sigma_n}\right) \frac{\Delta \tilde{X}_n^T [S_n^{-1}(\lambda)]^T e_i e_i^T S_n^{-1}(\lambda) \Delta \tilde{X}_n}{\sigma_n^2};$$

$$B_i^{(2)}(\theta, \sigma_n) = \begin{pmatrix} \frac{\partial^2 B_i}{\partial \lambda^2} & \frac{\partial^2 B_i}{\partial \beta^T \partial \lambda} \\ \ast & \frac{\partial^2 B_i}{\partial \beta^T \partial \beta} \end{pmatrix}.$$

Recall that

$$G_n(\theta; \sigma_n) = \frac{1}{n} \sum_{i=1}^{n} \Delta Y_i K\left(\frac{e_i^T S_n^{-1}(\lambda) \Delta X_n}{\sigma_n}\right) = \frac{1}{n} \sum_{i=1}^{n} \mathbb{1}\{Y_i \neq Y_{i2}\} (1 - 2 \cdot \mathbb{1}\{Y_i = 0, Y_{i2} = 1\}) B_i,$$

then we have

$$T_n(\theta, \sigma_n) = \frac{\partial G_n(\theta, \sigma_n)}{\partial \theta} = \frac{1}{n} \sum_{i=1}^{n} \mathbb{1}\{Y_i \neq Y_{i2}\} (1 - 2 \cdot \mathbb{1}\{Y_i = 0, Y_{i2} = 1\}) B_i^{(1)}(\theta, \sigma_n);$$

$$Q_n(\theta, \sigma_n) = \frac{\partial^2 G_n(\theta, \sigma_n)}{\partial \theta \partial \theta^T} = \frac{1}{n} \sum_{i=1}^{n} \mathbb{1}\{Y_i \neq Y_{i2}\} (1 - 2 \cdot \mathbb{1}\{Y_i = 0, Y_{i2} = 1\}) B_i^{(2)}(\theta, \sigma_n);$$
Appendix B: Proofs

Proof of Lemma 1. Without loss of generality, let $q' = q$ and consider the case in which $\beta_{0,q} > 0$ (the case $\beta_{0,q} < 0$ is symmetric). For any $(\lambda, \beta) \in \Lambda \times \mathbb{R}^q$, let $\bar{\beta} = (\beta_1, \ldots, \beta_{q-1})$ and $\bar{\beta}_0 = (\beta_{0,1}, \ldots, \beta_{0,q-1})$. To show that $R(\lambda, \beta) > 0$, it is sufficient to show that, for all $(\lambda, \beta) \in \Lambda \times \mathbb{R}^q$ with $\beta / ||\beta|| = \beta_0 / ||\beta_0||$, either $\Pr(e_i^T S_n^{-1}(\lambda) \Delta X_n \beta < 0 < e_i^T S_n^{-1} \Delta X_n \beta_0)$ or $\Pr(e_i^T S_n^{-1} \Delta X_n \beta_0 < 0 < e_i^T S_n^{-1}(\lambda) \Delta X_n \beta)$ or both.

Denote $e_i^T S_n^{-1}(\lambda) = (a_{i1}(\lambda), a_{i2}(\lambda), \ldots, a_{in}(\lambda))$ and $e_i^T S_n^{-1} = (b_{i1}(\lambda_0), b_{i2}(\lambda_0), \ldots, b_{in}(\lambda_0))$, as the inverse matrices $S_n^{-1}(\lambda)$ and $S_n^{-1}$ exist, so there exists at least one $a_{ij}(\lambda) \neq 0$ and one $b_{ij'}(\lambda_0) \neq 0$. Apparently, three are four possible index sets: $J, J', K, K'$, where $a_{ij}(\lambda) \neq 0, b_{ij}(\lambda_0) = 0$ for all $j \in J$; $a_{ij}(\lambda) = 0, b_{ij'}(\lambda_0) \neq 0$ for all $j' \in J'$; $a_{ik}(\lambda) \neq 0, b_{ik}(\lambda_0) = 0$ for all $k \in K$; $a_{ik}(\lambda) = 0, b_{ik'}(\lambda_0) = 0$ for all $k' \in K'$.

It is easy to see that

$$\Pr\left(e_i^T S_n^{-1}(\lambda) \Delta X_n \beta < 0 < e_i^T S_n^{-1} \Delta X_n \beta_0\right)$$

$$= \Pr\left(\sum_{j \in J} a_{ij}(\lambda) \Delta X_j \beta + \sum_{k \in K} a_{ik}(\lambda) \Delta X_k \beta < 0 < \sum_{j' = 1} b_{ij'}(\lambda_0) \Delta X_{j'} \beta_0 + \sum_{k \in K} b_{ik}(\lambda_0) \Delta X_k \beta_0\right)$$

$$\geq \Pr\left(A_{i,j \in J}, B_{i,j' \in J'}, C_{i,k \in K}\right)$$

(A.1)

where

$A_{i,j} = \{a_{ij}(\lambda) \Delta X_j \beta < 0\} = \{a_{ij}(\lambda) \Delta \bar{X}_j \bar{\beta} + a_{ij}(\lambda) \Delta X_{j,q} \beta_q < 0\}$,

$B_{i,j'} = \{b_{ij'}(\lambda_0) \Delta X_{j'} \beta_0 > 0\} = \{b_{ij'}(\lambda_0) \Delta \bar{X}_{j'} \bar{\beta}_0 + b_{ij'}(\lambda_0) \Delta X_{j',q} \beta_{0,q} > 0\}$,

$C_{i,k} = \{a_{ik}(\lambda) \Delta \bar{X}_k \bar{\beta} + a_{ik}(\lambda) \Delta X_{k,q} \beta_q < 0 < b_{ik}(\lambda_0) \Delta \bar{X}_k \bar{\beta}_0 + b_{ik}(\lambda_0) \Delta X_{k,q} \beta_{0,q}\}$

for all $j, j'$ and $k$.

In the proof of Lemma 2 in Manski (1985), three cases were considered to show the positive conditional probability of $C_{i,k}$ when regarding the different signs of $\beta_q$. In our proof, we have six different cases as we need to consider the different signs of $a_{ik}(\lambda)$ and $b_{ik}(\lambda_0)$:

(i) Case $a_{ik}(\lambda) \beta_q < 0$ and $b_{ik}(\lambda_0) > 0$: $C_{i,k} = \left[\Delta X_{k,q} > \max \left(-\frac{\Delta \bar{X}_k \bar{\beta}}{\beta_q}, -\frac{\Delta \bar{X}_k \bar{\beta}_0}{\beta_{0,q}}\right)\right];$

(ii) Case $a_{ik}(\lambda) \beta_q < 0$ and $b_{ik}(\lambda_0) < 0$: $C_{i,k} = \left[-\frac{\Delta \bar{X}_k \bar{\beta}}{\beta_q} < \Delta X_{k,q} < -\frac{\Delta \bar{X}_k \bar{\beta}_0}{\beta_{0,q}}\right];$
(iii) Case \( a_{ik}(\lambda)\beta_q = 0 \) and \( b_{ik}(\lambda_0) > 0 \): \( C_{i,k} = a_{ik}(\lambda)\Delta \tilde{X}_{k,\beta} < 0, \Delta X_{k,q} > -\frac{\Delta \tilde{X}_k\beta_0}{\beta_0 q} \); 
(iv) Case \( a_{ik}(\lambda)\beta_q = 0 \) and \( b_{ik}(\lambda_0) < 0 \): \( C_{i,k} = a_{ik}(\lambda)\Delta \tilde{X}_{k,\beta} < 0, \Delta X_{k,q} < -\frac{\Delta \tilde{X}_k\beta_0}{\beta_0 q} \); 
(v) Case \( a_{ik}(\lambda)\beta_q > 0 \) and \( b_{ik}(\lambda_0) > 0 \): \( C_{i,k} = -\frac{\Delta \tilde{X}_k\beta_0}{\beta_0 q} < \Delta X_{k,q} < -\frac{\Delta \tilde{X}_k\beta_0}{\beta_0 q} \); 
(vi) Case \( a_{ik}(\lambda)\beta_q > 0 \) and \( b_{ik}(\lambda_0) < 0 \): \( C_{i,k} = \Delta X_{k,q} < \min \left( -\frac{\Delta \tilde{X}_k\beta_0}{\beta_0 q}, \frac{\Delta \tilde{X}_k\beta_0}{\beta_0 q} \right) \).

Under Assumption 2 and using the same argument as in the proof of Lemma 2 in Manski (1985), we know that the conditional probabilities of \( A_{i,j}, B_{i,j'}, \) and \( C_{i,k} \) in case (i) and (vi) are always positive. In case (iii) and (iv), we can always find a positive constant \( D \) such that the conditional probability \( \Pr \left( a_{ik}(\lambda)\Delta \tilde{X}_{k,\beta} - D < 0 \right) > 0 \), therefore,

\[
\Pr(C'_{i,k}) = \Pr \left( a_{ik}(\lambda)\Delta \tilde{X}_{k,\beta} - D < 0, \Delta X_{k,q} > -\frac{\Delta \tilde{X}_k\beta_0}{\beta_0 q} \right) > 0.
\]

To make sure this adjustment does not change the original conditional probability in equation (A.1), we just need to change one element in \( A_{i,j} \) to \( A'_{i,j} = \{ a_{ij}(\lambda)\Delta X_{j,\beta} + D < 0 \} \), and the conditional probability of \( A'_{i,j} \) is still positive. Using the similar adjustment, we can always make sure that the conditional probabilities \( \Pr(C'_{i,k}) \) in case (ii) and (v) to be positive.

Recall that in equation (A.1),

\[
\Pr \left( e_i^\top S_n^{-1}(\lambda) \Delta X_{n,\beta} < 0 < e_i^\top S_n^{-1} \Delta X_{n,\beta_0} \right) \\
\geq \Pr \left( A'_{i,j} \cup B_{i,j'} \cup C'_{i,k} \in K \right) \\
= \Pr \left( A'_{i,j} \mid B_{i,j'} \cup C'_{i,k} \in K \right) \Pr \left( B_{i,j'} \mid C'_{i,k} \in K \right) \Pr \left( C'_{i,k} \in K \right)
\]

as we just argued that each conditional probability in equation (A.2) is positive under Assumption 2, so we have

\[
\Pr \left( e_i^\top S_n^{-1}(\lambda) \Delta X_{n,\beta} < 0 < e_i^\top S_n^{-1} \Delta X_{n,\beta_0} \right) > 0, \quad R(\lambda, \beta) > 0
\]

Therefore, \( (\lambda_0, \beta_0) \) is identified relative to \( (\lambda, \beta) \) except those \( \beta \) that are scalar multiples of \( \beta_0 \). \( \square \)

For the proof of Theorem 1, we need Proposition 1 and the following Lemmas.
Proof of Proposition 1. To prove the NED of \( \{Y_{it}\} \), we first show the NED of

\[
Y_{it}^* = e_i^T S_n^{-1}(\lambda) (X_{it}\beta + \alpha_n + \epsilon_{it}) = \sum_{j=1}^{n} a_{ij}(\lambda) (X_{jt}\beta + \alpha_j + \epsilon_{jt}).
\]

By Assumption 5 and Theorem 14.1 of Davidson (1994), the process \( V_{it} = X_{it}\beta + \alpha_i + \epsilon_{it} \) is strong mixing with \( \alpha \)-mixing coefficient \( \alpha(m) \). Then, by the Minkowski inequality,

\[
||Y_{it}^* - E(Y_{it}^*|\mathcal{Z}_{i,n}(m))||_2 = \left\| \sum_{j,d(l_i,l_j) > m} a_{ij}(\lambda) (V_{jt} - E(V_{jt}|\mathcal{Z}_{i,n}(m))) \right\|_2 \leq d_t \nu(m), \tag{A.3}
\]

where \( \nu(m) = \sup_l \sum_{j,d(l_i,l_j) > m} |a_{ij}(\lambda)| \), and \( d_t = 2\sup_l ||V_{jt}||_2 \). Therefore, \( \{Y_{it}^*\} \) is NED because \( \nu(m) \to 0 \) as \( m \to \infty \) by equation (12).

For any \( \epsilon > 0 \), let \( \delta_\epsilon(0) \) denote the \( \epsilon \)-neighborhood of 0, then we have the following inequality for the indicator function:

\[
\left| \mathbb{I}\{x_1 > 0\} - \mathbb{I}\{x_2 > 0\} \right| \leq \frac{|x_1 - x_2|}{\epsilon} \mathbb{I}\{x_1 \notin \delta_\epsilon(0) \text{ or } x_2 \notin \delta_\epsilon(0) \} + \mathbb{I}\{x_1 \in \delta_\epsilon(0), x_2 \in \delta_\epsilon(0)\}. \tag{A.4}
\]

Denote \( B = \{Y_{it}^* \in \delta_\epsilon(0), E(Y_{it}^*|\mathcal{Z}_{i,n}(m) \in \delta_\epsilon(0)\} \), then we have

\[
||\mathbb{I}\{Y_{it}^* > 0\} - E(\mathbb{I}\{Y_{it}^* > 0\}|\mathcal{Z}_{i,n}(m))||_2 \\
\leq ||\mathbb{I}\{Y_{it}^* > 0\} - \mathbb{I}\{E(Y_{it}^*|\mathcal{Z}_{i,n}(m) > 0)\}||_2 \\
= (E(\mathbb{I}\{Y_{it}^* > 0\} - \mathbb{I}\{E(Y_{it}^*|\mathcal{Z}_{i,n}(m) > 0)\})^2)^{1/2} \\
\leq \left( \frac{1}{\epsilon^2} \int_{B^c} |Y_{it}^* - E(Y_{it}^*|\mathcal{Z}_{i,n}(m))|^2 \, dP + \int_{B} \, dP \right)^{1/2} \\
\leq \frac{1}{\epsilon} \left( \int_{B^c} |Y_{it}^* - E(Y_{it}^*|\mathcal{Z}_{i,n}(m))|^2 \, dP \right)^{1/2} + \left( \int_{B} \, dP \right)^{1/2} \\
\leq \frac{1}{\epsilon} ||Y_{it}^* - E(Y_{it}^*|\mathcal{Z}_{i,n}(m))||_2 + \left( \int_{B} \, dP \right)^{1/2} \\
\leq \frac{1}{\epsilon} d_t \nu(m) + \left( \int_{B} \, dP \right)^{1/2},
\]

where the first inequality is followed by Theorem 10.12 of Davidson (1994), the third line is by definition, the fourth line is by equation (A.4), the last line is followed by equation (A.3). As these two terms converge to 0 when \( \epsilon \) converges to 0 at a slower rate than \( \nu(m) \),
so the process \{Y_{it}\} is near epoch dependent.

The NED of process \{\Delta Y_i\} follows from Theorem 17.8 of Davidson (1994), and the NED of \{\text{sgn}(e_i^T S_n^{-1}(\lambda)\Delta X_n \beta)\} could be shown similarly as \{Y_{it}\}. \hfill \Box

**Lemma 2.** Under Assumptions 1-3, and define

\[
G(\theta) = \frac{1}{n} \sum_{i=1}^{n} G_i(\theta) = \frac{1}{n} \sum_{i=1}^{n} E \left[ \Delta Y_i \text{sgn}(e_i^T S_n^{-1}(\lambda)\Delta X_n \beta) \right],
\]

then \(G(\theta_0) > G(\theta)\) for all \(\theta = (\lambda, \beta) \in \Lambda \times R^q\), where \(\beta/||\beta|| \neq \beta_0/||\beta_0||\) when \(\lambda = \lambda_0\).

**Proof of Lemma 2.** Given Lemma 1, a similar result of Manski (1987) Lemma 3 could be easily shown that \(G_i(\theta_0) > G_i(\theta)\) for all individual \(i\). That is, \(\theta_0\) uniquely maximizes each \(G_i(\theta)\), so it actually maximizes \(G(\theta) = \frac{1}{n} \sum_{i=1}^{n} G_i(\theta)\). To prove the uniqueness of \(\theta_0 = \operatorname{arg\,max}_\theta G(\theta)\), suppose there is a \(\theta' \neq \theta_0\) such that \(\theta'\) maximizes \(G(\theta)\), which means there exists at least one \(m\) such that \(G_m(\theta') \geq G_m(\theta_0)\), this contradicts with \(\theta_0\) is a unique maximizer of \(G_m(\theta)\). Therefore, we have \(\theta' = \theta_0\), and \(\theta_0\) is also the unique maximizer of \(G(\theta)\). \hfill \Box

**Lemma 3.** For all \(c \in \mathbb{R}\) if \((\Delta Y_i, \Delta X_i)\) is strong mixing, then

\[
\sup_{\theta \in \Theta} \left| \frac{1}{n} \sum_{i=1}^{n} \left( \Delta Y_i \mathbb{I}\{e_i^T S_n^{-1}(\lambda)\Delta X_n \beta \leq c\} - E\Delta Y_i \mathbb{I}\{e_i^T S_n^{-1}(\lambda)\Delta X_n \beta \leq c\} \right) \right| \rightarrow^p 0.
\]

In addition, if \(\alpha(m) \leq Cm^{-r}\) for positive constants \(C\) and \(r\), then convergence is almost surely.

**Proof of Lemma 3.** The proof is similar to the proof of Lemma 4 in de Jong and Woutersen (2011), except that \(\Delta Y_i\) is a heterogenous rather than stationary strong mixing process. We also apply the generic uniform law of large numbers of the Theorem of Andrews (1987). It requires compactness of the parameter space \(\Theta\), which is assumed by Assumptions 3 and 6; the summands \(q_i(w_i, \theta), q_i^*(w_i, \theta) = \sup\{q_i(w_i, \theta') : \theta' \in \Theta, d(\theta, \theta') < \rho\}\) and \(q_{si}(w_i, \theta) = \inf\{q_i(w_i, \theta') : \theta' \in \Theta, d(\theta, \theta') < \rho\}\) are well-defined and satisfy a (respectively weak or strong) law of large numbers; and for all \(\theta \in \Theta\),

\[
\lim_{\rho \rightarrow 0} \sup_i \left| \frac{1}{n} \sum_{i=1}^{n} (E q_i^*(w_i, \theta) - E q_{si}(w_i, \theta)) \right| = 0.
\]
Here we show the last result, denote $q_i^1(w_i, \theta) = e_i^T S_n^{-1}(\lambda) \Delta X_n \beta$ and $K = \sup_i \sup_{\theta^*, \theta^* < \theta} \frac{\partial q_i^1(w_i, \theta)}{\partial \theta}|_{\theta = \theta^*}$, we have

$$
\lim_{\rho \to 0} \frac{1}{n} \sum_{i=1}^n (E q_i^1(w_i, \theta) - E q_i^1(w_i, \theta)) = \lim_{\rho \to 0} \frac{1}{n} \sum_{i=1}^n \left( E \Delta Y_i \mathbb{1}\{q_i^1(w_i, \theta') \leq c\} - E \Delta Y_i \mathbb{1}\{q_i^1(w_i, \theta') \leq c\} \right)
$$

$$
\leq \lim_{K \to \infty} \lim_{\rho \to 0} \frac{1}{n} \sum_{i=1}^n \left( E \Delta Y_i \mathbb{1}\{q_i^1(w_i, \theta) \leq c + \rho K\} - E \Delta Y_i \mathbb{1}\{q_i^1(w_i, \theta') \leq c - \rho K\} \right)
$$

$$
+ \lim_{K \to \infty} \lim_{\rho \to 0} \frac{1}{n} \sum_{i=1}^n \left( E \Delta Y_i \mathbb{1}\{q_i^1(w_i, \theta') \leq c - \rho K\} \right)
$$

$$
\leq \lim_{K \to \infty} \lim_{\rho \to 0} \frac{1}{n} \sum_{i=1}^n \left( \Pr\{q_i^1(w_i, \theta) \leq c + \rho K\} - \Pr\{q_i^1(w_i, \theta') \leq c - \rho K\} \right) = 0,
$$

because $\Delta X_{i,q}$ has a continuous distribution. Moreover, note that $q_i(w_i, \theta)$, $q_i^1(w_i, \theta)$ and $q_{\ast 1}(w_i, \theta)$ are all well-defined strong mixing random variables and satisfy a strong law of large numbers of Theorem 4 of De Jong (1995) if $\alpha(m) + \nu(m) \leq Cm^{-r}$ for some positive constants $C$ and $r$.

**Lemma 4.** Under Assumptions 1-6, $G_n^*(\theta) \to_p G(\theta)$ uniformly over $\theta \in \Theta$, where $G_n^*(\theta) = \frac{1}{n} \sum_{i=1}^n \Delta Y_i \text{sgn} \left( e_i^T S_n^{-1}(\lambda) \Delta X_n \beta \right)$. In addition, if $\alpha(m) \leq Cm^{-r}$ for positive constants $C$ and $r$, then convergence is almost surely.

**Proof of Lemma 4.** By equation (8), we have

$$
G_n^*(\theta) = \frac{1}{n} \sum_{i=1}^n \Delta Y_i \text{sgn} \left( e_i^T S_n^{-1}(\lambda) \Delta X_n \beta \right)
$$

$$
= \frac{2}{n} \sum_{i=1}^n \Delta Y_i \mathbb{1}\{e_i^T S_n^{-1}(\lambda) \Delta X_n \beta \geq 0\} - \frac{1}{n} \sum_{i=1}^n \Delta Y_i,
$$

both terms satisfy a weak or strong uniform law of large numbers by Lemma 3.

**Lemma 5.** $G(\theta)$ is continuous at all $\theta = (\lambda, \beta^T)^T$ such that $\beta_q \neq 0$.

**Proof of Lemma 5.** The result follows from the Theorem of Andrews (1987) and Lemma 3.
Lemma 6. Under Assumptions 1-6, $|G_n(\theta; \sigma_n) - G_n^*(\theta)| \to^p 0$ uniformly over $\theta \in \Theta$. In addition, if $\alpha(m) \leq Cm^{-r}$ for positive constants $C$ and $r$, then convergence is almost surely.

Proof of Lemma 6. As in Charlier, Melenberg, and van Soest (1995), here we actually adjusted the definition of $G_n^*(\theta)$ such that $G_n^*(\theta) = \frac{1}{n} \sum_{i=1}^{n} \Delta Y_i \mathbb{1}\{e_i^T S_n^{-1} (\lambda) \Delta X_n \beta \geq 0\}$ as in equation (8).

$$
|G_n(\theta; \sigma_n) - G_n^*(\theta)| = \frac{1}{n} \sum_{i=1}^{n} \Delta Y_i \left| \mathbb{1}\{e_i^T S_n^{-1} (\lambda) \Delta X_n \beta \geq 0\} - K \left( \frac{e_i^T S_n^{-1} (\lambda) \Delta X_n \beta}{\sigma_n} \right) \right|
$$

Under the uniform weak or strong law of large numbers for $\frac{1}{n} \sum_{i=1}^{n} \mathbb{1}\{|e_i^T S_n^{-1} (\lambda) \Delta X_n \beta| < c\}$ converging to $\frac{1}{n} \sum_{i=1}^{n} \text{Pr}\{|e_i^T S_n^{-1} (\lambda) \Delta X_n \beta| < c\}$ (implied by Lemma 3), similar to Horowitz (1992 Lemma 4), we can easily show that $|G_n(\theta; \sigma_n) - G_n^*(\theta)| \to 0$ almost surely uniformly over $\theta \in \Theta$ as $n \to \infty$.

Proof of Theorem 1. For weak or strong consistency of $(\theta_n \to \theta_0)$, it is sufficient to verify the following conditions: (i) $G(\theta)$ has a unique maximum at $\theta_0$; (ii) The parameter space $\Theta$ is compact; (iii) $G(\theta)$ is continuous; (iv) $\{G_n(\theta)\}$ converges uniformly in probability to $G(\theta)$, and strong consistency can be obtained if this is replaced by $\sup_{\theta \in \Theta} |G_n(\theta) - G(\theta)| \to^{a.s.} 0$.

Condition (i) is satisfied by Lemmas 1 and 2, condition (ii) is provided by Assumptions 3 and 6, condition (iii) is proved by Lemma 5, and condition (iv) is obtained by Lemmas 4 and 6.

For the proofs of Theorems 2 and 3, we need the following Lemmas:

Lemma 7. Let Assumptions 1-11 and 14 hold. Then

$$(a) \quad \lim_{n \to \infty} \text{E} \left[ \sigma_n^{-h} T_n(\theta_0; \sigma_n) \right] = A; \quad (b) \quad \lim_{n \to \infty} \text{Var} \left[ (n\sigma_n)^{1/2} T_n(\theta_0; \sigma_n) \right] = D.$$
Proof of Lemma 7. As we know that

\[ T_n(\theta_0, \sigma_n) = \frac{1}{n} \sum_{i=1}^{n} \Delta Y_i B_i^{(1)}(\theta_0, \sigma_n) \]

\[ = \frac{1}{n} \sum_{i=1}^{n} \mathbb{1}\{Y_{i1} \neq Y_{i2}\} \left(1 - 2 \cdot \mathbb{1}\{Y_{i1} = 0, Y_{i2} = 1\}\right) B_i^{(1)}(\theta_0, \sigma_n), \]

then

\[ E_n(T) = E[\sigma_n^{-h} T_n(\theta_0, \sigma_n)] \]

\[ = \sigma_n^{-1} \frac{1}{n} \sum_{i=1}^{n} \Pr\{Y_{i1} \neq Y_{i2}\} \int \left[1 - 2F_i(-z_i\mid z_i, \tilde{Z}_i)\right] B_i^{(1)}(\theta_0, \sigma_n)p_i(z_i\mid \tilde{Z}_i)dz_i dP_i(\tilde{Z}_i). \]

By Assumption 1, condition (5) and the Corollary of Manski (1987), we could easily derive that Median \( (Y_{i1} - Y_{i2}) \Delta X_n, Y_{i1} \neq Y_{i2} = \text{sgn}\{z_i\} \), so we have Median \( (\tilde{\epsilon}_i \Delta X_n, Y_{i1} \neq Y_{i2}) = 0 \) and \( F_i(0)\{0, \tilde{Z}_i\} = 0.5 \) for almost every \( \tilde{Z}_i \) and \( i = 1, \ldots, n \).

The proof of part (a) is analogous to that of Lemma 5 in Horowitz (1992), the only adjustment is that we need the boundedness of matrices \( S_n^{-1} \) and \( W_n^{-1}S_n^{-1} \) to guarantee the boundedness of \( \tilde{B}_i \) for applying Lebesgue’s dominated convergence theorem. This is immediately from Assumption 6 and footnote 3.

To prove part (b), let first denote \( t_n(\theta_0, \sigma_n) = \frac{1}{n} \sum_{i=1}^{n} \mathbb{1}\{Y_{i1} \neq Y_{i2}\} B_i^{(1)}(\theta_0, \sigma_n) \), then

\[ V_n(T) = \text{Var} \left[ (n\sigma_n)^{1/2} T_n(\theta_0, \sigma_n) \right] \]

\[ = n\sigma_n E \left[ t_n(\theta_0, \sigma_n) t_n(\theta_0, \sigma_n)^\top \right] + o(1) \]

\[ = \frac{\sigma_n}{n} \sum_{i=1}^{n} E \left[ B_i^{(1)}(\theta_0, \sigma_n) B_i^{(1)}(\theta_0, \sigma_n)^\top \right] \Pr\{Y_{i1} \neq Y_{i2}\} \]

\[ + \frac{\sigma_n}{n} \sum_{i=1}^{n} \sum_{j \neq i} E \left[ B_i^{(1)}(\theta_0, \sigma_n) B_j^{(1)}(\theta_0, \sigma_n)^\top \right] \Pr\{Y_{i1} \neq Y_{i2}\} \Pr\{Y_{j1} \neq Y_{j2}\} + o(1) \]

\[ = D_{n1} + D_{n2} + o(1). \]

Similar to Lemma 5 of Horowitz (1992),

\[ D_{n1} = \frac{1}{n\sigma_n} \sum_{i=1}^{n} E \left[ K'(z_i\mid \sigma_n) \right] B_{1,i} B_{1,i}^\top \Pr\{Y_{i1} \neq Y_{i2}\} \]

\[ = \frac{1}{n\sigma_n} \sum_{i=1}^{n} \Pr\{Y_{i1} \neq Y_{i2}\} \int \left[ K'(z_i\mid \sigma_n) \right] B_{1,i} B_{1,i}^\top p_i(z_i\mid \tilde{Z}_i)dz_i dP_i(\tilde{Z}_i) \rightarrow D \]

27
by Lebesgue’s dominated convergence theorem and Assumptions 7-10. Lemma 7 of de Jong and Woutersen (2011) shows that $D_{n2}$ is asymptotically negligible. This finishes the proof of part (b).

**Lemma 8.** Let Assumptions 1-11 and 14 hold. (a) If $n\sigma_{n}^{2h+1} \to \infty$ as $n \to \infty$, $\sigma_{n}^{-h}T_{n}(\theta_{0}; \sigma_{n})$ converges in probability to $A$. (b) If $n\sigma_{n}^{2h+1} \to \infty$ has a finite limit $\kappa$ as $n \to \infty$, $(n\sigma_{n})^{1/2}T_{n}(\theta_{0}; \sigma_{n})$ converges in distribution to $MVN(\kappa^{1/2}A, D)$.

Analogously to Horowitz (1992) and de Jong and Woutersen (2011), define

$$g_{i}(\zeta) = \mathbb{1}\{Y_{i1} \neq Y_{i2}\} (2 \cdot \mathbb{1}\{Y_{i1} = 1, Y_{i2} = 0\} - 1) \tilde{B}_{1,i}K' \left( \frac{\hat{z}_{i}}{\sigma_{n}} + \zeta^{\top} \tilde{B}_{1,i} \right).$$

**Lemma 9.** If $(Y_{it}, X_{it})$ is strong mixing with strong mixing sequence $\alpha(m)$, and there exist a sequence $m_{n} \geq 1$ such that

$$\sigma_{n}^{-3(p+q-1)}\sigma_{n}^{-2}n^{1/s}\alpha(m_{n}) + (\log(nm_{n})) \left( n^{1-2/s}m_{n}^{-2} \right)^{-1} \to 0.$$

then

$$\sup_{\zeta} \left| \frac{1}{n\sigma_{n}^{2}} \sum_{i=1}^{n} [g_{i}(\zeta) - Eg_{i}(\zeta)] \right| \to p 0.$$

**Proofs of lemmas 8 and 9.** The proofs are identical to the proofs of Lemma 8 and 11 of de Jong and Woutersen (2011) except that we have a different score function.

**Lemma 10.** Let Assumptions 1-15 hold, and define $\phi_{n} = (\tilde{\theta}_{n} - \hat{\theta}_{0})/\sigma_{n}$, where $\theta_{n}$ is a smoothed spatial maximum score estimator. Then $\text{plim}_{n \to \infty} \phi_{n} = 0$.

**Proof of Lemma 10.** This follows from Lemma 9 and the reasoning of Lemma 8 in Horowitz (1992).

**Lemma 11.** Let Assumptions 1-15 hold. Let $\{\theta_{n}'\} = \{\tilde{\theta}_{n}', \beta_{n,q}'\}$ be any sequence in $\Theta$ such that $(\theta_{n}' - \theta_{0})/\sigma_{n} \to 0$ as $n \to \infty$. Then $\text{plim}_{n \to \infty} Q_{n}(\theta_{n}' ; \sigma_{n}) = Q$.

**Proof of Lemma 11.** We can separately show that the elements of $Q_{n}(\theta ; \sigma_{n})$ follow a uniformly law of larger numbers. The proof is then analogous to the proof of Lemma 13 in
de Jong and Woutersen (2011), except that we have different objective functions.

Proof of Theorem 2. The proof is identical to that of Theorem 2 in Horowitz (1992), where we need Lemmas 10 and 11 instead of Lemmas 8 and 9 in Horowitz (1992).

Proof of Theorem 3. The proof is identical to that of Theorem 7 in de Jong and Woutersen (2011), where we need Lemmas 10 and 11 instead of Lemmas 12 and 13 in de Jong and Woutersen (2011).

References


Beron, K., and W. Vlijmen (2004): “Probit in a spatial context: a Monte Carlo analysis.”


31