IRTG 1792 Discussion Paper 2020-017

Using generalized estimating equations to estimate nonlinear models with spatial data

Cuicui Lu [*]
Weining Wang [*2] [*3]
Jeffrey M. Wooldridge [*4]

[*] Nanjing University Business School, China
[*2] University of London, UK
[*3] Humboldt-Universität zu Berlin, Germany
[*4] Michigan State University, USA

This research was supported by the Deutsche Forschungsgesellschaft through the International Research Training Group 1792 "High Dimensional Nonstationary Time Series".

http://irtg1792.hu-berlin.de
ISSN 2568-5619

Using generalized estimating equations to estimate nonlinear models with spatial data *

Cuicui Lu†, Weining Wang ‡, Jeffrey M. Wooldridge§

Abstract

In this paper, we study estimation of nonlinear models with cross sectional data using two-step generalized estimating equations (GEE) in the quasi-maximum likelihood estimation (QMLE) framework. In the interest of improving efficiency, we propose a grouping estimator to account for the potential spatial correlation in the underlying innovations. We use a Poisson model and a Negative Binomial II model for count data and a Probit model for binary response data to demonstrate the GEE procedure. Under mild weak dependency assumptions, results on estimation consistency and asymptotic normality are provided. Monte Carlo simulations show efficiency gain of our approach in comparison of different estimation methods for count data and binary response data. Finally we apply the GEE approach to study the determinants of the inflow foreign direct investment (FDI) to China.

keywords: quasi-maximum likelihood estimation; generalized estimating equations; nonlinear models; spatial dependence; count data; binary response data; FDI equation

JEL Codes: C13, C21, C35, C51

*This paper is supported by the National Natural Science Foundation of China, No.71601094 and German Research Foundation.

†Department of Economics, Nanjing University Business School, Nanjing, Jiangsu 210093 China; email: lucuicui@nju.edu.cn

‡Department of Economics, City, U of London; Northampton Square, Clerkenwell, London EC1V 0HB. Humboldt-Universität zu Berlin, C.A.S.E. - Center for Applied Statistics and Economics; email: weining.wang@city.ac.uk

§Department of Economics, Michigan State University, East Lansing, MI 48824 USA; email: wooldri1@msu.edu

1 Introduction

In empirical economic and social studies, there are many examples of discrete data which exhibit spatial or cross-sectional correlations possibly due to the closeness of geographical locations of individuals or agents. One example is the technology spillover effect. The number of patents a firm received shows correlation with that received by other nearby firms (E.g. Bloom et al. (2013)). Another example is the neighborhood effect. There is a causal effect between the individual decision whether to own stocks and the average stock market participation of the individual's community (E.g.Brown et al. (2008)). These two examples involves dealing with discrete data. The first example is concerned with count data and the second one handles binary response data. Nonlinear models are more appropriate than linear models for discrete response data. With spatial correlation, these discrete variables are no longer independent. Both the nonlinearity and the spatial correlation make the estimation difficult.

In order to estimate nonlinear models, one way is to use maximum likelihood estimation (MLE). A full MLE specifies the joint distribution of spatial random variables. This includes correctly specifying the marginal and the conditional distributions, which impose very strong assumptions on the data generating processes. However, given a spatial data set, the dependence structure is generally unknown. If the joint distribution of the variables is misspecified, MLE is in general not consistent. One of the alternative MLE method is partial-maximum likelihood estimation (PMLE), which only uses marginal distributions. Wang et al. (2013) use a bivariate Probit partial MLE to improve the estimation efficiency with a spatial Probit model. Their approach requires to correctly specify the marginal distribution of the binary response variable conditional on the covariates and distance measures[1] There are two concerns with Wang et al. (2013). First the computation is already hard for a bivariate distribution. The multivariate marginal distribution of a higher dimensional variable, e.g., trivariate, is more computationally demanding; second it also requires the correct specification of the marginal bivariate distribution to obtain consistency. The bivariate marginal distribution of a spatial multivariate normal distribution is bivariate normal, thus the bivariate Probit model can be derived. But there are other distributions whose marginal

[1]A sample of spatial data is collected with a set of geographical locations. Spatial dependence is usually characterized by distances between observations. A distance measure is how one defines the distances between observations. Physical distance or economic distance could be two options. Information about agents locations is commonly imprecise, e.g. only zip code is known. Conley and Molinari (2007) deals with the inference problem when there exist distance errors. In this paper we assume there are no measurement errors in pairwise distances.

distribution is not the same anymore. For example, the marginal distribution of a multivariate Logit is not logistic. If the partial likelihood is misspecified, the estimation of the mean parameters could be not consistent. With less distributional assumptions, the quasi-maximum likelihood estimation (QMLE) can also be used to estimate nonlinear models. Using a density that belongs to a linear exponential family (LEF), QMLE is consistent if we correctly specify the conditional mean while other features of the density can be misspecified (Gourieroux et al. (1984)). Lee (2004) derives asymptotic distributions of quasi-maximum likelihood estimators for spatial autoregressive models by allowing not assuming normal distributions. In a panel data case, pooled or partial QMLE (PQMLE) which ignores serial correlations is consistent under some regularity conditions (Wooldridge (2010)).

We further relax distributional assumptions than those required in bivariate partial MLE as in Wang et al. (2013). Suppose we only assume correct mean function and one working variance covariance matrix[2] which may not be correct. Using QMLE in the LEF, we can consistently estimate the mean parameters as well as the average partial effects. The generalized estimating equations (GEE) approach is one of the QMLE methods. It is used in panel data models to account for serial correlation and thus get more efficient estimators. A generalized estimating equation is used to estimate the parameters of a generalized linear model with a possible unknown correlation between outcomes (Liang and Zeger (1986)). Parameter estimates from the GEE are consistent even when the variance and covariance structure is misspecified under mild regularity conditions. This is quite related to a different terminology, composite likelihood. Varin et al. (2011) provide a survey of developments in the theory and application of composite likelihood. The motivation for the use of composite likelihood is usually computational, to avoid computing or modelling the joint distributions of high dimensional random processes. One can find many related reference in the literature, such as Bhat et al. (2010). As a special case of composite likelihood methods, one way is to use partial conditional distribution, and maximize the summand of log likelihoods for each observation. It assumes a working independence assumption, which means that the estimators are solved by ignoring dependence between individual likelihoods. The parameters can be consistently estimated if the partial log likelihood function satisfies certain regularity assumptions. However, a consistent variance estimator should be provided for valid inference[3]. When there exists spatial correlation, the pooled maxi-

[2]The true variance covariance matrix is generally unknown. By specifying a working variance covariance matrix, one can capture some of the correlation structure between observations.

[3]Ignoring dependence in the estimation of parameters will result in wrong inferences if the variances are calculated in the way that independence is assumed. Dependence should be accounted for to the

mum likelihoods (composite likelihoods) can be considered as misspecified likelihoods because of the independence assumption.

Generalized least squares (GLS) could be used to improve the estimation efficiency in a linear regression model even if the variance covariance structure is misspecified. Lu and Wooldridge (2017) propose a quasi-GLS method to estimate the linear regression model with an spatial error component. By first estimating the spatial parameter in the error variance and then using estimated variance matrix for within group observations, the quasi-GLS is computationally easier and would not loose much efficiency compared to GLS. Similarly, the multivariate nonlinear weighted least squares estimator (MNWLS), see Chapter 12.9.2 in Wooldridge (2010), is essentially a GLS approach applied in nonlinear models to improve the estimation efficiency.

It is worth noting that the GEE approach discussed in this paper is a two-step method, which is essentially a special MNWLS estimator that uses a LEF variance assumption and a possibly misspecified working correlation matrix in the estimation. The GEE approach was first extended to correlated data by Liang and Zeger (1986), which propose a fully iterated GEE estimator in a panel data setting. In addition, Zeger and Liang (1986) fit the GEE method to discrete dependent variables. The iterated GEE method has solutions which are consistent and asymptotically Gaussian even when the temporal dependence is misspecified. The consistency of mean parameters only depends on the correct specification of the mean, not on the choice of working correlation matrix. GEE used in nonlinear panel data models and system of equations is supposed to obtain more efficient conditional mean parameters with covariance matrix accounting for the dependency structure of the data. In this paper, we apply a similar idea to grouped spatial data. We use the PQMLE as the initial estimator for the two-step GEE and study the efficiency properties of a two-step GEE estimator and expect that GEE can give more efficient estimators compared to PQMLE.

Moreover, we demonstrate theoretically how to use our GEE approach within the QMLE framework in a spatial data setting to obtain consistent estimators. We give a series of assumptions, based on which QMLE estimators are consistent for the spatial processes. To derive the asymptotics for the GEE estimator we have to use a uniform law of large numbers (ULLN) and a central limit theorem (CLT) for spatial data. These limit theorems are the fundamental building blocks for the asymptotic theory of nonlinear spatial M-estimators, for example, maximum likelihood estimators (MLE) and generalized method of moments estimators (GMM) (Jenish and Prucha (2012)). Conley (1999) makes an important contribution toward developing an asymptotic the-

extent of how much one ignores it in the estimation.

ory of GMM estimators for spatial processes. He utilizes Bolthausen (1982) CLT for stationary random fields. Jenish and Prucha (2009, 2012) provide ULLNs and a CLTs for near-epoch dependent spatial processes. Using theorems in Jenish and Prucha (2009, 2012), one can analyze more interesting economic phenomena. It should be noted that although GEE can be considered as a special case of M-estimation, we have carefully checked how the near-epoch dependence property of the underlying processes is translated to our responses and the partial sum processes involved in proving the asymptotics of the estimation. Our setup is different from the literature as it is with a grouped estimation structure. Finally, we have provided a consistency proof of the proposed semiparametric estimator of the variance covariance matrix.

We contribute to the literature in three aspects. First, we propose a simple method which uses less distributional assumptions by only specifying the conditional mean for spatial dependent data. The method is computationally easier by dividing data into small groups compared to using all information. We model the spatial correlation as a moving average (MA) type in the underlying innovations instead of the spatial autoregressive (SAR) model in the dependent variable. Second, we proved the theoretical property of our estimator by applying ULLN and CLT in Jenish and Prucha (2009, 2012) to the GEE estimator with careful checking the hyper assumptions. Third, we emphasize the possible efficiency gain from making use of spatial correlation from our simulation study, and we demonstrate how to use GEE with two types of data: count and binary response.

In Section 2, the GEE methodology in a QMLE framework under the spatial data context is proposed. In Section 3, we look in detail at a Poisson model and Negative Binomial II model for count data with a multiplicative spatial error term. We further study a Probit model for binary response data with spatial correlation in the latent error term. In Section 4, a series of assumptions are given based on Jenish and Prucha (2009, 2012) under which GEE-estimators are consistent and have an asymptotic normal distribution. The asymptotic distributions for GEE for spatial data are derived. Consistent variance covariance estimators are provided for the nonlinear estimators. Section 5 contains Monte Carlo simulation results which compare efficiency of different estimation methods for the nonlinear models explored in the previous section. Section 6 contains an application to study the determinants of the inflow FDI to China using city level data. The technical details are delegated to Section 7.

2 Methodology

2.1 Notation and definition

Unlike linear models, a very important feature of nonlinear models is that estimators cannot be obtained in a closed form, which requires new tools for asymptotic analysis: uniform law of large numbers (ULLN) and a central limit theorem (CLT). Jenish and Prucha (2009) develop ULLN and CLT for α-mixing random fields on unevenly spaced lattices that allow for nonstationary processes with trending moments. But the mixing property can fail for quite a few reasons, thus we adopt the notion of near-epoch dependence (NED) as in Jenish and Prucha (2012) which refers to a generalized class of random fields that is "closed with respect to infinite transformations." We consider spatial processes located on a unevenly spaced lattice $D \subseteq \mathbb{R}^d, d \geq 1$. The space $\mathbb{R}^d$ is endowed with the metric $\rho(i,j) = max_{1\leq l\leq d}|j_l - i_l|$ with the corresponding norm $|i|_\infty = max_{1\leq l\leq d}|i_l|$, where i_l is the l-th component of i. The distance between any subsets $U, V \in D$ is defined as $\rho(U,V) = \inf\{\rho(i,j) : i \in U \text{ and } j \in V\}$. Further, let $|U|$ denote the cardinality of a finite subset $U \subseteq D$. The setting is illustrated in Jenish and Prucha (2009, 2012).

Let $Z = \{Z_{n,i}, i \in D_n, n \geq 1\}$ and $\varepsilon = \{\varepsilon_{n,i}, i \in T_n, n \geq 1\}$ be triangular arrays of random fields defined on a probability space $(\Omega, \mathscr{F}, P)$ with $D_n \subseteq T_n \subseteq D$ where D satisfies A.1). The cardinality of D_n and T_n satisfy $\lim_{n\to\infty} |D_n| \to \infty$, $\lim_{n\to\infty} |T_n| \to \infty$. For any vector $v \in R^p$, $|v|_2$ denotes the L_2 norm of v. For any $n \times m$ matrix A with element a_{ij}, denote $|A|_1 = \max_{1\leq j\leq m} \sum_{i=1}^{n} |a_{ij}|$ and $|A|_\infty = \max_{1\leq i\leq n} \sum_{j=1}^{m} |a_{ij}|$, $|A|_2$ denotes the 2-norm. For any random vector X, denote $\|X_{n,i}\|_p = (\mathsf{E}\,|X_{n,i}|^p)^{1/p}$ as its L_p-norm, where the absolute pth moment exists. We brief $\|X_{n,i}\|_2$ as $\|X_{n,i}\|$. Let $\mathcal{F}_{n,i}(s) = \sigma(\varepsilon_{n,j} : j \in D_n, \rho(i,j) \leq s)$ as the σ- field generated by random vectors $\varepsilon_{n,j}$ located within distance s from i. Given two sequences of positive numbers x_n and y_n, write $x_n \lesssim y_n$ if there exists constant $C > 0$ such that $x_n/y_n \leq C$, also we can write $x_n = \mathcal{O}(y_n)$. A sequence x_n is said to be $o(y_n)$ if $x_n/y_n \to 0$, as $n \to \infty$. In a similar manner, The notation, $X_n = \mathcal{O}_p(a_n)$ means that the set of values X_n/a_n is stochastically bounded. That is, for any $\varepsilon > 0$, there exists a finite M > 0 and a finite N > 0 such that, $P(|X_n/a_n| > M) < \varepsilon, \forall n > N$. $|.|_a$ is the elementwise absolute value of a matrix $|A|_a$. $a \vee b$ is $\max(a,b)$.

Definition 1. *Let $Z = \{Z_{n,i}, i \in D_n, n \geq 1\}$ and $\varepsilon = \{\varepsilon_{n,i}, i \in D_n, n \geq 1\}$ be random fields with $\|Z_{n,i}\|_p < \infty, p \geq 1$, where $D_n \subseteq D$ and its cardinality $|D_n| = n$. Let $\{d_{n,i}, i \in D_n, n \geq 1\}$ be an array of finite positive constants. Then the random field Z*

is said to be L_p-near-epoch dependent on the random field ε if

$$\|Z_{n,i} - \mathsf{E}(Z_{n,i}|\mathcal{F}_{n,i}(s))\|_p < d_{n,i}\varphi(s)$$

for some sequence $\varphi(s) \geq 0$ with $\lim_{s\to\infty} \varphi(s) = 0$. $\varphi(s)$ are denoted as the NED coefficients, and $d_{n,i}$ are denoted as NED scaling factors. If $\sup_n \sup_{i\in D_n} d_{n,i} < \infty$, then Z is called as uniformly L_p-NED on ε.

A.1) The lattice $D \subseteq \mathbb{R}^d, d \geq 1$, is infinitely countable. The distance $\rho(i,j)$ between any two different individual units i and j in D is at least larger than a positive constant, i.e., $\forall i,j \in D : \rho(i,j) \geq \rho_0$, w.l.o.g. we assume $\rho_0 > 1$.

We will present the L_2-NED properties of a random field Z on some α-mixing random field ε. The definition of the α-mixing coefficient employed in the paper are stated as following.

Definition 2. *Let $\mathscr{A}$ and $\mathscr{B}$ be two σ-algebras of $\mathscr{F}$, and let*

$$\alpha(\mathscr{A},\mathscr{B}) = \sup(|P(A\cap B) - P(A)P(B)|, A \in \mathscr{A}, B \in \mathscr{B}),$$

For $U \subseteq D_n$ and $V \subseteq D_n$, let $\sigma_n(U) = \sigma(\varepsilon_{n,i}, i \in U)$ $(\sigma_n(V) = \sigma(\varepsilon_{n,i}, i \in V))$ and $\alpha_n(U,V) = \alpha(\sigma_n(U), \sigma_n(V))$. Then, the α-mixing coefficients for the random field ε are defined as:

$$\overline{\alpha}(u,v,h) = \sup_n \sup_{U,V}(\alpha_n(U,V), |U| \leq u, |V| \leq v, \rho(U,V) \geq h).$$

Note that we suppress the dependence on n from now on for the triangular array. Let $\{(\mathbf{x}_i, y_i), i = 1,2,...,n\}$, where $(\mathbf{x}_i, y_i)$ is the observation at location s_i. $\mathbf{x}_i$ is a row vector of independent variables which can be continuous, discrete or a combination. The dependent variable y_i can be continuous or discrete. Let $(\mathbf{x}_g, \mathbf{y}_g)$ be the observations in group g and B_g is the associated set of locations within the group g. We will focus on the case of a discrete dependent variable, a binary response and a count. Let $\theta \in \mathbf{R}^p$, $\gamma \in \mathbf{R}^q$ and $\theta \in \mathbf{\Theta}, \gamma \in \mathbf{\Gamma}$, where $\mathbf{\Theta} \times \mathbf{\Gamma}$ is a compact set, and (θ^0, γ^0) is the true parameter value.

2.2 The generalized estimating equations methodology

The GEE methodology proposed in equations (6) and (7) in Liang and Zeger (1986) is an iterated approach to estimate the mean parameters. We simplify the procedure

using a two-step method by first estimate the working correlation matrix and then apply MWNLS. In the following, we write the GEE methodology in the group level notation. Groups are divided according to geographical properties or other researcher defined economic (social) relationships. Our asymptotic analysis is based on large number of groups $g = 1, \cdots, G$. The notation D_G indicates the lattice containing group locations, each group location is denoted as vectorizing the elements in B_g. Let the total number of groups be $|D_G| = G$, while the total number of observations is still $|D_n| = n$. Let L_g be the number of observations in group g. For simplicity assume $L_g = L$, for all g. Let $\{(\mathbf{x}_g, \mathbf{y}_g)\}$ be the observations for group g, where $\mathbf{x}_g$ is an $L \times p$ matrix and $\mathbf{y}_g$ is an $L \times 1$ vector. There are two extreme cases of the group size. The first case is when the group size is 1, the resulting estimator is the usual PQMLE estimator, which means we ignore all of the pairwise correlations. The second case is when the group size is n, which means we are using all the pairwise information. If the group size is not equal to 1 or n, the estimation is actually a "partial" QMLE. By "partial", we mean that we do not use full information, but only the information within the same groups. Note that we work with the case with number of groups $G \to \infty$ in our theory, while the groupsize L is assumed to be fixed.

Assume that we correctly specify conditional mean of $\mathbf{y}_g$, that is, the expectation of $\mathbf{y}_g$ conditional on $\mathbf{x}_g$ is

$$\mathsf{E}\left(\mathbf{y}_g|\mathbf{x}_g\right) = \mathbf{m}_g\left(\mathbf{x}_g; \theta^0\right) = \mathbf{m}_g(\theta^0). \tag{1}$$

Assume the conditional variance-covariance matrix of $\mathbf{y}_g$ is $\mathbf{W}_g^*$ which is unknown in most cases, where $\mathbf{W}_g \overset{\text{def}}{=} \text{Cov}(y_g, y_g|\mathbf{x}_g) = \mathsf{E}(\mathbf{y}_g\mathbf{y}_g^\top|\mathbf{x}_g) - \mathsf{E}(\mathbf{y}_g|\mathbf{x}_g)\,\mathsf{E}(\mathbf{y}_g|\mathbf{x}_g)^\top$. Usually we parameterize a corresponding weight matrix $\mathbf{W}_g$ by $\mathbf{W}_g(\theta, \gamma)$, where $\theta \in \Theta \subset \mathbf{R}^q$ and $\gamma \in \Gamma \subset \mathbf{R}^p$ as a nuisance parameter involved only in the estimation of the variance covariance matrix. In practice, we usually preestimate γ and thus it is replaced by a consistent estimate of $\hat{\gamma}$, then $\mathbf{W}_g$ is denoted as $\mathbf{W}(\theta, \hat{\gamma})$.

The objective function for group g and the whole sample are given as follows:

$$q_g(\theta, \gamma) \overset{\text{def}}{=} \left(\mathbf{y}_g - \mathbf{m}_g(\theta)\right)^\top \mathbf{W}_g^{-1}\left(\theta, \gamma\right)\left(\mathbf{y}_g - \mathbf{m}_g(\theta)\right), \tag{2}$$

$$Q_G(\theta, \gamma) \overset{\text{def}}{=} (M_G G)^{-1} \sum_g q_g(\theta), \tag{3}$$

where M_G is a scaling constant defined in A.5) in section 4.

Theoretically, an estimator of θ^0, γ^0 is given by

$$(\hat{\theta}, \hat{\gamma}) = \mathbf{argmin}_{\theta,\gamma \in \mathbf{\Theta},\mathbf{\Gamma}} Q_G(\theta, \gamma). \tag{4}$$

In practice a GEE estimator is obtained by a two-step procedure, where the first step is to estimate the nuisance parameter γ and the second step is to have the parameter θ estimated with the plug-in estimator $\hat{\gamma}$ from step 1.

$$\hat{\theta}_{\text{GEE}} = \mathbf{argmin}_{\theta \in \mathbf{\Theta}} Q_G(\theta, \hat{\gamma}). \tag{5}$$

Because this only uses the groupwise information, it actually is a "quasi" or "pseudo" MWNLS. The quasi-score equation, which is the first order condition for GEE, is defined as follows:

$$\mathbf{S}_G(\theta, \gamma) = \frac{1}{GM_G} \sum_g \nabla \mathbf{m}_g(\theta)^\top \mathbf{W}_g^{-1}(\theta, \gamma) \left[\mathbf{y}_g - \mathbf{m}_g(\theta)\right], \tag{6}$$

where $\nabla_\theta \mathbf{m}_g(\theta)$ is the gradient of $\mathbf{m}_g(\theta)$. M_G is defined as the scaling constant in A.5) in section 4. The GEE estimator $(\hat{\theta}, \hat{\gamma}) = \text{argzero}_{\theta \in \Theta, \gamma \in \Gamma} \mathbf{S}_G(\theta, \gamma)$.

Denote the population version of loss as $\mathbf{S}_\infty(\theta, \gamma) = \lim_{G \to \infty} \mathsf{E}\, \mathbf{S}_G(\theta, \gamma)$, and $Q_\infty(\theta, \gamma) = \lim_{G \to \infty} (GM_G)^{-1} \sum_g \mathsf{E}\, q_g(\theta, \gamma)$. Thus the true parameter (θ^0, γ^0) $= \text{argzero}_{\theta \in \Theta, \gamma \in \Gamma} \mathbf{S}_\infty(\theta, \gamma) = \text{argmin}_{\theta \in \Theta, \gamma \in \Gamma} Q_\infty(\theta, \gamma)$.

Frequently we restrict our attention to the exponential family, which embraces many frequency encountered distributions, such as Bernoulli, Poisson and Gaussian, etc.

Now we write this estimation in a QMLE framework. We suppress the parameter γ for a moment. Assume the probability density function $f(\mathbf{y}_g | \mathbf{x}_g; \theta)$ is in the LEF.(See details in Appendix 7.7 .)

Without accounting for the spatial covariance, one characterization of QMLE in LEF is that the individual score function has the following form:

$$\mathbf{s}_i(\theta) = \nabla m_i(\theta)^\top \{y_i - m_i(\theta)\} / v_i(m_i(\theta)), \tag{7}$$

where $\nabla m_i(\mathbf{x}_i; \theta)$ is the $1 \times p$ gradient of the mean function and $v_i(m_i(\mathbf{x}_i, D_n; \theta))$ is the conditional variance function associated with the chosen LEF density. For Bernoulli distribution, $v_i(m_i(\mathbf{x}_i; \theta)) = m_i(\mathbf{x}_i; \theta)(1 - m_i(\mathbf{x}_i; \theta))$, and for Poisson distribution, $v_i(m_i(\mathbf{x}_i; \theta)) = m_i(\mathbf{x}_i; \theta)$. Note that (7) gives a consistent estimator but is not likely to be the most efficient estimator as it ignores the possible spatial correlations between observations. However, it accounts for possible heteroscedasticity.

We write the quasi-score function for a group. Let $\mathbf{v}_g(\mathbf{m}_g(\mathbf{x}_g;\theta))$ be the conditional variance covariance matrix for group g. Then score involved in the estimation is denoted as

$$\mathbf{S}_G(\theta) = \frac{1}{M_G G}\sum_g s_g(\theta) = \frac{1}{M_G G}\sum_g \nabla\mathbf{m}_g(\mathbf{x}_g;\theta)^\top \mathbf{v}_g(\mathbf{m}_g(\mathbf{x}_g;\theta))^{-1}[\mathbf{y}_g - \mathbf{m}_g(\mathbf{x}_g;\theta)], \tag{8}$$

where

$$s_g(\theta) = \nabla\mathbf{m}_g(\mathbf{x}_g;\theta)^\top \mathbf{v}_g(\mathbf{m}_g(\mathbf{x}_g;\theta))^{-1}[\mathbf{y}_g - \mathbf{m}_g(\mathbf{x}_g;\theta)]. \tag{9}$$

We specify a more general form of variance $\mathbf{v}_g(\theta)$ with the dependency of the nuisance parameter γ. The conditional mean vector is correctly specified for each individual $\mathrm{E}(y_i|\mathbf{x}_i) = m_i(\mathbf{x}_i;\theta^0)$. Thus for each group, $\mathbf{m}_g(\mathbf{x}_g;\theta^0) = \mathsf{E}(\mathbf{y}_g|\mathbf{x}_g)$. Let $s_g(\theta,\gamma)$ denote the $p\times 1$ vector of score for group g. Let $h_g(\theta,\gamma)$ be the $p\times p$ matrix of Hessian for group g. The score function for $Q_G(\theta,\gamma)$ can be defined as $\mathbf{S}_G(\theta,\gamma)$ and the Hessian can be defined as $\mathbf{H}_G(\theta,\gamma)$. The score function for GEE can be written as

$$\mathbf{S}_G(\theta,\gamma) = \frac{1}{M_G G}\sum_g s_g(\theta,\gamma) = \frac{1}{M_G G}\sum_g \nabla\mathbf{m}_g^\top(\theta)\mathbf{W}_g^{-1}(\theta,\gamma)[\mathbf{y}_g - \mathbf{m}_g(\theta)]. \tag{10}$$

and the Hessian is

$$\begin{aligned}
\mathbf{H}_G(\theta,\gamma) &\equiv \frac{1}{M_G G}\sum_g h_g(\theta,\gamma)\\
&= -\frac{1}{M_G G}\sum_g \nabla_\theta\mathbf{m}_g^\top(\theta)\mathbf{W}_g^{-1}(\theta,\gamma)\nabla_\theta\mathbf{m}_g(\theta)\\
&\quad +\frac{1}{M_G G}\sum_g[\{(\mathbf{y}_g - \mathbf{m}_g(\theta))^\top\mathbf{W}_g^{-1}(\theta,\gamma)\otimes I_q\}]\partial\mathrm{Vec}(\nabla\mathbf{m}_g^\top(\theta))/\partial\theta\\
&\quad +\frac{1}{M_G G}\sum_g\{(\mathbf{y}_g - \mathbf{m}_g(\theta))^\top\otimes\nabla\mathbf{m}_g^\top(\theta)\}\partial\mathrm{Vec}(\mathbf{W}_g(\theta,\gamma))/\partial\theta\\
&\stackrel{\mathrm{def}}{=} \mathbf{H}_{G,1}(\theta,\gamma) + \mathbf{H}_{G,2}(\theta,\gamma) + \mathbf{H}_{G,3}(\theta,\gamma),
\end{aligned} \tag{11}$$

where Vec is denoted as the vectorization of a matrix A.

2.3 The first-step estimation of the weight matrix

In this subsection, we demonstrate one way to find an estimator for γ involved in $\mathbf{W}_g(\theta,\gamma)$. $\mathbf{W}_g(\theta,\gamma)$ can be written as

$$\mathbf{W}_g(\theta,\gamma) = \mathbf{V}_g(\mathbf{x}_g;\theta)^{1/2}\mathbf{R}_g(\gamma,D_G)\mathbf{V}_g(\mathbf{x}_g;\theta)^{1/2}, \tag{12}$$

where $\mathbf{V}_g$ is the $L \times L$ diagonal matrix that only contains variances of $\mathbf{y}_g - \mathbf{m}_g(\mathbf{x}_g, \theta^0)$ and $\mathbf{R}_g$ is the $L \times L$ correlation matrix for group g.

Let

$$\mathbf{V}_g(\mathbf{x}_g; \theta) = \begin{pmatrix} v_{g1} & 0 & \cdots & 0 \\ 0 & v_{g2} & & \vdots \\ \vdots & & \ddots & 0 \\ 0 & ... & 0 & v_{gL} \end{pmatrix}, \tag{13}$$

where the lth element on the diagonal is $v_{gl} = \text{Var}(\mathbf{y}_{gl}|\mathbf{x}_{gl})$ in group g, $\mathbf{y}_{gl}$ is the lth element in the vector $\mathbf{y}_g$ and $\mathbf{x}_{gl}$ is the lth row in $\mathbf{x}_g$. And

$$\mathbf{R}_g(\gamma, D_G) = \begin{pmatrix} 1 & \pi_{g12} & \cdots & \pi_{g1L} \\ \pi_{g21} & 1 & & \vdots \\ \vdots & & \ddots & \pi_{gL-1,L} \\ \pi_{gL1} & ... & \pi_{gL,L-1} & 1 \end{pmatrix}. \tag{14}$$

Let d_{glm} be the distance between the lth and the mth observations in group g. An example of a parametrization of the correlation i.e. the l, mth, $l \neq m$, element of $\mathbf{R}_g$, as in Cressie (1992) is

$$\pi_{glm} = 1 - b - c\left[1 - \exp\left(-d_{glm}/\rho\right)\right], \tag{15}$$

where the spatial correlation parameters $\gamma = (b, c, \rho)$, $b \geq 0, c \geq 0, \rho \geq 0$, and $b + c \leq 2$.[4]Set $b = c = 1$ without loss of generality. Then

$$\pi_{glm} = \begin{cases} 1 & \text{if } l = m, \\ \exp\left(-d_{glm}/\rho\right) & \text{otherwise.} \end{cases} \tag{16}$$

Although the above specification does not represent all the possibilities, it at least provides a way of how to parameterize the spatial correlation, and therefore the basis for testing spatial correlation.

The following provides a way to estimate γ. Let $\check{\theta}$ be the first-step PQMLE estimator. $\check{u}_i = y_i - m_i\left(x_i; \check{\theta}\right)$ are the first-step residuals. $\check{v}_i = v\left(m_i\left(\mathbf{x}_i; \check{\theta}\right)\right)$ is the fitted variance of individual i corresponding to the chosen LEF density. Let $\check{r}_i = \check{u}_i/\sqrt{\check{v}_i}$ be

[4]See Cressie (1992) p.61 for more examples.

the standardized residual. Let $\check{\mathbf{r}}_g = (\check{r}_{g1}, \check{r}_{g2}, ..., \check{r}_{gL})^\top$. Then $\check{\mathbf{r}}_\mathbf{g}\check{\mathbf{r}}_\mathbf{g}{}^\top$ is the estimated sample correlation matrix for group g. Let $\mathbf{e}_g(\check{\theta})$ be a vector containing $L(L-1)/2$ different elements of the lower (or upper) triangle of $\check{\mathbf{r}}_\mathbf{g}\check{\mathbf{r}}_\mathbf{g}{}^\top$, excluding the diagonal elements. Let $\mathbf{z}_g(\gamma)$ be the vector containing the elements in $\mathbf{R}_g$ corresponding to the same entries of elements in $\check{\mathbf{r}}_\mathbf{g}\check{\mathbf{r}}_\mathbf{g}{}^\top$. We can follow Prentice (1988), who provides one way to find a consistent estimator for γ by solving:

$$\hat{\gamma} = \mathbf{argmin}_{\gamma\in\Gamma} \sum_g (\mathbf{e}_g(\check{\theta}) - \mathbf{z}_g(\gamma))^\top (\mathbf{e}_g(\check{\theta}) - \mathbf{z}_g(\gamma)). \tag{17}$$

3 Estimating nonlinear models with spatial error: two examples

The setup of nonlinear models with spatial data varies with different models. For each model, we need to incorporate the spatial correlated term in an appropriate way. In this Section, we will demonstrate how we incorporate the spatial correlated error term in two types of discrete data and how to use a GEE procedure to estimate the nonlinear models. The first example is for count data and the second one is for binary response data.

3.1 Example 1 Count data with a multiplicative spatial error

A count variable is a variable that takes on nonnegative integer values, such as the number of patents applied for by a firm during a year. Bloom et al. (2013) studies spillover effects of R&D between firms in terms of firm patents. Other examples include the number of times someone being arrested during a given year. Count data examples with upper bound include the number of children in a family who are high school graduates, in which the upper bound is number of children in the family (Wooldridge (2010)).

3.1.1 Poisson model

We first model the count data with a conditional Poisson density, $f(y|\mathbf{x}) = \exp[-\mu]\,\mu^y/y!$, where $y! = 1 \cdot 2 \cdot ... \cdot (y-1) \cdot y$ and $0! = 1$. μ is the conditional mean of y. The Poisson QMLE requires us only to correctly specify the conditional mean. A default assumption for the Poisson distribution is that the mean is equal to the variance. Note that even if y_i does not follow the Poisson distribution, the QMLE approach will give a

consistent estimator if you use the Poisson density function and a correctly specified conditional mean (Gourieroux et al. (1984)). Moreover, y_i even need not to be a count variable. The most common mean function in applications is the exponential form:

$$\mathsf{E}\left(y_i|\mathbf{x}_i\right) = \exp\left(\mathbf{x}_i\beta_0\right). \tag{18}$$

When spatial correlation exists, we can characterize count data model with a multiplicative spatial error. Silva and Tenreyro (2006) use the Poisson pseudo-maximum-likelihood (PPML), which is the Poisson QMLE in this paper, to estimate the gravity model for trade. They argue that constant elasticity models should be estimated in their multiplicative form, because using a log linear model can cause bias in coefficient estimates under heteroskedasticity. Now we further consider the Poisson regression model with spatial correlation in the multiplicative error,

$$\mathsf{E}\left(y_i|\mathbf{x}_i, v_i\right) = v_i \exp\left(\mathbf{x}_i\beta_0\right), \tag{19}$$

where v_i is the multiplicative spatial error term. Let $\mathbf{v}$ equal $\left(v_1, v_2, ..., v_n\right)^\top$. (Note that for this example we treat location i as an one dimensional object.) This model is characterized by the following assumptions:

(1) $\{(\mathbf{x}_i, v_i), i = 1, 2, ..., n\}$ is a mixing sequence on the sampling space D_n, with mixing coefficient α.

(2) $\mathsf{E}\left(y_i|\mathbf{x}_i, v_i\right) = v_i \exp\left(\mathbf{x}_i\beta_0\right)$.

(3) y_i, y_j are independent conditional on $\mathbf{x}_i, \mathbf{x}_j, v_i, v_j, i \neq j$.

(4) v_i has a conditional multivariate distribution, $\mathsf{E}\left(v_i|\mathbf{x}_i\right) = 1$. $\mathrm{Var}\left(v_i|\mathbf{x}_i\right) = \tau^2$, $\mathrm{Cov}\left(v_i, v_j|\mathbf{x}_i, \mathbf{x}_j\right) = \tau^2 \cdot c\left(d_{ij}, \rho\right)$, where $c\left(d_{ij}, \rho\right)$ is the correlation function of v_i and v_j.

Under the above assumptions, and again conditional on D_n is suppressed, we can integrate out v_i by using the law of iterated expectations.

$$\mathsf{E}\left(y_i|\mathbf{x}_i, D_n\right) = \mathsf{E}\left(\mathsf{E}\left(y_i|\mathbf{x}_i, v_i\right)|\mathbf{x}_i, D_n\right) = \exp\left(\mathbf{x}_i\beta_0\right). \tag{20}$$

If x_j is continuous, the partial effects on $\mathsf{E}\left(y_i|\mathbf{x}_i, D_n\right)$ is $\exp\left(\mathbf{x}_i\beta_0\right)\beta_j$. If x_j is discrete the partial effects is the change in $\mathsf{E}\left(y_i|\mathbf{x}_i, \mathbf{D}_n\right)$ when, say, x_K goes from a_K to $a_K + 1$ which is

$$\exp\left(\beta_1 + x_2\beta_2 + ... + \beta_K\left(a_K + 1\right)\right) - \exp\left(\beta_1 + x_2\beta_2 + ... + \beta_K a_K\right). \tag{21}$$

The pooled QMLE gives a consistent estimator for the mean parameters, which solves:

$$\hat{\beta}_{PQMLE} = \mathbf{arg}\max_{\theta\in\Theta}\sum_{i=1}^{n} l_i(\beta) = \sum_{i=1}^{n} y_i\mathbf{x}_i\beta - \sum_{i=1}^{n}\exp(\mathbf{x}_i\beta) - \sum_{i=1}^{n}\log(y_i!). \tag{22}$$

Its score function is

$$\sum_{i=1}^{n}\mathbf{x}_i^{\top}\left[y_i - \exp\left(\mathbf{x}_i\check{\beta}_{\text{QMLE}}\right)\right] = \mathbf{0}. \tag{23}$$

Since this estimator does not account for any heteroskedasticity or spatial correlation, a robust estimator for the asymptotic variance of partial QMLE estimator is provided as follows,

$$\begin{aligned}\widehat{\text{Avar}}\left(\check{\beta}_{\text{QMLE}}\right) &= \left[\sum_{i=1}^{n}\exp\left(-\mathbf{x}_i\check{\beta}_{\text{QMLE}}\right)\mathbf{x}_i^{\top}\mathbf{x}_i\right]^{-1} \\ &\quad \sum_{i=1}^{n}\sum_{j=1}^{n}k(d_{ij})\,\mathbf{x}_i^{\top}\hat{u}_i\hat{u}_j\mathbf{x}_j\left[\sum_{i=1}^{n}\exp\left(-\mathbf{x}_i\check{\beta}_{\text{QMLE}}\right)\mathbf{x}_i^{\top}\mathbf{x}_i\right]^{-1},\end{aligned} \tag{24}$$

where $k(d_{ij})$ is a kernel function depending on the distance between observations i and j.

Moreover, a very specific nature of the Poisson distribution is that we can write down the conditional variances and covariances of y:

$$\text{Var}(y_i|\mathbf{x}_i, D_n) = \exp(\mathbf{x}_i\beta_0) + \exp(2\mathbf{x}_i\beta_0)\cdot\tau^2. \tag{25}$$

The conditional variance of y_i given $\mathbf{x}_i$ is a function of both the level and the quadratic of the conditional mean. The traditional Poisson variance assumption is that the conditional variance should equal the conditional mean. That is, $\text{Var}(y_i|\mathbf{x}_i) = \exp(\mathbf{x}_i\beta_0)$. The Poisson GLM variance assumption is $\text{Var}(y_i|\mathbf{x}_i) = \sigma^2\exp(\mathbf{x}_i\beta_0)$ with an overdispersion or underdispersion parameter σ^2, which is a constant. Obviously, there is over-dispersion in (25) since $\exp(2\mathbf{x}_i\beta_0)\cdot\tau^2 \geq 0$, and the over-dispersion parameter is $1 + \exp(\mathbf{x}_i\beta_0)\cdot\tau^2$, which is changing with $\mathbf{x}_i$. This does not coincide with Poisson variance assumption and the GLM variance assumption. What is more, the conditional covariances can be written in the following form,

$$\text{Cov}(y_i, y_j|\mathbf{x}_i, \mathbf{x}_j, D_n) = \exp(\mathbf{x}_i\beta_0)\exp(\mathbf{x}_j\beta_0)\cdot\tau^2\cdot c(d_{ij}, \rho). \tag{26}$$

In the group level notation,

$$\mathsf{E}(\mathbf{y}_g|\mathbf{x}_g, D_G) = \exp(\mathbf{x}_g\beta_0). \tag{27}$$

Let $\mathbf{W}_g$ be the variance-covariance matrix for group g evaluated at the true value β_0, ρ_0. The variance of the lth element in group g is

$$v_{gl} = \exp\left(\mathbf{x}_{gl}\beta_0\right)\left(1 + \exp\left(\mathbf{x}_{gl}\beta_0\right)\cdot\tau^2\right), \tag{28}$$

and the covariance of the lth and mth elements in group g is

$$r_{glm} = \exp\left(\mathbf{x}_{gl}\beta_0\right)\exp\left(\mathbf{x}_{gm}\beta_0\right)\cdot\tau^2\cdot c\left(d_{glm},\rho\right). \tag{29}$$

Here $\gamma = \left(\tau^2,\rho\right)^\top$ and $\hat{\gamma} = \left(\hat{\tau}^2,\hat{\rho}\right)^\top$ is an estimator for γ. Let $\check{\beta}_{\text{PQMLE}}$ be the partial QMLE estimator in the first step. Then the elements in $\mathbf{W}_g$ can be estimated as

$$\hat{v}_{gl} = \exp\left(\mathbf{x}_{gl}\check{\beta}_{\text{PQMLE}}\right) + \exp\left(2\mathbf{x}_{gl}\check{\beta}_{\text{PQMLE}}\right)\cdot\hat{\tau}^2, \tag{30}$$

$$\hat{r}_{glm} = \exp\left(\mathbf{x}_{gl}\check{\beta}_{\text{PQMLE}}\right)\exp\left(\mathbf{x}_{gm}\check{\beta}_{\text{PQMLE}}\right)\cdot\hat{\tau}^2\cdot c\left(d_{ij},\hat{\rho}\right). \tag{31}$$

Based on the conditional distribution, the first order conditions for GEE is:

$$\sum_g \mathbf{x}_g^\top \mathbf{W}_g^{-1}\left(\hat{\gamma},\hat{\theta}\right)\left[\mathbf{y}_g - \exp\left(\mathbf{x}_g\hat{\beta}_{\text{GEE}}\right)\right] = 0. \tag{32}$$

$\hat{\beta}_{\text{GEE}}$ is consistent and follows a normal distribution asymptotically by Theorem 1 and 2. We will brief $\mathbf{W}_g^{-1}\left(\hat{\gamma},\hat{\theta}\right)$ as $\hat{\mathbf{W}}_g^{-1}$ in the following text. The variance estimator for the asymptotic variance that is robust to misspecification of spatial correlation is:

$$\begin{aligned}\widehat{\text{Avar}}\left(\hat{\beta}_{\text{GEE}}\right) \;=\; & \left(\sum_g \exp\left(2\mathbf{x}_g^\top\hat{\beta}_{\text{GEE}}\right)_g \mathbf{x}_g^\top\hat{\mathbf{W}}_g^{-1}\mathbf{x}_g\right)^{-1} \\ & \left(\sum_g\sum_{h(\neq g)} k(d_{gh})\exp\left(\mathbf{x}_g^\top\hat{\beta}_{\text{GEE}} + \mathbf{x}_h\hat{\beta}_{\text{GEE}}\right)\mathbf{x}_g^\top\hat{\mathbf{W}}_g^{-1}\hat{\mathbf{u}}_g\hat{\mathbf{u}}_h^\top\hat{\mathbf{W}}_h^{-1}\mathbf{x}_h^\top\right) \\ & \left(\sum_g \exp\left(2\mathbf{x}_g^\top\hat{\beta}_{\text{GEE}}\right)\mathbf{x}_g^\top\hat{\mathbf{W}}_g^{-1}\mathbf{x}_g\right)^{-1}\end{aligned} \tag{33}$$

where $k(d_{gh})$ is a kernel function depending on the distances between groups. The distances could be the smallest distance between two observations belonging to different groups.

The pivotal parameters, τ^2 and ρ, can be estimated using the Poisson QMLE residuals. Let $\check{u}_i^2 = \left[y_i - \exp\left(\mathbf{x}_i\check{\beta}_{\text{QMLE}}\right)\right]^2$ be the squared residuals from the Poisson QMLE. Based on equation (28), τ^2 can be estimated as the coefficient by regressing $\check{u}_i^2 - \exp\left(\mathbf{x}_i\check{\beta}_{\text{QMLE}}\right)$ on $\exp\left(2\mathbf{x}_i\check{\beta}_{\text{QMLE}}\right)$. The situation to estimate ρ depends on the spe-

cific form of $c(d_{ij}, \rho)$. We would like to assume a structure, though it might be wrong, to approximate the true covariance. For example, suppose the covariance structure of e_i and e_j is $\exp\left(\frac{\rho}{d_{ij}}\right) - 1$, and the correlation structure is $c(d_{ij}, \rho) = \frac{\exp\left(\frac{\rho}{d_{ij}}\right)-1}{\mathrm{e}-1}$, then an estimator for ρ is:

$$\hat{\rho} = \text{argmin}_{\rho} \sum_{i=1}^{n} \sum_{j \neq i}^{n} \left\{ \frac{\check{u}_i \check{u}_j}{\exp\left(\mathbf{x}_i \check{\beta}\right) \exp\left(\mathbf{x}_j \check{\beta}\right)} - \left[\exp\left(\frac{\rho}{d_{ij}}\right) - 1\right] \right\}^2. \tag{34}$$

Then $\hat{\mathbf{W}}_g$ is obtained by plugging $\hat{\tau}^2$ and $\hat{\rho}$ back in the variance-covariance matrix. We can also directly calculate $\hat{\rho}$ as

$$\hat{\rho} = \frac{1}{n \cdot (n-1)} \sum_{i=1}^{n} \sum_{j \neq i}^{n} \left[\log\left(\frac{\check{u}_i \check{u}_j}{\exp\left(\mathbf{x}_i \check{\beta}\right) \exp\left(\mathbf{x}_j \check{\beta}\right)} + 1 \right) \cdot d_{ij} \right]. \tag{35}$$

3.1.2 The negative binomial model

Since the conditional variances and covariances can be written in a specific form, we would consider NegBin II model of Cameron and Trivedi (1986) as a more appropriate model. The NegBin II model can be derived from a model of multiplicative error in a Poisson model. With an exponential mean, $y_i|\mathbf{x}_i, v_i, D_n \sim$Poisson$[v_i \exp(\mathbf{x}_i \beta_0)]$. Under the above assumptions for Poisson distribution, with the conditional mean (20) and conditional variance (25), $y_i|\mathbf{x}_i$ is shown to follow a negative binomial II distribution. It implies overdispersion, but where the amount of overdispersion increases with the conditional mean,

$$\text{Var}(y_i|\mathbf{x}_i, D_n) = \exp(\mathbf{x}_i \beta_0) \left(1 + \exp(\mathbf{x}_i \beta_0) \cdot \tau^2\right). \tag{36}$$

Now the log-likelihood function for observation i is

$$\begin{aligned} l_i(\beta, \tau) \;=\; & \left(\tau^2\right)^{-2} \log\left[\frac{(\tau^2)^{-2}}{(\tau^2)^{-2} + \exp(\mathbf{x}_i \beta)}\right] + y_i \log\left[\frac{\exp(\mathbf{x}_i \beta)}{(\tau^2)^{-2} + \exp(\mathbf{x}_i \beta)}\right] \\ & + \log\left[\Gamma\left(y_i + \left(\tau^2\right)^{-2}\right) / \Gamma\left(\left(\tau^2\right)^{-2}\right)\right], \end{aligned} \tag{37}$$

where $\Gamma(\cdot)$ is the gamma function defined for $r > 0$ by $\Gamma(r) = \int_0^\infty z^{r-1} \exp(-z)\, dz$. For fixed τ^2, the log likelihood equation in (37) is in the exponential family; see Gourieroux et al. (1984). Thus the negative binomial QMLE using (37) is consistent under conditional mean assumption only, which is the same as the Poisson QMLE. Since the negative binomial II likelihood captures the nature of the variance function,

it should deliver more efficient estimation when the data generating process is correctly specified, although the spatial correlation is not accounted. Again, we can use a GEE working correlation matrix to account for the spatial correlation.

3.2 Example 2. Binary response data with spatial correlation in the latent error

The Probit model is one of the popular binary response models. The dependent variable y has conditional Bernoulli distribution and takes on the values zero and one, which indicates whether or not a certain event has occurred. For example, $y = 1$ if a firm adopts a new technology, and $y = 0$ otherwise. The value of the latent variable y^* determines the outcome of y.

Assume the Probit model is

$$y_i = 1\left[y_i^* > 0\right], \tag{38}$$

$$y_i^* = \mathbf{x}_i\beta + e_i. \tag{39}$$

We do not observe y_i^*; we only observe y_i. Let $\Phi(\cdot)$ be the standard normal cumulative density function (CDF), and ϕ be the standard normal probability density function (PDF). Assume that the mean function $m_i(\mathbf{x}_i;\beta) \equiv \mathsf{E}(y_i|\mathbf{x}_i, D_n) = \Phi(\mathbf{x}_i\beta)$ is correctly specified. e is the spatial correlated latent error. Let $\mathbf{e} = (e_1, e_2, ..., e_n)^\top$. For example, Pinkse and Slade (1998) use the following assumption of $\mathbf{e}$:

$$\mathbf{e} = \rho W\mathbf{e} + \varepsilon, \tag{40}$$

where $\varepsilon = (\varepsilon_1, \varepsilon_2, ..., \varepsilon_n)$ which has a standard normal distribution. W is a $n \times n$ weight matrix with zeroes on the diagonal and inverse of distances off diagonal. ρ is a correlation parameter. We can see e can be written as a function of ε,

$$\mathbf{e} = (I - \rho W)^{-1}\varepsilon. \tag{41}$$

Thus the conditional expectation of is zero. The variance covariance matrix of is

$$\mathrm{Var}(\mathbf{e}|\mathbf{x}, D_n) = (I - \rho W)^{-1}(I - \rho W)^{-1\top}. \tag{42}$$

If we assume that $e|x$ has a multivariate normal distribution with mean zero and variance matrix specified in (42). Thus a much simpler specification is to directly model

$e|x$ as a multivariate distribution. Different from the usual multivariate distribution[5], the covariances of e should depend on the pairwise distances d_{ij}. We also let the covariances depend on a parameter ρ. The above equation can be written in a conditional mean form:

$$\mathsf{E}\left(y_i|\mathbf{x}_i, D_n\right) = \Phi\left(\mathbf{x}_i\beta\right). \tag{43}$$

It is very natural to write the variance function for a Bernoulli distribution,

$$\mathrm{Var}\left(y_i|\mathbf{x}_i, D_n\right) = \Phi\left(\mathbf{x}_i\beta\right)\left[1 - \Phi\left(\mathbf{x}_i\beta\right)\right]. \tag{44}$$

We are interested in the partial effects of x to y. For a continuous x_K the partial effect is

$$\frac{\partial \mathsf{E}\left(y_i|\mathbf{x}_i, D_n\right)}{\partial x_K} = \Phi\left(\mathbf{x}_i\beta\right)\beta_K. \tag{45}$$

For a discrete x_K, the partial effects when x_K changes from a_K to $a_K + 1$ is

$$\Phi\left(\beta_1 + x_2\beta_2 + ... + \beta_K\left(a_K + 1\right)\right) - \Phi\left(\beta_1 + x_2\beta_2 + ... + \beta_K a_K\right). \tag{46}$$

A simple one-step estimation is the pooled Bernoulli quasi-MLE (QMLE), which is obtained by maximizing the pooled Probit log-likelihood. The log likelihood function for each observation is

$$l_i\left(\beta\right) = y_i \log\Phi\left(\mathbf{x}_i\beta\right) + \left(1 - y_i\right)\log\left[1 - \Phi\left(\mathbf{x}_i\beta\right)\right]. \tag{47}$$

Let $\breve{u}_i = y_i - \Phi\left(\mathbf{x}_i\breve{\beta}\right), i = 1, 2, ..., n$ be the residuals from the partial QMLE estimation. At this stage, a robust estimator for the asymptotic variance of $\breve{\beta}_{\mathrm{PQMLE}}$ can be computed as follows:

$$\begin{aligned}\widehat{\mathrm{Avar}}\left(\breve{\beta}_{\mathrm{PQMLE}}\right) = & \left(\sum_{i=1}^{n}\frac{\phi^2\left(\mathbf{x}_i\breve{\beta}_{\mathrm{PQMLE}}\right)\mathbf{x}_i^\top\mathbf{x}_i}{\Phi\left(\mathbf{x}_i\breve{\beta}\right)\left[1 - \Phi\left(\mathbf{x}_i\breve{\beta}_{\mathrm{PQMLE}}\right)\right]}\right)^{-1} \\ & \left(\sum_{i=1}^{n}\sum_{j\neq i}^{n}k\left(d_{ij}\right)\frac{\phi\left(\mathbf{x}_i\breve{\beta}_{\mathrm{PQMLE}}\right)\phi\left(\mathbf{x}_j\breve{\beta}_{\mathrm{PQMLE}}\right)\mathbf{x}_i^\top\breve{u}_i\breve{u}_j\mathbf{x}_j}{\Phi\left(\mathbf{x}_i\breve{\beta}_{\mathrm{PQMLE}}\right)\left[1 - \Phi\left(\mathbf{x}_i\breve{\beta}_{\mathrm{PQMLE}}\right)\right]}\right) \\ & \left(\sum_{i=1}^{n}\frac{\phi^2\left(\mathbf{x}_i\breve{\beta}_{\mathrm{PQMLE}}\right)\mathbf{x}_i^\top\mathbf{x}_i}{\Phi\left(\mathbf{x}_i\breve{\beta}_{\mathrm{PQMLE}}\right)\left[1 - \Phi\left(\mathbf{x}_i\breve{\beta}_{\mathrm{PQMLE}}\right)\right]}\right)^{-1},\end{aligned} \tag{48}$$

where $k\left(d_{ij}\right)$ is the kernel weight function that depends on pairwise distances. This

[5]A multivariate normal distribution usually specifies the mean vector and correlation matrix. The correlations do not depend on the pairwise distance between two variables.

partial QMLE and its robust variance-covariance estimator provides a legitimate way of the estimation of the spatial Probit model.

We use partial QMLE as a first-step estimator. An estimator for the working variance matrix for each group is

$$\check{v}_{gl} = \Phi\left(\mathbf{x}_{gl}\check{\beta}_{\text{PQMLE}}\right)\left[1 - \Phi\left(\mathbf{x}_{gl}\check{\beta}_{\text{PQMLE}}\right)\right]. \tag{49}$$

And assume the working correlation function for lth and mth elements in group g is

$$r_{glm} = \mathbf{C}\left(d_{glm}, \rho\right). \tag{50}$$

For example, suppose that

$$\mathbf{C}\left(d_{glm}, \rho\right) = \frac{\rho}{d_{glm}} \text{ or } \exp\left(-\frac{d_{glm}}{\rho}\right). \tag{51}$$

Let $\check{u}_i$ be the partial QMLE residual and $\hat{r}_i = \check{u}_i/\sqrt{\check{v}_i}$, for $i = 1, 2, ..., n$, be the standardized residuals. $\hat{\mathbf{C}}_{ij}$ equals the sample correlation of $\check{u}_i/\sqrt{\check{v}_i}$ and $\check{u}_j/\sqrt{\check{v}_j}$. Using the correlations within groups, one estimator of ρ is

$$\hat{\rho} = \mathbf{argmin}_\rho \sum_g \sum_{l=1}^{L} \sum_{m<l} \left[\hat{r}_{gl}\hat{r}_{gm} - C\left(d_{glm}, \rho\right)\right]^2, \tag{52}$$

for $l < m$.

The second-step GEE estimator for β is

$$\hat{\beta}_{\text{GEE}} = \mathbf{argmin}_\beta \sum_g \left(\mathbf{y}_g - \Phi\left(\mathbf{x}_g\beta\right)\right)^\top \hat{\mathbf{W}}_g^{-1}\left(\mathbf{y}_g - \Phi\left(\mathbf{x}_g\beta\right)\right). \tag{53}$$

The first order condition is

$$\sum_g \phi\left(\mathbf{x}_g\hat{\beta}_{\text{GEE}}\right)^\top \hat{\mathbf{W}}_g^{-1}\left(\mathbf{y}_g - \Phi\left(\mathbf{x}_g\hat{\beta}_{\text{GEE}}\right)\right) = \mathbf{0}. \tag{54}$$

$\hat{\beta}_{\text{GEE}}$ is consistent and follows a normal distribution asymptotically by Theorem 2. $\hat{\beta}$ is consistent even for misspecified spatial correlation structure $\hat{\mathbf{W}}_g$. The asymptotic

variance estimator that is robust to misspecification of spatial correlation is:

$$\begin{aligned}\widehat{\text{Avar}}\left(\hat{\beta}_{\text{GEE}}\right) &= \left(\sum_g \phi^2\left(\mathbf{x}_g\hat{\beta}_{\text{GEE}}\right)\mathbf{x}_g^\top\hat{\mathbf{W}}_g^{-1}\mathbf{x}_g\right)^{-1} \\ &\left(\sum_g\sum_{h(\neq g)} k(d_{gh})\phi\left(\mathbf{x}_g\hat{\beta}_{\text{GEE}}\right)\phi\left(\mathbf{x}_h\hat{\beta}_{\text{GEE}}\right)\mathbf{x}_g^\top\hat{\mathbf{W}}_g^{-1}\hat{\mathbf{u}}_g\hat{\mathbf{u}}_h^\top\hat{\mathbf{W}}_h^{-1}\mathbf{x}_h\right) \\ &\left(\sum_g \phi^2\left(\mathbf{x}_g\hat{\beta}_{\text{GEE}}\right)\mathbf{x}_g^\top\hat{\mathbf{W}}_g^{-1}\mathbf{x}_g\right)^{-1},\end{aligned} \tag{55}$$

where $k(d_{gh})$ is a kernel function which depends on the distances between groups.

An alternative approach is to specify the specific distributions of the multivariate normal distribution of the latent error, and then find the estimator for the spatial correlation parameter for the latent error within a MLE framework. For example, see Wang et al. (2013).

4 Theorems

In this section, we provide the assumptions and results on the theoretical properties our GEE estimation.

4.1 Consistency and Normality

A.2) $\{y_i\}$ is L_4- uniformly NED on the $\alpha-$ mixing random field $\varepsilon = \{\varepsilon_i, i \in D_n\}$, where $\varepsilon_i = (x_i, \epsilon_i)$($\epsilon_i$s are some underlying innovation processes). With the $\alpha-$ mixing coefficient $\overline{\alpha}(u, v, r) \leq (u+v)^\tau\hat{\alpha}(r)$, and $\hat{\alpha}(r) \to 0$ as $r \to \infty$. Assume that $\sum_{r=1}^{\infty} r^{d-1}\hat{\alpha}(r) < \infty$. The NED constant is $d_{n,i}$, ($\sup_{n,i\in T_n} d_{n,i} < \infty$) and the NED coefficient is $\psi(s)$ with $\psi(s) \to 0$, where recall that L is the group size, and $\sum_{r=0}^{\infty} r^{d-1}\psi(r) \to 0$.

Remark: See section 7.6 for a detailed verification of the special cases. It should be noted that by the Lyapunov inequality, if $\{y_i\}$ is L_k-NED, then it is also L_l-NED with the same coefficients $d_{n,i}$ and $\psi(s)$ for any $l \leq k$.

A.3) The parameter space $\mathbf{\Theta} \times \mathbf{\Gamma}$ is a compact subset on $\mathcal{R}^{p+q}$ with metric $\nu(.,.)$.

A.4) $q_g(\theta, \gamma)$, $(s_g(\theta, \gamma))$, $(h_g(\theta, \gamma))$ are $\mathbf{R}^{p_w} \times \Theta \times \Gamma \to \mathbf{R}^1(\mathbf{R}^p), (\mathbf{R}^{p^2})$ measurable for each $\theta \in \Theta, \gamma \in \Gamma$, and Lipschitz continuous on $\mathbf{\Theta} \times \Gamma$.

A.5) $\mathsf{E}\sup_{\theta\in\Theta}|m_{g,i}|^r \leq C_1$, $\mathsf{E}\sup_{\theta\in\Theta,\gamma\in\Gamma}|w_{g,i,j}|^r \leq C_2$, $\mathsf{E}|y_{g,i}|^r \leq C_3$ $\mathsf{E}\sup_{\theta\in\Theta}|\nabla_\theta m_{g,i}|^r \leq C_4$, where C_1, C_2, C_3, C_4 are constants, where $w_{g,i,j}, y_{g,i}, m_{g,i}$ is the elementwise component for $\mathbf{W}_g^{-1}(\theta,\gamma)$, $\mathbf{y}_g$, $\mathbf{m}_g(\theta,\gamma)$. $r > 4p''\vee 4p'$. $m_{g,i}, w_{g,i,j}$ are continuously differentiable up to the third order derivatives, and its rth moment (the supreme over the parameter space) is bounded up to the second order derivatives. Define $d_g = \max_{i\in B_g} d_{n,i}$, $M_G \stackrel{\text{def}}{=} \max_g d_g \vee c_{g,q} \vee c_{g,s} \vee c_{g,h}$. Also assume that $\sup_G \sup_g (c_{g,q} \vee c_{g,s} \vee c_{g,h})/d_g \leq C_5$, where C_5 is a constant.

Remark: Condition A.5) guarantees that there exists non random positive constants such that $c_{g,q}, c_{g,s}, c_{g,h}, g \in D_G, n \geq 1$ such that $\mathsf{E}|q_g/c_{g,q}|^{p''} < \infty$, $\mathsf{E}|s_g/c_{g,s}|_2^{p''} < \infty$, $\mathsf{E}|h_g/c_{g,h}|_1^{p''} < \infty$.

From now on we work with group level asymptotics. Define the field $\tilde{\varepsilon} = \{\varepsilon_g : g \in 1, \cdots, G\}$ with grouped observations. First of all suppose that D_n is divided by G blocks with $\cup_1^G B_g = D_n \subset T_n$, and the group level lattice is denoted as D_G. Define the distance between two groups g, h as $\rho(g,h) = \mathbf{min}_{i\in B_g, j\in B_h}\rho(i,j)$. And the $\alpha-$ mixing coefficient between two union of groups for $U = \{g_1, \cdots, g_L\}$, $V = \{h_1, \cdots, h_M\}$, $\rho(U,V) = \mathbf{min}_{l\in 1\cdots L, m\in 1,\cdots,M}\rho(g_l, h_m)$ is thus $\tilde{\alpha}(u,v,r) = \tilde{\alpha}(L \leq u, M \leq v, \rho(U,V) \geq r) = \sup_{L\leq u, M\leq v, \rho(U,V)\geq r}\alpha(\sigma(U),\sigma(V))$. If the group size are the same, i.e. L, then the mixing coefficients of the grouped observations have the following relationship with respect to it in the original field $\tilde{\alpha}(u,v,r) = \alpha(uL, vL, r)$. We can assume $\tilde{\alpha}(u,v,r) = (uL+vL)^\tau\hat{\alpha}(r)$.

Assume that $L^\tau\hat{\alpha}(r) \to 0$ as $r \to \infty$, and $\tilde{\varepsilon}$ would maintain the $\alpha-$ mixing property. Define the ball around group g with radius s to be $\mathcal{F}_g(s) = \sigma\{\cup_{h:\rho(g,h)\leq s} B_h\}$.

A.6) The $\alpha-$ mixing coefficients of the input field $\tilde{\varepsilon}$ satisfy $\tilde{\alpha}(u,v,r) \leq \phi(uL, vL)\hat{\alpha}(r)$, with $\phi(uL, vL) = (u+v)^\tau L^\tau$ and for some $\hat{\alpha}(r)$, $\sum_{r=1}^\infty L^\tau r^{d-1}\hat{\alpha}(r) < \infty$.

A.7) We assume moment conditions on the objects involved to prove the NED property of $\mathbf{H}_G(\theta,\gamma)$. $b_{ij} \stackrel{\text{def}}{=} e_i^\top(\mathbf{1}^\top\mathbf{W}_g(\theta,\gamma)\otimes I_g)|\partial\text{Vec}(\nabla\mathbf{m}_g(\theta))/\partial\theta|_a e_j$. $c_{ij} = e_i^\top(\mathbf{1}^\top \otimes \nabla\mathbf{m}_g^\top(\theta))|\partial\text{Vec}(\nabla\mathbf{m}_g(\theta))/\partial\theta|_a e_j$. $\|\sup_{\theta\in\Theta,\gamma\in\Gamma} b_{ij}\|$ and $\|\sup_{\theta\in\Theta,\gamma\in\Gamma} c_{ij}\|$ are finite.

A.8) (Identifiability)Let $\overline{Q}_G(\theta,\gamma) \stackrel{\text{def}}{=} \frac{1}{|M_G||D_G|}\sum_g \mathsf{E}(q_g(\theta,\gamma))$. Recall that $Q_\infty(\theta,\gamma) \stackrel{\text{def}}{=} \lim_{G\to\infty}\bar{Q}_G(\theta,\gamma)$. Assume that θ^0, γ^0 are identified unique in a sense that $\liminf_{G\to\infty}\mathbf{inf}_{\theta\in\Theta:\nu(\theta,\theta^0)\geq\varepsilon} Q_G(\theta,\gamma) > c_0 > 0$, for any γ and a positive constant c_0.

Remark A.8) can be implied from positive definiteness of $\mathbf{W}_g(\theta,\gamma)$ and the same identification assumption $\liminf_{G\to\infty}\inf_{\theta\in\Theta:\nu(\theta,\theta_0)\geq\varepsilon} Q'_G(\theta,\gamma) > c_0 > 0$ on $Q'_G(\theta,\gamma) \stackrel{\text{def}}{=}$

$\frac{1}{M_G|D_G|}\sum_{g\in|D_G|}\mathsf{E}\left[\mathbf{y}_g-\mathbf{m}_g\left(\mathbf{x}_g;\theta\right)\right]^\top\left[\mathbf{y}_g-\mathbf{m}_g\left(\mathbf{x}_g;\theta\right)\right]$. As it can be seen that with probability $1-\mathcal{o}_p(1)$
$\liminf_{G\to\infty}\inf_{\theta\in\Theta:\nu(\theta,\theta_0)\geq\varepsilon}Q_G\left(\theta,\gamma\right)>\liminf_{G\to\infty}\inf_{\theta\in\Theta:\nu(\theta,\theta_0)\geq\varepsilon}\lambda_{min}\{\mathbf{W}_g(\theta,\gamma)\}Q'_\infty(\theta,\gamma)$, where $\lambda_{min}\{\mathbf{W}_g(\theta,\gamma)\}$ is the minimum eigenvalue of the matrix $\lambda_{min}\{\mathbf{W}_g(\theta,\gamma)\}$. If we assume that with probability $1-\mathcal{o}_p(1)$, $\lambda_{min}\{\mathbf{W}_g(\theta,\gamma)\}>c$ where c is a positive constant. We now comment on assumptions, Condition A.2) is concerning the L_2 NED property of our data generating processes. A.3) and A.4) are the standard regularities assumptions. A.5) is a few moment assumptions on the statistical objects involved in the estimation. A.6) is the mixing coefficients restrictions after grouping observations. A.7) is again moment conditions on the elementwise Hessian matrices. A.8) is a condition on identification of our estimator. Given the assumptions, we can provide the consistency property of our estimation.

Theorem 1. *(Consistency) Under A.1)-A.8) the GEE-estimator in (4) is consistent, that is,* $\nu(\hat{\theta},\theta^0)\to_p 0$ *as* $G\to\infty$.

Theorem 1 indicates the consistency of the estimation as long as the number of groups tends to infinity. The proof is in the Appendix. To prove further the asymptotic normality of the estimation we need in addition the following assumptions.

A.9) The true point θ^0,γ^0 lies in the interior point of Θ,Γ. $\hat{\gamma}$ is estimated with $|\hat{\gamma}-\gamma^0|_2=\mathcal{o}_p(G^{-1/2})$.
Remark Verification of this assumption is in Proposition 1 and its proof in the Appendix.

A.10) $c'<\lambda_{min}(M_G^{-2}\,\mathsf{E}\left(\nabla\mathbf{m}_g^\top(\theta^0)\mathbf{W}_g^{-1}(\theta^0,\gamma^0)\nabla\mathbf{m}_g(\theta^0)\right))$
$<\lambda_{max}(M_G^{-2}\,\mathsf{E}\left(\nabla\mathbf{m}_g^\top(\theta^0)\mathbf{W}_g^{-1}(\theta^0,\gamma^0)\nabla\mathbf{m}_g(\theta^0)\right)<C'$ is positive definite, and c' and C' are two positive constants.

Define $\mathbf{u}_g=\mathbf{y}_g-\mathbf{m}_g(\theta^0)$ and $\hat{\mathbf{u}}_g=\mathbf{y}_g-\mathbf{m}_g(\hat{\theta})$

$$\mathbf{S}_G\left(\theta,\hat{\gamma}\right)=\frac{1}{M_G|D_G|}\sum_g\nabla\mathbf{m}_g^\top\left(\theta\right)\mathbf{W}_g^{-1}\left(\theta,\hat{\gamma}\right)\left[\mathbf{y}_g-\mathbf{m}_g\left(\theta\right)\right]. \tag{56}$$

Define

$$\begin{aligned}AS_G &= \frac{1}{G}\sum_g\mathsf{E}\left[\nabla\mathbf{m}_g^\top\left(\theta^0\right)\mathbf{W}_g^{-1}\left(\theta^0,\gamma^0\right)\mathbf{u}_g\mathbf{u}_g^\top\mathbf{W}_g^{-1}\left(\theta^0,\gamma^0\right)\nabla\mathbf{m}_g\left(\theta^0\right)\right] \\ &+\frac{1}{G}\sum_g\sum_{h,h\neq g}\mathsf{E}\left[\nabla\mathbf{m}_g^\top\left(\theta^0\right)\mathbf{W}_g^{-1}\left(\theta^0,\gamma^0\right)\mathbf{u}_g\mathbf{u}_h^\top\mathbf{W}_h^{-1}\left(\theta^0,\gamma^0\right)\nabla\mathbf{m}_h\left(\theta^0\right)\right],\end{aligned} \tag{57}$$

and $AS_\infty = \lim_{G\to\infty} AS_G$.

A.11) $\mathbf{S}_G\left(\hat{\gamma}, \hat{\theta}\right) = \mathcal{O}_p(1)$. $\inf_G |D_G|^{-1} M_G^{-2} \lambda_{min}(\mathbf{AS}_\infty) > 0$, where $\mathbf{AS}_\infty$ is defined in equation (57). The mixing coefficients satisfy $\sum_{r=1}^{\infty} r^{(d\tau^*+d)-1} L^{\tau^*} \hat{\alpha}^{\delta/(2+\delta)}(r) < \infty$. $(\tau^* = \delta\tau/(4+2\delta))$.

A.9) is concerning the the pre-estimation of the nuisance parameter γ, and A.10), A.11) are two standard assumptions on the regularities of the estimation. Note that $\mathbf{S}_G\left(\hat{\theta}, \hat{\gamma}\right) = \mathcal{O}_p(1) = 0$ if $\hat{\theta}, \hat{\gamma}$ lies in the interior point of the parameter space. In the following, we verify that with our proposal of estimating $\hat{\gamma}$ in (17) in Section 2 , we will achieve A.9).

Proposition 1. *Under A.1)-A.3), A.5), A.6) and A.8)', A.9)', A.11)', (A.8)', A.9)', A.11)'are defined in the Appendix), the estimator solving equation (17) satisfies,*

$$|\hat{\gamma} - \gamma^0|_2 = \mathcal{O}_p(1/\sqrt{G}). \tag{58}$$

$\mathbf{H}_\infty \stackrel{\text{def}}{=} \lim_{G\to\infty} \mathsf{E}\, \mathbf{H}_G(\theta^0, \gamma^0)$, where $\mathbf{H}_G(\theta^0, \gamma^0)$ is defined in equation (11). It is not surprising to see that our estimation will be asymptotically normally distributed, with a variance covariance matrix of a sandwich form $AV(\theta^0)$, which involves the Hessian. The rate of convergence is shown to be $\sqrt{G}$.

Theorem 2. *Under A.1) - A.11), we have* $AV(\theta^0) \stackrel{\text{def}}{=} \mathbf{H}_\infty^\top \mathbf{AS}_\infty \mathbf{H}_\infty$.

$$\sqrt{G} AV(\theta^0)^{-1/2} (\hat{\theta} - \theta^0) \Rightarrow \mathbb{N}(0, I_p). \tag{59}$$

4.2 Consistency of variance covariance matrix estimation

In this subsection, we propose a semiparametric estimator of the asymptotic variance in Theorem 2, and prove its consistency. The estimation is tailored to account for the spatial dependency of the underlying process. This facilitates us to create a confidence interval for our estimation.

First let

$$\hat{\mathbf{A}} = \frac{1}{|D_G|} \sum_g \nabla \hat{\mathbf{m}}_g^\top \hat{\mathbf{W}}_g^{-1} \nabla \hat{\mathbf{m}}_g, \tag{60}$$

$$\hat{\mathbf{B}} = \frac{1}{|D_G|} \sum_g \sum_{h\neq g} k(d_{gh}) \nabla \hat{\mathbf{m}}_g^\top \hat{\mathbf{W}}_g^{-1} \hat{\mathbf{u}}_g \hat{\mathbf{u}}_h^\top \hat{\mathbf{W}}_h^{-1} \nabla \hat{\mathbf{m}}_h^\top, \tag{61}$$

where $\nabla \hat{\mathbf{m}}_g \equiv \nabla \hat{\mathbf{m}}_g\left(\hat{\theta}\right)$, $\hat{\mathbf{W}}_g \equiv \hat{\mathbf{W}}_g(\hat{\gamma}, \hat{\theta})$.

The estimator of AV (θ^0) which is robust to misspecification of the variance covariance matrix is

$$\begin{aligned}\widehat{\mathrm{AV}}\left(\hat{\theta}\right) &= |D_G|\left(\sum_g \nabla\hat{\mathbf{m}}_g^\top \hat{\mathbf{W}}_g^{-1}\nabla\hat{\mathbf{m}}_g\right)^{-1} \\ &\quad \left(\sum_g \sum_{h(\neq g)} \nabla\hat{\mathbf{m}}_g^\top \hat{\mathbf{W}}_g^{-1} k\left(d_{gh}\right)\hat{\mathbf{u}}_g\hat{\mathbf{u}}_h^\top \hat{\mathbf{W}}_h^{-1}\nabla\hat{\mathbf{m}}_h\right) \\ &\quad \left(\sum_g \nabla\hat{\mathbf{m}}_g^\top \hat{\mathbf{W}}_g^{-1}\nabla\hat{\mathbf{m}}_g\right)^{-1}, \qquad (62) \\ &= \hat{\mathbf{A}}^{-1}\hat{\mathbf{B}}\hat{\mathbf{A}}^{-1} \qquad (63)\end{aligned}$$

where $k\left(d_{gh}\right)$ is the kernel function depending on the distance between group g and h, i.e. $\rho(g,h)$, and a bandwidth parameter h_g. As noted in Kelejian and Prucha (2007), there are many choices for the kernel functions, such as rectangular kernel, Bartlett or triangular kernel, etc. In particular, without loss of generality, we can choose the Bartlett kernel function $k\left(d_{gh}\right) = 1-\rho(g,h)/h_g$, for $\rho(g,h) < h_g$ and $k\left(g,h\right) = 0$ for $\rho(g,h) \geq h_g$. Further, we can obtain the average partial effects (APE) of interest and carry on valid inference.

We now list the assumptions needed for the consistency of estimator of $AV(\theta^0)$.

B.1) $\hat{\mathbf{u}}_g - \mathbf{u}_g = C_g\Delta_g$, where C_g is a $L\times p$, and Δ_g is a $p\times 1$ dimensional vector, with the condition that $|C_g|_2 = \mathcal{O}_p(1)$, and $|\Delta_g|_2 = \mathcal{O}_p((pG)^{-1/2})$.

B.2) The moment is bounded by a constant $\max_{h:\rho(h,g)\leq h_g}\mathsf{E}\left|Z_h\right|^{q'} \leq ML^2$, $q'\geq 1$, and M is a constant, where $Z_h \overset{\text{def}}{=} \nabla\mathbf{m}_h^\top(\theta^0)\mathbf{W}_h^{-1}(\theta^0,\gamma^0)\mathbf{u}_h$.

B.3) $|k(d_{gh})-1| \leq C_k|d_{gh}/h_g|^{\rho_K}$ for $d_{gh}\leq 1$ for some constant $\rho_k \geq 1$ and $0 < C_k < \infty$ $M_G^{-2}|D_G|^{-1}\sum_g\sum_h |\rho(g,h)/h_g|^{\rho_k}\|e_i^\top Z_g^\top\|\|Z_h e_j\| = o(1)$.

B.4) Assume that $h_g^{d/q'}|D_G|^{-1}L^{d/q'}L^2 = o(1)$, $h_g^{2d}L^{2d}\sum_{r=1}^\infty r^{(d\tau^*+d)-1}\hat{\alpha}^{\delta/(2+\delta)}(r) = \mathcal{O}(G)$, and $h_g^{2d}\sum_{r=1}^\infty L^{2d}r^{d-1}\psi((r-h_g)_+) = \mathcal{O}(G)$, $((r-h_g)_+ = \max(r-h_g,0))$ where δ is a constant and $\delta^* = \delta\tau/(2+\delta)$.

B.1) is an assumption for decomposing the difference between the residuals and the true error, as in Kelejian and Prucha (2007). B.2) is about the moment bound and B.3) is on property of the kernel function. B.4) constrains on the spatial dependence coefficients and the bandwidth length. We provide in the following theorem the consistency of the $\widehat{\mathrm{AV}}\left(\hat{\theta}\right)$. It is worth noting that we prove an elementwise version of the

consistency, and the results below can be verified equivalently in any matrix norm, as we consider fixed dimension parameter.

Theorem 3. *Under assumption B.1)- B.4) and A.1) - A.8). The variance-covariance estimator in (62) is consistent.* $\widehat{\mathrm{AV}}\left(\hat{\theta}\right) \rightarrow_p \mathrm{AV}\left(\theta^0\right)$.

5 Monte Carlo Simulations

In this section, we use Monte Carlo simulations to investigate the finite sample performances of our proposed GEE approach with groupwise data compared to the partial QMLE. We simulated count data and binary response data separately. We show that our GEE method is very critical for improving the efficiency of our estimation.The simulation mechanism is described as follows.

5.1 Sampling Space

We use sample sizes of 400 or 1600. We sample observations on a lattice. For example, for sample size of 400, the sample space is a 20×20 square lattice. Each observation resides on the intersections of this lattice. The locations for the data are $\{(r, s) : r, s = 1, 2, ..., 20\}$. The distance d_{ij} between location i and j is chosen to be the Euclidean distance. Suppose $A(a_i, a_j)$ and $B(b_i, b_j)$ are the two points on the lattice; their distance d_{ij} is $\sqrt{(a_i - b_i)^2 + (a_j - b_j)^2}$. The spatial correlation is based on a given parameter ρ and d_{ij}. The data are divided into groups of 4 and the number of groups are set to be 100 for sample size 400. Similarly, for the sample size of 1600, we use a 40×40 lattice. We still use sample size of 4 in each group and there are 400 groups in total. For simplicity, we keep the pairwise distances in different groups the same.

5.2 Count data

5.2.1 Data generating process

In the count data case, for a Poisson distribution the variances and covariances of the count dependent variable can be written in closed forms given the spatial correlation in the underlying spatial error term. That is, by knowing the correlations in the spatial error term, we can derive the correlations in the count dependent variable as shown in (25) and (26). Consider the following spatial count data generating process: 1. v_i is simulated as a multivariate lognormal variable with $\mathrm{E}(v_i) = 1$, exponentiating an underlying multivariate normal distribution using with correlation matrix W. Let a_i

be the underlying multivariate normal distributed variable. Then $v_i = \exp(a_i)$ follows a multivariate lognormal distribution. We describe the underlying spatial process in Case 1, 2, and 3 as three special cases to demonstrate different spatial correlations. 2. The coefficient parameters and explanatory variables are set as follows: $\beta_1 = 0.5, \beta_2 = 1, \beta_3 = 1, \beta_4 = 1.$; $x_2 \sim \mathrm{N}(0, 0.25), x_3 \sim \text{Uniform}(0,1), x_5 \sim \mathrm{N}(0,1), x_4 = 1[x_5 > 0]$. 3. The mean function for individual i is $m_i = v_i \exp(\beta_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4)$; 4. Finally we draw the dependent variable from the Poisson distribution with mean m_i: $y_i \sim \text{Poisson}(m_i)$. Specifically, the underlying spatial error a_i has the following three cases.

Case 1. $a_i = (I - \rho W)^{-1} e_i, e_i \sim \mathrm{N}(0,1)$; W is the matrix with W_g on the diagonal, $g = 1, 2, ..., G$. Other elements in W are equal to zero. For group size equal to four,

$$W_g = \frac{1}{3}\begin{pmatrix} 0 & 1 & 1 & 1 \\ 1 & 0 & 1 & 1 \\ 1 & 1 & 0 & 1 \\ 1 & 1 & 1 & 0 \end{pmatrix}. \tag{64}$$

Case 2. $a_i = (I - \rho W)^{-1} e_i, e_i \sim \mathrm{N}(0,1)$; W is the matrix with W_g on the diagonal, $g = 1, 2, ..., G$. The (l, m)th element in W_g, $W_{g\_lm} = \frac{\rho}{6 * d_{g\_lm}}$, $\rho = 0, 0.5, 1, 1.5, l \neq m$; $W_{g\_lm} = 0, l = m$ for group g. Correlations are zero if observations are in different groups. For group size equal to four,

$$W_g = \frac{1}{6}\begin{pmatrix} 0 & \frac{\rho}{d_{g\_12}} & \frac{\rho}{d_{g\_13}} & \frac{\rho}{d_{g\_14}} \\ \frac{\rho}{d_{g\_21}} & 0 & \frac{\rho}{d_{g\_23}} & \frac{\rho}{d_{g\_24}} \\ \frac{\rho}{d_{g\_31}} & \frac{\rho}{d_{g\_32}} & 0 & \frac{\rho}{d_{g\_34}} \\ \frac{\rho}{d_{g\_41}} & \frac{\rho}{d_{g\_42}} & \frac{\rho}{d_{g\_43}} & 0 \end{pmatrix}. \tag{65}$$

Case 3. In this case, the DGP has the following differences from Case 1 and Case 2. a_i is simulated as a multivariate lognormal variable by exponentiating an underlying multivariate normal distribution $\mathrm{N}\left(-\frac{1}{2}, 1\right)$ using with correlation matrix W. $W_{ij} = \frac{\rho}{d_{ij}}$, $\rho = 0, 0.2, 0.4, 0.6, i \neq j$; $W_{ii} = 1$; $i, j = 1, 2, ..., N$. The underlying normal distribution implies that v_i follows a multivariate lognormal distribution with $\mathrm{E}(v_i) = 1$. We set $\beta_1 = -1, \beta_2 = 1, \beta_3 = 1, \beta_4 = 1$. x_2 follows a multivariate normal distribution $\mathrm{N}(0, W)$; In this case, the data has general spatial correlations for each pair of observations if

$\rho \neq 0$.

$$\mathbf{W} = \begin{pmatrix} 1 & \frac{\rho}{d_{12}} & \frac{\rho}{d_{13}} & \cdots & \frac{\rho}{d_{1N}} \\ \frac{\rho}{d_{21}} & 1 & & \vdots & \frac{\rho}{d_{2N}} \\ \frac{\rho}{d_{31}} & & 1 & & \vdots \\ \vdots & ... & & \ddots & \frac{\rho}{d_{N-1,N}} \\ \frac{\rho}{d_{N1}} & \frac{\rho}{d_{N2}} & \cdots & \frac{\rho}{d_{N,N-1}} & 1 \end{pmatrix} \tag{66}$$

5.2.2 Simulation results

Table 1, Table 2 and Table 3 show three cases of simulation results with 1000 replications with two different samples and group sizes: (1) $N = 400$, $G = 100$, $L = 4$ (2) $N = 1600$, $G = 400$, $L = 4$. There are four estimators, Poisson partial QMLE estimator, Poisson GEE, Negative Binomial II (NB II) partial QMLE, and NB II GEE. For simplicity, we use an exchangeable working correlation matrix for GEE estimators. We can see that, first as spatial correlation increases the GEE methods has smaller standard deviations than QMLE. Second, when there is little spatial correlation, GEE does not increase much finite sample bias due to accounting for possible spatial correlation.

In Case 1, when there is no spatial correlation, the Poisson QMLE should be as efficient as GEE asymptotically. We can see that when $\rho = 0$, the coefficient estimates and their standard deviations of Poisson QMLE and GEE are pretty close, which means that there is little finite sample bias due to accounting for possible spatial correlation when there is actually no spatial correlation. The standard deviations for the estimated coefficients of Poisson QMLE and GEE are almost the same. The standard deviation of $\hat{\beta}_2$ equals 0.259 for Poisson QMLE and 0.260 for Poisson GEE when $\rho = 0$ for a sample size of 400. As ρ grows larger. the GEE estimator shows more and more efficiency improvement over the partial QMLE. For example, for a sample size of 400, when $\rho = 1$, the standard deviation of $\hat{\beta}_2$ equals 0.267 for Poisson QMLE and 0.259 for Poisson GEE. When $\rho = 1.5$, the standard deviation of $\hat{\beta}_2$ equals 0.320 for Poisson GEE and 0.302 for Poisson PQMLE. The NB II GEE also has some improvement over NB II PQMLE. When $\rho = 1$, the standard deviation of $\hat{\beta}_2$ equals 0.234 for NB II PQMLE and 0.226 for NB II GEE. When $\rho = 1.5$, the standard deviation of $\hat{\beta}_2$ equals 0.276 for NB II PQMLE and 0.261 for NB II GEE. When sample size increases from 400 to 1600, we see the similar scenarios. Case 2 and Case 3 have shown similar efficiency

Table 1: Means and Standard Deviations for Count Case 1, averaged over 1000 samples.

		N=400,G=100,L=4				N=1600,G=400,L=4			
		Poisson	GEE-poisson	NB II	GEE-nb2	Poisson	GEE-poisson	NB II	GEE-nb2
$\rho = 0$	$\hat{\beta}_2$	1.000	0.999	1.002	1.002	0.994	0.994	0.997	0.997
	s.d.$\left(\hat{\beta}_2\right)$	**0.259**	0.260	**0.227**	0.228	**0.160**	**0.160**	**0.136**	**0.136**
	$\hat{\beta}_3$	1.000	0.999	1.002	1.002	0.999	1.000	0.998	0.998
	s.d.$\left(\hat{\beta}_3\right)$	**0.259**	0.260	**0.227**	0.228	**0.137**	**0.137**	**0.121**	**0.121**
	$\hat{\beta}_4$	0.998	0.998	0.996	0.996	1.003	1.003	1.003	1.003
	s.d.$\left(\hat{\beta}_4\right)$	**0.146**	0.147	**0.137**	**0.137**	**0.071**	**0.071**	**0.067**	**0.067**
$\rho = 0.5$	$\hat{\beta}_2$	0.985	0.985	0.993	0.994	1.000	1.000	1.000	0.999
	s.d.$\left(\hat{\beta}_2\right)$	0.256	**0.255**	0.216	**0.215**	**0.127**	**0.127**	0.110	**0.109**
	$\hat{\beta}_3$	1.006	1.006	1.004	1.005	1.005	1.004	1.003	1.003
	s.d.$\left(\hat{\beta}_3\right)$	0.211	**0.210**	0.180	**0.179**	**0.106**	**0.106**	**0.092**	**0.092**
	$\hat{\beta}_4$	1.002	1.002	1.003	1.003	1.003	1.003	1.002	1.002
	s.d.$\left(\hat{\beta}_4\right)$	**0.117**	**0.117**	0.111	**0.110**	**0.058**	**0.058**	**0.054**	**0.054**
$\rho = 1$	$\hat{\beta}_2$	0.987	0.988	0.991	0.991	0.998	0.997	0.997	0.997
	s.d.$\left(\hat{\beta}_2\right)$	0.267	**0.259**	0.234	**0.226**	0.130	**0.127**	0.130	**0.128**
	$\hat{\beta}_3$	1.003	1.003	1.004	1.004	1.000	0.999	1.000	0.999
	s.d.$\left(\hat{\beta}_3\right)$	0.220	**0.214**	0.195	**0.190**	0.105	**0.102**	0.094	**0.091**
	$\hat{\beta}_4$	0.995	0.996	0.996	0.997	1.000	1.000	1.000	1.000
	s.d.$\left(\hat{\beta}_4\right)$	0.120	**0.119**	0.113	**0.111**	0.060	**0.058**	0.056	**0.054**
$\rho = 1.5$	$\hat{\beta}_2$	0.980	0.982	0.995	0.998	0.988	0.988	0.995	0.997
	s.d.$\left(\hat{\beta}_2\right)$	0.320	**0.302**	0.276	**0.261**	0.183	**0.173**	0.154	**0.145**
	$\hat{\beta}_3$	0.997	0.995	0.992	0.992	0.992	0.994	0.997	0.999
	s.d.$\left(\hat{\beta}_3\right)$	0.288	**0.271**	0.250	**0.234**	0.143	**0.136**	0.126	**0.120**
	$\hat{\beta}_4$	0.997	0.999	1.000	0.998	1.002	1.001	1.003	1.003
	s.d.$\left(\hat{\beta}_4\right)$	0.146	**0.139**	0.139	**0.131**	0.077	**0.072**	0.073	**0.068**

Note: The estimates with smaller standard deviations are marked with bold.

results for the GEE estimators.

5.3 Binary response data

5.3.1 Data generating process

For the Probit model, the correlations of latent normal errors result in correlations of binary response variables, but we cannot easily find the specific form of the conditional variances and covariances for the binary dependent variables. The correlations in latent error do not reflect the exact correlations in the binary dependent variables. Consider the following cases of data generating process. 1. The latent variable $y^* = \beta_1+\beta_2 x_2+\beta_3 x_3+\beta_4 x_4+e_4$, where e_4 is the latent spatial error term, and the parameters are set to be $\beta_1 = \beta_2 = \beta_3 = \beta_4 = 1$. Then the binary dependent variable is generated as $y_i = 1$ if $y_i^* \geq 1.5$ and $y_i = 0$ if $y_i^* < 1.5$. The explanatory variables are set as follows: $x_1 = 1$; $x_2 \sim \mathrm{N}(1,1)$; $x_3 = 0.2x_2 - 1.2e_1, e_1 \sim \mathrm{N}(0,1)$; $x_5 = 0.2x_2 + 0.2x_3 + e_2, e_2 \sim \mathrm{N}(0,1)$; $x_4 = 1[x_5 > 0]$. We consider two cases of latent spatial error terms and the corresponding binary response variables are generated as follows.

Case 1. The vector of spatial error $\mathbf{e}_4 = (I - \rho W)^{-1}\mathbf{e}_3, e_3 \sim \mathrm{N}(0,1)$, where

Table 2: Means and Standard Deviations for Count Case 2, averaged over 1000 samples

		N=400,G=100,L=4				N=1600,G=400,L=4			
		Poisson	GEE-poisson	NB II	GEE-nb2	Poisson	GEE-poisson	NB II	GEE-nb2
$\rho = 0$	$\hat{\beta}_2$	0.990	0.9990	0.992	0.992	0.994	0.994	0.999	0.999
	s.d.$\left(\hat{\beta}_2\right)$	**0.322**	0.323	**0.267**	0.268	**0.162**	**0.162**	**0.139**	**0.139**
	$\hat{\beta}_3$	0.986	0.987	0.992	0.992	0.997	0.997	0.999	0.999
	s.d.$\left(\hat{\beta}_3\right)$	**0.281**	**0.281**	**0.244**	**0.244**	**0.137**	**0.137**	**0.119**	**0.119**
	$\hat{\beta}_4$	0.999	1.000	0.998	0.998	0.998	0.998	0.998	0.998
	s.d.$\left(\hat{\beta}_4\right)$	**0.140**	0.141	**0.133**	**0.133**	**0.076**	**0.076**	**0.071**	**0.071**
$\rho = 0.5$	$\hat{\beta}_2$	0.972	0.971	0.981	0.980	1.000	1.000	1.001	1.002
	s.d.$\left(\hat{\beta}_2\right)$	**0.330**	0.331	**0.285**	0.286	**0.164**	0.165	**0.136**	**0.136**
	$\hat{\beta}_3$	0.992	0.991	0.995	0.994	0.999	0.999	0.999	0.999
	s.d.$\left(\hat{\beta}_3\right)$	**0.276**	**0.276**	**0.243**	**0.243**	**0.141**	**0.141**	**0.120**	**0.120**
	$\hat{\beta}_4$	0.995	0.995	0.996	0.995	0.998	0.998	0.998	0.998
	s.d.$\left(\hat{\beta}_4\right)$	**0.151**	**0.151**	0.142	**0.141**	**0.077**	**0.077**	**0.073**	**0.073**
$\rho = 1$	$\hat{\beta}_2$	1.017	1.014	1.016	1.014	0.998	0.997	0.998	0.997
	s.d.$\left(\hat{\beta}_2\right)$	0.400	**0.396**	0.319	**0.316**	0.193	**0.191**	0.161	**0.159**
	$\hat{\beta}_3$	0.975	0.976	0.978	0.979	1.005	1.004	1.004	1.003
	s.d.$\left(\hat{\beta}_3\right)$	**0.331**	**0.331**	0.278	**0.276**	0.158	**0.157**	0.135	**0.134**
	$\hat{\beta}_4$	0.998	0.996	0.995	0.994	1.000	1.000	1.000	1.000
	s.d.$\left(\hat{\beta}_4\right)$	0.185	0.182	0.173	0.169	0.088	0.087	0.083	0.081
$\rho = 1.5$	$\hat{\beta}_2$	0.970	0.973	1.013	1.015	1.004	1.001	1.008	1.004
	s.d.$\left(\hat{\beta}_2\right)$	0.677	**0.662**	0.577	**0.570**	0.311	**0.302**	0.262	**0.255**
	$\hat{\beta}_3$	0.972	0.972	0.974	0.976	0.999	0.997	1.001	1.000
	s.d.$\left(\hat{\beta}_3\right)$	0.627	**0.611**	0.524	**0.504**	0.293	**0.286**	0.239	**0.233**
	$\hat{\beta}_4$	1.002	1.000	1.000	0.998	0.999	1.000	0.999	1.000
	s.d.$\left(\hat{\beta}_4\right)$	0.326	**0.318**	0.293	**0.284**	0.160	**0.156**	0.144	**0.141**

Note: The estimates with smaller standard deviations are marked with bold.

Table 3: Means and Standard Deviations for Count Case 3, averaged over 1000 samples

		N=400, G=100, L=4				N=1600, G=400, L=4			
		Poisson	GEE-poisson	NB II	GEE-nb2	Poisson	GEE-poisson	NB II	GEE-nb2
$\rho = 0$	$\hat{\beta}_2$	0.998	0.998	1.000	1.000	0.998	0.998	0.999	0.999
	s.d.$\left(\hat{\beta}_2\right)$	**0.330**	**0.330**	**0.266**	0.267	**0.165**	**0.165**	**0.144**	**0.144**
	$\hat{\beta}_3$	1.002	1.002	1.002	1.002	0.995	0.995	0.995	0.995
	s.d.$\left(\hat{\beta}_3\right)$	**0.273**	0.274	**0.240**	0.241	**0.138**	**0.138**	**0.126**	**0.126**
	$\hat{\beta}_4$	0.998	0.999	0.998	0.998	0.997	0.997	0.997	0.997
	s.d.$\left(\hat{\beta}_4\right)$	**0.152**	0.153	**0.142**	0.143	**0.073**	**0.073**	**0.069**	**0.069**
$\rho = 0.2$	$\hat{\beta}_2$	0.991	0.911	0.997	0.996	0.998	0.999	0.999	1.000
	s.d.$\left(\hat{\beta}_2\right)$	**0.312**	**0.312**	0.272	**0.271**	0.158	**0.157**	**0.137**	**0.137**
	$\hat{\beta}_3$	0.991	0.991	0.996	0.996	1.000	1.000	0.999	0.999
	s.d.$\left(\hat{\beta}_3\right)$	**0.265**	0.266	**0.234**	0.235	**0.130**	**0.130**	**0.116**	**0.116**
	$\hat{\beta}_4$	1.002	1.002	1.003	1.004	0.999	0.999	0.998	0.998
	s.d.$\left(\hat{\beta}_4\right)$	**0.143**	**0.143**	**0.137**	**0.137**	**0.073**	**0.073**	**0.069**	**0.069**
$\rho = 0.4$	$\hat{\beta}_2$	0.988	0.989	0.992	0.993	0.999	0.999	0.997	0.998
	s.d.$\left(\hat{\beta}_2\right)$	0.305	**0.303**	0.261	**0.260**	0.162	**0.160**	0.140	**0.138**
	$\hat{\beta}_3$	1.002	1.003	1.006	1.006	0.997	0.996	0.999	0.998
	s.d.$\left(\hat{\beta}_3\right)$	0.267	**0.265**	0.238	**0.237**	0.129	**0.128**	0.117	**0.116**
	$\hat{\beta}_4$	0.998	0.998	0.997	0.998	1.003	1.003	1.003	1.003
	s.d.$\left(\hat{\beta}_4\right)$	0.139	**0.138**	0.131	**0.130**	0.074	**0.073**	0.070	**0.069**
$\rho = 0.6$	$\hat{\beta}_2$	0.995	0.995	0.999	0.999	1.005	1.006	1.003	1.004
	s.d.$\left(\hat{\beta}_2\right)$	0.300	**0.292**	0.260	**0.251**	0.161	**0.156**	0.135	**0.130**
	$\hat{\beta}_3$	1.004	1.003	1.002	1.000	1.002	1.001	1.006	1.005
	s.d.$\left(\hat{\beta}_3\right)$	0.253	**0.247**	0.219	**0.213**	0.131	**0.128**	0.114	**0.110**
	$\hat{\beta}_4$	1.004	1.003	1.003	1.002	1.000	1.000	1.000	1.000
	s.d.$\left(\hat{\beta}_4\right)$	0.140	**0.138**	0.132	**0.130**	0.072	**0.071**	**0.068**	**0.068**

Note: The estimates with smaller standard deviations are marked with bold.

$\rho = 0, 0.5, 1, 1.5$ respectively. W is the matrix with W_g on the diagonal, $g = 1, 2, ..., G$. Other elements in W are equal to zero. In this case, only individuals within a group are correlated. For group size equal to four, W_g is the same as in (64) in Case 1 for count data.

Case 2. The latent spatial error $e_4 \sim$ MVN(0,W), that is, e_4 follows a standard multivariate normal distribution with expectation zero and W is $N \times N$ correlation matrix. $W_{ij} = \frac{\rho}{d_{ij}}$, $\rho = 0, 0.2, 0.4, 0.6, i \neq j$; $W_{ii} = 1$; $i, j = 1, 2, ..., N$. W is the same as in (66) in Case 3 for count data. Therefore, the data has general spatial correlations for each pair of observations if $\rho \neq 0$.

5.3.2 Simulation results

In the simulation, two estimators are compared, the Probit partial QMLE estimator, and the Probit GEE estimator with an exchangeable working correlation matrix. We show two cases of the simulation: (1) N=400, G=100, L=4; 2) N=1600, G=400, L=4. The replication times are 1000. The simulation results for Case 1 and Case 2 are in Table 4 and Table 5 separately. We find the following results.

First, in both cases, the GEE estimator is less biased than the partial QMLE estimator. For example, for N=400, in Case 1 when $\rho = 1, \hat{\beta}_2$ equals 1.252 for QMLE and 1.203 for GEE. In Case 2 when $\rho = 0.6, \hat{\beta}_2$ equals 1.148 for QMLE and 1.090 for GEE. Second, the GEE estimator has some obvious efficiency improvement over partial QMLE. For example, in case 1 when $\rho = 1$, the standard deviation of $\hat{\beta}_2$ equals 0.280 for QMLE and 0.173 for GEE. In Case 2 when $\rho = 0.6$, the standard deviation of $\hat{\beta}_2$ equals 0.270 for QMLE and 0.164 for GEE for a sample size of 400. Third, when we increase the sample size to 1600 and number of groups to 400 correspondingly, the same scenario applies. What is more, the bias and especially standard deviations for both the Probit QMLE and GEE reduces. For example, for N=1600, in Case 1 when $\rho = 1$, the standard deviations of $\hat{\beta}_2$ reduce to 0.121 for QMLE and 0.081 for GEE.

6 An empirical application of the inflow FDI to China

In the empirical FDI literature, the gravity equation specification was initially adopted from the empirical literature on trade flows. The gravity equation has been widely used and extended in international trade since Tinbergen (1962). Anderson and Van Wincoop

Table 4: Means and Standard Deviations for Probit Case 1, averaged over 1000 samples

		N=400, G=100, L=4		N=1600, G=400, L=4	
		Probit	GEE-probit	Probit	GEE-probit
$\rho = 0$	$\hat{\beta}_2$	1.076	1.033	1.016	1.007
	s.d.$\left(\hat{\beta}_2\right)$	0.230	**0.142**	0.103	**0.069**
	$\hat{\beta}_3$	1.070	1.031	1.016	1.018
	s.d.$\left(\hat{\beta}_3\right)$	0.205	**0.127**	0.084	**0.059**
	$\hat{\beta}_4$	1.069	1.021	1.019	1.011
	s.d.$\left(\hat{\beta}_4\right)$	0.304	**0.200**	0.136	**0.103**
$\rho = 0.5$	$\hat{\beta}_2$	1.310	1.252	1.229	1.213
	s.d.$\left(\hat{\beta}_2\right)$	0.293	**0.169**	0.124	**0.077**
	$\hat{\beta}_3$	1.310	1.256	1.229	1.214
	s.d.$\left(\hat{\beta}_3\right)$	0.259	**0.156**	0.111	**0.720**
	$\hat{\beta}_4$	1.297	1.243	1.227	1.213
	s.d.$\left(\hat{\beta}_4\right)$	0.364	**0.238**	0.165	**0.112**
$\rho = 1$	$\hat{\beta}_2$	1.254	1.203	1.180	1.167
	s.d.$\left(\hat{\beta}_2\right)$	0.280	**0.173**	0.121	**0.081**
	$\hat{\beta}_3$	1.236	1.192	1.176	1.164
	s.d.$\left(\hat{\beta}_3\right)$	0.236	**0.152**	0.105	**0.072**
	$\hat{\beta}_4$	1.238	1.196	1.175	1.165
	s.d.$\left(\hat{\beta}_4\right)$	0.356	**0.241**	0.156	**0.109**
$\rho = 1.5$	$\hat{\beta}_2$	1.022	0.982	0.963	0.949
	s.d.$\left(\hat{\beta}_2\right)$	0.230	**0.149**	0.102	**0.070**
	$\hat{\beta}_3$	1.013	0.979	0.966	0.953
	s.d.$\left(\hat{\beta}_3\right)$	0.196	**0.132**	0.086	**0.063**
	$\hat{\beta}_4$	1.003	0.963	0.968	0.953
	s.d.$\left(\hat{\beta}_4\right)$	0.309	**0.209**	0.139	**0.101**

Note: The estimates with smaller standard deviations are marked with bold.

Table 5: Means and Standard Deviations for Probit Case 2, averaged over 1000 samples

		N=400, G=100, L=4		N=1600, G=400, L=4	
		Probit	GEE-probit	Probit	GEE-probit
$\rho = 0$	$\hat{\beta}_2$	1.068	1.033	1.009	1.004
	s.d.$\left(\hat{\beta}_2\right)$	0.226	**0.143**	0.101	**0.067**
	$\hat{\beta}_3$	1.070	1.036	1.011	1.006
	s.d.$\left(\hat{\beta}_3\right)$	0.196	**0.127**	0.085	**0.061**
	$\hat{\beta}_4$	1.056	1.019	1.006	1.002
	s.d.$\left(\hat{\beta}_4\right)$	0.301	**0.214**	0.142	**0.099**
$\rho = 0.2$	$\hat{\beta}_2$	1.100	1.046	1.023	1.013
	s.d.$\left(\hat{\beta}_2\right)$	0.252	**0.139**	0.102	**0.070**
	$\hat{\beta}_3$	1.087	1.040	1.020	1.012
	s.d.$\left(\hat{\beta}_3\right)$	0.212	**0.124**	0.085	**0.060**
	$\hat{\beta}_4$	1.096	1.043	1.021	1.012
	s.d.$\left(\hat{\beta}_4\right)$	0.343	**0.210**	0.138	**0.103**
$\rho = 0.4$	$\hat{\beta}_2$	1.106	1.059	1.036	1.024
	s.d.$\left(\hat{\beta}_2\right)$	0.257	**0.153**	0.106	**0.071**
	$\hat{\beta}_3$	1.099	1.058	1.034	1.022
	s.d.$\left(\hat{\beta}_3\right)$	0.207	**0.133**	0.091	**0.065**
	$\hat{\beta}_4$	1.104	1.059	1.031	1.020
	s.d.$\left(\hat{\beta}_4\right)$	0.326	**0.213**	0.145	**0.103**
$\rho = 0.6$	$\hat{\beta}_2$	1.148	1.090	1.041	1.035
	s.d.$\left(\hat{\beta}_2\right)$	0.270	**0.164**	0.109	**0.077**
	$\hat{\beta}_3$	1.140	1.089	1.037	1.030
	s.d.$\left(\hat{\beta}_3\right)$	0.238	**0.157**	0.096	**0.072**
	$\hat{\beta}_4$	1.131	1.074	1.039	1.034
	s.d.$\left(\hat{\beta}_4\right)$	0.346	**0.232**	0.151	**0.106**

Note: The estimates with smaller standard deviations are marked with bold.

(2003) specify the gravity equation as

$$T_{ij} = \alpha_0 Y_i^{\alpha_1} Y_j^{\alpha_2} D_{ij}^{\alpha_3} \eta_{ij} \tag{67}$$

where T_{ij} is the trade flows between country i and country j. T_{ij} is proportional to the product of the two countries' GDPs, denoted by Y_i and Y_j, and inversely proportional to their distance. D_{ij} broadly represents trade resistance. Let η_{ij} be a stochastic error that represents deviations from the theory. As a tradition in the existing literature, by taking the natural logarithms of both sides and adding other control variables represented by Z_{ij}, the log-linearized equation is:

$$\ln T_{ij} = \ln \alpha_0 + \alpha_1 \ln Y_i + \alpha_2 \ln Y_j + \alpha_3 \ln D_{ij} + \beta Z_{ij} + \ln \eta_{ij} \tag{68}$$

For the above equation, a traditional estimation approach is to use ordinary least squares (OLS). However, there are two problems with the OLS estimation of the log linearized model. First, T_{ij} must be positive in order to take the logarithm. A transformation of $\log(T_{ij} + 1)$ can solve the problem of logarithm but it is not clear how to interpret the estimation results with respect to the original values. Second, the

estimation heavily depends on the independence assumption of η_{ij} and explanatory variables, which means the variance of η_{ij} cannot depend on the explanatory variables. Because of taking the logarithm, only under very specific conditions on η_{ij} is the log linear representation of the constant-elasticity model useful as a device to estimate the parameters of interest (Silva and Tenreyro (2006)). Jensen's inequality implies that $(\mathsf{E} \log Y)$ is smaller than $\log \mathsf{E}(Y)$, thus log-linearized models estimated by OLS as elasticities can be highly misleading in the presence of heteroscedasticity. If the variance of η_{ij} is dependent on the explanatory variables, ordinary least squares is not consistent any more.

We adopt this specification and augment it to the inflow FDI to cities of China. and use nonlinear estimation method, the GEE estimation. The estimating equation is specified as follows

$$\begin{aligned} \mathsf{E}\left(FDI_i|X_i\right) &= \exp[\beta_0 + \beta_1 \ln(GDP_i) + \beta_2 \ln\left(GDPPC_i\right) + \beta_3 \ln(WAGE_i) \\ &\quad + \beta_4 \ln\left(SCIEXP_i\right) + \beta_5 BORDER_i], \end{aligned} \tag{69}$$

where FDI_i is the inflow FDI in actual use for city i, X_i represents all explanatory variables. The control variables includes city level GDP, GDP per capita, the average wage, the government expenditure to science, and whether the city is on the border. We collect data of inflow FDI to 287 cities in 31 provincial administrative regions in 2007 in mainland China from the website of Development Research Center of the State Council of P. R. China [6]. Three cities, Jiayuguan (Gansu Province), Dingxi (Gansu Province) and Karamay (Xinjiang Province), are dropped because of missing data on FDI. Thus we are using 284 cities in total. We collect the latitudes and longitudes of the center of each city using Google map and calculated the geographical distance matrix between cities. The city center is defined as the location of the city government. We use provinces as natural grouping so there are 31 groups. Each group has one to twenty cities. The descriptive statistics are in Table 6. The grouping information is in Table 7.

For comparison, we also provide the OLS estimates of the log-linearized model:

$$\begin{aligned} \ln\left(FDI_i\right) &= \beta_0 + \beta_1 \ln(GDP_i) + \beta_2 \ln\left(GDPPC_i\right) + \beta_3 \ln(WAGE_i) \\ &\quad + \beta_4 \ln\left(SCIEXP_i\right) + \beta_5 BORDER_i + u_i. \end{aligned} \tag{70}$$

[6]The website of Development Research Center of the State Council of P. R. China is www.drcnet.com.cn

Table 6: Descriptive statistics

Variables	Obs	Average	Std.Dev.	Min	Max	Variable description
FDI	284	43571.94	99369.96	0	791954	10,000 dollars
lnFDI	275	9.28	1.81	3.14	13.58	
GDP	287	9451788	1.31e+07	618352	1.20e+08	10,000 yuan
lnGDP	287	15.58	0.92	13.34	18.60	
GDPPC	287	21566.76	16506.67	3398	98938	yuan
lnGDPPC	287	9.76	0.65	8.13	11.50	
WAGE	287	21228.01	5800.10	9523.21	49311.1	yearly, yuan.
lnWAGE	287	9.93	0.25	9.16	10.81	
SCIEXP	287	23513.22	91766.74	469	1100000	10,000 yuan
lnSCIEXP	287	8.86	1.25	6.15	13.91	
BORDER	287	0.06	0.24	0	1	=1 if on the border

The log linearized model suffers from two main problems, first the dependent variable cannot take log if it is zero; second as mentioned in Silva and Tenreyro (2006) the log linearization can cause bias in parameter estimates if there exists heteroskedasticity in the error term u_i.

To estimate the equation for FDI, we use OLS, Poisson QMLE, Poisson GEE with the exchangeable working matrix, NB QMLE, NB GEE with the exchangeable working matrix. In Table 8 the results show advantage of Poisson GEE estimation. All estimation results verifies the positive effect of GDP and GDP per capita in the gravity equation for FDI. These estimates are all significant at the 1% level. What is more, the standard error of GDP and GDP per capita for Poisson GEE is smaller than that for Poisson QMLE, which is smaller than that for OLS. The Poisson regression has significant results on the explanatory variables, log(wage), log(sciexp) and border, which are not significant in the OLS regression. The local average wage has a negative effect on inflow FDI to this city. Compared to other estimation methods, the Poisson GEE estimates on log(wage) is the most significant, at 1% level. It means that when the average wage increase by 1%, the inflow FDI would decrease by about 1%, which could due to the inhabiting effect of labor cost. Similarly, when local government increase science expenditure by 1%, the inflow FDI would increase by about 0.3%, which is shown by Poisson QMLE and Poisson GEE, and in which case the Poisson GEE estimate has smaller standard error than Poisson QMLE, which are 0.102 and 0.110 respectively.

Table 7: Grouping information

Group	Province	Freq.	Percent	Group	Province	Freq.	Percent
1	Beijing	1	0.35	17	Henan	17	5.92
2	Tianjin	1	0.35	18	Hubei	12	4.18
3	Hebei	11	3.83	19	Hunan	13	4.53
4	Shanxi	11	3.83	20	Guangdong	21	7.32
5	Guangxi	14	4.88	21	Hainan	2	0.70
6	Inner Mongolia	9	3.14	22	Chongqing	1	0.35
7	Liaoning	14	4.88	23	Sichuan	18	6.27
8	Jilin	8	2.79	24	Guizhou	4	2.07
9	Heilongjiang	12	4.18	25	Yunnan	8	2.76
10	Shanghai	1	0.35	26	Shaanxi	10	3.45
11	Jiangsu	13	4.53	27	Gansu	12	4.14
12	Zhejiang	11	3.83	28	Qinghai	1	0.34
13	Anhui	17	5.92	29	Ningxia	5	1.72
14	Fujian	9	3.14	30	Xinjiang	2	0.69
15	Jiangxi	11	3.83	31	Tibet	1	0.34
16	Shandong	17	5.92	Total		287	1.00

Table 8: Estimating the FDI equation

	OLS	Poisson	GEE _poisson	NB	GEE _nb2
lnGDP	1.099***	0.705***	0.746***	1.071***	0.982***
	(0.188)	(0.151)	(0.132)	(0.205)	(0.176)
lnGDPPC	0.570***	0.747***	0.687***	0.610***	0.533***
	(0.219)	(0.134)	(0.122)	(0.157)	(0.172)
lnWAGE	-0.123	-0.726*	-1.013***	-0.146	-0.111
	(0.393)	(0.384)	(0.390)	(0.400)	(0.331)
lnSCIEXP	0.186	0.289***	0.311***	0.094	0.137
	(0.142)	(0.110)	(0.102)	(0.111)	(0.106)
BORDER	-0.192	-0.593***	-0.197*	-0.556**	-0.037
	(0.187)	(0.166)	(0.128)	(0.185)	(0.273)
_cons	-13.894***	-3.884	-1.238	-12.360***	-11.021***
	(3.670)	(3.094)	(3.011)	(3.130)	(2.863)
Observations	275	284	284	284	284
F(5, 269)	152.03				
Wald Chi2(5)		701.24	269.58	602.66	495.67
p value	0.000	0.000	0.000	0.000	0.000

Note: Robust standard errors are in parentheses.

***, ** and * indicate significance at the 1%, 5%, and 10% level separately.

7 Appendix

7.1 Some Useful Lemmas

We verify the L_1 NED property of $q_g(\theta,\gamma)$, $h_g(\theta,\gamma)$ accordingly via the L_4 NED property of $\mathbf{y}_g$, and the L_2 NED property of $s_g(\theta,\gamma)$ for central limit theorem.

Lemma 1. *Under condition A.1)- A.8), $q_g(\theta,\gamma)$ is L_1 NED on $\tilde{\varepsilon}$, with the NED constant as $d_g \overset{\text{def}}{=} \max_{i\in B_g} d_{n,i}$, and with the NED coefficients $\psi(s)$. Moreover, we have ULLN for the partial sum $\{M_G G\}^{-1}\sum_g q_g(\theta,\gamma)$, namely*
$sup_{\theta\in\Theta,\gamma\in\Gamma}(M_G G)^{-1}\sum_g\{q_g(\theta,\gamma)-\mathsf{E}[\sum_g q_g(\theta^0,\gamma^0)]\}\to_p 0$.

Proof. We verify $q_g(\theta,\gamma)$ is L_1 NED on $\tilde{\varepsilon}$. From A.1), we work with increasing domain asymptotics, which essentially assume that the growth of the sample size is achieved by an unbounded expansion of the sample region. Namely $|D_G| = G \to \infty$.

The groupwise vector $\mathbf{y}_g$ satisfies $\|\mathbf{y}_g - \mathsf{E}(\mathbf{y}_g|\mathcal{F}_g(s))\|_2 \le \sum_{i\in B_g} d_{i,n}\psi(s) \le d_g L\psi(s)$ ($d_g = \max_{i\in B_g} d_{i,n}$) for $s\to\infty$ and $\psi(s)\to 0$ when $s\to\infty$. We abbreviate $W_{g,ij}$ as an element of $\mathbf{W}_g(\gamma,\theta)$. Thus $y_{g,i}$ is L_2 NED on $\tilde{\varepsilon}$ by A.2). As $\mathsf{E}\,|y_{gi}W_{g,ij}y_{gj} - \mathsf{E}\{y_{gi}|\mathcal{F}_g(s)\}W_{g,ij}\,\mathsf{E}\{y_{gj}|\mathcal{F}_g(s)\}| \le \|y_{gi}-\mathsf{E}\{y_{gi}|\mathcal{F}_g(s)\}\|_2\|W_{g,ij}y_{gj}\|_2$
$+\|y_{gj}-\mathsf{E}\{y_{gj}|\mathcal{F}_g(s)\}\|_2\|W_{g,ij}y_{gi}\|_2 \le C(d_{n,i}\vee d_{n,j})\psi(s)$, by the fact that $\mathcal{F}_i(s)\subset\mathcal{F}_g(s)$ should hold for any $i\in B_g$. Therefore we have $\mathsf{E}\,|(\mathbf{y}_g-\mathbf{m}_g)^\top\mathbf{W}_g(\mathbf{y}_g-\mathbf{m}_g) - \mathsf{E}\{(\mathbf{y}_g-\mathbf{m}_g)^\top\mathbf{W}_g(\mathbf{y}_g-\mathbf{m}_g)|\mathcal{F}_g(s)\}| \le \sum_i\sum_j \mathsf{E}\,|(y_g-m_g)_iW_{g,ij}(y_g-m_g)_j - \mathsf{E}\{(y_g-m_g)_iW_{g,ij}(y_g-m_g)_j|\mathcal{F}_g(s)\}| \le \sum_i\sum_j \mathsf{E}\,|(y_g-m_g)_iW_{g,ij}(y_g-m_g)_j - \mathsf{E}\{(y_g-m_g)_i|\mathcal{F}_g(s)\}W_{g,ij}\,\mathsf{E}\{(y_g-m_g)_j|\mathcal{F}_g(s)\}| \le CL^2d_g\psi(s)$, where $d_g=\max_{i\in B_g}d_{n,i}$ with $d_{n,i}=\mathcal{O}(L)$.

Given the L_1- NED property of $q_g(\theta,\gamma)$ regarding the ULLN, we first look at a pointwise convergence of the function $q_g(.,.)$. We need to verify the following assumptions:

i) There exists non random positive constants $c_g, g\in D_n, n\ge 1$ such that for any θ,γ, such that $\mathsf{E}\,|q_g/c_g|^{p'}<\infty$, where $p'>1$.

ii) The $\alpha-$ mixing coefficients of the input field ε satisfy $\tilde{\alpha}(u,v,r)\le\psi(uL,vL)\hat{\alpha}(r)$, and for some $\hat{\alpha}(r)$, $\sum_{r=1}^{\infty} r^{d-1}L^\tau\hat{\alpha}(r)<\infty$.

Condition i) is implied by A.5) with the moment assumptions on objects involved in $q_g(\gamma,\theta)$ with $c_{g,q}=\mathcal{O}(L^2)$. The reason is that $\mathsf{E}\,|q_g(\gamma,\theta)|^{p'}\le\mathsf{E}\sup_{\theta\in\Theta,\gamma\in\Gamma}|q_g(\gamma,\theta)|^{p'}$. For ii) we see that it is implied from A.6).

Moreover the uniform convergence needs in addition two assumptions:

i) $p'-$ dominance assumption. There exists an array of positive real constants $\{c_{g,q}\}$ such that $p \geq 1$.

$$\limsup_{G} \frac{1}{|D_G|} \sum_{g} \mathsf{E}\left(\mathbf{q}_g^{p'} 1\left(\mathbf{q}_g > k\right)\right) \to 0 \text{ as } k \to \infty, \tag{71}$$

where $\mathbf{q}_g = \sup_{\gamma \in \mathbf{\Gamma}, \theta \in \mathbf{\Theta}} |q_g(\gamma, \theta)| / c_{g,q}$. This is a revision form of the domination condition as Assumption 6 in Jenish and Prucha (2009). Uniform boundedness of $q_g(\gamma, \theta)$ is covered by setting $c_{g,q} = \mathcal{O}(L)$.

ii) Stochastic equicontinuity. We assume that $q_g(\theta, \gamma)$ to be L_0 stochastic equicontinuity on $\Gamma \times \Theta$ iff $\lim_{G\to\infty} 1/|D_G| \sum_{g\in D_G} \mathbb{P}(\sup_{(\gamma'\in\Gamma, \theta'\in\Theta)\in B(\theta',\gamma',\delta)} |q_g(\gamma,\theta) - q_g(\gamma',\theta')| > \varepsilon) \to 0$, where $B(\theta', \gamma', \delta)$ is a $\delta-$ ball around the point γ', θ' with $\nu(\theta, \theta') \leq \delta$ and $\nu(\gamma, \gamma') \leq \delta$.

i) is implied by condition A.5). Namely we would like to prove the condition i), which is implied by the L $-s$ for any constant $s > p'$ boundedness of $\mathbf{q}_g$. Then we need to verify $\sup_g \|\mathbf{q}_g\|_s < C$. As $\|\mathbf{q}_g\|_s = (\mathsf{E}\,|\sup_{\theta\in\Theta,\gamma\in\Gamma} q_g(\theta,\gamma)|^s)^{1/s} \leq \sum_l \sum_g \mathsf{E}\,|\tilde{\varepsilon}_{g,l}^s w_{g,l,m}^s \tilde{\varepsilon}_{g,m}^s|$ $\leq \sum_l \sum_g (\mathsf{E}\,\tilde{\varepsilon}_{g,l}^{2s} \tilde{\varepsilon}_{g,m}^{2s})^{1/(2s)} (\mathsf{E}\, w_{g,l,m}^{2s})^{1/(2s)}$, where $\tilde{\varepsilon}_{g,l} \stackrel{\text{def}}{=} \sup_{\theta\in\Theta} \nabla \mathbf{m}_{g,l}(\theta)$, and $w_{g,l,m} \stackrel{\text{def}}{=} \sup_{\theta\in\Theta,\gamma\in\Gamma} w_{g,l,m}(\gamma,\theta)$. Therefore it can be seen that this will be implied by A.5) with $s < r/4$, with $c_{g,q} = \mathcal{O}(L^2)$.

The stochastic equicontinuity can be guaranteed by $q_g(\theta, \gamma)$ to be Lipschitz in parameter. Namely for any $(\gamma, \theta) \in (\Gamma, \Theta)$ and $(\gamma', \theta') \in (\Gamma, \Theta)$

$$|q_g(\gamma, \theta) - q_g(\gamma', \theta')| \leq B_{g1} g(\nu(\gamma, \gamma')) + B_{g2} g(\nu(\theta, \theta')), \tag{72}$$

where $g(s) \to 0$ when $s \to \infty$, and B_{g1}, B_{g2} are random variables that do not depend on θ, γ. And $p' > 0$,

$$\text{limsup}_{n\to\infty} (|D_G||M_G|)^{-1} \sum_{g} \mathsf{E}\,|B_{gl}|_a^{p'} < \infty. \tag{73}$$

To verify this

$$\begin{aligned} &|c_{g,q}^{-1}(\mathbf{y}_g - \mathbf{m}_g(\theta))^\top \mathbf{W}_g^{-1}(\theta,\gamma)(\mathbf{y}_g - \mathbf{m}_g(\theta)) - c_{g,q}^{-1}(\mathbf{y}_g - \mathbf{m}_g(\theta'))^\top \mathbf{W}_g^{-1}(\theta',\gamma')(\mathbf{y}_g - \mathbf{m}_g(\theta'))| \\ &\leq |c_{g,q}^{-1} \sup_{\theta\in\Theta,\gamma\in\Gamma} s_g(\theta,\gamma)|_2 |\theta - \theta'|_2 + |c_{g,q}^{-1} \sup_{\theta\in\Theta,\gamma\in\Gamma} s_{g,\gamma}(\theta,\gamma)|_2 |\gamma - \gamma'|_2, \end{aligned} \tag{74}$$

where $s_{g,\gamma}(\theta,\gamma) \stackrel{\text{def}}{=} (\mathbf{y}_g - \mathbf{m}_g(\theta))^\top \otimes (\mathbf{y}_g - \mathbf{m}_g(\theta))^\top \partial(\mathbf{W}_g^{-1}(\theta,\gamma))/\partial\gamma$.

By A.5) we have $c_{g,q}^{-1} \| |\sup_{\theta\in\Theta,\gamma\in\Gamma} s_g(\theta,\gamma)|_2 \|_{p'} = \mathcal{O}(p)$, $c_{g,q}^{-1} \| |\sup_{\theta\in\Theta,\gamma\in\Gamma} s_{g,\gamma}(\theta,\gamma)|_2 \|_{p'} = \mathcal{O}(q)$. So we have ii) and the desired results $\sup_{\theta\in\Theta,\gamma\in\Gamma} |\mathbf{Q}_G(\theta,\gamma)_{i,j} - \overline{\mathbf{Q}}_{\infty,i,j}(\theta^0,\gamma^0)| \to$

$\mathcal{O}_p(1)$.

■

Lemma 2. *Under condition A.1)- A.8), $s_g(\theta,\gamma)$ is L_2 NED on $\tilde{\varepsilon}$, with the NED constant as $d_g = \max_{i\in B_g} d_{n,i}$, and with the NED coefficients $\psi(s)$. Moreover, we have a ULLN for the partial sums $(M_G D_G)^{-1}\sum_g s_g(\theta,\gamma)$.*

Proof. This proof is similarly proved as in lemma 1. It can be seen that $\|\|\nabla\mathbf{m}_g^\top\mathbf{W}_g(\theta,\gamma)(\mathbf{y}_g-\mathbf{m}_g(\theta)) - \mathsf{E}\{\nabla\mathbf{m}_g^\top\mathbf{W}_g(\theta,\gamma)(\mathbf{y}_g-\mathbf{m}_g(\theta))|\mathcal{F}_g(s)\}|_2\| \leq \sum_i\sum_j \|\|\nabla m_{gi}^\top W_{ij}(\theta,\gamma)(y_g-m_g)_j - \mathsf{E}\{\nabla m_{gi}^\top W_{ij}(\theta,\gamma)(y_g-m_g)_j|\mathcal{F}_g(s)\}|_2\| \leq \sum_i\sum_j \|\|\nabla m_{gi}^\top W_{ij}(\theta,\gamma)(y_g-m_g)_j - \nabla m_{gi}^\top W_{ij}(\theta,\gamma)\,\mathsf{E}\{(y_g-m_g)_j|\mathcal{F}_g(s)\}|_2\| \leq C'L^2 d_g\psi(s)$, where $d_g = \max_{i\in B_g} d_{n,i}$, and $\max_{i,j}\|\|\nabla m_{gi}^\top W_{ij}|_2\|_4 \lesssim C'pL$ according to A.5). The p' dominance assumption will be following from A.5) given the fact that $\sup_g \mathsf{E}\,|\sup_{\theta\in\Theta,\gamma\in\Gamma} s_g|^r < C$, for $p' < s < r/4$. This would imply the uniform integrability.

Regarding the Lipschitz condition needed for the stochastic equicontinuity property $M_G^{-1}|\nabla\mathbf{m}_g^\top\mathbf{W}_g(\theta,\gamma)(\mathbf{y}_g-\mathbf{m}_g) - \nabla\mathbf{m}_g^\top(\theta')\mathbf{W}_g(\theta',\gamma')(\mathbf{y}_g-\mathbf{m}_g(\theta'))| \leq M_G^{-1}|h_g|_2|\theta-\theta'|_2 + M_G^{-1}|H_{g,\gamma}|_2|\gamma-\gamma'|_2$, where $h_g \overset{\text{def}}{=} \sup_{\theta\in\Theta,\gamma\in\Gamma} h_g(\theta,\gamma)$ and $H_{g,\gamma} \overset{\text{def}}{=} \sup_{\theta\in\Theta,\gamma\in\Gamma}\partial s_g(\gamma,\theta)/\partial\gamma$. The finiteness of $\sup_g \mathsf{E}(|H_g|_2^{p'}), \sup_g \mathsf{E}(|H_{g,\gamma}|_2^{p'})$ will be implied by A.5).

■

Lemma 3. *Under condition A.1)- A.8), $h_g(\theta,\gamma)$ is L_1 NED on $\tilde{\varepsilon}$, with the NED constant as $d_g = \max_{i\in B_g} d_{n,i}$, and with the NED coefficient $\psi(s)$. Moreover, we have a ULLN for the partial sums $(M_G D_G)^{-1}\sum_g h_g(\theta,\gamma)$.*

Proof. Now we verify the component involved in the partial sums in $\mathbf{H}_G(\theta,\hat{\gamma})$ are also L_1 NED on $\tilde{\varepsilon}$.

Namely, $h_{1g} \overset{\text{def}}{=} \nabla_\theta\mathbf{m}_g^\top(\theta)\,\mathbf{W}_g(\gamma,\theta)^{-1}\nabla_\theta\mathbf{m}_g(\theta)$, $h_{2g} \overset{\text{def}}{=} [(\mathbf{y}_g-\mathbf{m}_g(\theta))^\top\mathbf{W}_g(\gamma,\theta)^{-1}\otimes I_q]\partial\text{Vec}(\nabla\mathbf{m}_g^\top)/\partial\theta$, $h_{3g} \overset{\text{def}}{=} \{(\mathbf{y}_g-\mathbf{m}_g(\theta))^\top\otimes\nabla\mathbf{m}_g^\top\}\partial\text{Vec}\{\mathbf{W}_g(\gamma,\theta)\}/\partial\theta$. It is obvious that h_{1g} is NED on $\tilde{\varepsilon}$ as a measurable function of $\mathbf{x}_g$. Define e_i as a $p\times 1$ vector with only the $i-$th component as 1, $|.|_a$ is taking the elementwise absolute value. And $b_{ij} \overset{\text{def}}{=} e_i^\top(\mathbf{1}^\top W_g\otimes I_g)|\partial\text{Vec}(\nabla\mathbf{m}_g)/\partial\theta|_a e_j$. We verify now h_{2g} for any fixed point γ and θ, it can seen that $\mathsf{E}\,|h_{2g,i,j} - \mathsf{E}\{h_{2g,i,j}|\mathcal{F}_g(s)\}| \leq \mathsf{E}\,|e_i^\top([\{\mathbf{y}_{g,i}-\mathsf{E}(\mathbf{y}_{g,i}|\mathcal{F}_g(s))\}^\top\mathbf{W}_g(\gamma,\theta)^{-1}\otimes I_q]\partial\text{Vec}(\nabla\mathbf{m}_g^\top)/\partial\theta)e_j| \leq \mathsf{E}(\max_{i\in B_g}|y_{g,i}-\mathsf{E}[y_{g,i}|\mathcal{F}_g(s)]|b_{ij}|) \leq L^{1/2}\|\sup_{\theta\in\Theta,\gamma\in\Gamma} b_{ij}\| d_g\psi(s)$, where for sufficiently large s and $d_g = \mathcal{O}(L^{1/2}\|b_{ij}\|d_g)$. Therefore we proved the L_1 NED of H_{2g}. Similarly for H_{3g}, define $c_{ij} = e_i^\top(\mathbf{1}^\top\otimes\nabla\mathbf{m}_g^\top(\theta))|\partial\text{Vec}(\nabla\mathbf{m}_g(\theta))/\partial\theta|_a e_j$. Then $\mathsf{E}\,|H_{2g,i,j} - \mathsf{E}\{H_{2g,i,j}|\mathcal{F}_l(s)\}| \leq {L_g}^{1/2}\|\sup_{\theta\in\Theta,\gamma\in\Gamma} c_{ij}\| d_g\psi(s)$, where for sufficiently large s and assume that $L^{1/2}\|\sup_{\theta\in\Theta,\gamma\in\Gamma} c_{ij}\| d_g\psi(s) \to 0$. We proved thus the L_1 NED of H_{3g}. Then we would have the pointwise convergence of $\mathbf{H}_{G,1}(\theta,\gamma)$, $\mathbf{H}_{G,2}(\theta,\gamma)$,

$\mathbf{H}_{G,3}(\theta,\gamma)$ any fixed point $\theta \in \Theta, \gamma \in \Gamma$. To ensure that with probability $1-\mathcal{O}_p(1)$, $|\mathbf{H}_G(\theta,\hat{\gamma}) - \mathbf{H}_\infty(\theta^{\mathbf{0}},\gamma^{\mathbf{0}})| \leq \sup_{\theta\in\Theta,\gamma\in\Gamma}|\mathbf{H}_G(\theta,\gamma) - \mathbf{H}_\infty(\theta^{\mathbf{0}},\gamma^{\mathbf{0}})| \to 0$, therefore we need a ULLN.

Moreover the uniform convergence needs in addition two assumptions:

i) There exists an array of positive real constants $\{c_{g,h}\}$ such that for constant $\delta > 0$:

$$\text{limsup}_G \frac{1}{|D_G|}\sum_g E\left(\mathbf{H}_{l,g,i,j}^{2+\delta}\mathbf{I}\left(\mathbf{H}_{l,g,i,j} > k\right)\right) \to 0 \text{ as } k \to \infty, \tag{75}$$

where $\mathbf{H}_{l,g,i,j} = \sup_{\theta\in\mathbf{\Theta},\gamma\in\mathbf{\Gamma}} |h_{gl,i,j}(\theta,\gamma)/c_{g,h}|_1$. This is a again revision form of the domination condition as Assumption 6 in Jenish and Prucha (2009). Uniform boundedness of $\mathbf{H}_g(\theta)$ is covered by setting $c_{g,h} = \mathcal{O}(L^2)$. $l = 1,2,3$.

ii) Stochastic equicontinuity. We assume that $H_{l,g}(\theta,\gamma)$ to be L_0 stochastic equicontinuity on Γ iff $\lim_{G\to\infty} 1/|D_G| \sum_g \mathbb{P}(\sup_{(\gamma'\in\Gamma,\theta'\in\Theta)\in B(\gamma',\theta',\delta)} |H_{g,i,j}(\gamma,\theta)$ $- H_{g,i,j}(\gamma',\theta')| > \varepsilon) \to 0$.

The stochastic equicontinuity can be guaranteed by $h_{g,i,j}(\theta,\gamma)$ to be Lipschitz in parameter, which is ensured by A.5).

Then we have $\sup_{\gamma\in\Gamma,\theta\in\Theta} |\mathbf{H}_{G,i,j}(\theta,\gamma) - \mathbf{H}_{\infty,i,j}(\gamma,\theta)]| \to \mathcal{O}_p(1)$. ■

7.2 Proof of Theorem 1

Two sufficient conditions for consistent estimators are i) identification implied by A.8) and ii) the objective function $Q_G(\theta,\gamma)$ satisfies the uniform law of large numbers (ULLN). By Lemma 1, we have the uniform LLN of $Q_G(\theta,\gamma)$.

Namely $\theta \in \mathbf{\Theta}, \gamma \in \mathbf{\Gamma}$, $\sup_{\theta\in\Theta,\gamma\in\Gamma} \frac{1}{M_G|D_G|}[Q_G(\theta,\gamma) - Q_\infty(\theta^{\mathbf{0}},\gamma^0)] \xrightarrow{p} 0$, as $G \to \infty$.

Thus we conclude that under A.1)-A.8), the GEE estimator is consistent.

7.3 Proof of Theorem 2

7.3.1 Step 1 : Main expansion step

Recall $\mu_g = \mathbf{y}_g - \mathbf{m}_g(\theta^0)$ and $\hat{\mu}_g = \mathbf{y}_g - \mathbf{m}_g(\hat{\theta})$

$$\mathbf{S}_G(\theta,\hat{\gamma}) = \frac{1}{M_G G}\sum_g \nabla\mathbf{m}_g^\top(\theta)\mathbf{W}_g^{-1}(\hat{\gamma},\theta)[\mathbf{y}_g - \mathbf{m}_g(\theta)]. \tag{76}$$

From the first order condition A.11).

$$\mathbf{S}_G\left(\hat{\gamma},\hat{\theta}\right) = \mathcal{O}_p(1).$$

To expand $\mathbf{S}_G\left(\hat{\gamma},\hat{\theta}\right)$ around the point γ^0,θ^0, we have,

$$\begin{aligned}
&\mathbf{S}_G\left(\hat{\gamma},\hat{\theta}\right)=\mathbf{S}_G\left(\gamma^{\mathbf{0}},\theta^{\mathbf{0}}\right)+\mathbf{H}_G(\tilde{\theta},\tilde{\gamma})(\hat{\theta}-\theta^0)+\nabla_\gamma\mathbf{S}_G(\tilde{\gamma},\tilde{\theta})(\hat{\gamma}-\gamma^0)\\
&=\mathbf{S}_G\left(\gamma^{\mathbf{0}},\theta^{\mathbf{0}}\right)+\mathbf{H}_\infty(\theta^0,\gamma^0)(\hat{\theta}-\theta^0)+\mathbf{F}_0(\hat{\gamma}-\gamma^0)\\
&+\{\mathbf{H}_G(\tilde{\theta},\tilde{\gamma})-\mathbf{H}_\infty(\theta^0,\gamma^0)\}(\hat{\theta}-\theta^0)+\{\nabla_\gamma\mathbf{S}_G(\tilde{\theta},\tilde{\gamma})-\mathbf{F}_0\}(\hat{\gamma}-\gamma^0)
\end{aligned}$$

where $\tilde{\theta},\tilde{\gamma}$ lie in the line segment between θ^0,γ^0 to $\hat{\theta},\hat{\gamma}$, $\mathbf{F}_0$ is a $L\times q$ matrix, $\mathbf{F}_0=\lim_{G\to\infty}\left\{\frac{1}{M_G|D_G|}\sum_g \mathsf{E}\left[\nabla_\gamma s_g\left(\theta^0;\gamma^0\right)\right]\right\}$. From the derivation below we see that $\mathbf{F}_0=\mathbf{0}$, the asymptotic distribution of the average score does not depend on the distribution of $\hat{\gamma}$, and the first-step estimation of $\hat{\gamma}$ will not affect the second-step estimation in terms of asymptotic variance.

$\mathbf{F}_0$ is the the limit of orthogonal score by construction. To identify this, we can see that $\nabla_\gamma\{\nabla\mathbf{m}_g^\top\left(\theta^0\right)\mathbf{W}_g^{-1}(\theta^0,\gamma^0)\left[\mathbf{y}_g-\mathbf{m}_g\left(\theta^0\right)\right]\}$
$=\left\{\mathbf{y}_g-\mathbf{m}_g\left(\theta^0\right)\right\}^\top\otimes\nabla\mathbf{m}_g^\top(\theta^0)\nabla_\gamma\mathrm{Vec}\{\mathbf{W}_g^{-1}(\theta^0,\gamma^0)\}$.
$\left[\mathbf{y}_g-\mathbf{m}_g\left(\theta^0\right)\right]^\top\otimes\nabla\mathbf{m}_g^\top(\theta^0)\nabla_\gamma\mathrm{Vec}\{\mathbf{W}_g^{-1}(\theta^0,\gamma^0)\}$
$=\mathsf{E}[\mathsf{E}[\{\mathbf{y}_g-\mathbf{m}_g\left(\theta^0\right)\}^\top|\mathbf{x}_g]\otimes\nabla\mathbf{m}_g^\top(\theta^0)\nabla_\gamma\mathrm{Vec}\{\mathbf{W}_g^{-1}(\theta^0,\gamma^0)\}]=0$.

To handle the term $\{\mathbf{H}_G(\tilde{\theta},\tilde{\gamma})-\mathbf{H}_\infty(\theta^0,\gamma^0)\}(\hat{\theta}-\theta^0)+\{\nabla_\gamma\mathbf{S}_G^\top(\tilde{\theta},\tilde{\gamma})-\mathbf{F}_0^\top\}(\hat{\gamma}-\gamma^0)$, we need the ULLN for $\mathbf{H}_G(\theta^0,\gamma^0)$ to derive $|e_i^\top(\mathbf{H}_G(\tilde{\theta},\tilde{\gamma})-\mathbf{H}_\infty(\theta^0,\gamma^0))e_j|\leq\sup_{\theta,\gamma}|e_i^\top(\mathbf{H}_G(\theta,\gamma)-\mathbf{H}_\infty(\theta^0,\gamma^0))e_j|\to_p 0$. Also for $\nabla_\gamma\mathbf{S}_G(\tilde{\theta},\gamma)$ to derive $|e_i^\top(\{\nabla_\gamma\mathbf{S}_G^\top(\tilde{\theta},\gamma)-\mathbf{F}_0^\top\})e_j|\leq\sup_{\theta\in\Theta,\gamma\in\Gamma}|e_i^\top\{\nabla_\gamma\mathbf{S}_G^\top(\theta,\gamma)-\mathbf{F}_0^\top\}e_j|\to_p 0$. This is already verified by Lemma 3. We arrive at the conclusion that for any vector $a\in\mathcal{R}^p,|a|_2=1$, $|a^\top\{\mathbf{H}_G(\tilde{\theta},\tilde{\gamma})-a^\top\mathbf{H}_\infty(\theta^0,\gamma^0)\}(\hat{\theta}-\theta^0)|\leq|\{a^\top\mathbf{H}_G(\tilde{\theta},\tilde{\gamma})-a^\top\mathbf{H}_\infty(\theta^0,\gamma^0)\}|_2|(\hat{\theta}-\theta^0)|_2=\mathcal{o}_p(1)\times\mathcal{O}_p(|(\hat{\theta}-\theta^0)|_2)=\mathcal{o}_p(|(\hat{\theta}-\theta^0)|_2)$ and $|a^\top\{\nabla_\gamma\mathbf{S}_G^\top(\tilde{\theta},\tilde{\gamma})-\mathbf{F}_0^\top\}(\hat{\gamma}-\gamma^0)|_2=\mathcal{o}_p(|(\hat{\gamma}-\gamma^0)|_2)=\mathcal{o}_p(G^{-1/2})$ by A.8).

Next we look at the invertibility of the matrix $\mathbf{H}_\infty(\theta^0,\gamma^0)$. Taking the expected value of the score function over the distribution of $\left(\mathbf{x}_g,\mathbf{y}_g\right)$ gives

$$\begin{aligned}
\mathsf{E}\left[h_g\left(\theta^0,\gamma^0\right)\right] &= \mathsf{E}[\mathsf{E}[\mathbf{h}_g\left(\mathbf{w}_g,\theta^0,\gamma^0\right)|\mathbf{x}_g]]\\
&= \mathsf{E}[\{(\mathbf{y}_g-\mathbf{m}_g(\theta^0))^\top\otimes\nabla\mathbf{m}_g^\top(\theta^0)\}\partial\mathrm{Vec}\{\mathbf{W}_g\left(\theta^0,\gamma^0\right)\}/\partial\theta]\\
&\quad-\mathsf{E}[\nabla\mathbf{m}_g^\top(\theta^0)\mathbf{W}_g^{-1}(\theta^0,\gamma^0)\nabla_\theta\mathbf{m}_g(\theta^0)]\\
&\quad+\mathsf{E}[[\{(\mathbf{y}_g-\mathbf{m}_g(\theta^0))^\top\mathbf{W}_g^{-1}\left(\theta^0,\gamma^0\right)\otimes I_q\}]\partial\mathrm{Vec}(\nabla\mathbf{m}_g^\top)/\partial\theta]\\
&= \mathsf{E}[-\nabla\mathbf{m}_g(\theta^0)^\top\mathbf{W}_g^{-1}\left(\theta^0,\gamma^0\right)\nabla\mathbf{m}_g(\theta^0)]\\
&\quad+\mathsf{E}[\{\mathsf{E}[(\mathbf{y}_g-\mathbf{m}_g(\theta^0))^\top|\mathbf{x}_g]\otimes\nabla\mathbf{m}_g^\top\}\partial\mathrm{Vec}(\mathbf{W}_g(\theta^0,\gamma^0))/\partial\theta]\\
&\quad+\mathsf{E}[\{\mathsf{E}[\{(\mathbf{y}_g-\mathbf{m}_g(\theta^0))^\top|\mathbf{x}_g\}\mathbf{W}_g^{-1}\left(\theta^0,\gamma^0\right)\otimes I_q]\}\partial\mathrm{Vec}(\nabla\mathbf{m}_g^\top)(\theta^0)/\partial\theta]\\
&= \mathsf{E}[-\nabla\mathbf{m}_g^\top(\theta^0)\mathbf{W}_g^{-1}\left(\theta^0,\gamma^0\right)\nabla\mathbf{m}_g(\theta^0)],
\end{aligned}$$

which is negative definite by assumption A.9).

The GEE estimator can be specifically written as

$$\sqrt{G}\left(\hat{\theta}-\theta_0\right)=\left[\mathbf{H}_\infty(\theta^0,\gamma^0)\right]^{-1}\frac{1}{\sqrt{G}}\sum_g s_g\left(\theta^0,\gamma^{\mathbf{0}}\right)+\mathcal{O}_p(1)+\mathcal{O}_p(\sqrt{G}|\hat{\theta}-\theta^0|_2). \quad (77)$$

Due to the L_2 NED property of s_g, $\mathsf{Var}(\sum_{g=1}^G s_g)=\mathcal{O}(G)$, thus

we have $\sqrt{G}|\hat{\theta}-\theta^0|_2 \lesssim |\mathbf{H}_\infty(\theta^0,\gamma^0)^{-1}|_2=\mathcal{O}_p(CM_G^2)$, as the order of $\frac{1}{\sqrt{G}}\sum_g s_g\left(\mathbf{w}_g,\theta^0;\gamma^{\mathbf{0}}\right)$ under assumption B.3) is $\mathcal{O}_p(G^{-1/2})$. This implies that $\mathcal{O}_p(\sqrt{G}|\hat{\theta}-\theta^0|_2)=\mathcal{O}_p(1)$.

7.3.2 Step 2 Central Limit Theorem

We derive the variance of $s_g\left(\mathbf{w}_g,\theta^0,\gamma^0\right)$ in this subsection.

$$\begin{aligned}
& AS_G=\mathrm{Var}\left[\frac{1}{\sqrt{G}}\sum_g s_g\left(\mathbf{w}_g,\theta^0,\gamma^0\right)\right] \\
= \;& \mathrm{Var}\left\{\frac{1}{\sqrt{G}}\sum_g \nabla\mathbf{m}_g^\top\left(\theta^0\right)\mathbf{W}_g^{-1}\left(\theta^0,\gamma^0\right)\left[\mathbf{y}_g-\mathbf{m}_g\left(\theta^0\right)\right]\right\} \\
= \;& \mathrm{Var}\left[\frac{1}{\sqrt{G}}\sum_g \nabla\mathbf{m}_g^\top\left(\theta^0\right)\mathbf{W}_g^{-1}\left(\theta^0,\gamma^0\right)\mathbf{u}_g\right] \\
= \;& \frac{1}{G}\sum_g \mathsf{E}\left[\nabla\mathbf{m}_g^\top\left(\theta^0\right)\mathbf{W}_g^{-1}\left(\theta^0,\gamma^0\right)\mathbf{u}_g\mathbf{u}_g^\top\mathbf{W}_g^{-1}\left(\theta^0,\gamma^0\right)\nabla\mathbf{m}_g\left(\theta^0\right)\right] \\
& +\frac{1}{G}\sum_g\sum_{h,h\neq g}\mathsf{E}\left[\nabla\mathbf{m}_g^\top\left(\theta^0\right)\mathbf{W}_g^{-1}\left(\theta^0,\gamma^0\right)\mathbf{u}_g\mathbf{u}_h^\top\mathbf{W}_h^{-1}\left(\theta^0,\gamma^0\right)\nabla\mathbf{m}_h\left(\theta^0\right)\right].
\end{aligned}$$

The next step is to apply the central limit theorem (Corollary 1 in Jenish and Prucha (2012)) the element $\mathbf{S}_G=\frac{1}{\sqrt{G}}\sum_g s_g\left(\mathbf{w}_g,\theta^0,\gamma^0\right)$, and $AS_\infty=\lim_{G\to\infty}AS_G$. For that we need to verify the following conditions:

i) s_g is uniform L_2 NED on the $\alpha-$ mixing random field $\tilde{\varepsilon}$ with coefficients d_gL and $\psi(s)$, $\sup_{G,g}d_gL<\infty$ and $\sum_{r=1}^\infty r^{d-1}\psi(r)<\infty$. Moreover $\sup_G\sup_g\|s_g\|_r$, where $r>2+\delta'$, with δ' as a constant.

ii) The input field $\tilde{\varepsilon}$ is $\alpha-$ mixing with coefficient $\sum_{r=1}^\infty r^{(d\tau^*+d)-1}L^{\tau^*}\hat{\alpha}^{\delta/(2+\delta')}(r)<\infty$. ($\tau^*=\delta'\tau/(4+2\delta')$)

iii) $\inf_G|D_G|^{-1}M_G^{-2}\lambda_{min}(\mathbf{AS}_\infty)>0$. (suppressed G for the triangular array.)

i) is proved in Lemma 2, ii) can be inferred by A.11), and iii) can be inferred from

A.10). Therefore under A.1)-A.11)

$$\mathbf{A}\mathbf{S}_{\infty}^{-1/2}\mathbf{S}_G \Rightarrow \mathbb{N}(0, I_p). \tag{78}$$

So we have $AV(\hat{\theta}) = \mathbf{H}_{\infty}^{\top}\mathbf{A}\mathbf{S}_{\infty}\mathbf{H}_{\infty}$

$$\sqrt{G}AV(\hat{\theta})^{-1/2}(\hat{\theta} - \theta^0) \Rightarrow \mathbb{N}(0, I_p). \tag{79}$$

7.4 Proof of Proposition 1

A.8)' (Identifiability) $\overline{E}_G(\theta, \gamma) \stackrel{\text{def}}{=} \sum_g (\mathbf{e}_g(\check{\theta}) - \mathbf{z}_g(\gamma))^{\top}(\mathbf{e}_g(\check{\theta}) - \mathbf{z}_g(\gamma))$. And $E_{\infty}(\gamma, \theta) \stackrel{\text{def}}{=} \lim_{G\to\infty} \overline{E}_G(\gamma, \theta)$. Assume that θ^0, γ^0 are identified unique in a sense that $\liminf_{G\to\infty} \inf_{\gamma\in\Gamma:\nu(\gamma,\gamma^0)\geq\varepsilon} \overline{E}_G(\theta, \gamma) > c_0 > 0$, for a positive constant c_0.

A.9)' The true point θ^0, γ^0 lies in the interior point of Θ, Γ. $\check{\theta}$ is estimated with $|\check{\theta} - \theta^0|_2 = \mathcal{O}_p(G^{-1/2})$.

A.11)' $(|D_G|M_G)^{-1} \sum_g \sum_l \sum_{m<l} (\mathbf{e}_{glm}(\check{\theta}) - \mathbf{z}_{glm}(\hat{\gamma}))\partial \mathbf{z}_{glm}(\hat{\gamma})/\partial\gamma = o_p(1)$.

In this subsection, we verify the consistency of the preestimator $\hat{\gamma}$. As we have

$$\hat{\gamma} = \mathbf{argmin}_{\gamma} \sum_g (\mathbf{e}_g(\check{\theta}) - \mathbf{z}_g(\gamma))^{\top}(\mathbf{e}_g(\check{\theta}) - \mathbf{z}_g(\gamma)), \tag{80}$$

which leads to $\mathbf{argzero}_{\gamma\in\Gamma} \sum_g \sum_l \sum_{m<l} (\mathbf{e}_{glm}(\check{\theta}) - \mathbf{z}_{glm}(\gamma))\partial \mathbf{z}_{glm}(\gamma)/\partial\gamma = 0$.

We can proceed with a similar expansion step as in Section 7.3.1. Therefore
$\sum_g \sum_l \sum_{m<l} \{\mathbf{e}_{glm}(\check{\theta}) - \mathbf{z}_{glm}(\hat{\gamma})\}\partial \mathbf{z}_{glm}(\hat{\gamma})/\partial\gamma$
$= \sum_g \sum_l \sum_{m<l} \{\mathbf{e}_{glm}(\theta^0) - \mathbf{z}_{glm}(\gamma^0)\}\partial \mathbf{z}_{glm}(\gamma^0)/\partial\gamma$
$+ \sum_g \sum_l \sum_{m<l} \{\mathbf{e}_{glm}(\tilde{\theta}) - \mathbf{z}_{glm}(\tilde{\gamma})\}\partial \mathbf{z}_{glm}(\tilde{\gamma})/\partial\gamma\partial\gamma^{\top}(\hat{\gamma} - \gamma^0)$
$- \sum_g \sum_l \sum_{m<l} \{\partial \mathbf{z}_{glm}(\tilde{\gamma})/\partial\gamma\}\partial \mathbf{z}_{glm}(\tilde{\gamma})/\partial\gamma^{\top}(\hat{\gamma} - \gamma^0)$
$+ \sum_g \sum_l \sum_{m<l} \{\partial \mathbf{z}_{glm}(\tilde{\gamma})/\partial\gamma\}\partial \mathbf{e}_{glm}(\tilde{\theta})/\partial\theta^{\top}(\check{\theta} - \theta^0)$, where $\tilde{\gamma}, \tilde{\theta}$ lies in the line segment between θ^0, γ^0 and $\check{\theta}, \hat{\gamma}$.

It is known that under proper NED assumptions a pooled estimation $\check{\theta}$ satisfying $|\check{\theta} - \theta^0|_2 = \mathcal{O}_p(1/\sqrt{n})$. The verification step would be similar to the proof in Section 7.3.1, where we also need ULLN for the term $G^{-1} \sum_g \sum_l \sum_{m<l} \partial \mathbf{z}_{glm}(\tilde{\gamma})/\partial\gamma\partial \mathbf{z}_{glm}(\tilde{\gamma})/\partial\gamma^{\top}$, $2G^{-1} \sum_g \sum_l \sum_{m<l} \{\mathbf{e}_{glm}(\tilde{\theta}) - \mathbf{z}_{glm}(\tilde{\gamma})\}\partial \mathbf{e}_{glm}(\tilde{\theta})/\partial\theta$ and $G^{-1} \sum_g \sum_l \sum_{m<l} \{\mathbf{e}_{glm}(\tilde{\theta}) - \mathbf{z}_{glm}(\tilde{\gamma})\}\partial \mathbf{z}_{glm}(\tilde{\gamma})/\partial\gamma\partial\gamma$
This will lead to $\sum_g \sum_l \sum_{m<l} \{\mathbf{e}_{glm}(\theta^0) - \mathbf{z}_{glm}(\gamma^0)\}\partial \mathbf{z}_{glm}(\gamma^0)/\partial\gamma = \mathcal{O}_p(\sqrt{G})$. (Lemma A.3 in Jenish and Prucha (2012)).

The desired results now follows from condition A.1) - A.3), A.5), A.6) and A.8)', A9)', A11)'.

7.5 Proof of Theorem 3

We prove that $\sup_{(\gamma,\theta)\in(\Gamma,\Theta)} e_i^\top \hat{\mathbf{A}}(\theta,\gamma) e_j \to_p e_i^\top \mathbf{A}_0 e_j$,
and $\sup_{(\gamma,\theta)\in(\Gamma,\Theta)} e_i^\top \hat{\mathbf{B}}(\theta,\gamma) e_j \to_p e_i^\top \mathbf{B}_0 e_j$. And by the Slutsky's theorem the variance covariance estimation is consistent. Firstly we prove that $e_i^\top(\hat{\mathbf{A}} - \mathbf{A}_0)e_j \to_p 0$. This is implied by uniform law of large numbers for near-epoch dependent sequences, as mentioned the NED property of the underlying sequence $(\mathbf{x}_g)$ is trivial under condition A.1) - A.5) as it is a measurable function of the input field $\tilde{\varepsilon}$.

$e_i^\top \hat{\mathbf{A}} e_j = \frac{1}{GM_G} \sum_g e_i^\top \nabla \hat{\mathbf{m}}_g^\top \hat{\mathbf{W}}_g^{-1} \nabla \hat{\mathbf{m}}_g e_j \to_p$
$\lim_{G\to\infty} \frac{1}{GM_G} \sum_g e_i^\top \mathsf{E}\left(\nabla \mathbf{m}_g^\top \mathbf{W}_g^{-1} \nabla \mathbf{m}_g\right) e_j = e_i^\top \mathbf{A}_0 e_j$.

We still need to prove that $e_i^\top \hat{\mathbf{B}} e_j \to_p e_i^\top \mathbf{B}_0 e_j$. We denote $\mathbf{W}_g = \mathbf{W}_g(\theta^0,\gamma^0)$ and $\hat{\mathbf{W}}_g = \mathbf{W}_g(\hat{\theta},\hat{\gamma})$.

Recall that $Z_g \overset{\text{def}}{=} \nabla \mathbf{m}_g^\top \mathbf{W}_g^{-1} \mathbf{u}_g$, and $\hat{Z}_g \overset{\text{def}}{=} \nabla \hat{\mathbf{m}}_g^\top \hat{\mathbf{W}}_g^{-1} \hat{\mathbf{u}}_g$.

$$
\begin{aligned}
\mathbf{B}_0 &= \lim_{G\to\infty} \mathrm{Var}\left[\frac{1}{\sqrt{M_G^2|D_G|}} \sum_g s_g\left(\theta^0,\gamma^0\right)\right] \\
&= \lim_{G\to\infty} \frac{1}{M_G^2|D_G|} \sum_g \mathsf{E}\left[\nabla \mathbf{m}_g^\top \mathbf{W}_g^{-1} \mathbf{u}_g \mathbf{u}_g^\top \mathbf{W}_g^{-1} \nabla \mathbf{m}_g\right] \\
&\quad + \frac{1}{M_G^2|D_G|} \sum_g \sum_{h(\neq g)} \mathsf{E}\left[\nabla \mathbf{m}_g^\top \mathbf{W}_g^{-1} \mathbf{u}_g \mathbf{u}_h^\top \mathbf{W}_h^{-1} \nabla \mathbf{m}_h\right] \\
&= \lim_{G\to\infty} \frac{1}{M_G^2|D_G|} \sum_g \mathsf{E}\left[Z_g^\top Z_g\right] + \frac{1}{M_G^2|D_G|} \sum_g \sum_{h(\neq g)\in D_G} \mathsf{E}\left[Z_g^\top Z_h\right]. \\
\hat{\mathbf{B}} &= \frac{1}{M_G^2|D_G|} \sum_g \sum_{h(\neq g)} k(d_{gh}) \nabla \hat{\mathbf{m}}_g^\top \hat{\mathbf{W}}_g^{-1} \hat{\mathbf{u}}_g \hat{\mathbf{u}}_h^\top \hat{\mathbf{W}}_h^{-1} \nabla \hat{\mathbf{m}}_h, \\
&= \frac{1}{M_G^2|D_G|} \sum_g \nabla \hat{\mathbf{m}}_g^\top \hat{\mathbf{W}}_g^{-1} \hat{\mathbf{u}}_g \hat{\mathbf{u}}_g^\top \hat{\mathbf{W}}_g^{-1} \nabla \hat{\mathbf{m}}_g \\
&\quad + \frac{1}{M_G^2|D_G|} \sum_g \sum_{h(\neq g)} k(d_{gh}) \nabla \hat{\mathbf{m}}_h \hat{\mathbf{W}}_g^{-1} \hat{\mathbf{u}}_g \hat{\mathbf{u}}_h^\top \hat{\mathbf{W}}_h^{-1} \nabla \hat{\mathbf{m}}_h \\
&= \frac{1}{M_G^2|D_G|} \sum_g \hat{Z}_g^\top \hat{Z}_g + \frac{1}{M_G^2|D_G|} \sum_g \sum_{h(\neq g)} k(d_{gh}) \hat{Z}_g^\top \hat{Z}_h.
\end{aligned}
$$

Define $\mathbf{B}_0^k$ and $\mathbf{B}^k$ as

$$\begin{aligned}
\mathbf{B}_0^k &= \frac{1}{M_G^2|D_G|}\sum_g \mathsf{E}\left[\nabla \mathbf{m}_g^\top \mathbf{W}_g^{-1}\mathbf{u}_g\mathbf{u}_g^\top \mathbf{W}_g^{-1}\nabla \mathbf{m}_g\right] \\
&\quad + \frac{1}{M_G^2|D_G|}\sum_g \sum_{h(\neq g)} k(d_{gh})\,\mathsf{E}\left[\nabla \mathbf{m}_g^\top \mathbf{W}_g^{-1}\mathbf{u}_g\mathbf{u}_h^\top \mathbf{W}_h^{-1}\nabla \mathbf{m}_h\right] \\
&= \frac{1}{M_G^2|D_G|}\sum_g \mathsf{E}\left(Z_g^\top Z_g\right) + \frac{1}{M_G^2|D_G|}\sum_{g\in D_G y}\sum_{h(\neq g)} k(d_{gh})\,\mathsf{E}\left(Z_g^\top Z_h\right). \\
\mathbf{B}^k &= \frac{1}{M_G^2|D_G|}\sum_{h(\neq g)} \left[\nabla \mathbf{m}_g^\top \mathbf{W}_g^{-1}\mathbf{u}_g\mathbf{u}_g^\top \mathbf{W}_g^{-1}\nabla \mathbf{m}_g\right] \\
&\quad + \frac{1}{M_G^2|D_G|}\sum_g \sum_{h(\neq g)} k(d_{gh})\left[\nabla \mathbf{m}_g^\top \mathbf{W}_g^{-1}\mathbf{u}_g\mathbf{u}_h^\top \mathbf{W}_h^{-1}\nabla \mathbf{m}_h\right] \\
&= \frac{1}{M_G^2|D_G|}\sum_g Z_g^\top Z_g + \frac{1}{M_G^2|D_G|}\sum_g \sum_{h(\neq g)} k(d_{gh}) Z_g^\top Z_h.
\end{aligned}$$

Next write the estimation error for $\mathbf{B}_0$ as in three parts, namely the part consists of generated errors (I_1), the variance (I_2) and the bias part (I_3). We need to prove that the generated error term is negligible, the variance term is small induced by the property NED, and the bias term is also small.

$$\begin{aligned}
& \left|e_i^\top(\hat{\mathbf{B}} - \mathbf{B}_0)e_j\right| \\
= & \left|e_i^\top(\hat{\mathbf{B}} - \mathbf{B}^k)e_j + e_i^\top(\mathbf{B}^k - \mathbf{B}_0^k)e_j + e_i^\top(\mathbf{B}_0^k - \mathbf{B}_0)e_j\right| \\
\leq & \left|e_i^\top(\hat{\mathbf{B}} - \mathbf{B}^k)e_j\right| + \left|e_i^\top(\mathbf{B}^{\mathbf{k}} - \mathbf{B}_{\mathbf{0}}^{\mathbf{k}})e_j\right| + \left|e_i^\top(\mathbf{B}_{\mathbf{0}}^{\mathbf{k}} - \mathbf{B}_0)e_j\right| \\
\overset{\text{def}}{=} & I_1 + I_2 + I_3
\end{aligned}$$

The following statement are what we need to to prove, and will lead to $\left|e_i^\top(\hat{\mathbf{B}} - \mathbf{B}_0)e_j\right| = o_p(1)$.

$$\begin{aligned}
I_1 &= |e_i^\top(\hat{\mathbf{B}} - \mathbf{B}^k)e_j| \\
&= |\frac{1}{M_G^2|D_G|}\sum_g e_i^\top \hat{Z}_g^\top \hat{Z}_g e_j + \frac{1}{M_G^2|D_G|}\sum_g \sum_{h(\neq g)} k(d_{gh}) e_i^\top \hat{Z}_g^\top \hat{Z}_h e_j \\
&\quad -[\frac{1}{M_G^2|D_G|}\sum_g e_i^\top Z_g^\top Z_g e_j + \frac{1}{M_G^2|D_G|}\sum_g \sum_{h(\neq g)} k(d_{gh}) e_i^\top Z_g^\top Z_h e_j]| = o_p(1)
\end{aligned}$$

$$
\begin{aligned}
I_2 &= |e_i^\top(\mathbf{B}^k - \mathbf{B}_0^k)e_j| \\
&= |\frac{1}{M_G^2|D_G|}\sum_g e_i^\top\left[Z_g^\top Z_g - \mathsf{E}\left(Z_g^\top Z_g\right)\right]e_j \\
&\quad + \frac{1}{M_G^2|D_G|}\sum_g\sum_{h(\neq g)} k(d_{gh})e_i^\top\left[Z_g^\top Z_h - \mathsf{E}\left(Z_g^\top Z_h\right)\right]e_j| \\
&= o_p(1)
\end{aligned}
$$

$$
\begin{aligned}
I_3 &= \left|e_i^\top(\mathbf{B}_0^k - \mathbf{B}_0)e_j\right| \\
&= |\frac{1}{|D_G|M_G^2}\sum_g\sum_{h(\neq g)} k(d_{gh})e_i^\top\,\mathsf{E}(Z_g^\top Z_h)e_j - \frac{1}{GM_G^2}\sum_g\sum_{h(\neq g)} e_i^\top\,\mathsf{E}\left[Z_g^\top Z_h\right]e_j| \\
&= \frac{1}{|D_G|M_G^2}\sum_g\sum_{h(\neq g)} |k(d_{gh})-1|e_i^\top\,\mathsf{E}\left(Z_g^\top Z_h\right)e_j| \\
&= o_p(1)
\end{aligned}
$$

To prove each of I_1, I_2, I_3 is $o_p(1)$, we define $p_{gh} = Z_g^\top Z_h - \mathsf{E}\left(Z_g^\top Z_h\right)$.

Step 1 We handle firstly I_1, $I_1 \leq |M_G^{-2}|D_G|^{-1}\sum_g\sum_h e_i^\top(\hat{Z}_g - Z_g)^\top Z_h e_j K(d_{gh})|$ $+|M_G^{-2}|D_G|^{-1}\sum_g\sum_h e_i^\top(\hat{Z}_h - Z_h^\top)(\hat{Z}_g - Z_g)e_j K(d_{gh})| + |M_G^{-2}|D_G|^{-1}\sum_g\sum_h e_i^\top Z_g^\top(\hat{Z}_h - Z_h)e_j K(d_{gh})| \stackrel{\text{def}}{=} I_{11} + I_{12} + I_{13}$. Assume that $\hat{Z}_g - Z_g = (\nabla\mathbf{m}_g^\top\mathbf{W}_g^{-1}(\hat{\mathbf{u}}_g - \mathbf{u}_g))$ $= (\nabla\mathbf{m}_g^\top\mathbf{W}_g^{-1}C_g\Delta_g)$. $\sum_g|C_g|_2 = \mathcal{O}_p(LG)$ and $|\Delta_g|_2 = \mathcal{O}_p(G^{-1/2})$, where recall that $|.|_2$ defined the Euclidean norm of a matrix. Thus we have $I_{11} = M_G^{-2}|D_G|^{-1}\sum_g\sum_h|e_i^\top(\hat{Z}_g - Z_g)^\top Z_h e_j K(d_{gh})| = M_G^{-2}|D_G|^{-1}\sum_g\sum_h|e_i^\top\nabla\mathbf{m}_g\mathbf{W}_g^{-1}C_g\Delta_g Z_h e_j K(d_{gh})|$ $\leq M_G^{-2}|D_G|^{-1}\sum_g|e_i^\top\nabla\mathbf{m}_g\mathbf{W}_g^{-1}C_g\Delta_g|_2|\max_{\rho(h,g)\leq h_g}Z_h e_j|_2$ $\leq M_G^{-2}|D_G|^{-1}\sum_g|e_i^\top\nabla\mathbf{m}_g\mathbf{W}_g^{-1}C_g|_2|\Delta_g|_2|\max_{h:\rho(h,g)\leq h_g}Z_h e_j|_2 = \mathcal{O}_p(h_g^{d/q'}L^{d/q'}/\sqrt{G})$, given the fact that the number of observations lying in a h_g ball is $\{\sharp h : \rho(h,g)\leq h_g\} \lesssim Ch_g^d L^d$, $(\mathsf{E}\,|\max_{h:\rho(h,g)\leq h_g}Z_h|^2)^{1/2} \leq Ch_g^{d/q'}\max_{h:\rho(h,g)\leq h_g}\|Z_h\|_{q'}L^{d/q'}$, where from B.2) we have that $\max_{h:\rho(h,g)\leq h_g}\|Z_h\|_{q'} \leq CL^2$.

$I_{12} = M_G^{-2}|D_G|^{-1}\sum_g\sum_h e_i^\top(\hat{Z}_g - Z_g)^\top(\hat{Z}_h - Z_h)K(d_{gh})e_j = M_G^{-2}|D_G|^{-1}\sum_g\sum_h e_i^\top(\hat{Z}_g - Z_g)^\top(\hat{Z}_h - Z_h)K(d_{gh})e_j \leq M_G^{-2}|D_G|^{-1}\sum_g\sum_h e_i^\top\nabla\mathbf{m}_g^\top\mathbf{W}_g^{-1}C_g\Delta_g(\nabla\mathbf{m}_h^\top\mathbf{W}_h^{-1}C_h\Delta_h)^\top K(d_{gh})e_j$ $\leq |e_i^\top\nabla\mathbf{m}_g^\top\mathbf{W}_g^{-1}C_g|_2|\Delta_g|_2|\Delta_h^\top|_2|C_h^\top\mathbf{W}_h^{-1}\nabla\mathbf{m}_h^\top e_j|_2 = \mathcal{O}_p(h_g^{d/q'}|D_G|^{-1}L^{d/q'}L^2)$. The rate of I_{13} is similarly derived as I_{11}. Then from B.1) $I_1 = o_p(1)$.

Step 2 Now we look at the variance case I_2,

$I_2 = \frac{1}{|D_G|M_G^2}\sum_g\sum_{h\neq g}|k(d_{gh})e_i^\top\left[Z_g^\top Z_h - \mathsf{E}\left(Z_g^\top Z_h\right)\right]e_j| = o_p(1)$.

As we can see that $\mathsf{E}\,I_2 = 0$ and we need to study

$\mathsf{Var}(I_2) = |D_G|{-}2M_G^{-4}\sum_{g1}\sum_{h1}\sum_{g2}\sum_{h2}k(d_{g1h1})k(d_{g2h2})$

$\mathsf{E}\{e_i^\top\left[Z_{g1}^\top Z_{h1} - \mathsf{E}\left(Z_{g1}^\top Z_{h1}\right)\right]e_j e_i^\top\left[Z_{g2}^\top Z_{h2} - \mathsf{E}\left(Z_{g2}^\top Z_{h2}\right)\right]e_j\}$.

Denote $p_{g1h1,ij} \stackrel{\text{def}}{=} e_i^\top\left[Z_{g1}^\top Z_{h1} - \mathsf{E}\left(Z_{g1}^\top Z_{h1}\right)\right]e_j$.

According to assumption A.1)- A.8), the underlying random field $\tilde{\varepsilon}$ with $\alpha-$ mixing $\tilde{\alpha}(u,v,r) \leq (uL+vL)^{\tau}\hat{\alpha}(r)$, with $\tau \geq 0$. We need to verify the NED property of $p_{g1h1,ij}$.

From Lemma 2, the NED property of $Z_g = s_g(\theta^0, \gamma^0)$ with $\psi(m)$ and NED constant bounded by $L^2 d_g C'$, where C' is a bound for the $\max_{i,j} \| |\nabla m_g(\theta^0) \mathbf{W}_{gij}^{-1}(\theta^0, \gamma^0)|_2 \|_4$. According to the definition of Bartlett kernel we focus on the pairs with $\rho(h1, g1) \leq h_g$ and $\rho(h2, g2) \leq h_g$, we see that $p_{g1h1,ij}$, $\| Z_{h1}^{\top} Z_{g_1} - \mathsf{E}[Z_{h1}^{\top} Z_{g1} | \mathcal{F}_{h1}(s+h_g)] \| \leq (\| |Z_{h1}|_2 \|_4 d_{g1} \vee \| |Z_{g1}|_2 \|_4 d_{h1}) \psi(s)$.

Therefore $p_{g1h1,ij}$ would be also L_2 NED with $\psi(m) = \tilde{\psi}(m + h_g)$, with $m > h_g$.

From the property of the L_2 NED, following from Lemma B.3 of Jenish and Prucha (2012),

$\mathrm{Cov}(p_{g1h1,ij}, p_{g2h2,ij}) = \mathsf{E}\{e_i^{\top} \left[Z_{g1}^{\top} Z_{h1} - \mathsf{E}\left(Z_{g1}^{\top} Z_{h1} \right) \right] e_j e_i^{\top} \left[Z_{g2}^{\top} Z_{h2} - \mathsf{E}\left(Z_{g2}^{\top} Z_{h2} \right) \right] e_j \}$
$\leq \|p_{g1h1,ij}\|_{2+\delta} \{ C_1 \|p_{g1h1,ij}\|_{2+\delta} [\rho(g1,g2)/3]^{d\tau^*} \hat{\alpha}^{\delta/(2+\delta)} (\rho(g1,g2)/3) + C_2 \tilde{\psi}([\rho(g1,g2)]/3) \}$,
where $\tau^* \stackrel{\text{def}}{=} \delta\tau/(2+\delta)$.

So $\mathsf{Var}(I_2) = M_G^{-4} |D_G|^{-2} h_g^{2d} L^{2d} \sum_{g1} \sum_{g2} \max_{h1,h2} k(d_{g1h1}) k(d_{g2h2}) \, \mathsf{E}\{ e_i^{\top} \left[Z_{g1}^{\top} Z_{h1} - \mathsf{E}\left(Z_{g1}^{\top} Z_{h1} \right) \right] e_j$
$e_i^{\top} \left[Z_{g2}^{\top} Z_{h2} - \mathsf{E}\left(Z_{g2}^{\top} Z_{h2} \right) \right] e_j \} \leq M_G^{-4} |D_G|^{-2} h_g^{2d} L^{2d} \max_{h1,h2} \sum_{g1,g2} \|p_{g1h1,ij}\|_{2+\delta} \{ C_1 \|p_{g1h1,ij}\|_{2+\delta}$
$\{\rho(g1,g2)/3\}^{d\tau^*} \hat{\alpha}^{\delta/(2+\delta)} (\rho(g1,g2)/3) + C_2 \tilde{\psi}(\rho(g1,g2)/3) \}$
$\leq M_G^{-4} G^{-2} h_g^{2d} \max_{h1,h2} \sum_{g1} \sum_{r=1}^{\infty} \sum_{g2 \in \{g_2 : \rho_{g1,g2} \in [r, r+1)\}} \|p_{g1h1,ij}\|_{2+\delta} \{ C_1 \|p_{g1h1,ij}\|_{2+\delta} [\rho(g1,g2)/3]^{d\tau^*}$
$\hat{\alpha}^{\delta/(2+\delta)} (\rho(g1,g2)/3) + C_2 \psi(([\rho(g1,g2)]/3 - h_g))_+ \}$
$\leq |D_G|^{-1} h_g^{2d} L^{2d} \sum_{r=1}^{\infty} \{ C_1' r^{(d\tau^*+d)-1} \hat{\alpha}^{\delta/(2+\delta)}(r) + C_2 r^{d-1} \psi((r-h_g)_+) \}$. From B.4) we assume that $h_g^{2d} L^{2d} \sum_{r=1}^{\infty} r^{(d\tau^*+d)-1} \hat{\alpha}^{\delta/(2+\delta)}(r) = \mathcal{O}(G)$, and $h_g^{2d} \sum_{r=1}^{\infty} L^{2d} r^{d-1} \psi((r - h_g)_+) = \mathcal{O}(G)$, then we have $\mathsf{Var}(I_2) = o(1)$.

Step 3

According to B.4), $|k(d_{gh}) - 1| \leq C_k |\rho(g,h)/h_g|^{\rho_K}$ for $\rho(g,h)/h_g \leq 1$ for some constant $\rho_k \geq 1$ and $0 < C_k < \infty$.

We handle the bias term I_3,
$M_G^{-2} |D_G|^{-1} \sum_g \sum_h |e_i^{\top} (k(\rho(g,h)/h_g) - 1) \, \mathsf{E}(Z_g^{\top} Z_h) e_j|$
$\leq M_G^{-2} |D_G|^{-1} \sum_g \sum_h C_k |\rho(g,h)/h_g|^{\rho_k} e_i^{\top} \, \mathsf{E}(Z_g^{\top} Z_h) e_j$
$\leq M_G^{-2} |D_G|^{-1} \sum_g \sum_h |\rho(g,h)/h_g|^{\rho_k} \|e_i^{\top} Z_g^{\top}\| \|Z_h e_j\|$.

Also according B.4), $M_G^{-2} |D_G|^{-1} \sum_g \sum_h |\rho(g,h)/h_g|^{\rho_k} \|e_i^{\top} Z_g^{\top}\| \|Z_h e_j\|$ is $o(1)$.

7.6 Two special cases

To justify the NED assumptions in A.2), we now verify the two L_2 NED properties in our example. (L_4 NED can be similarly verified.) In particular we would like to analyze how the underlying assumptions of the data innovation processes would induce the assumption of A.1).

7.6.1 Poisson Regression/ Negative Binomial

The focused model is $y_{n,i}$s are poisson counts observations, $\mathsf{E}(y_{n,i}|x_{n,i}, v_{n,i}) = \exp(x_{n,i}^\top \beta) v_{n,i}$. We suppose that $v_{i,n} = g(\eta_{i,n})$, where $g(.)$ is twice continuously differentiable function. For example $g(x) = \exp(x)$ and then $\mathsf{E}(y_{n,i}|x_{n,i}, v_{n,i}) = \exp(x_{n,i}^\top \beta + \eta_{n,i})$, and $x_{n,i}$ are controls with $p \times 1$ dimension.

We assume that $\eta_{n,i}$ follows a spatial autoregressive model. Namely $\eta_{n,i} = \lambda \sum_{j=1}^n w_{n,ij}\eta_{n,j} + \epsilon_{n,i}$. Suppose $\eta_n = \lambda W \eta_n + \epsilon_n$, and $\eta_n = (I - \lambda W)^{-1}\epsilon_n$, define $[a_{ij}] = (I - \lambda W)^{-1}$.

Then we have

$$v_{n,i} = g(\sum_{j=1} a_{ij}\epsilon_{n,j}).$$

For the moment we assume the decomposition: $y_{n,i} = \mathsf{E}(y_{n,i}|x_{n,i}, v_{n,i}) + \varepsilon_{n,i}$.

Assume that $\{\xi_{n,i} = (x_{n,i}, \epsilon_{n,i}, \varepsilon_{n,i})\}$ are mixing random field.

We now establish that $Y = \{y_{n,i}, s_i \in D_n, n \geq 1\}$ is uniform L_2 NED on $\xi = \{\xi_{n,i}, s_i \in D_n, n \geq 1\}$. Define $\mathcal{F}_{n,i}(s) = \sigma(\xi_{n,j} : j \in D_n, \rho(i,j) \leq s)$.

It can be seen that, for any $i \in D_n$,

$$\begin{aligned}\tilde{y}_{n,i} = y_{n,i} - \mathsf{E}(y_{n,i}|\mathcal{F}_{n,i}(s)) &= \exp(x_{n,i}^\top\beta)v_{n,i} + \varepsilon_{n,i} - \exp(x_{n,i}^\top\beta)\,\mathsf{E}(v_{n,i}|\mathcal{F}_{n,i}(s)) - \varepsilon_{n,i} \\ &= [v_{n,i} - \mathsf{E}\{v_{n,i}|\mathcal{F}_{n,i}(s)\}]\exp(x_{n,i}^\top\beta)\end{aligned}$$

As $v_{n,i} - \mathsf{E}(v_{n,i}|\mathcal{F}_{n,i}(s)) = g(\sum_j a_{ij}\epsilon_{n,j}) - \mathsf{E}\{g(\sum_j a_{ij}\epsilon_{n,j})|\mathcal{F}_{n,i}(s)\}$.

Taylor expansion to the first order yield,

$$g(\sum_j a_{ij}\epsilon_{n,j}) - \mathsf{E}\{g(\sum_j a_{ij}\epsilon_{n,j})|\mathcal{F}_{n,i}(s)\} = g'(\tilde{a}) \sum_{j \in B^c(s)} a_{ij}\epsilon_{n,j}, \tag{81}$$

where $\tilde{a}$ is a point between 0 and $\sum_j a_{ij}\epsilon_{n,j}$, $B^c(s)$ is the set of j with $\rho(i,j) \geq s$. Thus we have

$$(\mathsf{E}\,|\tilde{y}_{n,i}|^2)^{1/2} \leq C \sum_{j \in B^c(s)} |a_{ij}|, \tag{82}$$

where we assume that $\|g'(\tilde{a})\epsilon_{n,j}\|_2$ is uniformly bounded by C. Also we require that $\text{limsup}_{s\to\infty} \sup_{i \in D_n} \sum_{j \in B^c(s)} |a_{ij}| \to 0$. The proof is completed.

7.6.2 Probit Model

We now prove the case of probit model,

$$y_{n,i} = \mathbf{I}(y^*_{n,i} > 0)$$
$$y^*_{n,i} = x^\top_{n,i}\beta + e_{n,i}.$$

And $e_{n,i} = \lambda \sum_j w_{n,ij} e_{n,j} + v_{n,i}$. We now establish that $Y = \{y_{n,i}, s_i \in D_n, n \geq 1\}$ ($\|y^*_{n,i}\|_2 < \infty$) is L_2 NED on $\xi = \{(x_{n,i}, e_{n,i}), s_i \in D_n, n \geq 1\}$. Thus again similar to the previous case we can denote $e_{n,i} = \sum_j a_{ij} v_{n,i}$, where a_{ij} are the matrix entries of $(I - \lambda W)^{-1}$.

Proof. First of the latent process is $\{y^*_{n,i}\}$ is a special case of the Cliff-Ord type of process, and therefore would be L_2- unform NED if $\text{limsup}_{s\to\infty} \sup_{i\in D_n} \sum_{j\in B^c(s)} |a_{ij}| \to 0$, and $\|v_{n,i}\|_{r'} \leq \infty$, $r' = 2$.
For any $\epsilon > 0$, define the event $B = \{|y^*_{n,i}| < \epsilon, |\,\mathsf{E}[y^*_{n,i}|\mathcal{F}_{n,i}(s)]| < \epsilon\}$. Using $|\mathbf{I}(x_1 \geq 0) - \mathbf{I}(x_2 \geq 0) \leq \frac{|x_1 - x_2|}{\epsilon}\mathbf{I}(x_1 > \epsilon \text{ or } x_2 > \epsilon) + \mathbf{I}(x_1 < \epsilon, x_2 < \epsilon)$, we have

$$\begin{aligned}
&\|y_{n,i} - \mathsf{E}[y_{n,i}|\mathcal{F}_{n,i}(s)]\| = \|\mathbf{I}(y^*_{n,i} \geq 0) - \mathsf{E}[\mathbf{I}(y^*_{n,i} \geq 0)|\mathcal{F}_{n,i}(s)]\| \\
&\leq \|\mathbf{I}(y^*_{n,i} \geq 0) - \mathbf{I}\{\mathsf{E}[y^*_{n,i}|\mathcal{F}_{n,i}(s)] \geq 0\}\| = \left\{\mathsf{E}\left|\mathbf{I}(y^*_{n,i} \geq 0) - \mathbf{I}\{\mathsf{E}[y^*_{n,i}|\mathcal{F}_{n,i}(s)] \geq 0\}\right|^2\right\}^{\frac{1}{2}} \\
&\leq \left\{\frac{1}{\epsilon^2}\int_{B^c}\left|y^*_{n,i} - \mathsf{E}[y^*_{n,i}|\mathcal{F}_{n,i}(s)]\right|^2 d\mathbb{P} + \int_B d\mathbb{P}\right\}^{\frac{1}{2}} \\
&\leq \left\{\frac{1}{\epsilon^2}\int_{B^c}\left|y^*_{n,i} - \mathsf{E}[y^*_{n,i}|\mathcal{F}_{n,i}(s)]\right|^2 d\mathbb{P}\right\}^{\frac{1}{2}} + \left\{\int_B d\mathbb{P}\right\}^{\frac{1}{2}} \\
&\leq \frac{1}{\epsilon}\|y^*_{n,i} - \mathsf{E}[y^*_{n,i}|\mathcal{F}_{n,i}(s)]\|_2 + \pi_4\epsilon^{1/2}, \quad \text{for some constant } \pi_4 > 0,
\end{aligned}$$

where the first inequality is based on Therorem 10.12 of Davidson (1994) by taking $\mathbf{I}\{\mathsf{E}[y^*_{n,i}|\mathcal{F}_{n,i}(s)] \geq 0\}$ as an approximation of $\mathbf{I}(y^*_{n,i} \geq 0)$ with measure $\mathcal{F}_{n,i}(s)$. When taking $\epsilon = \|y^*_{n,i} - \mathsf{E}[y^*_{n,i}|\mathcal{F}_{n,i}(s)]\|^q, 0 < q < 1$, when ϵ converges to 0, both terms converge to 0 at a slower rate than $\|u^*_{n,i} - \mathsf{E}[y^*_{n,i}|\mathcal{F}_{n,i}(s)]\|$, therefore, the process $\{(y_{n,i})\}_{i=1}^n$ is uniform L_2 NED. ■

7.7 Exponential family

For parameter $\theta \in \mathbf{R}^p$, and a random variable X. $f(x,\theta) = h(x)\text{exp}\{\theta^\top T(x) - A(\theta)\}$, where $A(\theta) = \log \int h(x)\text{exp}\{\theta^\top T(x)\}dF(x)$ is the *cumulant function*, and $T(x)$ is referred to as the *sufficient statistics*. In particular, we know that $\partial A(\theta)/\partial\theta = \mathsf{E}(T(X))$

and $\partial A(\theta)/\partial\theta\partial\theta^\top = \mathsf{Var}(T(X)) = I(\theta)$ are regarded as the Fisher information matrix.

Suppose y_i is following an exponential family condition on x_i, then the conditional mean and conditional variance function will be both expressed as known function, which is the first and the second derivative of the cumulants generating function $A(\mu_i)$. In particular $\mathsf{E}(T(y_i)) = \partial A(\mu_i)/\partial\mu_i|_{\mu_i=v(x_i^\top\theta)}$, and the variance covariance $\mathsf{Var}(T(y_i)) = \partial A(\mu_i)/\partial\mu_i\partial\mu_i^\top|_{\mu_i=v(x_i^\top\theta)}$, where $v(\cdot)$ is a link function. Notably the variance covariance function is thus treated as a known function related to the conditional mean in this case as they are both related to $A(\cdot)$.

References

Anderson, J. E. and Van Wincoop, E. (2003). Gravity with gravitas: a solution to the border puzzle, *American economic review* **93**(1): 170–192.

Bhat, C. R., Varin, C. and Ferdous, N. (2010). A comparison of the maximum simulated likelihood and composite marginal likelihood estimation approaches in the context of the multivariate ordered-response model, *Maximum simulated likelihood methods and applications*, Emerald Group Publishing Limited, pp. 65–106.

Bloom, N., Schankerman, M. and Van Reenen, J. (2013). Identifying technology spillovers and product market rivalry, *Econometrica* **81**(4): 1347–1393.

Bolthausen, E. (1982). On the central limit theorem for stationary mixing random fields, *The Annals of Probability* pp. 1047–1050.

Brown, J. R., Ivković, Z., Smith, P. A. and Weisbenner, S. (2008). Neighbors matter: Causal community effects and stock market participation, *The Journal of Finance* **63**(3): 1509–1531.

Cameron, A. C. and Trivedi, P. K. (1986). Econometric models based on count data. comparisons and applications of some estimators and tests, *Journal of applied econometrics* **1**(1): 29–53.

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

Conley, T. G. and Molinari, F. (2007). Spatial correlation robust inference with errors in location or distance, *Journal of Econometrics* **140**(1): 76–96.

Cressie, N. (1992). Statistics for spatial data, *Terra Nova* **4**(5): 613–617.

Davidson, J. (1994). *Stochastic limit theory: An introduction for econometricians*, OUP Oxford.

Gourieroux, C., Monfort, A. and Trognon, A. (1984). Pseudo maximum likelihood methods: Theory, *Econometrica: Journal of the Econometric Society* pp. 681–700.

Jenish, N. and Prucha, I. R. (2009). Central limit theorems and uniform laws of large numbers for arrays of random fields, *Journal of Econometrics* **150**(1): 86–98.

Jenish, N. and Prucha, I. R. (2012). On spatial processes and asymptotic inference under near-epoch dependence, *Journal of Econometrics* **170**(1): 178–190.

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

Lee, L.-F. (2004). Asymptotic distributions of quasi-maximum likelihood estimators for spatial autoregressive models, *Econometrica* **72**(6): 1899–1925.

Liang, K.-Y. and Zeger, S. L. (1986). Longitudinal data analysis using generalized linear models, *Biometrika* **73**(1): 13–22.

Lu, C. and Wooldridge, J. M. (2017). Quasi-generalized least squares regression estimation with spatial data, *Economics Letters* **156**: 138–141.

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

Prentice, R. L. (1988). Correlated binary regression with covariates specific to each binary observation, *Biometrics* pp. 1033–1048.

Silva, J. S. and Tenreyro, S. (2006). The log of gravity, *The Review of Economics and statistics* **88**(4): 641–658.

Tinbergen, J. (1962). An analysis of world trade flows, *Shaping the world economy* **3**: 1–117.

Varin, C., Reid, N. and Firth, D. (2011). An overview of composite likelihood methods, *Statistica Sinica* pp. 5–42.

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

Wooldridge, J. M. (2010). *Econometric analysis of cross section and panel data*, MIT press.

Zeger, S. L. and Liang, K.-Y. (1986). Longitudinal data analysis for discrete and continuous outcomes, *Biometrics* pp. 121–130.

IRTG 1792 Discussion Paper Series 2020

For a complete list of Discussion Papers published, please visit http://irtg1792.hu-berlin.de.

001 "Estimation and Determinants of Chinese Banks' Total Factor Efficiency: A New Vision Based on Unbalanced Development of Chinese Banks and Their Overall Risk" by Shiyi Chen, Wolfgang K. Härdle, Li Wang, January 2020.

002 "Service Data Analytics and Business Intelligence" by Desheng Dang Wu, Wolfgang Karl Härdle, January 2020.

003 "Structured climate financing: valuation of CDOs on inhomogeneous asset pools" by Natalie Packham, February 2020.

004 "Factorisable Multitask Quantile Regression" by Shih-Kang Chao, Wolfgang K. Härdle, Ming Yuan, February 2020.

005 "Targeting Cutsomers Under Response-Dependent Costs" by Johannes Haupt, Stefan Lessmann, March 2020.

006 "Forex exchange rate forecasting using deep recurrent neural networks" by Alexander Jakob Dautel, Wolfgang Karl Härdle, Stefan Lessmann, Hsin-Vonn Seow, March 2020.

007 "Deep Learning application for fraud detection in financial statements" by Patricia Craja, Alisa Kim, Stefan Lessmann, May 2020.

008 "Simultaneous Inference of the Partially Linear Model with a Multivariate Unknown Function" by Kun Ho Kim, Shih-Kang Chao, Wolfgang K. Härdle, May 2020.

009 "CRIX an Index for cryptocurrencies" by Simon Trimborn, Wolfgang Karl Härdle, May 2020.

010 "Kernel Estimation: the Equivalent Spline Smoothing Method" by Wolfgang K. Härdle, Michael Nussbaum, May 2020.

011 "The Effect of Control Measures on COVID-19 Transmission and Work Resumption: International Evidence" by Lina Meng, Yinggang Zhou, Ruige Zhang, Zhen Ye, Senmao Xia, Giovanni Cerulli, Carter Casady, Wolfgang K. Härdle, May 2020.

012 "On Cointegration and Cryptocurrency Dynamics" by Georg Keilbar, Yanfen Zhang, May 2020.

013 "A Machine Learning Based Regulatory Risk Index for Cryptocurrencies" by Xinwen Ni, Wolfgang Karl Härdle, Taojun Xie, August 2020.

014 "Cross-Fitting and Averaging for Machine Learning Estimation of Heterogeneous Treatment Effects" by Daniel Jacob, August 2020.

015 "Tail-risk protection: Machine Learning meets modern Econometrics" by Bruno Spilak, Wolfgang Karl Härdle, October 2020.

016 "A data-driven P-spline smoother and the P-Spline-GARCH models" by Yuanhua Feng, Wolfgang Karl Härdle, October 2020.

017 "Using generalized estimating equations to estimate nonlinear models with spatial data" by Cuicui Lu, Weining Wang, Jeffrey M. Wooldridge, October 2020.

IRTG 1792, Spandauer Strasse 1, D-10178 Berlin
http://irtg1792.hu-berlin.de

This research was supported by the Deutsche Forschungsgemeinschaft through the IRTG 1792.