Partial Maximum Likelihood Estimation of a Spatial Probit Model

Honglin Wang* Michigan State University

Emma M. Iglesias† Michigan State University

Jeffrey M. Wooldridge‡ Michigan State University

May 14, 2009

Abstract

This paper analyzes a spatial Probit model for cross sectional dependent data in a binary choice context. Observations are divided by pairwise groups and bivariate normal distributions are specified within each group. Partial maximum likelihood estimators are introduced and they are shown to be consistent and asymptotically normal under some regularity conditions. Consistent covariance matrix estimators are also provided. Finally, a simulation study shows the advantages of our new estimation procedure in this setting. Our proposed partial maximum likelihood estimators are shown to be more efficient than the generalized method of moments counterparts.

Keywords: Spatial statistics, Maximum Likelihood, Probit Model.
JEL classification: C12, C13, C21, C24, C25.

1 Introduction

Most econometrics techniques on cross-section data are based on the assumption of independence of observations. However, economic activities become more and more correlated over space with modern

*Department of Economics, Michigan State University, 101 Marshall-Adams Hall, East Lansing, MI 48824-1038, USA. e-mail: wanghon7@msu.edu.

†Department of Economics, Michigan State University, 101 Marshall-Adams Hall, East Lansing, MI 48824-1038, USA. e-mail: iglesia5@msu.edu.

‡Department of Economics, Michigan State University, 101 Marshall-Adams Hall, East Lansing, MI 48824-1038, USA. e-mail: wooldri1@msu.edu.

communication and transportation improvements. On the other hand, technological advances in communications and the geographic information system (GIS) make spatial data more available than before. Spatial correlations among observations received more and more attentions in regional, real estate, agricultural, environmental and industrial organizations economics (Lee, 2004).

Econometricians began to pay more attention on spatial dependence problems in the last two decades and some important advances have been done in both theoretical and empirical studies[1]. Spatial dependence not only means lack of independence between observations, but also a spatial structure underlying these spatial correlations (Anselin and Florax, 1995). There are two ways to capture spatial dependence by imposing structures on a model: one is in the domain of geostatistics where the spatial index is continuous (Conley, 1999), the other is that spatial sites form a countable lattice (Lee, 2004). Among the lattice models, there are also two types of spatial dependence models according to spatial correlation between variables or error terms: the spatial autoregressive dependent variable model (SAR) and the spatial autoregressive error model (SAE). In most applications of spatial models, the dependent variables are continuous (Conley, 1999; Lee, 2004; Kelejian and Prucha; 1999, 2001; among others), and only few applications address the spatial dependence with discrete choice dependent variables (exceptions include: Case, 1991; McMillen 1995; Pinkse and Slade, 1998; Lesage 2000; Beron and Vijerberg 2003). This paper is designed to address this gap and we are concerned about the SAE model with discrete choice dependent variables.

As the name indicates, there are two aspects in the discrete choice model with spatial dependence. First, the dependent variable is discrete and the leading cases occur where the choice is binary. Probit and Logit are the two most popular non-linear models for binary choice problems. For the sake of brevity, in this study we focus on Probit model, but the approach developed here generalizes to other discrete choice models.

In discrete choice models, if the observations are independent, we use maximum likelihood estimation to get efficient estimators given the correct conditional distribution of dependent variables. The nice part of the maximum likelihood estimator (MLE) is that we can still get consistency, asymptotic normality but inefficient estimators in many situations (panel data or clustering) by pseudo MLE even when we ignore certain dependence among observations (Poirier and Rudd,1988). However, the non-linear property causes computation difficulties in estimation, and this computational difficulty becomes much worse when dependence occurs, which results in solving n-dimensional integration.

Dependence is the other aspect of this problem. General forms of dependence are rarely allowed for in cross-sectional data, although routinely allowed for in time-series data (Conley, 1999). For example, some scholars discussed discrete choice models with dependence in time-series data: Robinson (1982) relaxed Amemiya (1973) assumptions of independence in Tobit model, and proved that the MLE with dependent observations is strongly consistent and asymptotically normal under some regularity conditions. Poirier and Rudd (1988) discussed the Probit model with dependence in time-series

[1] Anselin, Florax and Rey (2004) wrote a comprehensive review about econometrics for spatial models.

data, and developed generalized conditional moment (GCM) estimators which are computational attractive and relatively more efficient.

However, dependence in space is more complicated than in the time setting because of four reasons: first, time is one dimensional whereas space has at least two dimensions; second, time has natural order (direction) whereas space has no natural direction; third, time is regularly divided because of regular astronomical phenomena whereas spatial observations are attached to geographic properties of the surface of the earth; fourth, time-series observations are draws from a continuous process whereas, with spatial data, it is common for the sample and the population to be the same (Pinkse et al., 2007).

Therefore, how to deal with dependence in space in estimation is the key to spatial econometricians. Inspired by works about dependence in time-series data, Conley (1999) uses metrics of economic distance to characterize dependence among agents, and shows that the GMM estimator is consistent and asymptotically normal under some assumptions similar to time-series data. He also provides how to get consistent covariance matrix estimator by an approach similar to Newey-West (1987). Pinkse and Slade (1998) use GMM in the discrete choice setting with the SAE model, and show that the GMM estimator remains consistent and asymptotically normal under some regularity conditions. Although Pinkse and Slade (1998) generated generalized residuals from the MLE as the basis of the GMM estimators, they do not take advantage of information from spatial correlations among observations, and hence the GMM estimator is much less efficient than full ML estimators. Lee (2004) examines carefully the asymptotic properties of MLE and quasi-MLE for the linear spatial autoregressive model (SAR), and he shows that the rate of convergence of those estimators may depend on some general features of the spatial weights matrix of the model. If each units are influenced by only a few neighboring units, the estimators may have $\sqrt{n}$-rate of convergence and asymptotic normality; otherwise, it may have lower rate of convergence and estimators could be inconsistent.

In this study, we choose to capture spatial dependence by considering spatial sites to form a countable lattice, and explore a middle-ground approach which trades off efficiency and computation burdensome. The idea is to divide spatial dependent observations into many small groups (clusters) in which adjacent observations belong to one group. The implicit rationale behind this is adjacent observations usually account for the most important spatial correlations between observations. If we can correctly specify the conditional joint distribution within groups, which allows us to utilize relatively more information of spatial correlations, estimating the model by partial MLE will give us consistent and more efficient estimators, which should be generally better than GMM estimators. However, this approach is subject to biased variance-covariance matrix estimators because of spatial correlations among groups. To deal with this problem, we follow the methods proposed by Newey-West (1987) and Conley (1999) to get consistent variance-covariance matrix estimators. Of course, this middle ground approach will not get the most efficient estimator. However, since information from adjacent observations usually capture important spatial correlations in the whole sample, we get a consistent and a relatively efficient estimators, and we avoid some tedious computations at

expense of a loss of a relatively small part of efficiency.

This paper is organized as follows. First, we review econometric techniques on discrete choice models. Second, the SAE model with discrete choice dependent variable is presented and regularity conditions are specified. Section 3 presents the bivariate spatial Probit model. In Section 4, we prove consistency and asymptotic normality of partial ML estimators under regularity assumptions, and discuss how to get consistent covariance matrix estimators. Section 5 presents a simulation study showing the advantages of our new estimation procedure in this setting. Finally, Section 6 concludes. The proofs are collected in Appendix 1, while the results for the simulation study are provided in Appendix 2.

2 Discrete Choice Models with Spatial Dependence

2.1 Probit Model without Dependence

We first review the standard Probit model without dependence and the underlying linear latent variable model is

$$Y_i^* = X_i\beta + \varepsilon_i, \tag{1}$$

where Y_i^* is the latent dependent variable and a scalar, X_i is a $1 \times K$ vector of regressors, β is a $K \times 1$ parameter vector to be estimated, and ε_i is a continuous random variable, independent of X_i, and it follows a standard normal distribution. However, we cannot observe Y_i^*, and we can only observe the indicator Y_i, which is related to Y_i^* as follows

$$Y_i = \left\{ \begin{array}{ll} 1 & \text{if } Y_i^* > 0, \\ 0 & \text{if } Y_i^* \leq 0. \end{array} \right\}. \tag{2}$$

Therefore, we can get the conditional distribution of Y_i given X_i as

$$P(Y_i = 1|X_i) = P(Y_i^* > 0|X_i) = P(\varepsilon_i > -X_i\beta|X_i) = \Phi(X_i\beta), \tag{3}$$

where Φ denotes the standard normal cumulative distribution function (cdf). It is easy to see we can get

$$P(Y_i = 0|X_i) = 1 - \Phi(X_i\beta). \tag{4}$$

Since Y_i is a Bernoulli random variable, we can write the conditional density function of Y_i conditional on X_i as

$$f(Y_i|X_i) = [\Phi(X_i\beta)]^{Y_i}[1 - \Phi(X_i\beta)]^{1-Y_i}, \qquad Y_i = 0, 1. \tag{5}$$

Also, given the independence assumption of random variables, the log likelihood function can be written as

$$Log(L) = \sum_{i=1}^{n} \{Y_i \log[\Phi(X_i\beta)] + (1 - Y_i) \log[1 - \Phi(X_i\beta)]\}, \tag{6}$$

and the sufficient condition for uniqueness of the global maximum of $Log(L)$ is that the function is strictly concave (Gourieroux, 2000). We can solve then $\widehat{\beta}$ from the first order condition

$$\frac{\partial Log(L)}{\partial \beta} = \sum_{i=1}^{n} \frac{Y_i - \Phi(X_i\beta)}{\Phi(X_i\beta)[1-\Phi(X_i\beta)]}\phi(X_i\beta)X_i' = 0, \tag{7}$$

where ϕ is the probability density function (pdf) of the standard normal distribution. However, the simple closed-form expressions for the MLE are not available because the cdf of the normal distribution has no close-form solution. So the MLE must be solved by using numerical algorithms[2]. In general, we can prove that the conditional MLE is consistent and the most efficient estimator given some regularity conditions[3] such as correctly specifying a parametric model, an identified β and a log-likelihood function that is continuous in β.

2.2 A Probit Model with Spatial Error Correlation

Consider the Probit model with spatial error correlation (SAE), where the underlying linear latent variable model is

$$Y_i^* = X_i\beta + \varepsilon_i, \tag{8}$$

$$\varepsilon_i = \lambda \sum_{j=1}^{n} W_{ij}\varepsilon_j + u_i. \tag{9}$$

where W_{ij} is an element in the spatial weights matrix W which can be defined by different spatial distances such as the Euclidean distance. λ is the spatial autoregressive error coefficient and we have a random variable $u_i \sim i.i.d\ N(0,1)$. We can write equations (8) and (9) in matrix form as follows

$$Y^* = X\beta + \varepsilon \tag{10}$$

$$\varepsilon = (I - \lambda W)^{-1} u, \tag{11}$$

so that the variance-covariance matrix for the model is

$$\Omega \equiv Var(\varepsilon|X) = [(I-\lambda W)'(I-\lambda W)]^{-1}. \tag{12}$$

If Y^* is observable, equation (10) becomes a linear function, and we can use the Jacobian transformation of u into Y^* and write the log likelihood function as

$$L(\beta,\lambda) = -\frac{n}{2}\ln(2\pi) - \frac{1}{2}(Y^* - X\beta)'A'A(Y^* - X\beta) + \ln|A| \tag{13}$$

where $A = I - \lambda W$, and then the estimate of β can be solved as $\widehat{\beta} = (X'A'AX)^{-1}X'A'AY^*$.

[2]Commonly used numerical solutions are all derived from Newton's method. (see Gourieroux, 2000 for details).

[3]See details in Wooldridge (2001, page 391).

However, in practice we cannot observe Y^*, and we can only observe Y_i, and it implies a non-linear Probit model because of the normal distributional assumption. Moreover the errors are correlated, and the full likelihood function becomes

$$L = P(Y_1 = y_1, Y_2 = y_2, \cdots Y_n = y_n) = \int_{-\infty}^{a_1} \cdots \int_{-\infty}^{a_n} \phi(u)du, \tag{14}$$

$$\phi(u) = (2\pi)^{-\frac{n}{2}} |\Omega|^{-1} e^{-\frac{1}{2}(u'\Omega^{-1}u)}. \tag{15}$$

Although theoretically, if we take the first derivatives subject to β and the spatial coefficient λ, we obtain

$$\frac{\partial L}{\partial \beta} = \frac{\partial\{\int_{-\infty}^{a_1} \cdots \int_{-\infty}^{a_n} (2\pi)^{-\frac{n}{2}} |(I-\lambda W)'(I-\lambda W)| e^{-\frac{1}{2}[u'(I-\lambda W)'(I-\lambda W)u]} du\}}{\partial \beta} = 0, \tag{16}$$

$$\frac{\partial L}{\partial \lambda} = \frac{\partial\{\int_{-\infty}^{a_1} \cdots \int_{-\infty}^{a_n} (2\pi)^{-\frac{n}{2}} |(I-\lambda W)'(I-\lambda W)| e^{-\frac{1}{2}[u'(I-\lambda W)'(I-\lambda W)u]} du\}}{\partial \lambda} = 0. \tag{17}$$

The expression of the first derivatives are quite complicated, but if we have sufficient computational ability and β and λ are identifiable, we can get consistent and efficient estimates of β and λ by using numerical methods. However, in practice, it would be a formidable computational task even for a moderate size sample. We now propose a more attractive procedure in the next sections.

2.3 Probit Models with Other Forms of Spatial Correlation

Generally, there is no reason to think that spatial correlation is properly modeled by (9). Other forms are possible. For example, one might assume that, outside of a certain geographic radius from a given observation i, ε_i is uncorrelated with shocks to the outlying regions. So, for example, we might assume a constant correlation with any unit within a given radius – similar to a random effects structure for unbalanced panel data.

Alternatively, we may prefer more of a moving average structure, such as

$$\varepsilon_i = u_i + \lambda \left(\sum_{h \neq i} W_{ih} u_h \right), \tag{18}$$

where the u_i are i.i.d. with unit variance. This formulation is attractive because it is relatively easy to find variances and pairwise correlations, which we will use in the partial MLEs described in the next section. For example,

$$Var(\varepsilon_i | W) = 1 + \lambda^2 \left(\sum_{h \neq i} W_{ih}^2 \right). \tag{19}$$

Clearly, methods that use only the variance in estimation can only identify λ^2 (but we almost always think $\lambda > 0$, anyway). Pairwise covariances can also be obtained from

$$Cov(\varepsilon_i, \varepsilon_j|W) = \lambda W_{ij} + \lambda W_{ji} + \lambda^2 \left(\sum_{h \neq i, h \neq j} W_{ih} W_{jh} \right). \tag{20}$$

Expressions like this for the covariance between different errors are important for applying grouped partial MLE methods.

3 Using Partial MLEs to Estimate General Spatial Probit Models

Estimating a probit spatial autocorrelation model by full MLE is a prodigious task, although several approaches have been applied. The EM algorithm can be used (McMillen, 1992), the RIS simulator (Beron and Vijverberg, 2003), and the Bayesian Gibbs sampler (Lesage, 2000). But each of these approaches is still computationally burdensome. To combine such approaches with simulation studies, or to be able to quickly estimate a range of models, is outside the abilities of even current computation capabilities for even moderate sample sizes.

To get an estimator that is computationally feasible, Pinkse and Slade (1998) proposed using generalized method of moments (GMM) using information on the marginal distributions of the binary responses. In particular, the generalized residuals from the marginal probit log likelihood are used to construct moment conditions for the GMM method. Pinske and Slade show that, under conditions very similar to those in this paper, the GMM estimator is consistent and asymptotically normal. The consistent variance-covariance matrix can also be obtained theoretically without a covariance stationary assumption, although Pinske and Slade do not discuss estimation of the asymptotic variance. Therefore, the GMM estimator is almost practically useful, but it is fundamentally based on the marginal probit models. Thus, while a GMM estimator can be obtained that is efficient given the information on the marginal likelihood, the method throws out much useful information. We describe a simplified version of this approach in Section 3.1, which, in effect, uses a heteroskedastic probit model to estimate the β_j along with any spatial autocorrelation parameter.

Using only the marginal distribution of Y_i, conditional on the covariates and weights, likely results in serious loss of information for estimating both β and the spatial autocorrelation parameters. Our key contribution in this paper is to explore the use of partial maximum likelihood where we group small numbers of nearby observations and obtain the joint distribution of those observations. Naturally, these distributions are determined by the fully specified spatial autocorrelation model – just as we must obtain the implied variance to apply marginal probit methods. Once the covariances between observations are found as a function of the weights and λ, we can use that information in multivariate probit estimation. Section 3.2 covers the case of where we describe a bivariate probit

approach, with heteroskedasticity and covariance implied by the particular spatial autocorrelation model. Using a single covariance in addition to the variance seems likely to improve efficiency of estimation.

3.1 Univariate Probit Partial MLE

One way to estimate the coefficients β along with spatial correlation parameters is to derive the marginal distributions, $P(Y_i = 1|X, W)$ as a function of all of the weights (and the parameters, β and λ, of course). Under the joint normality assumption, the model will be a form of probit with heteroskedasticity. In particular, given any spatial probit model such that the variances are well defined, we can find

$$P(Y_i = 1|X, W) = \Phi(X_i\beta/\sigma_i(\lambda)), \tag{21}$$

where $\sigma_i^2(\lambda) = Var(\varepsilon_i|X, W) = Var(\varepsilon_i|W)$ is a function of all weights, W, and the spatial correlation parameters, λ. As is well-known in time series contexts – see, for example, Poirier and Ruud (1988) or Robinson (1982) – using probit while ignoring the time series correlation leads to consistent estimation under standard regularity conditions, provided the data are weakly dependent. Thus, it is not surprising that pooled probit that accounts for the heteroskedasticity in the marginal distribution is generally consistent for spatially correlated data, too – provided, of course, we limit the amount of spatial correlation.

The log likelihood can be written generically as

$$Log(L) = \sum_{i=1}^{n} \{Y_i \log[\Phi(X_i\beta/\sigma_i(\lambda))] + (1 - Y_i) \log[1 - \Phi(X_i\beta/\sigma_i(\lambda))]\}, \tag{22}$$

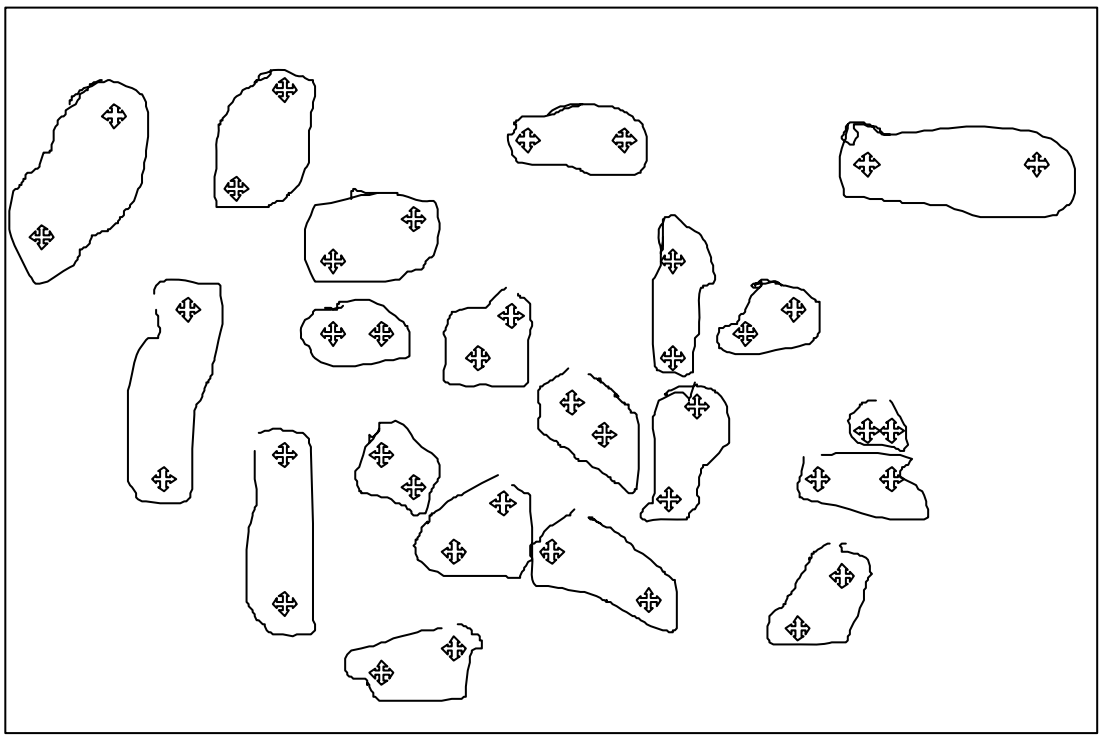
Assuming that β and λ are identified, and that the conditions in Section 4 hold, the pooled heteroskedastic probit is generally consistent and $\sqrt{n}$-asymptotically normal. But, for reasons we discussed above, it is likely to be very inefficient relative to the full MLE. Further, estimators that use some information on the spatial correlation across observations seem more promising in terms of increasing precision.

3.2 Bivariate Probit Partial MLE

We now turn to using information on pairs of "nearby" observations to identify β and λ. There is nothing special about using pairs; we could use, say, triplets, or even larger groups. But the bivariate case is easy to illustrate and is computationally quite feasible.

For illustration, assume a sample includes $2n$ observations, and we divide the $2n$ observations into n pairwise groups according to the spatial Euclidean distance between them (see Graph 1). In other words, each group includes two observations, with the idea being that the internal correlation between the two observations is more important than external correlations with observations in other

groups. Within a group, the two observations follow a conditional bivariate normal distribution because error terms are assumed to have a joint normal distribution.



Graph 1: ($2n$ observations $\Longrightarrow$ n groups)

In group g, we have

$$Y^*_{g1} = \beta_1 X_{g11} + \beta_2 X_{g12} + \ldots + \beta_k X_{g1k} + \varepsilon_{g1} \tag{23}$$

$$Y^*_{g2} = \beta_1 X_{g21} + \beta_2 X_{g22} + \ldots + \beta_k X_{g2k} + \varepsilon_{g2}, \qquad g = 1, 2 \ldots . n. \tag{24}$$

Rewrite the above equations in matrix form as

$$Y^*_{g1} = X_{g1}\beta + \varepsilon_{g1} \tag{25}$$

$$Y^*_{g2} = X_{g2}\beta + \varepsilon_{g2}, \qquad g = 1, 2 \ldots . n, \tag{26}$$

where X_{g1} and X_{g2} are $1 \times K$ vectors of regressors and β is a $K \times 1$ vector. ε_{g1} and ε_{g2} are scalars. In group g, observation A and observation B are not only correlated with each other, but also correlated with other observations over space. Therefore, the variances and covariance between ε_{g1} and ε_{g2} not only depend on the weight within group, but also weights with other observations out of the group, and the parameters, λ as well. See, for example equation (20).

It is easy to see that $E(\varepsilon_{g1}|X_{g1}, W) = E(\varepsilon_{g2}|X_{g2}, W) = 0$, and the covariance-variance matrix for group g is defined as $\Omega_g \equiv Var(\varepsilon_g|X_g, W)$ where

$$Var(\varepsilon_g|X_g, W) \equiv \Omega_g(W, \lambda) = \begin{bmatrix} \Omega_{g11} & \Omega_{g12} \\ \Omega_{g21} & \Omega_{g22} \end{bmatrix}, \tag{27}$$

where we suppress the dependence on W and λ in what follows for notational simplicity.

Note here that elements in Ω_g depend not only on the weight between two observations in group g, but also weights for every observation in the whole sample, because two observations in group g not only correlated with each other, but also correlated with other observations over space. Since we define two nearby observations as one group, we pick up the corresponding part (Ω_g) from the whole covariance-variance matrix (see equation (20)).

Since we cannot observe Y_{g1}^* and Y_{g2}^*, as we discussed in the univariate Probit model, we define

$$Y_g = \left\{ \begin{array}{ll} 1 & \text{if } Y_g^* > 0, \\ 0 & \text{if } Y_g^* \leq 0 \end{array} \right\}. \tag{28}$$

Therefore the conditional bivariate normal distribution of Y_{g1} and Y_{g2} given X_g is given as

$$\begin{aligned} P(Y_{g1} &= 1, Y_{g2} = 1|X_g) = P(X_{g1}\beta + \varepsilon_{g1} > 0, X_{g2}\beta + \varepsilon_{g2} > 0|X_g) & (29) \\ &= P(\varepsilon_{g1} < X_{g1}\beta, \varepsilon_{g2} < X_{g2}\beta|X_g) = \Phi_2(\frac{X_{g1}\beta}{\sqrt{\Omega_{g11}}}, \frac{X_{g2}\beta}{\sqrt{\Omega_{g22}}}, \rho_g), & (30) \\ \rho_g &= \frac{Cov(\varepsilon_{g1}, \varepsilon_{g2})}{\sqrt{Var(\varepsilon_{g1})}\sqrt{Var(\varepsilon_{g2})}} = \frac{\Omega_{g12}}{\sqrt{\Omega_{g11}\Omega_{g22}}}, & (31) \end{aligned}$$

where Φ_2 is the standard bivariate normal distribution, ϕ_2 is the standard density function of the bivariate normal distribution and ρ_g is the standardized covariance between two error terms.

Given that $(\varepsilon_{g1}, \varepsilon_{g2})$ has a joint normal distribution, we can write

$$\varepsilon_{g1} = \delta_{g1}\varepsilon_{g2} + e_{g1} \tag{32}$$

where

$$\delta_{g1} = \frac{Cov(\varepsilon_{g1}, \varepsilon_{g2})}{Var(\varepsilon_{g2})}, \tag{33}$$

and e_{g1} is independent of X_g and ε_{g2}.

Because of the joint normality of $(\varepsilon_{g1}, \varepsilon_{g2})$, e_{g1} is also normally distributed with $E(e_{g1}) = 0$, and

$$Var(e_{g1}) = Var(\varepsilon_{g1}) - \delta_{g1}^2 Var(\varepsilon_{g2}). \tag{34}$$

Thus, we can write the conditional distribution of e_{g1} as

$$(e_{g1}|X_g, \varepsilon_{g2}) \sim N(0, Var(e_{g1})). \tag{35}$$

Substitute equation (32) back to $Y_{g1}^* = X_{g1}\beta + \varepsilon_{g1}$, and we can get

$$Y_{g1}^* = X_{g1}\beta + \delta_{g1}\varepsilon_{g2} + e_{g1}. \tag{36}$$

Therefore

$$P(Y_{g1} = 1|X_g, \varepsilon_{g2}) = \Phi(\frac{X_{g1}\beta + \delta_{g1}\varepsilon_{g2}}{\sqrt{Var(e_{g1})}}). \tag{37}$$

The reason we want to find (37) is to retrieve $P(Y_{g1}=1,Y_{g2}=1|X_g)$. Since

$$P(Y_{g1}=1,Y_{g2}=1|X_g)=P(Y_{g1}=1|Y_{g2}=1,X_g)\times P(Y_{g2}=1|X_g) \tag{38}$$

it is easy to see that $P(Y_{g2}=1|X_g)=\Phi(\frac{X_{g2}\beta}{\sqrt{Var(\varepsilon_{g2})}})$, and thus it remains to get $P(Y_{g1}=1|Y_{g2}=1,X_g)$.

First, since $Y_{g2}=1$ if and only if $\varepsilon_{g2}>-X_{g2}\beta$, and ε_{g2} follows a normal distribution and it is independent of X_g, then the density of ε_{g2} given $\varepsilon_{g2}>-X_{g2}\beta$ is

$$\frac{\phi(\frac{\varepsilon_{g2}}{\sqrt{Var(\varepsilon_{g2})}})}{P(\varepsilon_{g2}>-X_{g2}\beta)}=\frac{\phi(\frac{\varepsilon_{g2}}{\sqrt{Var(\varepsilon_{g2})}})}{\Phi(\frac{X_{g2}\beta}{\sqrt{Var(\varepsilon_{g2})}})}. \tag{39}$$

Therefore,

$$\begin{aligned}
P(Y_{g1} &= 1|Y_{g2}=1,X_g)=E[P(Y_{g1}=1|X_g,\varepsilon_{g2})|Y_{g2}=1,X_g) && (40)\\
&= E[\Phi(\frac{X_{g1}\beta+\delta_{g1}\varepsilon_{g2}}{\sqrt{Var(e_{g1})}})|Y_{g2}=1,X_g] && (41)\\
&= \frac{1}{\Phi(\frac{X_{g2}\beta}{\sqrt{Var(\varepsilon_{g2})}})}\int_{-X_{g2}\beta}^{\infty}\Phi(\frac{X_{g1}\beta+\delta_{g1}\varepsilon_{g2}}{\sqrt{Var(e_{g1})}})\phi(\frac{\varepsilon_{g2}}{\sqrt{Var(\varepsilon_{g2})}})d\varepsilon_{g2} && (42)
\end{aligned}$$

and it is easy to see that $P(Y_{g1}=0|Y_{g2}=1,X_g)=1-P(Y_{g1}=1|Y_{g2}=1,X_g)$ because Y_{g1} is the binary variable.

Similarly, we can get

$$P(Y_{g1}=1|Y_{g2}=0,X_g)=\frac{1}{1-\Phi(\frac{X_{g2}\beta}{\sqrt{Var(\varepsilon_{g2})}})}\int_{-\infty}^{X_{g2}\beta}\Phi(\frac{X_{g1}\beta+\delta_{g1}\varepsilon_{g2}}{\sqrt{Var(e_{g1})}})\phi(\frac{\varepsilon_{g2}}{\sqrt{Var(\varepsilon_{g2})}})d\varepsilon_{g2} \tag{43}$$

and $P(Y_{g1}=0|Y_{g2}=0,X_g)=1-P(Y_{g1}=1|Y_{g2}=0,X_g)$.

Now we are ready to get $P(Y_{g1}=1,Y_{g2}=1|X_g)$ as follows

$$\begin{aligned}
P(Y_{g1} &= 1,Y_{g2}=1|X_g)=\frac{1}{\Phi(\frac{X_{g2}\beta}{\sqrt{Var(\varepsilon_{g2})}})}\int_{-X_{g2}\beta}^{\infty}\Phi(\frac{X_{g1}\beta+\delta_{g1}\varepsilon_{g2}}{\sqrt{Var(e_{g1})}})\phi(\frac{\varepsilon_{g2}}{\sqrt{Var(\varepsilon_{g2})}})d\varepsilon_{g2}\\
&\quad\times\Phi(\frac{X_{g2}\beta}{\sqrt{Var(\varepsilon_{g2})}}) && (44)\\
&= \int_{-X_{g2}\beta}^{\infty}\Phi(\frac{X_{g1}\beta+\delta_{g1}\varepsilon_{g2}}{\sqrt{Var(e_{g1})}})\phi(\frac{\varepsilon_{g2}}{\sqrt{Var(\varepsilon_{g2})}})d\varepsilon_{g2}, && (45)
\end{aligned}$$

and similarly we can obtain finally

$$\begin{aligned}
P(Y_{g1} &= 0,Y_{g2}=1|X_g)=\Phi(\frac{X_{g2}\beta}{\sqrt{Var(\varepsilon_{g2})}})-\int_{-X_{g2}\beta}^{\infty}\Phi(\frac{X_{g1}\beta+\delta_{g1}\varepsilon_{g2}}{\sqrt{Var(e_{g1})}})\phi(\frac{\varepsilon_{g2}}{\sqrt{Var(\varepsilon_{g2})}})d\varepsilon_{g2} && (46)\\
P(Y_{g1} &= 1,Y_{g2}=0|X_g)=\int_{-\infty}^{X_{g2}\beta}\Phi(\frac{X_{g1}\beta+\delta_{g1}\varepsilon_{g2}}{\sqrt{Var(e_{g1})}})\phi(\frac{\varepsilon_{g2}}{\sqrt{Var(\varepsilon_{g2})}})d\varepsilon_{g2} && (47)\\
P(Y_{g1} &= 0,Y_{g2}=0|X_g)\\
&= [1-\Phi(\frac{X_{g2}\beta}{\sqrt{Var(\varepsilon_{g2})}})]-\int_{-\infty}^{X_{g2}\beta}\Phi(\frac{X_{g1}\beta+\delta_{g1}\varepsilon_{g2}}{\sqrt{Var(e_{g\theta 1})}})\phi(\frac{\varepsilon_{g2}}{\sqrt{Var(\varepsilon_{g2})}})d\varepsilon_{g2}. && (48)
\end{aligned}$$

4 Partial Maximum Likelihood Estimation

As we discussed in the introduction, if the observations are independent, we can simplify the multivariate distribution into the product of univariate distributions, and then the ML estimator can be obtained easily. However, spatial correlations among observations do not allow the simplification any more. Under spatial correlation, the situation is kind of similar to the panel data case. In panel data, we cannot assume independence among observations over different periods for the same person (or firm), which means we are not likely to specify the full conditional density of Y given X correctly. Therefore, we need to relax the assumption in the panel data case. The way we deal with the problem is that if we have a correctly specified model for the density of Y_t given X_t, we can define the partial log likelihood function as

$$\underset{\theta\in\Theta}{Max}\sum_{i=1}^{N}\sum_{t=1}^{T}\log f_t(y_{it}|X_{it},\theta), \tag{49}$$

where $f_t(y_{it}|X_{it},\theta)$ is the density for y_{it} given x_{it} for each t. The partial log likelihood function works because θ_0 (the true value) maximizes the expected value of the above equation provided we have the densities $f_t(y_{it}|X_{it},\theta)$ correctly specified (Wooldridge, 2002).

We can apply a similar idea to the spatial Probit model: if we have the bivariate normal densities $\phi_{2g}(Y_{g1,}Y_{g2}|X_g,\theta)$ correctly specified for each group, we could get a consistent estimator by partial ML. However, there are several differences between panel data and spatial dependent data: first, the panel data model assumes that the cross section dimension (N) is sufficiently large relative to the time dimension (T), but in spatial data we do not have this assumption. Second, in the panel data model, we view the cross section observations as independent, while in the spatial data model, even though we divided the sample into n groups, however, we are definitely not assuming independence among groups. Observations in different groups are still correlated, but the correlations are assumed to decay as distances become further away. Third, as we discussed before, dependence in space is more complicated than dependence in time, and we need to assume that the correlations between groups die out quickly enough as distance goes further away. In short, we need to examine carefully how the weak law of large numbers (WLLN) and central limit theorem (CLT) can be applied in the spatial dependent case. We will discuss these issues and provide proofs in the following sections.

First, we can write the partial log likelihood function as

$$L=\sum_{g=1}^{n}\{Y_{g1}Y_{g2}\log P_g(Y_{g1}=1,Y_{g2}=1|X_g)+Y_{g1}(1-Y_{g2})\log P_g(Y_{g1}=1,Y_{g2}=0|X_g)$$
$$+(1-Y_{g1})Y_{g2}\log P_g(Y_{g1}=0,Y_{g2}=1|X_g)+(1-Y_{g1})(1-Y_{g2})\log P_g(Y_{g1}=0,Y_{g2}=0|X_g)\},\quad g=1,2...n \tag{50}$$

and for the sake of brevity, we define

$$P_g(1,1) \equiv \log P_g(Y_{g1}=1,Y_{g2}=1|X_g);\quad P_g(1,0)\equiv\log P_g(Y_{g1}=1,Y_{g2}=0|X_g); \tag{51}$$
$$P_g(0,1) \equiv \log P_g(Y_{g1}=0,Y_{g2}=1|X_g)\ \text{ and }\ P_g(0,0)\equiv\log P_g(Y_{g1}=0,Y_{g2}=0|X_g). \tag{52}$$

Therefore, we can rewrite the partial log likelihood function as

$$L = \sum_{g=1}^{n} \{Y_{g1}Y_{g2}P_g(1,1) + Y_{g1}(1-Y_{g2})P_g(1,0) + (1-Y_{g1})Y_{g2}P_g(0,1) + (1-Y_{g1})(1-Y_{g2})P_g(0,0)\}. \quad (53)$$

4.1 Consistency of Bivariate Probit Estimation

Consistent estimators $\widehat{\theta} \equiv (\widehat{\beta}, \widehat{\lambda})'$ are the ones that converge in probability to the true value $\theta_0 \equiv (\beta_0, \lambda_0)'$, i.e. $\widehat{\theta} \xrightarrow{p} \theta_0$, as the sample size goes to infinity for all possible true values. In this section, to make the asymptotic arguments formal, we distinguish between the true value, θ_0, and a generic parameter value θ.

In the bivariate probit estimation, the estimator $\widehat{\theta}$ is defined as: $\widehat{\theta}$ maximizes $Q_n(\theta)$ subject to $\theta \in \Theta$, where Θ is the parameters set. The objective function $Q_n(\theta)$ is defined as

$$\begin{aligned} Q_n(\theta) \equiv \frac{1}{n}\sum_{g=1}^{n} \{ & Y_{g1}Y_{g2}P_g(1,1) + Y_{g1}(1-Y_{g2})P_g(1,0) \\ & + (1-Y_{g1})Y_{g2}P_g(0,1) + (1-Y_{g1})(1-Y_{g2})P_g(0,0)\}, \end{aligned} \quad (54)$$

i.e, in other words,

$$\widehat{\theta} = \arg\max_{\theta \in \Theta} Q_n(\theta). \quad (55)$$

Remember that this objective function represents a partial log likelihood, not a fully log likelihood: we are only using information on the conditional distribution $D(Y_{g1}, Y_{g2}|X, W)$ across the groups g. We are not using $D(Y_1, Y_2, ..., Y_n|X, W)$ as in a full maximum likelihood setting.

The identification condition is that $Q(\theta)$ is uniquely maximized at the true value θ_0, where $Q(\theta)$ is defined as

$$Q(\theta) \equiv \lim_{n\to\infty} E[Q_n(\theta)]. \quad (56)$$

This condition typically holds for well-specified models when there is not perfect collinearity among the regressors. Further, one needs to be a little careful in parameterizing the spatial autocorrelation, but standard models of spatial autocorrelation cause no problems.

The following Theorem 1 states the main consistency result. We define $S(\theta) \equiv \frac{\partial Q_n}{\partial \theta}(\theta)$ and $\lim_{n\to\infty} E[S_n(\theta)] = S(\theta)$.

THEOREM 1. *If (i) θ_0 is the interior of a compact set Θ, which is the closure of a concave set, (ii) Q attains a unique maximum over the compact set Θ at θ_0,(iii) Q is continuous on Θ , (iv) the density of observations in any region whose area exceeds a fixed minimum is bounded, (v) as $n \to \infty, \sup_g(\left\| \frac{1}{\Pr(Y_{g1}=1,Y_{g2}=1|X_g)} + \frac{1}{\Pr(Y_{g1}=1,Y_{g2}=0|X_g)} + \frac{1}{\Pr(Y_{g1}=0,Y_{g2}=1|X_g)} + \frac{1}{\Pr(Y_{g1}=0,Y_{g2}=0|X_g)} \right\|) < \infty$, (vi) as $n \to \infty$, $\sup_g(\|X_g\| + \|Y_g\|) = O(1)$, (vii) $\sup_{ngj} |Cov(Y_{gi}, Y_{ji})| \leq \alpha(d_{gj}), i = 1, 2$ where d_{gj} denotes the distance between group g and j, and $\alpha(d) \to 0$ as $d \to \infty$, and (viii) $\lim_{n\to\infty} E[Q_n(\theta)]$ exists, (ix) $\sup_g \|W_g\| < \infty$, then $\widehat{\theta} - \theta_0 = o_p(1)$.*

Proof: Given in Appendix 1.

Condition (i) is a standard assumption from set theory. Condition (ii) is the identification condition for MLE. Condition (iii) assumes that the function Q is continuous in the metric space, which is a reasonable assumption and necessary for the proof that $Q_n(\theta)$ is stochastically equicontinuous. Condition (iv) simply excludes that an infinite number of observations crowd in one bounded area. The minimum area restriction is imposed because an infinitesimal area around a single observation has infinite density. Condition (v) makes sure any one of these four situations will be present in a sufficiently large sample. Condition (vi) makes sure the regressors are deterministic and uniformly bounded, which is not a strong assumption in this literature. Condition (vii) is the key assumption for this theorem, and it requires that the dependence among groups decays sufficiently quick when the distance between groups become further apart. This assumption employs the concept from α-mixing to define the rate of dependence decreasing as distance increases. Condition (viii) assumes the limit of $E[S_n(\theta)]$ exists as $n \to \infty$, which is not a strong assumption. Condition (ix) is actually implied by the rule of dividing groups, which just excludes that the two groups are exactly in the same location.

4.2 Asymptotic Normality

As we discussed in the introduction, the spatial dependence is more complicated than time-series dependence at least in four perspectives. These differences cause that central limit theorems (CLT) need stronger conditions for the spatial dependence case. To deal with general dependence problems, the common way in the literature is to use the so called "Bernstein Sums", which break up S_n into blocks (partial sums), and we consider the sequence of blocks. Each block must be so large, relative to the rate at which the memory of the sequence decays, that the degree to which the next block can be predicted from current information is negligible. But at the same time, the number of blocks must increase with n so that the CLT argument can be applied to this derived sequence (Davidson, 1994).

In this section, we show under what assumptions we are able to apply McLeish's central limit theorem (1974) to spatial dependence cases to get asymptotic normality for the spatial Probit estimator. This is presented in the following Theorem. A^T denotes the transpose of matrix A.

THEOREM 2: *If the assumptions of Theorem 1 hold, and in addition: (i) as* $d \to \infty$, $\frac{d^2\alpha(dd^*)}{\alpha(d^*)} = o(1)$ *for all fixed* $d^* > 0$, *(ii) the sampling area grows uniformly at a rate of* $\sqrt{n}$ *in two non-opposing directions, (iii)* $B(\theta_0) \equiv \lim_{n\to\infty} E[nS_n(\theta_0)S_n^T(\theta_0)]$ *and* $A(\theta_0) \equiv \lim_{n\to\infty} -E[H(\theta_0)]$ *are uniformly positive definite matrices; then* $\sqrt{n}(\widehat{\theta} - \theta_0) \to N[0, A(\theta_0)^{-1}B(\theta_0)A(\theta_0)^{-1}]$, *where* $S_n(\theta_0) \equiv \frac{\partial Q_n}{\partial \theta}(\theta_0)$ and $H(\theta_0) = \frac{\partial^2 Q_n}{\partial\theta\partial\theta^T}(\theta_0)$.

Proof: Given in Appendix 1.

Condition (i) is stronger than condition (vii) in Theorem 1, and it is also stronger than the usual condition in time series data because spatial dependent data has more dimension correlations than time series data. It shows that how dependence decays when distance between groups gets further away, and the dependence decays at the rate fast enough. Condition (ii) just repeats the assumption in the Bernstein's blocking method, the two non-opposing directions just exclude sampling area grows at two parallel directions, which does not make much sense in spatial dependent case. Conditions in (iii) are natural conditions about matrices, which are implied by the previous assumptions. Matrices are semidefinite if some extreme situations happen such as $\Pr(Y_{g1} = 1, Y_{g2} = 1|X_g) = 0$, which are assumed to be excluded in the previous assumptions.

4.3 Estimation of Variance-covariance Matrices

Consistent estimation of the asymptotic covariance matrix is important for the construction of asymptotic confidence intervals and hypothesis tests (Newey and West, 1987). The estimations of A (i.e. $\hat{A} = A(\widehat{\theta})$) are relatively easy, usually just obtaining sample analogues of θ_0 with $\widehat{\theta}$; but the estimation of B (i.e. $\widehat{B} = B(\widehat{\theta})$) is more difficult and more important because of the correlations among groups. Newey-West (1987) proposed a method to estimate the variance-covariance matrix in settings of dependence of infinite order under a covariance stationary condition, and they suggested modified Bartlett weights to make sure the estimated variance and test statistics were positive. Andrews (1991) established the consistency of kernel HAC (Heteroskedasticity and Autocorrelation Consistent) estimators under more general conditions. Pinkse and Slade (1998) also showed that we can obtain $B_n(\widehat{\theta}) - B(\theta_0) = o_p(1)$ under regularity assumptions, where $B_n(\theta) \equiv nE[S_n(\theta)S_n^T(\theta)]$ (see Lemma 9 in Appendix 1). This approach is feasible in practice only if we can get closed form expressions for $E[S_n(\theta)S_n^T(\theta)]$, which should be a function of θ, and then plug in $\widehat{\theta}$ for θ_0 in the function to get consistent covariance estimators. However, it is difficult to get closed form expressions for $B_n(\widehat{\theta})$ in practice, and hence we follow an alternative approach proposed by Conley (1999).

A feasible way to obtain a consistent estimate of a variance-covariance matrix that allows for a wider range of dependence is to apply the approach of Conley (1999) along the lines of Newey-West (1987). We follow this procedure in the following Theorem 3.

Let Ξ_Λ be the $\sigma-$algebra generated by a given random field $\psi_{s_m}, s_m \in \Lambda$ with Λ compact, and let $|\Lambda|$ be the number of $s_m \in \Lambda$. Let $\Upsilon(\Lambda_1, \Lambda_2)$ denote the minimum Euclidean distance from an element of Λ_1 to an element of Λ_2. There exists also a regular lattice index random field W_s^* that is equal to one if location $s \in Z^2$ is sampled and zero otherwise. W_s^* is assumed to be independent of the underlying random field and to have a finite expectation and to be stationary. The mixing coefficient is defined as

$$\begin{aligned} \alpha_{k,l}(n) &\equiv \sup\{|P(A \cap B) - P(A)P(B)|\},\ A \in \Xi_{\Lambda_1}, B \in \Xi_{\Lambda_2} \quad \text{and} \\ |\Lambda_1| &\leq k,\ |\Lambda_2| \leq l,\ \Upsilon(\Lambda_1, \Lambda_2) \geq n. \end{aligned}$$

We also define a new process $R_s(\theta)$ such as

$$R_s(\theta) = \left\{ \begin{array}{cc} S(\theta) & \text{if } W_s^* = 1, \\ 0 & \text{if } W_s^* = 0. \end{array} \right\}$$

Then

THEOREM 3. *If (i) Λ_τ grows uniformly in two non-opposing directions as $\tau \longrightarrow \infty$, (ii) $B(\theta_0) \equiv \lim_{n\to\infty} E[S_n(\theta_0)S_n^T(\theta_0)]$ and $A(\theta_0) \equiv \lim_{n\to\infty} -E[H(\theta_0)]$ are uniformly positive definite matrices, (iii) Y_{gi}, Y_{ji} as defined in Theorem 1, $i = 1, 2$ and W_s^* are mixing where $\alpha_{k,l}(n)$ converges to zero as $n \to \infty$; $S(\theta)$ is Borel measurable for all $\theta \in \Theta$, and continuous on Θ and first moment continuous on Θ, (iv) $\sum_{m=1}^{\infty} m\alpha_{k,l}(m) < \infty$ for $k + l \leq 4$, (v) $\alpha_{1,\infty}(m) = o(m^{-2})$, (vi) for some $\delta > 0$, $E(\|S(\theta_0)\|)^{2+\delta} < \infty$ and $\sum_{m=1}^{\infty} m\alpha_{1,1}(m)^{\delta/(2+\delta)} < \infty$, (vii) $H(\theta)$ is Borel measurable for all $\theta \in \Theta$, continuous on Θ and second moment continuous, $A(\theta_0)$ exists and is full rank, (viii) $\sum_{s\in Z^2} cov(R_0(\theta_0), R_s(\theta_0))$ is a non-singular matrix, (ix) the $K_{MP}(j,k)$ are uniformly bounded and $K_{MP}(j,k) \longrightarrow 1$, $n_\tau \longrightarrow \infty$ as $\tau \longrightarrow \infty$ $(M, P \longrightarrow \infty)$, $L_M = o\left(M^{1/3}\right)$ and $L_P = o\left(P^{1/3}\right)$, (x) for some $\delta > 0$, $E(\|S(\theta_0)\|)^{4+\delta} < \infty$ and Y_{gi}, Y_{ji} as defined in Theorem 1, $i = 1, 2$ and W_s^* are mixing where $\alpha_{\infty,\infty}(m)^{\delta/(2+\delta)} = o(m^{-4})$, (xi) $E\sup_\Theta \|R_{m,p}(\theta)\|^2 < \infty$ and $E\sup_\Theta \|(\partial/\partial\theta)[R_{m,p}(\theta)]\|^2 < \infty$, then*

$$\widehat{B}_\tau - B(\theta_0) = o_p(1) \text{ as } \tau \longrightarrow \infty,$$

where we split $s = [m, p]$, Λ_τ is a rectangle so that $m \in \{1, 2, ..., M\}$ and $p \in \{1, 2, ..., P\}$ and

$$\begin{aligned} \widehat{B}_\tau &= n_\tau^{-1} \sum_{j=0}^{L_M} \sum_{k=0}^{L_P} \sum_{m=j+1}^{M} \sum_{p=k+1}^{P} K_{MP}(j,k) \left(\begin{array}{c} R_{m,p}\left(\widehat{\theta}\right) R_{m-j,p-k}\left(\widehat{\theta}\right)^T + \\ R_{m-j,p-k}\left(\widehat{\theta}\right) R_{m,p}\left(\widehat{\theta}\right)^T \end{array} \right) \\ &\quad -n_\tau^{-1} \sum_{m=1}^{M} \sum_{p=1}^{P} R_{m,p}\left(\widehat{\theta}\right) R_{m,p}\left(\widehat{\theta}\right)^T. \end{aligned}$$

To ensure positive semi-definite covariance matrix estimates, we need to choose an appropriate two-dimensional weights function that is a Bartlett window in each dimension

$$K_{MP}(j,k) = \left\{ \begin{array}{ll} (1 - \frac{|j|}{L_M})(1 - \frac{|k|}{L_P}) & for |j| < L_M, |k| < L_P \\ 0 & \text{else} \end{array} \right\}.$$

Proof: It follows from Conley (1999), Proposition 3.

5 Simulation Study

In the previous section, we have proved that the partial maximum likelihood estimator (PMLE) based on the bivariate normal distribution is consistent and asymptotically normal. Moreover, one

of the most attractive properties of our new PMLE is that we can get a more efficient estimator compared to the GMM estimator, and the approach is much less computational demanding when compared to full information methods. In order to learn about the gains in efficiency that we obtain in the context of a Bivariate Spatial Probit model when using PML versus GMM, we conduct in this Section a simulation study to show the efficiency gains of PML.

5.1 Simulation Design and Results

Instead of comparing our PMLE to the GMM estimator of Pinkse and Slade (1998) directly, we choose to compare the PMLE to the heteroskedastic Probit estimator (HPE) because of two reasons: First, the HPE uses similar information with the GMM estimator because both methods use generalized residuals from the Probit estimation to construct the moment conditions, which means that both methods use the information from the heterogeneities of the diagonal terms of the variance-covariance matrix, while our PMLE uses both diagonal and off-diagonal correlations information between two closest neighbors. Second, the STATA[4] source codes for bivariate probit estimation and heteroskedastic Probit estimation are available online, and we can easily add the spatial parts into these existing source codes to compare PML estimators with Heteroskedastic Probit Estimators.

According to the theoretical framework given in previous sections, we could generate a dataset which allows a general correlation structure across groups as equations (8) and (9), and it requires to specify the exact formula (as functions of λ and W) for the elements of Ω_g. However, it is quite difficult to derive the pairwise covariances for a bivariate probit because the exact formula for Ω_{g12} (and of $\Omega_{g11}, \Omega_{g22}$) is very complicated, which is an element of the inverse matrix with $2n$ spatially correlated observations as follows

$$\Omega_g = \begin{bmatrix} \Omega_{g11} & \Omega_{g12} \\ \Omega_{g21} & \Omega_{g22} \end{bmatrix} = [(I-\lambda W)'(I-\lambda W)]_g^{-1} = \begin{bmatrix} \Omega_{111} & ... & ... & ... & ... \\ ... & ... & ... & ... & ... \\ ... & ... & \Omega_{g11} & \Omega_{g12} & ... \\ ... & ... & \Omega_{g21} & \Omega_{g22} & ... \\ ... & ... & ... & ... & \Omega_{n22} \end{bmatrix}. \tag{57}$$

Therefore, it seems reasonable to do the following. Let R be the weighting matrix which can be generated in STATA[5] according to the distance between observations

$$Y_i^* = X_{i1}\beta_1 + X_{i2}\beta_2 + X_{i3}\beta_3 + \varepsilon_i \tag{58}$$

$$\varepsilon = \lambda R u, \tag{59}$$

[4]See http://www.stata.com/

[5]The STATA command is "Spatwmat". Since the speed to calculate the inverse of a matrix is much slower as the size of matrix increases, and moreover the maximum matrix size in Stata is 800, we allow here each observation to be spatially correlated to nearby 99 observations.

where $u \sim Normal(0, I_n)$. The weighting matrix R is standardized so that the diagonal elements are ones, and then the elements of R shrink as distance is increasing . The reason we do this is because it is easier to determine $Var(\varepsilon_i)$ and $Cov(\ \varepsilon_i, \varepsilon_j)$ to apply the HP and the bivariate probit estimators. In this way, we still allow general correlation across groups, and we are able to compare the efficiency gains from only using the diagonal information (the HP approach) to using both diagonal and off-diagonal information (bivariate probit), and we do not require to know the exact formula for the elements in Ω_g (given in equation (57)) to reach the same goal.

Therefore, we generate the dataset according to equations (58) and (59), which allows spatial correlation between any two observations, and we set the true parameter values for β_1, β_2 and β_3 equal to 1, 1 and 1 respectively. Since our main focus in this study is on the estimation of the spatial parameter λ, we also set different λ true values for each simulated sample: $\lambda = 0.2; 0.4; 0.6$ and 0.8, to test for the performance of the two estimation methods (PML and HP). These values for λ are in the range of the estimated value in the empirical application of Pinkse and Slade (1998). In this setting and with 1000 replications, we consider a sample size of $N = 1000$ observations (where the sample size is divided into 500 pairwise groups). Finally, we also simulate samples of sizes 500 and 1500 (with 250 and 750 pairwise groups respectively) to check the performance of the two methods in different samples sizes. The simulation results are reported in Tables 1 (for the spatial parameter λ) and 2 (for the β_1, β_2 and β_3) in Appendix 2.

From Table 2, we can observe that both the HPE and the PMLE of β_1, β_2 and β_3 converge to true parameter values across the different parameter values as sample size increasing. Also the the PML estimator has much less bias than the HPE. Moreover, as expected, PML always provides smaller standard errors than the HP estimation method and bias and standard errors decrease in general when sample size increases.

Furthermore, it is in Table 1 where we can observe the largest advantages of using PML versus HP. We can see that the PMLE is much better than the HPE in terms of estimating the spatial parameter λ. The PMLE is always much closer to the true parameter values and with small standard errors across different sample sizes and parameter values (as expected from our theoretical results), while the HPE is much further away from true parameter values and it is has a much larger standard deviation over the different sample sizes, even though HPE also shows the trend to converge to the true values in general as the sample increases. The HPE has always much larger standard deviation than the PMLE, showing clearly the gains in efficiency of PML versus HPE/GMM as predicted by our theory. Since both the HPE and the GMM estimator use generalized residuals from Probit estimation to construct the moment conditions, we conjecture that the GMM estimator is subject to similar inefficiency problems in estimating the spatial coefficient. Also, as it is expected, the bias of the PMLE decreases when N increases.

In summary, from the simulation results of Tables 1 and 2, we see how the PMLE outperforms clearly the HPE (i.e., the GMM estimator of Pinkse and Slade (1998)), specially when estimating the spatial parameter λ, which implies that the PMLE is much more robust and efficient in the context

of the spatial probit model. The simulation results provide clear evidence of the gains in efficiency that can be obtained by PML versus GMM, as predicted by our theoretical results in the previous section.

6 Conclusions

The idea of this paper is simple and intuitive: instead of just using information in moment conditions (GMM), we divide observations into pairwise groups. Provided we correctly specify the conditional joint distribution within these pairwise groups, we show that the spatial bivariate Probit model allows us to use the most important information of spatial correlations among adjacent observations and to get more efficient estimators. We also prove that partial MLE is consistent and asymptotically normal under some regularity conditions. We also discuss how to get consistent covariance matrix estimators under general spatial dependence by following the approach of Conley (1999) and Newey-West (1987), which is more usable in practice compared to the proposal of Pinkse and Slade (1998). The attractive part of this study is that we can get a more efficient partial ML estimator without introducing stronger assumptions (in some sense, we need weaker assumptions than the GMM method), and the approach is much less computational demanding compared to full information methods. In order to learn about the gains in efficiency that we obtain in the bivariate Probit model with PMLE versus the GMM estimator, we provide a simulation study in Section 5. The advantages in terms of bias and efficiency of our new estimation procedure proposed in this paper are clearly demonstrated. Moreover, if we extend this method to the trivariate or higher dimensional multivariate Probit models, we can obtain even more efficient estimators, but it comes at the expense of more computational demands.

7 Appendix 1

7.1 Proofs to Theorems

Proof of Theorem 1. If we can prove that $Q_n(\theta) \xrightarrow{p} Q(\theta)$ uniformly, by the information inequality, $Q(\theta)$ has a unique maximum at the true parameter when θ_0 is identified. Then under technical conditions for the limit of the maximum to be the maximum of the limit, $\widehat{\theta}$ should converge in probability to θ_0. Sufficient conditions for the maximum of the limit to be the limit of maximum are that the convergence in probability is uniform and the parameter set is compact (Newey, 1994).

To prove consistency, the proof includes three parts:

(i) Q has a unique maximum at θ_0.

(ii) $Q_n(\theta) - Q(\theta) = o_p(1)$ at all $\theta \in \Theta$.

(iii) $Q_n(\theta)$ is stochastically equicontinuous and Q is continuous on Θ.

Condition (i) and Q to be continuous on Θ are assumed. The proof of condition (ii) is provided in Lemma 1, and the proof that $Q_n(\theta)$ is stochastically equicontinuous can be found in Lemma 2.

Q.E.D. ■

Proof of Theorem 2. To find out the asymptotic normality of the Partial MLE for spatial bivariate Probit model, we start the proof from mean value theorem. Since $\frac{\partial Q_n}{\partial \theta}(\widehat{\theta}) = 0$, and by using the mean value theorem

$$\begin{aligned} \frac{\partial Q_n}{\partial \theta}(\widehat{\theta}) &= 0 = \frac{\partial Q_n}{\partial \theta}(\theta_0) + \frac{\partial^2 Q_n}{\partial \theta \partial \theta^T}(\theta^*)(\widehat{\theta} - \theta_0) \quad (60) \\ &\Rightarrow (\widehat{\theta} - \theta_0) = -[\frac{\partial^2 Q_n}{\partial \theta \partial \theta^T}(\theta^*)]^{-1} \frac{\partial Q_n}{\partial \theta}(\theta_0), \quad (61) \end{aligned}$$

where θ^* lies between $\widehat{\theta}$ and θ_0.

First, let us discuss the term $\frac{\partial^2 Q_n}{\partial \theta \partial \theta^T}(\theta^*)$ to find out the asymptotic properties of $\frac{\partial^2 Q_n}{\partial \theta \partial \theta^T}(\theta^*)$. Recall that

$$\begin{aligned} Q_n(\theta) &= \frac{1}{n}\sum_{g=1}^{n}\{Y_{g1}Y_{g2}P_g(1,1) + Y_{g1}(1-Y_{g2})P_g(1,0) \\ &\quad +(1-Y_{g1})Y_{g2}P_g(0,1) + (1-Y_{g1})(1-Y_{g2})P_g(0,0)\}, \quad (62) \end{aligned}$$

where $P_g(1,1) \equiv \log P_g(Y_{g1} = 1, Y_{g2} = 1|X_g)$ etc. Also

$$\begin{aligned} \frac{\partial^2 Q_n}{\partial \theta \partial \theta^T}(\theta) &= \frac{1}{n}\sum_{g=1}^{n}\{Y_{g1}Y_{g2}\frac{\partial^2 P_g(1,1)}{\partial \theta \partial \theta^T} + Y_{g1}(1-Y_{g2})\frac{\partial^2 P_g(1,0)}{\partial \theta \partial \theta^T} \\ &\quad +(1-Y_{g1})(Y_{g2})\frac{\partial^2 P_g(0,1)}{\partial \theta \partial \theta^T} + (1-Y_{g1})(1-Y_{g2})\frac{\partial^2 P_g(0,0)}{\partial \theta \partial \theta^T}, \quad (63) \end{aligned}$$

where

$$\begin{aligned} \frac{\partial^2 P_g(1,1)}{\partial \theta \partial \theta^T} &= \frac{-1}{[\Pr(Y_{g1}=1, Y_{g2}=1|X_g)]^2}[\frac{\partial \Pr(Y_{g1}=1, Y_{g2}=1|X_g)}{\partial \theta}]^2 \\ &\quad +\frac{1}{\Pr(Y_{g1}=1, Y_{g2}=1|X_g)}\frac{\partial^2[\Pr(Y_{g1}=1, Y_{g2}=1|X_g)]}{\partial \theta \partial \theta^T}, \quad (64) \end{aligned}$$

and all other terms behave similar.

As before, we only discuss one of these terms, and the same logic applies to the other terms. We know that

$$\begin{aligned} &\frac{1}{n}\sum_{g=1}^{n}[Y_{g1}Y_{g2}\frac{\partial^2 P_g(1,1)}{\partial \theta \partial \theta^T}(\theta^*)] \\ = &\frac{1}{n}\sum_{g=1}^{n}Y_{g1}Y_{g2}\{\frac{-1}{[\Pr(Y_{g1}=1, Y_{g2}=1|X_g)]^2}[\frac{\partial \Pr(Y_{g1}=1, Y_{g2}=1|X_g)}{\partial \theta}(\theta^*)]^2 \\ &+\frac{1}{\Pr(Y_{g1}=1, Y_{g2}=1|X_g)}\frac{\partial^2[\Pr(Y_{g1}=1, Y_{g2}=1|X_g)]}{\partial \theta \partial \theta^T}(\theta^*)\}. \quad (65) \end{aligned}$$

Look at the first term of the above equation given by

$$\frac{1}{n}\sum_{g=1}^{n} Y_{g1}Y_{g2}\{\frac{-1}{[\Pr(Y_{g1}=1,Y_{g2}=1|X_g)]^2}[\frac{\partial \Pr(Y_{g1}=1,Y_{g2}=1|X_g)}{\partial\theta}(\theta^*)]^2\}. \tag{66}$$

Since $\left\|\frac{1}{[\Pr(Y_{g1}=1,Y_{g2}=1|X_g)]^2}\right\| < \infty$, we can write this term as

$$\frac{1}{n}\sum_{g=1}^{n} K_{g11}[\frac{\partial \Pr(Y_{g1}=1,Y_{g2}=1|X_g)}{\partial\theta}(\theta^*)]^2, \tag{67}$$

where $K_{g11} \equiv Y_{g1}Y_{g2}\frac{-1}{[\Pr(Y_{g1}=1,Y_{g2}=1|X_g)]^2}$.

In order to prove

$$\frac{1}{n}\sum_{g=1}^{n} K_{g11}[\frac{\partial \Pr(Y_{g1}=1,Y_{g2}=1|X_g)}{\partial\theta}(\theta^*)]^2 \xrightarrow{p} \frac{1}{n}\sum_{g=1}^{n} K_{g11}[\frac{\partial \Pr(Y_{g1}=1,Y_{g2}=1|X_g)}{\partial\theta}(\theta_0)]^2, \tag{68}$$

we need to show that it holds for all $\|\varpi\| = 1$. Set $\overline{K_{g11}} = \varpi^T K_g$ and then

$$\begin{aligned}
&\varpi^T\{\frac{1}{n}\sum_{g=1}^{n} K_{g11}[\frac{\partial \Pr(Y_{g1}=1,Y_{g2}=1|X_g)}{\partial\theta}(\widehat{\theta})]^2 \\
&-\frac{1}{n}\sum_{g=1}^{n} K_{g11}[\frac{\partial \Pr(Y_{g1}=1,Y_{g2}=1|X_g)}{\partial\theta}(\theta_0)]^2\} &(69)\\
=\ &\frac{1}{n}\sum_{g=1}^{n} \overline{K_{g11}}\{[\frac{\partial \Pr(Y_{g1}=1,Y_{g2}=1|X_g)}{\partial\theta}(\widehat{\theta})]^2 - [\frac{\partial \Pr(Y_{g1}=1,Y_{g2}=1|X_g)}{\partial\theta}(\theta_0)]^2\} &(70)\\
=\ &(\widehat{\theta}-\theta_0)\frac{2}{n}\sum_{g=1}^{n} \overline{K_{g11}}\frac{\partial \Pr(Y_{g1}=1,Y_{g2}=1|X_g)}{\partial\theta}(\theta^*) \times \frac{\partial^2 \Pr(Y_{g1}=1,Y_{g2}=1|X_g)}{\partial\theta\partial\theta^T}(\theta^*). &(71)
\end{aligned}$$

From the proof of Theorem 1, we know that $\sup_g \left\|\frac{\partial \Pr(Y_{g1}=1,Y_{g2}=1|X_g)}{\partial\theta}\right\| < \infty$. From Lemma 3, $\sup_g \left\|\frac{\partial^2 \Pr(Y_{g1}=1,Y_{g2}=1|X_g)}{\partial\theta\partial\theta^T}\right\| < \infty$. From Theorem 1, we also know that $\widehat{\theta}-\theta_0 = o_p(1)$ and hence

$$(\widehat{\theta}-\theta_0)\frac{2}{n}\sum_{g=1}^{n} \overline{K_{g11}}\frac{\partial \Pr(Y_{g1}=1,Y_{g2}=1|X_g)}{\partial\theta}(\theta^*) \times \frac{\partial^2 \Pr(Y_{g1}=1,Y_{g2}=1|X_g)}{\partial\theta\partial\theta^T}(\theta^*) = o_p(1) \tag{72}$$

$$\Longrightarrow \varpi^T\{\frac{1}{n}\sum_{g=1}^{n} K_{g11}[\frac{\partial \Pr(Y_{g1}=1,Y_{g2}=1|X_g)}{\partial\theta}(\widehat{\theta})]^2 - \frac{1}{n}\sum_{g=1}^{n} K_{g11}[\frac{\partial \Pr(Y_{g1}=1,Y_{g2}=1|X_g)}{\partial\theta}(\theta_0)]^2\} = o_p(1) \tag{73}$$

$$\Longrightarrow \frac{1}{n}\sum_{g=1}^{n} K_{g11}[\frac{\partial \Pr(Y_{g1}=1,Y_{g2}=1|X_g)}{\partial\theta}(\theta^*)]^2 \xrightarrow{p} \frac{1}{n}\sum_{g=1}^{n} K_{g11}[\frac{\partial \Pr(Y_{g1}=1,Y_{g2}=1|X_g)}{\partial\theta}(\theta_0)]^2. \tag{74}$$

By definition,

$$\lim_{n\to\infty}\frac{1}{n}\sum_{g=1}^{n} K_{g11}[\frac{\partial \Pr(Y_{g1}=1,Y_{g2}=1|X_g)}{\partial\theta}(\theta_0)]^2 = E\{K_{g11}[\frac{\partial \Pr(Y_{g1}=1,Y_{g2}=1|X_g)}{\partial\theta}(\theta_0)]^2\}, \tag{75}$$

and therefore,

$$\frac{1}{n}\sum_{g=1}^{n} Y_{g1}Y_{g2}\{\frac{-1}{[\Pr(Y_{g1}=1,Y_{g2}=1|X_g)]^2}[\frac{\partial \Pr(Y_{g1}=1,Y_{g2}=1|X_g)}{\partial\theta}(\theta^*)]^2\} \xrightarrow{p} \tag{76}$$

$$E\{Y_{g1}Y_{g2}\frac{-1}{[\Pr(Y_{g1}=1,Y_{g2}=1|X_g)]^2}[\frac{\partial \Pr(Y_{g1}=1,Y_{g2}=1|X_g)}{\partial\theta}(\theta_0)]^2\}. \tag{77}$$

Similarly, we can prove in relation to the second term that

$$\frac{1}{n}\sum_{g=1}^{n} Y_{g1}Y_{g2}\frac{1}{\Pr(Y_{g1}=1,Y_{g2}=1|X_g)}\frac{\partial^2[\Pr(Y_{g1}=1,Y_{g2}=1|X_g)]}{\partial\theta\partial\theta^T}(\theta^*) \tag{78}$$

$$\xrightarrow{p} E\{Y_{g1}Y_{g2}\frac{1}{\Pr(Y_{g1}=1,Y_{g2}=1|X_g)}\frac{\partial^2[\Pr(Y_{g1}=1,Y_{g2}=1|X_g)]}{\partial\theta\partial\theta^T}(\theta_0)\}. \tag{79}$$

As usual, we apply repeatedly the above arguments to the other terms. Finally, we can get that

$$\lim_{n\to\infty}\frac{\partial^2 Q_n}{\partial\theta\partial\theta^T}(\theta^*) \xrightarrow{p} E[\frac{\partial^2 Q_n}{\partial\theta\partial\theta^T}(\theta_0)]. \tag{80}$$

If we define

$$\begin{aligned} H \equiv\ & \{Y_{g1}Y_{g2}\frac{\partial^2 P_g(1,1)}{\partial\theta\partial\theta^T} + Y_{g1}(1-Y_{g2})\frac{\partial^2 P_g(1,0)}{\partial\theta\partial\theta^T} \\ & +(1-Y_{g1})(Y_{g2})\frac{\partial^2 P_g(0,1)}{\partial\theta\partial\theta^T} + (1-Y_{g1})(1-Y_{g2})\frac{\partial^2 P_g(0,0)}{\partial\theta\partial\theta^T}\} \end{aligned} \tag{81}$$

where H denotes the Hessian, equation (81) can be rewritten as

$$\lim_{n\to\infty}\frac{1}{n}\sum_{g=1}^{n} H(\theta^*) \xrightarrow{p} \lim_{n\to\infty} E[H(\theta_0)]. \tag{82}$$

Therefore, it remains to show the asymptotic normality of the score term: $\frac{\partial Q_n}{\partial\theta}(\theta_0)$. For the sake of brevity, redefine the score as: $S_n(\theta_0) \equiv \frac{\partial Q_n}{\partial\theta}(\theta_0)$. Then

$$\begin{aligned} S_n(\theta_0) =\ & \frac{1}{n}\sum_{g=1}^{n}\{Y_{g1}Y_{g2}\frac{\partial P_g(1,1)}{\partial\theta}(\theta_0) + Y_{g1}(1-Y_{g2})\frac{\partial P_g(1,0)}{\partial\theta}(\theta_0) \\ & +(1-Y_{g1})Y_{g2}\frac{\partial P_g(0,1)}{\partial\theta}(\theta_0) + (1-Y_{g1})(1-Y_{g2})\frac{\partial P_g(0,0)}{\partial\theta}(\theta_0)\}. \end{aligned} \tag{83}$$

We need to show that $B^{-\frac{1}{2}}(\theta_0)S_n(\theta_0) \to N(0, I_K)$, where $B(\theta) \equiv \lim_{n\to\infty} nE[S_n(\theta)S_n^T(\theta)]$. Note that the information matrix equality does not hold here, i.e. $-E[H(\theta_0)] \neq E[S_n(\theta)S_n^T(\theta)]$, because the score terms are correlated with each other over space. In this part, we follow Pinkse and Slade (1998) and we use Bernstein's blocking methods and the McLeish's (1974) central limit theorem for dependent processes. First, define $T_{na_n} \equiv \Pi_{j=1}^{a_n}(1+i\gamma D_{n,j})$, where $i^2=-1$, and $D_{n,j}(j=1,2...a_n)$ is

an array of random variables on the probability triple $(\Omega, \digamma, P)$.γ is a real number. McLeish's (1974) central limit theorem for dependent processes requires the following four conditions

(i) $\{T_{na_n}\}$is uniformly integrable,

(ii) $ET_{na_n} \to 1$,

(iii) $\sum_{j=1}^{a_n} D_{n,j}^2 \xrightarrow{p} 1$,

(iv) $\underset{j\leq a_n}{Max}|D_{n,j}| \xrightarrow{p} 0$.

Now we need to define $D_{n,j}$ in our case. Let $Y_{0n} \equiv \varpi^T\{\frac{\sqrt{n}S_g(\theta_0)}{\sqrt{B(\theta_0)}}\} = n^{-\frac{1}{2}}\sum_{t=1}^{n} A_{nt}$ for implicitly define A_{nt}. In order to prove $Y_{0n} \xrightarrow{d} N(0,1)$, we need to establish that the property holds for all $\|\varpi\| = 1$ using the Cramer-Wold device. As in the proof of Theorem 1, we split the region in which observations are located up to an a_n area of size $\sqrt{b_n} \times \sqrt{b_n}$. We also know that a_n increases faster than $\sqrt{n}$ and b_n slower, where a_n and b_n are integers such that $a_n b_n = n$. Let a_n and b_n be constructed such that $\alpha(\sqrt{b_n})a_n \to 0$. Let $n^{\tau-\frac{1}{2}} \times b_n < 1$, uniformly in n, for some fixed $0 < \tau < \frac{1}{2}$. Let Λ_{nj} denote the set of indices corresponding to the observations in area j. By assumption a number $C > 0$ exists such that $Max_j(\#\Lambda_{nj}) < Cb_n$. Define $D_{n,j} \equiv n^{-\frac{1}{2}}\sum_{t\in\Lambda_{nj}} A_{nt}$, and hence we can write $Y_{0n} = \sum_{j=1}^{a_n} D_{nj}$.

Now we are ready to discuss the four conditions for Mcleish's (1974) central limit theorem. First, look at condition (iv), which requires that $\underset{j\leq a_n}{Max}|D_{n,j}| = o_p(1)$

$$\underset{j\leq a_n}{Max}|D_{n,j}| = \underset{j\leq a_n}{Max}|n^{-\frac{1}{2}}\sum_{t\in\Lambda_{nj}} A_{nt}|. \tag{84}$$

Since by assumption

$$Max_j(\#\Lambda_{nj}) < Cb_n \Rightarrow \underset{j\leq a_n}{Max}|n^{-\frac{1}{2}}\sum_{t\in\Lambda_{nj}} A_{nt}| \leq Cb_n \times n^{-\frac{1}{2}}\sup\|A_{nt}\|, \tag{85}$$

where # denotes the number of objects, by definition we have that

$$\begin{aligned}\varpi^T\{\frac{\sqrt{n}S_g(\theta_0)}{\sqrt{B(\theta_0)}}\} &= n^{-\frac{1}{2}}\sum_{t=1}^{n} A_{nt}, \sum_{t=1}^{n} A_{nt} = \varpi^T\frac{1}{\sqrt{B_0}}\{Y_{g1}Y_{g2}\frac{\partial P_g(1,1)}{\partial\theta}(\theta_0) + Y_{g1}(1-Y_{g2})\frac{\partial P_g(1,0)}{\partial\theta}(\theta_0) \\ &\quad +(1-Y_{g1})Y_{g2}\frac{\partial P_g(0,1)}{\partial\theta}(\theta_0) + (1-Y_{g1})(1-Y_{g2})\frac{\partial P_g(0,0)}{\partial\theta}(\theta_0)\}. \end{aligned}\tag{86}$$

Since $B(\theta_0)$ is positive definite, $B(\theta_0)^{-\frac{1}{2}}$is bounded for sufficiently large n, and we have that $\sup_g\|Y_{gn}\| < \infty$ by assumption (vi) in Theorem 1. We have also proved that $\sup_g\left\|\frac{\partial P_g(1,1)}{\partial\theta}\right\| < \infty$ in Lemma 2. Therefore, we are able to prove that $\sup\|A_{nt}\| < \infty$. Then $Cb_n \times n^{-\frac{1}{2}}\sup\|A_{nt}\| = O_p(Cb_n \times n^{-\frac{1}{2}}) = o_p(1)$ by construction of b_n. Hence we can get that $\underset{j\leq a_n}{Max}|D_{n,j}| = o_p(1)$.

Second, let us discuss condition (i): $\{T_{na_n}\}$ is uniformly integrable. Following Davidson (1994), if a random variable is integrable, the contribution to the integer of extreme random variable values must be negligible. In other words, if $E|T_{na_n}| < \infty, E(|T_{na_n}|1_{|T_{na_n}|>K}) \to 0$, as $K \to \infty$, it is

equivalent to say $P[\sup_{n>N}|T_{na_n}|>K]=0$, for some $K>0$ as $n\to\infty$. Here we follow the proof of Lemma 10 in Pinkse and Slade (1998). We have that

$$P[\sup_{n>N}|T_{na_n}|>K]=P[\sup_{n>N}|\Pi_{j=1}^{a_n}(1+i\gamma D_{n,j})|>K] \tag{87}$$

$$\leq P[\sup_{n>N}|\Pi_{j=1}^{a_n}(\sqrt{1+\gamma^2D_{n,j}^2})|>K] \tag{88}$$

$$\begin{aligned}&=\{P[\sup_{n>N}|\Pi_{j=1}^{a_n}(\sqrt{1+\gamma^2D_{n,j}^2})|>K|(\sup_{n>N,j}n^\tau|D_{nj}|\leq C)]\times P[\sup n^\tau|D_{nj}|\leq C]\\&+P[\sup_{n>N}|\Pi_{j=1}^{a_n}(\sqrt{1+\gamma^2D_{n,j}^2})|>K|(\sup_{n>N,j}n^\tau|D_{nj}|>C)]\times P[\sup n^\tau|D_{nj}|>C]\}\end{aligned} \tag{89}$$

$$\leq\{P[\sup_{n>N}|\Pi_{j=1}^{a_n}(\sqrt{1+\gamma^2D_{n,j}^2})|>K|(\sup_{n>N,j}n^\tau|D_{nj}|\leq C)]+P[\sup n^\tau|D_{nj}|>C] \tag{90}$$

where C is a uniform upper bound to $\sum_{t\in\Lambda_{nj}}A_{nt}$. Therefore,

$$P[\sup n^\tau|D_{nj}| > C]=P[\sup n^\tau|n^{-\frac{1}{2}}\sum_{t\in\Lambda_{nj}}A_{nt}|>C] \tag{91}$$

$$= P[\sup n^{\tau-\frac{1}{2}}\sum_{t\in\Lambda_{nj}}|A_{nt}|>C]\leq P[\sup n^{\tau-\frac{1}{2}}b_n\sum_{t\in\Lambda_{nj}}|A_{nt}|>C]=0 \tag{92}$$

since $n^{\tau-\frac{1}{2}}b_n<1$ and by construction of b_n.Then,

$$P[\sup_{n>N}|\Pi_{j=1}^{a_n}(\sqrt{1+\gamma^2D_{n,j}^2})|>K|(\sup_{n>N,j}n^\tau|D_{nj}|\leq C)]\leq P[\sup_{n>N}|(1+\gamma^2n^{-2\tau}C^2)^{\frac{a_n}{2}}|>K]=0 \tag{93}$$

provided we set K sufficiently large. Therefore, we proved that $P[\sup_{n>N}|T_{na_n}|>K]=0\Rightarrow\{T_n\}$ is uniformly integrable.

Third, condition (ii) requires that $ET_{na_n}\to 1$, which is equivalent to say that $ET_{na_n}-1=o(1)$; see proof in Lemma 4.

Fourth, in order to prove (iii): $\sum_{j=1}^{a_n}D_{n,j}^2\xrightarrow{p}1$, by Lemma 8, $\sum_{j=1}^{a_n}D_{n,j}^2-1=\sum_{j=1}^{a_n}E(D_{n,j}^2)-1+o_p(1)$ and

$$\sum_{j=1}^{a_n}E(D_{n,j}^2)-1+o_p(1)=E(Y_{0n}^2)-1-\sum_{i\neq j}^{a_n}E(D_{ni}D_{nj})+o_p(1)=o_p(1), \tag{94}$$

by construction of Y_{0n}, since $E(Y_{0n}^2)=1$. It remains to show that $\sum_{i\neq j}^{a_n}E(D_{ni}D_{nj})=o(1)$. This condition is proved in Lemmas 5-7[6]. *Q.E.D.* ■

[6]Lemmas 5-8 are along the lines of those in Pinkse and Slade (1998), which are a simplified version of the proofs in Davidson (1994).

7.2 Technical Lemmas

The proofs of Theorems 1-2 require the use of the following Lemmas 1-8.

LEMMA 1: *Under the assumptions in Theorem 1,* $Q_n(\theta) - Q(\theta) = o_p(1)$ *for all* $\theta \in \Theta$.

Proof: we can rewrite $Q_n(\theta)$ as

$$\begin{aligned} Q_n(\theta) &= \frac{1}{n}\sum_{g=1}^{n}\{Y_{g1}Y_{g2}[P_g(1,1) - P_g(1,0) - P_g(0,1) + P_g(0,0)] \\ &\quad + Y_{g1}[P_g(1,0) - P_g(0,0)] + Y_{g2}[P_g(0,1) - P_g(0,0)] + P_g(0,0)\}. \end{aligned} \tag{95}$$

Since we assume that $\lim_{n\to\infty} E[Q_n(\theta)]$ exists, and by definition $Q(\theta) \equiv \lim_{n\to\infty} E[Q_n(\theta)]$, this implies that: $Q(\theta) - E[Q_n(\theta)] = o(1)$. In order to prove $Q_n(\theta) - Q(\theta) = o_p(1)$, we only need to show that $Q_n(\theta) - E[Q_n(\theta)] = o_p(1)$. That is equivalent to prove that the distance between $Q_n(\theta)$ and $E[Q_n(\theta)]$ is infinitely small as $n \to \infty$. That is: $E\|Q_n(\theta) - E[Q_n(\theta)]\|^2 \to 0$ as $n \to \infty$, and by definition, it is equivalent to $Var[Q_n(\theta)] \to 0$ as $n \to \infty$.

It is easy to see that

$$\begin{aligned} &Var_{ngj}[Q_n(\theta)] = \\ &\frac{1}{n^2}\sum_{g=1}^{n}\sum_{j=1}^{n}\{\gamma_{ng1}\gamma_{nj1}cov(Y_{g1}Y_{g2}, Y_{j1}Y_{j2}) + 2\gamma_{ng1}\gamma_{nj2}cov(Y_{g1}Y_{g2}, Y_{j1}) + 2\gamma_{ng1}\gamma_{nj3}cov(Y_{g1}Y_{g2}, Y_{j2}) \\ &\quad + \gamma_{ng2}\gamma_{nj2}cov(Y_{g1}, Y_{j1}) + 2\gamma_{ng2}\gamma_{nj3}cov(Y_{g1}, Y_{j2}) + \gamma_{ng3}\gamma_{nj3}cov(Y_{g2}, Y_{j2}), \end{aligned} \tag{96}$$

where $\gamma_{ng1} = [P_g(1,1) - P_g(1,0) - P_g(0,1) + P_g(0,0)], \gamma_{ng2} = [P_g(1,0) - P_g(0,0)]$, and $\gamma_{ng3} = [P_g(0,1) - P_g(0,0)]$. The same definition applies to γ_{nj1}, $\gamma_{nj2}$and γ_{nj3}.

Note that here

$$P_g(1,1) = \log\{\int_{-X_{g2}\beta}^{\infty} \Phi(\frac{X_{g1}\beta + \delta_{g1}\varepsilon_{g2}}{\sqrt{Var(e_{g1})}})\phi(\frac{\varepsilon_{g2}}{\sqrt{Var(\varepsilon_{g2})}})d\varepsilon_{g2}\} \tag{97}$$

which is not a function of Y_g or Y_j. Hence γ_{ng1} is not a function of Y_g or Y_j.The same logic applies to the other terms (γ_{ng2}, γ_{ng3}, $\gamma_{nj1}, \gamma_{nj2}$ and γ_{nj3}). Since $0 \le P_g(1,1) \le 1$, the same applies to $P_g(1,0)$, $P_g(0,1)$ and $P_g(0,0)$. Therefore, it is easy to see that $|\gamma_{ngi}| \le 2$, and the same $|\gamma_{nji}|$, and hence $|\gamma_{ngi}\gamma_{nji}| \le 4$, $i = 1, 2$.

Therefore, we can write

$$\begin{aligned} &Sup_{ngj}|Var[Q_n(\theta)]| = \\ &\frac{1}{n^2}\sum_{g=1}^{n}\sum_{j=1}^{n}\{4cov(Y_{g1}Y_{g2}, Y_{j1}Y_{j2}) + 8cov(Y_{g1}Y_{g2}, Y_{j1}) + 8cov(Y_{g1}Y_{g2}, Y_{j2}) \\ &\quad + 4cov(Y_{g1}, Y_{j1}) + 8cov(Y_{g1}, Y_{j2}) + 4cov(Y_{g2}, Y_{j2}). \end{aligned} \tag{98}$$

In the previous equation, firstly, let us look at the term $\frac{1}{n^2}\sum_{g=1}^{n}\sum_{j=1}^{n}4cov(Y_{g1},Y_{j1})$

$$\frac{1}{n^2}\sum_{g=1}^{n}\sum_{j=1}^{n}4cov(Y_{g1},Y_{j1}) \leq \frac{1}{n^2}\sum_{g=1}^{n}\sum_{j=1}^{n}4Sup|cov(Y_{g1},Y_{j1})| \leq \frac{4}{n^2}\sum_{g=1}^{n}\sum_{j=1}^{n}\alpha(d_{gj}) \tag{99}$$

by assumption (vii). Therefore, we need to prove that

$$\frac{4}{n^2}\sum_{g=1}^{n}\sum_{j=1}^{n}\alpha(d_{gj}) = o(1) \quad \text{as} \quad n \to \infty. \tag{100}$$

Following Pinkse and Slade (1998), we also use the Bernstein's (1927) blocking method to prove this as follows. We split the region in which observations are located up to an a_n area of size $c_1\sqrt{b_n} \times c_2\sqrt{b_n}$. We also know that a_n increases faster than $\sqrt{n}$ and b_n slower, where a_n and b_n are integers such that $a_n b_n = n$. Without loss of generality, we assume $c_1 = c_2 = 1$, and let a_n and b_n be constructed such that $\alpha(\sqrt{b_n})a_n \to 0$. Let $n^{\tau-\frac{1}{2}} \times b_n < 1$, uniformly in n, for some fixed $0 < \tau < \frac{1}{2}$. By construction of b_n, $O_p(n^{-\frac{1}{2}}b_n) = o_p(1)$. Then we are able to apply the same idea to our case. In our case, the groups g and j take the role of a_n and b_n,where one grows faster and the other grows slower than $\sqrt{n}$. We also know the d_{gj} is the distance between $|g-j|$. So we can find an upper bound for $|g-j|$ as the maximum between group g and j. Let us suppose that j is the one that grows faster than $\sqrt{n}$ and g is the one that grows slower than $\sqrt{n}$. Then we can cancel one of the summations corresponding to g with n^{-1}. Moreover, since j grows faster than $\sqrt{n}$ but slower than n^{-1}, one way is to define $\sum_{j=1}^{\sqrt{n}} j\alpha(j)$ as the one that grows faster than $\sqrt{n}$ but slower than n in such a way that

$$\sum_{g=1}^{n}\sum_{j=1}^{n}\alpha(d_{gj}) = O(\frac{1}{n}\sum_{j=1}^{\sqrt{n}} j\alpha(j)). \tag{101}$$

Finally, $\sum_{j=1}^{\sqrt{n}} j\alpha(j)$ grows slower than n and therefore, $O(\frac{1}{n}\sum_{j=1}^{\sqrt{n}} j\alpha(j)) = o(1)$. So, we can get

$$\frac{1}{n^2}\sum_{g=1}^{n}\sum_{j=1}^{n}4cov(Y_{g1},Y_{j1}) \leq \frac{4}{n^2}\sum_{g=1}^{n}\sum_{j=1}^{n}\alpha(d_{gj}) = o(1). \tag{102}$$

We can apply the same logic to $\frac{1}{n^2}\sum_{g=1}^{n}\sum_{j=1}^{n}8cov(Y_{g1},Y_{j2})$ and $\frac{1}{n^2}\sum_{g=1}^{n}\sum_{j=1}^{n}4cov(Y_{g2},Y_{j2})$. Let us consider $\frac{1}{n^2}\sum_{g=1}^{n}\sum_{j=1}^{n}4cov(Y_{g1}Y_{g2},Y_{j1}Y_{j2})$. If we define $Y_g = Y_{g1}Y_{g2}$ and $Y_j = Y_{j1}Y_{j2}$, we can apply the same logic to prove that $\frac{1}{n^2}\sum_{g=1}^{n}\sum_{j=1}^{n}4cov(Y_g,Y_j) \leq \frac{4}{n^2}\sum_{g=1}^{n}\sum_{j=1}^{n}\alpha(d_{gj}) = o(1)$. Therefore, we are able to show that

$$E\left\|Q_n(\theta) - E[Q_n(\theta)]\right\|^2 \leq Sup_{ngj}|Var[Q_n(\theta)]| \leq \frac{36}{n^2}\sum_{g=1}^{n}\sum_{j=1}^{n}\alpha(d_{gj}) = o(1). \tag{103}$$

Hence, $Q(\theta) - E[Q_n(\theta)] = o(1) \Longrightarrow Q_n(\theta) - Q(\theta) = o_p(1)$ at all $\theta \in \Theta$.*Q.E.D.* ■

LEMMA 2 *Under the assumptions in Theorem 1, $Q_n(\theta) - Q(\theta)$ is stochastically equicontinuous.*

Proof: The proof requires only to show that $Q_n(\theta)$ is stochastically equicontinuous because $Q(\theta)$ is continuous by assumption (iii). We have that

$$\begin{aligned} Q_n(\theta) - Q_n\left(\widetilde{\theta}\right) &= \frac{1}{n}\sum_{g=1}^{n}\{Y_{g1}Y_{g2}[P_g(1,1,\theta) - P_g(1,1,\widetilde{\theta})] \\ &+Y_{g1}(1-Y_{g2})[P_g(1,0,\theta) - P_g(1,0,\widetilde{\theta})] \\ &+(1-Y_{g1})Y_{g2}[P_g(0,1,\theta) - P_{g01}(0,1,\widetilde{\theta})] \\ &+(1-Y_{g1})(1-Y_{g2})[P_g(0,0,\theta) - P_g(0,0,\widetilde{\theta})]\}. \end{aligned} \tag{104}$$

By the mean value theorem

$$\begin{aligned} Q_n(\theta) - Q_n\left(\widetilde{\theta}\right) = \frac{1}{n}\sum_{g=1}^{n}\{Y_{g1}Y_{g2}[\frac{\partial P_g(1,1)}{\partial\theta^T}(\theta^*)(\theta-\widetilde{\theta})] + Y_{g1}(1-Y_{g2})[\frac{\partial P_g(1,0)}{\partial\theta^T}(\theta^*)(\theta-\widetilde{\theta})] \\ +(1-Y_{g1})Y_{g2}[\frac{\partial P_g(0,1)}{\partial\theta^T}(\theta^*)(\theta-\widetilde{\theta})] + (1-Y_{g1})(1-Y_{g2})[\frac{\partial P_g(0,0)}{\partial\theta^T}(\theta^*)(\theta-\widetilde{\theta})]\} \end{aligned} \tag{105}$$

where θ^* lies between θ and $\widetilde{\theta}$. In order to prove $Q_n(\theta)$ is stochastically equicontinuous, it is sufficient to show that

$$\sup_{\theta\in\Theta}|\frac{1}{n}Y_{g1}Y_{g2}\sum_{g=1}^{n}\frac{\partial P_g(1,1)}{\partial\theta^T}(\theta)| = O_p(1), \tag{106}$$

and the same requirement applies to other terms. For simplicity issues we just prove one of them and the rest follow the same argument. Recall that

$$P_g(1,1) \equiv \log P_g(Y_{g1}=1, Y_{g2}=1|X_g), \tag{107}$$

and note that $P_g(Y_{g1}=1, Y_{g2}=1|X_g) = \Phi_2(\frac{X_{g1}\beta}{\sqrt{\Omega_{g11}}}, \frac{X_{g2}\beta}{\sqrt{\Omega_{g22}}}, \rho_g|X_g)$, where Φ_2 is the bivariate normal distribution function. Also

$$\frac{\partial P_g(1,1)}{\partial\theta^T} = \frac{\partial[\log\Phi_2(\frac{X_{g1}\beta}{\sqrt{\Omega_{g11}}}, \frac{X_{g2}\beta}{\sqrt{\Omega_{g22}}}, \rho)]}{\partial\theta^T}. \tag{108}$$

and since $\theta \equiv (\beta, \lambda)$

$$\frac{\partial P_g(1,1)}{\partial\theta^T}(\theta) = \left\{ \begin{array}{c} \frac{\partial P_g(1,1)}{\partial\beta^T}(\beta) \\ \frac{\partial P_g(1,1)}{\partial\lambda}(\lambda) \end{array} \right\}. \tag{109}$$

We focus first on $\frac{\partial P_g(1,1)}{\partial\beta^T}(\beta)$, where

$$\frac{\partial P_g(1,1)}{\partial\beta^T} = \frac{\partial[\log\Phi_2(\frac{X_{g1}\beta}{\sqrt{\Omega_{g11}}}, \frac{X_{g2}\beta}{\sqrt{\Omega_{g22}}}, \rho)]}{\partial\beta^T} = \frac{\frac{s_{g1}X_{g1}}{\sqrt{\Omega_{g11}}} + \frac{s_{g2}X_{g2}}{\sqrt{\Omega_{g22}}}}{\Phi_2(\frac{X_{g1}\beta}{\sqrt{\Omega_{g11}}}, \frac{X_{g2}\beta}{\sqrt{\Omega_{g22}}}, \rho_g)}, \tag{110}$$

with

$$s_{g1} = \phi(\frac{X_{g1}\beta}{\sqrt{\Omega_{g11}}})\Phi(\frac{(\frac{X_{g2}\beta}{\sqrt{\Omega_{g22}}} - \rho\frac{X_{g1}\beta}{\sqrt{\Omega_{g11}}})}{\sqrt{1-\rho_g^2}}), \tag{111}$$

$$s_{g2} = \phi(\frac{X_{g2}\beta}{\sqrt{\Omega_{g22}}})\Phi(\frac{(\frac{X_{g1}\beta}{\sqrt{\Omega_{g11}}} - \rho\frac{X_{g2}\beta}{\sqrt{\Omega_{g22}}})}{\sqrt{1-\rho_g^2}}). \tag{112}$$

By assumption (v)

$$\sup_g \left\| \frac{1}{\Phi_2(\frac{X_{g1}\beta}{\sqrt{\Omega_{g11}}}, \frac{X_{g2}\beta}{\sqrt{\Omega_{g22}}}, \rho_g)} \right\| = \sup_g \left\| \frac{1}{\Pr(Y_{g1}=1, Y_{g2}=1|X_g)} \right\| < \infty. \tag{113}$$

and it is easy to see that $\left\| \frac{s_{g1}X_{g1}}{\sqrt{\Omega_{g11}}} + \frac{s_{g2}X_{g2}}{\sqrt{\Omega_{g22}}} \right\| < \infty$ provided that $\sup_g(\|X_g\|) < \infty$. Therefore,

$$\sup_g \left\| \frac{\partial P_g(1,1)}{\partial \beta^T}(\beta) \right\| < \infty. \tag{114}$$

We now discuss the second term $\frac{\partial P_g(Y_{g1}=1,Y_{g2}=1|X_g)}{\partial \lambda}(\lambda)$, where

$$\frac{\partial P_g(1,1)}{\partial \lambda} = \frac{\partial[\log \Phi_2(\frac{X_{g1}\beta}{\sqrt{\Omega_{g11}}}, \frac{X_{g2}\beta}{\sqrt{\Omega_{g22}}}, \rho)]}{\partial \lambda} \tag{115}$$

$$= \frac{\phi_2(\frac{\varepsilon_{g1}}{\sqrt{\Omega_{g11}}}, \frac{\varepsilon_{g2}}{\sqrt{\Omega_{g22}}}, \rho_g)}{\Phi_2(\frac{X_{g1}\beta}{\sqrt{\Omega_{g11}}}, \frac{X_{g2}\beta}{\sqrt{\Omega_{g22}}}, \rho_g)} \times \frac{\partial \phi_2(\frac{\varepsilon_{g1}}{\sqrt{\Omega_{g11}}}, \frac{\varepsilon_{g2}}{\sqrt{\Omega_{g22}}}, \rho_g)}{\partial \lambda} \tag{116}$$

and after some algebra, we can prove that $\sup_g \left\| \frac{\partial \phi_2(\frac{\varepsilon_{g1}}{\sqrt{\Omega_{g11}}}, \frac{\varepsilon_{g2}}{\sqrt{\Omega_{g22}}}, \rho_g)}{\partial \lambda} \right\| < \infty$ provided that $\sup_g \|W_g\| < \infty$.

Therefore, it easy to see when $\sup_g \left\| \frac{\partial P_g(1,1)}{\partial \lambda} \right\| < \infty$ and $\sup_g \left\| \frac{\partial P_g(1,1)}{\partial \beta^T} \right\| < \infty$, we can get

$$\sup_g \left\| \frac{\partial P_g(1,1)}{\partial \theta^T} \right\| < \infty. \tag{117}$$

We apply the same logic to the other terms, and we can prove that $\sup_g \left\| \frac{\partial P_g(1,0)}{\partial \theta^T}(\theta) \right\|$, $\sup_g \left\| \frac{\partial P_g(0,1)}{\partial \theta^T}(\theta) \right\|$ and $\sup_g \left\| \frac{\partial P_g(0,0)}{\partial \theta^T}(\theta) \right\|$ are also bounded.

Therefore, finally $\sup_{\theta\in\Theta} |\frac{1}{n} Y_{g1}Y_{g2} \sum_{g=1}^n \frac{\partial P_g(1,1)}{\partial \theta^T}(\theta)| = O_p(1)$ given $\sup_g(\|Y_g\|) = O(1)$, and hence we can prove that $Q_n(\theta) - Q(\theta)$ is stochastically equicontinuous. *Q.E.D.* ■

LEMMA 3: *Under the assumptions in Theorem 2,* $\sup_g \left\| \frac{\partial^2 \Pr(Y_{g1}=1,Y_{g2}=1|X_g)}{\partial \theta \partial \theta^T} \right\| < \infty$.

Proof: From Lemma 2, we know that

$$\begin{aligned}
&\frac{\partial \Pr(Y_{g1}=1,Y_{g2}=1|X_g)}{\partial \beta^T} \\
= &\frac{\phi(\frac{X_{g1}\beta}{\sqrt{\Omega_{g11}}})\Phi(\frac{(\frac{X_{g2}\beta}{\sqrt{\Omega_{g22}}}-\rho\frac{X_{g1}\beta}{\sqrt{\Omega_{g11}}})}{\sqrt{1-\rho_g^2}})X_{g1}}{\sqrt{\Omega_{g11}}}+\frac{\phi(\frac{X_{g2}\beta}{\sqrt{\Omega_{g22}}})\Phi(\frac{(\frac{X_{g1}\beta}{\sqrt{\Omega_{g11}}}-\rho\frac{X_{g2}\beta}{\sqrt{\Omega_{g22}}})}{\sqrt{1-\rho_g^2}})X_{g2}}{\sqrt{\Omega_{g22}}} \qquad (118)\\
\Longrightarrow &\frac{\partial^2 \Pr(Y_{g1}=1,Y_{g2}=1|X_g)}{\partial \beta \partial \beta^T} \\
= &\frac{X_{g1}\phi(\frac{X_{g1}\beta}{\sqrt{\Omega_{g11}}})\{X_{g1}\frac{X_{g1}\beta}{\sqrt{\Omega_{g11}}}\Phi[\frac{(\frac{X_{g2}\beta}{\sqrt{\Omega_{g22}}}-\rho\frac{X_{g1}\beta}{\sqrt{\Omega_{g11}}})}{\sqrt{1-\rho_g^2}}]+\phi[\frac{(\frac{X_{g2}\beta}{\sqrt{\Omega_{g22}}}-\rho\frac{X_{g1}\beta}{\sqrt{\Omega_{g11}}})}{\sqrt{1-\rho_g^2}}]\frac{(\frac{X_{g2}}{\sqrt{\Omega_{g22}}}-\rho\frac{X_{g1}}{\sqrt{\Omega_{g11}}})}{\sqrt{1-\rho_g^2}}\}}{\sqrt{\Omega_{g11}}} \\
&+\frac{X_{g2}\phi(\frac{X_{g2}\beta}{\sqrt{\Omega_{g22}}})\{X_{g2}\frac{X_{g2}\beta}{\sqrt{\Omega_{g22}}}\Phi[\frac{(\frac{X_{g1}\beta}{\sqrt{\Omega_{g11}}}-\rho\frac{X_{g2}\beta}{\sqrt{\Omega_{g22}}})}{\sqrt{1-\rho_g^2}}]+\phi[\frac{(\frac{X_{g1}\beta}{\sqrt{\Omega_{g11}}}-\rho\frac{X_{g2}\beta}{\sqrt{\Omega_{g22}}})}{\sqrt{1-\rho_g^2}}]\frac{(\frac{X_{g1}}{\sqrt{\Omega_{g11}}}-\rho\frac{X_{g2}}{\sqrt{\Omega_{g22}}})}{\sqrt{1-\rho_g^2}}\}}{\sqrt{\Omega_{g22}}}, \qquad (119)
\end{aligned}$$

and even though the above expression is complicated, it is easy to see that all the terms are bounded provided the assumptions in Theorem 2 hold. This is equivalent to

$$\sup_g \left\| \frac{\partial^2 \Pr(Y_{g1}=1,Y_{g2}=1|X_g)}{\partial \beta \partial \beta^T} \right\| < \infty, \qquad (120)$$

$$\begin{aligned}
&\frac{\partial \Pr(Y_{g1}=1,Y_{g2}=1|X_g)}{\partial \lambda} \\
= &\frac{\phi_2(\frac{\varepsilon_{g1}}{\sqrt{\Omega_{g11}}},\frac{\varepsilon_{g2}}{\sqrt{\Omega_{g22}}},\rho_g)}{\Phi_2(\frac{X_{g1}\beta}{\sqrt{\Omega_{g11}}},\frac{X_{g2}\beta}{\sqrt{\Omega_{g22}}},\rho_g)}\times\frac{\partial\phi_2(\frac{\varepsilon_{g1}}{\sqrt{\Omega_{g11}}},\frac{\varepsilon_{g2}}{\sqrt{\Omega_{g22}}},\rho_g)}{\partial\lambda}, \qquad (121)\\
\Longrightarrow &\frac{\partial^2 \Pr(Y_{g1}=1,Y_{g2}=1|X_g)}{\partial \lambda^2} \\
= &\frac{(\frac{\partial\phi_2}{\partial\lambda})^2(\Phi_2-\phi_2\phi_2)}{(\Phi_2)^2}+\frac{\phi_2}{\Phi_2}\frac{\partial^2\phi_2}{\partial\lambda^2}. \qquad (122)
\end{aligned}$$

It is easy to see that the first term of the above equation is bounded from previous results (i.e. $\sup_g \left\|\frac{\partial\phi_2}{\partial\lambda}\right\| < \infty$) and the second term can be also proved bounded since $\frac{\partial^2\phi_2}{\partial\lambda^2}$ can be proved to be bounded given that $\sup_g \|W_g\| < \infty$ after some algebra. Hence $\sup_g \left\|\frac{\partial^2 \Pr(Y_{g1}=1,Y_{g2}=1|X_g)}{\partial\theta\partial\theta^T}\right\| < \infty$. *Q.E.D.* ■

LEMMA 4. *Under the assumptions in Theorem 2,* $ET_{na_n}-1=o(1)$, where $T_{na_n}\equiv\Pi_{j=1}^{a_n}(1+i\gamma D_{n,j})$.

Proof: By definition, $T_{na_n}=\Pi_{j=1}^{a_n}(1+i\gamma D_{n,j})=T_{n,a_n-1}+i\gamma T_{n,a_n-1}D_{nn}$. By repeatedly multiplying out, we finally get $T_{na_n}=1+i\gamma\sum_{j=1}^{a_n}T_{n,j-1}D_{nj}$. Hence,

$$ET_{na_n}-1=E(i\gamma\sum_{j=1}^{a_n}T_{n,j-1}D_{nj}). \qquad (123)$$

In order to prove $ET_{na_n} - 1 = o(1)$, we just need to show that: $E(i\gamma \sum_{j=1}^{a_n} T_{n,j-1}D_{nj}) = o(1)$. This is equivalent to prove that $E(T_{n,j-1}D_{nj}) = o(a_n^{-1})$. We can rewrite $T_{n,j-1}$ as $T_{n,j-1} = \Pi_{k=1}^{j-1}(1 + i\gamma D_{n,k})$.We know there are $j-1$ groups of $D_{n,k}$ in $T_{n,j-1}$. We split these $j-1$ groups into two parts: groups adjacent to group j, and groups that are not adjacent to group j. We then define the area Ξ_{nj-1} as the area which is adjacent to group j. Therefore, $T_{n,j-1} = \Pi_{k\in\Xi_{nj-1}}(1+i\gamma D_{n,k})\Pi_{k\notin\Xi_{nj-1}}(1 + i\gamma D_{n,k}) = \Pi_{k\in\Xi_{nj-1}}(1 + i\gamma D_{n,k})T_{Rnj}$,where $T_{Rnj} \equiv \Pi_{k\notin\Xi_{nj-1}}(1 + i\gamma D_{n,k})$, which includes the groups which are not adjacent to group j.

Since $T_{n,j-1} = \Pi_{k\in\Xi_{nj-1}}(1 + i\gamma D_{n,k})T_{Rnj}$,we just need to prove

$$E[D_{nj}(\Pi_{k\in\Xi_{nj-1}}(1 + i\gamma D_{n,k})T_{Rnj})] = E[D_{nj}T_{Rnj}(\Pi_{k\in\Xi_{nj-1}}(1 + i\gamma D_{n,k}))] = o(a_n^{-1}). \tag{124}$$

We know that

$$\begin{aligned} E[D_{nj}T_{Rnj}(\Pi_{k\in\Xi_{nj-1}}(1 + i\gamma D_{n,k}))] &= E[D_{nj}T_{Rnj}(1 + i\gamma \sum_{k\in\Xi_{nj-1}} T_{n,k-1}D_{nk})] &(125)\\ &= E[D_{nj}T_{Rnj}] + E[D_{nj}T_{Rnj}(i\gamma \sum_{k\in\Xi_{nj-1}} T_{n,k-1}D_{nk})]. &(126)\end{aligned}$$

First, we look at the term $E[D_{nj}T_{Rnj}]$. Since $T_{Rnj} \equiv \Pi_{k\notin\Xi_{nj-1}}(1 + i\gamma D_{n,k})$, that means the group is not adjacent to group j. By Bernstein's method, we split the region in such a way that the distance between group j and non-adjacent group is at least $b_n^{\frac{1}{2}}$. Hence, $Max|E[D_{nj}T_{Rnj}]| = Max|cov(D_{nj}, T_{Rnj}) = \alpha(\sqrt{b_n})$ provided $E(D_{nj}) = 0$ and by assumption (vi) in Theorem 1. By construction of a_n and b_n, $\alpha(\sqrt{b_n})a_n = o(1)$, and hence we obtain $Max|E[D_{nj}T_{Rnj}]| = o(a_n^{-1})$.

Second, we look at the term $E[D_{nj}T_{Rnj}(i\gamma \sum_{k\in\Xi_{nj-1}} T_{n,k-1}D_{nk})]$. We have that

$$E[D_{nj}T_{Rnj}(i\gamma \sum_{k\in\Xi_{nj-1}} T_{n,k-1}D_{nk})] = i\gamma \sum_{k\in\Xi_{nj-1}} E[D_{nj}T_{Rnj}\Pi_{k\in\Xi_{nj-1}}D_{nk})]. \tag{127}$$

Consider $E[D_{nj}T_{Rnj}D_{nk})]$ first. We know that $E[D_{nj}T_{Rnj}D_{nk})] = cov(D_{nj}, T_{Rnj}D_{nk})$ provided $E(D_{nj}) = 0$. Since $cov(D_{nj}, T_{Rnj}D_{nk}) \to cov(D_{nj}, T_{Rnj})$ as $n \to \infty$, because T_{Rnj} gets more and more terms (all groups not adjacent to group j), while D_{nk} keeps the same amount. In the first step, we have proved that $cov(D_{nj}, T_{Rnj}) = o(a_n^{-1})$, and by the same argument $cov(D_{nj}, T_{Rnj}D_{nk}) = o(a_n^{-1})$.

Therefore, we can prove that $E(T_{n,j-1}D_{nj}) = o(a_n^{-1}) \Rightarrow ET_{na_n} - 1 = o(1)$. *Q.E.D.* ■

LEMMA 5. *Under the assumptions in Theorem 2,* $\sum_{i\neq j}^{a_n} E(D_{ni}D_{nj}) = o(1)$.

Proof: We know that $\sum_{i\neq j}^{a_n} E(D_{ni}D_{nj}) = \sum_{i=1}^{a_n}\sum_{j=1}^{a_n} E(D_{ni}D_{nj}) - \sum_{i=j}^{a_n} E(D_{ni}D_{nj}) = o(1)$ if we can show that $Max \sum_{i=1}^{a_n} |E(D_{ni}D_{nj})| = o(a_n^{-1})$. This is equivalent to prove $\sum_{i\neq j}^{a_n} E(D_{ni}D_{nj}) = o(1)$ because the summation over j contains $a_n - 1$ terms.

Define Ξ_{nil} as the set of indices corresponding to blocks that have l blocks removed from every direction from block l. In other words, we assume there are no more than $8l$ blocks within distance

l. Hence,

$$Max\sum_{i=1}^{a_n}|E(D_{ni}D_{nj})| \leq Max\sum_{l=1}^{\sqrt{a_n}}\sum_{j\in\Xi_{nil}}|E(D_{ni}D_{nj})| \quad (128)$$

$$\leq Max\sum_{j\in\Xi_{nil}}|E(D_{ni}D_{nj})| + Max\sum_{l=2}^{\sqrt{a_n}}\sum_{j\in\Xi_{nil}}|E(D_{ni}D_{nj})|. \quad (129)$$

The first term is proved to be $o(n^{-1}b_n) = o(a_n^{-1})$ in Lemma 6. The second term can be also proved to be $o(a_n^{-1})$ in Lemma 7. *Q.E.D.* ■

LEMMA 6: *Under the assumptions in Theorem 2,* $Max\sum_{i\neq j}|E(D_{ni}D_{nj})| = o(n^{-1}b_n) = o(a_n^{-1})$.

Proof: Since $D_{n,j} = n^{-\frac{1}{2}}\sum_{t\in\Lambda_{nj}}A_{nt}$ by definition

$$Max\sum_{i\neq j}|E(D_{ni}D_{nj})| = Max_{i\notin j}|n^{-1}\sum_{s\in\Lambda ni,t\in\Lambda nj}E(A_{ns}A_{nt})| \quad (130)$$

$$\leq Max_{i\notin j}C_1n^{-1}\sum_{s\in\Lambda ni,t\in\Lambda nj}\alpha(d_{ts}) \quad (131)$$

because $E(A_{ns}A_{nt}) = Cov(A_{ns}, A_{nt}) = C_1\alpha(d_{ts})$, where $C_1 > 0$.

To compute the upper bound of the correlation between i and j, we just need to consider the strongest case, e.g. the i and j are adjacent each other. By Bernsteins' blocking method, the number of (t, s) combinations that are within distance d is bounded by $C_2\sqrt{b_n}d^2$, where $C_2 > 0$. Hence we can get

$$Max_{i\notin j}C_1n^{-1}\sum_{s\in\Lambda ni,t\in\Lambda nj}\alpha(d_{ts}) \leq C_3Max_{i\notin j}n^{-1}\sqrt{b_n}\sum_{d=0}^{C_4\sqrt{b_n}}d^2\alpha(d), \quad (132)$$

where $C_3 = C_1C_2, C_4 > 0$.

By assumption (ii) in Theorem 2, $d^2\alpha(d) \to 0$, as $d \to \infty$. Therefore,

$$C_3Max_{i\notin j}n^{-1}\sqrt{b_n}\sum_{d=0}^{C_4\sqrt{b_n}}d^2\alpha(d) = o(n^{-1}b_n). \quad (133)$$

Since $a_nb_n = n$ by construction, $o(n^{-1}b_n) = o(a_n^{-1})$. *Q.E.D.* ■

LEMMA 7: *Under the assumptions in Theorem 2,* $Max\sum_{l=2}^{\sqrt{a_n}}\sum_{j\in\Xi_{nil}}|E(D_{ni}D_{nj})| = o(a_n^{-1})$.

Proof: Because $Max_{j\in\Xi_{nil}} \times Max_{s\in\Lambda ni} \times Max_{t\in\Lambda nj}|E(A_{ns}A_{nt}) = O(\alpha\sqrt{b_n}(l-1))$, we have that

$$Max\sum_{l=2}^{\sqrt{a_n}}\sum_{j\in\Xi_{nil}}|E(D_{ni}D_{nj})| \leq C_5Max\sum_{l=2}^{\sqrt{a_n}}\#\Xi_{nil}n^{-1} \times \#\Lambda_{ni} \times \#\Lambda_{nj}\alpha(\sqrt{b_n}(l-1)) \quad (134)$$

$$\leq C_6n^{-1}b_n^2l\sum_{l=1}^{\sqrt{a_n}}\alpha(\sqrt{b_n}l) = o(n^{-1}b_nl\sum_{l=1}^{\sqrt{a_n}}\alpha(l) = o(n^{-1}b_n) \quad (135)$$

$$= o(a_n^{-1}). \quad (136)$$

where # denotes the number of objects, and $o(n^{-1}b_n l\sum_{l=1}^{\sqrt{a_n}}\alpha(l)=o(n^{-1}b_n)$ follows from assumption (i): as $d\to\infty$, $\frac{d^2\alpha(dd^*)}{\alpha(d^*)}=o(1)$. *Q.E.D.* ■

LEMMA 8: *Under the assumptions in Theorem 2,* $\sum_{j=1}^{a_n}D_{n,j}^2=\sum_{j=1}^{a_n}E(D_{n,j}^2)+o_p(1)$.

Proof: In order to prove $\sum_{j=1}^{a_n}D_{n,j}^2=\sum_{j=1}^{a_n}E(D_{n,j}^2)+o_p(1)$, it suffices to show that

$$\sum_{i=1}^{a_n}\sum_{j=1}^{a_n}Cov(D_{n,i}^2,D_{n,j}^2)=o(1). \tag{137}$$

We have that

$$\begin{aligned}\sum_{i=1}^{a_n}\sum_{j=1}^{a_n}Cov(D_{n,i}^2,D_{n,j}^2) &= \sum_{i=1}^{a_n}\sum_{j=1}^{a_n}\{[D_{n,i}^2-E(D_{n,i}^2)][D_{n,j}^2-E(D_{n,j}^2)]\} &(138)\\ &\le C_7\sum_{l=0}^{C_8\sqrt{a_n}}(l+1)\alpha(\sqrt{b_n}l)MaxE(D_{ni}^4), &(139)\end{aligned}$$

where C_7, $C_8>0$ are large enough. Also

$$\begin{aligned}MaxE(D_{ni}^4) &\le n^{-2}Max\sum_{t1,t2,t3,t4\in\Lambda_{nj}}|E[A_{nt1},A_{nt2},A_{nt3},A_{nt4}] &(140)\\ &\le C_9n^{-2}Max_j\sum_{t1,t2,t3,t4\in\Lambda_{nj}}\{\alpha(d_{t1,t2})+...+\alpha(d_{t3,t4})\} &(141)\\ &\le C_{10}n^{-2}Max_j\sum_{t1,t2\in\Lambda_{nj}}\{\alpha(d_{t1,t2})\} &(142)\\ &\le C_{11}n^{-2}b_n^2Max_j\sum_{t1\in\Lambda_{nj}}\sum_{l=0}^{c_{12}\sqrt{b_n}}l\alpha(l)=O(n^{-2}b_n^3), &(143)\end{aligned}$$

where $C_9,C_{10},C_{11},C_{12}>0$, $Sup|\sum_{l=0}^{\infty}l\alpha(l)|<\infty$. Therefore finally

$$C_7\sum_{l=0}^{C_8\sqrt{a_n}}(l+1)\alpha(\sqrt{b_n}l)MaxE(D_{ni}^4)=O(n^{-2}b_n^3a_n)=o(1), \tag{144}$$

because $a_nb_n=n$ and $n^{-1}b_n^2\to0$ as $n\to\infty$. *Q.E.D.* ■

Finally, the following Lemma 9 generalizes Pinkse and Slade (1998) results as a way to obtain consistent estimates of the variance covariance matrix.

LEMMA 9: *If assumptions in Theorem 2 hold, and* $\sup_g\left\|\frac{\partial\Phi_4}{\partial\theta}+\frac{\partial\Phi_3}{\partial\theta}\right\|<\infty$, *then* $A_n(\widehat{\theta})-A(\theta_0)=o_p(1)$ *and* $B_n(\widehat{\theta})-B(\theta_0)=o_p(1)$; *where* $B_n(\theta)\equiv nE[S_n(\theta)S_n^T(\theta)]$ *and* $A_n(\theta)\equiv-E[H(\theta)]$.

Proof: First, we prove that $A_n(\widehat{\theta})-A(\theta_0)=o_p(1)$. We know that $A_n(\widehat{\theta})=-\frac{1}{n}\sum_{g=1}^{n}H_g(\widehat{\theta})$, and by definition, $\lim_{n\to\infty}A_n(\theta_0)=A(\theta_0)$. So we just need prove that $\varpi^T\{A_n(\widehat{\theta})-\lim_{n\to\infty}A_n(\theta_0)\}=o_p(1)$ for all $\|\varpi\|=1$. From the proof of Theorem 2, we have already proved that

$$\frac{1}{n}\sum_{g=1}^{n}H_g(\widehat{\theta})\to\frac{1}{n}\sum_{g=1}^{n}H_g(\theta_0) \tag{145}$$

as $n \to \infty$, provided that $\widehat{\theta} - \theta_0 = o_p(1)$ which is proved in Theorem 1. Therefore, we can get $A_n(\widehat{\theta}) - A(\theta_0) = o_p(1)$.

Second, we consider how to show $B_n(\widehat{\theta}) - B(\theta_0) = o_p(1)$. As before, it is sufficient to show that $B_n(\widehat{\theta}) - B_n(\theta_0) = o_p(1)$ as $n \to \infty$. We know that $B_n(\theta_0) = nE[S_n(\theta_0)S_n^T(\theta_0)] = nVar(S_n(\theta_0))$ given $S_n(\theta_0) = 0$. Recall from the proof of Theorem 2 that

$$\begin{aligned} S_n(\theta_0) &= \frac{1}{n}\sum_{g=1}^{n}\{Y_{g1}Y_{g2}\frac{\partial P_g(1,1)}{\partial\theta}(\theta_0) + Y_{g1}(1-Y_{g2})\frac{\partial P_g(1,0)}{\partial\theta}(\theta_0) \\ &\quad +(1-Y_{g1})Y_{g2}\frac{\partial P_g(0,1)}{\partial\theta}(\theta_0) + (1-Y_{g1})(1-Y_{g2})\frac{\partial P_g(0,0)}{\partial\theta}(\theta_0)\}, \end{aligned} \tag{146}$$

and we can rewrite it as

$$\begin{aligned} S_n(\theta_0) &= \frac{1}{n}\sum_{g=1}^{n}\{Y_{g1}Y_{g2}[\frac{\partial P_g(1,1)}{\partial\theta}(\theta_0) - \frac{\partial P_g(1,0)}{\partial\theta}(\theta_0) - \frac{\partial P_g(0,1)}{\partial\theta}(\theta_0) + \frac{\partial P_g(0,0)}{\partial\theta}(\theta_0)] \\ &\quad +Y_{g1}[\frac{\partial P_g(1,0)}{\partial\theta}(\theta_0) - \frac{\partial P_g(0,0)}{\partial\theta}(\theta_0] + Y_{g2}[\frac{\partial P_g(0,1)}{\partial\theta}(\theta_0) - \frac{\partial P_g(0,0)}{\partial\theta}(\theta_0)] \\ &\quad +\frac{\partial P_g(0,0)}{\partial\theta}(\theta_0)\}. \end{aligned} \tag{147}$$

For the sake of brevity, we redefine

$$\psi_{ng1} \equiv [\frac{\partial P_g(1,1)}{\partial\theta}(\theta_0) - \frac{\partial P_g(1,0)}{\partial\theta}(\theta_0) - \frac{\partial P_g(0,1)}{\partial\theta}(\theta_0) + \frac{\partial P_g(0,0)}{\partial\theta}(\theta_0)], \tag{148}$$

$$\psi_{ng2} \equiv [\frac{\partial P_g(1,0)}{\partial\theta}(\theta_0) - \frac{\partial P_g(0,0)}{\partial\theta}(\theta_0)], \tag{149}$$

$$\psi_{ng3} \equiv [\frac{\partial P_g(0,1)}{\partial\theta}(\theta_0) - \frac{\partial P_g(0,0)}{\partial\theta}(\theta_0)], \tag{150}$$

$$\psi_{ng4} \equiv \frac{\partial P_g(0,0)}{\partial\theta}(\theta_0). \tag{151}$$

Therefore,

$$\begin{aligned} Var(S_n(\theta_0)) &= n^{-1}B_n(\theta_0) \\ &= n^{-2}\sum_{g=1}^{n}\sum_{j=1}^{n}\{\psi_{ng1}\psi_{nj1}Cov(Y_{g1}Y_{g2}, Y_{j1}Y_{j2}) + 2\psi_{ng1}\psi_{nj2}Cov(Y_{g1}Y_{g2}, Y_{j1}) \\ &\quad +2\psi_{ng1}\psi_{nj3}Cov(Y_{g1}Y_{g2}, Y_{j2}) + \psi_{ng2}\psi_{nj2}Cov(Y_{g1}, Y_{j1}) \\ &\quad +2\psi_{ng2}\psi_{nj3}Cov(Y_{g1}, Y_{j2}) + \psi_{ng3}\psi_{nj3}Cov(Y_{g2}, Y_{j2}), \end{aligned} \tag{152}$$

where $\psi_{nj1}, \psi_{nj2}, \psi_{nj3}$ are defined similarly as $\psi_{ng1}, \psi_{ng2}, \psi_{ng3}$.

As before, we just need to provide the proof for one of these terms, and the same logic applies to

other terms. We consider the most complicated term and the rest follow the same argument

$$
\begin{aligned}
& n^{-1}\sum_{g=1}^{n}\sum_{j=1}^{n}[\psi_{ng1}\psi_{nj1}Cov(Y_{g1}Y_{g2},Y_{j1}Y_{j2})] \\
= \; & n^{-1}\sum_{g=1}^{n}\sum_{j=1}^{n}\psi_{ng1}\psi_{nj1}[E(Y_{g1}Y_{g2}Y_{j1}Y_{j2})-E(Y_{g1}Y_{g2})E(Y_{j1}Y_{j2})]. \qquad (153)\\
E(Y_{g1}Y_{g2}Y_{j1}Y_{j2}) = \; & \Pr(Y_{g1}=1,Y_{g2}=1,Y_{j1}=1,Y_{j2}=1|X_g) \qquad (154)\\
= \; & \Phi_4(y_{g1},y_{g2},y_{j1},y_{j2},\rho_{12},\rho_{13},\rho_{14},\rho_{23},\rho_{24},\rho_{34}) \qquad (155)
\end{aligned}
$$

where Φ_4 is the cdf for the quadvariate standard normal distribution, $y_{g1}=\frac{Y_{g1}}{\sqrt{Var(Y_{g1})}}$etc. Similarly,

$$
\begin{aligned}
E(Y_{g1}Y_{g2}) &= \Pr(Y_{g1}=1,Y_{g2}=1|X_g)=\Phi_2(y_{g1},y_{g2},\rho_{12}), \qquad (156)\\
E(Y_{j1}Y_{j2}) &= \Pr(Y_{j1}=1,Y_{j2}=1|X_g)=\Phi_2(y_{j1},y_{j2},\rho_{34}), \qquad (157)
\end{aligned}
$$

and therefore,

$$
\begin{aligned}
E(Y_{g1}Y_{g2})E(Y_{j1}Y_{j2}) &= \Phi_2(y_{g1},y_{g2},\rho_{12})\times\Phi_2(y_{j1},y_{j2},\rho_{34}) \qquad (158)\\
&= \Phi_4(y_{g1},y_{g2},y_{j1},y_{j2},\rho_{12},0,0,0,0,\rho_{34}), \qquad (159)
\end{aligned}
$$

so we can write the first term as

$$
\begin{aligned}
B_n(\theta_0) &= n^{-1}\sum_{g=1}^{n}\sum_{j=1}^{n}\psi_{ng1}\psi_{nj1}[E(Y_{g1}Y_{g2}Y_{j1}Y_{j2})-E(Y_{g1}Y_{g2})E(Y_{j1}Y_{j2})] \qquad (160)\\
&= n^{-1}\sum_{g=1}^{n}\sum_{j=1}^{n}\psi_{ng1}\psi_{nj1}[\Phi_4(y_{g1},y_{g2},y_{j1},y_{j2},\rho_{12}(\theta_0),\rho_{13}(\theta_0),\rho_{14}(\theta_0),\rho_{23}(\theta_0),\rho_{24}(\theta_0),\rho_{34}(\theta_0))\\
&\quad -\Phi_4(y_{g1},y_{g2},y_{j1},y_{j2},\rho_{12}(\theta_0),0,0,0,0,\rho_{34}(\theta_0))]. \qquad (161)
\end{aligned}
$$

Similarly, we can write the first term of $B_n(\widehat{\theta})$ as

$$
\begin{aligned}
& n^{-1}\sum_{g=1}^{n}\sum_{j=1}^{n}\psi_{ng1}\psi_{nj1}[\Phi_4(y_{g1},y_{g2},y_{j1},y_{j2},\rho_{12}(\widehat{\theta}),\rho_{13}(\widehat{\theta}),\rho_{14}(\widehat{\theta}),\rho_{23}(\widehat{\theta}),\rho_{24}(\widehat{\theta}),\rho_{34}(\widehat{\theta}))\\
& -\Phi_4(y_{g1},y_{g2},y_{j1},y_{j2},\rho_{12}(\widehat{\theta}),0,0,0,0,\rho_{34}(\widehat{\theta}))]. \qquad (162)
\end{aligned}
$$

By the mean value theorem, the first term of $B_n(\widehat{\theta})-B_n(\theta_0)$ is given as

$$
\begin{aligned}
& n^{-1}\sum_{g=1}^{n}\sum_{j=1}^{n}\psi_{ng1}\psi_{nj1}\{[\Phi_4(y_{g1},y_{g2},y_{j1},y_{j2},\rho_{12}(\widehat{\theta}),\rho_{13}(\widehat{\theta}),\rho_{14}(\widehat{\theta}),\rho_{23}(\widehat{\theta}),\rho_{24}(\widehat{\theta}),\rho_{34}(\widehat{\theta}))\\
& -\Phi_4(y_{g1},y_{g2},y_{j1},y_{j2},\rho_{12}(\theta_0),\rho_{13}(\theta_0),\rho_{14}(\theta_0),\rho_{23}(\theta_0),\rho_{24}(\theta_0),\rho_{34}(\theta_0))]\\
& -[\Phi_4(y_{g1},y_{g2},y_{j1},y_{j2},\rho_{12}(\widehat{\theta}),0,0,0,0,\rho_{34}(\widehat{\theta}))\\
& -\Phi_4(y_{g1},y_{g2},y_{j1},y_{j2},\rho_{12}(\theta_0),0,0,0,0,\rho_{34}(\theta_0))]\} \qquad (163)\\
= \; & n^{-1}(\widehat{\theta}-\theta_0)\sum_{g=1}^{n}\sum_{j=1}^{n}\psi_{ng1}\psi_{nj1}\{\frac{\partial\Phi_4(y_{g1},y_{g2},y_{j1},y_{j2},\rho_{12}(\theta^*),\rho_{13}(\theta^*),\rho_{14}(\theta^*),\rho_{23}(\theta^*),\rho_{24}(\theta^*),\rho_{34}(\theta^*)}{\partial\theta}\\
& -\frac{\partial\Phi_4(y_{g1},y_{g2},y_{j1},y_{j2},\rho_{12}(\theta^*),0,0,0,0,\rho_{34}(\theta^*))}{\partial\theta}\}. \qquad (164)
\end{aligned}
$$

Since $\sup_g \left\|\psi_{ng1}\right\| < \infty$ by the proof in Theorem 2, we just need to assume

$$\sup_g \left\| \frac{\partial \Phi_4(y_{g1}, y_{g2}, y_{j1}, y_{j2}, \rho_{12}(\theta^*), \rho_{13}(\theta^*), \rho_{14}(\theta^*), \rho_{23}(\theta^*), \rho_{24}(\theta^*), \rho_{34}(\theta^*)}{\partial \theta} \right\| < \infty, \tag{165}$$

and the same argument applies to

$$\sup_g \left\| \frac{\partial \Phi_4(y_{g1}, y_{g2}, y_{j1}, y_{j2}, \rho_{12}(\theta^*), 0, 0, 0, 0, \rho_{34}(\theta^*))}{\partial \theta} \right\| < \infty \tag{166}$$

so that

$$\begin{aligned} n^{-1}(\widehat{\theta} - \theta_0) \sum_{g=1}^{n} \sum_{j=1}^{n} \psi_{ng1} \psi_{nj1} \{ & \frac{\partial \Phi_4(y_{g1}, y_{g2}, y_{j1}, y_{j2}, \rho_{12}(\theta^*), \rho_{13}(\theta^*), \rho_{14}(\theta^*), \rho_{23}(\theta^*), \rho_{24}(\theta^*), \rho_{34}(\theta^*)}{\partial \theta}) \\ & - \frac{\partial \Phi_4(y_{g1}, y_{g2}, y_{j1}, y_{j2}, \rho_{12}(\theta^*), 0, 0, 0, 0, \rho_{34}(\theta^*))}{\partial \theta} \} \to 0, \end{aligned} \tag{167}$$

because $(\widehat{\theta} - \theta_0) \to 0$ and the other terms are bounded.

Repeat the proofs to the other terms, plus the new assumption about $\sup_g \left\| \frac{\partial \Phi_3}{\partial \theta} \right\| < \infty$, and then we can prove $B_n(\widehat{\theta}) - B(\theta_0) = o_p(1)$. *Q.E.D.* ■

8 Appendix 2

TABLE 1*: SIMULATION RESULTS OF DIFFERENT ESTIMATORS OF λ IN THE CONTEXT OF THE BIVARIATE SPATIAL PROBIT MODEL.

		$\lambda = 0.2$		$\lambda = 0.4$		$\lambda = 0.6$		$\lambda = 0.8$	
		HPE	PMLE	HPE	PMLE	HPE	PMLE	HPE	PMLE
$N = 500$	mean	3.938	0.514	6.177	0.519	7.698	0.571	7.735	0.634
	bias	3.738	0.314	5.777	0.319	7.098	-0.029	6.935	-0.166
	(s.d.)	(12.158)	(0.120)	(15.776)	(0.205)	(16.929)	(0.151)	(16.202)	(0.289)
$N = 1000$	mean	3.174	0.512	4.668	0.518	5.456	0.581	5.914	0.672
	bias	2.974	0.312	4.268	0.118	4.856	-0.019	5.114	-0.128
	(s.d)	(8.844)	(0.107)	(9.100)	(0.133)	(9.631)	(0.149)	(10.173)	(0.276)
$N = 1500$	mean	2.746	0.511	4.050	0.507	4.872	0.609	5.426	0.708
	bias	2.546	0.311	3.650	0.107	4.272	0.009	4.626	-0.092
	(s.d.)	(6.423)	(0.099)	(7.414)	(0.124)	(8.598)	(0.149)	(8.514)	(0.253)

*Results are presented for our new Partial Maximum Likelihood Estimator (PMLE) and the Heteroskedastic Probit Estimator (HPE) of λ. Numbers in brackets show standard deviations (s.d.).

TABLE 2*: SIMULATION RESULTS OF DIFFERENT ESTIMATORS OF β_1, β_2 AND β_3 IN THE CONTEXT OF THE BIVARIATE SPATIAL PROBIT MODEL.

			β_1= 1		β_2= 1		β_3= 1	
			HPE	PMLE	HPE	PMLE	HPE	PMLE
$\lambda = 0.2$	$N = 500$	mean	5.322	2.618	5.333	2.619	5.329	2.623
		(s.d.)	(8.844)	(0.839)	(8.872)	(0.855)	(8.863)	(0.870)
	$N = 1000$	mean	5.308	2.616	5.296	2.616	5.289	2.618
		(s.d)	(7.612)	(0.560)	(7.570)	(0.560)	(7.568)	(0.564)
	$N = 1500$	mean	5.247	2.604	5.239	2.602	5.235	2.604
		(s.d.)	(6.624)	(0.540)	(6.606)	(0.536)	(6.613)	(0.543)
$\lambda = 0.4$	$N = 500$	mean	3.610	1.329	3.614	1.329	3.608	1.328
		(s.d.)	(5.305)	(0.362)	(5.311)	(0.365)	(5.290)	(0.366)
	$N = 1000$	mean	3.600	1.318	3.593	1.316	3.588	1.315
		(s.d.)	(4.192)	(0.355)	(4.177)	(0.355)	(4.178)	(0.353)
	$N = 1500$	mean	3.456	1.281	3.441	1.281	3.438	1.278
		(s.d.)	(3.818)	(0.342)	(3.793)	(0.343)	(3.798)	(0.339)
$\lambda = 0.6$	$N = 500$	mean	2.898	0.972	2.876	0.966	2.885	0.969
		(s.d.)	(3.761)	(0.271)	(3.723)	(0.268)	(3.735)	(0.271)
	$N = 1000$	mean	2.669	0.981	2.669	0.979	2.657	0.978
		(s.d.)	(2.951)	(0.261)	(2.953)	(0.261)	(2.916)	(0.259)
	$N = 1500$	mean	2.508	1.016	2.499	1.015	2.501	1.016
		(s.d.)	(2.726)	(0.250)	(2.706)	(0.250)	(2.708)	(0.253)
$\lambda = 0.8$	$N = 500$	mean	2.246	0.805	2.237	0.801	2.249	0.802
		(s.d.)	(2.810)	(0.373)	(2.803)	(0.373)	(2.841)	(0.392)
	$N = 1000$	mean	2.098	0.843	2.096	0.843	2.082	0.843
		(s.d.)	(2.281)	(0.349)	(2.279)	(0.349)	(2.246)	(0.340)
	$N = 1500$	mean	2.086	0.884	2.096	0.886	2.094	0.886
		(s.d.)	(2.059)	(0.316)	(2.071)	(0.314)	(2.073)	(0.318)

*Results are presented for our new Partial Maximum Likelihood Estimator (PMLE) and the Heteroskedastic Probit Estimator (HPE) of β_1, β_2 and β_3. Numbers in brackets show standard deviations (s.d.).

References

[1] Andrews, D. W. K. (1991): "Heteroskedasticity and Autocorrelation Consistent Covariance Matrix Estimation", *Econometrica* 59, 3, 817-858.

[2] Amemiya, T. (1973): "Regression analysis when the dependent variable is truncated normal", *Econometrica* 41, 997-1016.

[3] Anselin, L. (1988): "Spatial econometrics: methods and Models", Kluwer Academic Publishers.

[4] Anselin, L. and Florax, R.J.G.M. (1995): "New direction in spatial econometrics", Springer-Verlag, Berlin, Germany.

[5] Anselin, L. Florax, R.J.G.M, and Rey, J.S. (2004): "Econometrics for spatial models: Recent advances", *Advances in Spatial econometrics*. Springer-Verlag, Berlin, Germany,1-28.

[6] Beron, K.J. and Vijverberg, W.P. (2003): "Probit in a spatial context: A Monte Carlo approach", *Advances in Spatial econometrics*. Springer-Verlag, Berlin, Germany, 169-196.

[7] Bernstein, S. (1927): "Sur l'Extension du Theoreme du Calcul des Probabilities aux Sommes de Quantities Dependantes", *Mathematische Annalen* 97, 1-59.

[8] Case, A.C. (1991): "Spatial patterns in household demand", *Econometrica* 59, 953-965.

[9] Case, A.C. (1992): " Neighborhood influence and technology change", *Regional Science and Urban Economics* 22, 491-508.

[10] Conley, T. G. (1999): "GMM estimation with cross sectional dependence", *Journal of Econometrics* 92, 1-45.

[11] Davidson, J. (1994): "Stochastic limit theory", Oxford: Oxford University Press.

[12] Fleming, M. M.(2005): "Techniques for estimation spatially dependent discrete choice models", *Advances in Spatial econometrics*. Springer-Verlag, Berlin, Germany, 145-168.

[13] Gourieroux, C. (2000): "Econometrics of qualitative dependent variables", Cambridge University Press.

[14] Greene, W.H. (2003): "Econometrics Analysis", 4th Edition, Prentice-Hall, Upper Saddle River, N.J.

[15] Harvey, A. (1976): " Estimating regression models with multiplicative heteroscedasticity", *Econometrica* 44, 461-465.

[16] Kelejian, H.H. and Prucha, I. R. (1999): "A generalized moments estimator for the autpregressive parametre in a spatial model", *International Economic Review* 40, 509-533.

[17] Kelejian, H.H. and Prucha, I. R. (2001): "On the asymptotic distribution of the Moran I test statistic with applications", *Journal of Econometrics* 104, 219-257.

[18] Kotz, S. Balakrishnan, N. and Johnson, N. (2000): " Continuous Multivariate Distributions", 2nd Edition. Wiley Series in Probability and Statistics.

[19] Lee, L.-F. (2004): "Asymptotic distribution of quasi-maximum likelihood estimators for spatial autoregressive models", *Econometrica* 72, 6, 1899-1925.

[20] Lesage, J. P. (2000): " Bayesian estimation of limit dependent variable spatial autoregressive models", *Geographical Analysis* 32, 19-35.

[21] McLeish, D. L. (1974): "Dependent Central Limit Theorems and Invariance Principals", *Annals of Probability* 2, 620-628.

[22] McMillan, D. P. (1995): "Spatial effects in Probit models: A Monte Carlo Investigation", *New directions in Spatial econometrics*. Springer-Verlag, Berlin, Germany, 189-228.

[23] McMillan, D. P. (1992): "Probit with spatial autocorrelation", *Journal of Regional Science* 32, 335-348.

[24] Mukherjea, A. and Stephens, R. (1990): "The problem of identification of parameters by the distribution of the maximum random variable: solution for the trivariate normal case", *Journal of Multivariate Analysis* 34, 95-115.

[25] Newey, W.K. and West, K. D. (1987): "A simple, positive semi-definite, Heteroskedasticity and autocorrelation consistent covariance matrix", *Econometrica* 55, 703-308.

[26] Newey, W.K and Mcfadden, D. (1994): " Large sample estimation and hypothesis testing", *Handbook of Econometrics*, Ch. 36, Vol 4, North-Holland, New York.

[27] Pinkse, J. Shen L. and Slade, M. E. (2007): "A central limit theorem for endogenous locations and complex spatial interactions", *Journal of Econometrics* 140, 215-225.

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

[29] Plackett, R.L. (1954): "A reduction formula for normal multivariate integrals", *Biometrika* 41, 351-360.

[30] Poirier, D. and Ruud, P. A. (1988): "Probit with dependent observations", *Review of Economic Studies* 55, 593-614.

[31] Robinson, P. M. (1982): "On the asymptotic properties of estimators of models containing limit dependent variables", *Econometrica* 50, 27-41.

[32] White, H. (2001): "Asymptotic theory for econometricians", 2nd Edition, Orlando, FL. Academic Press.

[33] Wooldridge, J. (2002): "Econometric analysis of cross section and panel data", The MIT Press, Cambridge, Massachusetts.