Partially Linear Spatial Probit Models

Mohamed-Salem Ahmed, Sophie Dabo-Niang, Michaël Genin, Alaa Ali Hassan

Pub. Inst. Stat. Univ. Paris
63, fasc. 2-3, 2019, 71-96
Numéro spécial en l'honneur des 80 ans de Denis Bosq /
Special issue in honour of Denis Bosq's 80th birthday

Partially Linear Spatial Probit Models*

M.S Ahmed[†], S Dabo-Niang[‡] and G Genin[†] and A.A Hassan[§]

Abstract: A partially linear probit model for spatially dependent data is considered. A triangular array setting is used to cover various patterns of spatial data. Conditional spatial heteroscedasticity and non-identically distributed observations and a linear process for disturbances are assumed, allowing various spatial dependencies. The estimation procedure is a combination of a weighted likelihood and a generalized method of moments. The procedure first fixes the parametric components of the model and then estimates the non-parametric part using weighted likelihood; the obtained estimate is then used to construct a GMM (Generalized Method of Moments) parametric component estimate. The consistency and asymptotic distribution of the estimators are established under sufficient conditions. Some numerical results are provided to investigate the finite sample performance of the estimators.

Introduction

Agriculture, economics, environmental sciences, urban systems, and epidemiology activities often utilize spatially dependent data. Therefore, modelling such activities requires one to find a type of correlation between some random variables in one location with other variables in neighbouring locations; see for instance [30]. This is a significant feature of spatial data analysis. Spatial/Econometrics statistics provides tools to perform such modelling. Many studies on spatial effects in statistics and econometrics using many diverse models have been published; see [10], [2], [3] and [4] for a review.

Two main methods of incorporating a spatially dependent structure [see for instance 10] can essentially be distinguished as between geostatistics and lattice data. In the domain of geostatistics, the spatial location is valued in a continuous set of $\mathbb{R}^N$, $N \geq 2$. However, for many activities, the spatial index or location does not vary continuously and may be of the lattice type, the baseline of this current work. In image analysis, remote sensing from satellites, agriculture etc., data are often received as a regular lattice and identified as the centroids of square pixels, whereas a mapping often forms an irregular lattice. Basically, statistical models for lattice data are linked to nearest neighbours to express the fact that data are nearby.

---

*Footnote to the title with the 'thankstext' command.

[†]Univ. Lille, CHU Lille, ULR 2694 - METRICS: Évaluation des technologies de santé et des pratiques médicales, F-59000 Lille, France

[‡]PAINLEVE-CNRS 8524, INRIA-MODAL, Université de Lille, Villeneuve d'ascq, France

[§]PAINLEVE-CNRS 8524, Université de Lille, Villeneuve d'ascq, France

AMS 2000 subject classifications: Primary 62G05, 62H11; secondary 62M30

Keywords and phrases: Binary choice model, GMM, Spatial semiparametric inference

Two popular spatial dependence models have received substantial attention for lattice data, the spatial autoregressive (SAR) dependent variable model and the spatial autoregressive error model (SAE, where the model error is an SAR), which extend the regression in a time series setting to spatial one.
From a theoretical point of view, various linear spatial regression SAR and SAE models as well as their identification and estimation methods, e.g., two-stage least squares (2SLS), three-stage least squares (3SLS), maximum likelihood (ML) or quasi-maximum likelihood (QML) and the generalized method of moments (GMM), have been developed and summarized by many authors such as [3], [16], [17], [9], [10], [8], [19], [20], [22], [41], [23], [13], [40]. Introducing nonlinearity into the field of spatial linear lattice models has attracted less attention; see for instance [32], who generalized kernel regression estimation to spatial lattice data. [38] proposed a semi-parametric GMM estimation for some semi-parametric SAR models. Extending these models and methods to discrete choice spatial models has seen less attention; only a few papers have been concerned with this topic in recent years. This may be, as noted by [12] (see also [36] and [6]), due to the "added complexity that spatial dependence introduces into discrete choice models". Estimating the model parameters with a full ML approach in spatially discrete choice models often requires solving a very computationally demanding problem of n-dimensional integration, where n is the sample size.
For linear models, many discrete choice models are fully linear and utilize a continuous latent variable; see for instance [36], [39] and [25], who proposed pseudo-ML methods, and [30], who studied a method based on the GMM approach. Also, others methodologies of estimation are emerged like, EM algorithm [27] and Gibbs sampling approach [21].

When the relationship between the discrete choice variable and some explanatory variables is not linear, a semi-parametric model may represent an alternative to fully parametric models. This type of model is known in the literature as *partially linear choice spatial models* and is the baseline of this current work. When the data are independent, these choice models can be viewed as special cases of the famous generalized additive models [14] and have received substantial attention in the literature, and various estimation methods have been explored [see for instance 7, 15, 34].
To the best of our knowledge, semi-parametric spatial choice models have not yet been investigated from a theoretical point of view. To fill this gap, this work addresses an SAE spatial probit model for when the spatial dependence structure is integrated in a disturbance term of the studied model.
We propose a semi-parametric estimation method combining the GMM approach and the weighted likelihood method. The method consists of first fixing the parametric components of the model and non-parametrically estimating the non-linear component by weighted likelihood [37]. The obtained estimator depending on the values at which the parametric components are fixed is used to construct a GMM estimator [30] of these components.
The remainder of this paper is organized as follows. In Section 1, we introduce the

studied spatial model and the estimation procedure. Section 2 is devoted to hypotheses and asymptotic results, while Section 3 reports a discussion and computation of the estimates. Section 4 gives some numerical results based on simulated data to illustrate the performance of the proposed estimators. The last section presents the proofs of the main results.

1. Model

We consider that at n spatial locations $\{s_1, s_2, \ldots, s_n\}$ satisfying $\|s_i - s_j\| > \rho$ with $\rho > 0$, observations of a random vector (Y, X, Z) are available. Assume that these observations are considered as triangular arrays [32] and follow the partially linear model of a latent dependent variable Y^*:

$$Y_{in}^* = X_{in}^T \beta_0 + g_0(Z_{in}) + U_{in}, \qquad 1 \leq i \leq n,\ n = 1, 2, \ldots \tag{1.1}$$

with

$$Y_{in} = \mathbb{I}(Y_{in}^* \geq 0), \qquad 1 \leq i \leq n,\ n = 1, 2, \ldots \tag{1.2}$$

where $\mathbb{I}(\cdot)$ is the indicator function; X and Z are explanatory random variables taking values in the two compact subsets $\mathcal{X} \subset \mathbb{R}^p (p \geq 1)$ and $\mathcal{Z} \subset \mathbb{R}^d (d \geq 1)$, respectively; the parameter β_0 is an unknown $p \times 1$ vector that belongs to a compact subset $\Theta_\beta \subset \mathbb{R}^p$; and $g_0(\cdot)$ is an unknown smooth function valued in the space of functions $\mathcal{G} = \left\{ g \in C^2(\mathcal{Z}) : \|g\| = \sup_{z \in \mathcal{Z}} |g(z)| < C \right\}$, with $C^2(\mathcal{Z})$ the space of twice differentiable functions from $\mathcal{Z}$ to $\mathbb{R}$ and C a positive constant. In model (1.1), β_0 and $g_0(\cdot)$ are constant over i (and n). Assume that the disturbance term U_{in} in (1.2) is modelled by the following spatial autoregressive process (SAR):

$$U_{in} = \lambda_0 \sum_{j=1}^{n} w_{ijn} U_{jn} + \varepsilon_{in}, \qquad 1 \leq i \leq n,\ n = 1, 2, \ldots \tag{1.3}$$

where, we assume that, for all $n = 1, 2, \ldots$, $\{\varepsilon_{in},\ 1 \leq i \leq n\}$ is independent of $\{X_{in},\ 1 \leq i \leq n\}$ and $\{Z_{in},\ 1 \leq i \leq n\}$, and $\{X_{in},\ 1 \leq i \leq n\}$ is independent of $\{Z_{in},\ 1 \leq i \leq n\}$.

λ_0 is the autoregressive parameter, valued in the compact subset $\Theta_\lambda \subset \mathbb{R}$, w_{ijn}, $j = 1, \ldots, n$ are the elements in the i–th row of a non-stochastic $n \times n$ spatial weight matrix W_n, which contains the information on the spatial relationship between observations. This spatial weight matrix is usually constructed as a function of the distances (with respect to some metric) between locations; see [30] for additional details. The $n \times n$ matrix $(I_n - \lambda_0 W_n)$ is assumed to be non-singular for all n, where I_n denotes the $n \times n$ identity matrix and $\{\varepsilon_{in},\ 1 \leq i \leq n\}$ are assumed to be independent random Gaussian variables; $\mathbb{E}(\varepsilon_{in}) = 0$ and $\mathbb{E}(\varepsilon_{in}^2) = 1$ for $i = 1, \ldots, n$ $n = 1, 2, \ldots$. Note that one can rewrite (1.3) as

$$U_n = (I_n - \lambda_0 W_n)^{-1} \varepsilon_n, \quad n = 1, 2, \ldots \tag{1.4}$$

where $U_n = (U_{n1}, \ldots, U_{nn})^T$ and $\varepsilon_n = (\varepsilon_{n1}, \ldots, \varepsilon_{nn})^T$. Therefore, the variance-covariance matrix of U_n is

$$V_n(\lambda_0) \equiv \mathrm{Var}(U_n) = (I_n - \lambda_0 W_n)^{-1} \left\{ (I_n - \lambda_0 W_n)^T \right\}^{-1}, \quad n = 1, 2, \ldots \tag{1.5}$$

This matrix allows one to describe the cross-sectional spatial dependencies between the n observations. Furthermore, the fact that the diagonal elements of $V_n(\lambda_0)$ depend on λ_0 and particularly on i and n allows some spatial heteroscedasticity. These spatial dependences and heteroscedasticity depend on the neighbourhood structure established by the spatial weight matrix W_n.
The elements w_{ijn} of W_n are usually considered as inversely proportional to the distance between spatial units i and j with respect to some metric [physical distance, social network or economic distance, see for instance 30]. The matrices W_n are usually classified into two groups: *Weights Based on Distance* and *Weights Based on Boundaries*. For Weights Based on Distance, the distance d_{ij} between each pair of spatial units (regions, cities, centroids,...) i and j are basically considered.

- *k-Nearest Neighbor weights*
  $w_{ij} = \begin{cases} 1 & \texttt{if } j \in N_k(i), \\ 0 & \texttt{Otherwise} \end{cases}$ where $N_k(i)$ is the set of the k closest units or regions to i for $k \in \{1, ..., n-1\}$
- *Power Distance Decay weights*
  $w_{ij} = \begin{cases} d_{ij}^{-\alpha} & \texttt{if } 0 \leq d_{ij} \leq \delta \\ 0 & \texttt{if } d_{ij} > \delta \end{cases}$, where α is any positive exponent, typically $\alpha = 1$ or $\alpha = 2$.

For Weights Based on Boundaries, spatial contiguity is often used to specify neighboring location in the sense of sharing a common border. There are different type of spatial contiguity but the classical cases are those referred to *Rook contiguity* (with only common boundaries), *Bishop contiguity* (with only common vertices) and *Queen contiguity* (with both Rook and Bishop contiguity).

$$w_{ij} = \begin{cases} 1 & \texttt{if } i \texttt{ and } j \texttt{ are contiguity} \\ 0 & \texttt{Otherwise} \end{cases}$$

In general, we can rewrite the last equation as:

$$w_{ij} = \begin{cases} 1 & \ell_{ij} > 0 \\ 0 & \ell_{ij} = 0 \end{cases},$$

with ℓ_{ij} denotes the length of shared boundary.
Before proceeding further, let us give some particular cases of the model.
If one consider i.i.d observations, that is, $V_n(\lambda_0) = \sigma^2 I_n$, with σ depending on λ_0, the obtained model may be viewed as a special case of classical generalized partially linear models [e.g. 34] or the classical generalized additive model [14]. Several

approaches for estimating this particular model have been developed; among these methods, we cite that of [34] based on the concept of the generalized profile likelihood [e.g 35]. This approach consists of first fixing the parametric parameter β and non-parametrically estimating $g_0(\cdot)$ using the weighted likelihood method. This last estimate is then used to construct a profile likelihood to estimate β_0.
When $g_0 \equiv 0$ (or is an affine function), that is, without a non-parametric component, several approaches have been developed to estimate the parameters β_0 and λ_0. The basic difficulty encountered is that the likelihood function of this model involves an n-dimensional normal integral; thus, when n is high, the computation or asymptotic properties of the estimates may present difficulties [e.g. 31]. Various approaches have been proposed to addressed this difficulty; among these approaches, we cite the following:

- Feasible Maximum Likelihood approach: this approach consists of replacing the true likelihood function by a pseudo-likelihood function constructed via marginal likelihood functions. [36] proposed a pseudo-likelihood function obtained by replacing $V_n(\lambda_0)$ by some diagonal matrix with the diagonal elements of $V_n(\lambda_0)$. Alternatively, [39] proposed to divide the observations by pairwise groups, where the latter are assumed to be independent with a bivariate normal distribution in each group, and estimate β_0 and λ_0 by maximizing the likelihood of these groups. Recently [25] proposed a pseudo-likelihood function defined as an approximation of the likelihood function where the latter is inspired by some univariate conditioning procedure.
- Generalized Method of Moments (GMM) approach used by [30]. These authors used the generalized residuals defined by $\tilde{U}_{in}(\beta, \lambda) = \mathbb{E}\left(U_{in} | Y_{in}, \beta, \lambda\right)$, $1 \leq i \leq n$, $n = 1, 2, \ldots$ with some instrumental variables to construct moment equations to define the GMM estimators of β_0 and λ_0.

In what follows, using the n observations (X_{in}, Y_{in}, Z_{in}), $i = 1, ..., n$, we propose parametric estimators of β_0, λ_0 and a non-parametric estimator of the smooth function $g_0(\cdot)$.
We give asymptotic results according to *increasing domain* asymptotic. This consists of a sampling structure whereby new observations are added at the edges (boundary points) compare to the *infill* asymptotic, which consists of a sampling structure whereby new observations are added in-between existing observations. A typical example of an increasing domain is lattice data. An infill asymptotic is appropriate when the spatial locations are in a bounded domain.

1.1. Estimation Procedure

We propose an estimation procedure based on a combination of a weighted likelihood method and a generalized method of moments. We first fix the parametric components β and λ of the model and estimate the non-parametric component using a weighted likelihood. The obtained estimate is then used to construct generalized

residuals, where the latter are combined with the instrumental variables to propose GMM parametric estimates. This approach will be described as follow.

By equation (1.2), we have

(1.6)
$$\mathbb{E}_0\left(Y_{in}|X_{in}, Z_{in}\right) = \Phi\left(\left(v_{in}(\lambda_0)\right)^{-1}\left(X_{in}^T\beta_0 + g_0(Z_{in})\right)\right), \quad 1 \le i \le n, \quad n = 1, 2, \ldots$$

where $\mathbb{E}_0$ denotes the expectation under the true parameters (i.e., β_0, λ_0 and $g_0(\cdot)$), $\Phi(\cdot)$ is the cumulative distribution function of a standard normal distribution, and $(v_{in}(\lambda_0))^2 = V_{iin}(\lambda_0)$, $1 \le i \le n$, $n = 1, 2, \cdots$ are the diagonal elements of $V_n(\lambda_0)$. For each $\beta \in \Theta_\beta$, $\lambda \in \Theta_\lambda$, $z \in \mathcal{Z}$ and $\eta \in \mathbb{R}$, we define the conditional expectation on Z_{in} of the log-likelihood of Y_{in} for $1 \le i \le n$, $n = 1, 2, \ldots$, as

$$H(\eta; \beta, \lambda, z) = \mathbb{E}_0\left(\mathcal{L}\left(\Phi\left(\left(v_{in}(\lambda)\right)^{-1}\left(\eta + X_{in}^T\beta\right)\right); Y_{in}\right)\middle| Z_{in} = z\right), \tag{1.7}$$

with $\mathcal{L}(u; v) = \log\left(u^v(1-u)^{1-v}\right)$. Note that $H(\eta; \beta, \lambda, z)$ is assumed to be constant over i (and n). For each fixed $\beta \in \Theta_\beta$, $\lambda \in \Theta_\lambda$ and $z \in \mathcal{Z}$, $g_{\beta,\lambda}(z)$ denotes the solution in η of

$$\frac{\partial}{\partial \eta} H(\eta; \beta, \lambda, z) = 0. \tag{1.8}$$

Then, we have $g_{\beta_0,\lambda_0}(z) = g_0(z)$ for all $z \in \mathcal{Z}$.

Now, using $g_{\beta,\lambda}(\cdot)$, we construct the GMM estimates of β_0 and λ_0 as in [30]. For that, we define the generalized residuals, replacing $g_0(Z_{in})$ in (1.1) by $g_{\beta,\lambda}(Z_{in})$:

$$\begin{aligned} \tilde{U}_{in}(\beta, \lambda, g_{\beta,\lambda}) &= \mathbb{E}\left(U_{in}|Y_{in}, \beta, \lambda\right) \\ &= \frac{\phi\left(G_{in}(\beta, \lambda, g_{\beta,\lambda})\right)\left(Y_{in} - \Phi\left(G_{in}(\beta, \lambda, g_{\beta,\lambda})\right)\right)}{\Phi\left(G_{in}(\beta, \lambda, g_{\beta,\lambda})\right)\left(1 - \Phi\left(G_{in}(\beta, \lambda, g_{\beta,\lambda})\right)\right)}, \end{aligned} \tag{1.9}$$

where $\phi(\cdot)$ is the density of the standard normal distribution and $G_{in}(\beta, \lambda, g_{\beta,\lambda}) = \left(v_{ni}(\lambda)\right)^{-1}\left(X_{in}^T\beta + g_{\beta,\lambda}(Z_{in})\right)$.
For simplicity of notation, we write $\theta = (\beta^T, \lambda)^T \in \Theta \equiv \Theta_\beta \times \Theta_\lambda$ when possible. Note that in (1.9), the generalized residual $\tilde{U}_{in}(\cdot, \cdot)$ is calculated by conditioning only on Y_{in} and not on the entire sample $\{Y_{in}, i = 1, 2, \ldots, n, n = 1, \ldots\}$ or a subset of it. This of course will influence the efficiency of the estimators of θ obtained by these generalized residuals, but it allows one to avoid a complex computation; see [31] for additional details. To address this loss of efficiency, let us follow [30]'s procedure, which consists of employing some instrumental variables to create some moment conditions, and use a random matrix to define a criterion function. Both the instrumental variables and the random matrix permit one to consider more information about the spatial dependences and heteroscedasticity characterizing the dataset. Let us now detail the estimation procedure. Let

$$S_n(\theta, g_\theta) = n^{-1}\xi^T\tilde{U}_n(\theta, g_\theta). \tag{1.10}$$

where $\tilde{U}_n(\theta, g_\theta)$ is an $n \times 1$ vector, composed of $\tilde{U}_{in}(\theta, g_\theta)$, $1 \leq i \leq n$ and ξ_n is an $n \times q$ matrix of instrumental variables, whose ith row is given by the $1 \times q$ random vector ξ_{in}. The latter may depend on $g_\theta(\cdot)$ and θ. We assume that ξ_{in} is $\sigma(X_{in}, Z_{in})$, measurable for each $i = 1, \ldots, n$, $n = 1, 2, \ldots$. We suppress the possible dependence of the instrumental variables on the parameters for notational simplicity. The GMM approach consists of minimizing the following sample criterion function:

$$Q_n(\theta, g_\theta) = S_n^T(\theta, g_\theta) M_n S_n(\theta, g_\theta), \tag{1.11}$$

where M_n is some positive-definite $q \times q$ weight matrix that may depend on the sample information. The choice of the instrumental variables and weight matrix characterizes the difference between GMM estimator and all pseudo-maximum likelihood estimators. For instance, if one takes

$$\xi_{in}(\theta, g_\theta) = \frac{\partial G_{in}(\theta, \eta_i)}{\partial \theta} + \frac{\partial G_{in}(\theta, \eta_i)}{\partial \eta} \frac{\partial g_\theta}{\partial \theta}(Z_{in}), \tag{1.12}$$

with $\eta_i = g_\theta(Z_{in})$, $G_{in}(\theta, \eta_i) = (v_{in}(\lambda))^{-1} \left(X_{in}^T \beta + \eta_i\right)$, and $M_n = I_q$ with $q = p + 1$, then the GMM estimator of θ is equal to a pseudo-maximum profile likelihood estimator of θ, accounting only for the spatial heteroscedasticity.
Now, let

$$S(\theta, g_\theta) = \lim_{n \to \infty} \mathbb{E}_0 \left(S_n(\theta, g_\theta) \right), \tag{1.13}$$

and

$$Q(\theta, g_\theta) = S^T(\theta, g_\theta) M S(\theta, g_\theta),$$

where M, the limit of the sequence M_n, is a nonrandom positive-definite matrix. The functions $S_n(\cdot, \cdot)$ and $Q_n(\cdot, \cdot)$ are viewed as empirical counterparts of $S(\cdot, \cdot)$ and $Q(\cdot, \cdot)$, respectively.
Clearly, $g_\theta(\cdot)$ is not available in practice. However, we need to estimate it, particularly by an asymptotically efficient estimate. By (1.8) and for fixed $\theta^T = (\beta^T, \lambda) \in \Theta$, an estimator of $g_\theta(z)$, for $z \in \mathcal{Z}$, can be given by $\hat{g}_\theta(z)$, which denotes the solution in η of

$$\sum_{i=1}^{n} \frac{\partial}{\partial \eta} \mathcal{L} \left(\Phi \left(G_{in}(\theta, \eta) \right); Y_{in} \right) K \left(\frac{z - Z_{in}}{b_n} \right) = 0, \tag{1.14}$$

where $K(\cdot)$ is a kernel from $\mathbb{R}^d$ to $\mathbb{R}_+$ and b_n is a bandwidth depending on n.

Now, replacing $g_\theta(\cdot)$ in (1.11) by the estimator $\hat{g}_\theta(\cdot)$ permits one to obtain the GMM estimator $\hat{\theta}$ of θ as

$$\hat{\theta} = \operatorname{argmin}_{\theta \in \Theta} Q_n(\theta, \hat{g}_\theta). \tag{1.15}$$

A classical inconvenience of the estimator $\hat{g}_\theta(z)$ proposed in (1.14) is that the bias of $\hat{g}_\theta(z)$ is high for z near the boundary of $\mathcal{Z}$. Of course, this bias will affect the estimator of θ given in (1.15) when some of the observations Z_{in} are near the boundary of $\mathcal{Z}$. A local linear method, or more generally the local polynomial method [11], can be used to reduce this bias. Another alternative is to use *trimming* [34], in which the function $S_n(\theta, g_\theta)$ is computed using only observations associated with Z_{in} that are away from the boundary. The advantage of this approach is that the theoretical results can be presented in a clear form, but it is less tractable from a practical point of view, in particular, for small sample sizes.

2. Large sample properties

We now turn to the asymptotic properties of the estimators derived in the previous section: $\hat{\theta}^T = (\hat{\beta}^T, \hat{\lambda})$ and $\hat{g}_{\hat{\theta}}(\cdot)$. Let us use the following notation: $\frac{d}{d\theta} S(\theta, g_\theta)$ means that we differentiate $S(.,.)$ with respect to θ, and $\frac{\partial}{\partial \theta} S(\theta, g_\theta)$ is the partial derivative of $S(\cdot,\cdot)$ w.r.t the first variable. The partial derivative of $S_n(\theta, g)$ w.r.t g, for any function $v \in \mathcal{G}$, is

$$\frac{\partial S_n}{\partial g}(\theta, g)(v) = n^{-1} \sum_{i=1}^{n} \xi_{in} \frac{\partial \tilde{U}_{in}}{\partial \eta}(\theta, \eta_i) v(Z_{in}).$$

Without ambiguity, $\|a\|$ denotes $\sup_t |a(t)|$ when a is a function, $\left(\sum a_i^2\right)^{1/2}$ when a is a vector, and $\left(\sum\sum a_{ij}^2\right)^{1/2}$ when a is a matrix.
Let the following matrices be needed in the asymptotic variance-covariance matrix of $\hat{\theta}$:

$$B_1(\theta_0) = \lim_{n\to\infty} \mathbb{E}_0 \left(n S_n(\theta_0, g_0) S_n^T(\theta_0, g_0) \right),$$

$$B_2(\theta_0) = \left\{ \frac{d}{d\theta} S^T(\theta, g_\theta) \Big|_{\theta=\theta_0} \right\} M \left\{ \frac{d}{d\theta} S(\theta, g_\theta) \Big|_{\theta=\theta_0} \right\},$$

with

$$\frac{d}{d\theta} S(\theta, g_\theta) = \frac{\partial S}{\partial \theta}(\theta, g_\theta) + \frac{\partial S}{\partial g}(\theta, g_\theta) \frac{\partial}{\partial \theta} g_\theta, \tag{2.1}$$

and

$$\Omega(\theta_0) = \{B_2(\theta_0)\}^{-1} \left\{ \frac{d}{d\theta} S^T(\theta, g_\theta) \Big|_{\theta=\theta_0} \right\} M B_1(\theta_0) M \left\{ \frac{d}{d\theta} S(\theta, g_\theta) \Big|_{\theta=\theta_0} \right\} \{B_2(\theta_0)\}^-$$

The following assumptions are required to establish the asymptotic results.
Assumption A1. (Smoothing condition). For each fixed $\theta \in \Theta$ and $z \in \mathcal{Z}$, let

$g_\theta(z)$ denote the unique solution with respect to η of

$$\frac{\partial}{\partial \eta} H(\eta; \theta, z) = 0.$$

For any $\varepsilon > 0$ and $g \in \mathcal{G}$, there exists $\gamma > 0$ such that

$$\sup_{\theta \in \Theta, z \in \mathcal{Z}} \left| \frac{\partial}{\partial \eta} H(g(z); \theta, z) \right| \leq \gamma \quad \Longrightarrow \quad \sup_{\theta \in \Theta, z \in \mathcal{Z}} |g(z) - g_\theta(z)| \leq \varepsilon. \tag{2.2}$$

Assumption A2. (Marginal distributions). The density $f_{in}(\cdot)$ of Z_{in} exists, is continuous on $\mathcal{Z}$ uniformly on i and n and satisfies

$$\liminf_{n \to \infty} \inf_{z \in \mathcal{Z}} \frac{1}{n} \sum_{i=1}^{n} f_{in}(z) > 0. \tag{2.3}$$

The joint probability density $f_{ijn}(.,.)$ of (Z_{in}, Z_{jn}) exists and is bounded on $\mathcal{Z} \times \mathcal{Z}$ uniformly on $i \neq j$ and n.

Assumption A3. (Spatial dependence). Let $h_{in}^{\theta, \eta_i}(\cdot|\cdot, \cdot)$ denote the conditional log likelihood function of Y_{in} given (X_{in}, Z_{in}), where $\eta_i = g(Z_{in})$. Let T_{in} be the vector (Y_{in}, X_{in}, Z_{in}), $i = 1, \ldots, n$, $n = 1, 2 \ldots$, $\tilde{p} = p + 1$, and assume that for all $i, l = 1, \ldots, n$,

$$|\mathrm{Cov}_0(\psi(T_{in}), \psi(T_{ln}))| \leq \{\mathrm{Var}_0(\psi(T_{in}))\, \mathrm{Var}_0(\psi(T_{ln}))\}^{1/2} \alpha_{iln}, \tag{2.4}$$

with

$$\psi(T_{in}) = K\left(\frac{z - Z_{in}}{b_n}\right) \quad \text{or}$$

$$\psi(T_{in}) = K\left(\frac{z - Z_{in}}{b_n}\right) \frac{\partial^{j_1 + \cdots + j_{\tilde{p}} + r}}{\partial \theta_1^{j_1} \cdots \partial \theta_{\tilde{p}}^{j_{\tilde{p}}} \partial \eta^r} h_{in}^{\theta, \eta}(Y_{in} | X_{in}, Z_{in} = z),$$

for all $z \in \mathcal{Z}$, $\theta \in \Theta, \eta = g(z)$ with $g \in \mathcal{G}$, and for all nonnegative integers $j_1, \ldots, j_{\tilde{p}} = 0, 1, 2$ and $r = 0, \ldots, 4$, such that $j_1 + \cdots + j_{\tilde{p}} + r \leq 6$.

We assume that

$$\left|\mathrm{Cov}_0\left(\xi_{itn}\tilde{U}_{in}(\theta, g_\theta), \xi_{jsn}\tilde{U}_{jn}(\theta, g_\theta)\right)\right| \leq \left\{\mathrm{Var}_0\left(\xi_{itn}\tilde{U}_{in}(\theta, g_\theta)\right) \mathrm{Var}_0\left(\xi_{jsn}\tilde{U}_{jn}(\theta, g_\theta)\right)\right\}^{1/2} \alpha_{ijn}, \tag{2.5}$$

for all $\theta \in \Theta$, $i, j = 1, \ldots, n$, $n = 1, 2, \ldots$ and for any $s, t = 1, \ldots, q$,

and

$$\left|\mathrm{Cov}_0\left(\xi_{in}^{(2)}(\theta_0, \eta_i^0), \xi_{jn}^{(2)}(\theta_0, \eta_j^0)\right)\right| \leq \left\{\mathrm{Var}_0\left(\xi_{in}^{(2)}(\theta_0, \eta_i^0)\right) \mathrm{Var}_0\left(\xi_{jn}^{(2)}(\theta_0, \eta_j^0)\right)\right\}^{1/2} \alpha_{ijn}, \tag{2.6}$$

with

$$\xi_{in}^{(2)}(\theta_0, \eta_i^0) := w^T \xi_i \Lambda\left(G_{in}(\theta_0, \eta_i^0)\right) \phi\left(G_{in}(\theta_0, \eta_i^0)\right) \frac{\partial G_{in}}{\partial \theta}(\theta_0, \eta_i^0),$$

where $\eta_i^0 = g_0(Z_i)$ for each $w \in \mathbb{R}^q$ such that $\|w\| = 1$.

In addition, assume that there is a decreasing (to 0) positive function $\varphi(\cdot)$ such that the "mixing" numbers verify $\alpha_{ijn} = O(\varphi(\|s_i - s_j\|))$, $r^2\varphi(rr^*)/\varphi(r^*) = o(1)$, as $r \to 0$, for all fixed $r^* > 0$, where s_i and s_j are spatial coordinates associated with observations i and j, respectively.

Assumption A4. The kernel K satisfies $\int K(u)du = 1$. It is Lipschitzian, i.e., there is a positive constant C such that

$$|K(u) - K(v)| \leq C\|u - v\| \qquad \text{for all } u, v \in \mathbb{R}^d.$$

Assumption A5. The bandwidth b_n satisfies $b_n \to 0$ and $nb_n^{3d+1} \to \infty$ as $n \to \infty$.

Assumption A6. The instrumental variables satisfy $\sup_{i,n} \|\xi_{in}\| = O_p(1)$, where ξ_{in} is the i-th column of the $n \times q$ matrix of instrumental variables ξ_n.

Assumption A7. $\theta^T = (\beta^T, \lambda)$ takes values in a compact and convex set $\Theta = \Theta_\beta \times \Theta_\lambda \subset \mathbb{R}^p \times \mathbb{R}$, and $\theta_0^T = (\beta_0^T, \lambda_0)$ is in the interior of Θ.

Assumption A8. $S(\cdot,\cdot)$ is continuous on both arguments θ and g, and $Q(\cdot, g.)$ attains a unique minimum over Θ at θ_0.

Assumption A9. The square root of the diagonal elements of $V_n(\lambda)$ are twice continuous differentiable functions with respect to λ and

$$\sup_{\lambda \in \Theta_\lambda} \left| v_{in}^{-1}(\lambda) + \frac{d}{d\lambda} v_{in}(\lambda) + \frac{d^2}{d\lambda^2} v_{in}(\lambda) \right| < \infty \text{ uniformly on } i \text{ and } n.$$

Assumption A10. $B_1(\theta_0)$ and $B_2(\theta_0)$ are positive-definite matrices, and $M_n - M = o_p(1)$.

Remark 1. *Assumption A1 ensures the smoothness of $H(.;.,.)$ around its extrema point $g_\theta(.)$; see [34]. Assumption A2 is a decay of the local independence condition of the covariates Z_{in}, meaning that these variables are not identically distributed; a similar condition can be find in [32]. Condition (2.3) generalizes the classical assumption $\inf_z f(z) > 0$ used in the case of estimating the density function $f(\cdot)$ with identically distributed or stationary random variables. This assumption has been used in [32] (**Assumption A7**(x), p. 8). Assumption A3 describes the spatial dependence structure, it is a particular case of the Assumption A in [28] and may be verified by mixing random variables, see [28] for more details. Note that the processes that we use are not assumed stationary; this allows for greater generalizability and the dependence structure to change with the sample size n (see [28] for more discussion). Conditions (2.4), (2.5) and (2.6) are not restrictive. When the regressors and instrumental variables are deterministic, conditions (2.4) and (2.5) are equivalent to $|Cov_0(Y_{in}, Y_{ln})| \leq \alpha_{iln}$. The condition on $\varphi(\cdot)$ is satisfied when the latter tends to zero at a polynomial rate, i.e., $\varphi(t) = O(t^{-\tau})$, for all $\tau > 2$, as in the case of mixing*

random variables.
Assumption A6 requires that the instruments and explanatory variables be bounded uniformly on i and n. In addition, when the instruments depend on θ and $g(\cdot)$, they are also uniformly bounded with respect to these parameters. The compactness condition in Assumption A7 is standard, and the convexity is somewhat unusual; however, it is reasonable in most applications. Condition A8 is necessary to ensure the identification of the true parameters θ_0. Assumption A9 requires the standard deviations of the errors to be uniformly bounded away from zero with bounded derivatives. This has been considered by [30]. Assumption A10 is classic ([30]) and required in the proof of Theorem 2.2. Those authors noted that in their model (without a non-parametric component), when the autoregressive parameter $\lambda_0 = 0$, $B_2(\theta_0)$ is not invertible, regardless of the choice of M_n. This is also the case in our context because for each $g_\theta(z)$ solution of (1.8), $\theta \in \Theta$ and $z \in \mathcal{Z}$, we have

$$\frac{\partial g_\theta}{\partial \beta}(z) = -\frac{E\left(\Gamma_{jn}(\theta, g_\theta(z)) X_{jn} \middle| Z_{jn} = z\right)}{E\left(\Gamma_{jn}(\theta, g_\theta(z)) \middle| Z_{jn} = z\right)},$$

and

$$\frac{\partial g_\theta}{\partial \lambda}(z) = \frac{v_{jn}'(\lambda)}{v_{jn}(\lambda)} \frac{E\left(\Gamma_{jn}(\theta, g_\theta(z)) \left(X_{jn}^T \beta + g_\theta(z)\right) \middle| Z_{jn} = z\right)}{E\left(\Gamma_{jn}(\theta, g_\theta(z)) \middle| Z_{jn} = z\right)} =$$

$$\frac{v_{jn}'(\lambda)}{v_{jn}(\lambda)} \left(g_\theta(z) - \beta^T \frac{\partial g_\theta}{\partial \beta}(z)\right),$$

where $v_{jn}'(\lambda) = \frac{d}{d\lambda} v_{jn}(\lambda) = v_{jn}(\lambda) \left[W_n S_n^{-1}(\lambda) V_n(\lambda)\right]_{jj}$,

$$\Gamma_{jn}(\cdot) = \Lambda'\left(G_{jn}(\cdot)\right)\left[Y_{jn} - \Phi(G_{jn}(\cdot))\right] - \Lambda\left(G_{jn}(\cdot)\right)\phi\left(G_{jn}(\cdot)\right)$$

and $\Lambda(\cdot) = \phi(\cdot)/(1 - \Phi(\cdot))\Phi(\cdot)$. *However*

$$\left.\frac{\partial g_\theta}{\partial \lambda}(z)\right|_{\lambda=0} = 0 \qquad \textit{because} \qquad v_{jn}'(0) = 0,$$

then $B_2(\theta_0)$ will be singular when $\lambda_0 = 0$.

With these assumptions in place, we are able to give some asymptotic results. The weak consistencies of the proposed estimators are given in the following two results. The first theorem and corollary below establish the consistency of our estimators, whereas the second theorem addresses the question of convergence to a normal distribution of the parametric component when it is properly standardized.

Theorem 2.1. *Under Assumptions A1-A10, we have*

$$\hat{\theta} - \theta_0 = o_p(1).$$

Corollary 2.1. *If the assumptions of Theorem 2.1 are satisfied, then we have*

$$\left\|\hat{g}_{\hat{\theta}} - g_0\right\| = o_p(1).$$

Proof of Corollary 2.1 Note that

$$\begin{aligned}\left\|\hat{g}_{\hat{\theta}} - g_0\right\| &\leq \left\|\hat{g}_{\hat{\theta}} - g_{\hat{\theta}}\right\| + \left\|g_{\hat{\theta}} - g_0\right\| \\ &\leq \sup_{\theta}\left\|\hat{g}_{\theta} - g_{\theta}\right\| + \sup_{\theta}\left\|\frac{\partial g_{\theta}}{\partial \theta}\right\| \left\|\hat{\theta} - \theta_0\right\| = o_p(1),\end{aligned}$$

since, by the assumptions of Theorem 2.1, $\sup_{\theta}\left\|\hat{g}_{\theta} - g_{\theta}\right\| = o_p(1)$ and $\sup_{\theta}\left\|\frac{\partial g_{\theta}}{\partial \theta}\right\| < \infty$.

The following gives an asymptotic normality result of $\hat{\theta}$.

Theorem 2.2. *Under assumptions A1-A10, we have*

$$\sqrt{n}\left(\hat{\theta} - \theta_0\right) \rightarrow_{law} \mathcal{N}\left(0, \Omega(\theta_0)\right)$$

Remark 2. *In practice, the previous asymptotic normality result can be used to construct asymptotic confidence intervals and build hypothesis tests when a consistent estimate of the asymptotic covariance matrix $\Omega(\theta_0)$ is available. To estimate this matrix, let us follow the idea of [30] and define the estimator*

$$\Omega_n(\hat{\theta}) = \left\{B_{2n}(\hat{\theta})\right\}^{-1}\left\{\frac{d}{d\theta}S_n^T(\theta, \hat{g}_{\theta})\Big|_{\theta=\hat{\theta}}\right\} M_n B_{1n}(\hat{\theta}) M_n \left\{\frac{d}{d\theta}S_n(\theta, \hat{g}_{\theta})\Big|_{\theta=\hat{\theta}}\right\}\left\{B_{2n}(\hat{\theta})\right\}^{-1},$$

with

$$B_{1n}(\theta) = nS_n(\theta, \hat{g}_{\theta})S_n^T(\theta, \hat{g}_{\theta}) \quad \text{and} \quad B_{2n}(\theta) = \left\{\frac{d}{d\theta}S_n^T(\theta, \hat{g}_{\theta})\right\} M_n \left\{\frac{d}{d\theta}S_n(\theta, \hat{g}_{\theta})\right\}.$$

The consistency of $\Omega_n(\hat{\theta})$ will be based on that of $B_{1n}(\hat{\theta})$ and $B_{2n}(\hat{\theta})$, the estimators of $B_1(\theta_0)$ and $B_2(\theta_0)$, respectively. Note that the consistency of $B_{2n}(\hat{\theta})$ is relatively easy to establish. On the other hand, that of $B_{1n}(\hat{\theta})$ asks for additional assumptions and an adaption of the proof of Theorem 3 of [30, p.134] to our case; this is of interest to future research.

3. Computation of the estimates

The aim of this section is to outline in detail how the regression parameters β, the spatial auto-correlation parameter λ and the non-linear function g_{θ} can be estimated. We begin with the computation of $\hat{g}_{\theta}(z)$, which will play a crucial role in what follows.

3.1. *Computation of the estimate of the non-parametric component*

An iterative method is needed to compute the $\hat{g}_\theta(z)$ solution of (1.14) for each fixed $\theta \in \Theta$ and $z \in \mathcal{Z}$. For fixed $\theta^T = (\beta, \lambda) \in \Theta$ and $z \in \mathcal{Z}$, let $\eta_\theta = g_\theta(z)$ and $\psi(\eta; \theta, z)$ denote the left-hand side of (1.14), which can be rewritten as

$$\psi(\eta;\theta,z) = \sum_{i=1}^{n} [v_{in}(\lambda)]^{-1} \Lambda\left(G_{in}(\theta,\eta)\right) \left[Y_{in} - \Phi(G_{in}(\theta,\eta))\right] K\left(\frac{z - Z_{in}}{b_n}\right). \tag{3.1}$$

Consider the Fisher information:

$$\begin{aligned}
\Psi(\eta_\theta;\theta,z) &= E_0\left(\frac{\partial}{\partial\eta}\psi(\eta;\theta,z)\Big|_{\eta=\eta_\theta} \Big| \{(X_{in}, Z_{in}),\ 1 \leq\ i \leq n,\ n = 1, \ldots\}\right) \\
&= -\sum_{i=1}^{n} [v_{in}(\lambda)]^{-2} \Lambda\left(G_{in}(\theta,\eta_\theta)\right) \phi\left(G_{in}(\theta,\eta_\theta)\right) K\left(\frac{z - Z_{in}}{b_n}\right) + \\
&\quad + \sum_{i=1}^{n} [v_{in}(\lambda)]^{-2} \Lambda'\left(G_{in}(\theta,\eta_\theta)\right) \left[\Phi\left(G_{in}(\theta_0,\eta_0)\right) - \Phi\left(G_{in}(\theta,\eta_\theta)\right)\right] K\left(\frac{z - Z_{in}}{b_n}\right) \qquad (3.2)
\end{aligned}$$

Note that the second term in the RHS (Right Hand Side) of (3.2) is negligible when θ is near the true parameter θ_0.
Because $\psi(\eta; \theta, z) = 0$ for $\eta = \hat{g}_\theta(z)$, an initial estimate $\tilde{\eta}$ can be updated to $\eta^\dagger$ using Fisher's scoring method:

$$\eta^\dagger = \tilde{\eta} - \frac{\psi(\tilde{\eta};\theta,z)}{\Psi(\tilde{\eta};\theta,z)}. \tag{3.3}$$

The iteration procedure (3.3) requests some starting value $\tilde{\eta} = \tilde{\eta}_0$ to ensure convergence of the algorithm. To this end, let us adapt the approach of [34], which consists of supposing that for fixed $\theta \in \Theta$, there exists a $\tilde{\eta}_0$ satisfying $G_{in}(\theta, \tilde{\eta}_0) = \Phi^{-1}(Y_{in})$ for $i = 1, \ldots, n$. Knowing that $G_{in}(\theta, \tilde{\eta}_0) = (v_{in}(\lambda))^{-1} \left(X_{ni}^T \beta + \tilde{\eta}_0\right)$, we have $\tilde{\eta}_0 = v_{in}(\lambda)\Phi^{-1}(Y_{in}) - X_{in}^T\beta$. Then, (3.3) can be updated using the following initial value:

$$\eta_0^\dagger = \tilde{\eta}_0 - \frac{\psi(\tilde{\eta}_0;\theta,z)}{\Psi(\tilde{\eta}_0;\theta,z)} = \frac{\sum_{i=1}^{n} [v_{in}(\lambda)]^{-1} \Lambda(C_{in})\phi(C_{in}) \left[C_{in} - [v_{in}(\lambda)]^{-1} X_{in}^T\beta\right] K\left(\frac{z - Z_{in}}{b_n}\right)}{\sum_{i=1}^{n} [v_{in}(\lambda)]^{-2} \Lambda(C_{in})\phi(C_{in}) K\left(\frac{z - Z_{in}}{b_n}\right)},$$

where $C_{in} = \Phi^{-1}(Y_{in})$, $i = 1, \ldots, n$, is computed using a slight adjustment because $Y_{in} \in \{0, 1\}$.
With this initial value, the algorithm iterates until convergence.

Selection of the bandwidth

A critical step (in non- or semi-parametric models) is the choice of the bandwidth parameter b_n, which is usually selected by applying some cross-validation approach.

The latter was adapted by [38] in the case of a spatial semi-parametric model. Because cross-validation may be very time consuming, which is true in the case of our model, we adapt the following approach used in [34] to achieve greater flexibility:

1. Consider the linear regression of C_{in} on X_{in}, $i = 1, \ldots, n$, without an intercept term, and let $R_{1n}, \ldots, R_{nn}$ denote the corresponding residuals.
2. Since we expect $\mathbb{E}(R_{in}|Z_{in} = z)$ to have similar smoothness properties as $g_0(.)$, the optimal bandwidth b_n is that of the non-parametric regression of the $\{R_{in}\}_{i=1,\cdots,n}$ on $\{Z_{in}\}_{i=1,\cdots,n}$, chosen by applying any non-parametric regression bandwidth selection method. For that, we use the cross-validation method in the *np* R Package.

3.2. Computation of $\hat{\theta}$

The parametric component β and the spatial autoregressive parameter λ are computed as mentioned above by a GMM approach based on some instrumental variables ξ_n and the weight matrix M_n. The choices of these instrumental variables and weight matrix M_n are as follows.
Because $\psi(\hat{g}_\theta(z); \theta, z) = 0$, if we differentiate the latter with respect to β and λ, we have

$$\frac{\partial}{\partial \beta}\hat{g}_\theta(z) = -\frac{\sum_{i=1}^{n}\left[v_{in}(\lambda)\right]^{-2}\Delta_{in}(\theta, z)X_{in}K\left(\frac{z-Z_{in}}{b_n}\right)}{\sum_{i=1}^{n}\left[v_{in}(\lambda)\right]^{-2}\Delta_{in}(\theta, z)K\left(\frac{z-Z_{in}}{b_n}\right)},$$

and

$$\frac{\partial}{\partial \lambda}\hat{g}_\theta(z) = \frac{\sum_{i=1}^{n}\left[v_{in}(\lambda)\right]^{-1}v_{in}'(\lambda)\Delta_{in}(\theta, z)\left[X_{in}^T\beta + \hat{g}_\theta(z)\right]K\left(\frac{z-Z_{in}}{b_n}\right)}{\sum_{i=1}^{n}\left[v_{in}(\lambda)\right]^{-2}\Delta_{in}(\theta, z)K\left(\frac{z-Z_{in}}{b_n}\right)}$$
$$+\frac{\sum_{i=1}^{n}\left[v_{in}(\lambda)\right]^{-2}v_{in}'(\lambda)\Lambda\left(G_{in}(\theta, \hat{g}_\theta(z))\right)\left[Y_{in} - \Phi\left(G_{in}(\theta, \hat{g}_\theta(z))\right)\right]K\left(\frac{z-}{b}\right.}{\sum_{i=1}^{n}\left[v_{in}(\lambda)\right]^{-2}\Delta_{in}(\theta, z)K\left(\frac{z-Z_{in}}{b_n}\right)}$$

with

$$\Delta_{in}(\theta, z) = \Lambda'\left(G_{in}(\theta, \hat{g}_\theta(z))\right)\left[Y_{in} - \Phi\left(G_{in}(\theta, \hat{g}_\theta(z))\right)\right] - \Lambda\left(G_{ni}(\theta, \hat{g}_\theta(z))\right)\phi\left(G_{in}(\theta, \hat{g}_\theta(\right.$$

Then, the previous result is used to define the following instrumental variables:

$$\xi_{in}(\theta, \hat{g}_\theta) = \frac{\partial G_{in}(\theta, \hat{\eta}_i)}{\partial \theta} + \frac{\partial G_{in}(\theta, \hat{\eta}_i)}{\partial \eta}\frac{\partial}{\partial \theta}\hat{g}_\theta(Z_{in}),$$

with $\hat{\eta}_i = \hat{g}_\theta(Z_{in})$.
For the weight matrix, one can use $M_n = I_q$ with $q = p + 1$ as in [30]. Then, the obtained GMM estimator of θ with this choice of M_n is equal to the pseudo-profile

maximum likelihood estimator of θ, accounting only for the spatial heteroscedasticity. Another empirical choice could be the idea of continuous updating GMM estimator (One step GMM) used in [29]:

$$M_n(\theta) = \left\{ n^{-1} \sum_{i,j=1}^{n} \delta_{ij} \xi_{ni} \xi_{jn}^T \tilde{U}_{in}(\theta, \hat{g}_\theta) \tilde{U}_{jn}(\theta, \hat{g}_\theta) \right\}^{-1} \tag{3.4}$$

with the weights

$$\delta_{ij} = \frac{\sum_{r=1}^{n} \tau_{ri} \tau_{rj}}{\left[\sum_{r=1}^{n} \tau_{ri}^2 \sum_{r=1}^{n} \tau_{rj}^2 \right]^{1/2}} \qquad \text{for } i, j = 1, \ldots, n,$$

where τ_{ij} is a number depending on w_{nij} such that the nearer location i is to location j, the larger τ_{ij} is. For instance, we expect to have more efficient estimators with this matrix.

4. Finite sample properties

In this section, we study the performance of the proposed model based on some numerical results, which highlight the importance of accounting for both the spatial dependence and the partial linearity. Random datasets from the following spatial semi-parametric models are generated and first we investigate the estimation quality of the proposed procedure which accounts both the spatial dependence and the partial linearity. The influences of the spatial dependence and the partial linearity are investigated by comparing the behavior of our model to that of the non-spatial partially linear probit (NSPLP) model and the fully linear SAE probit (LSAEP) model, respectively. The *GAM* and *ProbitSpatial* [24] R packages will be used to provide the estimates associated to NSPLP and LSAEP models respectively. We generate observations from the following spatial latent partial linear model:

$$\begin{aligned} Y_{in}^* &= \beta_1 X_{in}^{(1)} + \beta_2 X_{in}^{(2)} + g(Z_{in}) + U_{in}; \qquad Y_{in} = \mathbb{I}(Y_{in}^* > 0),\ i = 1, \ldots, n \\ U_n &= (I_n - \lambda W_n)^{-1} \varepsilon_n \end{aligned}$$

where $U_n \sim \mathcal{N}(0, I_n)$ and W_n is the spatial weight matrix associated to n locations chosen randomly in a 60×60 regular grid and with elements constructed in such way that each location has at least 6 neighbors. The explanatory variables $X^{(1)}$ and $X^{(2)}$ are generated as pseudo $\mathcal{B}(0.7)$ and $\mathcal{U}[-2, 2]$, respectively, and the other explanatory variable Z is equal to the sum of 48 independent random variables, each uniformly distributed over $[-0.25, 0.25]$. Here, we use the non-linear function $g(t) = t + 2\cos(0.5\pi t)$ and parameters $\beta_1 = -1$, $\beta_2 = 1$. Different spatial dependence parameters λ; 0.2 (weak spatial dependence) 0.5 and 0.8 (strong spatial dependence) are considered. Finally, the sample size effect is observed by considering n equals to

200, 400 and 800 with 300 replications of each simulation.
Our estimation procedure is applied with a Gaussian kernel $K(t) = (2\pi^{-1/2})\exp(-t^2$, and optimal bandwidth b_n selected by [34]'s approach detailed previously.
We consider the trivial instrumental variables and two choices of matrix $M_n = I_n$ which leads to the pseudo-maximum profile likelihood estimators (named PLSP 1) and a second choice M_n given in (3.4) with components $\tau_{ij} = w_{nij}$, the estimates obtained with this matrix choice are denoted PLSP 2. The second choice of the weight matrix allows to incorporate more information about the spatial dependance.
The results are given in Table 1, the columns titles Mean, Median and SD give the average, median and standard deviation, respectively, over these 300 replications associated with each estimation method.

In one hand, when we compare the estimators (PLSP 1 and PLSP2) based on our approach (PLSPM) with those based on the LSAEP model, we notice that the latter yields more biased estimators of the coefficients β_1 and β_2. It makes sense that ignoring the partial linearity (see also Figure 1) weakens the quality of the estimation of the coefficients β_1 and β_2.

On the other hand, note that the LSAEP and PLSP 1 estimates are similar in case of low spatial dependence ($\lambda = 0.2$) compare to large spatial dependence ($\lambda = 0.8$) framework. It makes sense that ignoring a high spatial dependence does not allow a model that does not account any spatial structure to find consistent estimates of the coefficients β_1 and β_2 and the smooth function $g(\cdot)$ (see Figure 1) .
Note that the second choice of the weight matrix (estimates PLSP 2) allowed to improve the efficiency of the proposed estimates particularly in case of high spatial dependence (see PLSP 2 estimates in case of $\lambda = 0.8$). In contrast, it is less appropriate in case of low spatial dependance. However, one may think of testing the intensity of the spatial dependence before applying the proposed model with a non identity weight matrix, using for instance Moran's test [18].

Discussion

In this manuscript, we have proposed a spatial semi-parametric probit model for identifying risk factors at onset and with spatial heterogeneity. The parameters involved in the models are estimated using weighted likelihood and generalized method of moment methods. A technique based on dependent random arrays facilitates the estimation and derivation of asymptotic properties, which otherwise would have been difficult to perform due to the complexity introduced by the spatial dependence to the model and high-dimensional integration required by a full maximum likelihood approach. Moreover, the technique yields consistent estimates through proper choices of the bandwidth, weight matrix, and instrumental variables. The proposed models provide a general framework and tools for researchers and practitioners when addressing binary semi-parametric choice models in the presence of spatial correlation. Although they provide significant contributions to the body of knowledge, additional investigations need to be done.

TABLE 1

The mean, median and standard deviation (SD) of the parameters β_1, β_2, and λ estimates, over the 300 replications

λ	n	Methods	$\beta_1 = -1$			$\beta_2 = 1$			λ		
			Mean	Median	SD	Mean	Median	SD	Mean	Median	SD
0.20	200	PLSP 1	-1.06	-1.00	0.40	1.05	0.98	0.26	0.12	0.00	0.40
		PLSP 2	-1.06	-1.07	0.28	1.06	1.04	0.19	0.24	0.16	0.43
		LSAEP	-0.65	-0.65	0.20	0.67	0.67	0.10	-0.14	0.01	0.5
		NSPLP	-1.02	-1.00	0.22	1.02	1.00	0.11			
	400	PLSP 1	-1.01	-0.99	0.23	1.01	0.99	0.15	0.05	0,00	0.32
		PLSP 2	-1.08	-1.06	0.22	1.06	1.05	0.15	0.21	0.08	0.40
		LSAEP	-0.64	-0.66	0.15	0.66	0.66	0.06	-0.02	0.09	0.37
		NSPLP	-1.02	-1.00	0.22	1.02	1.00	0.11			
	800	PLSP 1	-0.99	-1.01	0.16	0.99	0.98	0.09	0.05	0.00	0.23
		PLSP 2	-1.06	-1.06	0.21	1.06	1.04	0.13	0.27	0.24	0.42
		LSAEP	-0.62	-0.62	0.12	0.65	0.64	0.05	0.01	0.05	0.29
		NSPLP	-1.01	-1.00	0.16	0.98	0.99	0.07			
0.50	200	PLSP 1	-1.10	-1.04	0.42	1.08	1.00	0.34	0.24	0.01	0.43
		PLSP 2	-1.06	-1.06	0.32	1.12	1.09	0.24	0.33	0.49	0.45
		LSAEP	-0.62	-0.62	0.12	0.65	0.64	0.05	0.01	0.05	0.29
		NSPLP	-1.00	-1.00	0.30	0.98	0.97	0.16			
	400	PLSP 1	-1.04	-1.01	0.30	1.04	0.98	0.23	0.23	0.01	0.36
		PLSP 2	-1.03	-1.01	0.25	1.06	1.03	0.18	0.33	0.42	0.42
		LSAEP	-0.62	-0.61	0.17	0.65	0.64	0.08	0.15	0.27	0.37
		NSPLP	-0.96	-0.94	0.24	0.97	0.97	0.11			
	800	PLSP 1	-0.96	-0.94	0.16	1,00	0.97	0.13	0.24	0.06	0.29
		PLSP 2	-1.02	-1.00	0.18	1.05	1.00	0.15	0.36	0.47	0.40
		LSAEP	-0.62	-0.60	0.12	0.65	0.65	0.05	0.27	0.30	0.19
		NSPLP	-0.98	-0.98	0.15	0.97	0.96	0.07			
0.80	200	PLSP 1	-1.11	-1.03	0.53	1.12	1,00	0.4	0.54	0.79	0.41
		PLSP 2	-0.99	-1.01	0.31	0.99	0.95	0.23	0.45	0.65	0.44
		LSAEP	-0.67	-0.67	0.26	0.65	0.65	0.12	0.47	0.54	0.24
		NSPLP	-0.86	-0.87	0.30	0.85	0.84	0.15			
	400	PLSP 1	-1.03	-0.97	0.36	1.06	0.95	0.35	0.52	0.70	0.39
		PLSP 2	-0.97	-0.93	0.26	0.98	0.96	0.19	0.54	0.74	0.39
		LSAEP	-0.62	-0.62	0.19	0.67	0.66	0.08	0.56	0.57	0.11
		NSPLP	-0.81	-0.81	0.21	0.82	0.81	0.11			
	800	PLSP 1	-0.97	-0.95	0.26	1.00	0.92	0.27	0.49	0.60	0.38
		PLSP 2	-1,00	-0.97	0.23	1,00	0.97	0.20	0.57	0.76	0.39
		LSAEP	-0.63	-0.61	0.13	0.67	0.66	0.06	0.60	0.60	0.07
		NSPLP	-0.80	-0.81	0.15	0.83	0.83	0.08			

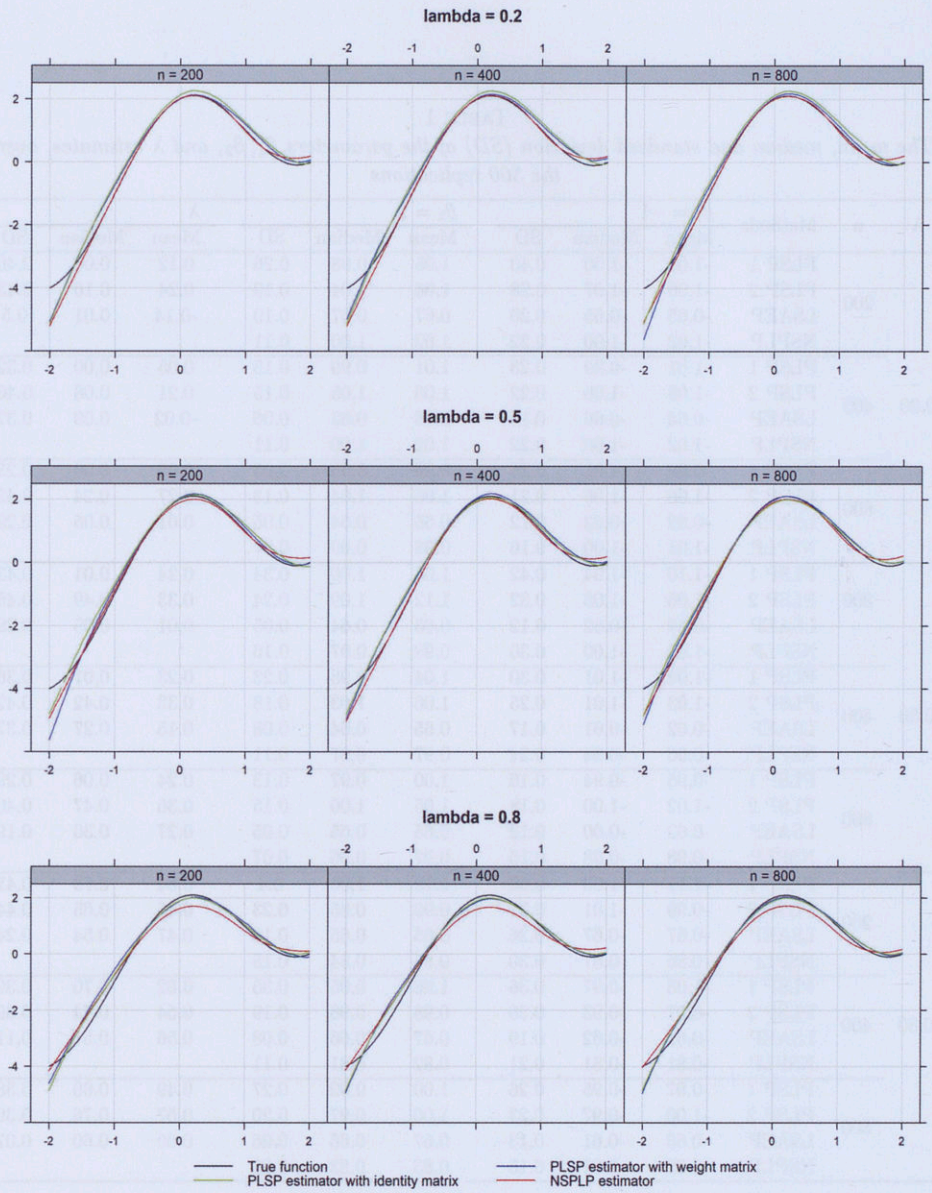

FIGURE 1. *The true function $g(\cdot)$ and the average of its estimates, over the 300 replications*

As indicated previously, weights are used to improve the efficiency and convergence of the GMM procedure. For instance, the finite sample properties section shown that the kind of weight matrix defined in 3.4 with elements τ_{ij} may improve the efficiency of the proposed estimator but is less appropriate in case of weak spatial dependence. Then, it would be interesting to develop other choices of weights τ_{ij} toward achieving a better performance. Another topic of future research is to allow some spatial dependency in the covariates (SAR models) and the response (endogenous models) for more generality.

A. Appendix section

Proposition A.1. *Under Assumptions A1-A6, for $\theta \in \Theta$ and $z \in \mathcal{Z}$, the functions $g_\theta(z)$ and $\hat{g}_\theta(z)$, solutions of (1.8) and (1.14), respectively, satisfy*

1. *for all $i, j = 0, 1, 2$, $i + j \leq 2$,*

$$\frac{\partial^{i+j}}{\partial\theta_l^i \partial\theta_r^j} g_\theta(z) \quad \text{and} \quad \frac{\partial^{i+j}}{\partial\theta_l^i \partial\theta_r^j} \hat{g}_\theta(z) \quad \text{exist and are finite for all } 1 \leq l, r \leq p+1.$$

2. $\sup_{\theta\in\Theta} \|\hat{g}_\theta - g_\theta\|$, $\sup_{\theta\in\Theta} \max_{j=1,\ldots,p+1} \left\| \frac{\partial}{\partial\theta_j} (\hat{g}_\theta - g_\theta) \right\|$ *and* $\sup_{\theta\in\Theta} \max_{1\leq i,j\leq p+1} \left\| \frac{\partial^2}{\partial\theta_i \partial\theta_j} (\hat{g}_\theta - g_\theta) \right\|$,

are all order $o_p(1)$ as $n \to \infty$.

Without loss of generality, the proof of this proposition is ensured by Lemma A.2 in the univariate case i.e., Θ, $\mathcal{Z} \subset \mathbb{R}$.

The following lemma is useful in the proof of Lemma A.2. It is an extension of Lemma 8 in [35] to spatially dependent data.

Lemma A.1. *Let $\zeta_\theta(Y_i)$ denote a scalar function of Y_{in}, $i = 1, \ldots, n$, $n = 1, 2, \ldots$, depending on a scalar parameter $\theta \in \Theta$, and for $j = 0, 1, 2$, let*

$$\zeta_\theta^{(j)}(Y_{in}) = \frac{\partial^j}{\partial\theta^j} \zeta_\theta(Y_{in}), \qquad i = 1, \ldots, n, \; n = 1, 2, \ldots$$

Let $f_i(\cdot)$ denote the density of Z_{in} (given in Assumption A2), and let $\bar{f}(z) = \frac{1}{n}\sum_{i=1}^n f_i(z)$.
Assume that

H.1 $\sup_\theta \sup_{1\leq i\leq n,n} \left|\zeta_\theta^{(j)}(Y_{in})\right| < \infty$ *for $j = 0, \ldots, 3$.*

H.2 *For all $\theta \in \Theta$, $j = 0, 1, 2$, and $1 \leq i, l \leq n$:*

$$\text{(A.1)} \quad |\text{Cov}\left(K_{in}(z), K_{ln}(z)\right)| \leq \left\{\text{Var}(K_{in}(z))\text{Var}(K_{in}(z))\right\}^{1/2} \varphi\left(\|s_i - s_l\|\right),$$

$$\left|\mathrm{Cov}\left(\zeta_\theta^{(j)}(Y_{in})K_{in}(z), \zeta_\theta^{(j)}(Y_{ln})K_{ln}(z)\right)\right| \leq$$

(A.2)

$$\left\{\mathrm{Var}\left(\zeta_\theta^{(j)}(Y_{in})K_{in}(z)\right)\mathrm{Var}\left(\zeta_\theta^{(j)}(Y_{ln})K_{ln}(z)\right)\right\}^{1/2}\varphi\left(\|s_i - s_l\|\right),$$

with $K_{in}(z) = K\left((z - Z_{in})/b\right)$.

Let $m_\theta(z) = \mathbb{E}\left(\zeta_\theta(Y_{in})|Z_{in} = z\right)$ *for* $z \in \mathcal{Z}$, *and assume that* $\dfrac{\partial^j}{\partial\theta^j}m_\theta(\cdot)$ *is continuous on* $\mathcal{Z}$, $j = 0, 1, 2$.

For each fixed $\theta \in \Theta$ *and* $z \in \mathcal{Z}$, *let the kernel estimator* $\widehat{m}_\theta(z)$ *of* $m_\theta(z)$ *be defined by*

$$\widehat{m}_\theta(z) = \frac{\sum_{i=1}^n \zeta_\theta(Y_{in})K_{in}(z)}{\sum_{i=1}^n K_{in}(z)}.$$

If Assumptions A2, A4, and A5 are satisfied, then

$$\sup_{\theta\in\Theta}\sup_{z\in\mathcal{Z}}\left|\frac{\partial^j}{\partial\theta^j}\widehat{m}_\theta(z) - \frac{\partial^j}{\partial\theta^j}m_\theta(z)\right| = o_p(1),$$

for $j = 0, 1, 2$.

Lemma A.1 generalizes Lemma 8 in [35] to spatially dependent data.

Lemma A.2. *For each* $\theta \in \Theta$ *and* $z \in \mathcal{Z}$, *let*

$$H(\eta; \theta, z) = \mathbb{E}_0\left(h_{in}^{\theta,\eta}(Y_{in}|X_{in}, Z_{in})|Z_{in} = z\right),\ 1 \leq i \leq n,\ n = 1, 2, \ldots$$

where $\eta = g(z)$, $g \in \mathcal{G}$ *and* $h_{in}^{\theta,\eta}(\cdot|\cdot,\cdot)$ *is defined in Assumption A3.*

Condition I*: For fixed but arbitrary* $\theta_1 \in \Theta$ *and* $\eta_1 \in \Pi$ *with* $\Pi = g_0(\mathcal{Z})$, *let*

$$\vartheta(\theta, \eta) = \int h_{in}^{\theta,\eta}(y|x\,, z)\exp(h_{in}^{\theta_1,\eta_1}(y|x\,, z))dy, \qquad \theta \in \Theta,\ \eta \in \Pi, (x, z) \in \mathcal{Z} \times \mathcal{Z}$$

where $\{\exp(h_{in}^{\theta,\eta}(y|x\,, z)), \theta \in \Theta, \eta \in \Pi\}$ *denotes the family of conditional density functions (indexed by the parameters* θ *and* η*) of* Y_{in} *given* $(X_{in}, Z_{in}) = (x, z) \in \mathcal{X} \times \mathcal{Z}$. *For each* $\theta \neq \theta_1$, *assume that*

$$\vartheta(\theta, \eta) < \vartheta(\theta_1, \eta_1).$$

Condition S*: Let* $\tilde{p} = p + 1$, *and for all nonnegative integers* $j_1, \ldots, j_{\tilde{p}} = 0, 1, 2$ *and* $r = 0, \ldots, 4$, *such that* $j_1 + \cdots + j_{\tilde{p}} + r \leq 6$, *assume that the derivative*

$$\frac{\partial^{j_1+\cdots+j_{\tilde{p}}+r}h_{in}^{\theta,\eta}}{\partial\theta_1^{j_1}\cdots\partial\theta_{\tilde{p}}^{j_{\tilde{p}}}\partial\eta^r}(y|x\,, z),$$

exists for almost all y and that

$$E_0\left(\sup_{i,n}\sup_{\theta\in\Theta}\sup_{g\in\mathcal{G}}\left|\frac{\partial^{j_1+\cdots+j_{\tilde{p}}+r}h_{in}^{\theta,\eta_i}}{\partial\theta_1^{j_1}\cdots\partial\theta_{\tilde{p}}^{j_{\tilde{p}}}\partial\eta^r}(Y_{in}|X_{in}\,,Z_{in})\right|^2\right)<\infty,\qquad\text{with}\qquad\eta_i=g(Z_{in}).$$

Assume that

$$\sup_z\sup_\theta\sup_\eta\left|\frac{\partial^j}{\partial\theta^j}H^{(k)}(\eta;\theta,z)\right|<\infty, \tag{A.3}$$

for $j=0,1,2$ and $k=2,3,4$ such that $j+k\leq 4$, with

$$H^{(k)}(\eta;\theta,z)=\frac{\partial^k}{\partial\eta^k}H(\eta;\theta,z).$$

Let

$$\widehat{H}(\eta;\theta,z)=\frac{\sum_{i=1}^n h_{in}^{\theta,\eta}(Y_{in}|X_{in},z)K_{in}(z)}{\sum_{i=1}^n K_{in}(z)};$$

then, $\widehat{g}_\theta(z)$ is a solution of $\widehat{H}^{(1)}(\eta;\theta,z)=0$ with respect to η for each fixed $\theta\in\Theta$ and $z\in\mathcal{Z}$.

If we assume that Assumptions A1-A6 are satisfied, then we have, for all $j=0,1,2$,

$$\sup_\theta\sup_z\left|\frac{\partial^j}{\partial\theta^j}\left(\widehat{g}_\theta(z)-g_\theta(z)\right)\right|=o_p(1). \tag{A.4}$$

The assumptions used in the previous lemma are satisfied under the conditions used in the main results. **Condition I** is needed to ensure the identifiability of the arbitrary parameter θ_1 (it plays the role of the true parameter θ_0). This condition is verified when $\theta_1=\theta_0$ by the identifiability of our model (1.1). **Condition S** allows integrals to be interchanged with differentiation; this will be combined with the implicit function theorem [see 33] to ensure the differentiability of $\hat{g}_\theta(z)$ with respect to θ.

Knowing that $\Phi(\cdot)$ is a smooth function on $\mathbb{R}$ and $h_{in}^{\theta,\eta}(\cdot|\cdot\,,\cdot)$ is

$$h_{in}^{\theta,\eta_i}(Y_{in}|X_{in}\,,Z_{in})=Y_{in}\log\left(\frac{\Phi(G_{in}(\theta,\eta_i))}{1-\Phi(G_{in}(\theta,\eta_i))}\right)-\log\left(1-\Phi(G_{in}(\theta,\eta_i))\right),$$

Condition S and Assumption (A.3) are satisfied under the continuity condition of $\Phi(\cdot)$ and $\phi(\cdot)$, Assumption A9 and the compactness of $\mathcal{X}$ and $\mathcal{Z}$.

Let the following notation and Lemmas:

$$\eta_i=g(Z_{in});\quad\tilde{U}_{in}=\tilde{U}_{in}(\theta,\eta_i);\qquad\Phi_{in}=\Phi(G_{in}(\theta,g_\theta));\qquad\Lambda_{in}=\Lambda(G_{in}(\theta,g_\theta)),$$

for all $\theta\in\Theta$, $1\leq i\leq n$, $n=1,2,\ldots$, with $\Lambda(\cdot)=\phi(\cdot)/\Phi(\cdot)(1-\Phi(\cdot))$.

The partial derivatives of $S_n(\theta,g)$ with respect to g of order $s=1,2,\ldots$, for any functions $v_1,\ldots,v_s$ in $\mathcal{G}$, are given by

$$\frac{\partial^s S_n}{\partial g^s}(\theta,g)(v_1,\cdots,v_s)=n^{-1}\sum_{i=1}^n\xi_{in}\frac{\partial^s\tilde{U}_{in}}{\partial\eta^s}(\theta,\eta_i)v_1(Z_{in})\cdots v_s(Z_{in}).$$

Lemma A.3. *Under Assumptions A3, A6 and A9, we have for all* $\theta \in \Theta$,

$$S_n(\theta, g_\theta) - S(\theta, g_\theta) = o_p(1). \tag{A.5}$$

In addition, we have

$$Q_n(\theta, g_\theta) - Q(\theta, g_\theta) = o_p(1), \tag{A.6}$$

if $M_n - M = o_p(1)$.

Note that if Assumption A10 is satisfied, then $M_n - M = o_p(1)$.

Lemma A.4. *Under Assumptions A6-A9, we have* $S_n(\cdot, g_\cdot) - S(\cdot, g_\cdot)$ *is stochastically equicontinuous on* Θ.
In addition, if $M_n - M = o_p(1)$, *then we have* $Q_n(\cdot, g_\cdot) - Q(\cdot, g_\cdot)$ *is also stochastically equicontinuous on* Θ.

Lemma A.5. *Under the assumptions of Proposition A.1 and Assumptions A6 and A9, we have*

$$\sup_{\theta \in \Theta} \|S_n(\theta, \hat{g}_\theta) - S_n(\theta, g_\theta)\| = o_p(1). \tag{A.7}$$

If in addition $M_n - M = o_p(1)$, *then we have*

$$\sup_{\theta \in \Theta} |Q_n(\theta, \hat{g}_\theta) - Q_n(\theta, g_\theta)| = o_p(1). \tag{A.8}$$

The proof of previous lemmas can be obtained on request from the authors.

Proof of Theorem 2.1

By Lemmas A.3 and A.4, Q_n converges to Q in probability uniformly, i.e.,

$$\sup_{\theta \in \Theta} |Q_n(\theta, g_\theta) - Q(\theta, g_\theta)| = o_p(1). \tag{A.9}$$

This result allows one to obtain

$$\left|Q(\hat{\theta}, g_{\hat{\theta}}) - Q(\theta_0, g_0)\right| = o_p(1). \tag{A.10}$$

Indeed, using $|\sup a - \sup b| \leq \sup |a - b|$, we have

$$\begin{aligned}
\left|Q(\hat{\theta}, g_{\hat{\theta}}) - Q(\theta_0, g_0)\right| &\leq \left|Q_n(\hat{\theta}, \hat{g}_{\hat{\theta}}) - Q(\hat{\theta}, g_{\hat{\theta}})\right| + \left|Q_n(\hat{\theta}, \hat{g}_{\hat{\theta}}) - Q(\theta_0, g_0)\right| \\
&\leq \sup_\theta |Q_n(\theta, \hat{g}_\theta) - Q(\theta, g_\theta)| + \left|\sup_\theta Q_n(\theta, \hat{g}_\theta) - \sup_\theta Q(\theta, g_\theta)\right| \\
&\leq 2 \sup_\theta |Q_n(\theta, \hat{g}_\theta) - Q(\theta, g_\theta)| \\
&\leq 2 \sup_\theta |Q_n(\theta, \hat{g}_\theta) - Q_n(\theta, g_\theta)| + 2 \sup_\theta |Q_n(\theta, g_\theta) - Q(\theta, g_\theta)| \\
&= o_p(1),
\end{aligned}$$

by Lemma A.5, (A.9) and $\sup_\theta Q(\theta, g_\theta) = Q(\theta_0, g_0)$ (see Assumption A8).
By Assumption A8, we have for a given $\theta \in \Theta$ that there exists $\varepsilon > 0$ and an open neighbourhood N_θ such that

$$\inf_{\theta_1 \in N_\theta} |Q(\theta_1, g_{\theta_1}) - Q(\theta_0, g_0)| > \varepsilon. \tag{A.11}$$

This and (A.10) imply that

$$\mathbb{P}_0\left(\hat{\theta} \in N_\theta\right) \leq \mathbb{P}_0\left(\left|Q(\hat{\theta}, g_{\hat{\theta}}) - Q(\theta_0, g_0)\right| > \varepsilon\right) \to 0, \text{ as } n \to \infty. \tag{A.12}$$

Let N_0 be an open neighborhood of θ_0, and consider the compact set $\Theta_0 = \Theta \setminus N_0$. Let $\{N_\theta : \theta \in \Theta, \theta \neq \theta_0\}$ denote the open covering of Θ_0 by the procedure given above (each neighbourhood N_θ satisfies (A.11)). By the compactness of Θ_0, let $\{N_{\theta_1}, \ldots, N_{\theta_r}\}$ be a finite sub-covering; then,

$$\mathbb{P}_0\left(\hat{\theta} \notin N_0\right) = \mathbb{P}_0\left(\hat{\theta} \in \Theta_0\right) \leq \sum_{j=1}^{r} \mathbb{P}_0\left(\hat{\theta} \in N_{\theta_j}\right) \to 0, \text{ as } n \to \infty,$$

by (A.12). Therefore, we can conclude that

$$\hat{\theta} - \theta_0 = o_p(1), \qquad \text{as} \qquad n \to \infty.$$

This yields the proof of Theorem 2.1. □

Proof of Theorem 2.2

The proof is based on the following lemmas, proof details can be obtained on request from the authors.

Lemma A.6. *Under the assumptions of Theorem 2.2 and for any $\tilde{\theta}$ such that $\tilde{\theta} - \theta_0 = o_p(1)$, we have*

$$\frac{\partial S_n}{\partial \theta}(\tilde{\theta}, g_{\tilde{\theta}}) - \frac{\partial S}{\partial \theta}(\theta_0, g_0) = o_p(1) \tag{A.13}$$

and

$$\frac{\partial S_n}{\partial g}(\tilde{\theta}, g_{\tilde{\theta}}) g'_{\tilde{\theta}} - \frac{\partial S}{\partial g}(\theta_0, g_0) g'_0 = o_p(1), \tag{A.14}$$

with $g'_{\tilde{\theta}}(.) = \left.\frac{g_\theta}{\partial \theta^T}(.)\right|_{\theta=\tilde{\theta}}$.

Lemma A.7. *Under the assumptions of Theorem 2.2, we have*

$$(i) \qquad \frac{d}{d\theta}\frac{\partial Q_n}{\partial g}(\theta, g_\theta)\Big|_{\theta=\theta_0} (\hat{g}_0 - g_0) = o_p(1)$$

$$(ii) \qquad \frac{\partial Q_n}{\partial g}(\theta, g_\theta)\Big|_{\theta=\theta_0} (\hat{g}_0^{'} - g_0^{'}) = o_p(1),$$

where

$$\hat{g}_0^{'}(.) = \frac{\partial \hat{g}_\theta}{\partial \theta}(.)\Big|_{\theta=\theta_0} \qquad \text{and} \qquad g_0^{'}(.) = \frac{\partial g_\theta}{\partial \theta}(.)\Big|_{\theta=\theta_0}.$$

Proof of Lemma A.7 can be obtained from request to the authors.

Lemma A.8. *Under the assumptions of Theorem 2.2, we have*

$$S_n(\theta, \hat{g}_\theta) - S_n(\theta, g_\theta) = r_n^{(1)}(\theta),$$

where

$$\sup_\theta \left\| \frac{\partial}{\partial \theta} r_n^{(1)}(\theta) \right\| = o_p(1), \qquad \text{and} \qquad \sup_\theta \left\| \frac{\partial^2}{\partial \theta \partial \theta^T} r_n^{(1)}(\theta) \right\| = o_p(1)$$

References

[1] Andrews, D. W. (1992). Generic uniform convergence. *Econometric theory*, *8*, 241–257.

[2] Anselin, L. (2010). Thirty years of spatial econometrics. *Papers in regional science*, *89*, 3–25.

[3] Anselin, L. (2013). *Spatial econometrics: methods and models* volume 4. Springer Science & Business Media.

[4] Arbia, G. (2006). *Spatial econometrics: statistical foundations and applications to regional convergence*. Springer Science & Business Media.

[5] Bernstein, S. (1927). Sur l'extension du théorème limite du calcul des probabilités aux sommes de quantités dépendantes. *Mathematische Annalen*, *97*, 1–59.

[6] Billé, A. G. (2014). Computational issues in the estimation of the spatial probit model: A comparison of various estimators. *The Review of Regional Studies*, *43*, 131–154.

[7] Carroll, R. J., Fan, J., Gijbels, I., & Wand, M. P. (1997). Generalized partially linear single-index models. *Journal of the American Statistical Association*, *92*, 477–489.

[8] Case, A. (1993). Spatial patterns in household demand. *Econometrica*, *52*, 285–307.

[9] Conley, T. G. (1999). Gmm estimation with cross sectional dependence. *Journal of econometrics*, *92*, 1–45.

[10] Cressie, N. (2015). *Statistics for spatial data*. John Wiley & Sons.

[11] Fan, J., & Gijbels, I. (1996). *Local polynomial modelling and its applications: monographs on statistics and applied probability 66* volume 66. CRC Press.

[12] Fleming, M. M. (2004). Techniques for estimating spatially dependent discrete choice models. In *Advances in spatial econometrics* (pp. 145–168). Springer.

[13] Garthoff, R., & Otto, P. (2017). Control charts for multivariate spatial autoregressive models. *AStA Adv. Stat. Anal.*, *101*, 67–94. URL: http://dx.doi.org/10.1007/s10182-016-0276-x. doi:.

[14] Hastie, T. J., & Tibshirani, R. J. (1990). *Generalized additive models* volume 43. CRC Press.

[15] Hunsberger, S. (1994). Semiparametric regression in likelihood-based models. *Journal of the American Statistical Association*, *89*, 1354–1365.

[16] Kelejian, H. H., & Prucha, I. R. (1998). A generalized spatial two-stage least squares procedure for estimating a spatial autoregressive model with autoregressive disturbances. *The Journal of Real Estate Finance and Economics*, *17*, 99–121.

[17] 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. URL: http://dx.doi.org/10.1111/1468-2354.00027. doi:.

[18] Kelejian, H. H., & Prucha, I. R. (2001). On the asymptotic distribution of the moran i test statistic with applications. *Journal of Econometrics*, *104*, 219–257.

[19] Lee, L.-F. (2004). Asymptotic distributions of quasi-maximum likelihood estimators for spatial autoregressive models. *Econometrica*, *72*, 1899–1925. URL: http://dx.doi.org/10.1111/j.1468-0262.2004.00558.x. doi:.

[20] Lee, L.-f. (2007). GMM and 2SLS estimation of mixed regressive, spatial autoregressive models. *J. Econometrics*, *137*, 489–514. URL: http://dx.doi.org/10.1016/j.jeconom.2005.10.004. doi:.

[21] LeSage, J. P. (2000). Bayesian estimation of limited dependent variable spatial autoregressive models. *Geographical Analysis*, *32*, 19–35.

[22] Lin, X., & Lee, L.-f. (2010). GMM estimation of spatial autoregressive models with unknown heteroskedasticity. *J. Econometrics*, *157*, 34–52. URL: http://dx.doi.org/10.1016/j.jeconom.2009.10.035. doi:.

[23] Malikov, E., & Sun, Y. (2017). Semiparametric estimation and testing of smooth coefficient spatial autoregressive models. *J. Econometrics*, *199*, 12–34. URL: http://dx.doi.org/10.1016/j.jeconom.2017.02.005. doi:.

[24] Martinetti, D., & Geniaux, G. (2016). Probitspatial: Probit with spatial dependence, sar and sem models. *CRAN*, . URL: https://CRAN.R-project.org/package=ProbitSpatial.

[25] Martinetti, D., & Geniaux, G. (2017). Approximate likelihood estimation of spatial probit models. *Regional Science and Urban Economics*, *64*, 30–45.

[26] Mátyás, L. (1999). *Generalized method of moments estimation* volume 5. Cambridge University Press.

[27] McMillen, D. P. (1992). Probit with spatial autocorrelation. *Journal of Regional Science*, *32*, 335–348.

[28] Pinkse, J., Shen, L., & Slade, M. (2007). A central limit theorem for endogenous locations and complex spatial interactions. *Journal of Econometrics*, *140*, 215–225.

[29] Pinkse, J., Slade, M., & Shen, L. (2006). Dynamic spatial discrete choice using

one-step gmm: an application to mine operating decisions. *Spatial Economic Analysis*, *1*, 53–99.

[30] Pinkse, J., & Slade, M. E. (1998). Contracting in space: An application of spatial statistics to discrete-choice models. *Journal of Econometrics*, *85*, 125–154.

[31] Poirier, D. J., & Ruud, P. A. (1988). Probit with dependent observations. *The Review of Economic Studies*, *55*, 593–614.

[32] Robinson, P. M. (2011). Asymptotic theory for nonparametric regression with spatial data. *Journal of Econometrics*, *165*, 5–19.

[33] Saaty, T. L., & Bram, J. (2012). *Nonlinear mathematics*. Courier Corporation.

[34] Severini, T. A., & Staniswalis, J. G. (1994). Quasi-likelihood estimation in semiparametric models. *Journal of the American statistical Association*, *89*, 501–511.

[35] Severini, T. A., & Wong, W. H. (1992). Profile likelihood and conditionally parametric models. *The Annals of statistics*, (pp. 1768–1802).

[36] Smirnov, O. A. (2010). Modeling spatial discrete choice. *Regional science and urban economics*, *40*, 292–298.

[37] Staniswalis, J. G. (1989). The kernel estimate of a regression function in likelihood-based models. *Journal of the American Statistical Association*, *84*, 276–283.

[38] Su, L. (2012). Semiparametric gmm estimation of spatial autoregressive models. *Journal of Econometrics*, *167*, 543–560.

[39] Wang, H., Iglesias, E. M., & Wooldridge, J. M. (2013). Partial maximum likelihood estimation of spatial probit models. *Journal of Econometrics*, *172*, 77–89.

[40] Yang, K., & Lee, L.-f. (2017). Identification and QML estimation of multivariate and simultaneous equations spatial autoregressive models. *J. Econometrics*, *196*, 196–214. URL: http://dx.doi.org/10.1016/j.jeconom.2016.04.019. doi:.

[41] Zheng, Y., & Zhu, J. (2012). On the asymptotics of maximum likelihood estimation for spatial linear models on a lattice. *Sankhya A*, *74*, 29–56. URL: http://dx.doi.org/10.1007/s13171-012-0009-5. doi:.