unibz

Fakultät für Wirtschaftswissenschaften
Facoltà di Economia
Faculty of Economics and Management

BEMPS –
Bozen Economics & Management
Paper Series

NO 44 / 2017

Quasi-ML estimation, Marginal Effects and Asymptotics for Spatial Autoregressive Nonlinear Models

Anna Gloria Billé, Samantha Leorato

Quasi–ML estimation, Marginal Effects and Asymptotics for Spatial Autoregressive Nonlinear Models

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

[a] *School of Economics and Management, Free University of Bozen–Bolzano, Bolzano, Italy*
[b] *Department of Economics and Finance, University of Rome Tor Vergata, Rome, Italy*

Abstract

In this paper we propose a Partial–MLE 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 the SAE(1)–probit model, considered by Wang et al. (2013), and the more interesting SAR(1)–probit model. We perform a complete asymptotic analysis, and account for the possible finite sum approximation of the covariance matrix (Quasi–MLE) to speed the computation. 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. Finally, we provide appropriate definitions of marginal effects for this setting. Finite sample properties of the estimator are studied through a simulation exercise and a real data application. In our simulations, we also consider both sparse and dense matrices for the specification of the true spatial models, and cases of model misspecifications due to different assumed weighting matrices.

Keywords: Spatial autoregressive–regressive probit model, Nonlinear modeling, SARAR, Partial Maximum Likelihood, Quasi Maximum Likelihood, Marginal effects.
JEL codes: C13,C31,C35,C51.

1. Introduction

Estimation theory and inference for econometric models which deal with spatially–distributed data differ substantially from the usual techniques used in standard statistics/econometrics, see Whittle (1954), Besag (1972), Besag (1974), Ord (1975), Cliff and Ord (1981). In spatial econometrics (Anselin, 1988), a large number of theoretical papers face instead the added difficulties in deriving the asymptotic properties of a series of extremum estimators, i.e. GMM, Quasi–MLE, etc., see Kelejian and Prucha (1998), Lee (2003), Lee (2004), Kelejian and Prucha (2010). The bi–directionality nature of spatial dependence, leading to a simultaneous specification rather than the conditional specification of spatial autoregressive models (Sain and Cressie, 2007)

*Corresponding author. E–mail: annagloria.bille@unibz.it

is an example. Anyway, being spatial dependence simply a special 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 are able to exploit the information gathered not only about the observed values but also on the locations of the endogenous random variables. For instance, the uniform boundedness assumption of the weighting matrices is 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.

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 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). Nonlinear models, like probit/logit models, are useful to analyse endogenous dichotomous dependent variables, but the specified functional form is nonlinear in parameters and their estimation requires iterative optimization procedures. Differently from the linear case, spatial dependence adds a further complexity in the estimation of parameters of spatial discrete choice (SDC) models.

From a computational point of view, 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). The optimization of the objective function requires repeated calculations of the inverses of n–dimensional matrices, i.e. $(\mathbf{I}_n - \rho\mathbf{W}_n)^{-1}$, which also preclude an easy extension to panel data applications, whose diffusion is experiencing a massive increase, see e.g. Smith and LeSage (2004), Lee and Yu (2010), Kapoor et al. (2007), Lee and Yu (2016), Baltagi et al. (2017). Approximate and conditional maximum likelihood estimators,in the works by Pace and LeSage (2011), Mozharovskyi and Vogler (2016), Martinetti and Geniaux (2017), are one way to cope with this problem.

Another 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 due to 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 recognise the works by Pinkse and Slade (1998), Klier and McMillen (2008), where the latter, in particular, proposed a Linearized GMM estimator which is feasible even with moderate to large sample sizes but it is reasonable as long as the autocorrelated coefficient is relatively small.

Composite MLEs have been proved to be computationally efficient and statistically consistent, see Heagerty and Lele (1998), Gao and Song (2010), Bhat (2011), Bai et al. (2014). In spatial econometrics, Wang et al. (2013) have recently proposed a Partial–MLE (a particular Quasi–MLE) for a spatial (first–order)

autoregressive probit error (SAE(1)–probit) model by dividing observations into many 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 (in the linear case see the work by Arbia (2014)). 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. Moreover, 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 in which 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 more interesting, because a direct information on interactions among economic agents' choices is measurable.

In this paper we propose a Partial–MLE specifically based on bivariate joint distributions to deal with spatial dependences within groups (couples in our case) of random variables for a SAR(1)–probit model and for its extension to the more general spatial (first–order) autoregressive-regressive probit model with (first–order) autoregressive disturbances (SARAR(1,1)–probit). For very large data sets we consider a truncated series as a reasonable approximation of $(\mathbf{I}_n - \rho\mathbf{W}_n)^{-1}$ as in Kelejian et al. (2004), which defines our Quasi–MLE (asymptotically equivalent to the Partial–MLE). We perform its finite sample properties and derive the increasing domain asymptotic results. A Kullback–Leibler divergence approach is used to choose couples by controlling for the loss of statistical information. We also propose proper definitions of the marginal effects, discussed through Monte Carlo simulations. In our simulations, we consider both *sparse* and *dense* matrices for the specification of the true spatial models. Robustness checks on model misspecification of the weighting matrices $\mathbf{W}_n$ are also included. All these figures make our work substantially different from that proposed by Wang et al. (2013).

The rest of the paper is organized as follow. Section 2 specifies a general spatial probit model, i.e. a spatial (first–order) autoregressive-regressive probit model with (first–order) autoregressive disturbances (SARAR(1,1)–probit) and explains the related problem of inconsistency due to unobserved spatially autocorrelated shocks. Section 3 describes our Partial and Quasi–maximum likelihood estimator (PMLE and QMLE) based on bivariate distributions. In Section 4 we propose an algorithm for the choice of couples based on a minimum expected information loss. Section 5 reports the asymptotic results based on the increasing domain assumption. Section 6 defines the marginal effects. Section 7 evaluates the finite sample properties of our QMLE 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). Finally, Section 9 concludes.

2. Model specification

Let $\mathbf{y}_n$ be a 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

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

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 $\mathbb{I}_n \left(\mathbf{y}^* > \mathbf{0}\right) = \left(\mathbb{I}(y_1^* > 0), \ldots, \mathbb{I}(y_n^* > 0)\right)'$, $\mathbf{X}_n$ 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 $\sigma_\varepsilon^2 < \infty$, such that $\boldsymbol{\Sigma}_\varepsilon = \sigma_\varepsilon^2 \mathbf{I}_n$. Latent variables are then assumed to be linear functions of the regressors, but they are observed through the use of a binary variable that makes the overall model nonlinear in parameters. In the nonlinear case, 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 $\mathbf{W}_n = \mathbf{M}_n$, then distinguishing among the two spatial effects may be difficult, with possible identification problems of the autoregressive parameters. In this particular case, sufficient conditions to ensure identifiability of the linear model is that the covariates make a material contribution towards explaining variation in the dependent variable.

The inclusion of spatially-lagged dependent variables $\mathbf{W}_n \mathbf{y}_n^*$ typically causes an endogeneity problem, which in turn produces inconsistency of least squares estimators. This problem is referred to the bi-directionality 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 (see Hamilton (1994)). These type 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.

In order to ensure stable spatial processes we have to introduce some assumptions in line with Kelejian and Prucha (2010). Let us first recall the following result

Lemma 2.1. *Let $\overline{\tau}$ denote the spectral radius of the square n–dimensional $\mathbf{W}_n$ (resp. $\mathbf{M}_n$) matrix, i.e.: $\overline{\tau} = max\{|\omega_1|, ..., |\omega_n|\}$, where $\omega_1, ..., \omega_n$ are the eigenvalues of $\mathbf{W}_n$ (resp. $\mathbf{M}_n$). Then, $(\mathbf{I}_n - \rho\mathbf{W}_n)^{-1}$ $\left(resp.\ \ (\mathbf{I}_n - \lambda\mathbf{M}_n)^{-1}\right)$ is non singular for all values of ρ (resp. λ) in the interval $(-1/\overline{\tau}, 1/\overline{\tau})$.*

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

Assumption 1(a) means that each spatial unit is not viewed as its own neighbor, whereas Assumption 1(b) ensures that the model in (1) can be uniquely defined by Lemma 2.1. Then, if we interpret model in (1) as an equilibrium relationship, 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).

Assumption 2. *Matrices $\mathbf{W}_n$ and $\mathbf{M}_n$ and $(\mathbf{I} - \rho\mathbf{W}_n)^{-1}$ and $(\mathbf{I} - \lambda\mathbf{M}_n)^{-1}$ 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 2, is equivalent to Assumption 5 in Lee (2004) and it ensures that the following infinite series expansions are well defined

$$\begin{aligned} \mathbf{A}_\rho^{-1} &= (\mathbf{I}_n - \rho\mathbf{W}_n)^{-1} = \mathbf{I}_n + \rho\mathbf{W}_n + \rho^2\mathbf{W}_n^2 + \cdots + \rho^q\mathbf{W}_n^q + \ldots \\ \mathbf{B}_\lambda^{-1} &= (\mathbf{I}_n - \lambda\mathbf{M}_n)^{-1} = \mathbf{I}_n + \lambda\mathbf{M}_n + \lambda^2\mathbf{M}_n^2 + \cdots + \lambda^q\mathbf{M}_n^q + \ldots \quad . \end{aligned} \tag{2}$$

It amounts at having rows and columns of both $\mathbf{W}_n$ and $\mathbf{M}_n$ before normalization uniformly bounded in absolute value as n goes to infinity, ensuring that the correlation between two spatial units should converge to zero as the distance separating them increases to infinity. General normalization rules exist. A spectral–normalization rule is generally recommended to guarantee 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, we should note that the resulting spatial interaction coefficient corresponding to the normalized weights matrix will in general depend on the sample size because the normalizing factor (e.g. the spectral radius of $\mathbf{W}_n$ or $\mathbf{M}_n$) depends on it as well.

The use of spectral–normalisation may raise some concerns related to Assumption 2: while Assumption 1 and Lemma 2.1 guarantee that both $(\mathbf{I}_n - \rho\mathbf{W}_n)$ and $(\mathbf{I}_n - \lambda\mathbf{M}_n)$ are bounded in row and column sum, it is not completely clear what is the effect of spectral–normalization on $\mathbf{W}_n$ and $\mathbf{M}_n$. In fact, if the spectral radius $\bar{\tau}_n$ decreases with n, the normalised matrices $\mathbf{W}_n$ (or $\mathbf{M}_n$) would have bounded entries, but could be unbounded in row or column sum. A deep investigation over the conditions that a weight matrix has to fulfil in order that the spectral–normalised matrix satisfies Assumption 2 is beyond the scope of this work and will

be a subject of further research.

Besides the spectral–normalization that we only consider for dense weight matrices, in this paper we use the typical row–normalization (i.e. $\mathbf{W}_n$, or $\mathbf{M}_n$, is a row–stochastic matrix) of the weighting matrices, so that $\sum_j w_{ij} = 1$ (resp. $\sum_j m_{ij} = 1$), which still have the appealing interpretation of considering spatial effects as a weighted average of neighboring spatial random variables. The resulting weighting matrices are used to define the data generating processes (DGPs), that will change according to the type of *criterion* and *distance* used. More details on the definition of the weight matrices are given in Section 7.

Due to the 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

$$\begin{aligned} \mathbf{y}_n^* &= \mathbf{A}_\rho^{-1}\mathbf{X}_n\boldsymbol{\beta} + \mathbf{A}_\rho^{-1}\mathbf{u}_n = \mathbf{A}_\rho^{-1}\mathbf{X}_n\boldsymbol{\beta} + \mathbf{A}_\rho^{-1}\mathbf{B}_\lambda^{-1}\boldsymbol{\varepsilon}_n = \mathbf{A}_\rho^{-1}\mathbf{X}_n\boldsymbol{\beta} + \nu_n, \quad \nu_n \sim \mathcal{N}_n\left(\mathbf{0}_n, \boldsymbol{\Sigma}_\nu\right) \\ \mathbf{y}_n &= \mathbb{I}_n\left(\mathbf{y}_n^* > \mathbf{0}_n\right) \end{aligned} \tag{3}$$

where $\nu_n = \mathbf{A}_\rho^{-1}\mathbf{B}_\lambda^{-1}\boldsymbol{\varepsilon}_n$ and $\boldsymbol{\Sigma}_\nu := \boldsymbol{\Sigma}_{\nu(\rho,\lambda)} = \mathbb{E}\left[\nu_n\nu_n'\right] = \sigma_\varepsilon^2\mathbf{A}_\rho^{-1}\mathbf{B}_\lambda^{-1}\mathbf{B}_\lambda^{-1'}\mathbf{A}_\rho^{-1'}$ with $\sigma_\varepsilon^2 = 1$ for identification. From the reduced form in equation (3), we finally obtain expected value and variance–covariance matrix, for all $i = 1, \ldots, n$:

$$\begin{aligned} \mathbb{E}\left[y_i \mid \mathbf{X}_n\right] &= \mathrm{P}\left\{y_i = 1 \mid \mathbf{X}_n\right\} = \mathrm{P}\left\{i\text{–th term of } \nu_n > i\text{–th term of } \left\{\mathbf{A}_\rho^{-1}\mathbf{X}_n\boldsymbol{\beta}\right\} \mid \mathbf{X}_n\right\} \\ &= \Phi\left(\{\boldsymbol{\Sigma}_{\nu(\rho,\lambda)}\}_{ii}^{-1/2}\{\mathbf{A}_\rho^{-1}\mathbf{X}_n\boldsymbol{\beta}\}_i\right). \end{aligned} \tag{4}$$

The robust (to heteroskedasticity) variance–covariance matrix is obtained from

$$\mathbb{VC}\left[y_i \mid \mathbf{X}_n\right] = \Phi\left(\{\boldsymbol{\Sigma}_{\nu(\rho,\lambda)}\}_{ii}^{-1/2}\{\mathbf{A}_\rho^{-1}\mathbf{X}_n\boldsymbol{\beta}\}_i\right)\left[1 - \Phi\left(\{\boldsymbol{\Sigma}_{\nu(\rho,\lambda)}\}_{ii}^{-1/2}\{\mathbf{A}_\rho^{-1}\mathbf{X}_n\boldsymbol{\beta}\}_i\right)\right]. \tag{5}$$

where $\{\cdot\}_{ii}$ is the $i-$th diagonal component of the matrix in brackets.

2.1. Nested–model specifications

Two widely used sub–models can be specified starting from equation (1): (a) spatial (first–order) autoregressive probit (SAR(1)–probit) model by letting $\lambda = 0$; (b) spatial (first–order) autoregressive error probit (SAE(1)–probit) model by letting $\rho = 0$.

(a)

$$\mathbf{y}_n^* = \rho\mathbf{W}_n\mathbf{y}_n^* + \mathbf{X}_n\boldsymbol{\beta} + \boldsymbol{\varepsilon}_n, \quad \boldsymbol{\varepsilon}_n \sim \mathcal{N}_n\left(\mathbf{0}_n, \boldsymbol{\Sigma}_\varepsilon\right), \quad \mathbf{y}_n = \mathbb{I}_n\left(\mathbf{y}_n^* > \mathbf{0}_n\right) \tag{6}$$

(b)

$$\mathbf{y}_n^* = \mathbf{X}_n\boldsymbol{\beta} + \mathbf{u}_n, \quad \mathbf{u}_n = \lambda\mathbf{M}_n\mathbf{u}_n + \boldsymbol{\varepsilon}_n, \quad \boldsymbol{\varepsilon}_n \sim \mathcal{N}_n\left(\mathbf{0}_n, \boldsymbol{\Sigma}_\varepsilon\right), \quad \mathbf{y}_n = \mathbb{I}_n\left(\mathbf{y}_n^* > \mathbf{0}_n\right) \tag{7}$$

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 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 structure. For example, in network economics literature, the main goal is to measure the direct interactions among economic/social agent choices, and a SAR(1)–probit model is usually preferred. However, a SAE(1)–probit model is at least interesting to possibly avoid the inconsistecy problem[1] (see subsection 2.2) which does not arise in the spatial linear case for the same model specification[2].

In this paper we develop the theory of a type of Partial–MLE for a SARAR(1,1)–probit model with a particular emphasis on the SAR(1)–probit case (see Section 3). We also study the properties of a Quasi–MLE, deriving from a finite order approximation of the spatial covariance structure entering the partial loglikelihood function.

2.2. The problem of inconsistency

The error term in a simple probit model summarizes the unknown information coming from other regressors (i.e. omitted variables) which we assume to be uncorrelated with those in $\mathbf{X}_n$. In this case, extremum estimators, such as likelihood based estimators, are consistent, see Amemiya (1977), Amemiya (1978) and Amemiya (1985). However, unknown forms of misspecification of the functional form (Yatchew and Griliches, 1985), for example when heteroskedastic errors are incorrectly assumed to be homoskedastic, lead to inconsistency of the maximum likelihood estimators in a nonlinear setting (Poirier and Ruud, 1988). Indeed, MLE is consistent if the conditional density of $\mathbf{y}_n|\mathbf{X}_n$ is correctly specified. Misspecification of the functional form in a probit context is equivalent to have a misspecification of the Bernoulli probability for each $y_i, 1 \leq i \leq n$.

In a SAE(1)–probit setting, heteroskedasticity will arise whenever the weights $\mathbf{M}_n$ induce non–constant diagonal terms of the matrix $\mathbf{\Sigma_u} = [\mathbf{B}_\lambda' \mathbf{B}_\lambda]^{-1}$. 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. $\operatorname{diag}(\mathbf{\Sigma_u}) = \operatorname{diag}[\mathbf{B}_\lambda' \mathbf{B}_\lambda]^{-1}$, 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).

[1] 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), Kelejian (2016).

[2] Apart from information that comes from economic theory, a SAE(1) model produces only more efficient estimates in the linear case.

In the general case, let $\mathbf{A}_\rho = (\mathbf{I}_n - \rho\mathbf{W}_n)$ and $\mathbf{B}_\lambda = (\mathbf{I}_n - \lambda\mathbf{M}_n)$. So we get

$$\begin{aligned}
\mathbf{y}_n^* &= \rho\mathbf{W}_n\mathbf{y}_n^* + \mathbf{X}_n\boldsymbol{\beta} + \mathbf{u}_n, \quad \mathbf{B}_\lambda\mathbf{u}_n = \boldsymbol{\varepsilon}_n \\
\mathbf{B}_\lambda\mathbf{y}_n^* &= \rho\mathbf{B}_\lambda\mathbf{W}_n\mathbf{y}_n^* + \mathbf{B}_\lambda\mathbf{X}_n\boldsymbol{\beta} + \boldsymbol{\varepsilon}_n \\
\mathbf{y}_n^* &= \lambda\mathbf{M}_n\mathbf{y}_n^* + \rho\mathbf{B}_\lambda\mathbf{W}_n\mathbf{y}_n^* + \mathbf{B}_\lambda\mathbf{X}_n\boldsymbol{\beta} + \boldsymbol{\varepsilon}_n, \quad \boldsymbol{\varepsilon}_n \sim \mathcal{N}_n\left(\mathbf{0}_n, \boldsymbol{\Sigma}_\varepsilon\right)
\end{aligned} \tag{8}$$

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

$$\mathbb{E}\left[\mathbf{y}_n^*\boldsymbol{\varepsilon}_n'\right] = \mathbf{A}_\rho^{-1}\mathbb{E}\left[\mathbf{u}_n\boldsymbol{\varepsilon}_n'\right] = \mathbf{A}_\rho^{-1}\mathbf{B}_\lambda^{-1} \tag{9}$$

and these correlations rule out the use of nonlinear least squares methods due to their inconsistency. For the SARAR(1,1)–probit model in equation (1), and its sub–specification SAR(1)–probit by letting $\lambda = 0$, we have $\mathbb{E}\left(\left(\mathbf{W}_n\mathbf{y}_n^*\right)\mathbf{u}_n'\right) \neq \mathbf{0}_n$ where $\mathbf{u}_n = \mathbf{B}_\lambda^{-1}\boldsymbol{\varepsilon}_n$ and $\mathbb{E}\left(\left(\mathbf{W}_n\mathbf{y}_n^*\right)\boldsymbol{\varepsilon}_n'\right) \neq \mathbf{0}_n$, respectively, see Kelejian and Prucha (1998) and Kelejian and Prucha (1999) in the linear case. Therefore, consistency can only be achieved by correctly specifying the conditional expected value of model in equation (1)

3. Partial–ML estimation

The main problem with estimating model (1) – or its sub–specifications – 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, GMM approach is preferred to MLE due to 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 propose a computationally feasible estimation procedure for the SARAR(1,1)–probit model. Our estimator is based on the principle of a Partial–MLE. In order to reduce the burden of inverting $\mathbf{A}_\rho$ and $\mathbf{B}_\lambda$, we also consider an asymptotically equivalent estimator, based on a finite order approximation (we refer to the Quasi–MLE estimator in this case). We give details on the definition of the estimator in Sections 3.1–3.2. Throughout this Section, all indexes n in vectors and matrices are omitted, to ease the notation.

3.1. SARAR(1,1)–probit model

We start by considering the SAR(1)–probit model specified in equation (6), and extend later the results to the SARAR(1,1)–probit model in equation (1). Similarly to Wang et al. (2013), we define a Partial–MLE, thus avoiding the problem of n–dimensional integration induced by these models. As already pointed out in Section 2, the major difference relative to the model considered in Wang et al. (2013) consists in 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[3] of this Section and Section 5 can be straightforwardly adapted to an $r-$dimensional partial distribution, with $r > 2$.

Throughout this Section, we are assuming that the couples $g = 1, \ldots, G$ are given (for example, $g_1 = 2g-1$, $g_2 = 2g$). We will discuss in more details criteria for the choice of couples in Section 4. Now, consider groups (couples) indexed by $g = 1, \ldots, G$. From model in equation (6), 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

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

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

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

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

Further, we write $\mathbf{X}_\rho = \mathbf{A}_\rho^{-1}\mathbf{X}$. Now, we can use arguments similar to those in Wang et al. (2013) to find, for all $d_1, d_2 \in \{0,1\}^2$, the probabilities:

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

For any $g = 1, \ldots, G$, let us define the functions (implicit in ρ and β)

$$\varphi_{1,g}(u) = \frac{\mathbf{x}_{\rho,g_1}\boldsymbol{\beta} + u\frac{\sigma_{g_1,g_2}}{\sigma^2_{g_2}}}{\sqrt{\sigma^2_{g_1} - \sigma^2_{g_1,g_2}/\sigma^2_{g_2}}} \quad \text{and} \quad \varphi_{2,g}(u) = \frac{\mathbf{x}_{\rho,g_2}\boldsymbol{\beta} + u\frac{\sigma_{g_1,g_2}}{\sigma^2_{g_1}}}{\sqrt{\sigma^2_{g_2} - \sigma^2_{g_1,g_2}/\sigma^2_{g_1}}}, \tag{10}$$

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

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

$$\begin{aligned} p_g(d_1, d_2) &= \int_{\{s_{g_1}u > -s_{g_1}\mathbf{x}_{\rho,g_1}\boldsymbol{\beta}\}} \frac{1}{\sigma_{g_1}}\phi\left(\frac{u}{\sigma_{g_1}}\right)\Phi\left(s_{g_2}\varphi_{2,g}(u)\right)du \\ &= \Pr\left\{s_{g_1}Z_1 > s_{g_1}\mathbf{x}_{\rho,g_1}\boldsymbol{\beta}, s_{g_2}Z_2 > s_{g_2}\mathbf{x}_{\rho,g_2}\boldsymbol{\beta}\right\} \end{aligned} \tag{11}$$

[3]The algorithm presented in Section 4 and the computation of the score vector given in Appendix E are instead specific to couples.

where $\mathbf{Z} = (Z_1, Z_2) \sim \mathcal{N}(0, \mathbf{\Sigma}_g)$[4].

The proof of Theorem 3.1 is given in Appendix C. Using Theorem 3.1, we can write the partial loglikelihood function of the spatial probit model as

$$\begin{aligned}\ell_n(\boldsymbol{\theta}; \mathbf{y}, \mathbf{X}) = \frac{1}{G}\sum_{g=1}^{G} [&y_{g_1} y_{g_2} \log(p_g(1,1)) + y_{g_1}(1-y_{g_2}) \log(p_g(1,0)) \\ &+ (1-y_{g_1}) y_{g_2} \log(p_g(0,1)) + (1-y_{g_1})(1-y_{g_2}) \log(p_g(0,0))]\end{aligned} \quad . \tag{12}$$

The loglikelihood for estimating a SARAR(1,1)–probit model is also given by equation (12) with probabilities defined in equation (11). The difference in the formula is given by the elements of the matrix $\mathbf{\Sigma}_g$, which now depends on both ρ and λ:

$$\mathbf{\Sigma} := \mathbf{\Sigma}_{\nu(\rho,\lambda)} = \mathbf{A}_\rho^{-1} \mathbf{B}_\lambda^{-1} \mathbf{B}_\lambda^{-1\prime} \mathbf{A}_\rho^{-1\prime}.$$

The score vector $\nabla(\boldsymbol{\theta}; \mathbf{y}, \mathbf{X}) = \left(\nabla_\beta(\boldsymbol{\theta})', \nabla_\rho(\boldsymbol{\theta})\right)'$ is equal to

$$\nabla_\beta(\boldsymbol{\theta}) = \frac{1}{G}\sum_g \nabla_\beta^g(\boldsymbol{\theta}), \qquad \nabla_\rho(\boldsymbol{\theta}) = \frac{1}{G}\sum_g \nabla_\rho^g(\boldsymbol{\theta})$$

where

$$\begin{aligned}\nabla_\beta^g(\boldsymbol{\theta}) = {} & y_{g_1} y_{g_2} \frac{\partial p_g(1,1)/\partial \boldsymbol{\beta}}{p_g(1,1)} + y_{g_1}(1-y_{g_2}) \frac{\partial p_g(1,0)/\partial \boldsymbol{\beta}}{p_g(1,0)} + (1-y_{g_1}) y_{g_2} \frac{\partial p_g(0,1)/\partial \boldsymbol{\beta}}{p_g(0,1)} \\ & + (1-y_{g_1})(1-yg_2) \frac{\partial p_g(0,0)/\partial \boldsymbol{\beta}}{p_g(0,0)} \\ \nabla_\rho^g(\boldsymbol{\theta}) = {} & y_{g_1} y_{g_2} \frac{\partial p_g(1,1)/\partial \rho}{p_g(1,1)} + y_{g_1}(1-y_{g_2}) \frac{\partial p_g(1,0)/\partial \rho}{p_g(1,0)} + (1-y_{g_1}) y_{g_2} \frac{\partial p_g(0,1)/\partial \rho}{p_g(0,1)} \\ & + (1-y_{g_1})(1-y_{g_2}) \frac{\partial p_g(0,0)/\partial \rho}{p_g(0,0)}\end{aligned} \tag{13}$$

and where formulas for $\partial p_g(d_1, d_2)/\partial \boldsymbol{\beta}$ and $\partial p_g(d_1, d_2)/\partial \rho$ are given in Appendix E. The score vector of a SARAR(1,1)–probit model is simply $\nabla(\boldsymbol{\theta}; \mathbf{y}, \mathbf{X}) = \left(\nabla_\beta(\boldsymbol{\theta})', \nabla_\rho(\boldsymbol{\theta}), \nabla_\lambda(\boldsymbol{\theta})\right)'$ with

$$\nabla_\beta(\boldsymbol{\theta}) = \frac{1}{G}\sum_g \nabla_\beta^g(\boldsymbol{\theta}), \qquad \nabla_\rho(\boldsymbol{\theta}) = \frac{1}{G}\sum_g \nabla_\rho^g(\boldsymbol{\theta}), \qquad \nabla_\lambda(\boldsymbol{\theta}) = \frac{1}{G}\sum_g \nabla_\lambda^g(\boldsymbol{\theta})$$

where all $\nabla_\cdot^g(\boldsymbol{\theta})$ follow from (13). The terms $\frac{p_g(d_1,d_2)}{\partial \boldsymbol{\beta}}$, $\frac{p_g(d_1,d_2)}{\partial \rho}$ and $\frac{p_g(d_1,d_2)}{\partial \lambda}$ are given in Appendix E.2.

To introduce our QML estimator, we note that (11), (12) and (13) are the exact formulas for the bivariate probabilities and partial loglikelihood of model (6) or (1) as well as for the score vector. In practice, however, 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}. A possible way to avoid the inversion of $\mathbf{A}_\rho$ (and $\mathbf{B}_\lambda$), when n is large, is through the approximation

[4]Note that in (11) the role of g_1 and g_2 may be inverted, thus, for example $p_g(1,1)$ may be equivalently written as:

$$p_g(1,1) = \int_{-\mathbf{x}_{\rho,g_2}\boldsymbol{\beta}}^{\infty} \frac{1}{\sigma_{g_2}} \phi\left(\frac{u}{\sigma_{g_2}}\right) \Phi(\varphi_{1,g}(u))\, du.$$

of $\mathbf{A}_\rho^{-1}$ by a finite sum: $\tilde{\mathbf{A}}_\rho = \sum_{k=0}^{q} \rho^k \mathbf{W}^k$, $q < \infty$. We denote the corresponding function based on equation (12) as the quasi loglikelihood, $\tilde{\ell}(\boldsymbol{\theta}; \mathbf{y}, \mathbf{X})$ and its optimal vector $\tilde{\boldsymbol{\theta}}$ the Quasi–MLE. Details on the finite order approximation are given in Appendix D and Appendix E. The use of finite sum approximation for the inverses of either $\mathbf{A}_\rho$ and $\mathbf{B}_\lambda$ is not new in the literature and indeed some authors suggest it as a good practise to reduce the time of computation, see for instance Martinetti and Geniaux (2017). Despite this, no great attention has been given to its asymptotic behavior, relatively to the cross–sectional sample size n. Intuitively, if the number of terms q of the finite order approximation is large enough, the difference between the Partial–MLE and the Quasi–MLE is negligible. Asymptotically, this accounts to assuming q to increase with n at the proper rate. Conditions on this rate are given in Section 5, where we study the asymptotic behavior of both the Partial–MLE and Quasi–MLE.

3.2. 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

$$\mathbf{h}\left(\overset{\circ}{\boldsymbol{\theta}}\right) : \begin{cases} \rho = \underline{\omega}_\rho^{-1} + \frac{\overline{\omega}_\rho^{-1} - \underline{\omega}_\rho^{-1}}{1+\exp\left(-\overset{\circ}{\rho}\right)}, \\ \lambda = \underline{\omega}_\lambda^{-1} + \frac{\overline{\omega}_\lambda^{-1} - \underline{\omega}_\lambda^{-1}}{1+\exp\left(-\overset{\circ}{\lambda}\right)}, \\ \boldsymbol{\beta} = \mathbf{h}_\beta\left(\overset{\circ}{\boldsymbol{\beta}}\right), \quad \text{for} \quad j = 1, \ldots, n \end{cases} \tag{14}$$

where $\left(\underline{\omega}_\rho, \overline{\omega}_\rho\right)$ and $(\underline{\omega}_\lambda, \overline{\omega}_\lambda)$ 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

$$\overset{\circ}{\nabla}\left(\overset{\circ}{\boldsymbol{\theta}}; \mathbf{y}, \mathbf{X}\right) = \mathcal{J}\left(\overset{\circ}{\boldsymbol{\theta}}; \mathbf{y}, \mathbf{X}\right)' \nabla(\boldsymbol{\theta}; \mathbf{y}, \mathbf{X}) \tag{15}$$

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

$$\mathcal{J}\left(\overset{\circ}{\boldsymbol{\theta}}\right):\begin{cases}\mathcal{J}\left(\overset{\circ}{\rho}\right)=\dfrac{\left(\overline{\omega}_{\rho}^{-1}-\underline{\omega}_{\rho}^{-1}\right)\exp\left(-\overset{\circ}{\rho}\right)}{\left(1+\exp\left(-\overset{\circ}{\rho}\right)\right)^{2}},\\ \mathcal{J}\left(\overset{\circ}{\lambda}\right)=\dfrac{\left(\overline{\omega}_{\lambda}^{-1}-\underline{\omega}_{\lambda}^{-1}\right)\exp\left(-\overset{\circ}{\lambda}\right)}{\left(1+\exp\left(-\overset{\circ}{\lambda}\right)\right)^{2}},\\ \mathcal{J}\left(\overset{\circ}{\boldsymbol{\beta}}\right)=\mathcal{J}\left(\boldsymbol{\beta}\right),\quad \text{for}\quad j=1,\ldots,n.\end{cases} \tag{16}$$

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 (as well as of its *quasi* counterpart) 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. 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.

One way to minimize the loss of information, is to consider the partial loglikelihood functions corresponding to different pair choices, in order to subsequently obtain a single estimation by an efficient minimum distance procedure as suggested by Wang et al. (2013) in a similar framework. Since the number of possible ways to choose couples from $n = 2G$ units corresponding to different partial loglikelihood functions is huge[5], the number of partitions that can be considered is necessarily very small compared to it, thus the question of ranking the best partitions is extremely important.

A particular choice of G couples from $n = 2G$ units can be obtained with two dual procedures: either we keep the order of units fixed and pick without replacement two units at a time, thus obtaining $(i_{g,1}, i_{g,2})$, $g = 1, \ldots, G$, or we pick consecutive couples $(2g-1, 2g)$ from different permutations of the units. According to the latter approach, finding the best selection of couples amounts at finding the best permutation of units $(i_1, \ldots, i_n)$ relatively to a specified optimality criterion. In line with this, it is convenient to introduce the following notation. Let $\pi : \pi(1, \ldots, n) = (i_1, \ldots, i_n)$ be a permutation map. Each π defines a unique set of couples by

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

We further denote by $\mathbf{P}_\pi$ the permutation matrix corresponding to π, namely

$$\mathbf{P}_\pi = (\mathbf{e}_{\pi(1)}, \ldots, \mathbf{e}_{\pi(n)})',$$

[5]Specifically, it is $(2G-1)!! = (2G)!/G!2^G$, because partial loglikelihood is invariant under permutations of couples and changes of the order of units within each pair.

where $\mathbf{e}_j$ is the $j-$th canonical column vector. Thus, $\mathbf{P}_\pi$ transforms a vector $\mathbf{z} = (z_1, \ldots, z_n)'$ into $\mathbf{P}_\pi \mathbf{z} = (z_{\pi(1)}, \ldots, z_{\pi(n)})'$. With this notation, the reduced model in equation (3) can be rewritten as

$$\begin{aligned} \mathbf{P}_\pi \mathbf{y}^* &= \mathbf{P}_\pi \mathbf{A}_\rho^{-1} \mathbf{X}\boldsymbol{\beta} + \mathbf{P}_\pi \nu, \qquad \nu \sim \mathcal{N}_n\left(\mathbf{0}_n, \boldsymbol{\Sigma}\right) \\ \mathbf{P}_\pi \mathbf{y} &= \mathbb{I}_n\left(\mathbf{P}_\pi \mathbf{y}^* > \mathbf{0}_n\right) \end{aligned} \tag{17}$$

Note that from the assumptions of the model in equation (1) defined in Section 2, we obtain $\mathbf{P}_\pi \nu \sim \mathcal{N}(\mathbf{0}, \mathbf{P}_\pi \boldsymbol{\Sigma} \mathbf{P}'_\pi)$, 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 of $\mathbf{P}_\pi \boldsymbol{\Sigma} \mathbf{P}'_\pi$[6].

We propose a criterion that gives us a (not necessarily unique) permutation map π^* solving a minimum KL–divergence problem. In order to explain the procedure, let us denote by P_θ the probability distributions of the $n-$tuple $(y_1, \ldots, y_n)$ (conditional on X), from model in equation (1). Using the notation introduced in Theorem 3.1, we get

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

Similarly, let $P_\theta^\pi = p_{1,\theta}^\pi \times p_{2,\theta}^\pi \times \cdots \times p_{G,\theta}^\pi$, where each $p_{g,\theta}^\pi$, consistently with equation (11), 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 denote by P_0 and P_0^π the probability distributions corresponding to $\boldsymbol{\theta} = \boldsymbol{\theta}_0$. Our idea is to find a permutation that minimizes the KL divergence between P_0^π and the true probability distribution P_0 of the whole vector $\mathbf{P}_\pi \mathbf{y}$, namely that minimizes:

$$KL(P_0^\pi \| P_0) = \sum_{\mathbf{d} \in \{0,1\}^n} P_0^\pi(\mathbf{d}) \log \frac{P_0^\pi(\mathbf{d})}{P_0(\mathbf{d})}, \tag{18}$$

over all possible permutations π.

Since the computation of the term $\Pr(\mathbf{y} = \mathbf{d} \mid \mathbf{X}) = P_0(\mathbf{d})$ involves a $n-$dimensional integration, 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 model (1)), respectively. Let $\mathcal{P}_n$ be the set of all permutations of n units corresponding to distinct bivariate distributions. Our algorithm is based on the following result:

[6]The matrix $\boldsymbol{\Sigma}_\pi$ can be written compactly as

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

where $\mathbf{E}_g$ 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.

Theorem 4.1. *(i) For every $\pi \in \mathcal{P}_n$ and $\boldsymbol{\theta} \in \Theta$,*

$$KL(P_\theta^\pi || P_\theta) \leq KL(f_\theta^\pi || f_\theta). \tag{19}$$

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

$$\arg\min_{\pi} KL(f_\theta^\pi || f_\theta) = \arg\min_{\pi \in \mathcal{P}} \sum_{g=1}^{G} (b(\pi(2g-1), \pi(2g)) - \log(\bar{\sigma}(\pi(2g-1), \pi(2g))) \tag{20}$$

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 suggests the following procedure based on the solution of a maximum weighted matching problem in a general graph:

Step 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): $(\tilde{\rho}, \tilde{\lambda})$ and compute $\tilde{\boldsymbol{\Sigma}}$ from it

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

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

Stap 4. Use Edmonds' *blossom algorithm* (see Galil (1986) and references therein) for the computation of the maximum weighted matching[7]

The procedure introduced in this section is a way for controlling the information loss, which tends to be higher: (i) when the weight matrix is dense; (ii) for large values of ρ (in absolute value). For this reason, we expect the use of the algorithm to improve the estimation in one of those two cases.

5. Asymptotics

In this section we study the asymptotic properties of the QML/PML estimators of the SARAR(1,1)–probit model. The asymptotic analysis performed here enters in the context of the increasing domain asymptotics, consistently with the literature. In order to prove consistency and asymptotic normality of the QML estimator, we shall need the consistency and asymptotic normality of the PML estimator, plus some condition on the rate of the sequence q_n of finite terms in the approximation of $\mathbf{A}_\rho^{-1}$ or $\mathbf{B}_\lambda^{-1}$, thus ensuring asymptotic equivalence of the QML and PML estimators.

Following the notation introduced in Section 3, let

$$\tilde{\boldsymbol{\theta}}_n = \arg\max_{\boldsymbol{\theta} \in \boldsymbol{\Theta}} \tilde{\ell}_n(\boldsymbol{\theta}; \mathbf{y}, \mathbf{X}),$$

[7] A maximum weighted matching is the set of edges of a graph, with no nodes in common, that maximizes the total weights.

where $\tilde{\ell}_n$ is the *quasi*-loglikelihood defined there. Moreover, to discriminate between the exact bivariate probabilities and those based on the $q-$th order finite sum approximation, we denote the last by $\tilde{p}_g(d_1, d_2)$, and the former by $p_g(d_1, d_2)$[8]. There is the following link between the $\tilde{\ell}_n$ and ℓ_n:

$$\begin{aligned}\mathrm{E}\left(\tilde{\ell}_n(\boldsymbol{\theta}; \mathbf{y}, \mathbf{X}) \mid \mathbf{X}\right) &= \sum_g \mathrm{E}\left(\sum_{d=(d_1,d_2)} \frac{1}{G}\mathbb{I}\{y_{g_1} = d_1, y_{g_2} = d_2\} \log \frac{\tilde{p}_{g,}(d_1, d_2)}{p_{g,}(d_1, d_2)} \mid \mathbf{X}\right) + \mathrm{E}\left(\ell_n(\boldsymbol{\theta}; \mathbf{y}, \mathbf{X}) \mid \mathbf{X}\right) \\ &= \mathrm{E}\left(\ell_n(\boldsymbol{\theta}; \mathbf{y}, \mathbf{X}) \mid \mathbf{X}\right) - \frac{1}{G}\sum_g KL(p_{g,}||\tilde{p}_{g,}).\end{aligned} \tag{21}$$

Thus, consistency and asymptotic normality come from the analogous properties of the PML estimator, and from negligibility of the term $\frac{1}{G}\sum_g KL(p_{g,}\|\tilde{p}_{g,})$.

In line with Wang et al. (2013), we need to add the following assumptions.

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

Assumption 5. *The density of observations in any region whose area exceeds a fixed minimum is bounded. Moreover,*

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

Assumption 6. *$\sup_{n,g,h} |\mathbb{C}ov(y_{gi}, y_{hi}| \leq \alpha(d_{gh})$, where d_{gh} is the distance between group 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) $\lim_{n\to\infty} \log n/q_n = 0$.*

Assumptions 4–6 are taken from Wang et al. (2013) and are used to prove consistency of the PML estimator. The first is a standard assumption for M–type estimators and also gives an identification condition. Assumption 5 is the same as (iv) of Theorem 1 in Wang et al. (2013) and it rules out the possibility that, for some couples, one (or more) of the 4 outcomes has conditional probability equal to zero. Assumption 6 is the *mixing* condition given in Wang et al. (2013), ensuring that dependence between observations rapidly decays with their distance. Finally, Assumption 7 is necessary for the asymptotic behavior of the QML estimator. 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 requires $q_n = O(n^\varepsilon)$ for some $\varepsilon > 0$.

[8]Note that both probabilities follow equations (11), the only difference being on the computation of terms $\mathbf{X}_\rho$ and of the elements of $\boldsymbol{\Sigma}_g$.

Theorem 5.1. *Under Assumptions 1–6, then,* $\|\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0\| = o_p(1)$. *If further Assumption 7(a) holds, then,* $\|\tilde{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0\| = o_p(1)$.

In order 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*

$$J(\boldsymbol{\theta}_0) = \lim_n G_n E\left(\frac{\partial \ell_n}{\partial \boldsymbol{\theta}}(\boldsymbol{\theta}_0)\frac{\partial \ell_n}{\partial \boldsymbol{\theta}'}(\boldsymbol{\theta}_0)\right)$$

and

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

are positive definite.

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

$$\sqrt{n}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0) \to \mathcal{N}\left(0, \mathbb{H}(\boldsymbol{\theta}_0)^{-1} J(\boldsymbol{\theta}_0)\mathbb{H}(\boldsymbol{\theta}_0)^{-1}\right) \tag{22}$$

If moreover Assumption 7 (a)-(b) holds, then

$$\sqrt{n}(\tilde{\boldsymbol{\theta}} - \boldsymbol{\theta}_0) \to \mathcal{N}\left(0, \mathbb{H}(\boldsymbol{\theta}_0)^{-1} J(\boldsymbol{\theta}_0)\mathbb{H}(\boldsymbol{\theta}_0)^{-1}\right) \tag{23}$$

Assumptions 8–10 are those used by Wang et al. (2013) to prove Theorem 2. Assumption 10 is quite standard in a MLE framework, while 8 and 9 are necessary to apply Bernstein's blocking method, used in McLeish's central limit theorem for dependent processes, see McLeish (1974).

Consistent estimation of $\mathbb{H}(\boldsymbol{\theta}_0)$ and $J(\boldsymbol{\theta}_0) = \lim_n G_n \mathrm{E}[\nabla(\boldsymbol{\theta}_0)\nabla(\boldsymbol{\theta}_0)']$, yelds a consistent estimator for the covariance matrix of $\hat{\boldsymbol{\theta}}$, In particular, $\mathbb{H}(\boldsymbol{\theta}_0)$ can be estimated through the average of the negative Hessian matrix at $\hat{\boldsymbol{\theta}}$ or $\tilde{\boldsymbol{\theta}}$. One possible way to estimate $J(\boldsymbol{\theta}_0)$ is to use the approach by Conley (1999) and its adaptation to non–stationary spatial processes by Kelejian and Prucha (2007). The resulting estimator for $J(\boldsymbol{\theta}_0)$ is basically the same as Theorem 3 in Wang et al. (2013), under conditions that are a simple modification of those therein. Another estimator could be obtained from the computation of $\frac{1}{G}\sum_{g=1}^{G} \mathrm{E}\left[\nabla^g(\hat{\boldsymbol{\theta}})\nabla^g(\hat{\boldsymbol{\theta}})'\right]$, by using the explicit formulas for $\boldsymbol{\nabla}^g(\boldsymbol{\theta})$ given in the Appendix. A third approach, that is the one we follow here, consists in a parametric bootstrap estimation: given $\hat{\boldsymbol{\theta}}$, we resample iid errors and use the estimated reduced latent model (3) to generate the latent variables $\mathbf{y}_b^*$ and the corresponding bootstrap sample $(\mathbf{y}_b, \mathbf{X})$, for $b = 1, \ldots, B$. We finally estimate the covariance matrix of $\hat{\boldsymbol{\theta}}$ through the empirical covariance matrix of the bootstrap estimates $\hat{\boldsymbol{\theta}}_b$ (see Section 8).

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), 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 $\mathbf{x}_{\cdot h} = (x_{1h}, x_{2h}, ..., x_{ih}, ..., x_{nh})'$ an n–dimensional vector of units referred to the h–th regressor, $h = 1, \ldots, k$, and $\mathbf{x}_{i\cdot} = (x_{i1}, x_{i2}, ..., x_{ih}, ..., x_{ik})'$ a k–dimensional vector of regressors referred to unit i. By considering equations (4) and (5), we propose the following specifications of the marginal effects

$$\begin{aligned}
&\frac{\partial \mathrm{P}\left(y_i = 1 \mid \mathbf{x}'_{i\cdot}, \sum_j w_{ij} y_j^*\right)}{\partial \mathbf{x}'_{\cdot h}} \mid_{\bar{\mathbf{x}}} = \phi\left(\{\boldsymbol{\Sigma}_{\nu(\rho,\lambda)}\}_{ii}^{-1/2} \left\{\mathbf{A}_\rho^{-1}\bar{\mathbf{X}}\right\}_{i\cdot} \boldsymbol{\beta}\right) \{\boldsymbol{\Sigma}_{\nu(\rho,\lambda)}^{-1/2}\}_{ii} \{\mathbf{A}_\rho^{-1}\}_{i\cdot}\beta_h \\
&\frac{\partial \mathrm{P}\left(y_i = 1 \mid \mathbf{x}'_{i\cdot}, \sum_j w_{ij} y_j^*\right)}{\partial \mathbf{x}'_{\cdot h}} \mid_{\mathbf{x}} = \phi\left(\{\boldsymbol{\Sigma}_{\nu(\rho,\lambda)}\}_{ii}^{-1/2} \left\{\mathbf{A}_\rho^{-1}\mathbf{X}\right\}_{i\cdot} \boldsymbol{\beta}\right) \{\boldsymbol{\Sigma}_{\nu(\rho,\lambda)}^{-1/2}\}_{ii} \{\mathbf{A}_\rho^{-1}\}_{i\cdot}\beta_h
\end{aligned} \tag{24}$$

where $\boldsymbol{\Sigma}_{\nu(\rho,\lambda)}$ is the variance–covariance matrix implied by the reduced form of a SARAR(1,1)–probit model and $\boldsymbol{\Sigma}_{\nu(\rho,\lambda)}^{-1/2} = \left\{{\sigma_{\nu_i}}^{-1}\right\}$, $\mathbf{A}_\rho^{-1} = (\mathbf{I} - \rho\mathbf{W})^{-1}$, $\bar{\mathbf{X}}$ is an n by k matrix of regressor–means, $(\,\cdot\,)_{i\cdot}$ considers the i–th row of the matrix inside, and $(\,\cdot\,)_{ii}$ the i–th diagonal element of a square matrix. Note that $\boldsymbol{\Sigma}_{\nu(\rho,\lambda)}$ reduces to $\boldsymbol{\Sigma}_{\mathbf{u}(\rho)}$ for a SAR(1)–probit specification as in equation (6) with $\mathbf{u} = \mathbf{A}_\rho^{-1}\boldsymbol{\varepsilon}$.

The first specification of equations (24) explains the impact of a marginal change in the mean of the h–th regressor, i.e. $\bar{\mathbf{x}}_{\cdot h}$, on the conditional probability of $\{y_i = 1\}$, i.e. $\mathrm{P}\left(y_i = 1 \mid \mathbf{x}'_{i\cdot}, \sum_j w_{ij} y_j^*\right)$, setting $\bar{\mathbf{x}}_{\cdot h'}$ for all the remaining regressors, $h' = 1, \ldots, k-1$. The second specification of equations (24) considers, instead, the marginal impact evaluated at each single value of $\mathbf{x}_{\cdot h}$. 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 due to 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 their own regions). 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 them (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 $n-1$ other 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 in equation (2). 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* due to the variation over space of the marginal impacts with respect to the spatial distribution of the regressors[9]. Within nonlinear models, the possibility of evaluating a marginal impact with respect to a particular value $\mathbf{x}_{ih}$ have the same meaning of considering a marginal impact in a particular region/site for regressor h. 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 a SAR(1)–probit model specified in equation (6) and a SARAR(1,1)-probit model specified in equation (1). 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 $\mathbf{X} = [\mathbf{x}_{\cdot 0}, \mathbf{x}_{\cdot 1}, \mathbf{x}_{\cdot 2}]$ of dimension $n \times 3$, which is composed by two regressors $\mathbf{x}_{\cdot 1}$, $\mathbf{x}_{\cdot 2}$ and a constant $\mathbf{x}_{\cdot 0}$, with $\mathbf{x}_{ij} = (\mathbf{x}_{1j}, \mathbf{x}_{2j}, \ldots, \mathbf{x}_{nj})'$ and $j = 0, 1, 2$. The regressor $\mathbf{x}_{\cdot 1}$ is drawn from a $\mathcal{U}(-1, 1)$ distribution and $\mathbf{x}_{\cdot 2}$ 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 non–negative weight and then normalized n–dimensional weight matrix. Finally, the number of simulation runs are 1000 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 $n = 900$, (c) 50×50 with $n = 2500$. The latter is an inverse distance–based matrix built on randomly generated coordinates from $\mathcal{U}(0, 50)$ and $\mathcal{U}(-70, 20)$ of the same dimension n as before. 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.

[9] See Billé et al. (2017) for a two–step approach specifically thought to account for unobserved discrete spatial heterogeneity in the beta's coefficients via iterated local estimation procedures.

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 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). 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[10] 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 (i.e. spectral radius), in order to ensure: (i) a proper parameter space for ρ (see lemma 2.1), (ii) the equivalence of the spatial models before and after normalization of the weights.

7.2. Finite sample results: SAR model

In this Section we show the finite sample properties of the PMLE and the marginal effects calculated as in equation (24). 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, 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 unbiassedness 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 from negative values to the positive ones. 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 equations (24), with

[10]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.

respect to the mean value and to each observation, respectively. In both cases, 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 due to differences in the indirect effects.

Subsections 7.2.1 and 7.2.2 discuss two issues, related to modeling and to the estimation procedure, respectively. Specifically, in Subsection 7.2.1 we try to measure the effect that model misspecification has on the estimates of both parameters and marginal effects, when misspecification is induced by an incorrect choice of the weight matrix. The second issue concerns the loss of information due to the approximation of the exact loglikelihood by partial loglikelihood, which is naturally expected to increase as the weight matrix becomes more dense. In Subsection 7.2.2 we thus investigate the ability of the algorithm given in Section 4 to mitigate the effect of this information loss.

7.2.1. Misspecification of 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 $\rho = (-0.6, 0.6)$ 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 ($\mathbf{W}_{sn}$) case. Whereas, the assumed sparse weighting matrix is based on a k–nn approach, with $k = 11$ as before (W_{knn}).

Results are reported in Tables A.4, A.5 and Figure B.2. Table A.4 shows that the PML estimator of the $\boldsymbol{\beta}$ coefficients is 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 due to 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 seems to be no significant differences in terms of the distributions when considering the two type of normalization rules, i.e. $\mathbf{W}_{sn}$ and $\mathbf{W}_{rn}$. 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[11]. Data are simulated from a SAR(1)–probit with $\boldsymbol{\beta} = (0, 1, -0.5)'$ and $\rho = 0.6$, using either a k–nn

[11] We use the R library `Rpython` (`https://cran.r-project.org/web/packages/rPython/index.html`) to run a program using the function `networkx.max_weight_matching` from the package `networkx` (`https://pypi.python.org/pypi/networkx/2.0`), which is a Python package for the creation and manipulation of graphs and networks.

matrix with $k = 11, 25, 50, 100$ or an inverse distance matrix. We compute both the PML and QML 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/QML 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, in order to figure out 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 more dense. 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, *sd* of both estimators increases rapidly[12].

Table A.3 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 there is a smaller increase of negative bias[13], with an overall variation of RMSE of -30%. There is a slight improvement in the *sd* of the $\hat{\beta}$'s and no effect on their means. The initial guess $\tilde{\rho}$ appears to have a negligible effect.

7.3. Finite sample results: SARAR 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 sample size ($n = 900$). We draw samples from 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 number of 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 weight matrices prevents possible misbehavior of the estimator due to identifiability issues, it being understood that the $\boldsymbol{\beta}$ coefficients are significant[14].

Table A.6 presents the results, for different values of the parameter λ, namely $\lambda = (0.8, 0.6, 0.4, 0.2)$. Similarly 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 only 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

[12] The tables are available upon request.

[13] This seems to be a consequence of a better behavior of the loglikelihood function that reduces drastically the occurrence of an optimum value $\hat{\rho}$ near the boundary ($\hat{\rho} \approx 1$).

[14] In the nonlinear case, it is possible that the two requirements must be contemporaneously satisfied to ensure identifications.

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 get worse in terms of RMSE (mainly due to 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[15].

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 QMLE. The model specification is referred to a SAR(1)–probit in equation (6). 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* and details are referred to LeSage et al. (2011). We have found some points/units to have 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 mechanmism 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 to 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, (c) 0–12 months. In each time horizon 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 size firms, with medium size firms representing the omitted class, two dummy variables reflecting low and high socio–economic class of the store *clientèle* (with the middle socio-economic 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).

The weighting matrix is built on a k-nn criterion with $k = 11$ for time horizon (a) and $k = 15$ for time horizons (b), (c). We obtain standard errors of our PML estimates by sampling from the latent variable

[15] Results available upon request.

distributions in the estimated reduced form model in equation (3), and then define the corresponding sample of binary variables $\mathbf{y}_b$, $b = 1, \ldots, B$, $B = 1000$. For each sample $(\mathbf{y}_b, \mathbf{X})$, a new vector of PML estimates $\hat{\boldsymbol{\theta}}_b$ is computed and its distribution over the $B = 1000$ samples is used to calculate the standard errors.

Table A.7 shows the PML estimates and their standard errors to be compared with those in Table 3 in LeSage et al. (2011). Table A.8 provides marginal effects, see equations (24), for each time horizon to be compared with the effects reported in Tables 4,5,6 in LeSage et al. (2011). All tables show that PML estimates are consistent with Bayesian estimates; in particular, the PLM estimate of the spatial correlation coefficient ρ is positive and significant, and 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).

9. Conclusions

In this paper we derive the asymptotic properties and evaluate the finite sample properties of a Partial Maximum Likelihood Estimator (Partial–MLE) for Spatial Autoregressive Probit Model with Autoregressive disturbances (SARAR(1,1)–probit model). The work is mainly based on the paper proposed by Wang et al. (2013), although substantial differences can be found. In our paper we consider the more general and interesting case of correlation among the dependent variables, which specifies at least the SAR(1)–probit rather than a simple SAE(1)–probit model. In addition, we propose a Kullback–Leibler approach for choosing the couples that maximize the partial log–likelihood function and we suggest exact formulas for defining the marginal effects in spatial binary contexts. Finally, we derive explicit expressions of the score vector, which can also can be used in the approach of Mozharovskyi and Vogler (2016), to improve computations.

The Partial–MLE and the Quasi–MLE are both asymptotically consistent given some regularity conditions. Our simulations suggest that the estimator performs well even with small sample sizes. The results rapidly and substantially improve as the sample size increases both in terms of unbiasedness and consistency. All the distributions are, moreover, bell–shaped from moderate-to-large samples and for all the true values of correlations. 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 to the true values. Also in the SARAR(1,1)–probit case, the estimator performs reasonably well, although there is a loss in efficiency in particular for $\hat{\rho}$ and $\hat{\beta}_0$.

The estimator is also computationally efficient, giving the opportunity to estimate the model even with large data set. Finally, in our empirical application we use a data set freely available in the R package. Results suggest that our Partial ML estimator gives parameter estimates and standard deviations quite similar to those obtained by the Bayesian approach typically used for this type of models. We consider two cases of model misspecification due to a different assumed weighting matrix: in both these cases the estimator properties and the direct marginal effects are quite robust in terms of the beta coefficients. This analysis on the other hand

confirms that an incorrect choice of $\mathbf{W}_n$ has a great impact on the estimation of ρ and, as a consequence, of the indirect effects, thus suggesting that great care must be paid to model selection.

Finally, the criterion proposed for the choice of couples deserves further investigation since it proves to be a viable and promising method to approximate a complex model with a simple one, with a limited loss in efficiency and accuracy.

Acknowledgements

The first Author is indebted to Professor Paul Elhorst for having emphasized the inconsistency problem during a brief discussion of a previous paper and for encouraging her to explore more this issue. The Authors wish to thank all the participants to the 9th Jean Paelinck Seminar on "Non–parametric Spatial Econometrics: Theory and Empirical Issues" for their insightful comments. This research has been carried out within the project "AIDA"from Free University of Bozen–Bolzano and within the project "Estimation of transformation models" at Bocconi Institute for Data Science and Analytics.

References

Takeshi Amemiya. The maximum likelihood and the nonlinear three-stage least squares estimator in the general nonlinear simultaneous equation model. *Econometrica: Journal of the Econometric Society*, pages 955–968, 1977.

Takeshi Amemiya. The estimation of a simultaneous equation generalized probit model. *Econometrica: Journal of the Econometric Society*, pages 1193–1205, 1978.

Takeshi Amemiya. *Advanced Econometrics*. Harvard university press, 1985.

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

Giuseppe Arbia. Pairwise likelihood inference for spatial regressions estimated on very large datasets. *Spatial Statistics*, 7 (Supplement C):21–39, 2014. ISSN 2211–6753.

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

Badi H. Baltagi, Peter H. Egger, and Michaela Kesina. *Bayesian Spatial Bivariate Panel Probit Estimation*, chapter 4, pages 119–144. 2017. doi: 10.1108/S0731-905320160000037011.

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

Kurt J. Beron and Wim P. M. Vijverberg. *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, 2004.

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

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

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

Chandra R Bhat. 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, 2011.

Anna Gloria Billé. 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, 2014.

Anna Gloria Billé, Roberto Benedetti, and Paolo Postiglione. A two–step approach to account for unobserved spatial heterogeneity. *Spatial Economic Analysis*, 0(0):1–20, 2017. doi: 10.1080/17421772.2017.1286373.

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

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

Leopoldo Catania and Anna Gloria Billé. Dynamic spatial autoregressive models with autoregressive and heteroskedastic disturbances. *Journal of Applied Econometrics*, 2017.

Andrew David Cliff and J Keith Ord. *Spatial processes: models & applications*. Taylor & Francis, 1981.

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

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

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

Zvi Galil. Efficient algorithms for finding maximum matching in graphs. *ACM Comput. Surv.*, 18(1):23–38, March 1986. doi: 10.1145/6462.6502.

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

James Douglas Hamilton. *Time series analysis*, volume 2. Princeton university press Princeton, 1994.

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

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

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

Harry H Kelejian. 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, 2016.

Harry H Kelejian and Ingmar R Prucha. 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, 1998.

Harry H Kelejian and Ingmar R Prucha. A generalized moments estimator for the autoregressive parameter in a spatial model. *International economic review*, 40(2):509–533, 1999.

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

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

Harry H Kelejian, Ingmar R Prucha, and Yevgeny Yuzefovich. 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, 2004.

Thomas Klier and Daniel P McMillen. 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, 2008.

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

Dayton M Lambert, Jason P Brown, and Raymond JGM Florax. 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, 2010.

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

Lung-Fei Lee. Asymptotic distributions of quasi-maximum likelihood estimators for spatial autoregressive models. *Econometrica*, pages 1899–1925, 2004.

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

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

James P LeSage, R Kelley Pace, Nina Lam, Richard Campanella, and Xingjian Liu. 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, 2011.

JP LeSage and R Kelley Pace. Introduction to spatial econometrics. *Boca Raton, FL: Chapman & Hall/CRC*, 2009.

Charles F. Manski. *Alternative Estimators and Sample Designs for Discrete Choice Analysis*. The MIT Press, 1981.

Davide Martinetti and Ghislain Geniaux. Approximate likelihood estimation of spatial probit models. *Regional Science and Urban Economics*, 64:30–45, 2017.

Daniel McFadden. Economic choices. *American Economic Review*, pages 351–378, 2001.

D. L. McLeish. Dependent central limit theorems and invariance principles. *Ann. Probab.*, 2(4):620–628, 08 1974. doi: 10.1214/aop/1176996608. URL `http://dx.doi.org/10.1214/aop/1176996608`.

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

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

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

Keith Ord. Estimation methods for models of spatial interaction. *Journal of the American Statistical Association*, 70(349): 120–126, 1975.

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

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

Dale J Poirier and Paul A Ruud. Probit with dependent observations. *The Review of Economic Studies*, 55(4):593–614, 1988.

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

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

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

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

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

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

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

Adonis Yatchew and Zvi Griliches. Specification error in probit models. *The Review of Economics and Statistics*, pages 134–139, 1985.

Appendix A. Tables

	n = 100					**n = 900**					**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 for the QML estimates of 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 $\hat{\rho}$. $\bar{\mathbf{X}}$ is referred to marginal impacts as in equation (27), and $\mathbf{X}$ as in equation (28). 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 $\mathbf{x}_{\cdot 1} \sim \mathcal{U}(-1, 1)$, $\mathbf{x}_{\cdot 2} \sim \mathcal{N}(0, 1)$.

$n = 900$	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.3: Summary statistics for the PML estimates of 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).

	β_0	β_1	β_2	ρ	β_0	β_1	β_2	ρ
True Matrix/Value	0	1	-0.5	0.6	0	1	-0.5	-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 for the QML estimates of 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
$\mathbf{x}_{\cdot 1}$																
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		
$\mathbf{x}_{\cdot 2}$																
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		
$\rho = -0.6$																
$\mathbf{x}_{\cdot 1}$																
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		
$\mathbf{x}_{\cdot 2}$																
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 $\mathbf{x}_{\cdot 1} \sim \mathcal{U}(-1, 1)$, $\mathbf{x}_{\cdot 2} \sim \mathcal{N}(0, 1)$.

$n = 900$

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.6: Summary statistics for the QML estimates of 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 Queen adjacency matrix. 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				Second				Third			
Regressors	Bayes	*sd*	PMLE	*sd*	Bayes	*sd*	PMLE	*sd*	Bayes	*sd*	PMLE	*sd*
constant	-7.616	2.595	-5.272	3.246	-2.978	2.730	-2.069	3.762	-4.336	2.723	-2.198	3.523
flood depth	-0.168	0.044	-0.136	0.062	-0.110	0.035	-0.112	0.095	-0.089	0.034	-0.102	0.097
log(median income)	0.733	0.252	0.510	0.319	0.311	0.268	0.238	0.368	0.484	0.268	0.287	0.345
small size	-0.276	0.140	-0.340	0.163	-0.109	0.149	-0.223	0.179	-0.214	0.154	-0.240	0.192
large size	-0.329	0.321	-0.361	0.368	-0.372	0.332	-0.442	0.433	-0.357	0.298	-0.424	0.435
low status customers	-0.329	0.166	-0.453	0.186	-0.342	0.161	-0.446	0.198	-0.321	0.162	-0.512	0.219
high status customers	0.085	0.131	0.034	0.149	0.041	0.153	-0.006	0.156	-0.101	0.165	-0.241	0.175
sole proprietorship	0.551	0.196	0.560	0.236	0.359	0.181	0.289	0.261	0.146	0.189	0.078	0.298
national chain	0.068	0.378	0.059	0.412	0.295	0.381	-0.099	0.443	-0.120	0.389	-0.621	0.498
Wy	0.382	0.094	0.515	0.158	0.578	0.084	0.621	0.146	0.584	0.093	0.664	0.130

Table A.7: Estimates and standard errors 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. Mean and *sd* are the mean and standard deviations based on 1000 different binary vectors, by drawing a different vector of innovations ε from a standard normal distribution.

	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.8: Marginal Effects respect to $\mathbf{X}$ 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.

Appendix B. Figures

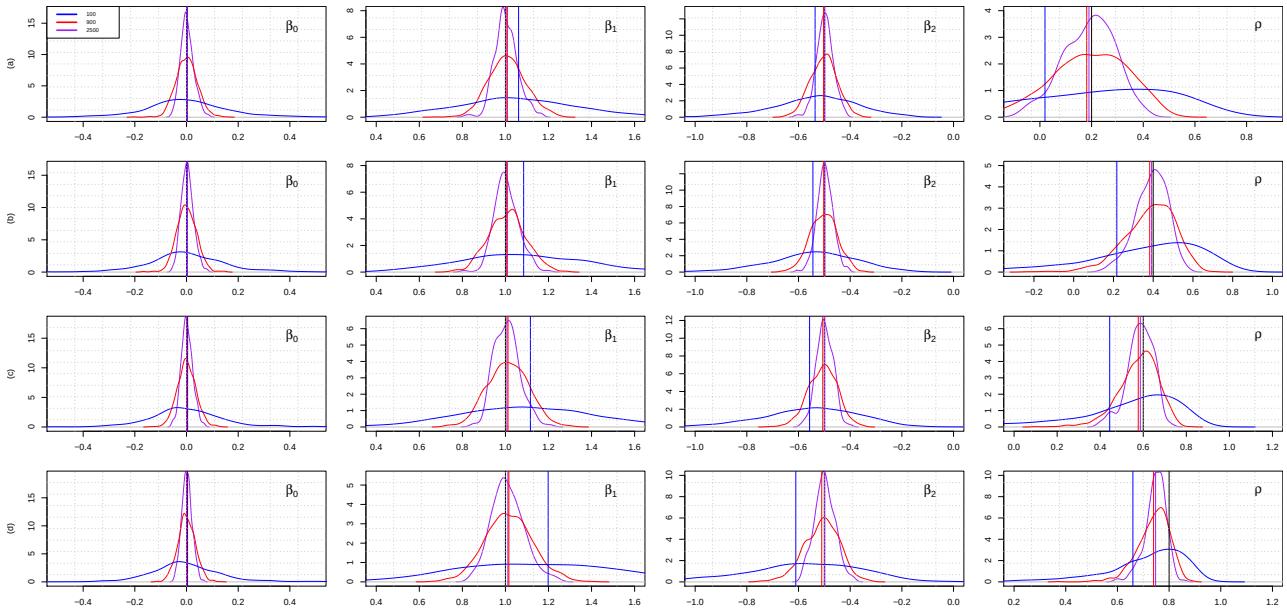

Figure B.1: Gaussian Kernel density for the PML estimated coefficients of the SAR(1)–probit model for different true values of ρ: (a) $\rho = 0.2$, (b) $\rho = 0.4$, (c) $\rho = 0.6$ (d) $\rho = 0.8$. The sample sizes are 100 (in blue), 900 (in red) and 2500 (in purple), while blue, red and purple vertical lines are the mean values, respectively. Vertical black lines are the true values of the parameters. The number of Monte Carlo replications are fixed to 1,000.

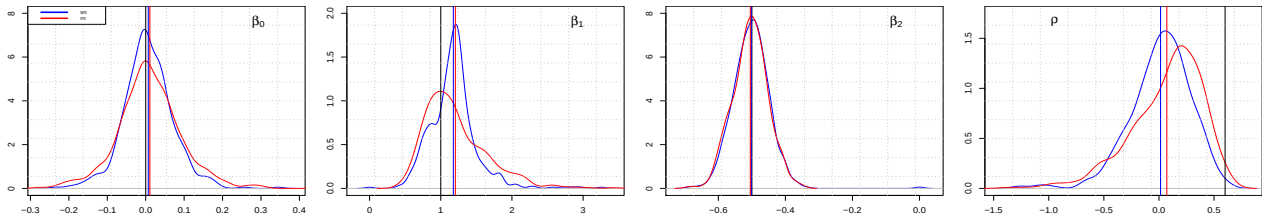

Figure B.2: Gaussian Kernel density for the PML estimated coefficients of the SAR(1)–probit model when $\mathbf{W}_n$ is misspecified. Two cases of misspecification: (i) $\mathbf{W}_{true} = \mathbf{W}_{sn}$ (in blue), (ii) $\mathbf{W}_{true} = \mathbf{W}_{rn}$ (in red). The assumed weighting matrix is $\mathbf{W}_{k-nn}$, $n = 900$ 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.

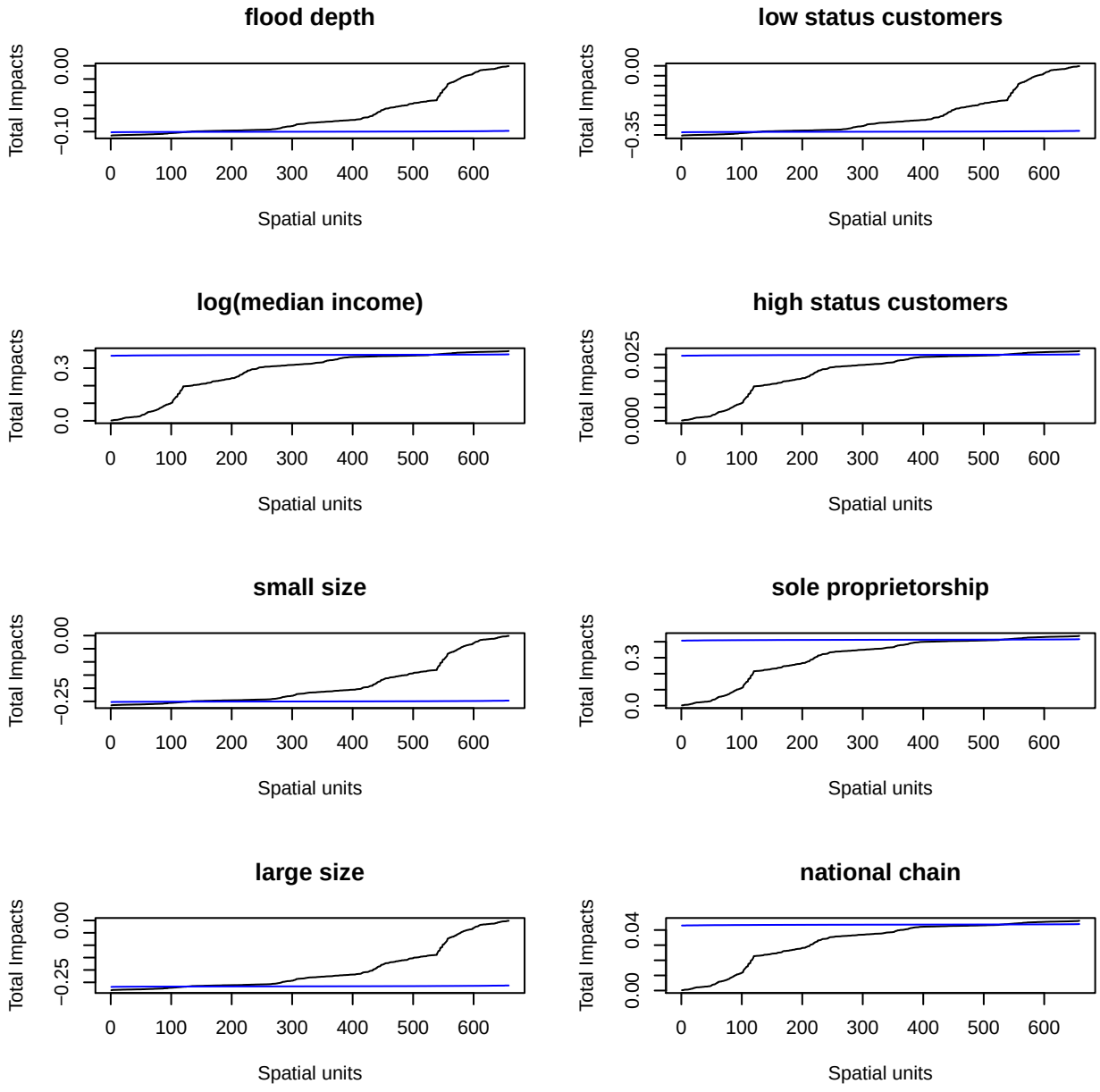

Figure B.3: Spatial heterogeneity of the total marginal impacts for each regressor during the first time horizon. Blue lines represent marginal impacts relative to $\hat{\mathbf{X}}$.

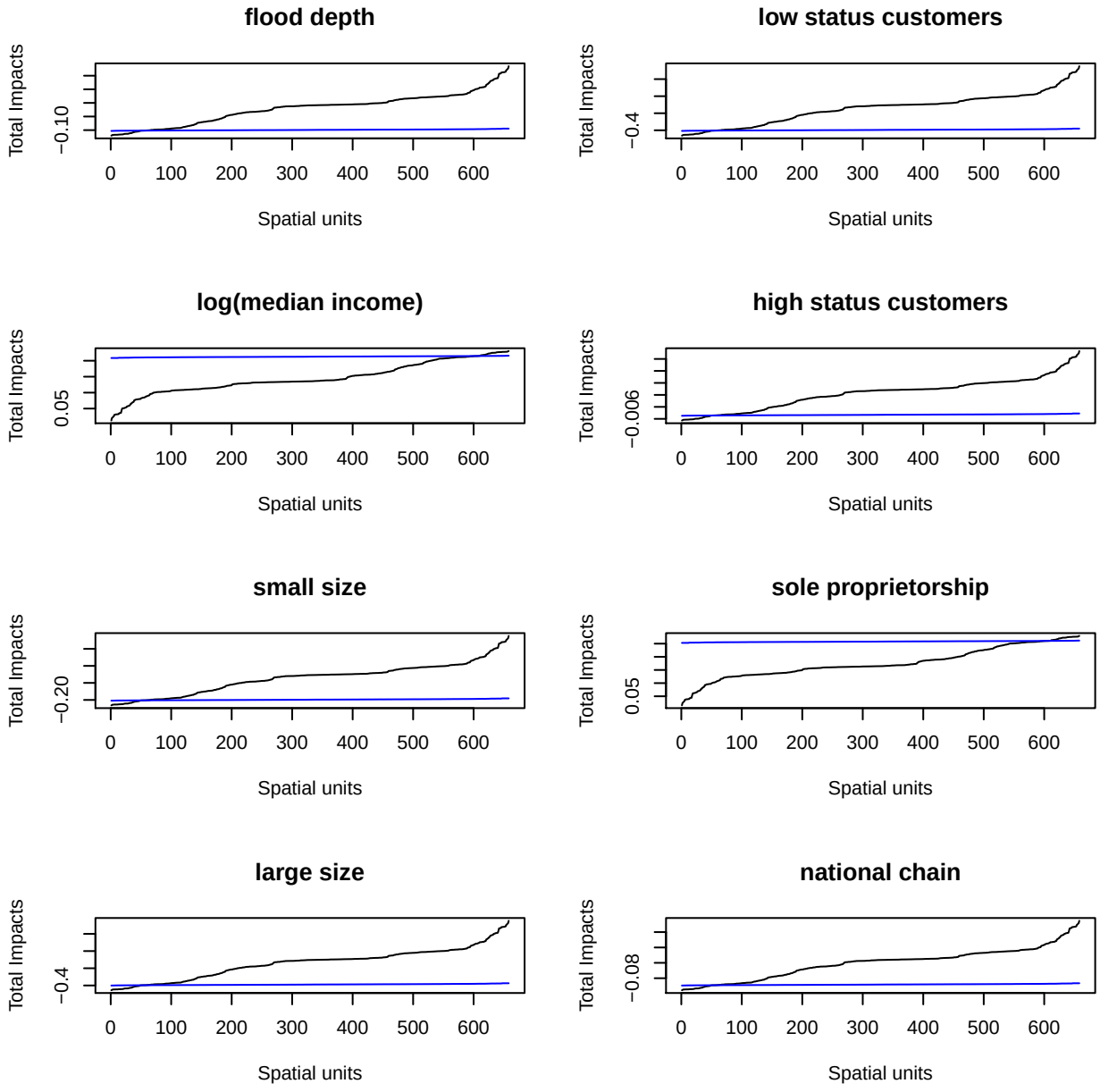

Figure B.4: Spatial heterogeneity of the total marginal impacts for each regressor during the second time horizon. Blue lines represent marginal impacts relative to $\hat{\mathbf{X}}$.

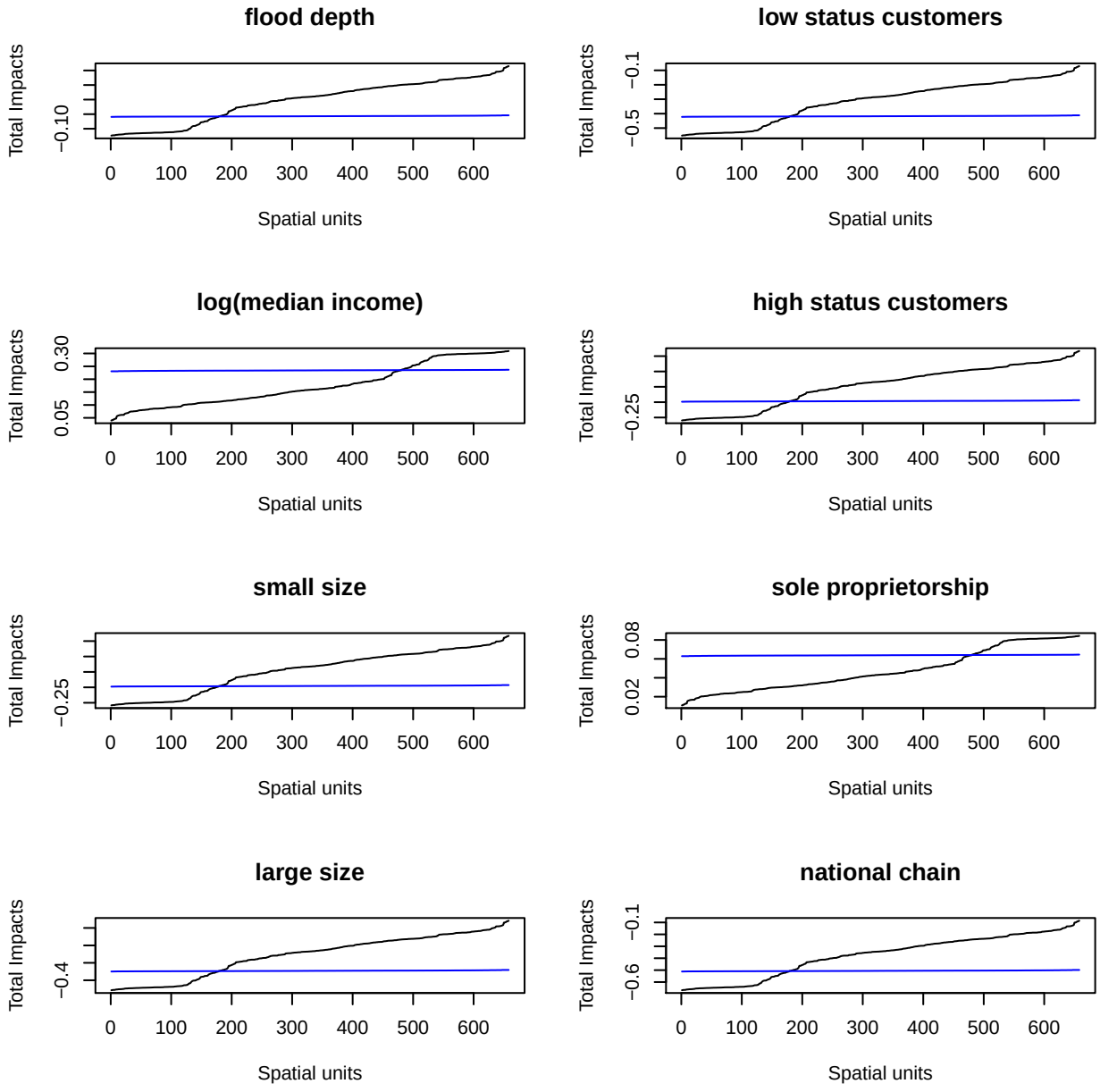

Figure B.5: Spatial heterogeneity of the total marginal impacts for each regressor during the third time horizon. Blue lines represent marginal impacts relative to $\hat{\mathbf{X}}$.

Appendix C. Proof of Theorems

Proof of Theorem 3.1

Repeating the steps in Wang et al. (2013), we can easily get

$$\begin{aligned} p_g(1,1) &= P\left(y_{g_1}=1, y_{g_2}=1 \mid \mathbf{X}\right) = P(y_{g_1}=1 \mid \mathbf{X}) P\left(y_{g_2}=1 \mid y_{g_1}=1, \mathbf{X}\right) \\ &= \Phi\left(\frac{\mathbf{x}_{\rho,g_1}\boldsymbol{\beta}}{\sigma_{g_1}}\right) \mathrm{E}_{u_{g_1}}\left(P\left(y_{g_2}=1 \mid \mathbf{X}, u_{g_1}\right) \mid y_{g_1}=1, \mathbf{X}\right) \\ &= \Phi\left(\frac{\mathbf{x}_{\rho,g_1}\boldsymbol{\beta}}{\sigma_{g_1}}\right) \mathrm{E}_{u_{g_1}|\mathbf{X},y_{g_1}=1}\left[\Phi\left(\frac{\mathbf{x}_{\rho,g_2}\boldsymbol{\beta}+\frac{\sigma_{g_1,g_2}}{\sigma^2_{g_1}}u_{g_1}}{\sqrt{\sigma^2_{g_2}-\frac{\sigma^2_{g_1,g_2}}{\sigma^2_{g_1}}}}\right)\right] \\ &= \int_{-\mathbf{x}_{\rho,g_1}\boldsymbol{\beta}}^{\infty} \frac{1}{\sigma_{g_1}}\phi\left(\frac{u_{g_1}}{\sigma_{g_1}}\right) \Phi\left(\frac{\mathbf{x}_{\rho,g_2}\boldsymbol{\beta}+\frac{\sigma_{g_1,g_2}}{\sigma^2_{g_1}}u_{g_1}}{\sqrt{\sigma^2_{g_2}-\frac{\sigma^2_{g_1,g_2}}{\sigma^2_{g_1}}}}\right) du_{g_1} \end{aligned} \tag{C.1}$$

from $u_{g_2} \mid u_{g_1}, \mathbf{X} \sim \mathcal{N}\left(\tau_g \frac{\sigma_{g_2}}{\sigma_{g_1}} u_{g_1}, (1-\tau_g^2)\sigma^2_{g,2}\right)$ and noting that the conditional density of $u_{g_1} \mid \mathbf{X}, y_{g_1}=1$ is:

$$p(u_{g_1} \mid \mathbf{X}, y_{g_1}=1) = \mathbb{I}\{u_{g_1} \geq -\mathbf{x}_{\rho,g_1}\boldsymbol{\beta}\} \frac{\frac{1}{\sigma_{g_1}}\phi\left(\frac{u_{g_1}}{\sigma_{g_1}}\right)}{\Phi\left(\frac{\mathbf{x}_{\rho,g_1}\boldsymbol{\beta}}{\sigma_{g_1}}\right)}.$$

In a similar way:

$$\begin{aligned} p_g(1,0) &= \Phi\left(\frac{\mathbf{x}_{\rho,g_1}\boldsymbol{\beta}}{\sigma_{g_1}}\right) \mathrm{E}_{u_{g_1}|\mathbf{X},y_{g_1}=1}\left(1-\Phi\left(\frac{\mathbf{x}_{\rho,g_2}\boldsymbol{\beta}+\frac{\sigma_{g_1,g_2}}{\sigma^2_{g_1}}u_{g_1}}{\sqrt{\sigma^2_{g_2}-\frac{\sigma^2_{g_1,g_2}}{\sigma^2_{g_1}}}}\right)\right) \\ &= \Phi\left(\frac{\mathbf{x}_{\rho,g_1}\boldsymbol{\beta}}{\sigma_{g_1}}\right) - p_g(1,1), \end{aligned} \tag{C.2}$$

$$\begin{aligned} p_g(0,1) &= \left(1-\Phi\left(\frac{\mathbf{x}_{\rho,g_1}\boldsymbol{\beta}}{\sigma_{g_1}}\right)\right) \mathrm{E}_{u_{g_1}|\mathbf{X},y_{g_1}=0}\left(\Phi\left(\frac{\mathbf{x}_{\rho,g_2}\boldsymbol{\beta}+\frac{\sigma_{g_1,g_2}}{\sigma^2_{g_1}}u_{g_1}}{\sqrt{\sigma^2_{g_2}-\frac{\sigma^2_{g_1,g_2}}{\sigma^2_{g_1}}}}\right)\right) \\ &= \int_{-\infty}^{-\mathbf{x}_{\rho,g_1}\boldsymbol{\beta}} \frac{1}{\sigma_{g_1}}\phi\left(\frac{u_{g_1}}{\sigma_{g_1}}\right) \Phi\left(\frac{\mathbf{x}_{\rho,g_2}\boldsymbol{\beta}+\frac{\sigma_{g_1,g_2}}{\sigma^2_{g_1}}u_{g_1}}{\sqrt{\sigma^2_{g_2}-\frac{\sigma^2_{g_1,g_2}}{\sigma^2_{g_1}}}}\right) du_{g_1}, \end{aligned} \tag{C.3}$$

$$\begin{aligned} p_g(0,0) &= \left(1-\Phi\left(\frac{\mathbf{x}_{\rho,g_1}\boldsymbol{\beta}}{\sigma_{g_1}}\right)\right) - \int_{-\infty}^{-\mathbf{x}_{\rho,g_1}\boldsymbol{\beta}} \frac{1}{\sigma_{g_1}}\phi\left(\frac{u_{g_1}}{\sigma_{g_1}}\right) \Phi\left(\frac{\mathbf{x}_{\rho,g_2}\boldsymbol{\beta}+\frac{\sigma_{g_1,g_2}}{\sigma^2_{g_1}}u_{g_1}}{\sqrt{\sigma^2_{g_2}-\frac{\sigma^2_{g_1,g_2}}{\sigma^2_{g_1}}}}\right) du_{g_1} \\ &= \left(1-\Phi\left(\frac{\mathbf{x}_{\rho,g_1}\boldsymbol{\beta}}{\sigma_{g_1}}\right)\right) - p_g(0,1). \end{aligned} \tag{C.4}$$

The identity of all the formulas with the two equivalent expressions in (**??**) is straightforward.

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 $\pi \in \mathcal{P}_n$ and its permutation matrix $\mathbf{P}_\pi$, the densities f_0^π and f_0 are $n-$variate Gaussian random vectors,

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

where

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

(see Section 4 for the definition of $\mathbf{E}_g$). From the formula of 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 $\log|\mathbf{A}| = \mathrm{tr}\log(\mathbf{A})$:

$$\begin{aligned}
KL(f_0^\pi||f_0) &= \frac{1}{2}\left[\mathrm{tr}(\mathbf{P}_\pi\mathbf{\Sigma}^{-1}\mathbf{P}'_\pi\mathbf{\Sigma}_\pi) - n - \log\frac{|\mathbf{P}_\pi\mathbf{\Sigma}^{-1}\mathbf{P}'_\pi|}{|\mathbf{\Sigma}_\pi|}\right] \\
&= \frac{1}{2}\left[\mathrm{tr}(\mathbf{P}_\pi\mathbf{\Sigma}^{-1}\mathbf{P}'_\pi\mathbf{\Sigma}_\pi) - \log\left|\mathbf{P}_\pi\mathbf{\Sigma}^{-1}\mathbf{P}'_\pi\mathbf{\Sigma}_\pi\right|\right] - \frac{n}{2} \\
&= \frac{1}{2}\left[\mathrm{tr}(\mathbf{A}^{-1}) + \log|\mathbf{A}| - n\right] \qquad \text{(C.5)}
\end{aligned}$$

with

$$\mathbf{A} = \mathbf{P}_\pi\mathbf{\Sigma}\mathbf{P}'_\pi\mathbf{\Sigma}_\pi^{-1} = \mathbf{P}_\pi(\mathbf{A}_\rho)^{-1}\mathbf{P}'_\pi\mathbf{P}_\pi(\mathbf{A}'_\rho)^{-1}\mathbf{P}'_\pi\mathbf{\Sigma}_\pi^{-1}$$

But because of $\mathrm{tr}(\mathbf{AB}) = \mathrm{tr}(\mathbf{BA})$ 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 $\mathbf{P}'_\pi\boldsymbol{\Sigma}^{-1}\mathbf{P}_\pi\boldsymbol{\Sigma}_\pi$, that is, because of $\boldsymbol{\Sigma}_\pi$ is block diagonal,

$$c(2g-1,2g-1)=\sigma^*(\pi(2g-1),\pi(2g-1))\sigma(\pi(2g-1),\pi(2g-1))+\sigma^*(\pi(2g-1),\pi(2g))\sigma(\pi(2g),\pi(2g-1))$$
$$c(2g,2g)=\sigma^*(\pi(2g),\pi(2g))\sigma(\pi(2g),\pi(2g))+\sigma^*(\pi(2g),\pi(2g-1))\sigma(\pi(2g-1),\pi(2g))$$

where $\sigma^*(i,j)$ and $\sigma(i,j)$ are the $(i,j)-$th components of $\boldsymbol{\Sigma}^{-1}$ and $\boldsymbol{\Sigma}$ respectively.

The term $\log|\mathbf{A}|$ can be written as a sum of G components as well: $\log|\mathbf{A}|=\log|\boldsymbol{\Sigma}|+\log|\boldsymbol{\Sigma}_\pi^{-1}|=\log|\boldsymbol{\Sigma}|-\log|\sum_g\mathbf{E}_g\mathbf{P}_\pi\boldsymbol{\Sigma}\mathbf{P}'_\pi\mathbf{E}_g|$. Since the matrix $\boldsymbol{\Sigma}_\pi=\sum_g\mathbf{E}_g\mathbf{P}_\pi\boldsymbol{\Sigma}\mathbf{P}'_\pi\mathbf{E}_g$ is a block diagonal matrix, its determinant is equal to theproduct of determinants of blocks in the diagonal, namely,

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

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 $\log|\boldsymbol{\Sigma}|$ is invariant under permutations.

From the above computations, one gets that

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

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

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

for all π, we can write

$$\begin{aligned}&\arg\min_\pi KL(f_0^\pi||f_0)\\&=\arg\min_\pi\sum_g\big[\sigma^*(\pi(2g-1),\pi(2g))\sigma(\pi(2g),\pi(2g-1))+\sigma^*(\pi(2g),\pi(2g-1))\sigma(\pi(2g-1),\pi(2g))\\&\quad-\log\left(\sigma(\pi(2g-1),\pi(2g-1))\sigma(\pi(2g),\pi(2g))-\sigma(\pi(2g),\pi(2g-1))\sigma(\pi(2g-1),\pi(2g))\right)\big]\end{aligned}\tag{C.6}$$

Proof of Theorem 5.1

We first prove consistency of the PML estimator $\hat{\boldsymbol{\theta}}_n=\arg\max\ell_n(\boldsymbol{\theta};\mathbf{y},\mathbf{X})$. This can be proved by using the same arguments of Wang et al. (2013): in particular, given Assumption 4, we need to prove $\ell_n(\boldsymbol{\theta})-\ell(\boldsymbol{\theta})=o_p(1)$ and stochastic equicontinuity of $\ell_n(\boldsymbol{\theta})$. The first result follows by repeating exactly the same arguments as those of Lemma 2 in Wang et al. (2013).

In order to prove stochastic equicontinuity, following Wang et al. (2013), we need to show that

$$\sup_{\boldsymbol{\theta}}\left|\frac{1}{G}\sum_{g=1}^G y_{g_1}y_{g_2}\frac{\partial p_g(d_1,d_2)/\partial\boldsymbol{\theta}}{p_g(d_1,d_2)}\right|=O_p(1)$$

for all d_1,d_2.

The term at the denominator of the above equation is bounded away from zero because of Assumption 5. Moreover, all derivatives $\partial p_g(d_1, d_2)/\partial \boldsymbol{\theta}$ are $O_p(1)$ from Lemma Appendix F.2.

Thus, stochastic equicontinuity of ℓ_n follows as in Lemma 3 of Wang et al. (2013) and this implies consistency of $\hat{\boldsymbol{\theta}}$.

The proof of the consistency of $\tilde{\boldsymbol{\theta}}$ follows from the asymptotic equivalence of $\tilde{\boldsymbol{\theta}}_n$ with the PML estimator $\hat{\boldsymbol{\theta}}_n = \arg\max \ell_n(\boldsymbol{\theta}; \mathbf{y}, \mathbf{X})$, which is a consequence of Lemma Appendix F.1. In fact, Assumption 3 implies, for all finite n/G, that $\|\mathbf{X}\boldsymbol{\beta}\|_2 = O(1)$ for all $\boldsymbol{\beta}$ in the interior of the compact parameter space Θ. Thus, by Assumption 7 (a), $\frac{1}{G}\sum_{g=1}^{G} KL(\tilde{f}_g||f_g) = o(1)$ which implies, because of (21) and Lemma **??**–(i), that $\|\hat{\boldsymbol{\theta}}_n - \tilde{\boldsymbol{\theta}}_n\| = o_p(1)$.

Proof of Theorem 5.2

The result is proven in two steps. First, we decompose $\sqrt{n}(\tilde{\boldsymbol{\theta}} - \boldsymbol{\theta}_0) = \sqrt{n}(\tilde{\boldsymbol{\theta}} - \hat{\boldsymbol{\theta}}) + \sqrt{n}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0)$, and show that the first term is negligible with respect to the second, where $\hat{\boldsymbol{\theta}}$ is the pairwise ML estimator based on the exact computation of ℓ_n.

Second, we prove that $\sqrt{n}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0)$ has the asymptotic Gaussian distribution (22). This second part follows the lines of the proof of Theorem 2 in Wang et al. (2013) and those of Pinkse and Slade (1998).

We start by showing that $\sqrt{n}(\tilde{\boldsymbol{\theta}} - \hat{\boldsymbol{\theta}}) = o_p(1)$. Using the mean value theorem,

$$\frac{\partial \ell_n(\tilde{\boldsymbol{\theta}})}{\partial \boldsymbol{\theta}} = \frac{\partial \ell_n(\hat{\boldsymbol{\theta}})}{\partial \boldsymbol{\theta}} + \frac{\partial^2 \ell_n(\boldsymbol{\theta}^*)}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}'}(\tilde{\boldsymbol{\theta}} - \hat{\boldsymbol{\theta}}) = \frac{\partial^2 \ell_n(\boldsymbol{\theta}^*)}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}'}(\tilde{\boldsymbol{\theta}} - \hat{\boldsymbol{\theta}})$$

implies

$$\sqrt{n}(\tilde{\boldsymbol{\theta}} - \hat{\boldsymbol{\theta}}) = \left(\frac{\partial^2 \ell_n(\boldsymbol{\theta}^*)}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}'}\right)^{-1} \frac{\partial \ell_n(\tilde{\boldsymbol{\theta}})}{\partial \boldsymbol{\theta}}.$$

Boundedness of $\left(\frac{\partial^2 \ell_n(\boldsymbol{\theta}^*)}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}'}\right)^{-1}$ follows from Lemma Appendix F.4, then if

$$\frac{\partial \ell_n(\tilde{\boldsymbol{\theta}})}{\partial \boldsymbol{\theta}} = \frac{\partial \tilde{\ell_n}(\tilde{\boldsymbol{\theta}})}{\partial \boldsymbol{\theta}} + o_p(1) = o_p(1) \tag{C.7}$$

the negligibility of $\sqrt{n}(\tilde{\boldsymbol{\theta}} - \hat{\boldsymbol{\theta}})$ follows. The proof of (C.7) is in Lemma Appendix F.5.

In order to prove

$$\sqrt{n}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0) \to \mathcal{N}\left(0, \mathbb{H}(\boldsymbol{\theta}_0)^{-1} \mathrm{J}(\boldsymbol{\theta}_0) \mathbb{H}(\boldsymbol{\theta}_0)^{-1}\right),$$

we can repeat the same steps in Wang et al. (2013).

We sketch the main steps of the proof. For more details we refer to Wang et al. (2013).

Using the mean value theorem,

$$0 = \frac{\partial \ell_n(\hat{\boldsymbol{\theta}})}{\partial \boldsymbol{\theta}} = \frac{\partial \ell_n(\boldsymbol{\theta}_0)}{\partial \boldsymbol{\theta}} + \frac{\partial^2 \ell_n(\boldsymbol{\theta}^*)}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}'}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0).$$

Thus, for some $\boldsymbol{\theta}^*$ such that $\|\boldsymbol{\theta}^* - \boldsymbol{\theta}_0\| \leq \|\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0\|$,

$$\sqrt{n}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0) = -\left(\frac{\partial^2 \ell_n(\boldsymbol{\theta}^*)}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}'}\right)^{-1} \frac{\partial \ell_n(\hat{\boldsymbol{\theta}})}{\partial \boldsymbol{\theta}}.$$

Then, we first need to prove that all terms composing $\frac{\partial^2 \ell_n(\boldsymbol{\theta}^*)}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}'}$ are bounded (therefore integrable), in order to conclude, by invoking consistency of $\hat{\boldsymbol{\theta}}$ and the law of large numbers, that

$$\lim_{n\to\infty} \frac{\partial^2 \ell_n(\boldsymbol{\theta}^*)}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}'} = \mathbb{H}(\boldsymbol{\theta}_0). \tag{C.8}$$

To prove this, the same exact arguments of Theorem 2 of Wang et al. (2013) apply. First, the bounds $\left\| \frac{\partial p_g(d_1,d_2)}{\partial \boldsymbol{\theta}} \right\| < \infty$ and $\left\| \frac{\partial^2 p_g(d_1,d_2)}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}'} \right\| < \infty$ come from Lemma Appendix F.2 and Lemma Appendix F.4 respectively.

In order to have the weak limit of $\sqrt{n}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0)$, we finally need to show that

$$\mathrm{J}^{-1/2}(\boldsymbol{\theta}_0) \frac{\partial \ell_n(\hat{\boldsymbol{\theta}})}{\partial \boldsymbol{\theta}} \to_d \mathcal{N}(0, \mathbf{I}).$$

Following Wang et al. (2013), and as in Theorem 1 of Pinkse and Slade (1998), we invoke Bernsteins blocking methods and the McLeishs (1974) central limit theorem for dependent processes (see McLeish (1974)). This states that, if, for the triangular array $T_{nk_n} = \prod_{j=1}^{k_n}(1 + \imath\gamma D_{n,j})$, where $\imath^2 = -1$ and γ is a real constant, the following conditions are satisfied: (i) $\{T_{n,k_n}\}$ is uniformly integrable; (ii) $\mathrm{E}\, T_{n,k_n} \to_n 1$; (iii) $\sum_{j=1}^{k_n} D_{n,j}^2 \to_p 1$; (iv) $\max_{j \leq k_n} |D_{n,j}| \to 0$, then $\sum_{j=1}^{k_n} D_{n,j} \to_d N(0,1)$.

Following the reasoning in Wang et al. (2013), the (sequence of) regions where the observations are located is split into a_n areas of size $\sqrt{b_n} \times \sqrt{b_n}$, with a_n growing at a faster rate than b_n and such that $a_n b_n = n$. Moreover, a_n and b_n are chosen so that $b_n < n^{1/2-\varepsilon}$ uniformly in n and $\alpha(\sqrt{b_n})a_n \to 0$. Let $\Lambda_{n,j}$ represents the set of indices of observations falling into the $j-$th area, and write $D_{n,j} = G^{-1/2} \sum_{g \in \Lambda_{n,j}} A_{n,g}$ where $A_{n,g}$ is implicitly defined by $\mathbf{z}' \sqrt{G} \mathrm{J}(\boldsymbol{\theta}_0)^{-1/2} \left(\frac{\partial \ell_n(\boldsymbol{\theta}_0)}{\partial \boldsymbol{\theta}} \right) = G^{-1/2} \sum_{g=1}^{G} A_{n,g}$, for an arbitrary vector s.t. $\|\mathbf{z}\| = 1$. It then remains to prove conditions (i)–(iv) to ascertain that the sum $\sum_{j=0}^{k_n} D_{n,j} = G^{-1/2} \sum_{g=1}^{G} A_{n,g}$ is asymptotically normal.

For the proofs of conditions (iv) and (i), we just follow Wang et al. (2013). Conditions (ii)-(iii) follow from Lemmas 4–7 in Wang et al. (2013).

Appendix D. Approximation of $\boldsymbol{\Sigma}_g$ and $\mathbf{X}_\rho$.

As already pointed out by Wang et al. (2013), the terms σ_{g_1}, σ_{g_2} and σ_{g_1,g_2}, that are essential for the computation of the probabilities p_g, can not be easily written in closed form as functions of ρ and the weight matrix. In our case, things are made even worse by the fact that the vectors $\mathbf{x}_{\rho,g_i}$ are complex transformations of the whole design matrix that also depends on ρ.

In this Section we give details on the approximation of the terms $\mathbf{X}_{\rho,g}$ and $\sigma_{g,\cdot}$ based on finite series expansion for $(\mathbf{I}-\rho\mathbf{W})^{-1}$, mentioned in Section 3.

The contribution of each pair g on the loglikelihood depends on the rows g_1, g_2 of the matrix $\mathbf{A}_\rho^{-1}$ and on the submatrix $\mathbf{\Sigma}_g$. Under Assumption 1,

$$\mathbf{A}_\rho^{-1} = \sum_{k=0}^{\infty} \rho^k \mathbf{W}^k,$$

implies,

$$\mathbf{X}_\rho = \sum_{k=0}^{\infty} \rho^k \mathbf{W}^k \mathbf{X}$$

and, for $i = 1, 2$,

$$\mathbf{x}_{\rho,g_i} = \sum_{j=1}^{n} \sum_{k=0}^{\infty} \rho^k w_{g_i,j}^{(k)} \mathbf{x}_j$$

where $w_{l,j}^{(k)}$ is the $(l,j)-$term of the matrix $\mathbf{W}^k$. Then $\mathbf{x}_{\rho,g_i}$ can be approximated by truncating the series expansion to the $q-$th term.

Similarly, since

$$\mathbf{\Sigma} = \mathbf{A}_\rho^{-1}(\mathbf{A}_\rho')^{-1} = \sum_{k=0}^{\infty} \rho^k \mathbf{W}^k \sum_{h=0}^{\infty} \rho^h (\mathbf{W}')^h,$$

we could approximate,

$$\tilde{\mathbf{\Sigma}} = \tilde{\mathbf{A}_\rho}^{-1}(\tilde{\mathbf{A}_\rho}')^{-1} = \sum_{k=0}^{q} \rho^k \mathbf{W}^k \sum_{h=0}^{q} \rho^h (\mathbf{W}')^h = \sum_{k=0}^{2q} \rho^k \sum_{h=0}^{\min(k,q)} \mathbf{W}^h (\mathbf{W}')^{k-h}.$$

This approach can be convenient in the case of large samples, to avoid inversion of large matrices and is especially useful in the dense matrix case.

Appendix E. Score vector

In this Section we are going to derive the score vectors of the SAR(1) and SARAR(1,1) probit models. These formulas will be used to easily study the behavior of the score vector, and to perform more efficient computation of the maximum likelihood estimators (pairwise and quasi).

Appendix E.1. SAR(1)–probit

In order to compute the score vector for the optimization of the quasi pairwise loglikelihood, we need to compute the derivatives of $p_g(d_1, d_2)$ with respect to β and ρ.

We first consider differencing with respect to β:

$$\begin{aligned}\frac{\partial p_g(1,1)}{\partial \boldsymbol{\beta}} &= \frac{\partial}{\partial \boldsymbol{\beta}}\frac{1}{\sigma_{g_1}}\int_{-\mathbf{x}_{\rho,g_1}\boldsymbol{\beta}}^{\infty}\phi\left(\frac{u}{\sigma_{g_1}}\right)\Phi\left(\varphi_{2,g}(u)\right)du \\ &= \frac{1}{\sigma_{g_1}}\phi\left(\frac{-\mathbf{x}_{\rho,g_1}\boldsymbol{\beta}}{\sigma_{g_1}}\right)\Phi\left(\varphi_{2,g}(-\mathbf{x}_{\rho,g_1}\boldsymbol{\beta})\right)\mathbf{x}'_{\rho,g_1} + \int_{-\mathbf{x}_{\rho,g_1}\boldsymbol{\beta}}^{\infty}\frac{1}{\sigma_{g_1}}\phi\left(\frac{u}{\sigma_{g_1}}\right)\frac{\partial}{\partial \boldsymbol{\beta}}\Phi\left(\varphi_{2,g}(u)\right)du \\ &= \frac{1}{\sigma_{g_1}}\phi\left(\frac{-\mathbf{x}_{\rho,g_1}\boldsymbol{\beta}}{\sigma_{g_1}}\right)\Phi\left(\varphi_{2,g}(-\mathbf{x}_{\rho,g_1}\boldsymbol{\beta})\right)\mathbf{x}'_{\rho,g_1} + \frac{1}{\sigma_{g_1}}\int_{-\mathbf{x}_{\rho,g_1}\boldsymbol{\beta}}^{\infty}\phi\left(\frac{u}{\sigma_{g_1}}\right)\phi\left(\varphi_{2,g}(u)\right)\frac{\mathbf{x}'_{\rho,g_2}}{\sqrt{\sigma^2_{g_2}-\frac{\sigma^2_{g_1,g_2}}{\sigma^2_{g_1}}}}du\end{aligned}$$

where $\varphi_{2,g}$ is given in (10)

After some algebra, we obtain:

$$\begin{aligned}\int_{-\mathbf{x}_{\rho,g_1}\boldsymbol{\beta}}^{\infty}\phi\left(\frac{u}{\sigma_{g_1}}\right)\phi\left(\varphi_{2,g}(u)\right)du &= \phi\left(\frac{\mathbf{x}_{\rho,g_2}\boldsymbol{\beta}}{\sigma_{g_2}}\right)\int_{-\mathbf{x}_{\rho,g_1}\boldsymbol{\beta}}^{\infty}\frac{1}{\sqrt{2\pi}}\exp\left\{-\frac{\left(u\sigma_{g_2}+\mathbf{x}_{\rho,g_2}\boldsymbol{\beta}\sigma_{g_1,g_2}/\sigma_{g_2}\right)^2}{2\left(\sigma^2_{g_1}\sigma^2_{g_2}-\sigma^2_{g_1,g_2}\right)}\right\}du \\ &= \phi\left(\frac{\mathbf{x}_{\rho,g_2}\boldsymbol{\beta}}{\sigma_{g_2}}\right)\phi\left(\varphi_{1,g}(-\mathbf{x}_{\rho,g_2}\boldsymbol{\beta})\right)\sqrt{\sigma^2_{g_1}-\sigma^2_{g_1,g_2}/\sigma^2_{g_2}}, \qquad \text{(E.1)}\end{aligned}$$

where, also $\varphi_{1,g}$ follows from (10). Thus,

$$\frac{\partial p_g(1,1)}{\partial \boldsymbol{\beta}} = \frac{1}{\sigma_{g_1}}\phi\left(\frac{\mathbf{x}_{\rho,g_1}\boldsymbol{\beta}}{\sigma_{g_1}}\right)\Phi\left(\varphi_{2,g}(-\mathbf{x}_{\rho,g_1}\boldsymbol{\beta})\right)\mathbf{x}'_{\rho,g_1} + \frac{1}{\sigma_{g_2}}\phi\left(\frac{\mathbf{x}_{\rho,g_2}\boldsymbol{\beta}}{\sigma_{g_2}}\right)\Phi\left(\varphi_{1,g}(-\mathbf{x}_{\rho,g_2}\boldsymbol{\beta})\right)\mathbf{x}'_{\rho,g_2}. \qquad \text{(E.2)}$$

Similarly,

$$\begin{aligned}\frac{\partial p_g(1,0)}{\partial \boldsymbol{\beta}} &= \frac{\partial}{\partial \boldsymbol{\beta}}\Phi\left(\frac{\mathbf{x}_{\rho,g_1}\boldsymbol{\beta}}{\sigma_{g_1}}\right) - \frac{\partial p_g(1,1)}{\partial \boldsymbol{\beta}} \\ &= \phi\left(\frac{\mathbf{x}_{\rho,g_1}\boldsymbol{\beta}}{\sigma_{g_1}}\right)\frac{\mathbf{x}'_{\rho,g_1}}{\sigma_{g_1}} - \frac{\partial p_g(1,1)}{\partial \boldsymbol{\beta}} \\ &= \frac{1}{\sigma_{g_1}}\phi\left(\frac{\mathbf{x}_{\rho,g_1}\boldsymbol{\beta}}{\sigma_{g_1}}\right)\left(1-\Phi\left(\varphi_{2,g}(-\mathbf{x}_{\rho,g_1}\boldsymbol{\beta})\right)\right)\mathbf{x}'_{\rho,g_1} - \frac{1}{\sigma_{g_2}}\phi\left(\frac{\mathbf{x}_{\rho,g_2}\boldsymbol{\beta}}{\sigma_{g_2}}\right)\Phi\left(\varphi_{1,g}(-\mathbf{x}_{\rho,g_2}\boldsymbol{\beta})\right)\mathbf{x}'_{\rho,g_2} \qquad \text{(E.3)}\end{aligned}$$

and, by repeating the same steps,

$$\frac{\partial p_g(0,1)}{\partial \boldsymbol{\beta}} = -\frac{1}{\sigma_{g_1}}\phi\left(\frac{\mathbf{x}_{\rho,g_1}\boldsymbol{\beta}}{\sigma_{g_1}}\right)\Phi\left(\varphi_{2,g}(-\mathbf{x}_{\rho,g_1}\boldsymbol{\beta})\right)\mathbf{x}'_{\rho,g_1} + \frac{1}{\sigma_{g_2}}\phi\left(\frac{\mathbf{x}_{\rho,g_2}\boldsymbol{\beta}}{\sigma_{g_2}}\right)\left(1-\Phi\left(\varphi_{1,g}(-\mathbf{x}_{\rho,g_2}\boldsymbol{\beta})\right)\right)\mathbf{x}'_{\rho,g_2} \qquad \text{(E.4)}$$

$$\frac{\partial p_g(0,0)}{\partial \boldsymbol{\beta}} = -\frac{1}{\sigma_{g_1}}\phi\left(\frac{\mathbf{x}_{\rho,g_1}\boldsymbol{\beta}}{\sigma_{g_1}}\right)\left(1-\Phi\left(\varphi_{2,g}(-\mathbf{x}_{\rho,g_1}\boldsymbol{\beta})\right)\right)\mathbf{x}'_{\rho,g_1} - \frac{1}{\sigma_{g_2}}\phi\left(\frac{\mathbf{x}_{\rho,g_2}\boldsymbol{\beta}}{\sigma_{g_2}}\right)\left(1-\Phi\left(\varphi_{1,g}(-\mathbf{x}_{\rho,g_2}\boldsymbol{\beta})\right)\right)\mathbf{x}'_{\rho,g_2} \qquad \text{(E.5)}$$

In order to compute the derivatives with respect to ρ, we need to define:

$$\dot{\mathbf{X}}_\rho := \frac{\partial \mathbf{X}_\rho}{\partial \rho} = \frac{\partial \mathbf{A}_\rho^{-1}\mathbf{X}}{\partial \rho} = -\mathbf{A}_\rho^{-1}\frac{\partial \mathbf{A}_\rho}{\partial \rho}\mathbf{A}_\rho^{-1}\mathbf{X} = \mathbf{A}_\rho^{-1}\mathbf{W}\mathbf{A}_\rho^{-1}\mathbf{X}, \qquad \text{(E.6)}$$

$$\begin{aligned}\dot{\boldsymbol{\Sigma}} := \frac{\partial \boldsymbol{\Sigma}}{\partial \rho} &= \frac{\partial}{\partial \rho}(\mathbf{I}-\rho\mathbf{W})^{-1}(\mathbf{I}-\rho\mathbf{W}')^{-1} = \left(\frac{\partial}{\partial \rho}\mathbf{A}_\rho^{-1}\right)(\mathbf{I}-\rho\mathbf{W}')^{-1} + (\mathbf{I}-\rho\mathbf{W})^{-1}\left(\frac{\partial}{\partial \rho}(\mathbf{A}'_\rho)^{-1}\right) \\ &= -\mathbf{A}_\rho^{-1}\frac{\partial \mathbf{A}_\rho}{\partial \rho}\mathbf{A}_\rho^{-1}(\mathbf{A}'_\rho)^{-1} - \mathbf{A}_\rho^{-1}(\mathbf{A}'_\rho)^{-1}\frac{\partial \mathbf{A}'_\rho}{\partial \rho}(\mathbf{A}'_\rho)^{-1} \qquad \text{(E.7)} \\ &= \mathbf{A}_\rho^{-1}\mathbf{W}\boldsymbol{\Sigma} + \boldsymbol{\Sigma}\mathbf{W}'(\mathbf{A}'_\rho)^{-1}.\end{aligned}$$

Note that, the matrix $\tilde{\mathbf{A}}_\rho$, described in Appendix D, can be plugged in (E.6) and (E.7) to approximate $\dot{\mathbf{X}}_\rho$, $\dot{\mathbf{\Sigma}}$ and the score vector.

Now, we denote by $\dot{\mathbf{X}}_g = (\dot{\mathbf{x}}'_{g_1}, \dot{\mathbf{x}}'_{g_2})'$ and

$$\dot{\mathbf{\Sigma}}_g = \begin{pmatrix} \dot{\sigma}^2_{g_1} & \dot{\sigma}_{g_1,g_2} \\ \dot{\sigma}_{g_1,g_2} & \dot{\sigma}^2_{g_2} \end{pmatrix}$$

the submatrix corresponding to rows g_1, g_2 and the gth diagonal block matrix, respectively of $\dot{\mathbf{X}} = \dot{\mathbf{X}}_\rho$ and $\dot{\mathbf{\Sigma}} = \frac{\partial \mathbf{\Sigma}}{\partial \rho}$.

We further note that, from

$$\dot{\sigma}^2_{g_1} = \frac{\partial \sigma^2_{g_1}}{\partial \rho} = \frac{\partial \sigma^2_{g_1}}{\partial \sigma_{g_1}} \frac{\partial \sigma_{g_1}}{\partial \rho} = 2\sigma_{g_1} \frac{\partial \sigma_{g_1}}{\partial \rho}$$

we have, $\frac{\partial \sigma_{g_1}}{\partial \rho} = \frac{\dot{\sigma}^2_{g_1}}{2\sigma_{g_1}}$ and thus, $\frac{\partial}{\partial \rho}\frac{1}{\sigma_{g_1}} = -\frac{1}{\sigma^2_{g_1}}\frac{\dot{\sigma}^2_{g_1}}{2\sigma_{g_1}}$, and

$$\frac{\partial}{\partial \rho}\frac{1}{\sigma_{g_1}}\phi\left(\frac{u}{\sigma_{g_1}}\right) = \frac{u}{\sigma^2_{g_1}}\phi\left(\frac{u}{\sigma_{g_1}}\right)\frac{u\,\dot{\sigma}^2_{g_1}}{2\sigma^3_{g_1}} - \phi\left(\frac{u}{\sigma_{g_1}}\right)\frac{\dot{\sigma}^2_{g_1}}{2\sigma^3_{g_1}} = \phi\left(\frac{u}{\sigma_{g_1}}\right)\left(\frac{u^2}{\sigma^2_{g_1}} - 1\right)\frac{\dot{\sigma}^2_{g_1}}{2\sigma^3_{g_1}}$$

Then, we can write down the derivatives with respect to ρ,

$$\begin{aligned} \frac{\partial}{\partial \rho} p_g(1,1) &= \frac{1}{\sigma_{g_1}}\phi\left(\frac{\mathbf{x}_{\rho,g_1}\boldsymbol{\beta}}{\sigma_{g_1}}\right)\Phi\left(\varphi_{2,g}(-\mathbf{x}_{\rho,g_1}\boldsymbol{\beta})\right)\dot{\mathbf{x}}_{g_1}\boldsymbol{\beta} \\ &\quad + \int_{-\mathbf{x}_{\rho,g_1}\boldsymbol{\beta}}^{\infty} \phi\left(\frac{u}{\sigma_{g_1}}\right)\left(\frac{u^2}{\sigma^2_{g_1}} - 1\right)\frac{\dot{\sigma}^2_{g_1}}{2\sigma^3_{g_1}}\Phi\left(\varphi_{2,g}(u)\right)du \\ &\quad + \frac{1}{\sigma_{g_1}}\int_{-\mathbf{x}_{\rho,g_1}\boldsymbol{\beta}}^{\infty} \phi\left(\frac{u}{\sigma_{g_1}}\right)\phi\left(\varphi_{2,g}(u)\right)\frac{\partial}{\partial \rho}\varphi_{2,g}(u)du \\ &= \frac{1}{\sigma_{g_1}}\phi\left(\frac{\mathbf{x}_{\rho,g_1}\boldsymbol{\beta}}{\sigma_{g_1}}\right)\Phi\left(\varphi_{2,g}(-\mathbf{x}_{\rho,g_1}\boldsymbol{\beta})\right)\dot{\mathbf{x}}_{g_1}\boldsymbol{\beta} + B + C \end{aligned} \tag{E.8}$$

The integral in B can be computed by parts and we obtain, after some computation:

$$\begin{aligned} B = \frac{\dot{\sigma}^2_{g_1}}{2\sigma^2_{g_1}}\Bigg[&-\frac{\mathbf{x}_{\rho,g_1}\boldsymbol{\beta}}{\sigma_{g_1}}\phi\left(\frac{\mathbf{x}_{\rho,g_1}\boldsymbol{\beta}}{\sigma_{g_1}}\right)\Phi\left(\varphi_{2,g}(-\mathbf{x}_{\rho,g_1}\boldsymbol{\beta})\right) - \frac{\mathbf{x}_{\rho,g_2}\boldsymbol{\beta}}{\sigma_{g_2}}\phi\left(\frac{\mathbf{x}_{\rho,g_2}\boldsymbol{\beta}}{\sigma_{g_2}}\right)\Phi\left(\varphi_{1,g}(-\mathbf{x}_{\rho,g_2}\boldsymbol{\beta})\right) \\ &+\frac{\sigma_{g_1,g_2}|\mathbf{\Sigma}_g|^{1/2}}{\sigma^2_{g_1}\sigma^2_{g_2}}\phi\left(\frac{\mathbf{x}_{\rho,g_1}\boldsymbol{\beta}}{\sigma_{g_1}}\right)\phi\left(\varphi_{2,g}(-\mathbf{x}_{\rho,g_1}\boldsymbol{\beta})\right)\Bigg] \end{aligned}$$

where $|\mathbf{\Sigma}_g| = \sigma^2_{g_1}\sigma^2_{g_2} - \sigma^2_{g_1,g_2}$.

Note moreover that

$$\frac{\phi\left(\frac{\mathbf{x}_{\rho,g_1}\boldsymbol{\beta}}{\sigma_{g_1}}\right)\phi\left(\varphi_{2,g}(-\mathbf{x}_{\rho,g_1}\boldsymbol{\beta})\right)}{|\mathbf{\Sigma}_g|^{1/2}} = \frac{\phi\left(\frac{\mathbf{x}_{\rho,g_2}\boldsymbol{\beta}}{\sigma_{g_2}}\right)\phi\left(\varphi_{1,g}(-\mathbf{x}_{\rho,g_2}\boldsymbol{\beta})\right)}{|\mathbf{\Sigma}_g|^{1/2}} = f(\mathbf{x}_{\rho,g_1}\boldsymbol{\beta}, \mathbf{x}_{\rho,g_2}\boldsymbol{\beta}),$$

is the bivariate density of $(u_{g_1}, u_{g_2}) \sim N(\mathbf{0}, \mathbf{\Sigma}_g)$ at $(\mathbf{x}_{\rho,g_1}\boldsymbol{\beta}, \mathbf{x}_{\rho,g_2}\boldsymbol{\beta})$:

$$f(\mathbf{x}_{\rho,g_1}\boldsymbol{\beta}, \mathbf{x}_{\rho,g_2}\boldsymbol{\beta}) = \frac{1}{|\mathbf{\Sigma}_g|^{1/2}2\pi}\exp\left\{-\frac{1}{2}(\mathbf{x}_{\rho,g_1}\boldsymbol{\beta}, \mathbf{x}_{\rho,g_2}\boldsymbol{\beta})'\mathbf{\Sigma}_g^{-1}(\mathbf{x}_{\rho,g_1}\boldsymbol{\beta}, \mathbf{x}_{\rho,g_2}\boldsymbol{\beta})\right\} \tag{E.9}$$

Thus,

$$B = \frac{\dot{\sigma}^2_{g_1}}{2\sigma^2_{g_1}}\left[-\frac{\mathbf{x}_{\rho,g_1}\boldsymbol{\beta}}{\sigma_{g_1}}\phi\left(\frac{\mathbf{x}_{\rho,g_1}\boldsymbol{\beta}}{\sigma_{g_1}}\right)\Phi\left(\varphi_{2,g}(-\mathbf{x}_{\rho,g_1}\boldsymbol{\beta})\right) - \frac{\mathbf{x}_{\rho,g_2}\boldsymbol{\beta}}{\sigma_{g_2}}\phi\left(\frac{\mathbf{x}_{\rho,g_2}\boldsymbol{\beta}}{\sigma_{g_2}}\right)\Phi\left(\varphi_{1,g}(-\mathbf{x}_{\rho,g_2}\boldsymbol{\beta})\right) + \frac{\sigma_{g_1,g_2}|\mathbf{\Sigma}_g|}{\sigma^2_{g_1}\sigma^2_{g_2}}f(\mathbf{x}_{\rho,g_1}\boldsymbol{\beta}, \mathbf{x}_{\rho,g_2}\boldsymbol{\beta})\right] \tag{E.10}$$

In order to compute the intergal in C, we first note that

$$\begin{aligned}\frac{\partial \varphi_{2,g}(u)}{\partial \rho} &= \frac{\dot{\mathbf{x}}_{g_2}\boldsymbol{\beta} + \left(\dot{\sigma}_{g_1,g_2}\sigma_{g_1}^2 - \sigma_{g_1,g_2}\dot{\sigma}_{g_1}^2\right)u/(\sigma_{g_1}^2)^2}{(\sigma_{g_2}^2 - \sigma_{g_1,g_2}^2/\sigma_{g_1}^2)^{1/2}} - \frac{1}{2}\frac{\mathbf{x}_{\rho,g_2}\boldsymbol{\beta} + \frac{\sigma_{g_1,g_2}}{\sigma_{g_1}^2}u}{(\sigma_{g_2}^2 - \sigma_{g_1,g_2}^2/\sigma_{g_1}^2)^{3/2}}\left(\dot{\sigma}_{g_2}^2 - \frac{2\sigma_{g_1,g_2}\dot{\sigma}_{g_1,g_2}\sigma_{g_1}^2 - \sigma_{g_1,g_2}^2\dot{\sigma}_{g_1}^2}{(\sigma_{g_1}^2)^2}\right) \\ &= a + bu\end{aligned} \tag{E.11}$$

where

$$\begin{aligned}a &= \frac{\sigma_{g_1}}{|\boldsymbol{\Sigma}_g|^{1/2}}\dot{\mathbf{x}}_{g_2}\boldsymbol{\beta} + \frac{2\sigma_{g_1,g_2}\dot{\sigma}_{g_1,g_2}\sigma_{g_1}^2 - \sigma_{g_1,g_2}^2\dot{\sigma}_{g_1}^2 - \sigma_{g_1}^4\dot{\sigma}_{g_2}^2}{2\sigma_{g_1}|\boldsymbol{\Sigma}_g|^{3/2}} \\ b &= \frac{1}{2|\boldsymbol{\Sigma}_g|^{3/2}}\left[2\frac{\sigma_{g_1}^2\sigma_{g_2}^2}{\sigma_{g_1}}\dot{\sigma}_{g_1,g_2} - \dot{\sigma}_{g_1}^2\left(2\frac{\sigma_{g_2}^2\sigma_{g_1,g_2}}{\sigma_{g_1}} - \frac{\sigma_{g_1,g_2}^3}{\sigma_{g_1}^3}\right) - \frac{\sigma_{g_1,g_2}\sigma_{g_1}^2}{\sigma_{g_1}}\dot{\sigma}_{g_2}^2\right]\end{aligned}$$

Then, by noting that (performing a change of variable)

$$\begin{aligned}C &= \int_{-\mathbf{x}_{\rho,g_1}\boldsymbol{\beta}/\sigma_{g_1}}^{\infty} \phi(u)\,\phi\left(\varphi_{2,g}(u\sigma_{g_1})\right)\,(a + b\sigma_{g_1}u)du \\ &= \phi\left(\frac{\mathbf{x}_{\rho,g_2}\boldsymbol{\beta}}{\sigma_{g_2}}\right)\int_{-\mathbf{x}_{\rho,g_1}\boldsymbol{\beta}}^{\infty}(a + b\sigma_{g_1}u)\frac{1}{\sqrt{2\pi}}\exp\left\{-\frac{1}{2}\frac{(u\sigma_{g_2} + \mathbf{x}_{\rho,g_2}\boldsymbol{\beta}\sigma_{g_1,g_2}/\sigma_{g_1}\sigma_{g_2})}{|\boldsymbol{\Sigma}_g|/\sigma_{g_1}^2}\right\}du \\ &= \phi\left(\frac{\mathbf{x}_{\rho,g_2}\boldsymbol{\beta}}{\sigma_{g_2}}\right)\int_{-\varphi_{1,g}(\mathbf{x}_{\rho,g_2}\boldsymbol{\beta})}^{\infty}\frac{|\boldsymbol{\Sigma}_g|^{1/2}}{\sigma_{g_1}\sigma_{g_2}}\left[a + b\sigma_{g_1}\left(\frac{|\boldsymbol{\Sigma}_g|^{1/2}}{\sigma_{g_1}\sigma_{g_2}}v - \frac{\mathbf{x}_{\rho,g_2}\boldsymbol{\beta}}{\sigma_{g_2}}\frac{\sigma_{g_1,g_2}}{\sigma_{g_1}\sigma_{g_2}}\right)\right]\phi(v)dv\end{aligned}$$

we get

$$C = \frac{|\boldsymbol{\Sigma}|^{3/2}b\sigma_{g_1}}{\sigma_{g_1}^2\sigma_{g_2}^2}f(\mathbf{x}_{\rho,g_1}\boldsymbol{\beta}, \mathbf{x}_{\rho,g_2}\boldsymbol{\beta}) + \frac{|\boldsymbol{\Sigma}|^{1/2}}{\sigma_{g_1}\sigma_{g_2}}\left(a - b\sigma_{g_1}\frac{\mathbf{x}_{\rho,g_2}\boldsymbol{\beta}}{\sigma_{g_2}}\frac{\sigma_{g_1,g_2}}{\sigma_{g_1}\sigma_{g_2}}\right)\phi\left(\frac{\mathbf{x}_{\rho,g_2}\boldsymbol{\beta}}{\sigma_{g_2}}\right)\Phi\left(\varphi_{1,g}(-\mathbf{x}_{\rho,g_2}\boldsymbol{\beta})\right).$$

Finally, by putting all terms together and after some tedious calculations, we get

$$\begin{aligned}\frac{\partial p_g(1,1)}{\partial \rho} = \frac{f(\mathbf{x}_{\rho,g_1}\boldsymbol{\beta}, \mathbf{x}_{\rho,g_2}\boldsymbol{\beta})}{2}&\left(2\dot{\sigma}_{g_1,g_2} - \dot{\sigma}_{g_1}^2\frac{\sigma_{g_1,g_2}}{\sigma_{g_1}^2} - \dot{\sigma}_{g_2}^2\frac{\sigma_{g_1,g_2}}{\sigma_{g_2}^2}\right) \\ &+ \phi\left(\frac{\mathbf{x}_{\rho,g_1}\boldsymbol{\beta}}{\sigma_{g_1}}\right)\Phi\left(\varphi_{2,g}(-\mathbf{x}_{\rho,g_1}\boldsymbol{\beta})\right)\left(\frac{\dot{\mathbf{x}}_{g_1}\boldsymbol{\beta}}{\sigma_{g_1}} - \frac{\dot{\sigma}_{g_1}^2}{2\sigma_{g_1}^2}\frac{\mathbf{x}_{\rho,g_1}\boldsymbol{\beta}}{\sigma_{g_1}}\right) \\ &+ \phi\left(\frac{\mathbf{x}_{\rho,g_2}\boldsymbol{\beta}}{\sigma_{g_2}}\right)\Phi\left(\varphi_{1,g}(-\mathbf{x}_{\rho,g_2}\boldsymbol{\beta})\right)\left(\frac{\dot{\mathbf{x}}_{g_2}\boldsymbol{\beta}}{\sigma_{g_2}} - \frac{\dot{\sigma}_{g_2}^2}{2\sigma_{g_2}^2}\frac{\mathbf{x}_{\rho,g_2}\boldsymbol{\beta}}{\sigma_{g_2}}\right)\end{aligned} \tag{E.12}$$

Similar steps lead to,

$$\begin{aligned}\frac{\partial p_g(0,1)}{\partial \rho} = -\frac{f(\mathbf{x}_{\rho,g_1}\boldsymbol{\beta}, \mathbf{x}_{\rho,g_2}\boldsymbol{\beta})}{2}&\left(2\dot{\sigma}_{g_1,g_2} - \dot{\sigma}_{g_1}^2\frac{\sigma_{g_1,g_2}}{\sigma_{g_1}^2} - \dot{\sigma}_{g_2}^2\frac{\sigma_{g_1,g_2}}{\sigma_{g_2}^2}\right) \\ &- \phi\left(\frac{\mathbf{x}_{\rho,g_1}\boldsymbol{\beta}}{\sigma_{g_1}}\right)\Phi\left(\varphi_{2,g}(-\mathbf{x}_{\rho,g_1}\boldsymbol{\beta})\right)\left(\frac{\dot{\mathbf{x}}_{g_1}\boldsymbol{\beta}}{\sigma_{g_1}} - \frac{\dot{\sigma}_{g_1}^2}{2\sigma_{g_1}^2}\frac{\mathbf{x}_{\rho,g_1}\boldsymbol{\beta}}{\sigma_{g_1}}\right) \\ &+ \phi\left(\frac{\mathbf{x}_{\rho,g_2}\boldsymbol{\beta}}{\sigma_{g_2}}\right)\left(1 - \Phi\left(\varphi_{1,g}(-\mathbf{x}_{\rho,g_2}\boldsymbol{\beta})\right)\right)\left(\frac{\dot{\mathbf{x}}_{g_2}\boldsymbol{\beta}}{\sigma_{g_2}} - \frac{\dot{\sigma}_{g_2}^2}{2\sigma_{g_2}^2}\frac{\mathbf{x}_{\rho,g_2}\boldsymbol{\beta}}{\sigma_{g_2}}\right)\end{aligned} \tag{E.13}$$

$$
\begin{aligned}
\frac{\partial p_g(1,0)}{\partial \rho} &= \frac{\partial}{\partial \rho}\left\{\Phi\left(\frac{\mathbf{x}_{\rho,g_1}\boldsymbol{\beta}}{\sigma_{g_1}}\right) - p_g(1,1)\right\} \\
&= -\frac{f(\mathbf{x}_{\rho,g_1}\boldsymbol{\beta}, \mathbf{x}_{\rho,g_2}\boldsymbol{\beta})}{2}\left(2\dot{\sigma}_{g_1,g_2} - \dot{\sigma}^2_{g_1}\frac{\sigma_{g_1,g_2}}{\sigma^2_{g_1}} - \dot{\sigma}^2_{g_2}\frac{\sigma_{g_1,g_2}}{\sigma^2_{g_2}}\right) \\
&\quad + \phi\left(\frac{\mathbf{x}_{\rho,g_1}\boldsymbol{\beta}}{\sigma_{g_1}}\right)\left(1 - \Phi\left(\varphi_{2,g}(-\mathbf{x}_{\rho,g_1}\boldsymbol{\beta})\right)\right)\left(\frac{\dot{\mathbf{x}}_{g_1}\boldsymbol{\beta}}{\sigma_{g_1}} - \frac{\dot{\sigma}^2_{g_1}}{2\sigma^2_{g_1}}\frac{\mathbf{x}_{\rho,g_1}\boldsymbol{\beta}}{\sigma_{g_1}}\right) \\
&\quad - \phi\left(\frac{\mathbf{x}_{\rho,g_2}\boldsymbol{\beta}}{\sigma_{g_2}}\right)\Phi\left(\varphi_{1,g}(-\mathbf{x}_{\rho,g_2}\boldsymbol{\beta})\right)\left(\frac{\dot{\mathbf{x}}_{g_2}\boldsymbol{\beta}}{\sigma_{g_2}} - \frac{\dot{\sigma}^2_{g_2}}{2\sigma^2_{g_2}}\frac{\mathbf{x}_{\rho,g_2}\boldsymbol{\beta}}{\sigma_{g_2}}\right)
\end{aligned} \tag{E.14}
$$

$$
\begin{aligned}
\frac{\partial p_g(0,0)}{\partial \rho} &= \frac{\partial}{\partial \rho}\left(1 - \Phi\left(\frac{\mathbf{x}_{\rho,g_1}\boldsymbol{\beta}}{\sigma_{g_1}}\right)\right) - \frac{\partial}{\partial \rho}p_g(0,1) \\
&= \frac{f(\mathbf{x}_{\rho,g_1}\boldsymbol{\beta}, \mathbf{x}_{\rho,g_2}\boldsymbol{\beta})}{2}\left(2\dot{\sigma}_{g_1,g_2} - \dot{\sigma}^2_{g_1}\frac{\sigma_{g_1,g_2}}{\sigma^2_{g_1}} - \dot{\sigma}^2_{g_2}\frac{\sigma_{g_1,g_2}}{\sigma^2_{g_2}}\right) \\
&\quad - \phi\left(\frac{\mathbf{x}_{\rho,g_1}\boldsymbol{\beta}}{\sigma_{g_1}}\right)\left(1 - \Phi\left(\varphi_{2,g}(-\mathbf{x}_{\rho,g_1}\boldsymbol{\beta})\right)\right)\left(\frac{\dot{\mathbf{x}}_{g_1}\boldsymbol{\beta}}{\sigma_{g_1}} - \frac{\dot{\sigma}^2_{g_1}}{2\sigma^2_{g_1}}\frac{\mathbf{x}_{\rho,g_1}\boldsymbol{\beta}}{\sigma_{g_1}}\right) \\
&\quad - \phi\left(\frac{\mathbf{x}_{\rho,g_2}\boldsymbol{\beta}}{\sigma_{g_2}}\right)\left(1 - \Phi\left(\varphi_{1,g}(-\mathbf{x}_{\rho,g_2}\boldsymbol{\beta})\right)\right)\left(\frac{\dot{\mathbf{x}}_{g_2}\boldsymbol{\beta}}{\sigma_{g_2}} - \frac{\dot{\sigma}^2_{g_2}}{2\sigma^2_{g_2}}\frac{\mathbf{x}_{\rho,g_2}\boldsymbol{\beta}}{\sigma_{g_2}}\right)
\end{aligned} \tag{E.15}
$$

Appendix E.2. SARAR(1,1)–probit

In the SARAR(1,1) spcification, the probabilities $p_g(d_1, d_2)$ follow the same equations (**??**) as in the SAR(1) case. However, when computing all the quantities in (**??**), one has to bear in mind that the components $\sigma^2_{g_1}, \sigma^2_{g_2}, \sigma_{g_1,g_2}$ now depend on both ρ and λ through the variance covariance matrix:

$$
\boldsymbol{\Sigma} = (\mathbf{I} - \rho\mathbf{W})^{-1}(\mathbf{I} - \lambda\mathbf{M})^{-1}(\mathbf{I} - \lambda\mathbf{M}')^{-1}(\mathbf{I} - \rho\mathbf{W}')^{-1}. \tag{E.16}
$$

Thus, also in computing the derivatives of each $p_g(d_1, d_2)$ the simultaneous dependence of $\boldsymbol{\Sigma} = \boldsymbol{\Sigma}_{\nu(\rho,\lambda)}$ on ρ and λ has to be considered. This, however, does not in general alter the structure of the derivatives with respect to $\boldsymbol{\beta}$ and to ρ.

It is in fact easy to see that $\partial p_g(d_1, d_2)/\partial\boldsymbol{\beta}$ follows (E.2), (E.3), (E.4) and (E.5). Similarly, $\partial p_g(d_1, d_2)/\partial\rho$ follows equations (E.12), (E.14), (E.13) and (E.15).

Note moreover that, by writing

$$
\begin{aligned}
\frac{\partial \boldsymbol{\Sigma}}{\partial \rho} &= \frac{\partial}{\partial \rho}\left(\mathbf{A}_\rho^{-1}\mathbf{B}_\lambda^{-1}(\mathbf{B}_\lambda^{-1})'(\mathbf{A}_\rho^{-1})'\right) \\
&= \mathbf{A}_\rho^{-1}\mathbf{W}\mathbf{A}_\rho^{-1}\mathbf{B}_\lambda^{-1}(\mathbf{B}_\lambda^{-1})'(\mathbf{A}_\rho^{-1})' + \mathbf{A}_\rho^{-1}\mathbf{B}_\lambda^{-1}(\mathbf{B}_\lambda^{-1})'(\mathbf{A}_\rho^{-1})'\mathbf{W}'(\mathbf{A}_\rho^{-1})' \\
&= \mathbf{A}_\rho^{-1}\mathbf{W}\boldsymbol{\Sigma} + \boldsymbol{\Sigma}\mathbf{W}'(\mathbf{A}_\rho^{-1})'
\end{aligned}
$$

we can use the same equation as in the right-had-side of (E.7) to compute all the components of $\dot{\boldsymbol{\Sigma}}_\rho$.

We now focus on the derivative $\dot{\boldsymbol{\Sigma}}_\lambda = \partial\boldsymbol{\Sigma}/\partial\lambda$:

$$
\begin{aligned}
\dot{\boldsymbol{\Sigma}}_\lambda &= \mathbf{A}_\rho^{-1}\mathbf{B}_\lambda^{-1}\mathbf{M}\mathbf{B}_\lambda^{-1}(\mathbf{B}_\lambda^{-1})'(\mathbf{A}_\rho^{-1})' + \mathbf{A}_\rho^{-1}\mathbf{B}_\lambda^{-1}(\mathbf{B}_\lambda^{-1})'\mathbf{M}'(\mathbf{B}_\lambda^{-1})'(\mathbf{A}_\rho^{-1})' \\
&= \mathbf{A}_\rho^{-1}\mathbf{B}_\lambda^{-1}\mathbf{M}\mathbf{A}_\rho\boldsymbol{\Sigma} + \boldsymbol{\Sigma}\mathbf{A}_\rho'\mathbf{M}'(\mathbf{B}_\lambda^{-1})'(\mathbf{A}_\rho^{-1})'.
\end{aligned} \tag{E.17}
$$

Finally, using (E.17), we can compute the elements of $\dot{\boldsymbol{\Sigma}}_\lambda$, namely $\dot{\sigma}^2_{g_1}(\lambda), \dot{\sigma}^2_{g_2}(\lambda), \dot{\sigma}_{g_1,g_2}(\lambda)$ to be used in the following derivatives $\partial p_g(d_1,d_2)/\partial\lambda$:

$$\begin{aligned}
\frac{\partial p_g(1,1)}{\partial\lambda} &= \int_{-\mathbf{x}_{\rho,g_1}\boldsymbol{\beta}}^{\infty} \frac{1}{\sigma_{g_1}}\phi\left(\frac{u}{\sigma_{g_1}}\right)\left[\left(\frac{u^2}{\sigma^2_{g_1}}-1\right)\frac{\dot{\sigma}^2_{g_1}(\lambda)}{2\sigma^2_{g_1}}\Phi\left(\varphi_{2,g}(u)\right)+\phi\left(\varphi_{2,g}(u)\right)\frac{\partial\varphi_{2,g}(u)}{\partial\lambda}\right]du \\
\frac{\partial p_g(1,0)}{\partial\lambda} &= -\phi\left(\frac{\mathbf{x}_{\rho,g_1}\boldsymbol{\beta}}{\sigma_{g_1}}\right)\frac{\dot{\sigma}^2_{g_1}(\lambda)\mathbf{x}_{\rho,g_1}\boldsymbol{\beta}}{2\sigma^3_{g_1}}-\frac{\partial p_g(1,1)}{\partial\lambda} \\
\frac{\partial p_g(0,1)}{\partial\lambda} &= \int_{-\infty}^{-\mathbf{x}_{\rho,g_1}\boldsymbol{\beta}} \frac{1}{\sigma_{g_1}}\phi\left(\frac{u}{\sigma_{g_1}}\right)\left[\left(\frac{u^2}{\sigma^2_{g_1}}-1\right)\frac{\dot{\sigma}^2_{g_1}(\lambda)}{2\sigma^2_{g_1}}\Phi\left(\varphi_{2,g}(u)\right)+\phi\left(\varphi_{2,g}(u)\right)\frac{\partial\varphi_{2,g}(u)}{\partial\lambda}\right]du \\
\frac{\partial p_g(0,0)}{\partial\lambda} &= \phi\left(\frac{\mathbf{x}_{\rho,g_1}\boldsymbol{\beta}}{\sigma_{g_1}}\right)\frac{\dot{\sigma}^2_{g_1}(\lambda)\mathbf{x}_{\rho,g_1}\boldsymbol{\beta}}{2\sigma^3_{g_1}}-\frac{\partial p_g(0,1)}{\partial\lambda},
\end{aligned} \tag{E.18}$$

with

$$\frac{\partial\varphi_{2,g}(u)}{\partial\lambda} = \frac{\dot{\sigma}_{g_1,g_2}(\lambda)\sigma^2_{g_1}-\sigma_{g_1,g_2}\dot{\sigma}^2_{g_1}(\lambda)}{\sigma_{g_1}\sqrt{\sigma^2_{g_1}\sigma^2_{g_2}-\sigma^2_{g_1,g_2}}}\,u-\frac{1}{2}\frac{\mathbf{x}_{\rho,g_2}\boldsymbol{\beta}+u\frac{\sigma_{g_1,g_2}}{\sigma^2_{g_1}}}{\left(\sigma^2_{g_2}-\sigma^2_{g_1,g_2}/\sigma^2_{g_1}\right)^{3/2}}\left(\dot{\sigma}^2_{g_2}(\lambda)-\frac{2\sigma_{g_1,g_2}\dot{\sigma}_{g_1,g_2}(\lambda)\sigma^2_{g_1}-\sigma^2_{g_1,g_2}\dot{\sigma}^2_{g_1}(\lambda)}{\sigma^4_{g_1}}\right).$$

Formulas in (E.18) can be simplified through integration, as for the other terms of the score. Some calculations lead to a formula very similar to (E.12):

$$\begin{aligned}
\frac{\partial p_g(1,1)}{\partial\lambda} = \frac{f(\mathbf{x}_{\rho,g_1}\boldsymbol{\beta},\mathbf{x}_{\rho,g_2}\boldsymbol{\beta})}{2}&\left(2\dot{\sigma}_{g_1,g_2}(\lambda)-\dot{\sigma}^2_{g_1}(\lambda)\frac{\sigma_{g_1,g_2}}{\sigma^2_{g_1}}-\dot{\sigma}^2_{g_2}(\lambda)\frac{\sigma_{g_1,g_2}}{\sigma^2_{g_2}}\right) \\
&-\phi\left(\frac{\mathbf{x}_{\rho,g_1}\boldsymbol{\beta}}{\sigma_{g_1}}\right)\Phi\left(\varphi_{2,g}(-\mathbf{x}_{\rho,g_1}\boldsymbol{\beta})\right)\frac{\dot{\sigma}^2_{g_1}(\lambda)}{2\sigma^2_{g_1}}\frac{\mathbf{x}_{\rho,g_1}\boldsymbol{\beta}}{\sigma_{g_1}}-\phi\left(\frac{\mathbf{x}_{\rho,g_2}\boldsymbol{\beta}}{\sigma_{g_2}}\right)\Phi\left(\varphi_{1,g}(-\mathbf{x}_{\rho,g_2}\boldsymbol{\beta})\right)\frac{\dot{\sigma}^2_{g_2}(\lambda)}{2\sigma^2_{g_2}}\frac{\mathbf{x}_{\rho,g_2}\boldsymbol{\beta}}{\sigma_{g_2}}
\end{aligned} \tag{E.19}$$

The other derivatives can be easily derived adjusting equations (E.13)–(E.15) in the same way.

Appendix F. Technical Lemmas

Lemma Appendix F.1. *Under Assumptions 1–7,*

$$\frac{1}{G}\sum_{g=1}^{G} KL(f_g||\tilde{f}_g) \leq \left(1+\|\mathbf{X}\boldsymbol{\beta}\|_2^2\right) O\left(|\tau\rho|^{2(q+1)}\right).$$

Lemma Appendix F.2. *Under Assumptions 1–6,* $\frac{\partial\ell_n(\hat{\boldsymbol{\theta}})}{\partial\boldsymbol{\theta}} = o_p(1)$.

Lemma Appendix F.3. *Under Assumptions 1–6, fon any* $\gamma\in\mathbb{R}^k$, $\gamma\neq 0$,

$$\sup_{g\leq G}\max_{i=1,2}\|\mathbf{X}_{\rho,g}\mathbf{X}'_{\rho,g}\|_2\phi(\mathbf{X}_{\rho,g_i}\gamma)<\infty$$

Lemma Appendix F.4. *Under Assumptions 1–6 and 8–10, for all* $\boldsymbol{\theta}\in\Theta$, *and for all* $g=1,\ldots,G$ *and* $d_1,d_2\in\{0,1\}^2$,

$$\left\|\frac{\partial^2 p_g(d_1,d_2)}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}'}\right\|_2<\infty$$

If further Assumption 5 holds, then

$$\left\| \frac{\partial^2 \ell_n(\boldsymbol{\theta})}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}'} - \frac{\partial^2 \tilde{\ell}_n(\boldsymbol{\theta})}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}'} \right\|_2 = o(1)$$

Lemma Appendix F.5. *Under Assumptions 1–6*

$$\frac{\partial \ell_n(\tilde{\boldsymbol{\theta}})}{\partial \boldsymbol{\theta}} = \frac{\partial \tilde{\ell}_n(\tilde{\boldsymbol{\theta}})}{\partial \boldsymbol{\theta}} + o_p(1) = o_p(1)$$