Partial ML Estimation for Spatial Autoregressive Nonlinear Probit Models with Autoregressive Disturbances

Anna Gloria Billé[a,*], Samantha Leorato[b]

[a]Faculty of Economics and Management, Free University of Bozen–Bolzano, Bolzano, Italy
[b]Department of Economics, Management, and Quantitative Methods, University of Milan, Italy

Abstract

In this paper, we propose a Partial MLE (PMLE) for a general spatial nonlinear probit model, i.e., SARAR(1,1) probit, defined through a SARAR(1,1) latent linear model. This model encompasses both the SAE(1) probit and the more interesting SAR(1) probit models, already considered in the literature. We provide a complete asymptotic analysis of our PMLE as well as appropriate definitions of the marginal effects. Moreover, we address the issue of the choice of the groups (couples, in our case) by proposing an algorithm based on a minimum KL divergence problem. Finite sample properties of the PMLE are studied through extensive Monte Carlo simulations. In particular, we consider both *sparse* and *dense* matrices for the true spatial model specifications, and cases of model misspecification given wrong assumed weighting matrices. In a real data example, we finally also compare our estimator with different MLE–based estimators and with the Bayesian approach.

Keywords: Spatial autoregressive–regressive probit model, Nonlinear modeling, SARAR, Partial maximum likelihood, Marginal effects.

JEL codes: C13,C31,C35,C51.

1. Introduction

Estimation theory and inference for econometric models that deal with spatially–distributed data differ substantially from the usual techniques of standard statistics/econometrics; see Whittle (1954), Besag (1972), Besag (1974), Ord (1975), and Cliff and Ord (1981). Further, in spatial econometrics (Anselin, 1988), a large number of theoretical papers highlight the added difficulties in deriving the asymptotic properties of a sequence of extremum estimators, i.e., GMM, quasi–MLE, etc.; see Kelejian and Prucha (1998), Lee (2003), Lee (2004), and Kelejian and Prucha (2010). The bidirectional nature of spatial dependence, leading to a simultaneous specification rather than the conditional specification typical of spatial autoregressive models (Sain and Cressie, 2007), is one of the sources of these complications. Anyway, being that spatial dependence is simply a special

[*]Corresponding author. E–mail: annagloria.bille@unibz.it, web page: https://www.unibz.it/it/faculties/economics-management/academic-staff/person/38038-anna-gloria-bille

case of cross–sectional dependence (Conley, 1999), the way by which spatial econometric models are typically specified and parametrized is convenient as long as we can exploit the information gathered not only about the observed values but also on the locations of the endogenous random variables.

Probabilistic choice theory and random utility models (RUM) have a long history in economics – see Manski (1981) – with, in particular, the important Nobel contribution by McFadden (2001). Modeling spatial discrete choice (and, in general, limited dependent) variables is becoming a challenging work in economics, see Wang et al. (2013), Qu and Lee (2013), Qu and Lee (2012), Lambert et al. (2010), Smirnov (2010), and Xu and Lee (2015). Nonlinear models, like probit/logit models, are useful to analyze endogenous dichotomous dependent variables, but the specified functional form is nonlinear in parameters, and their estimation requires iterative optimization procedures. To make matters worse, spatial dependence adds further complexity in the estimation of parameters.

In fact, one issue is that the unknown form of spatial dependence produces inconsistent structural estimates in a discrete choice framework; see, e.g., McMillen (1995) and Breslaw (2002). Indeed, the parametrization of spatial autoregressive models with a finite unknown number of parameters (i.e., the autocorrelated coefficients) implies at least (spatial) heteroskedasticity, which in turn leads to inconsistency of the standard probit estimator because of misspecification of the functional form (i.e., Bernoulli distributions). First attempts to deal with the implied heteroskedasticity are the contributions by Case (1992) and McMillen (1992). Within a generalized method of moments (GMM) framework, we recognize the works by Pinkse and Slade (1998) as well as Klier and McMillen (2008), where the latter proposed a linearized GMM estimator that is feasible even with moderate to large sample sizes, but it is reasonable as long as the autocorrelated coefficient is relatively small.

From a computational point of view, two major problems must be dealt with. First, direct optimization procedures require maximum simulated likelihood (MSL) estimators (Beron et al., 2003), which are time–consuming in large data sets because of the implied computational burden in evaluating an n–dimensional integral; see Fleming (2004). Second, the optimization of the objective function requires repeated calculations of the inverses of n–dimensional matrices. These also preclude an easy extension to panel data applications, whose diffusion is recently experiencing a massive increase; see, e.g., Smith and LeSage (2004), Lee and Yu (2010), Kapoor et al. (2007), Lee and Yu (2016), and Baltagi et al. (2017). Approximate and conditional maximum likelihood estimators – in the works by Pace and LeSage (2011), Mozharovskyi and Vogler (2016), and Martinetti and Geniaux (2017) – are one way to cope with these problems. However, current computational solutions do not provide an asymptotic analysis of their MLE–based estimators. Moreover, they heavily rely on the sparsity of weight matrices, thus cutting out different spatial patterns. One of the aims of the present paper is to fill this gap: on the one hand, we propose a likelihood–based estimator, whose implementation is not subordinate to the sparsity of the weight matrix; on the other hand, we provide a comprehensive analysis of its asymptotic properties.

Composite MLEs have been proved to be computationally tractable and statistically consistent; see Heagerty and Lele (1998), Gao and Song (2010), Bhat (2011), and Bai et al. (2014). In spatial econometrics, Wang et al. (2013) have recently proposed a partial maximum likelihood estimator (PMLE) for a spatial (first–order) autoregressive error probit (SAE(1) probit) model by dividing observations into several small groups (i.e., couples of spatially distributed random variables) in which adjacent observations belonged to a single group, and bivariate normal distributions were specified within each group; see Arbia (2014) in the linear case. Two limits in the work of Wang et al. (2013) are worth mentioning. First, as Ibragimov and Müller (2010) stressed, some a priori knowledge about the correlation structure is required to make a reasonable partition (i.e., clustering) of the data in a finite number of groups. However, statistically speaking, the optimal choice of groups is not known a priori. Second, a spatial (first–order) autoregressive probit (SAR(1) probit) model, i.e., with lagged dependent variables, is generally recognized to be a more interesting spatial model specification because the autocorrelation coefficient enters in both the mean and the covariance structure when considering the implied reduced form model. For instance, in empirical applications within social networks/interactions, a SAR(1) probit is often preferable, because a direct information on interactions among economic agents' choices is measurable.

Therefore, another aim of our work is to address precisely these points. First of all, we generalize the PMLE approach of Wang et al. (2013) to the wider family of spatial (first–order) autoregressive–regressive probit models with (first–order) autoregressive disturbances (SARAR(1,1) probit), with a particular focus on the SAR(1) probit nested specification. Further, we propose a Kullback-Leibler (KL) divergence approach for the choice of couples aimed at reducing the expected loss of statistical information. The definition of this criterion can be adjusted to apply to different models provided the error distribution is Gaussian. We point out that, despite the fact that SARAR(1,1) probit has been known for a while in the specialized literature, to the best of our knowledge, very little attention has been paid to it mainly because of theoretical and computational complications. Nevertheless, it is not uncommon that real data suggest the presence of an autoregressive structure both in the errors and in the latent dependent variable, as we show in our application (see Section 8).

We assess the finite sample properties of our PMLE and derive asymptotic results under the increasing domain assumption. In particular, we propose two direct estimation procedures of the asymptotic variance–covariance matrix, and parametric bootstrap approaches. We also present proper definitions of the marginal effects, discussed through extensive Monte Carlo simulations. In our simulations, we consider both *sparse* and *dense* matrices for the specification of the true spatial models. Robustness checks on the misspecification of the spatial weighting matrices are also included. Finally, a comparison between our PMLE and alternative MLE–based estimators and the Bayesian approach of LeSage et al. (2011) is also included in the empirical application. All these figures make our work substantially different from that proposed by Wang et al. (2013).

The rest of the paper is organized as follows. Section 2 specifies a SARAR(1,1) probit model, the assumptions behind it, and its nested model specifications. Section 3 describes our PMLE based on bivariate distributions. In section 4, we propose our algorithm for the choice of couples. Section 5 reports the asymptotic properties. Section 6 defines the marginal effects. Section 7 evaluates the finite sample properties of our PMLE with respect to both the parameters and the marginal impacts. Section 8 proposes to replicate the empirical application of business recovery in the aftermath of Hurricane Katrina by LeSage et al. (2011) and compare our PMLE with different MLE–based and Bayesian estimators. Finally, section 9 concludes this study.

2. Model specification

Let $\mathbf{y}_n$ be an n–dimensional stochastic vector of spatial binary variables located on a possibly unevenly spaced lattice $Z \subseteq \Re^n$. A spatial (first–order) autoregressive–regressive probit model with (first–order) autoregressive disturbances (SARAR(1,1) probit) is defined as

$$\begin{aligned} &\mathbf{y}_n^* = \rho \mathbf{W}_n \mathbf{y}_n^* + \mathbf{X}_n \boldsymbol{\beta} + \mathbf{u}_n, \quad \mathbf{u}_n = \lambda \mathbf{M}_n \mathbf{u}_n + \boldsymbol{\varepsilon}_n, \quad \boldsymbol{\varepsilon}_n \sim \mathcal{N}_n\left(\mathbf{0}_n, \sigma_\varepsilon^2 \mathbf{I}_n\right) \\ &\mathbf{y}_n = \mathbb{I}_n\left(\mathbf{y}_n^* > \mathbf{0}_n\right) \end{aligned} \tag{1}$$

where $\mathbf{y}_n^*$ is the n–dimensional vector of latent continuous dependent variables, $\mathbf{y}_n$ is the n–dimensional vector of observed binary dependent variables defined by the n–dimensional indicator function $\mathbb{I}_n\left(\mathbf{y}_n^* > \mathbf{0}\right) = \left(\mathbb{I}(y_1^* > 0), \ldots, \mathbb{I}(y_n^* > 0)\right)'$, $\mathbf{X}_n$ is the n by k matrix of exogenous variables including a constant term, $\mathbf{W}_n$ and $\mathbf{M}_n$ are n–dimensional spatial weighting matrices of known constants, $\boldsymbol{\theta} = \left(\boldsymbol{\beta}', \rho, \lambda\right)'$ is a $(k+2)$–dimensional parameter vector with autoregressive coefficients ρ and λ, and $\boldsymbol{\varepsilon}_n$ is a multivariate normal vector of innovations with zero mean and finite variance $\sigma_\varepsilon^2 < \infty$. Latent variables are then assumed to be linear functions of the regressors, but only a binary transformation is observed that makes the overall model nonlinear in parameters. The variance σ_ε^2 is usually set to 1 for identification. Additional conditions are needed for the identification of (ρ, λ) in a SARAR(1,1) probit model. Specifically, $\mathbf{M}_n$ and $\mathbf{W}_n$ are assumed to be different, thus allowing for different mechanisms to govern spatial correlation between shocks affecting the latent model and spatial dependence of the latent variables themselves. Then the entire spatial dependence can be easily disentangled. It is notable that when $\mathbf{W}_n = \mathbf{M}_n$, distinguishing among the two spatial effects may be difficult, with possible identification problems of the autoregressive parameters. In this particular case, a necessary condition to ensure identifiability of the linear model is that the covariates make a material contribution toward explaining variation in the dependent variable, i.e., at least one coefficient β_j $j = 2, \ldots, k$ is statistically significant.

The inclusion of spatially lagged dependent variables $\mathbf{W}_n \mathbf{y}_n^*$ typically causes an endogeneity problem. This problem is referred to the bidirectional nature of spatial dependence in which each site – say i – is a second–order neighbor of itself, implying that spatial spillover effects have the important meaning of feedback/indirect effects also on the site where the shock may have had origin. The problem also makes the overall model a system of n *simultaneous* equations (one for each random variable in space), with the consequence that spatial autoregressive models cannot be viewed as simple extensions of natural *recursive* time–series econometric models. These types of spatial models are then multivariate by definition, with the peculiarity of having statistical information coming from one observation for each random variable in space in a cross–sectional framework.

To ensure stable spatial processes, we must introduce some assumptions in line with Kelejian and Prucha (2010). Let us first recall the following result (see Lemma 1 in Kelejian and Prucha (2010)).

Lemma 2.1. *Let $\overline{\tau}_{\mathbf{W}_n}$ and $\overline{\tau}_{\mathbf{M}_n}$ denote the spectral radius of the square n–dimensional $\mathbf{W}_n$ and $\mathbf{M}_n$ matrices, i.e.:*
$\overline{\tau}_{\mathbf{W}_n} = max\{|\omega_1|, ..., |\omega_n|\}$ and $\overline{\tau}_{\mathbf{M}_n} = max\{|m_1|, ..., |m_n|\}$, where $\omega_1, ..., \omega_n$ and $m_1, ..., m_n$ are the eigenvalues of $\mathbf{W}_n$ and $\mathbf{M}_n$, respectively. Then, $\mathbf{A}_\rho := (\mathbf{I}_n - \rho\mathbf{W}_n)$ and $\mathbf{B}_\lambda := (\mathbf{I}_n - \lambda\mathbf{M}_n)$ are nonsingular for all values of ρ in the interval $(-1/\overline{\tau}_{\mathbf{W}_n}, 1/\overline{\tau}_{\mathbf{W}_n})$ and λ in the interval $(-1/\overline{\tau}_{\mathbf{M}_n}, 1/\overline{\tau}_{\mathbf{M}_n})$.

Assumption 1. *(a) All diagonal elements of $\mathbf{W}_n$ and $\mathbf{M}_n$ are zero. (b) $\rho \in (-1/\overline{\tau}_{\mathbf{W}_n}, 1/\overline{\tau}_{\mathbf{W}_n})$ and $\lambda \in (-1/\overline{\tau}_{\mathbf{M}_n}, 1/\overline{\tau}_{\mathbf{M}_n})$.*

Assumption 1(a) means that each spatial unit is not viewed as its own neighbor, Assumption 1(b) defines the parameter spaces of the autoregressive coefficients as functions of the spectral radius defined by Lemma 2.1. Under Assumption 1(b) the matrices $\mathbf{A}_\rho$ and $\mathbf{B}_\lambda$ admit an infinite series representation (f.i. $\mathbf{A}_\rho^{-1} = \sum_{k=0}^{\infty} \rho^k \mathbf{W}_n^k$), whereas Assumption 1(b) and Lemma 2.1 ensure that the model in equation (1) has a reduced form. Then if we interpret the model in (1) as an equilibrium relationship – see Billé and Arbia (2019) – this choice of the parameter space rules out unstable Nash equilibria. Note that, if all the eigenvalues of $\mathbf{W}_n$ (resp. $\mathbf{M}_n$) are real, which is the case for symmetric weighting matrices, and $(\underline{\omega} < 0, \overline{\omega} > 0)$, where $\underline{\omega} = \min\{\omega_1, ..., \omega_n\}$ and $\overline{\omega} = \max\{\omega_1, ..., \omega_n\}$, we are in the particular case in which ρ (resp. λ) lies in the interval $(1/\underline{\omega}, 1/\overline{\omega})$ (see Kelejian and Prucha (2010), note 6). Let us now recall that row and column sum norms of a matrix $\mathbf{A}$ are given, respectively, by $\|\mathbf{A}\|_\infty = \max_i \sum_j |a_{ij}|$ and $\|\mathbf{A}\|_1 = \max_j \sum_i |a_{ij}|$.

Assumption 2. *Matrices $\mathbf{W}_n$ and $\mathbf{M}_n$ and $(\mathbf{I}_n - \rho\mathbf{W}_n)^{-1}$ and $(\mathbf{I}_n - \lambda\mathbf{M}_n)^{-1}$ are uniformly bounded in both row and column sum norms.*

Assumption 3. *Elements of $\mathbf{X}_n$ are uniformly bounded constants, $\mathbf{X}_n$ has full column rank, and $lim_{n\to\infty} (\mathbf{X}_n'\mathbf{X}_n)/n$ exists and is nonsingular.*

Assumption 3 assumes the regressors to be fixed bounded constant, which is not very common in applications. The standard way to cope with randomness is to request the limit in Assumption 3 to be satisfied in mean. In this case, all results must be read as conditional on a bounded realization $\mathbf{X}(\omega) = \{\mathbf{X}_n\}_{n\geq 1}$ of the multivariate spatial process. Assumption 2 is equivalent to Assumption 5 in Lee (2004) and plays a fundamental role in the asymptotic properties of estimators, by guaranteeing, e.g., the boundedness of the variances of the latent variables $\mathbf{y}_n^*$.

Having both rows and columns of $\mathbf{W}_n$ and $\mathbf{M}_n$ uniformly bounded in absolute value as n goes to infinity ensures that the correlation between two spatial units should converge to zero as the distance separating them increases to infinity. This uniform boundedness assumption is generally a condition to limit the spatial correlation to a manageable degree and to ensure that the spatial process is not explosive. It is further simply a way to shrink some parameters of the variance–covariance matrix to zero, especially if a sparse matrix is assumed to be the true one generating the underlying spatial process.

It must be pointed out that, if $\mathbf{A}$ has a row sum norm equal to 1, then $\|\mathbf{A}^k\|_\infty = 1$, for every $k \geq 1$, and therefore, a standardized row sum norm of, e.g., $\mathbf{W}_n$, readily implies the same property for $\mathbf{A}_\rho^{-1}$. This is one of the reasons why the matrices $\mathbf{W}_n$ and $\mathbf{M}_n$ are often row–standardized, i.e., they are row–stochastic matrices. Sometimes an alternative standardization rule based on spectral normalization could be preferable because it guarantees the equivalence between the original spatial structural model and the model obtained from normalizing the $\mathbf{W}_n$ and $\mathbf{M}_n$ weighting matrices; see Kelejian and Prucha (2010). However, it must be pointed out that a matrix can be bounded in a spectral norm but unbounded in row or column sum norms, which means that spectral normalization of $\mathbf{W}_n$ is not in general a sufficient condition for Assumption 2. In our paper, we focus on row–standardized weight matrices, but we also consider spectral normalization. More details on the definition of the weight matrices are given in section 7.

Given the aforementioned simultaneous nature of spatial autoregressive processes, spatial models are typically specified in reduced forms. Under the above regularity conditions and assumptions, the structural model in (1) can be written in reduced form as

$$\begin{aligned}
&\mathbf{y}_n^* = \mathbf{A}_\rho^{-1}\mathbf{X}_n\boldsymbol{\beta} + \mathbf{A}_\rho^{-1}\mathbf{u}_n = \mathbf{A}_\rho^{-1}\mathbf{X}_n\boldsymbol{\beta} + \mathbf{A}_\rho^{-1}\mathbf{B}_\lambda^{-1}\boldsymbol{\varepsilon}_n = \mathbf{A}_\rho^{-1}\mathbf{X}_n\boldsymbol{\beta} + \boldsymbol{\nu}_n, \quad \boldsymbol{\nu}_n \sim \mathcal{N}_n\left(\mathbf{0}_n, \boldsymbol{\Sigma}_\nu\right) \\
&\mathbf{y}_n = \mathbb{I}_n\left(\mathbf{y}_n^* > \mathbf{0}_n\right) \qquad (2)
\end{aligned}$$

where $\boldsymbol{\nu}_n = \mathbf{A}_\rho^{-1}\mathbf{B}_\lambda^{-1}\boldsymbol{\varepsilon}_n$ and $\boldsymbol{\Sigma}_\nu := \boldsymbol{\Sigma}_{\nu(\rho,\lambda)} = \mathbb{E}\left[\boldsymbol{\nu}_n\boldsymbol{\nu}_n'\right] = \sigma_\varepsilon^2\mathbf{A}_\rho^{-1}\mathbf{B}_\lambda^{-1}\mathbf{B}_\lambda^{-1'}\mathbf{A}_\rho^{-1'}$ with $\sigma_\varepsilon^2 = 1$ for identification.

From the reduced form in equation (2), we finally obtain conditional expected value and variances, for all $i = 1, \ldots, n$:

$$\begin{aligned}
\mathbb{E}\left(y_i\right) &= \mathrm{P}\left(y_i = 1\right) = \mathrm{P}\left(\left\{\boldsymbol{\nu}_n\right\}_i > -\left\{\mathbf{A}_\rho^{-1}\mathbf{X}_n\boldsymbol{\beta}\right\}_i\right) = \Phi\left(\{\boldsymbol{\Sigma}_{\nu(\rho,\lambda)}\}_{ii}^{-1/2}\{\mathbf{A}_\rho^{-1}\mathbf{X}_n\boldsymbol{\beta}\}_i\right) \\
\mathbb{V}\mathrm{ar}\left(y_i\right) &= \Phi\left(\{\boldsymbol{\Sigma}_{\nu(\rho,\lambda)}\}_{ii}^{-1/2}\{\mathbf{A}_\rho^{-1}\mathbf{X}_n\boldsymbol{\beta}\}_i\right)\left[1 - \Phi\left(\{\boldsymbol{\Sigma}_{\nu(\rho,\lambda)}\}_{ii}^{-1/2}\{\mathbf{A}_\rho^{-1}\mathbf{X}_n\boldsymbol{\beta}\}_i\right)\right] \qquad (3)
\end{aligned}$$

where $\{\cdot\}_i$ is the i–th element of the vector in brackets and $\{\cdot\}_{ii}$ is the i–th diagonal component of the matrix in brackets.

2.1. Nested model specifications

Two widely used submodels can be specified starting from equation (1): the SAR(1) probit model by letting $\lambda = 0$ and the SAE(1) probit model by letting $\rho = 0$.

$$\text{(SAR)} \qquad \mathbf{y}_n^* = \rho \mathbf{W}_n \mathbf{y}_n^* + \mathbf{X}_n \boldsymbol{\beta} + \boldsymbol{\varepsilon}_n, \quad \boldsymbol{\varepsilon}_n \sim \mathcal{N}_n\left(\mathbf{0}_n, \mathbf{I}_n\right), \quad \mathbf{y}_n = \mathbb{I}_n\left(\mathbf{y}_n^* > \mathbf{0}_n\right) \tag{4}$$

$$\text{(SAE)} \qquad \mathbf{y}_n^* = \mathbf{X}_n \boldsymbol{\beta} + \mathbf{u}_n, \quad \mathbf{u}_n = \lambda \mathbf{M}_n \mathbf{u}_n + \boldsymbol{\varepsilon}_n, \quad \boldsymbol{\varepsilon}_n \sim \mathcal{N}_n\left(\mathbf{0}_n, \mathbf{I}_n\right), \quad \mathbf{y}_n = \mathbb{I}_n\left(\mathbf{y}_n^* > \mathbf{0}_n\right) \tag{5}$$

The former is generally considered more interesting for several reasons. From a statistical point of view, the autocorrelation coefficient ρ summarizes the information of a "direct" dependence/interaction structure among the random variables of interest, whereas λ captures the intensity of the dependence structure implied by the disturbances/shocks, so that they "indirectly" have an impact on the latent dependent variables. Moreover, for linear specifications, ρ enters in both the mean and the variance–covariance structure of the model, whereas λ enters only in the variance–covariance matrix. However, a SAE(1) probit model can possibly avoid the inconsistency problem, which does not arise in the spatial linear case for the same model specification. Indeed, apart from information that comes from the economic theory, a SAE(1) model produces only more efficient estimates in the linear case. We refer to Appendix D for more details on this issue. Finally, it is worth noting that alternative non–nested model specifications, e.g., spatial Durbin models, within nonlinear specifications can be defined, and the reader is referred to Billé and Arbia (2019).

3. Partial ML estimation

The main problem in estimating the model in equation (1) – or its subspecifications – via MLE is the need of numerical approximation of n–dimensional integrals, which are time–consuming even with moderate sample sizes. In spatial linear autoregressive models, the GMM approach is preferred to MLE because of computational tractability. However, current GMM approaches for spatial nonlinear models are either computationally intractable (Pinkse and Slade, 1998) or based on a linear approximation (Klier and McMillen, 2008), which is not feasible for higher autocorrelation coefficients. In this section, we develop the theory of the partial MLE for a SARAR(1,1) probit model with a particular emphasis on the SAR(1) probit case, which is consistent and computationally more feasible than the above numerical approximations. Computational issues on the estimation procedure can be found in Appendix C. Throughout this section, all indices n in vectors and matrices are omitted to ease the notation.

We start by considering the SARAR(1,1) probit model specified in equations (1) and (2) and show later the results to the SAR(1) probit model in equation (4). As already pointed out in section 2, the major difference relative to the model considered in Wang et al. (2013) consists of the fact that both the mean and the variance of the bivariate distribution of the latent variables depend on the parameter ρ through the matrix

$\mathbf{A}_\rho^{-1} = (\mathbf{I} - \rho\mathbf{W})^{-1}$. Thus, the probabilities $\Pr(y_{g_1} = d_1, y_{g_2} = d_2 \mid \mathbf{X})$ for every couple $g \equiv \{g_1, g_2\}$ and $d_1, d_2 \in \{0,1\}^2$, depend in a much more complex way on the weight matrix and on the parameter. Although we explicitly refer to partial loglikelihood based on bivariate marginals, most of the results of this section and section 5 can be straightforwardly adapted to an r–dimensional partial distribution, with $r > 2$. The algorithm presented in section 4 and the formulas of the score vectors given in the supplementary appendix are instead specific to couples.

Throughout this section, we are assuming that the couples $g = 1, \dots, G$ are given (for example, $g_1 = 2g-1$, $g_2 = 2g$). We will discuss in more detail criteria for the choice of couples in section 4. Now consider groups (couples) indexed by $g = 1, \dots, G$. From the model in equation (4), for the units (g_1, g_2) of a generic group g, we have $y_{g_1} = \mathbb{I}\{y^*_{g_1} > 0\}$ and $y_{g_2} = \mathbb{I}\{y^*_{g_2} > 0\}$, where

$$y^*_{g_1} = \{\mathbf{A}_\rho^{-1}\mathbf{X}\boldsymbol{\beta}\}_{g_1} + \nu_{g_1}$$
$$y^*_{g_2} = \{\mathbf{A}_\rho^{-1}\mathbf{X}\boldsymbol{\beta}\}_{g_2} + \nu_{g_2}$$

and where $\boldsymbol{\nu} = \mathbf{A}_\rho^{-1}\mathbf{B}_\lambda^{-1}\boldsymbol{\varepsilon} \sim \mathcal{N}\left(\mathbf{0}, \boldsymbol{\Sigma}_{\nu(\rho,\lambda)}\right)$.

In the following, we write the shortened form $\boldsymbol{\Sigma}$ for $\boldsymbol{\Sigma}_{\nu(\rho,\lambda)}$, leaving the dependence on $\boldsymbol{\nu}$, and so on (ρ, λ), implicit in the formula. Moreover, we denote by $\boldsymbol{\Sigma}_g$ the 2×2 block corresponding to the variance covariance matrix of $\boldsymbol{\nu}_g$:

$$\boldsymbol{\Sigma}_g = \begin{pmatrix} \sigma^2_{g_1} & \sigma_{g_1,g_2} \\ \sigma_{g_1,g_2} & \sigma^2_{g_2} \end{pmatrix}.$$

Further, we write $\mathbf{X}_\rho = \mathbf{A}_\rho^{-1}\mathbf{X}$ with rows $\mathbf{x}_{\rho,\cdot}$ and $\mathbf{X}_{\rho,g} = [\mathbf{x}'_{\rho,g_1}, \mathbf{x}'_{\rho,g_2}]'$. It is now easy to find, for all $d_1, d_2 \in \{0,1\}^2$, the probabilities:

$$p_g(d_1, d_2) = P\left(y_{g_1} = d_1, y_{g_2} = d_2\right) = P(y_{g_1} = d_1)P\left(y_{g_2} = d_2 \mid y_{g_1} = d_1\right).$$

For any $g = 1, \dots, G$, let us define the functions (implicit in ρ, λ, and $\boldsymbol{\beta}$)

$$\varphi_{1,g}(u) = \frac{\mathbf{x}_{\rho,g_1}\boldsymbol{\beta} + u\frac{\sigma_{g_1,g_2}}{\sigma^2_{g_2}}}{\sqrt{\sigma^2_{g_1} - \sigma^2_{g_1,g_2}/\sigma^2_{g_2}}} \quad \text{and} \quad \varphi_{2,g}(u) = \frac{\mathbf{x}_{\rho,g_2}\boldsymbol{\beta} + u\frac{\sigma_{g_1,g_2}}{\sigma^2_{g_1}}}{\sqrt{\sigma^2_{g_2} - \sigma^2_{g_1,g_2}/\sigma^2_{g_1}}}, \tag{6}$$

and $s_{g_i} = 2(d_i - 1/2)$.

Theorem 3.1. *The joint probabilities $p_g(d_1, d_2)$ are given by:*

$$\begin{aligned} p_g(d_1, d_2) &= \int_{\{s_{g_1}u > -s_{g_1}\mathbf{x}_{\rho,g_1}\boldsymbol{\beta}\}} \frac{1}{\sigma_{g_1}}\phi\left(\frac{u}{\sigma_{g_1}}\right)\Phi\left(s_{g_2}\varphi_{2,g}(u)\right)du \\ &= \Pr\left\{s_{g_1}Z_{g_1} > s_{g_1}\mathbf{x}_{\rho,g_1}\boldsymbol{\beta}, s_{g_2}Z_{g_2} > s_{g_2}\mathbf{x}_{\rho,g_2}\boldsymbol{\beta}\right\} \end{aligned} \tag{7}$$

where $\mathbf{Z} = (Z_{g_1}, Z_{g_2}) \sim \mathcal{N}(0, \boldsymbol{\Sigma}_g)$.

The proof of Theorem 3.1 can be found in the supplemental material. Using Theorem 3.1, we can write the partial loglikelihood function of the spatial probit model as

$$\ell_n\left(\boldsymbol{\theta};\mathbf{y},\mathbf{X}\right)=\frac{1}{G}\sum_{g=1}^{G}\log\left(p_g\left(y_{g_1}y_{g_2}\right)\right). \tag{8}$$

The partial loglikelihood for estimating a SAR(1) probit model or the SAE(1) probit considered in Wang et al. (2013) are also given by equation (8), with probabilities defined through equation (7) in the particular cases of $\lambda=0$ or $\rho=0$, respectively. Specifically, in a SAR(1) probit model, the matrix $\boldsymbol{\Sigma}$ now depends only on ρ, i.e., $\boldsymbol{\Sigma}:=\boldsymbol{\Sigma}_{\mathbf{u}(\rho)}=\mathbf{A}_\rho^{-1}\mathbf{A}_\rho^{-1\prime}$. The score vector is $\nabla\left(\boldsymbol{\theta};\mathbf{y},\mathbf{X}\right)=\left(\nabla_\beta\left(\boldsymbol{\theta}\right)',\nabla_\rho\left(\boldsymbol{\theta}\right),\nabla_\lambda\left(\boldsymbol{\theta}\right)\right)'$. In both cases, we write $\nabla\left(\boldsymbol{\theta};\mathbf{y},\mathbf{X}\right)=\frac{1}{G}\sum_g\nabla_\theta^g\left(\boldsymbol{\theta};\mathbf{y}_g\right)$, where

$$\nabla_\theta^g\left(\boldsymbol{\theta}\right)=\frac{\partial p_g\left(y_{g_1}y_{g_2}\right)/\partial\boldsymbol{\theta}}{p_g\left(y_{g_1}y_{g_2}\right)}, \tag{9}$$

and where formulas for $\frac{p_g(d_1,d_2)}{\partial\boldsymbol{\beta}}$, $\frac{p_g(d_1,d_2)}{\partial\rho}$ and $\frac{p_g(d_1,d_2)}{\partial\lambda}$ are can be found in the supplementary material for both the SAR and SARAR–probit specifications.

Equations (7), (8), and (9) give the exact formulas for the bivariate probabilities, the partial loglikelihood, and the score vector for the models in equations (1) or (4). By definition, they all depend on implicit functions of the matrix $\mathbf{A}_\rho^{-1}$ (and $\mathbf{B}_\lambda^{-1}$) through both $\mathbf{X}_\rho$ and the elements $\sigma_{g_1},\sigma_{g_2}$, and σ_{g_1,g_2}. It is a rather common practice to approximate the inversion of $\mathbf{A}_\rho$ (and $\mathbf{B}_\lambda$) by a truncated sum: $\mathbf{A}_\rho^{-1}\approx\sum_{k=0}^{q}\rho^k\mathbf{W}^k$, $q<\infty$; see, e.g., Kelejian et al. (2004). Despite that this has become one of the conventional approaches, no great attention has been paid so far to conditions ensuring the computation of a finite sum approximation and of the exact inverse to give (asymptotically) the same estimates. Intuitively, since the approximation error $\|\mathbf{A}_\rho^{-1}-\sum_{k=0}^{q}\rho^k\mathbf{W}^k\|\leq O(|\rho\bar{\tau}_W|^{q+1})$, if the number of terms q of the finite order approximation is large enough, the difference between the estimates obtained by these two approaches should be negligible. We will address this issue in more detail in section 5, where we study the asymptotic behavior of the PMLE.

4. The choice of couples of the spatial data

The choice of the G couples to be considered in the computation of the partial ML estimation is a potentially critical part of the procedure. In fact, the definition of the partial MLE only exploits the limited information of the two–dimensional distribution of the latent variables. Different associations of couples can, in principle, determine relevant differences in terms of information loss. In principle, it is auspicable to select couples that guarantee a minimal loss. However, since the number of possible ways to choose couples from $n=2G$ units corresponding to different partial loglikelihood functions is huge (specifically, it is $(2G-1)!!=(2G)!/G!2^G$), a *brute force approach*, based on comparing the partial loglikelihood for all different groupings, is clearly unmanageable. The aim of this section is to propose an algorithm for the choice of G couples for which the expected information loss is the lowest possible value. The procedure we propose is based on an algorithm from graph theory that is known to have a complexity equal, at most, to G^3.

Any selection of couples can be seen as a permutation problem: instead of extracting without replacement the elements of each couples, we can always choose the consecutive couples $(2g-1, 2g)$ but change their composition through the permutation of the units. Then finding the best selection of couples amounts to finding the best permutation of n units relatively to a specified optimality criterion. In line with this, it is convenient to introduce the following notation. Let $\pi : \pi\left(1, \ldots, n\right) = \left(i_1, \ldots, i_n\right)$ be a permutation map. Each π defines a unique set of couples by

$$\{\left(\pi\left(1\right), \pi\left(2\right)\right), \ldots, \left(\pi\left(2g-1\right)\pi\left(2g\right)\right), \ldots, \left(\pi\left(2G-1\right), \pi\left(2G\right)\right)\} = \{\left(i_1, i_2\right), \ldots, \left(i_{2g-1}, i_{2g}\right), \ldots, \left(i_{2G-1}, i_{2G}\right)\}. \quad (10)$$

We further denote by $\mathbf{P}_\pi$ the permutation matrix corresponding to π, namely, $\mathbf{P}_\pi = (\mathbf{e}_{\pi(1)}, \ldots, \mathbf{e}_{\pi(n)})'$, where $\mathbf{e}_j$ is the jth canonical column vector. Thus, $\mathbf{P}_\pi$ transforms a vector $\mathbf{z} = (z_1, \ldots, z_n)'$ into $\mathbf{P}_\pi\mathbf{z} = (z_{\pi(1)}, \ldots, z_{\pi(n)})'$. Then the reduced model in equation (2) can be rewritten as

$$\begin{aligned} \mathbf{P}_\pi\mathbf{y}^* &= \mathbf{P}_\pi\mathbf{A}_\rho^{-1}\mathbf{X}\boldsymbol{\beta} + \mathbf{P}_\pi\boldsymbol{\nu}, \qquad \boldsymbol{\nu} \sim \mathcal{N}_n\left(\mathbf{0}_n, \boldsymbol{\Sigma}\right) \\ \mathbf{P}_\pi\mathbf{y} &= \mathbb{I}_n\left(\mathbf{P}_\pi\mathbf{y}^* > \mathbf{0}_n\right) \end{aligned} \quad (11)$$

Note that from the assumptions of the model in equation (1) defined in section 2, we obtain $\mathbf{P}_\pi\boldsymbol{\nu} \sim \mathcal{N}(\mathbf{0}, \mathbf{P}_\pi\boldsymbol{\Sigma}\mathbf{P}_\pi')$, where $\mathbf{P}_\pi' = \mathbf{P}_{\pi^{-1}} = \mathbf{P}_\pi^{-1}$, and we use the short notation $\boldsymbol{\Sigma}$ for the SARAR(1,1) probit covariance matrix. Finally, we will use the notation $\boldsymbol{\Sigma}_\pi$ for the diagonal block matrix with diagonal blocks of size 2×2 as in $\mathbf{P}_\pi\boldsymbol{\Sigma}\mathbf{P}_\pi'$.

In this section, we propose a criterion that gives us a (not necessarily unique) permutation map π^* solving a minimum KL divergence problem. Let P_θ be the conditional probability of the n–tuple $(y_1, \ldots, y_n)$ and P_θ^π the conditional probability obtained by assuming that consecutive couples from (10) are independent. Specifically, using the notation introduced in Theorem 3.1,

$$P_\theta(\mathbf{d}) = \Pr\left(y_1 = d_1, \ldots, y_n = d_n\right) = \Pr\left(s_1 Z_1 > s_1\mathbf{x}_{\rho,1}\boldsymbol{\beta}, \ldots, s_n Z_n > s_n\mathbf{x}_{\rho,n}\boldsymbol{\beta}\right).$$

and $P_\theta^\pi = p_{1,\theta}^\pi \times p_{2,\theta}^\pi \times \cdots \times p_{G,\theta}^\pi$, where each $p_{g,\theta}^\pi$, consistently with equation (7), is equal to

$$\Pr\left\{s_{\pi(2g-1)}Z_1 > s_{\pi(2g-1)}\mathbf{x}_{\rho,\pi(2g-1)}\boldsymbol{\beta},\ s_{\pi(2g)}Z_2 > s_{\pi(2g)}\mathbf{x}_{\rho,\pi(2g)}\boldsymbol{\beta}\right\}.$$

In particular, we write P_0 and P_0^π if $\boldsymbol{\theta} = \boldsymbol{\theta}_0$. Our idea is to find a permutation that minimizes the KL divergence between P_0^π and P_0, namely, to minimize

$$KL(P_0^\pi \| P_0) = \sum_{\mathbf{d}\in\{0,1\}^n} P_0^\pi(\mathbf{d}) \log\frac{P_0^\pi(\mathbf{d})}{P_0(\mathbf{d})}, \quad (12)$$

over all the possible permutations π.

The computation of the term $\Pr(\mathbf{y} = \mathbf{d}) = P_0(\mathbf{d})$ is unfeasible because it involves an n–dimensional integration. Thus, we propose to minimize the KL divergence between the continuous Gaussian distributions of the latent variables that generate P_0^π and P_0, which we denote by f_0^π (n–variate Gaussian density with pairwise independent components) and f_0 (the full n–variate Gaussian density from the model in equation (1)), respectively. Let Π_n be the set of all permutations of n units corresponding to distinct bivariate distributions. Our algorithm is based on the following result:

Theorem 4.1. *(i) For every $\pi \in \Pi_n$ and $\boldsymbol{\theta} \in \Theta$, $KL(P_\theta^\pi || P_\theta) \leq KL(f_\theta^\pi || f_\theta)$.*

(ii) For any $\boldsymbol{\theta} = (\beta, \rho, \lambda) \in \Theta$, under model (1),

$$\arg\min_{\pi} KL(f_\theta^\pi || f_\theta) = \arg\min_{\pi \in \Pi} \sum_{g=1}^{G} \left(b(\pi(2g-1), \pi(2g)) - \log(\bar{\sigma}(\pi(2g-1), \pi(2g)))\right) \tag{13}$$

where $b(i,j) = \sigma^*(i,j)\sigma(j,i) + \sigma^*(j,i)\sigma(j,i)$, $\bar{\sigma}(i,j) = \sigma(i,i)\sigma(j,j) - \sigma(i,j)\sigma(j,i)$, $\sigma(i,j)$ is the (i,j)–th component of $\boldsymbol{\Sigma}$ and $\sigma^*(i,j)$ is the (i,j)–th component of $\boldsymbol{\Sigma}^{-1}$.

Theorem 4.1(i) suggests that (13) can be viewed as a minimax solution to the unfeasible problem of minimizing (12). Theorem 4.1(ii) instead transforms the objective function into the sum of the contributions of all couples. This helps define a procedure based on the solution of a maximum weighted matching problem in a general graph. Such a matching is the set of edges of a graph, with no nodes in common, that maximizes the total weights. Appendix F provides details on the maximum matching problem and briefly describes the blossom algorithm used to solve it. Our procedure is based on the following steps:

1) Start from a *guess* for the value of (ρ, λ) (only ρ or λ in the case of a SAR(1) or SAE(1) probit model, respectively), $\left(\tilde{\rho}, \tilde{\lambda}\right)$, and compute $\tilde{\boldsymbol{\Sigma}}$ from it.

2) For all couples (i,j), $i,j = 1, \ldots, n$, compute $b\,(i,j)$, $\bar{\sigma}\,(i,j)$ and $u\,(i,j) = b\,(i,j) - \log\left(\bar{\sigma}\,(i,j)\right)$ using $\tilde{\boldsymbol{\Sigma}}$.

3) Build a complete weighted graph $\mathcal{G}$, with n nodes and weights equal to $-u(i,j)$, for edge $\{i,j\}$.

4) Use Edmonds' *blossom algorithm* for the computation of the maximum weighted matching.

This procedure is a way to control the information loss, which tends to be higher (i) when the weight matrix is dense and (ii) for large values of (ρ, λ) (in absolute value). For this reason, we expect the use of the algorithm to improve the estimation in those cases. The algorithm requires the definition of a starting value for the autocorrelation parameters. In the Monte Carlo section, we explore how sensitive the algorithm is to the choice of the initial value, and finite sample performances do not seem to be affected by it. For this reason, we suggest the rule of thumb of choosing the arbitrary value of 0.5 if the spatial autocorrelation is expected to be positive (and -0.5 otherwise).

5. Asymptotics

In this section, we study the asymptotic properties of the PMLE for the SARAR(1,1) probit model. The analysis performed here enters in the context of the increasing domain asymptotics, consistent with the literature. Throughout the section, the number of groups (couples) is denoted by G_n to make clear its dependence on n. In what follows, the sequence of couples is considered as given. In line with Wang et al. (2013), we need to add the following assumptions.

Assumption 4. *$\ell = \lim_n \mathbb{E}\ell_n$ exists, and ℓ attains a unique maximum over the compact set Θ at the interior point $\boldsymbol{\theta}_0$.*

Assumption 5. *(a) Every subset of the sampling area of size c_n contains at most m_n units, where $\lim_n m_n/c_n < C < \infty$. (b) Moreover,*

$$\sup_{1\leq g\leq G_n} \left| \sum_{d_1,d_2=0}^{1} \frac{1}{p_g(d_1,d_2)} \right| < \infty.$$

Assumption 6. *$\sup_{n,g,h} |\mathbb{C}ov(y_{gi}, y_{hi})| \leq \alpha(d_{gh})$, where d_{gh} is the distance between groups g and h and $\alpha(c) \to 0$ as $c \to \infty$.*

Assumption 7. *(a) There exists a sequence $\{q_n\}$, with $\lim_{n\to\infty} q_n = \infty$, such that the matrix $\sum_{h=0}^{q_n} \rho^h \mathbf{W}_n^h$ is nonsingular (and $\sum_{h=0}^{q_n} \lambda^h \mathbf{M}_n^h$ is nonsingular) for all n and for all $\rho \in (-1/\overline{\tau}, 1/\overline{\tau})$ (and $\lambda \in (-1/\overline{\tau}, 1/\overline{\tau})$). (b) There exists a $\delta > 0$ such that $\lim_{n\to\infty} n^\delta/q_n < \infty$.*

Assumptions 4–6 are taken from Wang et al. (2013) and are used to prove consistency of the PMLE. The first is a standard assumption for M–type estimators and is an implicit identification condition. Finding explicit primitive conditions is an extremely difficult task, even for models simpler than those considered in this paper. However, in section 5.2, we consider some special cases attempting to give some better understanding of the implications of Assumption 4. Assumption 5 is the same as (iv) and (v) of Theorem 1 in Wang et al. (2013). The first part guarantees that observations do not tend to concentrate in an infinitesimal area, and it is a natural assumption within the increasing domain asymptotics. The second part rules out the possibility that, for some couples, one (or more) of the four outcomes has a conditional probability equal to zero. Assumption 6 is the *mixing* condition given in Wang et al. (2013), ensuring that the dependence between observations rapidly decays with their distance. We remark that since

$$\mathbb{C}\text{ov}(y_1, y_2) \leq O(\mathbb{C}\text{ov}(y_1^*, y_2^*)) \tag{14}$$

(the proof of this claim may be found in the supplementary material), Assumption 6 (and Assumption 8, defined below) can be conveniently related to the spatial structure of the latent Gaussian process. Consider, e.g., a SAR(1) probit model, and let us assume that $\mathbf{W}_n$ is a sparse matrix such that all elements i, j of

$\mathbf{W}_n^k$ are zero if the distance between units i and j is bounded below by $d(i,j) > \delta(k)$ and δ is a monotone non–decreasing function.[1] In this case, one can use the approximate bound $\mathbb{C}\mathrm{ov}(y_i, y_j) = O(\rho^{k_{ij}})$, with $k_{ij} = \min\{k : \delta(k) \geq d(i,j)\}$.

Finally, Assumption 7 is added to assess the validity of the PML estimates obtained under the finite sum approximation of the matrices $\mathbf{A}_\rho^{-1}$ and $\mathbf{B}_\lambda^{-1}$. In particular, Assumption 7(a) guarantees invertibility of the approximating sum $\sum_{h=0}^{q_n} \rho^h \mathbf{W}_n^h$ for all q_n and is therefore necessary for identification. Assumption 7(b) defines the minimum rate at which the number of approximating terms q_n has to increase with the sample size. This is a mild assumption since it basically requires the rate of q_n to be faster than the $\log n$.

Theorem 5.1. *Under Assumptions 1–6, $\|\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0\| = o_p(1)$. If further Assumption 7(a) holds, then the estimator obtained by a finite sum approximation of $\mathbf{A}_\rho^{-1}$ (and $\mathbf{B}_\lambda^{-1}$) is asymptotically equivalent to $\hat{\boldsymbol{\theta}}_n$.*

To prove asymptotic normality, we need the following further assumptions.

Assumption 8. *For all fixed $d > 0$,*

$$\lim_{k\to\infty} \frac{k^2 \alpha(kd)}{\alpha(d)} = 0.$$

Assumption 9. *The sampling area grows uniformly at a rate of $\sqrt{n}$ in two non–opposing directions.*

Assumption 10. *The matrices $J(\boldsymbol{\theta}_0) = \lim_n G_n \mathbb{E}\left(\frac{\partial \ell_n}{\partial \boldsymbol{\theta}}(\boldsymbol{\theta}_0) \frac{\partial \ell_n}{\partial \boldsymbol{\theta}'}(\boldsymbol{\theta}_0)\right)$ and*

$$\mathbf{H}(\boldsymbol{\theta}_0) = -\lim_{n\to\infty} \mathbb{E}\, H(\boldsymbol{\theta}_0) = -\lim_{n\to\infty} \mathbb{E}\left(\frac{\partial^2}{\partial \boldsymbol{\theta}_0 \partial \boldsymbol{\theta}_0'} \ell_n\right)$$

are positive definite.

Theorem 5.2. *Under Assumptions 1–6 and 8–10,*

$$\sqrt{G_n}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0) \to \mathcal{N}\left(0, \mathbf{H}(\boldsymbol{\theta}_0)^{-1} J(\boldsymbol{\theta}_0) \mathbf{H}(\boldsymbol{\theta}_0)^{-1}\right) \tag{15}$$

If further Assumption 7 holds, then the same asymptotic distribution is obtained if $\mathbf{A}_\rho^{-1}$ and $\mathbf{B}_\lambda^{-1}$ are approximated by a finite sum of q_n terms.

Proofs of both Theorems 5.1 and 5.2 are in the supplemental material. Assumptions 8–10 are those used by Wang et al. (2013) to prove Theorem 2. Assumption 10 is quite standard in an MLE framework, while Assumptions 8 and 9 are necessary to apply Bernstein's blocking method, used in McLeish's central limit theorem for dependent processes; see McLeish (1974).

[1] A typical example when this occurs is when $\mathbf{W}_n$ is built with a contiguity criterion, for which the elements of $\mathbf{W}_n^k$ are zero whenever the number of steps necessary to go from unit i to unit j is larger than k.

5.1. Estimation of the asymptotic variance-covariance matrix

Consistent estimation of $\mathbf{H}(\boldsymbol{\theta}_0)$ and $J(\boldsymbol{\theta}_0) = \lim_n G_n \mathbb{E}[\nabla(\boldsymbol{\theta}_0)\nabla(\boldsymbol{\theta}_0)']$, yields a consistent estimator for the covariance matrix of $\hat{\boldsymbol{\theta}}$. The most difficult part is the estimation of $J(\boldsymbol{\theta}_0)$, because $\mathbf{H}(\boldsymbol{\theta}_0)$ can be estimated through the average of the negative Hessian matrix at $\hat{\boldsymbol{\theta}}$. In the following, we propose two different approaches to estimate $J(\boldsymbol{\theta}_0)$ and two parametric bootstrap approaches as alternatives to directly obtain the standard errors.

Since we have the explicit formulas of the score vectors, the first approach consists of a direct estimation approach similar to that suggested by Pinkse and Slade (1998), which is based on the computation of $\frac{1}{G_n}\sum_{g=1}^{G_n}\mathbb{E}\left[\nabla_\theta^g(\hat{\boldsymbol{\theta}})\nabla_\theta^g(\hat{\boldsymbol{\theta}})'\right]$. Let us define, for two different couple indices $g \neq j$, the 4×4 submatrix $\boldsymbol{\Sigma}_{[gj]}$ obtained by extracting from $\boldsymbol{\Sigma}_\nu$ the rows and columns (g_1, g_2, j_1, j_2), and let $f_{[gj]}(u_1, u_2, u_3, u_4)$ be the density of the four–variate Gaussian $\mathcal{N}\left(\mathbf{0}, \boldsymbol{\Sigma}_{[gj]}\right)$ distribution. By using the sign functions $s_i = 2(d_i - 1/2)$, $i = 1, 2, 3, 4$, we can propose the following estimator:

$$\begin{aligned}\hat{J}(\hat{\boldsymbol{\theta}}_n) &= G_n\ \mathbb{E}\frac{\partial \ell_n}{\partial \boldsymbol{\theta}}(\hat{\boldsymbol{\theta}}_n)\frac{\partial \ell_n}{\partial \boldsymbol{\theta}'}(\hat{\boldsymbol{\theta}}_n)\\ &= \frac{1}{G_n}\sum_{(g,j):j\neq g}\sum_{i=1}^{4}\sum_{d_i=\{0,1\}}\frac{\partial p_g(d_1,d_2;\hat{\boldsymbol{\theta}}_n)}{\partial \boldsymbol{\theta}}\frac{\partial p_j(d_3,d_4;\hat{\boldsymbol{\theta}}_n)}{\partial \boldsymbol{\theta}'}\frac{p_{[gj]}\left(d_1,d_2,d_3,d_4;\hat{\boldsymbol{\theta}}_n\right)}{p_g(d_1,d_2;\hat{\boldsymbol{\theta}}_n)p_j(d_3,d_4;\hat{\boldsymbol{\theta}}_n)} \\ &\quad + \frac{1}{G_n}\sum_g\sum_{i=1}^{2}\sum_{d_i=\{0,1\}}\frac{\partial p_g(d_1,d_2;\hat{\boldsymbol{\theta}}_n)}{\partial \boldsymbol{\theta}}\frac{\partial p_g(d_1,d_2;\hat{\boldsymbol{\theta}}_n)}{\partial \boldsymbol{\theta}'}\frac{1}{p_g(d_1,d_2;\hat{\boldsymbol{\theta}}_n)},\end{aligned} \tag{16}$$

where

$$p_{[gj]}(d_1, d_2, d_3, d_4; \boldsymbol{\theta}) = \Pr_\theta\{y_{g_1} = d_1, y_{g_2} = d_2, y_{j_1} = d_3, y_{j_2} = d_4\} = \iiiint_{\mathcal{V}} f_{[gj]}(u_1, u_2, u_3, u_4)du_1du_2du_3du_4$$

and where $\mathcal{V} = \{(u_1, u_2, u_3, u_4) : s_iu_i > s_i\mathbf{x}_{\rho,g_i}\boldsymbol{\beta},\ i = 1\dots, 4\}$. In (16), we denoted the probabilities from equation (7) by $p_g(d_1, d_2; \boldsymbol{\theta})$, making explicit reference to the dependence on the parameter vector.

Theorem 5.3. *Under Assumptions 1–6 and 8–10 and if, for all couples $g = (g_1, g_2)$, $j = (j_1, j_2)$, $g \neq j$, we have*

$$\inf_{(d_1,\dots,d_4)\in\{0,1\}^4} p_{[gj]}(d_1, d_2, d_3, d_4; \boldsymbol{\theta}_0) > \delta,$$

$\|\hat{J}_n(\hat{\boldsymbol{\theta}}_n) - J(\boldsymbol{\theta}_0)\| = o_p(1)$.

Note that the condition on the joint distributions $p_{[gj]}(\cdot; \boldsymbol{\theta}_0)$ prevents the existence of too-strong dependences between any couples g, j: if, for example, conditional on a particular couple – say, $(y_{j_1}, y_{j_2}) = (1, 0)$ – only the value $(y_{g_1}, y_{g_2}) = (1, 0)$ could occur with probability 1, the condition would be violated, because of

$$p_{[gj]}(d_1, d_2, 1, 0; \boldsymbol{\theta}_0) = p_j(1, 0; \boldsymbol{\theta}_0) \cdot \Pr_{\theta_0}\{y_{g_1} = d_1, y_{g_2} = d_2 \mid y_{j_1} = 1, y_{j_2} = 0\} = 0$$

for all $(d_1, d_2) \neq (1, 0)$.

A second approach consists in using the estimator of $J(\boldsymbol{\theta}_0)$ proposed by Conley (1999). Since we have the pairs' contributions to the score in equation (9), we can modify the Conley's estimator of $J(\boldsymbol{\theta}_0)$ (equation 3.13 page 12 in his paper) in the following way

$$\hat{J}_\tau(\hat{\boldsymbol{\theta}}) = \tfrac{1}{n_\tau} \textstyle\sum_{j=0}^{L_M} \sum_{m=j+1}^{M} K_M(j) \left(\nabla_m \left(\hat{\boldsymbol{\theta}}_\tau\right) \nabla_{m-j} \left(\hat{\boldsymbol{\theta}}_\tau\right)' + \nabla_{m-j} \left(\hat{\boldsymbol{\theta}}_\tau\right) \nabla_m \left(\hat{\boldsymbol{\theta}}_\tau\right)' \right) - \tfrac{1}{n_\tau} \textstyle\sum_{m=1}^{M} \nabla_m \left(\hat{\boldsymbol{\theta}}_\tau\right) \nabla_m \left(\hat{\boldsymbol{\theta}}_\tau\right)' \quad (17)$$

where

$$K_M(j) = \begin{cases} \left(1 - \frac{|j|}{L_M}\right) & \text{if } |j| < L_M \\ 0 & \text{else} \end{cases} \quad (18)$$

and where $K_M(j)$ are uniformly bounded weights such that $K_M(0) = 1$ and $K_M(j) \to 1$ as $M \to \infty$, n_τ is the random number of selected spatial units in the sample, see Conley (1999, page 5), $L_M = o\left(M^{1/3}\right)$, $\nabla_m \left(\hat{\boldsymbol{\theta}}_\tau\right)$ is the score of the m–th pair of the selected spatial units with m the mean coordinates, and the subscript τ refers to the fact that n_τ and $\hat{\boldsymbol{\theta}}_\tau$ depend on the dimension of the subsample region which increases in area as $\tau \to \infty$. The dimension of the subsample region depends on the values of M and L_M. Our modification of the Conley's estimator is such that we directly work with pairs rather than single units in space.

A third approach, explored in the application, consists of using a parametric bootstrap procedure. Given the estimator $\hat{\boldsymbol{\theta}}_n$ and the matrices $\mathbf{W}_n$ and $\mathbf{X}_n$, a procedure can be described as follows:

(1) For each $b = 1, \ldots, B$, a vector $\mathbf{y}_b^\star = \{\mathbf{y}_1^\star, \ldots, \mathbf{y}_g^\star, \ldots, \mathbf{y}_{G_n}^\star\}$ of G_n independent couples of binary variables is generated through the distribution $\mathbf{y}_g = (d_1, d_1)$, with probability $p_g(d_1, d_2; \hat{\boldsymbol{\theta}}_n)$, $d_i = 0, 1$, $i = 1, 2$.

(2) $\ell_n^\star = \ell_n(\boldsymbol{\theta}; \mathbf{y}^\star)$ is computed from equation (8), and its maximizer over $\boldsymbol{\Theta}$, $\hat{\boldsymbol{\theta}}_b^\star$ is found.

(3) The variance $\mathbb{V}\mathrm{ar}^\star \left(\hat{\boldsymbol{\theta}}^\star\right)$ is the bootstrap estimator of the variance of $\hat{\boldsymbol{\theta}}_n$.

This approach provides consistent estimates of the sampling distribution of $\hat{\boldsymbol{\theta}}_n$, provided that all the assumptions required for asymptotic normality of $\hat{\boldsymbol{\theta}}_n$ are met.

Theorem 5.4. *Under Assumptions 1–6 and 8–10, the parametric bootstrap described in steps (1)–(3) is consistent for the distribution of $\hat{\boldsymbol{\theta}}_n$.*

Finally, an alternative parametric bootstrap procedure can also be defined, relying on the Gaussian latent model: (i) Generate iid Gaussian errors; (ii) Compute the bootstrap latent variables from (2) with $\boldsymbol{\theta} = \hat{\boldsymbol{\theta}}_n$ and $\mathbf{X}_n$ fixed and the corresponding bootstrap sample for the binary vector $\mathbf{y}^\star$; (iii) Obtain the bootstrap estimate $\hat{\boldsymbol{\theta}}_b^\star$ by maximizing equation (8) w.r.t. $\boldsymbol{\theta}$; (iv) Repeat (i)–(iii) B times and use the sampling variance of $\hat{\boldsymbol{\theta}}_b^\star$, $b = 1, \ldots, B$ to estimate the variance of $\hat{\boldsymbol{\theta}}_n$.

5.2. Choice of couples and identification assumptions

For each fixed n, we can implement the procedure in section 4 and define the groups accordingly. This will generate a sequence of groupings that potentially depend on the chosen initial values for (ρ, λ), and that in turn might affect some of the assumptions for the consistency of our PMLE. However, we claim this problem is not likely to affect Assumption 4, which is the key identification condition.

To justify our claim, note that a necessary condition for Assumption 4 is clearly that $\boldsymbol{\theta}$ must be equal to $\boldsymbol{\theta}_0$ if and only if

$$\lim_{n\to\infty} \mathbb{E}_{\theta_0} \frac{\partial \ell_n(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} = \lim_{n\to\infty} \sum_{g=1}^{G_n} \sum_{d_{g_1}, d_{g_2} \in \{0,1\}^2} \frac{\partial p_g(d_1, d_2; \boldsymbol{\theta})}{\partial \boldsymbol{\theta}} \frac{p_g(d_1, d_2; \boldsymbol{\theta}_0)}{p_g(d_1, d_2; \boldsymbol{\theta})} = 0. \tag{19}$$

While the *if* implication is trivially verified, because $\sum_{d_{g_1}, d_{g_2}} \frac{\partial p_g(d_1, d_2; \boldsymbol{\theta}_0)}{\partial \boldsymbol{\theta}} = 0$ by construction, identification requires also proving that no $\boldsymbol{\theta} \neq \boldsymbol{\theta}_0$ exists, for which the expected score in equation (19) is zero. Ensuring this by finding primitive conditions is an unbelievably difficult problem, even under simple structures of the weight matrices $\mathbf{W}_n$ and $\mathbf{M}_n$, and it goes beyond the scopes of the present paper. However, we discuss two very particular cases that might shed some light on situations when identification fails:

(i) Under a SARAR(1,1) probit specification, the identity $\mathbf{W}_n = \mathbf{M}_n$ puts at risk the identification of the spatial autocorrelation parameters, when the true regressor coefficients are near zero given that $\ell_n(0, \rho, \lambda) = \ell_n(0, \lambda, \rho)$.

(ii) Consider a SAR(1) probit model with a single regressor and assume that each unit has a unique neighbor, so that $\mathbf{W}_n$ is a Boolean matrix. By choosing the appropriate couples (w.l.o.g. consecutive couples), equation (8) coincides with the log–likelihood, $\mathbf{A}_\rho$ and $\boldsymbol{\Sigma}$ are block diagonal with the same 2×2 blocks, and further, $\mathbf{x}'_{\rho,2g-1} = (1, (1-\rho)\, x_{2g-1} + \rho x_{2g}) = \mathbf{x}'_{1-\rho,2g}$. In this setting, the parameter $\rho > 0$ fails to be identified if there is no variability in the regressors $\mathbf{X}$ within each correlated couple – i.e., if $x_{2g-1} = x_{2g}$, for all $g = 1, \ldots, n/2$ – because in this case, $\mathbb{E}_0 \ell_n(\boldsymbol{\beta}_0, \rho_0) = \mathbb{E}_0 \ell_n(\boldsymbol{\beta}_0, 1 - \rho_0)$.

The above–explained two examples are only very special cases, but from both of them, we understand that the failure of the identification assumption is possible when there is a pathological behavior of the partial loglikelihood function that is invariant under permutations of the units. In particular, in (i), the elements of $\mathbf{W}_n$ and $\mathbf{M}_n$ coincide for all units, whereas, in (ii), the identity $\mathbb{E}_0 \ell_n(\boldsymbol{\beta}_0, \rho_0) = \mathbb{E}_0 \ell_n(\boldsymbol{\beta}_0, 1 - \rho_0)$ depends on the fact that the values of the regressor X at all spatially correlated locations is the same.

Among the assumptions necessary for the consistency of the estimator, Assumption 5(b) seems to be another condition potentially affected by the couple selection criterion. It is, however, rather easy to guarantee that it holds for all possible permutations by replacing it with the following stronger condition: $\sup_{i,j} | \sum_{d_i, d_j} (\Pr_{\theta_0}(y_i = d_i, y_j = d_j))^{-1} | < \infty$. This condition is clearly stronger than Assumption 5(b), and it basically excludes correlations that are too strong between couples of units.

6. Marginal effects

In nonlinear regressions, the interpretation of the marginal effects in terms of the change in the conditional mean of $\mathbf{y}$ when regressors $\mathbf{X}$ change by one unit is no longer possible. The effects arising from changes in the explanatory variables depend in a nonlinear way on the levels of these variables, i.e., changes in the explanatory variable near the mean have a very different impact on decision probabilities than changes in very low or high values. For spatial autoregressive probit models, the nonlinearity increases in the evaluation of the marginal effects; see Beron and Vijverberg (2004) and LeSage et al. (2011). Recently, Billé (2014) has also pointed out the main consequences in evaluating marginal effects with and without the consideration of heteroskedasticity implied by the spatial autocorrelation coefficient.

Let $\mathbf{x}_{.h} = (x_{1h}, x_{2h}, ..., x_{ih}, ..., x_{nh})'$ be an n–dimensional vector of units referred to the h–th regressor, $h = 1, \ldots, k$, and $\mathbf{x}_{i.} = (x_{i1}, x_{i2}, ..., x_{ih}, ..., x_{ik})'$ be a k–dimensional vector of regressors referred to unit i. By considering the equations in (3), we propose the following specifications of the marginal effects

$$\begin{aligned}
\frac{\partial \mathrm{P}\left(y_i = 1\right)}{\partial \mathbf{x}'_{.h}} \,|_{\bar{\mathbf{x}}} &= \phi\left(\{\mathbf{\Sigma}_{\nu(\rho,\lambda)}\}_{ii}^{-1/2}\left\{\mathbf{A}_\rho^{-1}\bar{\mathbf{X}}\right\}_{i.}\boldsymbol{\beta}\right)\{\mathbf{\Sigma}_{\nu(\rho,\lambda)}\}_{ii}^{-1/2}\{\mathbf{A}_\rho^{-1}\}_{i.}\beta_h \\
\frac{\partial \mathrm{P}\left(y_i = 1\right)}{\partial \mathbf{x}'_{.h}} \,|_{\mathbf{x}} &= \phi\left(\{\mathbf{\Sigma}_{\nu(\rho,\lambda)}\}_{ii}^{-1/2}\left\{\mathbf{A}_\rho^{-1}\mathbf{X}\right\}_{i.}\boldsymbol{\beta}\right)\{\mathbf{\Sigma}_{\nu(\rho,\lambda)}\}_{ii}^{-1/2}\{\mathbf{A}_\rho^{-1}\}_{i.}\beta_h
\end{aligned} \tag{20}$$

where $\mathbf{\Sigma}_{\nu(\rho,\lambda)}$ is the variance–covariance matrix implied by the reduced form of a SARAR(1,1) probit model, $\bar{\mathbf{X}}$ is an n by k matrix of regressor means, $\{\cdot\}_{i.}$ is the i–th row of the matrix inside, and $\{\cdot\}_{ii}$ is the i–th diagonal element of a square matrix. Note that $\mathbf{\Sigma}_{\nu(\rho,\lambda)}$ reduces to $\mathbf{\Sigma}_{\mathbf{u}(\rho)}$ for a SAR(1) probit specification as in equation (4) with $\mathbf{u} = \mathbf{A}_\rho^{-1}\boldsymbol{\varepsilon}$.

The first specification of the equations in (20) explains the impact of a marginal change in the mean of the h–th regressor, i.e., $\bar{\mathbf{x}}_{.h}$, on the conditional probability of $\{y_i = 1\}$, i.e., $\mathrm{P}\left(y_i = 1\right)$, setting $\bar{\mathbf{x}}_{.h'}$ for all the remaining regressors, $h' = 1, \ldots, k-1$. The second specification of the equations in (20) considers instead the marginal impact evaluated at each single value of $\mathbf{x}_{.h}$. This is particularly informative in space since the possibility of evaluating a marginal impact with respect to a particular value $\mathbf{x}_{ih}$ has the same meaning of considering a marginal impact in a particular region/site for regressor h. The results are two n–dimensional square matrices for $\{y_1, y_2, \ldots, y_n\}$. Both the specifications should be evaluated with consistent estimates of the spatial autocorrelation coefficients $\left(\hat{\rho}, \hat{\lambda}\right)$. In section 7.2.1, we report results on the robustness of the marginal effects in the case of model misspecification implied by wrong assumed weighting matrices.

Spatial marginal effects are then split into an *average direct impact* and an *average indirect impact*. The average of the main diagonal elements of the n–dimensional matrix in both the equations is the average direct effect (i.e., the impact from the same region). The average of the cumulated off–diagonal elements is the average indirect effect – due to spatial spillover effects (i.e., the impact from other regions). Finally, the average total effects is the sum of these two (LeSage and Pace, 2009). Changes in the value of an explanatory variable in a single observation (i.e., a spatial unit) i may influence all the other $n-1$ observations. The scalar summary measure of indirect effects cumulates the spatial spillovers falling on all other observations, but the magnitude of impact will be greatest for nearby neighbors and declines in magnitude for higher–order neighbors. This comes out from the infinite series expansion. LeSage et al. (2011) pointed out the need to calculate measures of dispersion for these estimates. In section 7, we give some results on the marginal effects and their measures of dispersion based on our Monte Carlo simulations.

Observation–level total effects estimates, sorted from low to high values of each regressors, can be also viewed as an important measure of spatial variation in the impacts (Lacombe and LeSage, 2013). This kind of interpretation permits also to account for *spatial heterogeneity* given the variation over space of the marginal impacts with respect to the spatial distribution of the regressors. See Billé et al. (2017) for a two–step approach specifically thought to account for unobserved groupwise (discrete) spatial heterogeneity in the $\boldsymbol{\beta}$ coefficients via iterated local estimation procedures. We show some results on this issue in the empirical application in section 8. Finally, note that the specification of our marginal effects are different compared with those proposed by LeSage et al. (2011) and Beron and Vijverberg (2004).

7. Finite sample properties

In this section, we study the finite sample properties of our PMLE for the SARAR(1,1) probit model specified in equation (1) and the SAR(1) probit model specified in equation (4). For the finite sample properties of the linear SAR(1) model, see, e.g., Bao and Ullah (2007).

We plan different Monte Carlo experiments. All the DGPs are based on a fixed matrix $\mathbf{X} = [\mathbf{x}_{.0}, \mathbf{x}_{.1}, \mathbf{x}_{.2}]$ of dimension $n \times 3$, which is composed of two regressors $\mathbf{x}_{.1}$, $\mathbf{x}_{.2}$ and a constant $\mathbf{x}_{.0}$, with $\mathbf{x}_{ij} = (\mathbf{x}_{1j}, \mathbf{x}_{2j}, \ldots, \mathbf{x}_{nj})'$ and $j = 0, 1, 2$. The regressor $\mathbf{x}_{.1}$ is drawn from a $\mathcal{U}(-1, 1)$ distribution, and $\mathbf{x}_{.2}$ is drawn from a $\mathcal{N}(0, 1)$, whereas the true beta vector of the parameters is fixed to $\boldsymbol{\beta} = (0, 1, -0.5)'$. The autoregressive parameter ρ in the SAR(1) probit experiment takes the values $\{-0.8, -0.6, -0.4, -0.2, 0, 0.2, 0.4, 0.6, 0.8\}$, and $\mathbf{W}_n$ is a nonnegative weight and then normalized n–dimensional weight matrix. Finally, the number of simulation runs is $1,000$ each.

7.1. Weighting matrices

In our Monte Carlo experiment, we consider both *sparse* and *dense* matrices. The former is a k–nearest neighbor matrix built on regular square lattice grids of dimensions (a) 10×10 with $n = 100$, (b) 30×30 with $n = 900$, and (c) 50×50 with $n = 2,500$. The latter is an inverse distance–based matrix built on randomly generated coordinates from $\mathcal{U}(0, 50)$ and $\mathcal{U}(-70, 20)$, for $n = 900$ only. The coordinates are then used to define (Euclidean) distances among couples of units, and they can also be interpreted as centroids of areal units in the case of a discrete space.

It is worth noting that in the case of the k–nn criterion, the spatial information does not depend on how much units are distant from each other, but it guarantees a constant spatial statistical information, ensuring no difference between simulations built on regular/irregular grids and randomly generated coordinates. Regular grids are also suitable to avoid the problem of selecting more distant observations in the neighboring set $\mathcal{N}_k$ since they are somehow realistic for homogeneous point patterns.

The weighting matrix $\mathbf{W}_n$ must be normalized to obtain a proper parameter space of its corresponding autoregressive coefficient ρ. In the majority of the experiments, we consider the row normalization rule (i.e., $\mathbf{W}_n$ is a row–stochastic matrix). Row normalization has the appealing role of interpreting the spatial lag function as a weighted average of the (first–order) neighbors for each site in space. With inverse distance–based matrices, the row normalization does not lead to an easy economic interpretation of the spatial impacts. In particular, when considering distance decay or negative exponential functions rather than first–order contiguity matrices (e.g., queen criterion), the interpretation of the absolute role of the distance metric is usually lost. Moreover, as emphasized by Kelejian and Prucha (2010), the model with row–normalized weight matrices is no more equivalent to the original spatial one, with the exception of the k–nn approach.

For some experiments, we then consider the spectral normalization rule by rescaling the weighting matrix using its largest eigenvalue in absolute value, to ensure the following: (i) a proper parameter space for ρ (see lemma 2.1), and (ii) the equivalence of the spatial models before and after normalization of the weights.

7.2. Finite sample results: the SAR(1) probit model

In this section, we show the finite sample properties of the PMLE and the marginal effects calculated as in equation (20). The DGPs are built on the SAR(1) probit model with a fixed k–nn weighting matrix ($k = 11$), distinguishing between different true values of ρ and sample sizes n. Results are reported in Tables A.1 and A.2 and Figure B.1.

Table A.1 reports the summary statistics of our PMLE. The estimates of the $\boldsymbol{\beta}$ vector are good in terms of both unbiasedness and consistency in finite samples, aside from the different true values assumed by the autocorrelation ρ. We slightly underestimate the autocorrelation parameter ρ, especially as the true value approaches its upper limit, while the standard deviation (*sd*) and RMSE decreases as ρ increases. Figure B.1 shows the Gaussian kernel density functions for different sample sizes. The empirical distributions for all the parameters highly improve as the sample size increases. The Monte Carlo distribution of the estimators of the $\boldsymbol{\beta}$ parameters is approximately bell–shaped, whereas the distribution of $\hat{\rho}$ is quite asymmetric for $n = 100$, although the asymmetry rapidly tends to disappear for larger sample sizes.

Table A.2 shows the direct, indirect, and total impacts for $n = 900$ calculated as in the equations in (20) with regard to the mean value and to each observation, respectively. In both cases, estimated mean impacts, $m(\hat{\rho})$, are highly close to their true values $m(\rho)$ for different values of ρ. Slight differences can be found as the value of ρ increases in absolute value, manly because of differences in the indirect effects.

Table A.3 reports the finite sample properties of our PMLE when the sample size is fixed to $n = 658$ (the same as the empirical application in section 8) and $\rho = (0.4, 0.8)$. A comparison among the empirical standard deviations, i.e., sd, based on 1,000 Monte Carlo replications, the asymptotic standard deviations estimated through equation (16), and the average of the negative Hessian matrix at $\hat{\boldsymbol{\theta}}$, i.e., $\widehat{sd}_a$, and the true asymptotic standard deviations, i.e., sd_a, is reported. We can observe that the values of the estimated asymptotic standard deviations $\widehat{sd}_a$ are very closed to the true values, especially when $\rho = 0.4$.

7.2.1. Misspecification of $\mathbf{W}$

In this section, we provide some Monte Carlo results to check the robustness of our PMLE with a misspecification of the SAR(1) probit model by assuming a *sparse* weighting matrix rather than a *dense* one. We fixed $n = 900$, while $\rho \in \{-0.6, 0.6\}$ and $\boldsymbol{\beta} = (0, 1, -0.5)'$. The true dense matrix is built on inverse distance–based functions, distinguishing between the row normalization ($\mathbf{W}_{rn}$) and the spectral normalization ($\mathbf{W}_{sn}$) cases, whereas the assumed sparse weighting matrix is based on a k–nn approach, with $k = 11$ as before ($\mathbf{W}_{knn}$).

Results are reported in Tables A.4 and A.5 and Figure B.2. Table A.4 shows that the PML estimates of the $\boldsymbol{\beta}$ coefficients are quite robust with misspecified $\mathbf{W}_n$ matrices. The misspecification of the ρ coefficient is more evident, as expected. Table A.5 reports the main empirical results on the robustness of the marginal impacts. The indirect effects are not well accounted for because of the estimation of ρ, but the direct effects are robust. Finally, Figure B.2 shows the Gaussian kernel density functions for both types of misspecification, which are quite symmetric around the true values, with the exception of ρ. There seem to be no significant differences in terms of the distributions when considering the two types of normalization rules, i.e., $\mathbf{W}_{sn}$ and $\mathbf{W}_{rn}$. One notable exception is the case of β_1, where the row normalization has higher probability density on the true value of the parameter, while the spectral normalization is more symmetric around its mean.

7.2.2. The choice of couples and sparsity of $\mathbf{W}$

We run some Monte Carlo experiments aimed at assessing the performance of the algorithm introduced in section 4. We use the R library called `Rpython` to run a program able to create and manipulate graphs and networks by exploiting the function `networkx.max_weight_matching` inside the package `networkx`. Data are simulated from a SAR(1) probit with $\boldsymbol{\beta} = (0, 1, -0.5)'$ and $\rho = 0.6$, using either a k–nn matrix with $k = 11, 25, 50, 100$ or an inverse distance matrix. We compute the PML estimates using an initial guess for the parameter $\tilde{\rho}$ equal in sign to the true value ρ and then compare these estimates with the PML estimates obtained without the application of the maximum matching algorithm (the *default* pair choice corresponds to coupling units $(2g - 1, 2g)$ for all g). In the case of distance weight matrices, to determine the sensitivity of the procedure to the initial guess, we use two different values of $\tilde{\rho}$, both equal in sign to the true value ρ, one exactly equal to ρ and the other significantly smaller.

We expect the impact of the pair choice to increase as $\mathbf{W}_n$ becomes denser. Indeed, in the k–nn case, the maximum matching method proves to be slightly inefficient compared to the *default* pair choice until $k = 50$, when they are pretty much the same in terms of both *sd* and RMSE. For $k = 100$, the situation is reversed, with the maximum–matching *sd* and RMSE about 10% smaller relative to the default case. However, as k increases, the *sd* of both the estimators increases rapidly. The tables are available upon request.

Table A.6 reports the main summary statistics of the MC distribution of the *default* and maximum–matching estimators when $\mathbf{W}_n$ is an inverse distance weight matrix. The gain in terms of *sd* is quite relevant (-36% for the *sd* of $\hat{\rho}$ in the case of the spectral normalization), while a smaller increase of negative bias occurs, with an overall variation of RMSE of -30%. This seems to be a consequence of better behavior of the loglikelihood function that reduces drastically the occurrence of an optimum value $\hat{\rho}$ near the boundary ($\hat{\rho} \approx 1$). There exists a slight improvement in the *sd* of the $\hat{\boldsymbol{\beta}}$s and no effect on their means. The initial guess $\tilde{\rho}$ appears to have a negligible effect.

7.3. Finite sample results: the SARAR(1,1) probit model

We conclude our simulation analysis by showing some results of the estimation of 200 repeated draws of SARAR(1,1) probit samples of medium size ($n = 900$). We draw samples from the model in equation (1), assuming $\boldsymbol{\beta} = (0, 1, -0.5)'$ and $\rho = 0.6$ fixed. The weight matrix $\mathbf{W}_n$ is a k–nn with the number of nearest neighbors equal to 11. For the weighting matrix $\mathbf{M}_n$, we choose a queen contiguity criterion to define the weights inside, and then we row–standardize. The choice of the two very different weighting matrices prevents potential problems of identification.

Table A.7 presents the results, for different values of the parameter λ, namely, $\lambda \in \{0.8, 0.6, 0.4, 0.2\}$. Similar to what happens in the SAR case, the estimates of the $\boldsymbol{\beta}$ parameters are quite precise, while both the autocorrelation coefficients tend to be downward biased. The bias of ρ seems to be slightly increasing with λ; similarly, the lower the true value of λ, the lower the bias of $\hat{\lambda}$.

The standard deviation of the estimators of all the parameters (except λ itself) is monotonically increasing with λ: the relative increment of the standard deviations from case $\lambda = 0.2$ to $\lambda = 0.8$ is between 70% and 242%. Further, a comparison of the RMSE from Table A.1 (case $\rho = 0.6$) shows that $\hat{\rho}$ and $\hat{\beta}_0$ are particularly sensitive to the introduction of spatial autocorrelation in the errors, showing an increment of about 50% in the case of minimum autocorrelation ($\lambda = 0.2$), whereas the RMSE of the other estimators remains almost unchanged.

Finally, to get an intuition of the behavior in the case of the dense weight matrix, we make some simulations by using an inverse distance matrix $\mathbf{M}_n$. Although the performance of $\hat{\lambda}$ dramatically worsens in terms of RMSE (mainly because of a boost in *sd*), switching from a sparse to a dense weight matrix governing the error spatial correlation structure has almost no effect on all the other parameters both in terms of bias and *sd*. This also implies that the estimation of the marginal effects is not affected by this change. Results are available upon request.

8. Empirical application

In this section, we propose to replicate the empirical application in LeSage et al. (2011) by estimating the parameter sets $\boldsymbol{\theta} = \left(\boldsymbol{\beta}', \rho\right)'$ with our PMLE. The model specification is referred to a SAR(1) probit in equation (4). The data set used for this exploration entails 673 establishments tracked weekly during the year following Hurricane Katrina and then seasonally and annually in subsequent years. The data set is freely available in the R package *ProbitSpatial* (Martinetti and Geniaux, 2016), and details are referred to LeSage et al. (2011). We have found some points/units with the same coordinates. To avoid "zero distance" problems, we eliminate 15 observations from the data set, with a final sample dimension of $n = 658$.

The economic aim was to evaluate which factors have influenced decisions of establishments in reopening in the aftermath of Hurricane Katrina. A probabilistic decision mechanism is then easily described by a probit model, where each decision to reopen is defined by the event $\{y_i = 1\}$. Spatial effects are accounted for to consider potential endogenous network effects among these decisions so that the utility associated with an establishment reopening directly depends on the neighboring utilities, which in turn have effects on reopening decisions.

Coherently with their analysis, a SAR(1) probit model is estimated for three different time horizons: (a) 0–3 months, (b) 0–6 months, and (c) 0–12 months. The weighting matrix is built on a k–nn criterion, with $k = 11$ for time horizon (a) and $k = 15$ for time horizons (b) and (c). In each time horizon the firms' decisions are supposed to be simultaneous. Explanatory variables are the flood depth (measured in feet) at the location of the individual establishments, (log) median income for the census block group in which the store was located, two dummy variables reflecting small and large firms (with medium size firms representing the omitted class), two dummy variables reflecting the low and high socioeconomic class of the store *clientèle* (with the middle socioeconomic class excluded), and two dummy variables for type of store ownership, one reflecting sole proprietorships and the other representing national chains (with regional chains representing the excluded class).

Table A.8 shows the PML estimates and their standard errors to be compared with those in Table 3 in LeSage et al. (2011). As we can observe, the bootstrap standard errors are close to the Bayesian ones, and most of the time our standard errors are slightly smaller. We obtain standard errors of our PML estimates by using the first parametric bootstrap approach proposed in subsection 5.1, which we call for convenience the "probit" bootstrap. As a comparison, we also include alternative MLE–based estimators and their standard errors (if available): (i) the approximate MLE (AMLE) by Martinetti and Geniaux (2017) and (ii) the composite (univariate) MLE (CMLE) by Mozharovskyi and Vogler (2016). The estimates from the CML estimator are quite different, while the ones from the AMLE are very close to the others. Finally, we also estimate a

SARAR(1,1) probit model with our PMLE finding that the λ coefficient is negative and statistically significant, while the ρ coefficient is increased. Our estimates of the SARAR(1,1) probit model thus evidence a residual spatial correlation unaccounted for by the spatial autoregressive component on the latent variables.

Table A.9 provides marginal effects – see the equations in (20) – for each time horizon to be compared with the effects reported in Tables 4, 5, and 6 in LeSage et al. (2011). All tables show that PML estimates are consistent with Bayesian estimates; in particular, the PML estimate of the spatial correlation coefficient ρ is positive and significant as well as higher than the corresponding Bayesian estimate, for all the three time horizons. As a consequence, our estimates of the indirect effects are generally higher, in absolute value, compared to the corresponding indirect effects reported by LeSage et al. (2011). As stressed in section 6, potential spatial heterogeneity in terms of the marginal effects should be accounted for in empirical applications. Figure B.3 shows an interesting variability of the total impacts for the first time horizon, revealing that the total marginal impacts is even around zero for some spatial units. The same figures for the second and third time horizons can be found in supplementary material.

Finally, we computed in Table A.10 alternative estimates of the standard deviations of $\hat{\boldsymbol{\beta}}$: we obtained estimates of the variances by using the two asymptotic estimators presented in section 5.1 as well as the second bootstrap estimator based on the reduced form latent model described after Theorem 5.4, which we call for convenience the "latent" bootstrap. The two asymptotic estimators are very similar, and tend to have values higher than the bootstrap, especially for the intercept and (particularly in the case of Conley's formula) for $\hat{\rho}$. The "latent" bootstrap estimates are all slightly higher than the "probit" bootstrap ones, but lower than the asymptotic values.

9. Conclusions

In this paper, we derive the asymptotic properties and evaluate the finite sample properties of a partial maximum likelihood estimator (PMLE) for the spatial (first–order) autoregressive probit model with (first–order) autoregressive disturbances, i.e., an SARAR(1,1) probit model. Different from Wang et al. (2013), we consider the more general and interesting case of direct correlation among the dependent variables, which specifies at least the SAR(1) probit rather than a simple SAE(1) probit model. We propose a Kullback–Leibler approach for choosing the couples that maximize the partial loglikelihood function, and we suggest exact formulas for defining the marginal effects in spatial binary contexts. Cases of model misspecifications are also included. In addition, methods for estimating the asymptotic variance–covariance matrix and directly obtaining the standard errors through bootstrap approaches are also reported. Finally, the derivation of explicit expressions of the score vector (given in the supplementary material) can also be of interest in itself, for example used in the approach of Mozharovskyi and Vogler (2016) to improve the computations.

The PMLE is consistent given some regularity conditions. Unlike Wang et al. (2013), our simulations suggest that the estimator performs very well even with small sample sizes. The results substantially improve as the sample size increases both in terms of bias and standard deviation. All the distributions are bell–shaped from moderate to large samples and for all the values of correlations considered. The marginal effects calculated on the simulated data with respect to the mean and with respect to individual observations are also consistent and quite near the true values. In the SARAR(1,1) probit case, the estimator performs reasonably well, although a slight loss in efficiency exists, in particular for $\hat{\rho}$ and $\hat{\beta}_0$. However, this efficiency loss relative to the estimates from the SAR(1)-probit specification is not found in the empirical application.

We consider model misspecification given the assumption of an incorrect weighting matrix: in these cases, the estimator properties and the direct marginal effects are robust in terms of the $\boldsymbol{\beta}$ coefficients. This analysis confirms that an incorrect choice of the spatial weighting matrix greatly impacts on the estimation of the autocorrelation coefficient and, as a consequence, of the indirect effects, thus suggesting that great care must be paid to model selection. In our empirical application, results suggest that our PMLE estimator gives parameter estimates and standard deviations quite similar to those obtained by the Bayesian (LeSage et al., 2011) approach and by the AMSLE (Martinetti and Geniaux, 2017), but it tends in general to give higher estimates for ρ. Moreover, a comparison with other MLE–based estimators in terms of the estimated parameters is also included. Finally, the KL–based criterion proposed for choosing the couples deserves further investigation since it proves to be a promising method that could be applied to approximate a complex models with a simpler one, controlling the information loss.

Acknowledgments

The first author is indebted to Prof. Paul Elhorst for having emphasized the inconsistency problem during a brief discussion of a previous paper and for encouraging her to further explore this issue. The authors wish to thank all the participants of the 9th Jean Paelinck Seminar on “Nonparametric Spatial Econometrics: Theory and Empirical Issues”, the First Italian Workshop of Econometrics and Empirical Economics (IWEEE 2018), the XIIth World Conference of the Spatial Econometrics Association (SEA), and the 11th International Conference of the ERCIM WG on Computational and Methodological Statistics (CM Statistics 2018) for their insightful comments. We also thank the anonimous reviewers and the associate editor for their comments, that helped to substantially improve the final version of the paper. All remaining errors are ours. This research has been carried out within the project “Aggregation functions for innovation and data analysis” (AIDA) at the Free University of Bozen–Bolzano and within the project “Estimation of transformation models” at the Bocconi Institute for Data Science and Analytics.

References

Anselin, L. (1988). *Spatial econometrics: methods and models*, volume 4. Springer Science & Business Media.

Arbia, G. (2014). Pairwise likelihood inference for spatial regressions estimated on very large datasets. *Spatial Statistics*, 7(Supplement C):21–39.

Bai, Y., Kang, J., and Song, P. X.-K. (2014). Efficient pairwise composite likelihood estimation for spatial–clustered data. *Biometrics*, 70(3):661–670.

Baltagi, B. H., Egger, P. H., and Kesina, M. (2017). *Bayesian Spatial Bivariate Panel Probit Estimation*, chapter 4, pages 119–144.

Bao, Y. and Ullah, A. (2007). Finite sample properties of maximum likelihood estimator in spatial models. *Journal of Econometrics*, 137(2):396–413.

Beron, K. J., Murdoch, J. C., and Vijverberg, W. P. (2003). Why cooperate? public goods, economic power, and the montreal protocol. *Review of Economics and Statistics*, 85(2):286–297.

Beron, K. J. and Vijverberg, W. P. M. (2004). *Advances in Spatial Econometrics: Methodology, Tools and Applications*, chapter Probit in a Spatial Context: A Monte Carlo Analysis, pages 169–195. Springer Berlin Heidelberg, Berlin, Heidelberg.

Besag, J. E. (1972). Nearest-neighbour systems and the auto-logistic model for binary data. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 75–83.

Besag, J. E. (1974). Spatial interaction and the statistical analysis of lattice systems. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 192–236.

Bhat, C. R. (2011). The maximum approximate composite marginal likelihood (macml) estimation of multinomial probit-based unordered response choice models. *Transportation Research Part B: Methodological*, 45(7):923–939.

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(2, 3):131–154.

Billé, A. G. and Arbia, G. (2019). Spatial limited dependent variable models: A review focused on specification, estimation, and health economics applications. *Journal of Economic Surveys*.

Billé, A. G., Benedetti, R., and Postiglione, P. (2017). A two–step approach to account for unobserved spatial heterogeneity. *Spatial Economic Analysis*, 0(0):1–20.

Breslaw, J. A. (2002). Multinomial probit estimation without nuisance parameters. *The Econometrics Journal*, 5(2):417–434.

Case, A. (1992). Neighborhood influence and technological change. *Regional Science and Urban Economics*, 22(3):491–508.

Catania, L. and Billé, A. G. (2017). Dynamic spatial autoregressive models with autoregressive and heteroskedastic disturbances. *Journal of Applied Econometrics*.

Cliff, A. D. and Ord, J. K. (1981). *Spatial processes: models & applications*. Taylor & Francis.

Cochrane, D. and Orcutt, G. H. (1949). Application of least squares regression to relationships containing auto-correlated error terms. *Journal of the American Statistical Association*, 44(245):32–61.

Conley, T. G. (1999). Gmm estimation with cross sectional dependence. *Journal of Econometrics*, 92(1):1–45.

Edmonds, J. (1965a). Maximum matching and a polyhedron with 0,1-vertices. *Journal of Research of the National Bureau of Standards Section B*, 69:125–130.

Edmonds, J. (1965b). Paths, trees, and flowers. *Can. J. Math*, 17:449–467.

Fleming, M. M. (2004). Techniques for estimating spatially dependent discrete choice models. In *Advances in Spatial Econometrics*, pages 145–168. Springer.

Galil, Z. (1986). Efficient algorithms for finding maximum matching in graphs. *ACM Comput. Surv.*, 18(1):23–38.

Gao, X. and Song, P. X.-K. (2010). Composite likelihood bayesian information criteria for model selection in high-dimensional data. *Journal of the American Statistical Association*, 105(492):1531–1540.

Heagerty, P. J. and Lele, S. R. (1998). A composite likelihood approach to binary spatial data. *Journal of the American Statistical Association*, 93(443):1099–1111.

Ibragimov, R. and Müller, U. K. (2010). t-statistic based correlation and heterogeneity robust inference. *Journal of Business & Economic Statistics*, 28(4):453–468.

Kapoor, M., Kelejian, H. H., and Prucha, I. R. (2007). Panel data models with spatially correlated error components. *Journal of Econometrics*, 140(1):97–130.

Kelejian, H. H. (2016). Critical issues in spatial models: error term specifications, additional endogenous variables, pre-testing, and bayesian analysis. *Letters in Spatial and Resource Sciences*, 9(1):113–136.

Kelejian, H. H. and 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(1):99–121.

Kelejian, H. H. and Prucha, I. R. (2007). Hac estimation in a spatial framework. *Journal of Econometrics*, 140(1):131–154.

Kelejian, H. H. and Prucha, I. R. (2010). Specification and estimation of spatial autoregressive models with autoregressive and heteroskedastic disturbances. *Journal of Econometrics*, 157(1):53–67.

Kelejian, H. H., Prucha, I. R., and Yuzefovich, Y. (2004). Instrumental variable estimation of a spatial autoregressive model with autoregressive disturbances: Large and small sample results. *Advances in Econometrics: Spatial and Spatio–Temporal econometrics*, pages 163–198.

Klier, T. and McMillen, D. P. (2008). Clustering of auto supplier plants in the united states: Generalized method of moments spatial logit for large samples. *Journal of Business & Economic Statistics*, 26(4):460–471.

Lacombe, D. J. and LeSage, J. P. (2013). Use and interpretation of spatial autoregressive probit models. *The Annals of Regional Science*, pages 1–24.

Lambert, D. M., Brown, J. P., and Florax, R. J. (2010). A two-step estimator for a spatial lag model of counts: Theory, small sample performance and an application. *Regional Science and Urban Economics*, 40(4):241–252.

Lee, L.-f. (2003). Best spatial two–stage least squares estimators for a spatial autoregressive model with autoregressive disturbances. *Econometric Reviews*, 22(4):307–335.

Lee, L.-F. (2004). Asymptotic distributions of quasi-maximum likelihood estimators for spatial autoregressive models. *Econometrica*, pages 1899–1925.

Lee, L.-f. and Yu, J. (2010). Estimation of spatial autoregressive panel data models with fixed effects. *Journal of Econometrics*, 154(2):165–185.

Lee, L.-f. and Yu, J. (2016). Identification of spatial durbin panel models. *Journal of Applied Econometrics*, 31(1):133–162.

LeSage, J. and Pace, R. K. (2009). Introduction to spatial econometrics. *Boca Raton, FL: Chapman & Hall/CRC.*

LeSage, J. P., Kelley Pace, R., Lam, N., Campanella, R., and Liu, X. (2011). New orleans business recovery in the aftermath of hurricane katrina. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 174(4):1007–1027.

Mammen, E. (1992). Bootstrap, wild bootstrap, and asymptotic normality. *Probability Theory and Related Fields*, 93(4):439–455.

Manski, C. (1981). *Alternative Estimators and Sample Designs for Discrete Choice Analysis.* The MIT Press.

Martinetti, D. and Geniaux, G. (2016). Probitspatial r package: Fast and accurate spatial probit estimations. In *22. International Conference on Computational Statistics (COMPSTAT)*, Oviedo, Spain.

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

McFadden, D. (2001). Economic choices. *American Economic Review*, pages 351–378.

McLeish, D. L. (1974). Dependent central limit theorems and invariance principles. *Ann. Probab.*, 2(4):620–628.

McMillen, D. P. (1992). Probit with spatial autocorrelation. *Journal of Regional Science*, 32(3):335–348.

McMillen, D. P. (1995). Selection bias in spatial econometric models. *Journal of Regional Science*, 35(3):417–436.

Mozharovskyi, P. and Vogler, J. (2016). Composite marginal likelihood estimation of spatial autoregressive probit models feasible in very large samples. *Economics Letters*, 148:87–90.

Ord, K. (1975). Estimation methods for models of spatial interaction. *Journal of the American Statistical Association*, 70(349):120–

126.

Pace, R. K. and LeSage, J. P. (2011). *Fast Simulated Maximum Likelihood Estimation of the Spatial Probit Model Capable of Handling Large Samples*, chapter 1, pages 3–34.

Pinkse, J. and Slade, M. E. (1998). Contracting in space: An application of spatial statistics to discrete-choice models. *Journal of Econometrics*, 85(1):125–154.

Qu, X. and Lee, L.-f. (2012). Lm tests for spatial correlation in spatial models with limited dependent variables. *Regional Science and Urban Economics*, 42(3):430–445.

Qu, X. and Lee, L.-f. (2013). Locally most powerful tests for spatial interactions in the simultaneous sar tobit model. *Regional Science and Urban Economics*, 43(2):307–321.

Sain, S. R. and Cressie, N. (2007). A spatial model for multivariate lattice data. *Journal of Econometrics*, 140(1):226–259.

Smirnov, O. A. (2010). Modeling spatial discrete choice. *Regional Science and Urban Economics*, 40(5):292–298.

Smith, T. E. and LeSage, J. P. (2004). A bayesian probit model with spatial dependencies. In *Spatial and Spatiotemporal Econometrics*, pages 127–160. Emerald Group Publishing Limited.

Wang, H., Iglesias, E. M., and Wooldridge, J. M. (2013). Partial maximum likelihood estimation of spatial probit models. *Journal of Econometrics*, 172(1):77–89.

Whittle, P. (1954). On stationary processes in the plane. *Biometrika*, pages 434–449.

Wooldridge, J. M. (2014). Quasi-maximum likelihood estimation and testing for nonlinear models with endogenous explanatory variables. *Journal of Econometrics*, 182(1):226–234.

Xu, X. and Lee, L.-f. (2015). Maximum likelihood estimation of a spatial autoregressive tobit model. *Journal of Econometrics*, 188(1):264–280.

Appendix A. Tables

	$\mathbf{n = 100}$					$\mathbf{n = 900}$					$\mathbf{n = 2,500}$				
True Value	Mean	Median	*sd*	RMSE	MAD	Mean	Median	*sd*	RMSE	MAD	Mean	Median	*sd*	RMSE	MAD
$\beta_0 = 0.0$	-0.016	-0.007	0.323	0.323	0.145	-0.003	-0.003	0.066	0.066	0.043	0.002	0.002	0.037	0.037	0.024
$\beta_1 = 1.0$	1.058	1.043	0.289	0.294	0.189	1.009	1.003	0.093	0.094	0.065	1.007	1.002	0.054	0.054	0.031
$\beta_2 = -0.5$	-0.526	-0.517	0.161	0.164	0.108	-0.503	-0.503	0.052	0.052	0.034	-0.499	-0.496	0.034	0.034	0.021
$\rho = \mathbf{-0.8}$	-0.996	-0.927	0.737	0.763	0.547	-0.867	-0.866	0.262	0.270	0.170	-0.833	-0.823	0.159	0.162	0.121
$\beta_0 = 0.0$	-0.009	-0.005	0.300	0.300	0.127	-0.001	0.001	0.061	0.061	0.039	0.003	0.001	0.036	0.037	0.022
$\beta_1 = 1.0$	1.056	1.042	0.286	0.292	0.190	1.010	0.999	0.094	0.094	0.062	1.006	1.003	0.054	0.054	0.033
$\beta_2 = -0.5$	-0.528	-0.519	0.163	0.165	0.104	-0.503	-0.500	0.051	0.051	0.032	-0.499	-0.498	0.033	0.033	0.022
$\rho = \mathbf{-0.6}$	-0.785	-0.664	0.747	0.769	0.537	-0.644	-0.640	0.258	0.262	0.169	-0.624	-0.619	0.156	0.158	0.102
$\beta_0 = 0.0$	-0.008	-0.008	0.279	0.279	0.118	0.001	-0.001	0.056	0.056	0.037	0.003	0.001	0.034	0.034	0.022
$\beta_1 = 1.0$	1.055	1.046	0.279	0.284	0.181	1.008	1.000	0.091	0.091	0.060	1.005	1.003	0.054	0.054	0.033
$\beta_2 = -0.5$	-0.528	-0.522	0.161	0.163	0.099	-0.503	-0.501	0.052	0.052	0.034	-0.498	-0.500	0.033	0.033	0.024
$\rho = \mathbf{-0.4}$	-0.601	-0.447	0.733	0.760	0.479	-0.437	-0.421	0.242	0.245	0.149	-0.416	-0.403	0.147	0.148	0.093
$\beta_0 = 0.0$	-0.004	-0.007	0.259	0.259	0.104	0.001	-0.000	0.052	0.052	0.032	0.003	0.001	0.031	0.031	0.018
$\beta_1 = 1.0$	1.056	1.058	0.283	0.288	0.185	1.007	0.996	0.090	0.090	0.059	1.006	1.002	0.054	0.054	0.033
$\beta_2 = -0.5$	-0.527	-0.521	0.162	0.164	0.100	-0.503	-0.501	0.051	0.051	0.033	-0.498	-0.498	0.034	0.034	0.023
$\rho = \mathbf{-0.2}$	-0.397	-0.223	0.688	0.716	0.426	-0.232	-0.214	0.223	0.225	0.140	-0.215	-0.209	0.132	0.133	0.093
$\beta_0 = 0.0$	-0.001	-0.007	0.243	0.243	0.097	0.002	-0.000	0.047	0.048	0.029	0.002	-0.002	0.028	0.028	0.017
$\beta_1 = 1.0$	1.059	1.053	0.287	0.293	0.190	1.007	1.001	0.100	0.100	0.058	1.005	1.002	0.053	0.053	0.035
$\beta_2 = -0.5$	-0.529	-0.524	0.163	0.165	0.103	-0.501	-0.501	0.057	0.057	0.033	-0.497	-0.496	0.032	0.032	0.022
$\rho = \mathbf{0.0}$	-0.209	-0.026	0.659	0.691	0.354	-0.030	-0.003	0.200	0.202	0.132	-0.012	-0.008	0.112	0.113	0.081
$\beta_0 = 0.0$	0.002	-0.007	0.220	0.220	0.093	0.002	0.001	0.042	0.042	0.028	0.003	0.000	0.025	0.025	0.016
$\beta_1 = 1.0$	1.061	1.050	0.289	0.296	0.183	1.008	1.002	0.088	0.089	0.058	1.004	1.000	0.054	0.054	0.033
$\beta_2 = -0.5$	-0.536	-0.524	0.165	0.169	0.104	-0.501	-0.497	0.053	0.053	0.032	-0.498	-0.498	0.032	0.032	0.020
$\rho = \mathbf{0.2}$	0.020	0.178	0.574	0.601	0.280	0.175	0.188	0.165	0.167	0.111	0.190	0.200	0.100	0.100	0.072
$\beta_0 = 0.0$	0.001	-0.005	0.236	0.236	0.090	0.001	-0.001	0.040	0.040	0.025	0.003	0.002	0.023	0.024	0.016
$\beta_1 = 1.0$	1.085	1.069	0.298	0.310	0.199	1.009	1.010	0.089	0.090	0.055	1.005	1.000	0.057	0.057	0.036
$\beta_2 = -0.5$	-0.544	-0.535	0.172	0.177	0.109	-0.500	-0.502	0.054	0.054	0.038	-0.498	-0.498	0.031	0.031	0.020
$\rho = \mathbf{0.4}$	0.217	0.376	0.537	0.568	0.213	0.378	0.396	0.131	0.133	0.088	0.392	0.400	0.080	0.080	0.056
$\beta_0 = 0.0$	0.004	-0.009	0.236	0.236	0.082	0.001	0.000	0.036	0.036	0.025	0.002	0.001	0.022	0.022	0.014
$\beta_1 = 1.0$	1.116	1.097	0.329	0.349	0.220	1.009	1.010	0.098	0.098	0.068	1.007	1.005	0.059	0.060	0.041
$\beta_2 = -0.5$	-0.557	-0.543	0.194	0.202	0.123	-0.503	-0.501	0.059	0.059	0.041	-0.498	-0.500	0.032	0.032	0.023
$\rho = \mathbf{0.6}$	0.444	0.572	0.444	0.470	0.150	0.574	0.580	0.095	0.098	0.060	0.586	0.591	0.061	0.063	0.040
$\beta_0 = 0.0$	0.004	-0.013	0.226	0.226	0.076	0.001	0.001	0.034	0.034	0.023	0.001	0.001	0.020	0.020	0.012
$\beta_1 = 1.0$	1.198	1.161	0.428	0.472	0.279	1.013	1.007	0.110	0.111	0.073	1.011	1.004	0.075	0.076	0.048
$\beta_2 = -0.5$	-0.610	-0.576	0.397	0.412	0.149	-0.508	-0.502	0.070	0.071	0.049	-0.498	-0.500	0.039	0.039	0.026
$\rho = \mathbf{0.8}$	0.659	0.743	0.306	0.337	0.095	0.738	0.747	0.065	0.090	0.040	0.748	0.750	0.041	0.066	0.025

Table A.1: Summary statistics of the PML estimates for the SAR(1) probit coefficients, considering different n sample sizes for the simulated spatial series of observations on regular grids. The weighting matrix $\mathbf{W}_n$ is a row–normalized k–nn matrix with $k = 11$. The number of Monte Carlo replications are fixed to 1,000. The rows *sd*, RMSE and MAD report the empirical standard deviations, empirical root mean square errors of the estimated coefficients from the true values, and empirical median absolute deviations, respectively.

	$\rho=-0.8$		$\rho=-0.6$		$\rho=-0.4$		$\rho=-0.2$		$\rho=0.2$		$\rho=0.4$		$\rho=0.6$		$\rho=0.8$	
Regressors	$m(\rho)$	$m(\hat{\rho})$	$m(\rho)$	$m(\hat{\rho})$	$m(\rho)$	$m(\hat{\rho})$	$m(\rho)$	$m(\hat{\rho})$	$m(\rho)$	$m(\hat{\rho})$	$m(\rho)$	$m(\hat{\rho})$	$m(\rho)$	$m(\hat{\rho})$	$m(\rho)$	$m(\hat{\rho})$
$\bar{\mathbf{X}}$, $\mathbf{x}_{\cdot 1}$																
Direct																
Mean	0.392	0.394	0.395	0.397	0.397	0.400	0.398	0.401	0.398	0.400	0.394	0.396	0.384	0.387	0.351	0.367
sd		0.035		0.035		0.035		0.035		0.035		0.035		0.036		0.038
Indirect																
Mean	-0.184	-0.189	-0.154	-0.155	-0.117	-0.116	-0.067	-0.066	0.098	0.103	0.251	0.258	0.530	0.523	1.207	0.980
sd		0.041		0.047		0.054		0.063		0.095		0.125		0.185		0.304
Total																
Mean	0.208	0.205	0.240	0.242	0.280	0.283	0.331	0.335	0.496	0.504	0.646	0.654	0.914	0.911	1.558	1.347
sd		0.039		0.047		0.055		0.066		0.101		0.133		0.196		0.318
$\bar{\mathbf{X}}$, $\mathbf{x}_{\cdot 2}$																
Direct																
Mean	-0.196	-0.197	-0.197	-0.198	-0.198	-0.199	-0.199	-0.200	-0.199	-0.200	-0.197	-0.197	-0.192	-0.194	-0.176	-0.184
sd		0.020		0.020		0.020		0.020		0.021		0.021		0.021		0.024
Indirect																
Mean	0.092	0.094	0.077	0.077	0.058	0.058	0.034	0.033	-0.049	-0.052	-0.126	-0.128	-0.265	-0.262	-0.603	-0.492
sd		0.021		0.024		0.027		0.032		0.048		0.063		0.095		0.155
Total																
Mean	-0.104	-0.103	-0.120	-0.121	-0.140	-0.142	-0.165	-0.167	-0.248	-0.251	-0.323	-0.326	-0.457	-0.456	-0.779	-0.676
sd		0.021		0.025		0.030		0.035		0.054		0.070		0.103		0.166
$\mathbf{X}$, $\mathbf{x}_{\cdot 1}$																
Direct																
Mean	0.311	0.311	0.313	0.313	0.315	0.315	0.315	0.316	0.315	0.315	0.311	0.312	0.303	0.304	0.277	0.287
sd		0.021		0.021		0.021		0.021		0.020		0.021		0.022		0.023
Indirect																
Mean	-0.146	-0.149	-0.122	-0.122	-0.093	-0.091	-0.053	-0.052	0.077	0.081	0.198	0.202	0.419	0.410	0.953	0.765
sd		0.031		0.036		0.042		0.049		0.074		0.097		0.140		0.222
Total																
Mean	0.165	0.162	0.191	0.191	0.222	0.223	0.262	0.264	0.392	0.396	0.510	0.514	0.722	0.714	1.231	1.052
sd		0.030		0.036		0.042		0.051		0.076		0.100		0.143		0.225
$\mathbf{X}$, $\mathbf{x}_{\cdot 2}$																
Direct																
Mean	-0.156	-0.155	-0.157	-0.157	-0.157	-0.157	-0.158	-0.158	-0.157	-0.157	-0.156	-0.155	-0.152	-0.152	-0.139	-0.144
sd		0.014		0.013		0.013		0.013		0.014		0.014		0.014		0.016
Indirect																
Mean	0.073	0.074	0.061	0.061	0.046	0.045	0.027	0.026	-0.039	-0.041	-0.099	-0.101	-0.209	-0.205	-0.477	-0.384
sd		0.016		0.018		0.021		0.025		0.037		0.049		0.072		0.113
Total																
Mean	-0.083	-0.081	-0.095	-0.096	-0.111	-0.112	-0.131	-0.132	-0.196	-0.198	-0.255	-0.256	-0.361	-0.357	-0.615	-0.528
sd		0.016		0.020		0.023		0.027		0.041		0.053		0.076		0.118

Table A.2: Marginal effects summary statistics for different estimated coefficients $\hat{\rho}$. $\bar{\mathbf{X}}$ and $\mathbf{X}$ are referred to the first and second specifications of the marginal impacts in equation (20), respectively. The total impacts are split into the direct and indirect effects and compared with the true ones $m(\rho)$. The simulated spatial series are referred to Table A.1 with $n = 900$, $\mathbf{W}_n = \mathbf{W}_{k-nn}$, and the regressors are $\mathbf{x}_{\cdot 1} \sim \mathcal{U}(-1, 1)$, $\mathbf{x}_{\cdot 2} \sim \mathcal{N}(0, 1)$.

True Value	Mean	Median	sd	$\widehat{sd}_a$	sd_a	RMSE	MAD
$\beta_0 = 0.0$	0.000	-0.001	0.049	0.041	0.045	0.049	0.030
$\beta_1 = 1.0$	1.016	1.013	0.104	0.105	0.107	0.105	0.069
$\beta_2 = -0.5$	-0.509	-0.508	0.066	0.063	0.064	0.067	0.044
$\rho = \mathbf{0.4}$	0.360	0.396	0.178	0.102	0.114	0.183	0.105
$\beta_0 = 0.0$	0.002	0.002	0.041	0.040	0.054	0.041	0.026
$\beta_1 = 1.0$	1.028	1.023	0.145	0.149	0.169	0.148	0.090
$\beta_2 = -0.5$	-0.516	-0.515	0.090	0.087	0.099	0.092	0.060
$\rho = \mathbf{0.8}$	0.730	0.736	0.059	0.046	0.068	0.092	0.037

Table A.3: Summary statistics of the PML estimates for the SAR(1) probit coefficients with the data set Katrina (first horizon) in section 8 and $n = 658$. The weighting matrix $\mathbf{W}_n$ is a row–normalized k–nn matrix with $k = 11$. The number of Monte Carlo replications are fixed to 1,000. The rows sd, $\widehat{sd}_a$, sd_a, RMSE and MAD report the empirical standard deviations, estimated asymptotic standard deviations, true asymptotic standard deviations, empirical root mean square errors of the estimated coefficients from the true values, and empirical median absolute deviations, respectively.

True Matrix/Value	$\beta_0 = 0$	$\beta_1 = 1$	$\beta_2 = -0.5$	$\rho = 0.6$	$\beta_0 = 0$	$\beta_1 = 1$	$\beta_2 = -0.5$	$\rho = -0.6$
$\mathbf{W}_{sn}$								
Mean	0.007	1.177	-0.502	0.014	0.002	0.978	-0.499	-0.104
Median	0.003	1.175	-0.501	0.039	-0.002	0.902	-0.496	-0.058
sd	0.065	0.339	0.057	0.279	0.045	0.299	0.054	0.306
RMSE	0.066	0.382	0.057	0.649	0.045	0.300	0.054	0.583
MAD	0.038	0.152	0.034	0.165	0.024	0.191	0.033	0.216
$\mathbf{W}_{rn}$								
Mean	0.011	1.205	-0.505	0.071	0.002	0.936	-0.498	-0.116
Median	0.005	1.117	-0.503	0.136	-0.001	0.879	-0.494	-0.077
sd	0.086	0.439	0.053	0.324	0.043	0.288	0.055	0.311
RMSE	0.087	0.485	0.053	0.620	0.043	0.295	0.055	0.575
MAD	0.047	0.256	0.034	0.198	0.023	0.197	0.035	0.214

Table A.4: Summary statistics of the PML estimates for the SAR(1) probit coefficients when $\mathbf{W}_n$ is misspecified. The weighting matrix used to estimate the model is $\mathbf{W}_n = \mathbf{W}_{k-nn}$ with $k = 11$. The sample size is fixed to $n = 900$ and $\rho = (-0.6, 0.6)$.

	$\mathbf{W}_{sn}$								$\mathbf{W}_{rn}$							
$\rho = 0.6$	$\bar{\mathbf{X}}$				$\mathbf{X}$				$\bar{\mathbf{X}}$				$\mathbf{X}$			
Regressors	$m(\rho)$	$m(\hat{\rho})$	Lower	Upper	$m(\rho)$	$m(\hat{\rho})$	Lower	Upper	$m(\rho)$	$m(\hat{\rho})$	Lower	Upper	$m(\rho)$	$m(\hat{\rho})$	Lower	Upper
$\mathbf{x}_{.1}$																
Direct																
Mean	0.399	0.467	0.242	0.731	0.301	0.350	0.184	0.557	0.398	0.477	0.240	0.864	0.294	0.349	0.177	0.635
sd		0.133				0.099				0.173				0.125		
Indirect																
Mean	0.329	0.009	-0.264	0.217	0.252	0.005	-0.207	0.165	0.595	0.032	-0.349	0.267	0.439	0.023	-0.260	0.198
sd		0.135				0.101				0.168				0.124		
Total																
Mean	0.728	0.476	0.351	0.616	0.553	0.355	0.284	0.419	0.993	0.508	0.423	0.591	0.732	0.373	0.333	0.410
sd		0.067				0.035				0.043				0.020		
$\mathbf{x}_{.2}$																
Direct																
Mean	-0.199	-0.199	-0.240	-0.160	-0.151	-0.149	-0.176	-0.120	-0.199	-0.199	-0.239	-0.160	-0.147	-0.146	-0.170	-0.119
sd		0.023				0.016				0.021				0.013		
Indirect																
Mean	-0.165	-0.018	-0.165	0.076	-0.126	-0.013	-0.125	0.057	-0.297	-0.038	-0.219	0.084	-0.219	-0.028	-0.163	0.063
sd		0.063				0.048				0.078				0.057		
Total																
Mean	-0.364	-0.217	-0.372	-0.108	-0.276	-0.162	-0.276	-0.086	-0.496	-0.237	-0.427	-0.107	-0.366	-0.174	-0.311	-0.081
sd		0.067				0.049				0.080				0.058		
$\rho = -0.6$																
$\mathbf{x}_{.1}$																
Direct																
Mean	0.399	0.389	0.215	0.651	0.325	0.315	0.169	0.530	0.399	0.467	0.242	0.731	0.301	0.350	0.184	0.557
sd		0.118				0.093				0.133				0.099		
Indirect																
Mean	-0.123	-0.040	-0.300	0.129	-0.101	-0.032	-0.250	0.106	0.329	0.009	-0.264	0.217	0.252	0.005	-0.207	0.165
sd		0.111				0.090				0.135				0.101		
Total																
Mean	0.276	0.349	0.287	0.421	0.224	0.283	0.240	0.323	0.728	0.476	0.351	0.616	0.553	0.355	0.284	0.419
sd		0.034				0.022				0.067				0.035		
$\mathbf{x}_{.2}$																
Direct																
Mean	-0.199	-0.198	-0.242	-0.158	-0.162	-0.161	-0.190	-0.134	-0.199	-0.199	-0.240	-0.160	-0.151	-0.149	-0.176	-0.120
sd		0.022				0.014				0.023				0.016		
Indirect																
Mean	0.061	0.006	-0.136	0.088	0.050	0.005	-0.111	0.070	-0.165	-0.018	-0.165	0.076	-0.126	-0.013	-0.125	0.057
sd		0.055				0.045				0.063				0.048		
Total																
Mean	-0.138	-0.192	-0.347	-0.101	-0.112	-0.156	-0.278	-0.083	-0.364	-0.217	-0.372	-0.108	-0.276	-0.162	-0.276	-0.086
sd		0.058				0.046				0.067				0.049		

Table A.5: Marginal effects when $\mathbf{W}_n$ is misspecified. The Table reports results related to two *true* weighting matrices: (i) based on inverse distance with spectral normalisation $\mathbf{W}_{sn}$, (ii) based on inverse distance with row normalisation $\mathbf{W}_{rn}$. The total impacts are split into the direct and indirect effects and compared with the true ones $m(\rho)$. The simulated spatial series are referred to Table A.1 with $n = 900$, a k–nn weighting matrix $\mathbf{W}_{k-nn}$, $\rho = (-0.6, 0.6)$, and the regressors are $\mathbf{x}_{.1} \sim \mathcal{U}(-1, 1)$, $\mathbf{x}_{.2} \sim \mathcal{N}(0, 1)$.

$n = 900$	Default pairs				quasi-max-matching pairs				max-matching pairs			
$\mathbf{W}_n = \mathbf{W}_{sn}$	β_0	β_1	β_2	ρ	β_0	β_1	β_2	ρ	β_0	β_1	β_2	ρ
Mean	0.004	1.091	–0.497	0.433	0.004	1.088	-0.499	0.397	0.004	1.088	-0.498	0.401
Median	-0.003	1.044	-0.496	0.591	-0.003	1.071	-0.497	0.505	-0.003	1.065	-0.497	0.504
sd	0.076	0.181	0.054	0.528	0.061	0.140	0.054	0.339	0.059	0.139	0.054	0.338
RMSE	0.076	0.203	0.054	0.554	0.061	0.165	0.054	0.395	0.059	0.164	0.054	0.392
$\mathbf{W}_n = \mathbf{W}_{rn}$	β_0	β_1	β_2	ρ	β_0	β_1	β_2	ρ	β_0	β_1	β_2	ρ
Mean	0.006	1.129	–0.500	0.483	0.005	1.119	-0.499	0.468	0.005	1.119	-0.499	0.469
Median	-0.003	1.051	-0.498	0.711	-0.002	1.082	-0.499	0.561	-0.002	1.082	-0.499	0.568
sd	0.083	0.241	0.055	0.558	0.073	0.196	0.054	0.459	0.074	0.198	0.054	0.463
RMSE	0.083	0.273	0.055	0.570	0.074	0.229	0.054	0.478	0.074	0.231	0.054	0.481

Table A.6: Summary statistics of the PML estimates for the SAR(1) probit coefficients using alternative choices of pairs. The first columns correspond to the default choice, i.e. $g \equiv (2g-1, 2g)$; the other two sets of estimates refer to the algorithm proposed in Section 4, with different initial guess of the parameter ρ (namely, $\tilde{\rho} = 0.2$, in the *quasi–max–matching*, $\tilde{\rho} = 0.6$ in the *max-matching* case). Here, $\boldsymbol{\theta}_0 = (0, 1, -0.5, 0.6)$ and the two panels refer to $\mathbf{W}_n = \mathbf{W}_{sn}$ (inverse distance matrix with spectral normalization) and $\mathbf{W}_n = \mathbf{W}_{rn}$ respectively (inverse distance matrix with row normalization).

True Value	Mean	Median	*sd*	RMSE	MAD	True Value	Mean	Median	*sd*	RMSE	MAD
$\beta_0 = 0.0$	0.006	-0.001	0.178	0.178	0.145	$\beta_0 = 0.0$	0.010	-0.002	0.115	0.116	0.058
$\beta_1 = 1.0$	0.983	0.992	0.198	0.199	0.189	$\beta_1 = 1.0$	1.008	1.001	0.144	0.145	0.086
$\beta_2 = -0.5$	-0.488	-0.484	0.106	0.106	0.108	$\beta_2 = -0.5$	-0.498	-0.484	0.085	0.085	0.054
$\rho = 0.6$	0.527	0.611	0.323	0.332	0.547	$\rho = 0.6$	0.542	0.583	0.255	0.261	0.153
$\lambda = \mathbf{0.8}$	0.658	0.717	0.246	0.284	0.547	$\lambda = \mathbf{0.6}$	0.531	0.561	0.220	0.231	0.137
True Value	Mean	Median	*sd*	RMSE	MAD	True Value	Mean	Median	*sd*	RMSE	MAD
$\beta_0 = 0.0$	0.005	0.002	0.071	0.071	0.039	$\beta_0 = 0.0$	0.003	-0.000	0.052	0.052	0.034
$\beta_1 = 1.0$	1.014	1.010	0.123	0.123	0.082	$\beta_1 = 1.0$	1.019	1.008	0.112	0.113	0.067
$\beta_2 = -0.5$	-0.501	-0.489	0.074	0.074	0.049	$\beta_2 = -0.5$	-0.501	-0.497	0.062	0.062	0.042
$\rho = 0.6$	0.557	0.589	0.192	0.197	0.109	$\rho = 0.6$	0.564	0.592	0.150	0.155	0.089
$\lambda = \mathbf{0.4}$	0.355	0.376	0.224	0.229	0.156	$\lambda = \mathbf{0.2}$	0.165	0.162	0.233	0.236	0.151

Table A.7: Summary statistics of the PML estimates for the SARAR(1,1) probit coefficients from simulated spatial series of observations on regular grids. The weighting matrix $\mathbf{W}_n$ is a row–normalized k–nn matrix with $k = 11$, while $\mathbf{M}_n$ is a row–normalized weighting matrix based on the Queen contiguity criterion. The number of Monte Carlo replications are fixed to 200. The rows *sd*, RMSE and MAD report the empirical standard deviations, empirical root mean square errors of the estimated coefficients from the true values, and empirical median absolute deviations, respectively.

First Horizon	SAR(1)–probit								SARAR(1)–probit	
Regressors	Bayes	*sd*	AMLE	*sd*	CMLE	*sd*	PMLE	*sd*	PMLE	*sd*
constant	-7.616	2.595	-7.111	1.868	-3.069	–	-5.272	2.435	-2.539	1.033
flood depth	-0.168	0.044	-0.185	0.030	-0.071	–	-0.136	0.048	-0.062	0.026
log(median income)	0.733	0.252	0.691	0.181	0.281	–	0.510	0.238	0.243	0.099
small size	-0.276	0.140	-0.318	0.138	-0.323	–	-0.340	0.147	-0.428	0.144
large size	-0.329	0.321	-0.321	0.312	0.024	–	-0.361	0.328	-0.472	0.334
low status customers	-0.329	0.166	-0.486	0.147	-0.351	–	-0.453	0.154	-0.301	0.127
high status customers	0.085	0.131	0.057	0.127	-0.003	–	0.034	0.125	0.041	0.116
sole proprietorship	0.551	0.196	0.562	0.194	0.610	–	0.560	0.202	0.528	0.174
national chain	0.068	0.378	0.085	0.356	0.116	–	0.059	0.385	0.058	0.415
Wy	0.382	0.094	0.346	–	0.783	–	0.515	0.143	0.802	0.074
Mu	–	–	–	–	–	–	–	–	-0.579	0.130
Second Horizon	SAR(1)–probit								SARAR(1)–probit	
Regressors	Bayes	*sd*	AMLE	*sd*	CMLE	*sd*	PMLE	*sd*	PMLE	*sd*
constant	-2.978	2.730	-3.604	1.798	-0.927	–	-2.069	1.886	-0.898	1.031
flood depth	-0.110	0.035	-0.159	0.021	-0.080	–	-0.112	0.038	-0.060	0.031
log(median income)	0.311	0.268	0.402	0.176	0.125	–	0.238	0.190	0.114	0.104
small size	-0.109	0.149	-0.174	0.140	-0.222	–	-0.223	0.142	-0.283	0.133
large size	-0.372	0.332	-0.456	0.291	-0.442	–	-0.442	0.321	-0.356	0.265
low status customers	-0.342	0.161	-0.524	0.133	-0.382	–	-0.446	0.133	-0.271	0.120
high status customers	0.041	0.153	0.026	0.141	-0.029	–	-0.006	0.136	-0.025	0.117
sole proprietorship	0.359	0.181	0.264	0.175	0.188	–	0.289	0.180	0.230	0.158
national chain	0.295	0.381	-0.027	0.347	-0.486	–	-0.099	0.365	-0.481	0.363
Wy	0.578	0.084	0.433	–	0.768	–	0.621	0.129	0.833	0.092
Mu	–	–	–	–	–	–	–	–	-0.445	0.186
Third Horizon	SAR(1)–probit								SARAR(1)–probit	
Regressors	Bayes	*sd*	AMLE	*sd*	CMLE	*sd*	PMLE	*sd*	PMLE	*sd*
constant	-4.336	2.723	-3.240	1.658	0.348	–	-2.198	2.262	-0.898	1.122
flood depth	-0.089	0.034	-0.126	0.017	-0.040	–	-0.102	0.034	-0.060	0.024
log(median income)	0.484	0.268	0.403	0.163	0.044	–	0.287	0.233	0.114	0.115
small size	-0.214	0.154	-0.205	0.142	0.037	–	-0.240	0.148	-0.283	0.130
large size	-0.357	0.298	-0.439	0.288	-1.157	–	-0.424	0.311	-0.356	0.309
low status customers	-0.321	0.162	-0.586	0.125	-0.602	–	-0.512	0.141	-0.271	0.135
high status customers	-0.101	0.165	-0.233	0.143	-0.641	–	-0.241	0.146	-0.025	0.130
sole proprietorship	0.146	0.189	0.039	0.170	-0.257	–	0.078	0.169	0.230	0.152
national chain	-0.120	0.389	-0.532	0.347	-2.365	–	-0.621	0.401	-0.481	0.376
Wy	0.584	0.093	0.554	–	0.963	–	0.664	0.127	0.833	0.080
Mu	–	–	–	–	–	–	–	–	-0.445	0.197

Table A.8: Alternative ML–based and Bayesian estimates and empirical standard deviations for the first, second and third time horizons of the data set Katrina. A SAR(1)–probit model is assumed. The column PMLE refer to our estimator, Bayes refers to LeSage's Bayesian estimates, the column AMLE refers to the approximate MLE by Martinetti and Geniaux (2017) and the column CMLE refers to the composite (univariate) MLE by Mozharovskyi and Vogler (2016). The last column shows the estimates and standard deviations of our PMLE for a SARAR(1)–probit specification. Our *sd* are based on the "probit" bootstrap approach.

	PMLE			Bayes		
Impacts	First	Second	Third	First	Second	Third
Direct						
flood depth	-0.038	-0.027	-0.022	-0.048	-0.028	-0.020
log(median income)	0.141	0.058	0.062	0.212	0.078	0.111
small size	-0.094	-0.054	-0.052	-0.080	-0.028	-0.050
large size	-0.100	-0.107	-0.092	-0.095	-0.094	-0.082
low status customers	-0.126	-0.108	-0.111	-0.095	-0.086	-0.074
high status customers	0.009	-0.002	-0.052	0.025	0.010	-0.023
sole proprietorship	0.155	0.070	0.017	0.160	0.091	0.033
national chain	0.016	-0.024	-0.134	0.020	0.074	-0.029
Indirect						
flood depth	-0.037	-0.041	-0.040	-0.030	-0.034	-0.027
log(median income)	0.140	0.088	0.113	0.128	0.097	0.154
small size	-0.093	-0.082	-0.094	-0.050	-0.035	-0.072
large size	-0.099	-0.163	-0.167	-0.061	-0.121	-0.116
low status customers	-0.125	-0.164	-0.202	-0.058	-0.110	-0.102
high status customers	0.009	-0.002	-0.095	0.015	0.012	-0.034
sole proprietorship	0.154	0.107	0.031	0.099	0.118	0.050
national chain	0.016	-0.036	-0.244	0.012	0.100	-0.037
Total						
flood depth	-0.075	-0.068	-0.062	-0.078	-0.062	-0.048
log(median income)	0.282	0.146	0.175	0.340	0.174	0.265
small size	-0.188	-0.136	-0.146	-0.130	-0.063	-0.122
large size	-0.200	-0.270	-0.259	-0.156	-0.251	-0.199
low status customers	-0.250	-0.272	-0.313	-0.153	-0.195	-0.176
high status customers	0.019	-0.004	-0.147	0.040	0.023	-0.057
sole proprietorship	0.309	0.176	0.048	0.259	0.209	0.083
national chain	0.033	-0.060	-0.378	0.032	0.174	-0.067

Table A.9: Marginal effects with respect to $\mathbf{X}$ (second specification of equation (20)) for the first, second and third time horizons of the data set Katrina. The column Bayes refers to LeSage's Bayesian estimates, while the column PMLE refers to our PML estimates.

	First					Second					Third				
Regressors	Estimates	sd_{b_1}	sd_{b_2}	sd_a	sd_c	Estimates	sd_{b_1}	sd_{b_2}	sd_a	sd_c	Estimates	sd_{b_1}	sd_{b_2}	sd_a	sd_c
constant	-5.272	2.435	3.246	4.672	4.299	-2.069	1.886	3.762	4.838	3.359	-2.198	2.262	3.523	5.461	3.830
flood depth	-0.136	0.048	0.062	0.059	0.045	-0.112	0.038	0.095	0.050	0.027	-0.102	0.034	0.097	0.038	0.018
log(median income)	0.510	0.238	0.319	0.456	0.423	0.238	0.190	0.368	0.486	0.342	0.287	0.233	0.345	0.556	0.395
small size	-0.340	0.147	0.163	0.158	0.147	-0.223	0.142	0.179	0.178	0.173	-0.240	0.148	0.192	0.187	0.143
large size	-0.361	0.328	0.368	0.380	0.379	-0.442	0.321	0.433	0.398	0.293	-0.424	0.311	0.435	0.399	0.218
low status customers	-0.453	0.154	0.186	0.167	0.126	-0.446	0.133	0.198	0.154	0.131	-0.512	0.141	0.219	0.168	0.106
high status customers	0.034	0.125	0.149	0.148	0.155	-0.006	0.136	0.156	0.184	0.224	-0.241	0.146	0.175	0.208	0.181
sole proprietorship	0.560	0.202	0.236	0.217	0.168	0.289	0.180	0.261	0.217	0.164	0.078	0.169	0.298	0.246	0.173
national chain	0.059	0.385	0.412	0.408	0.383	-0.099	0.365	0.443	0.502	0.307	-0.621	0.401	0.498	0.530	0.279
Wy	0.515	0.143	0.158	0.212	0.504	0.621	0.129	0.146	0.236	0.509	0.664	0.127	0.130	0.223	0.508

Table A.10: Estimates and standard deviations for the first, second and third time horizons of the data set Katrina with the PMLE. A SAR(1) probit model is assumed. sd_{b_1} refers to the standard deviations from the first parametric bootstrap approach (”probit” bootstrap), sd_{b_2} refers to the standard deviations from the second parametric bootstrap approach (”latent” bootstrap), sd_a refers to the standard deviations from the estimation of the asymptotic variance–covariance matrix, and sd_c refers to the standard deviations from the Conley’s approach (Conley, 1999).

Appendix B. Figures

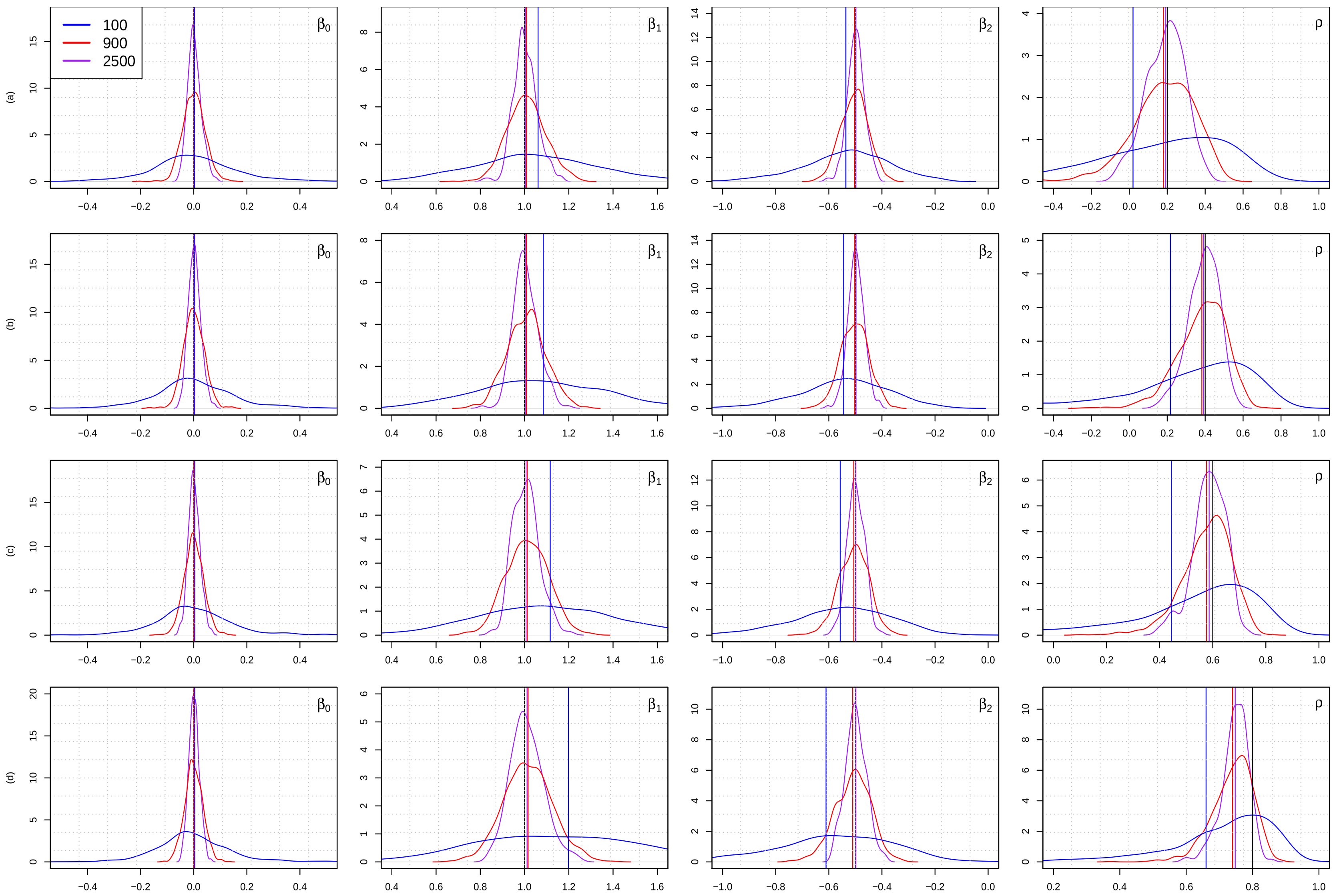

Figure B.1: Gaussian Kernel density for the PML estimated coefficients of the SAR(1) probit model for different true values of ρ: (a) $\rho = 0.2$, (b) $\rho = 0.4$, (c) $\rho = 0.6$ (d) $\rho = 0.8$. The sample sizes are 100 (in blue), 900 (in red) and 2,500 (in purple), while blue, red and purple vertical lines are the mean values, respectively. Vertical black lines are the true values of the parameters. The number of Monte Carlo replications is fixed to 1,000.

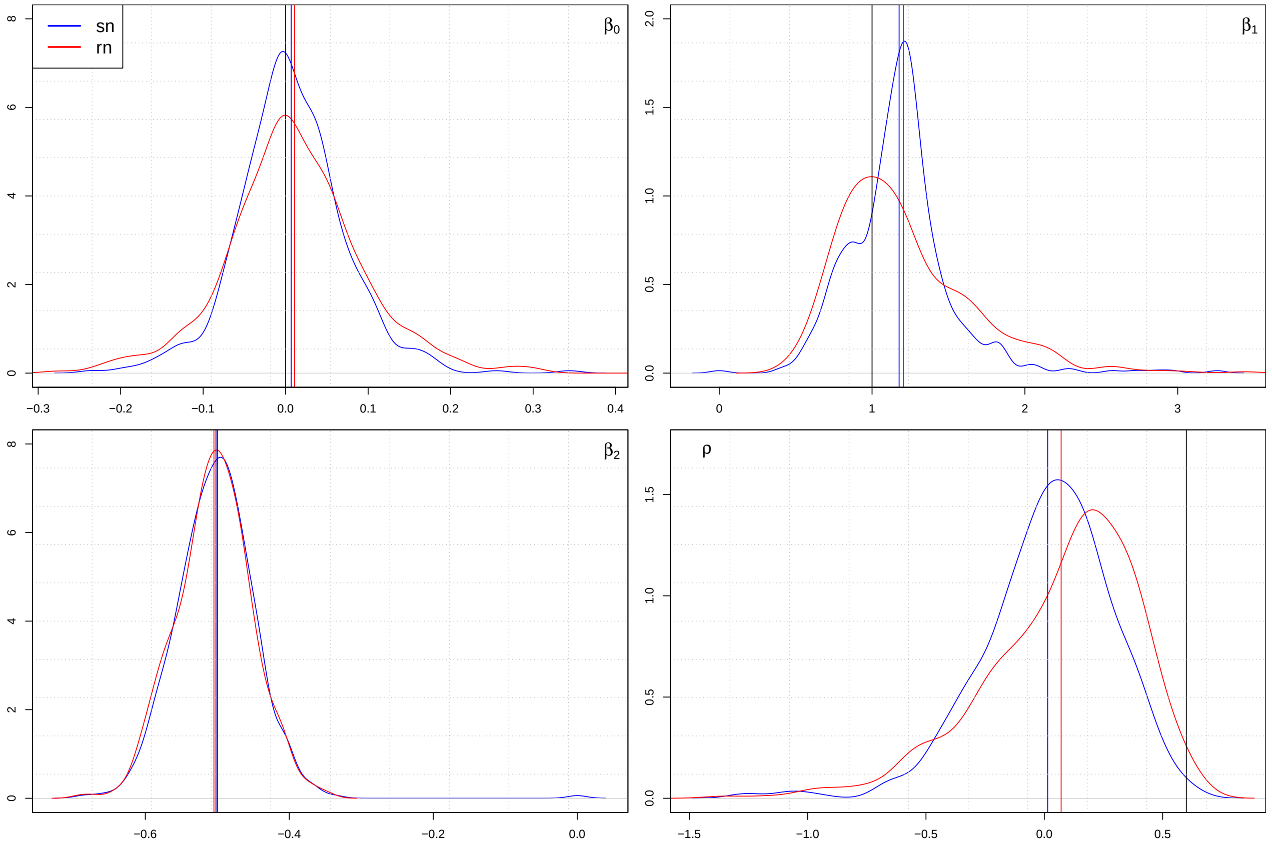

Figure B.2: Gaussian Kernel density for the PML estimated coefficients of the SAR(1) probit model when $\mathbf{W}_n$ is misspecified. Two cases of misspecification: (i) $\mathbf{W}_{true} = \mathbf{W}_{sn}$ (in blue), (ii) $\mathbf{W}_{true} = \mathbf{W}_{rn}$ (in red), where sn and rn refer to spectral–normalization and row–normalization, respectively. The assumed weighting matrix is based on a k–nearest neighbor approach $\mathbf{W}_{k-nn}$, with $k = 11$, whereas $n = 900$ and $\rho = 0.6$ are fixed. Red (blue) vertical and red dashed (blue dashed) vertical lines are the mean values, respectively. Vertical black lines are the true values of the parameters.

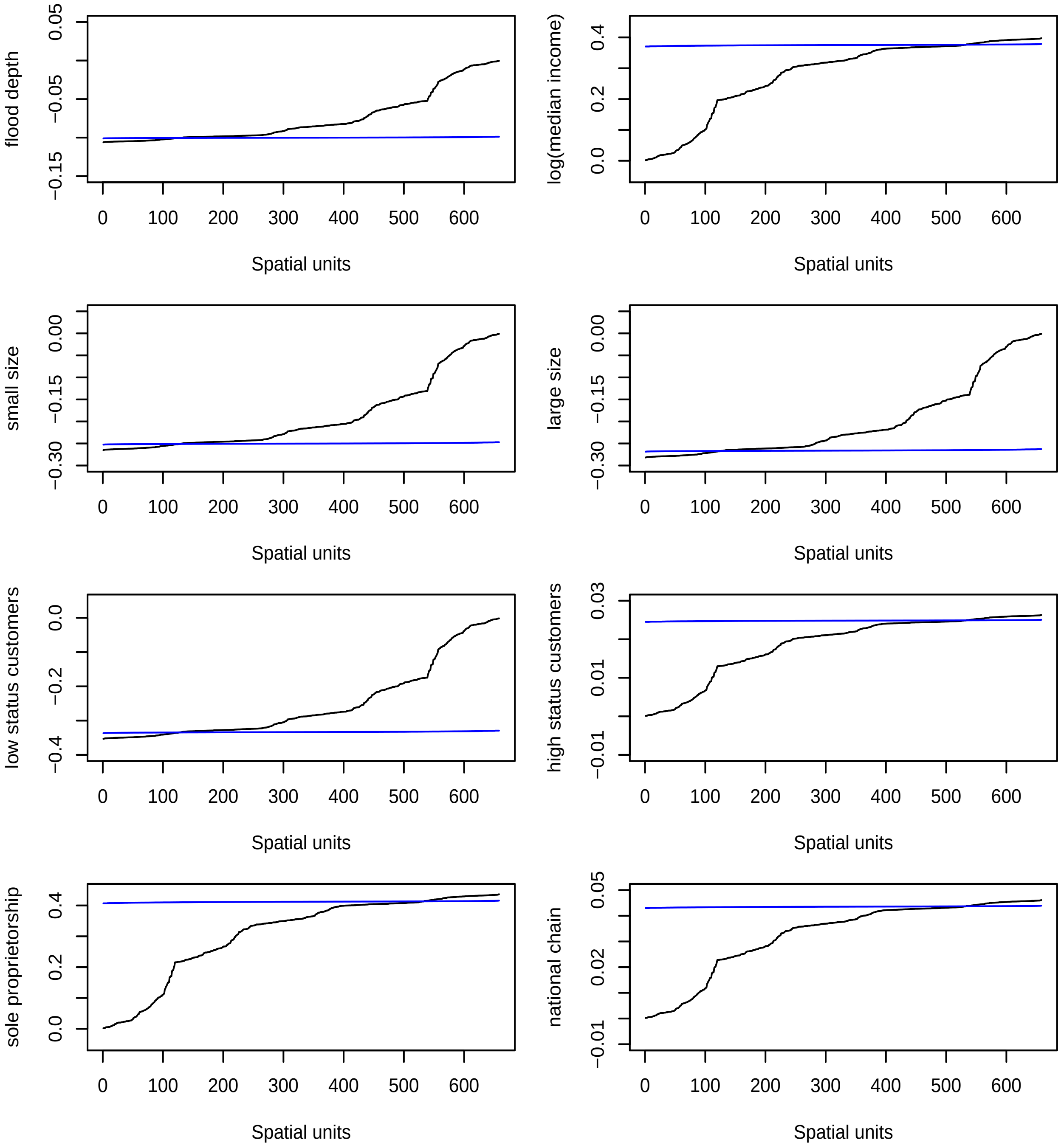

Figure B.3: Spatial heterogeneity of the total marginal impacts for each regressor during the first time horizon. Blue lines represent marginal impacts relative to the mean value.

Appendix C. Computational aspects

The computational optimization procedure is based on *unconstrained* minimization of the negative log–likelihood function with respect to the vector of parameters as in Catania and Billé (2017). So let $\mathbf{h}: \Re^{k+2} \to \Omega$ be a measurable vector valued mapping function such that $\mathbf{h} \in \mathcal{C}^2$ and $\mathbf{h}\left(\overset{\circ}{\boldsymbol{\theta}}\right) = \boldsymbol{\theta}$, where $\overset{\circ}{\boldsymbol{\theta}} = \left(\overset{\circ}{\boldsymbol{\beta}}{}', \overset{\circ}{\rho}, \overset{\circ}{\lambda}\right)'$ is the unconstrained vector of parameters defined in $\Re^{k+2}$. Given the necessary conditions on the parameter spaces for ρ and λ, we define the following mapping functions

$$\mathbf{h}\left(\overset{\circ}{\boldsymbol{\theta}}\right) : \begin{cases} \rho = \underline{\omega}_\rho^{-1} + \frac{\overline{\omega}_\rho^{-1} - \underline{\omega}_\rho^{-1}}{1+\exp\left(-\overset{\circ}{\rho}\right)}, \\ \lambda = \underline{\omega}_\lambda^{-1} + \frac{\overline{\omega}_\lambda^{-1} - \underline{\omega}_\lambda^{-1}}{1+\exp\left(-\overset{\circ}{\lambda}\right)}, \\ \boldsymbol{\beta} = \mathbf{h}_\beta\left(\overset{\circ}{\boldsymbol{\beta}}\right), \quad \text{for} \quad j = 1, \ldots, n \end{cases} \tag{C.1}$$

where $\left(\underline{\omega}_\rho, \overline{\omega}_\rho\right)$ and $\left(\underline{\omega}_\lambda, \overline{\omega}_\lambda\right)$ are the minimum and maximum eigenvalues of the weighting matrices $\mathbf{W}$ and $\mathbf{M}$, respectively. To obtain working parameters $\overset{\circ}{\boldsymbol{\theta}}$ from initial starting values of the natural parameters $\boldsymbol{\theta}$, inverse functions $\mathbf{h}^{-1}(\boldsymbol{\theta})$ are used. In the same way, let $\nabla(\boldsymbol{\theta}; \mathbf{y}, \mathbf{X})$ be the score vector of a specified log–likelihood function. By exploiting the chain rule we can define, $\overset{\circ}{\nabla}\left(\overset{\circ}{\boldsymbol{\theta}}; \mathbf{y}, \mathbf{X}\right) = \mathcal{J}\left(\overset{\circ}{\boldsymbol{\theta}}; \mathbf{y}, \mathbf{X}\right)' \nabla(\boldsymbol{\theta}; \mathbf{y}, \mathbf{X})$, where $\mathcal{J}\left(\overset{\circ}{\boldsymbol{\theta}}; \mathbf{y}, \mathbf{X}\right) = \left(\mathcal{J}\left(\overset{\circ}{\boldsymbol{\beta}}\right)', \mathcal{J}\left(\overset{\circ}{\rho}\right), \mathcal{J}\left(\overset{\circ}{\lambda}\right)\right)'$ is the Jacobian matrix with respect to the working/unconstrained parameters, and it is equal to

$$\mathcal{J}\left(\overset{\circ}{\boldsymbol{\theta}}\right) : \begin{cases} \mathcal{J}\left(\overset{\circ}{\rho}\right) = \frac{\left(\overline{\omega}_\rho^{-1} - \underline{\omega}_\rho^{-1}\right)\exp\left(-\overset{\circ}{\rho}\right)}{\left(1+\exp\left(-\overset{\circ}{\rho}\right)\right)^2}, \\ \mathcal{J}\left(\overset{\circ}{\lambda}\right) = \frac{\left(\overline{\omega}_\lambda^{-1} - \underline{\omega}_\lambda^{-1}\right)\exp\left(-\overset{\circ}{\lambda}\right)}{\left(1+\exp\left(-\overset{\circ}{\lambda}\right)\right)^2}, \\ \mathcal{J}\left(\overset{\circ}{\boldsymbol{\beta}}\right) = \mathcal{J}(\boldsymbol{\beta}), \quad \text{for} \quad j = 1, \ldots, n. \end{cases} \tag{C.2}$$

Appendix D. The problem of inconsistency

Maximum likelihood estimators are consistent if the density of $\mathbf{y}_n^*$ is correctly specified. Misspecification of the functional form in a probit context is equivalent to have misspecification each conditional probability of $y_i = 1$, $1 \leq i \leq n$.

In a SAE(1) probit setting, heteroskedasticity will arise whenever the weights $\mathbf{M}_n$ induce non–constant diagonal terms of the matrix $\boldsymbol{\Sigma}_\mathbf{u} = [\mathbf{B}_\lambda' \mathbf{B}_\lambda]^{-1}$. Indeed, this usually happens even for rather *simple* choices of $\mathbf{M}_n$, such as a k–nearest neighbor matrix. Heteroskedastic probit estimators (Case, 1992) that explicitly consider the diagonal elements of the variance–covariance matrix, i.e. $\operatorname{diag}(\boldsymbol{\Sigma}_\mathbf{u}) = \operatorname{diag}[\mathbf{B}_\lambda' \mathbf{B}_\lambda]^{-1}$, remain consistent. However, the form of heteroskedasticity is generally unknown if it is implied by the spatial autocorrelation coefficient, see McMillen (1995) and Pinkse and Slade (1998). Note that, if the true model includes spatial effects in the endogenous variables $\mathbf{y}_n^*$, the SAE(1) probit model still produces inconsistent estimates. For nonparametric estimation and general specifications of spatial error processes see Kelejian and Prucha (2007) and Kelejian (2016). See also Wooldridge (2014) for a Quasi–MLE in nonlinear models with endogenous regressors.

In order to briefly explain, let $\mathbf{A}_\rho = (\mathbf{I}_n - \rho \mathbf{W}_n)$ and $\mathbf{B}_\lambda = (\mathbf{I}_n - \lambda \mathbf{M}_n)$, as above, we get

$$\mathbf{y}_n^* = \lambda \mathbf{M}_n \mathbf{y}_n^* + \rho \mathbf{B}_\lambda \mathbf{W}_n \mathbf{y}_n^* + \mathbf{B}_\lambda \mathbf{X}_n \boldsymbol{\beta} + \boldsymbol{\varepsilon}_n, \quad \boldsymbol{\varepsilon}_n \sim \mathcal{N}_n\left(\mathbf{0}_n, \boldsymbol{\Sigma}_\varepsilon\right) \tag{D.1}$$

which is known as the Cochrane–Orcutt type transformation (Cochrane and Orcutt, 1949), a model in which the resulting disturbances are innovations. Even after the Cochrane–Orcutt transformation, both $\mathbf{W}_n \mathbf{y}_n^*$ and $\mathbf{M}_n \mathbf{y}_n^*$ are correlated with $\boldsymbol{\varepsilon}_n$ because

$$\mathbb{E}\left[\mathbf{y}_n^* \boldsymbol{\varepsilon}_n'\right] = \mathbf{A}_\rho^{-1} \mathbb{E}\left[\mathbf{u}_n \boldsymbol{\varepsilon}_n'\right] = \mathbf{A}_\rho^{-1} \mathbf{B}_\lambda^{-1} \tag{D.2}$$

and these correlations rule out the use of nonlinear least squares methods due to their inconsistency. Therefore, consistency can only be achieved by correctly specifying the conditional expected value of model in equation (1).

Appendix E. Proof of Theorems

Appendix E.1. Proof of Theorem 4.1

(i) For all $2n$tuple $\mathbf{d} = (d_1, \ldots, d_{2n})$, $d_i \in \{0, 1\}$, we denote by $E(\mathbf{d}) = \{\mathbf{y}^* = (y_1^*, \ldots, y_{2n}^*) : 2(d_j - 0.5) y_j^* < 0\}$. The 2^{2n} sets $E(\mathbf{d})$ for a partition of $\mathbb{R}^{2n}$, and we can thus write

$$\begin{aligned}
KL(f_\pi || f_\theta) &= \int_{\mathbb{R}^{2n}} f_\pi(\mathbf{y}^*) \log \frac{f_\pi(\mathbf{y}^*)}{f_\theta(\mathbf{y}^*)} d\mathbf{y}^* \\
&= \sum_{\mathbf{d}} P_\pi(\mathbf{d}) \int_{E(\mathbf{d})} \frac{f_\pi(\mathbf{y}^*)}{P_\pi(\mathbf{d})} \log \frac{f_\pi(\mathbf{y}^*)}{f_\theta(\mathbf{y}^*)} d\mathbf{y}^* \\
&= \sum_{\mathbf{d}} P_\pi(\mathbf{d}) \int_{E(\mathbf{d})} \frac{f_\pi(\mathbf{y}^*)}{P_\pi(\mathbf{d})} \log \frac{f_\pi(\mathbf{y}^*)/P_\pi(\mathbf{d})}{f_\theta(\mathbf{y}^*)/P_\theta(\mathbf{d})} d\mathbf{y}^* + \sum_{\mathbf{d}} P_\pi(\mathbf{d}) \log \frac{P_\pi(\mathbf{d})}{P_\theta(\mathbf{d})} \\
&= -\sum_{\mathbf{d}} P_\pi(\mathbf{d}) \int_{E(\mathbf{d})} \frac{f_\pi(\mathbf{y}^*)}{P_\pi(\mathbf{d})} \log \frac{f_\theta(\mathbf{y}^*)/P_\theta(\mathbf{d})}{f_\pi(\mathbf{y}^*)/P_\pi(\mathbf{d})} d\mathbf{y}^* + \sum_{\mathbf{d}} P_\pi(\mathbf{d}) \log \frac{P_\pi(\mathbf{d})}{P_\theta(\mathbf{d})} \\
&\geq -\sum_{\mathbf{d}} P_\pi(\mathbf{d}) \log \int_{E(\mathbf{d})} \frac{f_\pi(\mathbf{y}^*)}{P_\pi(\mathbf{d})} \frac{f_\theta(\mathbf{y}^*)/P_\theta(\mathbf{d})}{f_\pi(\mathbf{y}^*)/P_\pi(\mathbf{d})} d\mathbf{y}^* + KL(P_\pi || P_\theta) \\
&= KL(P_\pi || P_\theta)
\end{aligned}$$

where we used convexity of the map $f(x) = -\log x$ and Jensen's inequality.

(ii) For any given permutation $\pi \in \Pi_n$ and its permutation matrix $\mathbf{P}_\pi$, the densities f_0^π and f_0 are n–variate Gaussian random vectors,

$$\begin{aligned} f_0^\pi &\sim \mathcal{N}\left(\mathbf{P}_\pi\left(\mathbf{I}-\rho\mathbf{W}_\pi\right)^{-1}\mathbf{X}\beta, \mathbf{\Sigma}_\pi\right) \\ f_0 &\sim \mathcal{N}\left(\mathbf{P}_\pi\left(\mathbf{I}-\rho\mathbf{W}_\pi\right)^{-1}\mathbf{X}\beta, \mathbf{P}_\pi\mathbf{\Sigma}\mathbf{P}_\pi\right) \end{aligned}$$

where

$$\mathbf{\Sigma}_\pi = \sum_{g=1}^{G} \mathbf{E}_g\mathbf{P}_\pi\mathbf{\Sigma}\mathbf{P}'_\pi\mathbf{E}_g,$$

with $\mathbf{E}_g$ is the $n \times n$ matrix with all zero row vectors, except for rows $2g-1, 2g$, that are equal to $\mathbf{e}'_{2g-1}$ and $\mathbf{e}'_{2g}$ respectively.

From the formula of the KL–divergence of two multivariate Gaussian distributions with the same mean, and by using the properties $\mathrm{tr}(\mathbf{AB}) = \mathrm{tr}(\mathbf{BA})$, $|\mathbf{AB}| = |\mathbf{A}| \cdot |\mathbf{B}|$ and the fact that $\log|\mathbf{A}| = \mathrm{tr}\log(\mathbf{A})$, we have:

$$\begin{aligned} KL\left(f_0^\pi || f_0\right) &= \frac{1}{2}\left[\mathrm{tr}\left(\mathbf{P}_\pi\mathbf{\Sigma}^{-1}\mathbf{P}'_\pi\mathbf{\Sigma}_\pi\right) - n - \log\frac{|\mathbf{P}_\pi\mathbf{\Sigma}^{-1}\mathbf{P}'_\pi|}{|\mathbf{\Sigma}_\pi|}\right] \\ &= \frac{1}{2}\left[\mathrm{tr}\left(\mathbf{P}_\pi\mathbf{\Sigma}^{-1}\mathbf{P}'_\pi\mathbf{\Sigma}_\pi\right) - \log\left|\mathbf{P}_\pi\mathbf{\Sigma}^{-1}\mathbf{P}'_\pi\mathbf{\Sigma}_\pi\right|\right] - \frac{n}{2} \\ &= \frac{1}{2}\left[\mathrm{tr}\left(\mathbf{A}^{-1}\right) + \log|\mathbf{A}| - n\right] \end{aligned} \tag{E.1}$$

with $\mathbf{A} = \mathbf{P}_\pi\mathbf{\Sigma}\mathbf{P}'_\pi\mathbf{\Sigma}_\pi^{-1} = \mathbf{P}_\pi(\mathbf{A}_\rho)^{-1}\mathbf{P}'_\pi\mathbf{P}_\pi(\mathbf{A}'_\rho)^{-1}\mathbf{P}'_\pi\mathbf{\Sigma}_\pi^{-1}$. Since that $\mathrm{tr}\left(\mathbf{AB}\right) = \mathrm{tr}\left(\mathbf{BA}\right)$, we can compute the trace of $\mathbf{A}^{-1}$ as:

$$\mathrm{tr}(\mathbf{A}^{-1}) = \mathrm{tr}(\mathbf{P}_\pi\mathbf{\Sigma}^{-1}\mathbf{P}'_\pi\mathbf{\Sigma}_\pi) = \sum_{g=1}^{G}\mathrm{tr}\left(\mathbf{E}_g\mathbf{P}_\pi\mathbf{\Sigma}^{-1}\mathbf{P}'_\pi\mathbf{\Sigma}_\pi\mathbf{E}_g\right)$$

and for each g, the trace is equal to the sum $c(2g-1, 2g-1) + c(2g, 2g)$, where $c(i,j)$ is the i,j–th term of $\mathbf{P}'_\pi\mathbf{\Sigma}^{-1}\mathbf{P}_\pi\mathbf{\Sigma}_\pi$, since that $\mathbf{\Sigma}_\pi$ is block diagonal,

$$\begin{aligned} c(2g-1,2g-1) &= \sigma^*(\pi(2g-1),\pi(2g-1))\sigma(\pi(2g-1),\pi(2g-1)) + \sigma^*(\pi(2g-1),\pi(2g))\sigma(\pi(2g),\pi(2g-1)) \\ c(2g,2g) &= \sigma^*(\pi(2g),\pi(2g))\sigma(\pi(2g),\pi(2g)) + \sigma^*(\pi(2g),\pi(2g-1))\sigma(\pi(2g-1),\pi(2g)) \end{aligned}$$

where $\sigma^*(i,j)$ and $\sigma(i,j)$ are the (i,j)–th components of $\mathbf{\Sigma}^{-1}$ and $\mathbf{\Sigma}$, respectively. The term $\log|\mathbf{A}|$ can be written as a sum of G components as well: $\log|\mathbf{A}| = \log|\mathbf{\Sigma}| + \log|\mathbf{\Sigma}_\pi^{-1}| = \log|\mathbf{\Sigma}| - \log|\sum_g \mathbf{E}_g\mathbf{P}_\pi\mathbf{\Sigma}\mathbf{P}'_\pi\mathbf{E}_g|$. Since the matrix $\mathbf{\Sigma}_\pi = \sum_g \mathbf{E}_g\mathbf{P}_\pi\mathbf{\Sigma}\mathbf{P}'_\pi\mathbf{E}_g$ is a block diagonal matrix, its determinant is equal to the product of determinants of blocks in the main diagonal, namely,

$$\log|\sum_g \mathbf{E}_g\mathbf{P}_\pi\mathbf{\Sigma}\mathbf{P}'_\pi\mathbf{E}_g| = \log\prod_g |\mathbf{E}'_{g,g}\mathbf{P}_\pi\mathbf{\Sigma}\mathbf{P}'_\pi\mathbf{E}_{g,g}| = \sum_g \log|\mathbf{E}'_{g,g}\mathbf{P}_\pi\mathbf{\Sigma}\mathbf{P}'_\pi\mathbf{E}_{g,g}| = \sum_g \log|\mathbf{C}(\pi(2g-1),\pi(2g))|$$

with $\mathbf{E}_{g,g} = (\mathbf{e}_{2g-1}, \mathbf{e}_{2g})$, the $(2g-1, 2g)$–th columns of $\mathbf{E}_g$, while the determinant of $\log|\mathbf{\Sigma}|$ is invariant under permutations. Finally, we obtain

$$\arg\min_\pi KL(f_0^\pi||f_0) = \arg\min_\pi \sum_g \left(b(\pi(2g-1),\pi(2g)) - \log|\bar{\sigma}(\pi(2g-1),\pi(2g))|\right),$$

where $b(i,j) = c(i,i) + c(j,j)$ and $\bar{\sigma}(i,j) = \sigma(i,i)\sigma(j,j) - \sigma(i,j)\sigma(j,i)$. Now, because of

$$\sum_g \left[\sigma^*(\pi(2g-1),\pi(2g-1))\sigma(\pi(2g-1),\pi(2g-1)) + \sigma^*(\pi(2g),\pi(2g))\sigma(\pi(2g),\pi(2g))\right] = \sum_{i=1}^n \sigma^*(i,i)\sigma(i,i)$$

for all π, we can write

$$\begin{aligned}
&\arg\min_\pi KL(f_0^\pi || f_0) \\
&= \arg\min_\pi \sum_g \big[\sigma^*(\pi(2g-1),\pi(2g))\sigma(\pi(2g),\pi(2g-1)) + \sigma^*(\pi(2g),\pi(2g-1))\sigma(\pi(2g-1),\pi(2g)) \\
&\qquad - \log\left(\sigma(\pi(2g-1),\pi(2g-1))\sigma(\pi(2g),\pi(2g)) - \sigma(\pi(2g),\pi(2g-1))\sigma(\pi(2g-1),\pi(2g))\right)\big].
\end{aligned} \tag{E.2}$$

Appendix E.2. Proof of Theorem 5.3

Note that

$$\begin{aligned}
\hat{J}_n(\hat{\boldsymbol{\theta}}_n) - J(\hat{\boldsymbol{\theta}}_n) &= \frac{1}{G}\sum_g \sum_{j\neq g} \sum_{i=1}^4 \sum_{d_i=\{0,1\}} \nabla_\theta^g(\hat{\boldsymbol{\theta}}_n; d_1, d_2)\nabla_\theta^j(\hat{\boldsymbol{\theta}}_n; d_3, d_4)' \left(p_{[gj]}(d_1,d_2,d_3,d_4;\, \hat{\boldsymbol{\theta}}_n) - p_{[gj]}(d_1,d_2,d_3,d_4;\, \boldsymbol{\theta}_0)\right) \\
&\quad + \frac{1}{G}\sum_g \sum_{i=1}^2 \sum_{d_i=\{0,1\}} \nabla_\theta^g(\hat{\boldsymbol{\theta}}_n; d_1, d_2)\nabla_\theta^g(\hat{\boldsymbol{\theta}}_n; d_1, d_2)' \left(p_g(d_1,d_2;\, \hat{\boldsymbol{\theta}}_n) - p_g(d_1,d_2;\, \boldsymbol{\theta}_0)\right)
\end{aligned}$$

where $\nabla_\theta^g(\hat{\boldsymbol{\theta}}_n; d_1, d_2)$ is defined in equation (9). We then have,

$$\|\hat{J}_n(\hat{\boldsymbol{\theta}}_n) - J(\hat{\boldsymbol{\theta}}_n)\| \leq \sup_{1\leq g,j\leq G} \sup_{d_1,\ldots,d_4} \left|\frac{p_{[gj]}(d_1,d_2,d_3,d_4;\, \hat{\boldsymbol{\theta}}_n)}{p_{[gj]}(d_1,d_2,d_3,d_4;\, \boldsymbol{\theta}_0)} - 1\right| \left\|\mathbb{E}_{\mathbf{y}_g,\mathbf{y}_j} \frac{1}{G}\sum_g\sum_{j\neq g} \nabla_\theta^g(\hat{\boldsymbol{\theta}}_n; y_{g_1}, y_{g_2})\nabla_\theta^j(\hat{\boldsymbol{\theta}}_n; y_{j_1}, y_{j_2})'\right\|$$

where

$$\mathbb{E}_{\mathbf{y}_g,\mathbf{y}_j} \nabla_\theta^g(\hat{\boldsymbol{\theta}}_n; y_{g_1}, y_{g_2})\nabla_\theta^j(\hat{\boldsymbol{\theta}}_n; y_{j_1}, y_{j_2})' = \sum_{i=1}^4 \sum_{d_i=\{0,1\}} \nabla_\theta^g(\hat{\boldsymbol{\theta}}_n; d_1, d_2)\nabla_\theta^j(\hat{\boldsymbol{\theta}}_n; d_3, d_4)' p_{[gj]}(d_1,d_2,d_3,d_4;\, \boldsymbol{\theta}_0)$$

and we write $p_{[gj]}(d_1,d_2,d_3,d_4;\, \boldsymbol{\theta}) = p_g(d_1,d_2;\boldsymbol{\theta})\mathbb{I}(d_1 = d_3)\mathbb{I}(d_2 = d_4)$ whenever $g = j$. Clearly, because of $\sqrt{G}\frac{\partial \ell_n(\hat{\boldsymbol{\theta}}_n)}{\partial \boldsymbol{\theta}} \to_d \mathcal{N}\left(0, J\left(\boldsymbol{\theta}_0\right)\right)$, we have that

$$\mathbb{E}_{\mathbf{y}_g,\mathbf{y}_j} \frac{1}{G}\sum_g\sum_{j\neq g} \nabla_\theta^g(\hat{\boldsymbol{\theta}}_n; y_{g_1}, y_{g_2})\nabla_\theta^j(\hat{\boldsymbol{\theta}}_n; y_{j_1}, y_{j_2})' = \mathbb{E}_{\mathbf{y}} \sqrt{G}\frac{\partial \ell_n(\hat{\boldsymbol{\theta}}_n)}{\partial \boldsymbol{\theta}}\sqrt{G}\frac{\partial \ell_n(\hat{\boldsymbol{\theta}}_n)}{\partial \boldsymbol{\theta}'} = O_p(1)$$

while

$$\sup_{1\leq g,j\leq G} \sup_{d_1,\ldots,d_4} \left|\frac{p_{[gj]}(d_1,d_2,d_3,d_4;\, \hat{\boldsymbol{\theta}}_n)}{p_{[gj]}(d_1,d_2,d_3,d_4;\, \boldsymbol{\theta}_0)} - 1\right| = o_p(1)$$

from the continuity of $p_{[gj]}(\cdot;\boldsymbol{\theta})$ in $\boldsymbol{\theta}$ and because of the assumption $p_{[gj]}(d_1,d_2,d_3,d_4;\, \boldsymbol{\theta}_0) > \delta$ for all g, j and $\mathbf{d} = (d_1,d_2,d_3,d_4)$. The claim $\|\hat{J}_n(\hat{\boldsymbol{\theta}}_n) - J(\boldsymbol{\theta}_0)\| = o_p(1)$ finally follows if we prove that $\|J(\hat{\boldsymbol{\theta}}_n) - J(\boldsymbol{\theta}_0)\| = o_p(1)$. This is a consequence of the continuous mapping theorem, implying that $J\left(\boldsymbol{\theta}_0\right)^{-1/2} \hat{J}_n\left(\hat{\boldsymbol{\theta}}_n\right) J\left(\boldsymbol{\theta}_0\right)^{-1/2} \to_d \mathcal{W}\left(\mathbf{I}, 1\right)$, where $\mathcal{W}$ is a $(k+1)$–dimensional ($k+2$ in the SARAR probit case) Wishart distribution with scale matrix $\mathbf{I}$ and 1 d.f., and $J\left(\hat{\boldsymbol{\theta}}_n\right) = \mathbb{E}_{\theta_0} J_n\left(\hat{\boldsymbol{\theta}}_n\right)$, which implies that $J\left(\hat{\boldsymbol{\theta}}_n\right) \to_p J\left(\boldsymbol{\theta}_0\right)\mathbb{E}\left(\mathcal{W}\left(\mathbf{I}, 1\right)\right) = 1$.

Appendix E.3. Proof of Theorem 5.4

The proof follows by adapting Theorem 2 in Mammen (1992). Let $(\mathbf{Z}_1, \ldots, \mathbf{Z}_G)$ be a vector of random independent couples of binary variables $\mathbf{Z}_g = (z_{g_1}, z_{g_2})$ with distributions $p_g(\cdot, \cdot; \boldsymbol{\theta}_0)$, that is, the same marginal distributions as in equation (7), corresponding to the *true* parameter vector. We can imagine the vector $\mathbf{Z}$ as a random sample from a fictitious DGP, where all couples of binary variables are actually independent. Its corresponding exact likelihood function would be, using equation (8), $\ell_n(\boldsymbol{\theta}; \mathbf{Z}) = G^{-1} \sum_g \log\left(p_g\left(z_{g_1}, z_{g_2}; \boldsymbol{\theta}\right)\right)$. Let us denote $\hat{\boldsymbol{\theta}}_n(\mathbf{Z})$ the maximizer of $\ell_n(\boldsymbol{\theta}; \mathbf{Z})$, to distinguish it from $\hat{\boldsymbol{\theta}}_n$. We refer to the notation in Mammen (1992) and define the triangular array of vectors:

$$Y_{n,g} = \mathbf{H}(\boldsymbol{\theta}_0)^{-1} \frac{1}{G} \frac{\partial p_g\left(z_{g_1}, z_{g_2}; \hat{\boldsymbol{\theta}}_n\right)}{p_g\left(z_{g_1}, z_{g_2}; \hat{\boldsymbol{\theta}}_n\right)}, \quad g = 1, \ldots, G.$$

Note that Assumptions 1–10 hold if the data come from $\mathbf{Z}$ (the only Assumptions that have to be checked are Assumptions 6 and 8, that are trivially satisfied). Therefore, all conclusions drawn for $\hat{\boldsymbol{\theta}}_n$ hold for $\hat{\boldsymbol{\theta}}_n(\mathbf{Z})$ too, and we have:

$$\sqrt{n}\left(\hat{\boldsymbol{\theta}}_n(\mathbf{Z}) - \boldsymbol{\theta}_0\right) = \sum_g Y_{n,g} + o_p(1)$$

and $\sqrt{n}\left(\hat{\boldsymbol{\theta}}_n(\mathbf{Z}) - \boldsymbol{\theta}_0\right) J^{-1/2}(\boldsymbol{\theta}_0)$ is asymptotically standard normal (multivariate). Due to the independence of $\mathbf{Z}_g$'s, we can exploit a bivariate version of Mammen's Theorem 2, by concluding that

$$d_\infty\left(\mathcal{L}^\star_z\left(\sqrt{n}\left(\hat{\boldsymbol{\theta}}^\star_n - \hat{\boldsymbol{\theta}}_n(\mathbf{Z})\right)\right), \mathcal{L}_z\left(\sqrt{n}\left(\hat{\boldsymbol{\theta}}_n(\mathbf{Z}) - \boldsymbol{\theta}_0\right)\right)\right) \to_p 0$$

where we denoted by $\mathcal{L}_z$ the sampling probability distribution of the statistic induced by the DGP of the $\mathbf{Z}$ vector and by $\mathcal{L}^\star_z$ the distribution conditional on the sample $\mathbf{Z}$. Further, we have

$$d_\infty\left(\mathcal{L}_y\left(\sqrt{n}\left(\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0\right)\right), \mathcal{L}_z\left(\sqrt{n}\left(\hat{\boldsymbol{\theta}}_n(\mathbf{Z}) - \boldsymbol{\theta}_0\right)\right)\right) \to_p 0,$$

because they share the same limiting Gaussian distribution (clearly, the MLE estimator $\hat{\boldsymbol{\theta}}_n(\mathbf{Z})$ and the PMLE estimator $\hat{\boldsymbol{\theta}}$ are algebraically the same if the observed values of $\mathbf{Z}$ and $\mathbf{y}$ coincide). It remains to prove that the bootstrap distribution $\mathcal{L}^\star_y\left(\sqrt{n}\left(\hat{\boldsymbol{\theta}}^\star_n - \hat{\boldsymbol{\theta}}_n\right)\right)$ conditional on the sample $\mathbf{y}$, is asymptotically the same as $\mathcal{L}^\star_z\left(\sqrt{n}\left(\hat{\boldsymbol{\theta}}^\star_n - \hat{\boldsymbol{\theta}}_n(\mathbf{Z})\right)\right)$. By noting that the distribution of $\hat{\boldsymbol{\theta}}^\star_n$ conditional on the sample does not directly depend neither on $\mathbf{y}$ nor on $\mathbf{Z}$, but it is completely determined by $\hat{\boldsymbol{\theta}}_n$ or $\hat{\boldsymbol{\theta}}_n(\mathbf{Z})$, we have that, for any $\boldsymbol{\theta}_n \in \boldsymbol{\Theta}$, $\mathcal{L}^\star_z\left(\sqrt{n}\left(\hat{\boldsymbol{\theta}}^\star_n - \boldsymbol{\theta}_n\right)\right) = \mathcal{L}^\star_y\left(\sqrt{n}\left(\hat{\boldsymbol{\theta}}^\star_n - \boldsymbol{\theta}_n\right)\right)$ because of

$$\Pr\left\{\sqrt{n}\left(\hat{\boldsymbol{\theta}}^\star_n - \boldsymbol{\theta}_n\right) \le t \mid \ \mathbf{Z} : \hat{\boldsymbol{\theta}}_n(\mathbf{Z}) = \boldsymbol{\theta}_n\right\} = \Pr\left\{\sqrt{n}\left(\hat{\boldsymbol{\theta}}^\star_n - \hat{\boldsymbol{\theta}}_n\right) \le \mathbf{y} \mid \ \mathbf{y} : \hat{\boldsymbol{\theta}}_n(\mathbf{y}) = \boldsymbol{\theta}_n\right\}.$$

This, together with consistency of both $\hat{\boldsymbol{\theta}}_n$ and $\hat{\boldsymbol{\theta}}_n(\mathbf{Z})$ for $\boldsymbol{\theta}_0$, implies that

$$d_\infty\left(\mathcal{L}_y\left(\sqrt{n}\left(\hat{\boldsymbol{\theta}}^\star_n - \hat{\boldsymbol{\theta}}_n\right)\right), \mathcal{L}_z\left(\sqrt{n}\left(\hat{\boldsymbol{\theta}}^\star_n - \hat{\boldsymbol{\theta}}_n(\mathbf{Z})\right)\right)\right) \to_p 0.$$

Appendix F. Maximum matching

Maximum matching is a problem in graph theory consisting in finding the best way to match two nodes in a (weighted or unweighted) graph. Given a graph $\mathcal{G} = (V, E)$, a matching is a set M of pairwise non–adjacent edges, such that no couple of edges shares the same vertex. A maximum matching is a set M that is not a subset of any other matching of the same graph. If the graph is weighted, we can define a weighted maximum matching as a set M that produces a matching of maximum (minimum) total weight. The maximum matching problem has been solved by the Edmonds' blossom algorithm for unweighted matchings and later extended to weighted matchings, see Edmonds (1965b) and Edmonds (1965a). To explain the main ideas of the blossom algorithm, it is necessary to introduce some concepts from graph theory. A vertex v is exposed if there is no edge in M that is incident with v. A path P is called an M–augmenting path if it is a path that starts and ends in two exposed vertices and its edges are alternatively in and outside M. The maximum matching algorithm exploits a result, known as Berge's lemma: a matching M is maximal if there is no augmenting path in $\mathcal{G}$. The algorithm proceeds by starting from an initial matching and looking for an augmenting path. If the augmenting path P exists, then the matching is updated. If it does not exists, then the algorithm stops, see Figure F.4. The core of the blossom algorithm is a particular structure called "blossom", and contractions.

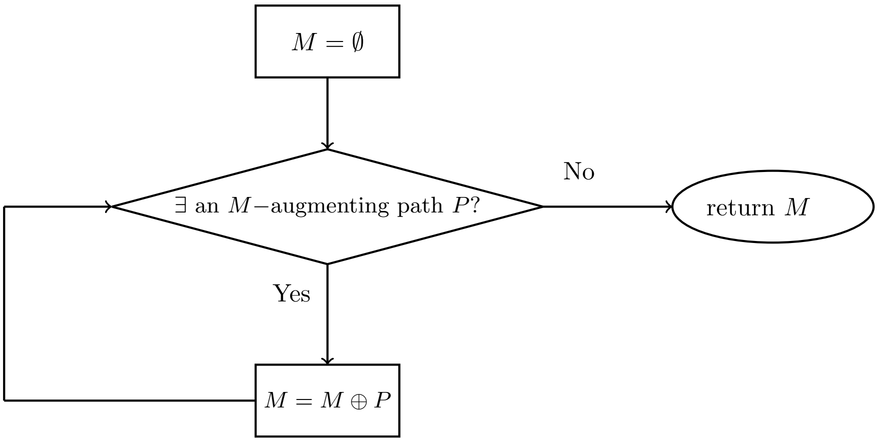

Figure F.4: Flow chart of a maximum matching algorithm

Given a graph $\mathcal{G}$, and a matching M of $\mathcal{G}$, a blossom B is a cycle of $2k + 1$ edges k of which belong to M, and such that there is one vertex v of the cycle (called the base) that has an alternating path of even length to an exposed vertex u. Blossoms are used to contract the graph: whenever a blossom is found in $\mathcal{G}$, the whole blossom is contracted into the base vertex v and this results into e new (smaller) graph $\mathcal{G}'$ and matching M'. Then, finding M–augmenting paths in $\mathcal{G}$ is transformed into the problem of finding M'–augmenting paths in the reduced graph $\mathcal{G}'$. Readers can refer to Galil (1986) (and references therein) for a deeper introduction to maximum matching algorthms.