LINEAR AND NONLINEAR ESTIMATION WITH SPATIAL DATA

By

Cuicui Lu

A DISSERTATION

Submitted to
Michigan State University
in partial fulfillment of the requirements
for the degree of

Economics – Doctor of Philosophy

2013

ABSTRACT

LINEAR AND NONLINEAR ESTIMATION WITH SPATIAL DATA

By

Cuicui Lu

In some economic situations, observations are cross-sectionally correlated. One example of cross-sectional correlation is spatial correlation, which means the correlation comes from the spatial closeness of different individuals. Spillover effect, externalities, network issues and so on are common causes for spatial correlation. For example, a supply shock in one region will result in production shocks in the regions nearby. This type of correlation reflects the correlations among individuals' unobservables.

In Chapter 1, I study a linear regression model with a spatially correlated error term. Most current literature in econometrics assumes cluster sampling (independence between different clusters) in the population; however, this could be easily violated. I study the case in which spatial correlation exists between each pair of observations without assuming independent clusters. Generalized least squares (GLS) can be applied to the cross-sectional dimension but it is hard to account for all pairwise correlations for a large sample of spatial data. It is because the calculation of the huge error covariance matrix generally needs large computer memory. Instead I use a pseudo generalized least squares (PGLS) approach, which means it is a GLS procedure but uses a "tapered" error covariance matrix. Data could be divided into groups according to natural geographic areas, only correlations within groups are accounted for while ignoring the correlations between groups. Since correlations within groups account for most of the correlations among observations, the resulting PGLS estimator will not lose much efficiency compared to GLS. The PGLS estimator is consistent, asymptotically normal, and computationally easier than GLS. A spatial heteroskedasticity and autocorrelation consistent (HAC) covariance estimator for PGLS which is robust to both heteroskedasticity and spatial correlation is provided. Monte Carlo simulations show that PGLS becomes more efficient than ordinary least squares (OLS) as spatial correlation increases.

Chapter 2 studies nonlinear estimation with spatial data.Generalized estimating equations (GEE) is applied to cross section data with spatial correlations in nonlinear models. I use a partial quasi-maximum likelihood estimator (PQMLE) in the first step and use GEE approach in the second step. Given some regularity conditions and assumptions, the asymptotic distribution of the two-step estimator is derived in the framework of M-estimation. I use a Probit model for binary data with a latent spatially error and a Poisson model for count data with a multiplicative spatial error to demonstrate the GEE procedures. As the spatial correlations in the underlying error term increase, those in the dependent variable also increase. Monte Carlo simulations show efficiency comparison of the PQMLE and GEE. The results show that correctly modeling the structure of the working correlation matrix is important in nonlinear models, which is quite different from the linear model. In addition, as spatial correlation increases, more efficiency estimation can be obtained by the GEE approach.

Chapter 3 studies conditions for the Numerical Equality of the OLS, GLS and Amemiya-Cragg Estimators. Conditions under which the ordinary least squares (OLS) and generalized least squares (GLS) estimators are equal are well known. This chapter extends these results in two ways. First, it give conditions under which GLS based on one assumed error variance matrix equals GLS based on a different assumed variance matrix. Second, it give conditions under which GLS equals the GMM estimator of Amemiya (1983) and Cragg (1983).

To my mother and father

ACKNOWLEDGMENTS

I was born in a beautiful town on the east coast of China. As I grew up, I witnessed the dramatic improvement in people's life of this town, as well as rapid growth of the Chinese economy. Thus I became very interested in the comparatively new "market economy" in China. I realized in high school that my dream was to become an economist. And after college, I came to Michigan State University for my graduate study. Behind me the whole time were my kind and diligent parents, who show their deep love to me and have faith in me as always. I am very grateful to my parents for supporting my education both financially and mentally, especially during graduate school.

At Michigan State University, I have had great fortune to be taught by Jeffrey Wooldridge, Peter Schmidt, Timothy Vogelsang, and Young Chae Lim. First and foremost, I thank my adviser, Jeffrey Wooldridge, for years of professional guidance. He gave me numerous suggestions and comments on my research. He also taught me how to do independent research. Peter Schmidt taught me how to do rigorous proofs. I learned very interesting time series knowledge from Tim Vogelsang. I also gained a deeper understanding of spatial statistics from Young Chae Lim. I would like to thank all of them who have been guiding me during my doctoral study.

I thank the economics department for giving me the opportunity to be a teaching assistant. I also thank Jeff, Tim and the University for providing conference funding. My fellow classmates and the friends I have made are treasures to me. I appreciate their help in both research and life during my study.

I am glad I am able to explore econometrics, which is an exciting area worthy of effort. I am on my way to my dream.

TABLE OF CONTENTS

LIST OF TABLES

Chapter 1

Pseudo Generalized Least Squares Regression Estimation with Spatial Data

1.1 Introduction

In some economic situations, observations are cross-sectionally correlated. One instance of cross-sectional correlation is spatial correlation, which means the correlations come from the spatial closeness of different individuals. Spillover effect, externalities, network issues, and so on are common causes for spatial correlation. For example, a supply shock in one region will result in production shocks in the regions nearby. This type of correlation reflects the correlations among individuals' unobservables. Similar to time series correlation which depends on the "distance" in time[1], spatial correlation depends on the "distance" in space. In economic study, spatial data are commonly observed in the two-dimensional Euclidean space $\mathcal{R}^2$. Each observation resides in the specific location on Earth's surface. We would expect the larger the distance is between two individuals, the smaller their correlation. A legitimate distance measure is the physical distance. In economics, a distance measure has

[1] The time passed between two time points.

a wider meaning. Conley (1999) uses a metric of "Economic Distance" to measure cross-sectional dependence. For example, *transportation cost* and *whether two countries are in a free trade area* are economic distances. Thus spatial correlation based on spatial distance is more complicated than time series dependence.

Researchers emphasize the importance of correct distance measurement. Measurement errors are often considered in the estimation, for example, Conley (1999) and Kelejian and Prucha (2007). Here we do not discuss the measurement errors problem and assume that we have correctly measured distances. Clustered data have a very similar structure to spatial data. The difference is that data in one cluster are correlated through a common cluster effect, and they do not depend on distances. Due to the complicated correlation relationship among observations, for spatial data, accounting for all pairwise correlations in the estimation is very difficult when the sample size is large. If the sample size is 100, there are 4950 pairwise correlations to account for. Although one can still do the estimation, ignoring the pairwise correlations and getting consistent estimators, one can improve estimation efficiency by using more information. Spatial data are commonly collected from different geographic areas such as unemployment rates in different states, foreign investment amounts in different cities in a developing country, and trade volume between different US states and Canadian provinces. Thus there is usually a natural division of groups for the data. Since observations far away have small or no correlations, we can account for the correlations only within the same groups. By doing so, we actually account for most of the total correlations. Therefore, in this paper, I propose a pseudo generalized least squares estimator (PGLS). It can account for most correlations and is more efficient than the usual OLS estimator, while the computation burden is reduced by dividing data into groups and only using the information within groups.

In spatial statistics, "increasing domain" and "fixed domain" asymptotics are popularly used (Cressie 1993). Under increasing domain asymptotics, the sampling space increases without bound, while the minimum distance between locations is bounded below by a positive constant. Under fixed domain asymptotics, the sampling space is fixed and bounded, and sampling locations become increasingly dense within this region. Zhang (2004) shows that some parameters cannot be consistently estimated by maximum likelihood estimation (MLE) under the fixed domain asymptiotics. Considering the properties of economic data, which can be obtained by sampling in a large Euclidean space, we will focus on increasing domain asymptotics in this paper. That is, given that the number in each group is fixed, when the sampling space gets larger, the number of the groups increases as the sample size increases.

In this chapter, Section 1.2 presents a linear regression model with possible spatial correlation in the error term. Section 1.3 presents the OLS, GLS and PGLS estimation methods. Section 1.4 discusses how to estimate the spatial correlation parameter and provides consistent covariance estimators. Section 1.5 explores quasi-MLE method for estimating a linear regression model. In Section 1.6, Monte Carlo simulation results show the advantages of the PGLS estimation procedure. Section 1.7 provides the conclusions.

1.2 A Linear Regression Model

Let $\mathcal{S}$ be the space the population resides. s_i , $i = 1, 2, ...$ represents a location in $\mathcal{S}$. Let d_{ij} be the distance between location $s_i \in \mathcal{S}$ and location $s_j \in \mathcal{S}$. The space can be one dimensional (like time series), two-dimensional (a Euclidean space) or multidimensional. Let $(\mathbf{x}_i, y_i)$ denote the data point sampled at location s_i. Let u_i denote the underlying unobservable at

s_i. Consider a linear regression model:

$$y_i = \mathbf{x}_i \beta + u_i, \quad i = 1, 2, ..., N, \tag{1.1}$$

where $\mathbf{x}_i$ is $1 \times K$ regressors with the first element $x_{i1} = 1$, and $\beta \equiv (\beta_1, \beta_2, ..., \beta_K)'$ is a $K \times 1$ unknown vector of parameters. In a matrix form, the above equation reads

$$\mathbf{y} = \mathbf{X}\beta + \mathbf{u}, \tag{1.2}$$

where $\mathbf{y} = (y_1, y_2, ..., y_N)'$, $\mathbf{X}$ is an $N \times K$ matrix, with the ith row equal to $\mathbf{x}_i$, and $\mathbf{u} = (u_1, u_2, ..., u_N)'$.

Generally we do not know the specific form of spatial correlation. In this paper, we focus on spatial correlations in the error term. The error term exhibits spatial correlation in the sense that

$$\text{Var}\left(\mathbf{u}|\mathbf{X}, \mathbf{D}\right) = \mathbf{\Omega}\left(\mathbf{D}, \lambda\right), \tag{1.3}$$

where $\mathbf{D} = \left\{d_{ij}, i, j = 1, 2, ..., N\right\}$ which contains all pairwise distances between observations. λ is a vector of variance covariance parameters. That is, the spatial correlations between different observations are commonly assumed to depend on the pairwise distances and some other fixed parameters. Note that under random sampling, $\mathbf{\Omega}$ would be a scalar matrix with all off-diagonal elements equal to zero. For simplicity in what follows, I often drop the conditioning on the explanatory variables and location indicators.

1.2.1 Spatial Error Model

Spatial error model is one of the common models based on underlying random fields. There are different cases of spatial error model. Instead of writing the specific error covariance matrix as a function of distance matrix $\mathbf{D}$, we can use a spatial weight matrix $\mathbf{W}$ which is defined as a function of $\mathbf{D}$. Let w_{ij} be the ijth element of $\mathbf{W}$. Let $\epsilon = (\varepsilon_1, \varepsilon_2, ..., \varepsilon_i, ..., \varepsilon_N)'$ be a vector of spatial white noise with mean zero and constant variance.

The spatial autoregressive (SAR) error model is most widely used in spatial econometrics (Anselin, 1988; Anselin *et al.*, 2004). Its error term is modeled as simultaneous autoregressive random field,

$$u_i = \rho \sum_{j=1}^{N} w_{ij} u_j + \varepsilon_i, \tag{1.4}$$

In a matrix form, the above equation reads,

$$\begin{aligned} \mathbf{u} &= \rho \mathbf{W} \mathbf{u} + \epsilon, & (1.5) \\ \mathbf{\Omega}(\lambda) &= \sigma^2 (I - \rho \mathbf{W})^{-1} \left(I - \rho \mathbf{W}'\right)^{-1}. & (1.6) \end{aligned}$$

In the spatial moving average (SMA) error model, the error term is modeled as moving average random fields,

$$u_i = \lambda \sum_{j=1}^{N} w_{ij} \varepsilon_j. \tag{1.7}$$

Equivalently,

$$\begin{aligned} \mathbf{u} &= \rho \mathbf{W} \epsilon, & (1.8) \\ \mathbf{\Omega}(\lambda) &= \rho^2 \sigma^2 \mathbf{W} \mathbf{W}'. & (1.9) \end{aligned}$$

And in the spatial autoregressive moving average (SARMA) error model,

$$u_i = \rho \sum_{j=1}^{N} w_{ij}^{(1)} u_j + \lambda \sum_{j=1}^{N} w_{ij}^{(2)} \varepsilon_j, \tag{1.10}$$

which can be written in matrix notation as

$$\begin{aligned} \mathbf{u} &= \rho \mathbf{W}^{(1)} \mathbf{u} + \gamma \mathbf{W}^{(2)} \epsilon, & (1.11) \\ \mathbf{\Omega}(\lambda) &= \rho^2 \sigma^2 \left(I - \rho \mathbf{W}^{(1)} \right)^{-1} \left(I - \rho \mathbf{W}^{(1)\prime} \right)^{-1} \mathbf{W}^{(2)} \mathbf{W}^{(2)\prime}. & (1.12) \end{aligned}$$

As an alternative to starting with specific models of spatial correlation, we can directly specify the variances and covariances for the error term. For example, the variance can be a constant, and the covariance can be a function $c(\cdot)$ of the variance, a distance between two locations, and an unknown parameter which is to be estimated. In the matrix form,

$$\begin{aligned} \mathbf{\Omega}(\lambda)_{ii} &= \sigma_i^2, \\ \mathbf{\Omega}(\lambda)_{ij} &= \sigma_i^2 c\left(d_{ij,}\rho\right). \end{aligned}$$

1.2.2 Positive-Definiteness of Spatial Covariance Matrix

The error covariance matrix $\mathbf{\Omega}$ must be positive-definite. For each function $c(\cdot)$ specified, one need to check whether $\mathbf{\Omega}$ is positive-definite. Suppose $h \in \mathcal{R}$ which is used as the distance temporarily in order to distinguish from the derivative sign. According to Christakos (1984), $\mathbf{\Omega}$ is positive-definite as long as function $c(\cdot)$ satisfy the following conditions:

i. At $h = 0$, $dc(h)/dh < 0$.

ii. $\lim_{h \to \infty} c(h) = 0$.

iii. $d^2c\left(h\right)/dh^2 \geq 0$.

Derivates at $h = 0$ are one-sided, taken from right.

1.3 Estimation

1.3.1 Ordinary Least Squares and Generalized Least Squares

First, consider two estimators. Conditioning on the explanatory variables and location indicators, the OLS estimator is

$$\hat{\beta}_{OLS} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}, \tag{1.13}$$

and the GLS estimator is

$$\hat{\beta}_{GLS} = \left(\mathbf{X}'\mathbf{\Omega}^{-1}\mathbf{X}\right)^{-1}(\mathbf{X}'\mathbf{\Omega}^{-1}\mathbf{y}). \tag{1.14}$$

Suppose $\mathbf{\Omega}$ depends on fixed pairwise distances and a parameter vector λ. Then we can write $\mathbf{\Omega}$ as a function of λ, $\mathbf{\Omega}\left(\lambda\right)$. Once we find an estimator for λ, say $\hat{\lambda}$, we can obtain the estimated covariance matrix $\hat{\mathbf{\Omega}} \equiv \mathbf{\Omega}\left(\hat{\lambda}\right)$. The feasible GLS estimator (FGLS) is

$$\hat{\beta}_{FGLS} = (\mathbf{X}'\hat{\mathbf{\Omega}}^{-1}\mathbf{X})^{-1}(\mathbf{X}'\hat{\mathbf{\Omega}}^{-1}\mathbf{y}). \tag{1.15}$$

The following regularity conditions are assumed for deriving the asymptotics.

Assumption 1. The pairwise distance $d_{ij} < \infty$ between two locations s_i and s_j are lower bounded by $d_0 > 0$, $i = 1, 2, ...$;

Assumption 2. $\{(\mathbf{x}_i, u_i)\}$ is a mixing sequence on the sampling space, with mixing coefficient α of size $-r/2\left(r-1\right), r \geq 2$ or ϕ of size $-r/\left(r-2\right), r > 2$;

Assumption 3.

(a) $\mathrm{E}\left(\mathbf{x}_i' u_i\right) = \mathbf{0}$,$i = 1, 2, ...$;

(b) $\mathrm{E}\left|\mathbf{x}_i' u_i\right|^r < \Delta < \infty$ for all i;

Assumption 4.

(a) $\mathbf{A}_N^1 \equiv \mathrm{E}\left(\frac{1}{N}\sum_{i=1}^N \mathbf{x}_i' \mathbf{x}_i\right)$ is uniformly positive definite and has full rank K;

(b) $p\lim\left(\mathbf{A}_N^1\right) = \mathbf{A}_1$ as $N \to \infty$;

Assumption 5.

(a) $\mathbf{B}_N^1 \equiv \mathrm{Var}\left(\frac{1}{\sqrt{N}}\sum_{i=1}^N \mathbf{x}_i' u_i\right)$ $=\mathrm{E}\left(\frac{1}{N}\mathbf{X}'\mathbf{\Omega}_N\mathbf{X}\right)$ is uniformly positive definite, where $\mathbf{\Omega}_N$ is a positive-definite symmetric covariance matrix;

(b) $p\lim\mathbf{B}_N^1 = \mathbf{B}_1$ as $N \to \infty$;

Assumption 6.

(a) $\mathrm{E}\left(\mathbf{u}|\mathbf{X}\right) = \mathbf{0}$. This is the so-called strict exogeneity assumption. It says every u_i is not correlated with any function of $\mathbf{X}$. This condition implies **Assumption 3.(a)** and $\mathrm{E}\left(\frac{1}{N}\mathbf{X}'\mathbf{\Omega}_N^{-1}\mathbf{u}\right) = \mathbf{0}$;

Assumption 7.

(**a**) $\mathbf{B}_N^2 \equiv \mathrm{Var}\left(\frac{1}{\sqrt{N}}\mathbf{X}'\mathbf{\Omega}_N^{-1}\mathbf{u}\right) = \mathrm{E}\left(\frac{1}{N}\mathbf{X}'\mathbf{\Omega}_N^{-1}\mathbf{u}\mathbf{u}'\mathbf{\Omega}_N^{-1}\mathbf{X}\right)$ $=\mathrm{E}\left(\frac{1}{N}\mathbf{X}'\mathbf{\Omega}_N^{-1}\mathbf{X}\right)$;

(**b**) $\mathbf{B}_N^2$ is uniformly positive definite and has full rank K;

(c) $p\lim\left(\mathbf{B}_N^2\right) = \mathbf{B}_2$ as $N \to \infty$;

Proposition 1 *Under **Assumption 1, 2, 3, 4,** and **5,** the OLS estimator $\hat{\beta}_{OLS}$ is consistent and $\sqrt{N}\left(\hat{\beta}_{OLS} - \beta\right) \to^d \mathrm{N}\left(\mathbf{0}, \mathbf{A}_1^{-1}\mathbf{B}_1\mathbf{A}_1^{-1}\right)$. See proof in Chapter 4.*

Proposition 2 *Under **Assumption 1, 2, 5, 6,** and **7,** the FGLS estimator $\hat{\beta}_{FGLS}$ is consistent and $\sqrt{N}\left(\hat{\beta}_{FGLS} - \beta\right) \to^d \mathrm{N}\left(0, \mathbf{B}_2^{-1}\right)$. See proof in Chapter 4.*

OLS is easier than GLS but it is a less efficient estimator. GLS is efficient but could be computationally hard if the sample size is very large since it involves computing the inverse of a large covariance matrix. The following section demonstrates another way to get an efficient estimator.

1.3.2 Pseudo Generalized Least Squares

A lot of spatial data are collected with geographical location information such as firms in a county and schools in a school district. We can consider the firms in a county or schools in a school district as a group when we deal with empirical data. Spatial correlation can generally exist among any of the individuals, whether they are in the same group. The correlations in a group are much easier to deal with than those not in the same group and they should represent most of the correlations among individuals in the sample. PGLS is only based on within-group correlations. The estimation is still weighted by the error covariance matrix but it is a *tapered* matrix in the sense that the correlations are set as zero if individuals are not in the same group no matter if their true values are zero. The asymptotics depend on the mechanism that the size of each group is fixed while the groups increases as the total number of observations increases. If we have a large sample data set, it is hard to account for all pairwise correlations. The reason is that there is insufficient information to estimate the $N \times N$ covariance matrix directly from the data. Even asymptotics are not helpful since the number of covariances increases with N^2, whereas the sample size only grows with N (Anselin 1999). But with fixed group size, it is possible to account for the pairwise correlations within the group. Therefore, it is more realistic that we only use the information within a group while ignoring the cross-group correlations. Since within-group correlations will take into account most correlations of the observations in that group, it is possible to get an estimator that is quite close to the FGLS estimator, which is the psuedo genralized least squares (PGLS) estimator.

For panel data, a group is a cross section, and within a group observations are possibly serial correlated. Panel data assumes random sampling in the cross-sectional dimension; however, in this paper, we discuss a situation in which there is only one pure cross section. We estimate the parameters as if there are no groupwise correlations although there are correlations between observations in different groups. This is not the most efficient estimator, but by using this procedure, we can obtain a consistent and *almost* efficient estimator.

For notational convenience, we write the the linear regression model for a group in the population as

$$\mathbf{y}_g = \mathbf{X}_g\beta + \mathbf{u}_g, \tag{1.16}$$

where g denotes the gth group, $g = 1, 2, ...$ The number of observations in group g is L_g. $\mathbf{y}_g$ and $\mathbf{u}_g$ are $L_g \times 1$ vectors. $\mathbf{X}_g$ is $L_g \times K$. Now we can state the assumptions using the group notation.

Assumption P1.

(**a**) The number of observations L_g in each group is fixed. L_g/G is a small number. For simplicity assume that there are the same number of observations in each group. Thus the group size is fixed as $L = N/G$.

(**b**) The number of groups and the sample size both increase as the sampling domain increases.

Assumption P2. Let $d_{gh} < \infty$ denote the pairwise distance between two groups g and h. d_{gh} is lower bounded by $d_* > 0, \quad i = 1, 2, ..., N.$ $d_{gh} \in \mathbf{D}_*$, where $\mathbf{D}_*$ is the space containing all pairwise group distances[2].

Assumption P3(Assumption 2). $\{(\mathbf{x}_i, u_i)\}$ is a mixing sequence on the sampling space, with mixing coefficient α of size $-r/2\,(r-1)\,, r \geq 2$ or ϕ of size $-r/\,(r-2)\,, r > 2$;

Assumption P4. Let $\mathbf{\Lambda}_g$ be the $L \times L$ within-group covariance matrix for group g. Let $\mathbf{\Lambda}_N$ be the $N \times N$ matrix that only contains within-group variances and covariances. In other words, $\mathbf{\Lambda}_g$ is the gth diagonal matrix of $\mathbf{\Lambda}_N$. Let $\mathbf{\Omega}_N$ be the true covariance matrix of the sample. Notice that $\mathbf{\Lambda}_N$ is part of $\mathbf{\Omega}_N$. Define a $N \times N$ tapering matrix Γ. If observation i and j are in the same group, the ijth entry of $\Gamma, \Gamma(i,j)$ is equal to one. If observation i and j are in different groups, $\Gamma(i,j)$ is equal to zero. Thus $\mathbf{\Lambda}_N = \mathbf{\Omega}_N \cdot \mathbf{\Gamma}$ element by element**.**

Assumption P5.

(**a**) $\mathrm{E}\left(\mathbf{u}_g|\mathbf{X}_g\right) = 0$. This implies that for any positive definite $L \times L$ dimensional matrix $\mathbf{\Lambda}_g$, $\mathrm{E}\left(\mathbf{X}_g'\mathbf{\Lambda}_g^{-1}\mathbf{u}_g\right) = \mathbf{0}$ and $p\mathrm{lim}\frac{1}{G}\sum_{g=1}^{G}\mathbf{X}_g'\mathbf{\Lambda}_g^{-1}\mathbf{u}_g = \mathbf{0}$. Note that when the number of groups is equal to one, **Assumption P5** becomes $\mathrm{E}\left(\mathbf{u}|\mathbf{X}\right) = 0$, which is the strict exogeneity assumption for GLS.

(**b**) $\mathrm{E}\left|\mathbf{X}_g'\mathbf{\Lambda}_g^{-1}\mathbf{u}_g\right|^r$ exists for all g.

Assumption P6.

(**a**) $\mathbf{S}_G \equiv \mathrm{Var}\left(\frac{1}{\sqrt{G}}\sum_{g=1}^{G}\mathbf{X}_g'\mathbf{\Lambda}_g^{-1}\mathbf{u}_g\right)$ $=\mathrm{E}\left(\frac{1}{G}\sum_{g=1}^{G}\sum_{h=1}^{G}\mathbf{X_g'\Lambda_g^{-1}u}_g\mathbf{u}_h'\mathbf{\Lambda_h^{-1}X}_h\right)$
$= \mathrm{E}\left(\frac{1}{G}\sum_{g=1}^{G}\sum_{h=1}^{G}\mathbf{X_g'\Lambda_g^{-1}\Omega}_{gh}\mathbf{\Lambda_h^{-1}X}_h\right)$, where $\mathbf{\Omega}_{gh}$ is the ghth block of matrix $\mathbf{\Omega}_N$;

(**b**) $p\mathrm{lim}\mathbf{S}_G = \mathbf{S}$ as $G \to \infty$;

Assumption P7.

(**a**) $\mathbf{Q}_g \equiv \mathbf{X}_g'\mathbf{\Lambda}_g^{-1}\mathbf{X}_g$ is uniformly positive definite and has full rank K;

[2] One way to define groupwise distance is to use the smallest distance between two observations belonging to two different groups. Another way is to use the average distance of all pairwise distances of observations belonging to different groups. What we use for a group distance could depend on the empirical needs. The importance of the way researchers define the groupwise distance is unknown. We will leave this problem for future discussion.

(**b**) $p\lim_{G\to\infty}\frac{1}{G}\sum_{g=1}^{G}\mathbf{Q}_g=\mathbf{Q}$, and $\operatorname{rank}(\mathbf{Q})=K$. This condition implies that $p\text{lim}\frac{1}{G}\sum_{g=1}^{G}\mathbf{X}_g'\boldsymbol{\Lambda}_g^{-1}\mathbf{X}_g$ exists and has full rank.

The PGLS estimator is

$$\hat{\beta}_{PGLS}=\left(\sum_{g=1}^{G}\mathbf{X}_g'\boldsymbol{\Lambda}_g^{-1}\mathbf{X}_g\right)^{-1}\left(\sum_{g=1}^{G}\mathbf{X}_g'\boldsymbol{\Lambda}_g^{-1}\mathbf{y}_g\right). \tag{1.17}$$

If we can obtain a consistent estimator for the spatial correlation parameter λ, we can get a feasible PGLS estimator. Let $\hat{\boldsymbol{\Lambda}}_g=\boldsymbol{\Lambda}_g\left(\hat{\lambda}\right)$. Note that λ should be the same as it is in GLS if the assumptions are true.

$$\hat{\beta}_{FPGLS}=\left(\sum_{g=1}^{G}\mathbf{X}_g'\hat{\boldsymbol{\Lambda}}_g^{-1}\mathbf{X}_g\right)^{-1}\left(\sum_{g=1}^{G}\mathbf{X}_g'\hat{\boldsymbol{\Lambda}}_g^{-1}\mathbf{y}_g\right). \tag{1.18}$$

Proposition 3 *Under **Assumption P1-P7,** $\hat{\beta}_{FPGLS}$ is consistent and has an asymptotic normal distribution,* $\sqrt{G}\left(\hat{\beta}_{FPGLS}-\beta\right)\xrightarrow{d}N\left(\mathbf{0},\mathbf{Q}^{-1}\mathbf{S}\mathbf{Q}^{-1}\right)$. *See proof in Chapter 4.*

The variance covariance estimator under **Assumption P6** for FPGLS can be given as

$$\begin{aligned}\widehat{\text{AVar}}\left(\hat{\beta}_{FPGLS}\right) = & \left(\sum_{g=1}^{G}\mathbf{X}_g'\hat{\boldsymbol{\Lambda}}_g^{-1}\mathbf{X}_g\right)^{-1} \\ & \left(\sum_{g=1}^{G}\sum_{h=1}^{G}\mathbf{X}_g'\hat{\boldsymbol{\Lambda}}_g^{-1}\boldsymbol{\Omega}_{gh}\left(\hat{\lambda}\right)\hat{\boldsymbol{\Lambda}}_g^{-1}\mathbf{X}_h\right) \\ & \left(\sum_{g=1}^{G}\mathbf{X}_g'\hat{\boldsymbol{\Lambda}}_g^{-1}\mathbf{X}_g\right)^{-1}.\end{aligned} \tag{1.19}$$

The above expression in a concise form reads:

$$\widehat{\text{AVar}}\left(\hat{\beta}_{FPGLS}\right)=\left(\mathbf{X}'\hat{\boldsymbol{\Lambda}}^{-1}\mathbf{X}\right)^{-1}\left(\mathbf{X}'\hat{\boldsymbol{\Lambda}}^{-1}\hat{\boldsymbol{\Omega}}\hat{\boldsymbol{\Lambda}}^{-1}\mathbf{X}\right)\left(\mathbf{X}'\hat{\boldsymbol{\Lambda}}^{-1}\mathbf{X}\right)^{-1}. \tag{1.20}$$

Usually the specific form of the true covariance matrix $\mathbf{\Omega}$ is not known. Thus equation (1.19) can not be obtained.

1.3.3 Discussion of the Strict Exogeneity Condition

PGLS depends on **Assumption P5. (a)** $\mathrm{E}\left(\mathbf{u}_g|\mathbf{X}_g\right) = 0$. This assumption means the error term is not correlated with any of the explanatory variables within the group, but can be correlated with explanatory variables that are not in the group. This assumption is less likely to hold if we divide observations into arbitrary groups. Thus the strict exogeneity condition $\mathrm{E}\left(\mathbf{u}|\mathbf{X}\right) = 0$, which holds for errors and explanatory variables are more reasonable. There are cases in which the strict exogeneity condition is violated. For example, one case is the spatial lag model in which the strict exogeneity assumption is necessarily false. The spatial lag model can be written as

$$\mathbf{Y} = \mathbf{AY} + \mathbf{BX} + \mathbf{u}, \tag{1.21}$$

where $\mathbf{u}$ and explanatory variables must correlated with each other. In this case, even OLS is not consistent. This model can be solved in the reduced form by quasi-maximum likelihood estimation. For example, Lee (2004). Another case is when the explanatory variable has spatial correlation. If a shock to one region is correlated with explanatory variables in other regions, the strict exogeneity can fail.

1.4 Consistent Variance-Covariance Estimators

1.4.1 Spatial HAC Estimator

Failure to account for spatial dependence may result in inconsistent standard errors using standard techniques. Thus getting robust standard errors to spatial dependence is very important for hypothesis testing and statistical inference. In time series literature, Newey and West (1994) use a Bartlett kernel to consistently estimate the covariance matrix for a time series process. This is called the heterskedasticity and autocorrelation consistent (HAC) variance covariance estimator. Driscoll and Kraay (1998) provide a nonparametric covariance matrix estimation technique which yields standard error estimates that are robust to very general forms of spatial and temporal dependence as the time dimension becomes large. Conley (1999) uses a Bartlett window to estimate the variance matrix robust to cross-sectional correlation, so a rectangular *window* of correlations are used in this estimation. Kelejian and Prucha (2007) suggest a spatial HAC estimation. In this section, I suggest HAC variance-covariance estimators for PGLS, and provide a proof that the HAC estimator for PGLS is consistent.

Define a kernel function $k\left(d_{ij}\right)$ which depends on pairwise distance d_{ij}. Let d_* be a cutoff point such that if $d_{ij} < d_*$, $\text{Cov}\left(u_i, u_j\right)$ will be estimated and used. Otherwise, treat $\text{Cov}\left(u_i, u_j\right)$ equal to zero. $k\left(d_{ij}\right)$ decreases as d_{ij} increases. Researchers can use different kernel functions. In this paper, I use a Bartlett kernel, but I do not constrain the correlations within a rectangular window as in Conley (1999). Instead we could call the kernel function a *bartlett circle*. That is, as the center of one circle with its radius equal to d_*, one observation's covariance with all observations within this circle will be estimated.

The Bartlett circle kernel weighting function is

$$k\left(d_{ij}\right)=\begin{cases}1-d_{ij}/d_{*} & d_{ij}\leq d_{*}\\ 0 & d_{ij}>d_{*}\end{cases}. \tag{1.22}$$

A consistent variance covariance estimator for OLS that is robust to heteroskedasticity and spatial correlation is

$$\widehat{\text{Avar}}\left(\hat{\beta}_{OLS}\right)_{rob}=\left(\sum_{i=1}^{N}\mathbf{x}_i'\mathbf{x}_i\right)^{-1}\left(\sum_{i=1}^{N}\sum_{j=1}^{N}k\left(d_{ij}\right)\hat{v}_i\hat{v}_j\right)\left(\sum_{i=1}^{N}\mathbf{x}_i'\mathbf{x}_i\right)^{-1}, \tag{1.23}$$

where $\hat{v}_i=\mathbf{x}_i\hat{u}_i$ and $\hat{u}_i$ is OLS residual.

A consistent variance covariance estimator for PGLS that is robust to misspecification of variance-covariance structure is

$$\widehat{\text{Avar}}\left(\hat{\beta}_{FGLS}\right)_{rob}=\left[\mathbf{X}'\hat{\mathbf{\Omega}}^{-1}\mathbf{X}\right]^{-1}\mathbf{K}\left(\mathbf{D}\right)\cdot\check{\mathbf{v}}\check{\mathbf{v}}'\left[\mathbf{X}'\hat{\mathbf{\Omega}}^{-1}\mathbf{X}\right]^{-1}, \tag{1.24}$$

where $\hat{\mathbf{\Omega}}\equiv\mathbf{\Omega}\left(\hat{\lambda}\right)$, $\check{\mathbf{v}}=\mathbf{X}'\mathbf{\Omega}\left(\hat{\lambda}\right)^{-1}\check{\mathbf{u}}$ and $\check{\mathbf{u}}=\mathbf{y}-\mathbf{X}\hat{\beta}_{FGLS}$. $\mathbf{K}\left(\mathbf{D}\right)$ is an $N\times N$ kernel matrix. Here we use $\mathbf{K}\left(\mathbf{D}\right)\cdot\check{\mathbf{v}}\check{\mathbf{v}}'$ denotes the product of $\mathbf{K}\left(\mathbf{D}\right)$ and $\check{\mathbf{v}}\check{\mathbf{v}}'$ element by element. The ijth element of $\mathbf{K}\left(\mathbf{D}\right)$ is $k\left(d_{ij}\right)$, which is the same as given above. $\check{\mathbf{v}}\check{\mathbf{v}}'$ is a matrix of products of pairwise residuals. Because of possible misspecification of the variance covariance structure $\mathbf{\Omega}$, the spatial correlation may not be fully accounted for, or be treated incorrectly. Therefore there might still be spatial correlation left and we use the FGLS residuals to get a robust estimator.

In the PGLS estimation, correlations among observations within the same groups are used and in different groups are not. Thus intuitively we would want the HAC estimator to only account for the correlations across different groups. Therefore the kernel weights should only be put on the correlations among observations across different groups.

Let $\tilde{\mathbf{v}}_g = \mathbf{X}_g' \mathbf{\Lambda}_g \left(\hat{\lambda}\right)^{-1} \tilde{\mathbf{u}}$ and $\tilde{\mathbf{u}} = \mathbf{y} - \mathbf{X}\hat{\beta}_{FPGLS}$. Define kernel function for PFGLS:

$$p\left(d_{gh}\right) = \begin{cases} 1 - d_{gh}/d_{**} & d_{gh} \leq d_{**} \\ 0 & d_{gh} > d_{**} \end{cases} \tag{1.25}$$

where d_{gh} is the distance between group g and h. Let $P\left(d_{gh}\right)$ be the $G \times G$ matrix with the ghth element $P_{gh} = p\left(d_{gh}\right)$ for $g, h = 1, 2, ..., G$. Let O be a square matrix with each element equal to one. The dimension of O is equal to N/G, which is the number of individuals within each group. Then the full $N \times N$ kernel matrix is $P \otimes O$.

The HAC estimator that is robust to groupwise spatial correlation and misspecification is

$$\begin{aligned} \widehat{\text{Avar}}\left(\hat{\beta}_{FPGLS}\right)_{rob} = & \left[\sum_{g=1}^{G}\left(\mathbf{X}_g'\hat{\mathbf{\Lambda}}_g^{-1}\mathbf{X}_g\right)\right]^{-1}\left[\sum_{g=1}^{G}\sum_{h=1}^{G} p\left(d_{gh}\right)\tilde{\mathbf{v}}_g\tilde{\mathbf{v}}_h'\right] \\ & \left[\sum_{g=1}^{G}\left(\mathbf{X}_g'\hat{\mathbf{\Lambda}}_g^{-1}\mathbf{X}_g\right)\right]^{-1}. \end{aligned} \tag{1.26}$$

Following the proof of Theorem 2 in Kelejian and Prucha (2007), we can complete the proof for consistency of the PGLS estimator. Kelejian and Prucha (2007) provide a proof for the consistency of the spatial HAC estimator for the linear regression model while their scenario is based on ordinary least squares estimator, which is a special case of PGLS when the group size is equal to one. Thus, the HAC estimator for PGLS is an extension to the one in Kelejian and Prucha (2007).

Proposition 4 $G \cdot \widehat{\text{Avar}}\left(\hat{\beta}_{FPGLS}\right)_{rob} - \mathbf{Q}^{-1}\mathbf{S}\mathbf{Q}^{-1} \xrightarrow{p} \mathbf{0}$. *See proof in Chapter 4.*

1.4.2 Estimation of the Spatial Correlation Parameter

A consistent estimator for ρ can be obtained by a minimum distance estimator,

$$\hat{\rho} = \arg\min \sum_{i=1}^{N} \sum_{j \neq i}^{N} [\hat{u}_i \hat{u}_j - \Omega_{ij}(d_{ij}, \rho)]^2.$$

A common structure of Ω_{ij} in spatial statistics is the exponential form,

$$\Omega_{ij} = \sigma^2 \exp\left(-\frac{d_{ij}}{\rho}\right).$$

An estimator for σ^2 is

$$\hat{\sigma}^2 = N^{-1} \sum_{i=1}^{N} \hat{u}_i^2.$$

1.5 Alternative Estimation Approach: Quasi-Maximum Likelihood Estimator

The quasi-maximum likelihood estimator (QMLE) using the estimated covariance matrix derived from OLS residuals is the same as PGLS. Supposing the same assumptions hold as before, $\text{E}\left(\mathbf{u}_g|\mathbf{X}_g\right) = \mathbf{y_g} - \mathbf{X}_g\beta$. $\text{Var}\left(\mathbf{u}_g|\mathbf{X}_g\right) = \sigma^2\mathbf{\Lambda}_g(\rho)$. We can write the quasi log likelihood function for group g using a multivariate normal distribution as

$$\begin{aligned} l_g\left(\beta,\rho,\sigma^2\right) &= \log f_g\left(\mathbf{y}_g|\mathbf{X}_g;\beta,\lambda\right) \qquad (1.27) \\ &= -\frac{L}{2}\log\left(2\pi\right)-\frac{L}{2}\log\sigma^2-\frac{1}{2}\log\left|\Lambda_g\left(\rho\right)\right| \\ &\quad -\frac{1}{2\sigma^2}\left(\mathbf{y}_g-\mathbf{X}_g\beta\right)'\Lambda_g^{-1}\left(\rho\right)\left(\mathbf{y}_g-\mathbf{X}_g\beta\right). \end{aligned}$$

The QMLE estimators $\hat{\beta}_{QMLE}, \hat{\rho}_{QMLE}$ and $\hat{\sigma}^2_{QMLE}$ jointly solve

$$\begin{aligned} \max_{\theta\in\Theta} L_G\left(\beta,\rho,\sigma^2\right) &= \sum_{g=1}^{G} l_g\left(\beta,\lambda\right) = -\frac{G\times L}{2}\log\left(2\pi\right)-\frac{G\times L}{2}\log\sigma^2 \qquad (1.28) \\ &\quad -\frac{1}{2}\sum_{g=1}^{G}\log\left|\Lambda_g\left(\rho\right)\right|-\frac{1}{2\sigma^2}\sum_{g=1}^{G}\left(\mathbf{y}_g-\mathbf{X}_g\beta\right)'\boldsymbol{\Lambda}_g^{-1}\left(\rho\right)\left(\mathbf{y}_g-\mathbf{X}_g\beta\right). \end{aligned}$$

Since the first term of log likelihood is a constant, we can eliminate it in the maximization problem. The log likelihood becomes

$$\begin{aligned} L_G\left(\beta,\rho,\sigma^2\right) &= -\frac{G\times L}{2}\log\sigma^2-\frac{1}{2}\sum_{g=1}^{G}\log\left|\Lambda_g\left(\rho\right)\right| \qquad (1.29) \\ &\quad -\frac{1}{2\sigma^2}\sum_{g=1}^{G}\left(\mathbf{y}_g-\mathbf{X}_g\beta\right)'\boldsymbol{\Lambda}_g^{-1}\left(\rho\right)\left(\mathbf{y}_g-\mathbf{X}_g\beta\right). \qquad (1.30) \end{aligned}$$

For a given ρ, by maximizing the log likelihood function we can get

$$\hat{\beta}_{QMLE} = \left[\sum_{g=1}^{G}\mathbf{X}_g'\boldsymbol{\Lambda}_g\left(\rho\right)^{-1}\mathbf{X}_g\right]^{-1}\left[\sum_{g=1}^{G}\mathbf{X}_g'\boldsymbol{\Lambda}_g\left(\rho\right)^{-1}\mathbf{y}_g\right] \qquad (1.31)$$

and

$$\hat{\sigma}^2_{QMLE} = \frac{1}{N}\sum_{g=1}^{G}\left(\mathbf{y}_g - \mathbf{X}_g\hat{\beta}_{QMLE}\right)' \Lambda_g^{-1}(\rho)\left(\mathbf{y}_g - \mathbf{X}_g\hat{\beta}_{QMLE}\right). \tag{1.32}$$

By plugging $\hat{\beta}_{QMLE}$ and $\hat{\sigma}^2_{QMLE}$ into the log likelihood we can get a concentrated likelihood. By maximizing the concentrated log likelihood $L(\rho)$, we get $\hat{\rho}_{QMLE}$ and further $\hat{\beta}_{QMLE}$ and $\hat{\sigma}^2_{QMLE}$.

$$\begin{aligned} L(\rho) &= -\frac{1}{2}\log\left[\frac{1}{N}\sum_{g=1}^{G}\left(\mathbf{y}_g - \mathbf{X}_g\hat{\beta}_{QMLE}\right)' \mathbf{\Lambda}_g^{-1}(\rho)\left(\mathbf{y}_g - \mathbf{X}_g\hat{\beta}_{QMLE}\right)\right] \\ &\quad -\frac{1}{2}\sum_{g=1}^{G}\log\left|\mathbf{\Lambda}_g(\rho)\right| - \frac{1}{2} \\ &= \frac{1}{2}\left[\log(N) - 1\right] - \frac{1}{2}\log\left[\sum_{g=1}^{G}\left(\mathbf{y}_g - \mathbf{X}_g\hat{\beta}_{QMLE}\right)' \mathbf{\Lambda}_g^{-1}(\rho)\left(\mathbf{y}_g - \mathbf{X}_g\hat{\beta}_{QMLE}\right)\right] \\ &\quad -\frac{1}{2}\sum_{g=1}^{G}\log\left|\mathbf{\Lambda}_g(\rho)\right| \end{aligned} \tag{1.33}$$

Note that if we plug $\hat{\rho}$ in $\hat{\beta}_{QMLE}$ we get $\hat{\beta}_{FPGLS}$.

We can also get a HAC estimator for $\hat{\beta}_{QMLE}$. The Hessian for each group g is $\mathbf{H}_g = \begin{pmatrix} \mathbf{H}_{g11} & \mathbf{H}_{g12} \\ \mathbf{H}_{g21} & \mathbf{H}_{g22} \end{pmatrix}$. Let $\mathbf{a}_g = -\mathrm{E}\left(\mathbf{H}_g|\mathbf{X}_g, \mathbf{D}_g\right)$, then

$$\mathbf{a}_g = \begin{pmatrix} \mathbf{X}_g'\Lambda_g^{-1}(\lambda)\mathbf{X}_g & 0 \\ 0 & \frac{1}{2}\nabla_\lambda\Lambda_g(\lambda)'\left[\Lambda_g^{-1}(\lambda)\otimes\Lambda_g^{-1}(\lambda)\right]\nabla_\lambda\Lambda_g(\lambda) \end{pmatrix}$$

$$\begin{aligned} \mathbf{H} &\equiv \frac{1}{G}\sum_{g=1}^{G}\mathbf{H}_g \\ \mathbf{A} &\equiv \mathrm{E}\left[-\mathbf{H}\right] = -\frac{1}{G}\sum_{g=1}^{G}\mathrm{E}\left(\mathbf{H}_g|\mathbf{X}_g,\mathbf{D}_g\right) = \frac{1}{G}\sum_{g=1}^{G}\mathbf{a}_g \end{aligned}$$

$$\begin{aligned} \hat{\mathbf{A}} &\equiv \frac{1}{G}\sum_{g=1}^{G}\mathbf{a}_g\left(\hat{\beta},\hat{\lambda}\right) \\ &= \frac{1}{G}\sum_{g=1}^{G}[\begin{array}{cc} \mathbf{X}_g'\mathbf{\Lambda}_g^{-1}\left(\hat{\lambda}\right)\mathbf{X}_g & 0 \\ 0 & \frac{1}{2}\nabla_{\lambda}\mathbf{\Lambda}_g\left(\hat{\lambda}\right)'\left[\mathbf{\Lambda}_g^{-1}\left(\hat{\lambda}\right)\otimes\mathbf{\Lambda}_g^{-1}\left(\hat{\lambda}\right)\right]\nabla_{\lambda}\mathbf{\Lambda}_g\left(\hat{\lambda}\right) \end{array}] \end{aligned}$$

Note that there is no $\hat{\beta}$ in $\hat{\mathbf{A}}$.

$$\mathbf{B}_G \equiv \mathrm{Var}\left(\frac{1}{\sqrt{G}}\sum_{g=1}^{G}\mathbf{s}_g\right) = \frac{1}{G}\sum_{g=1}^{G}\sum_{g=1}^{H}\mathrm{E}\left(\mathbf{s}_g\mathbf{s}_h'\right),$$

and let $\hat{\mathbf{s}}_g = \mathbf{s}_g\left(\hat{\beta},\hat{\lambda}\right)$,

$$\hat{\mathbf{B}}_{HAC} = \left\{\frac{1}{G}\sum_{g=1}^{G}\sum_{g=1}^{H}K\left[dist\left(g,h\right)\right]\hat{\mathbf{s}}_g\hat{\mathbf{s}}_h'\right\}.$$

Thus the HAC estimator for asymptotic variance of $\hat{\theta}$ is $\frac{1}{G}\hat{\mathbf{A}}^{-1}\hat{\mathbf{B}}_{HAC}\hat{\mathbf{A}}^{-1}$.

Proposition 5 *Under certain conditions, $\hat{\beta}_{QMLE}$ is consistent and has an asymptotic normal distribution. See proof in Chapter 4.*

1.6 A Monte Carlo Simulation Study

In this section, we provide Monte Carlo simulation results which shows that the PGLS estimator performs reasonably well for finite samples. In the simulation, we study the properties of the PGLS estimator compared to the OLS and GLS estimators. We also provide robust standard errors for OLS and PGLS.

1.6.1 Data Generating Process

The total number of observations is N which takes 400 and 1600 separately. The data are generated on a $\sqrt{N} \times \sqrt{N}$ lattice. The locations for the data are represented by coordinates $\left\{(r, s) : r, s = 1, 2, ..., \sqrt{N}\right\}$. The distance d_{ij} between location i and j is Euclidean distance. Suppose $A(a_i, a_j)$ and $B(b_i, b_j)$ are the two points on the lattice, then their Euclidean distance is $\sqrt{(a_i - b_i)^2 + (a_j - b_j)^2}$. Besides the Euclidean distance, other distance measures may also be considered. Commonly seen distance measures include a time series distance (as a special case of Euclidean distance) which means the locations are on a straight line, Manhattan distance $|a_i - b_i| + \left|a_j - b_j\right|$, and maximum coordinate-wise distance $\max\left(|a_i - b_i|, \left|a_j - b_j\right|\right)$. For simplicity, only Euclidean distance is used in this paper. We consider three cases of data-generating processes for different correlation structures.

1.6.1.1 Case I

In the first case, the exponential form of spatial correlation is used. We generate data as follows:

(1) $y_i = \alpha + x_i\beta + u_i, \quad \alpha = 1, \quad \beta = 1.$

(2) Specify the covariance matrix Ω for the error term:

$\Omega_{ij} = \sigma^2 \exp\left(-\frac{d_{ij}}{\rho}\right), \quad \sigma^2 = 1, \quad \rho = 0.1, 0.5, 1, 2, 5.$ In this case, the corresponding pairwise correlations for the error term when the pairwise distance is one are: 0.00, 0.14, 0.37, 0.61, 0.82.

(3) Generate the error term: $\mathbf{u} = \Omega^{1/2}\epsilon$, where ϵ is a vector of i.i.d. standard normal random numbers. u_i is the ith element in the vector $\mathbf{u}$. $\Omega^{1/2}$ is obtained by Cholesky decomposition.

(4) x_i is a spatially correlated variable. The vector $\mathbf{x} = \Omega^{1/2}\xi$, where ξ is a vector of i.i.d. standard normal random numbers and Ω is the same as in (2) and (3) but with $\rho = 1$ without loss of generality (w.l.o.g.).

The simulation results can be found in Table 1.1, 1.2, and 1.3. In this estimation process, the spatial correlation parameter is estimated by a minimum distance estimator $\hat{\rho} = \min \sum[\hat{u}_i\hat{u}_j - \sigma^2 \exp(-\frac{d_{ij}}{\rho})]^2$ for $i \neq j$.

1.6.1.2 Case II

In the second case, the spatial correlation is a function of the inverse of distance. The data generating process is as follows:

(1) $y_i = \alpha + x_i\beta + u_i, \quad \alpha = 1, \quad \beta = 1.$

(2) Specify the covariance matrix Ω for the error term:

$\Omega_{ij} = \sigma^2 \frac{\rho}{d_{ij}}, \quad \sigma^2 = 1, \quad \rho = 0, 0.2, 0.4, 0.6.$ In this case, the corresponding pairwise correlation for the error term when the pairwise distance is one is equal to ρ.

(3) Generate the error term: $\mathbf{u} = \Omega^{1/2}\epsilon$, where ϵ is a vector of i.i.d. standard normal random numbers. u_i is the ith element in the vector $\mathbf{u}$. $\Omega^{1/2}$ is obtained by Cholesky decomposition.

(4) x_i is a spatially correlated variable. The vector $\mathbf{x} = \Omega^{1/2}\xi$, where ξ is a vector of i.i.d.

standard normal random numbers and Ω is the same as in (2) and (3) but $\rho = 1$ w.l.o.g..

The simulation results can be found in Table 1.4 and 1.5, 1.6. In this estimation process, the spatial correlation parameter is estimated as the average of $\hat{u}_i \hat{u}_j / \hat{\sigma}^2$ for pairwise distance $d_{ij} = 1$, which means, ρ is estimated only using the pairs of observations whose distances are one.

1.6.1.3 Case III

For completeness, another data generating process which uses a spatial weight matrix that is popular in spatial econometrics is as follows:

(1) $y_i = \alpha + x_i \beta + u_i, \quad \alpha = 1, \quad \beta = 1.$

(2) There are N/L independent districts, where L is the number of observations within a district. Specify the spatial weight matrix $\mathbf{W}$ for the error term: $w_{ij} = 1/(L-1)$ if $i \neq j$ and i and j are in the same district. $w_{ij} = 0$ if $i = j$ or i and j are in different districts. In this case, the corresponding pairwise correlation for the error term within the district is equal to $\rho/(L-1)$.

(3) Generate the error term: $\mathbf{u} = \rho \mathbf{W} \mathbf{u} + \epsilon, \quad \sigma^2 = 1, \quad \rho = 0, 0.2, 0.4, 0.6$, where ϵ is a vector of i.i.d. standard normal random numbers. u_i is the ith element in the vector $\mathbf{u}$.

(4) x_i is a standard normal variable $\mathrm{N}(0,1)$.

The simulation results can be found in Table 1.7 and 1.8. In this estimation process, the spatial correlation parameter is estimated in the following way:

$$\hat{\rho} = \left[(\mathbf{W}\hat{\mathbf{u}})' (\mathbf{W}\hat{\mathbf{u}}) \right]^{-1} (\mathbf{W}\hat{\mathbf{u}})' \hat{\mathbf{u}}. \tag{1.34}$$

1.6.1.4 Robust Standard Errors

The simulations also provide robust standard errors for the OLS and PGLS estimators since these estimators ignore spatial correlations to some extent. We do not discuss the optimal kernal variance estimators and we only provide feasible variance estimators for the estimation of robust standard errors. For the OLS estimator, we use the formula in equation (1.23). The cutoff point d_* is set as $\sqrt[3]{N}$. When $N = 400, d_* = 7.368$. Thus, for observation i, its covariances with any observation that is within 7.4 distance units are accounted for, similar to the case when $N = 1600$. For the PGLS estimators, we use the formula in equation (1.26). The kernel function in the simulation is specified as in equation (1.25). First, the groupwise distance d_{gh} is specified as the distance between centers of the location. Let the cutoff point d_{**} be $\sqrt[3]{N}$. If $d_{gh} \leq d_{**}$, the covariances of the observations in group g and h will be used. To keep it simple, we put the same weight $1 - d_{gh}/d_{**}$ on each of those covariances. Like the lags in time series, how to choose the cutoff point needs consideration of economics of the problems.

1.6.2 Monte Carlo Simulation Results

For Case I, Table 1.1 shows that as ρ increases, the standard deviation of the OLS estimator increases, which means OLS becomes less efficient as spatial correlation increases in the sample. PGLS performance becomes better when compared to OLS as ρ increases. Using a group size equal to four, which is very small, we can gain back most of the efficiency. As we increase the group size, the feasible PGLS gains more efficiency. When we use a group size equal to 16, the PGLS estimator is almost as efficient as the GLS estimator. Table 1.2 provides robust standard errors for OLS, PGLS with group size 4 and 16. The spatial correlation parameter is calculated using a minimum distance estimator. Simulations show that this estimator is biased when the spatial correlation is high. However, even if the spatial correlation parameter is not consistently estimated, the feasible PGLS estimator can still gain efficiency back. This is because the estimated covariance (even though not consistent) captures some structure of the true covariance.

Case II and Case I have very similar results regarding PGLS behavior. But the estimation of spatial correlation in Case II is less biased and more efficient than in Case I.

In Case III, independent groups are generated, thus GLS is PGLS in this case. The estimation of spatial correlation as in equation 1.34 is an OLS estimator using the residuals. As in Table 1.7, the estimator for ρ is biased upwards, though the GLS estimator still behaves well.

Table 1.1: Case I: Mean Parameters and Standard Deviation

	N=400,	T=2000,	$\beta = 1$					
ρ	$\hat{\beta}_{ols}$	s.d.($\hat{\beta}_{ols}$)	$\hat{\beta}_{gls}$	s.d.($\hat{\beta}_{gls}$)	$\hat{\beta}^4_{pgls}$	s.d.($\hat{\beta}^4_{pgls}$)	$\hat{\beta}^{16}_{pgls}$	s.d.($\hat{\beta}^{16}_{pgls}$)
0.1	1.000	0.050	1.000	0.050	1.000	0.050	1.000	0.050
0.5	1.000	0.056	1.000	0.054	1.000	0.055	1.000	0.054
1	1.000	0.068	0.999	0.051	1.000	0.057	1.000	0.053
2	1.002	0.081	1.000	0.040	1.001	0.051	1.000	0.044
5	1.002	0.088	1.000	0.028	1.001	0.041	1.002	0.033
	N=1600,	T=2000,	$\beta = 1$					
ρ	$\hat{\beta}_{ols}$	s.d.($\hat{\beta}_{ols}$)	$\hat{\beta}_{gls}$	s.d.($\hat{\beta}_{gls}$)	$\hat{\beta}^4_{pgls}$	s.d.($\hat{\beta}^4_{pgls}$)	$\hat{\beta}^{16}_{pgls}$	s.d.($\hat{\beta}^{16}_{pgls}$)
0.1	0.999	0.025	0.999	0.025	0.999	0.025	0.999	0.025
0.5	0.999	0.028	0.999	0.027	0.999	0.027	0.999	0.027
1	0.999	0.034	0.999	0.025	0.999	0.028	0.999	0.026
2	1.000	0.043	0.999	0.020	0.999	0.025	0.999	0.022
5	0.999	0.049	0.999	0.013	0.999	0.018	0.999	0.015

s.d.() means standard deviation in the simulation.

$\hat{\beta}^4_{pgls}$ and $\hat{\beta}^{16}_{pgls}$ are the PGLS estimators using group size equal to 4 and 16 separately.

Table 1.2: Case I: Robust Standard Errors

	N=400,	T=2000		
ρ	s.e.*($\hat{\beta}_{ols}$)	r.s.e.($\hat{\beta}_{ols}$)	r.s.e.($\hat{\beta}^4_{pgls}$)	r.s.e.($\hat{\beta}^{16}_{pgls}$)
0.1	0.038	0.047	0.047	0.047
0.5	0.038	0.050	0.049	0.050
1	0.037	0.058	0.051	0.051
2	0.035	0.066	0.052	0.048
5	0.030	0.068	0.048	0.041
	N=1600,	T=2000		
ρ	s.e.*($\hat{\beta}_{ols}$)	r.s.e.($\hat{\beta}_{ols}$)	r.s.e.($\hat{\beta}^4_{pgls}$)	r.s.e.($\hat{\beta}^{16}_{pgls}$)
0.1	0.017	0.024	0.024	0.024
0.5	0.017	0.026	0.025	0.025
1	0.017	0.032	0.027	0.027
2	0.017	0.038	0.027	0.026
5	0.015	0.042	0.026	0.023

s.e.*($\hat{\beta}_{ols}$) is the average usual OLS standard error.

r.s.e.() is the average spatial correlation robust standard error.

Table 1.3: Case I: Estimated Error Variance Parameters

	N=400,	T=2000,	$\sigma^2 = 1$	
ρ	$\hat{\rho}$	s.d.$(\hat{\rho})$	$\hat{\sigma}^2$	s.d.$\left(\hat{\sigma}^2\right)$
0.1	0.134	0.125	0.995	0.072
0.5	0.480	0.079	0.992	0.074
1	0.919	0.166	0.980	0.095
2	1.554	0.357	0.943	0.150
5	2.297	0.563	0.825	0.248
	N=1600,	T=2000,	$\sigma^2 = 1$	
ρ	$\hat{\rho}$	s.d.$(\hat{\rho})$	$\hat{\sigma}^2$	s.d.$\left(\hat{\sigma}^2\right)$
0.1	0.121	0.102	0.999	0.035
0.5	0.495	0.040	0.998	0.037
1	0.979	0.097	0.995	0.048
2	1.855	0.310	0.984	0.084
5	3.511	0.824	0.930	0.175

$\hat{\rho}$ is the average of estimates for ρ.
$\hat{\sigma}^2$ is the average of estimates for σ^2.

Table 1.4: Case II: Mean Parameters and Standard Deviation

	N=400,	T=2000,	$\beta = 1$					
ρ	$\hat{\beta}_{ols}$	s.d.$(\hat{\beta}_{ols})$	$\hat{\beta}_{gls}$	s.d.$(\hat{\beta}_{gls})$	$\hat{\beta}^4_{pgls}$	s.d.$(\hat{\beta}^4_{pgls})$	$\hat{\beta}^{16}_{pgls}$	s.d.$(\hat{\beta}^{16}_{pgls})$
0	1.000	0.051	1.002	0.051	1.000	0.051	1.000	0.053
0.2	1.000	0.057	0.999	0.051	1.000	0.053	1.000	0.052
0.4	1.000	0.062	0.999	0.044	1.000	0.050	1.000	0.047
0.6	1.001	0.068	1.000	0.038	1.000	0.042	1.000	0.035
	N=1600,	T=2000,	$\beta = 1$					
ρ	$\hat{\beta}_{ols}$	s.d.$(\hat{\beta}_{ols})$	$\hat{\beta}_{gls}$	s.d.$(\hat{\beta}_{gls})$	$\hat{\beta}^4_{pgls}$	s.d.$(\hat{\beta}^4_{pgls})$	$\hat{\beta}^{16}_{pgls}$	s.d.$(\hat{\beta}^{16}_{pgls})$
0	0.999	0.025	0.999	0.087	0.999	0.025	0.999	0.025
0.2	0.999	0.029	0.999	0.025	0.999	0.027	0.999	0.026
0.4	0.999	0.033	0.999	0.022	0.999	0.025	0.999	0.023
0.6	0.999	0.037	1.002	0.018	1.000	0.021	1.000	0.017

s.d.() means standard deviation in the simulation.
$\hat{\beta}^4_{pgls}$ and $\hat{\beta}^{16}_{pgls}$ are the PGLS estimators using group size equal to 4 and 16 separately.

Table 1.5: Case II: Robust Standard Errors

	N=400,	T=2000		
ρ	s.e.*$(\hat{\beta}_{ols})$	r.s.e.$(\hat{\beta}_{ols})$	r.s.e.$(\hat{\beta}^{4}_{pgls})$	r.s.e.$(\hat{\beta}^{16}_{pgls})$
0	0.037	0.047	0.067	0.095
0.2	0.036	0.050	0.068	0.093
0.4	0.035	0.053	0.067	0.086
0.6	0.034	0.056	0.062	0.069
	N=1600,	T=2000		
ρ	s.e.*$(\hat{\beta}_{ols})$	r.s.e.$(\hat{\beta}_{ols})$	r.s.e.$(\hat{\beta}^{4}_{pgls})$	r.s.e.$(\hat{\beta}^{16}_{pgls})$
0	0.017	0.024	0.037	0.050
0.2	0.017	0.027	0.038	0.051
0.4	0.017	0.030	0.037	0.047
0.6	0.016	0.032	0.035	0.038

s.e.*$(\hat{\beta}_{ols})$ is the average usual OLS standard error.
r.s.e.() is the average spatial correlation robust standard error.

Table 1.6: Case II: Estimated Error Variance Parameters

	N=400,	T=2000,	$\sigma^2 = 1$	
ρ	$\hat{\rho}$	s.d.$(\hat{\rho})$	$\hat{\sigma}^2$	s.d.$\left(\hat{\sigma}^2\right)$
0	-0.003	0.036	0.995	0.072
0.2	0.171	0.048	0.968	0.078
0.4	0.352	0.055	0.943	0.098
0.6	0.541	0.051	0.919	0.127
	N=1600,	T=2000,	$\sigma^2 = 1$	
ρ	$\hat{\rho}$	s.d.$(\hat{\rho})$	$\hat{\sigma}^2$	s.d.$\left(\hat{\sigma}^2\right)$
0	-0.001	0.018	0.999	0.035
0.2	0.186	0.027	0.985	0.041
0.4	0.377	0.032	0.972	0.057
0.6	0.573	0.029	0.959	0.077

$\hat{\rho}$ is the average of estimates for ρ.
$\hat{\sigma}^2$ is the average of estimates for σ^2.

Table 1.7: Case III: Mean Parameters and Standard Deviation

Case III	N=400,	T=2000,	$\beta = 1$	
ρ	$\hat{\beta}_{ols}$	s.d.$(\hat{\beta}_{ols})$	$\hat{\beta}_{gls}$	s.d.$(\hat{\beta}_{gls})$
0	1.000	0.051	0.999	0.051
0.2	0.999	0.053	1.001	0.052
0.4	1.000	0.057	1.000	0.050
0.6	1.001	0.072	1.000	0.048
	N=1600,	T=2000,	$\beta = 1$	
ρ	$\hat{\beta}_{ols}$	s.d.$(\hat{\beta}_{ols})$	$\hat{\beta}_{gls}$	s.d.$(\hat{\beta}_{gls})$
0	1.000	0.026	1.000	0.026
0.2	1.001	0.026	1.001	0.026
0.4	1.000	0.029	1.001	0.025
0.6	1.001	0.036	1.002	0.024

s.d.() means standard deviation in the simulation. $\hat{\beta}_{ols}$ and $\hat{\beta}_{fgls}$ are the average of OLS and FGLS estimators.

Table 1.8: Case III: Estimated Error Variance Parameters

	N=400,	T=2000,	$\sigma^2 = 1$	
ρ	$\hat{\rho}$	s.d.$(\hat{\rho})$	$\hat{\sigma}^2$	s.d.$\left(\hat{\sigma}^2\right)$
0	-0.018	0.128	0.992	0.070
0.2	0.353	0.092	0.982	0.071
0.4	0.648	0.055	0.948	0.072
0.6	0.852	0.024	0.896	0.071
	N=1600,	T=2000,	$\sigma^2 = 1$	
ρ	$\hat{\rho}$	s.d.$(\hat{\rho})$	$\hat{\sigma}^2$	s.d.$\left(\hat{\sigma}^2\right)$
0	-0.004	0.062	0.998	0.036
0.2	0.365	0.044	0.998	0.036
0.4	0.656	0.026	0.995	0.036
0.6	0.856	0.015	0.985	0.036

$\hat{\rho}$ is the average of estimates for ρ.
$\hat{\sigma}^2$ is the average of estimates for σ^2.

1.7 Conclusions

In a linear regression model with spatial data, we can use a weighted least squares to improve efficiency, even if the weight might be misspecified. As long as the misspecified structure can capture some properties of the true variance structure of the spatial data, efficiency can be improved, which is the idea of ”pseudo GLS”.

Chapter 2

Estimation of Nonlinear Models in a Quasi-Maximum Likelihood Framework with Spatial Data

2.1 Introduction

In a lot of empirical economic and social studies, there are discrete data examples which exhibit spatial correlations due to the geographical locations of individuals or agents of interest. For instance, the number of patents a firm received shows correlation with that received by other firms near by. This may be due to a technology spillover effect or a common policy aiming at encouraging new technology in this place. 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 (Brown, Smith, & Weisbenner 2008). The first example is a count data example and the second one is a binary response example. Both examples deals with discrete data. Nonlinear models are suitable for the study on discrete variables. Unfortunately there are not much literature on the estimation of nonlinear models with discrete spatial data. Because of the spatial correlation, the discrete variables are not independent. Both the nonlinearity and the correlation make the estimation difficult.

Maximum likelihood estimation (MLE) is a widely used method in estimating nonlinear models. In order to use MLE, one needs to specify the joint distributions of spatial random variables. This includes correctly specifying the marginal and the conditional distributions. However, given a spatial data set, the dependence structure is generally unknown. If the joint distribution of the variables is misspecified, MLE is generally not consistent. Another estimation method is quasi-maximum likelihood estimation (QMLE). Using a density that belongs to a linear exponential family (LEF), QMLE is consistent if we correctly specify the conditional mean with other features of the density misspecified. In a panel data case, pooled (partial) QMLE which ignores serial correlations is consistent under some regularity conditions (Wooldridge 2010).

In their 2009 working paper, Wang, Iglesias and Wooldridge use a bivariate Probit partial MLE to improve the estimation efficiency with a spatial Probit model. Using their approach we would need to correctly specify the marginal distribution of the binary response variable conditional on the covariates and distance measures[1]. Since the bivariate marginal distribution of a spatial multivariate normal distribution is bivariate normal, one can derive the bivariate normal distribution under some distributional assumptions. There are two problems with this paper by Wang, Iglesias and Wooldridge (2012). First the computation is already hard for a bivariate distribution. The multivariate marginal distribution of a higher dimension is more computationally demanding; second it also requires the correct specifi-

[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 physical locations is commonly imprecise, eg. only zip code is known. Conley and Molinari (2007) deals with the inference problem when there exist distance errors. In this chapter I assume there are no measurement errors in pairwise distances.

cation of the marginal bivariate normal distribution to obtain consistency. In fact, we can have less restrictive distributional assumptions than those required in bivariate partial MLE. Suppose we only specify the mean and the working variances and covariances[2]. Using QMLE in the LEF, we can get consistent estimators. Even if the the variances and covariances are not correctly specified, we can still consistently estimate the mean parameters as well as the average partial effects, which are more interesting.

In the literature, the QMLE and GEE approach is used in panel data models to get more efficient estimators (Gourierous, Monfort, & Trongnon 1984). In this paper, I will demonstrate how to use the QMLE and GEE approach in a spatial data setting to get a consistent and more efficient estimator. Generalized least squares (GLS) can be used to improve the estimation efficiency in a linear regression model even if the variance covariance structure is misspecified. Similarly, generalized estimating equations (GEE) or weighted multivariate nonlinear least squares (WMNLS) are used in nonlinear panel data models and system of equations to obtain more efficient conditional mean parameters. Generally we expect that GEE can give more efficient estimators compared to PMLE, which uses only the marginal density to get the consistent estimators.

To use the QMLE in the spatial data setting, I first give a series of assumptions, based on which M-estimators are consistent for the spatial processes. To derive the asymptotics for the M-estimators we have to use a uniform law of large numbers (ULLN) and a central limit theorem (CLT). These limit theorems are the fundamental building blocks for the asymptotic theory of nonlinear spatial M-estimators, e.g. maximum likelihood estimators (MLE) and generalized method of moments estimators (GMM) (Jenish and Prucha, 2009).

[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.

Conley (1999) makes an important contribution towards developing an asymptotic theory of GMM estimators for spatial processes. He utilizes Bolthausen's (1982) CLT for stationary random fields. Jenish and Prucha (2009) provide a ULLN and a CLT for spatial data including nonstationary spatial processes. Using theorems in Jenish and Prucha (2009), one can analyze more interesting economic phenomena. For example, real estate prices usually shoot up as one moves from the periphery to the center of a big city. While I will not discuss trending processes in this paper, Cressie (1993) provides numerous examples of trending spatial processes.

In Section 2, the M-estimator framework under the spatial data context is established. A series of assumptions are given based on Jenish and Prucha (2009) under which M-estimators are consistent and have an asymptotic normal distribution. In Section 3, I propose a two-step GEE estimator in a QMLE framework. In Section 4, the asymptotic distributions for QMLE and GEE for spatial data are derived. In Section 5, consistent variance covariance estimators are provided for the nonlinear estimators for spatial data. In Section 6, we look in detail at a Probit model with spatial correlation in the error term of the latent variable and a count data model with a multiplicative spatial error term. Section 7 contains Monte Carlo simulation results which compare efficiency of different estimation methods for the two nonlinear models explored in the previous section. Section 8 concludes. Section 10 is the appendix.

2.2 M-estimation

In this section, I will examine the M-estimation framework of nonlinear models with spatial data. 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: we need uniform law of large numbers (ULLN) and a central limit theorem (CLT). The GEE procedure is a two-step M-estimation method within the QMLE framework. The M-estimator with spatial data is proved to be consistent and asymptotically normal under certain assumptions.

2.2.1 M-estimation

Assume that spatial processes reside on a regular lattice[3] $\mathbf{D}$ in a Euclidean space, $\mathcal{R}^d, d \geq 1$.[4] Let s denote a location in $\mathbf{D}$. Suppose we have a sample of N observations. Let $\mathbf{D}_N \subseteq \mathbf{D}$ contains the location information for this sample. Let s_i denote the location of the observation i, $i = 1, 2, ..., N$. Let d_{ij} be the pairwise distance between location s_i and location s_j. That is, d_{ij} is the pairwise distance between observation i and observation j. I first give some regularity conditions for the spatial processes I am studying and then I give a general framework for M-estimators with spatial data.

Following Jenish and Prucha (2009) **Definition 1**, I adopt the follwing definitions of mixing conditions for underlying random field. For $U \subseteq \mathbf{D}_N$ and $V \subseteq \mathbf{D}_N$, define σ-algebras $\sigma(U) = \sigma(\mathbf{x}_i; i \in U)$ and $\sigma(V) = \sigma(\mathbf{x}_i; i \in V)$. $|U|$ and $|V|$ denote the cardinality of U and V. The two commonly used mixing conditions are α-mixing and ϕ-mixing which are introduced separately by Rosenblatt and Ibragimov. The α-mixing and ϕ-mixing conditions are:

$\alpha_N(U, V) = \sup\left(|P(u \cap v) - P(u)P(v)|, u \in U, v \in V\right),$

[3] A lattice is a collection of spatial sites (locations) supplemented with neighborhood information (Cressie 1993, p. 383).

[4] A two-dimensional Euclidean space is called the Cartesian plane. Spatial processes can also reside in a higher dimension of space, $\mathcal{R}^n, n > 2$.

and $\phi_N(U,V) = \sup(|P(u|v) - P(u)|, u \in U, v \in V, P(u) > 0)$.

Define a metric $\rho(i,j) = \max_{1 \leq l \leq d}|i_l|$, where i_l denotes the l-th component of i. The mixing conditions for the underlying random fields are defined as follows:

$\alpha_{k,l,N}(r) = \sup(\alpha_N(U,V), |U| \leq k, |V| \leq l, \rho(U,V) \geq r),$

$\phi_{k,l,N}(r) = \sup(\phi_N(U,V), |U| \leq k, |V| \leq l, \rho(U,V) \geq r)$, with k, l, r, N natural numbers. Further let $\bar{\alpha}_{k,l,N}(r) = \sup_N \alpha_{k,l,N}(r)$ and $\bar{\phi}_{k,l,N}(r) = \sup_N \phi_{k,l,N}(r)$.

Let $\{\mathbf{w}_N\} = \{(\mathbf{x}_i, y_i)\}$, $i = 1, 2, ..., N$. $(\mathbf{x}_i, y_i)$ is the observation obtained at location s_i. $\mathbf{x}_i$ is a row vector of independent variables and y_i is a scalar dependent variable. $\theta \in \boldsymbol{\Theta}$ is a $P \times 1$ parameter vector, and θ_0 is the true parameter value. Let θ be a general notation for the parameter vector. An objective function Q_N depends on a sample of realizations of variables $\mathbf{w}_N$, location information $\mathbf{D}_N$, parameter θ and the sample size N. An M-estimator of θ_0 is given by minimizing the objective function Q_N as follows,

$$\hat{\boldsymbol{\theta}}_N = \arg\min_{\theta \in \boldsymbol{\Theta}} Q_N(\mathbf{w}_N, \mathbf{D}_N; \theta). \tag{2.1}$$

In particular, I will address the case when $Q_N(\mathbf{w}_N, \mathbf{D}_N; \theta)$ can be expressed as a sample average. An example of this type of M-estimator is partial (pooled) maximum likelihood (PMLE) estimator. The objective function can be written as

$$Q_N(\mathbf{w}_N, \mathbf{D}_N; \theta) = \frac{1}{N}\sum_{i=1}^{N} q_i(\mathbf{w}_i, \mathbf{D}_N; \theta), \tag{2.2}$$

where $q_i(\mathbf{w}_i, \mathbf{D}_N; \theta)$ is some real valued function defined on $\boldsymbol{\Theta}$. $\mathbf{w}_i$ is the observed data obtained at location s_i. $\mathbf{D}_N$ contains the location information of $\mathbf{w}_i$ and other observations. For simplicity reason, let $q_i(\theta) \equiv q_i(\mathbf{w}_i, \mathbf{D}_N; \theta)$ and $Q_N(\theta) \equiv Q_N(\mathbf{w}_N, \mathbf{D}_N; \theta)$. I will drop $\mathbf{w}$ and $\mathbf{D}$ unless they are needed for clarity.

Suppose that in a parametric model, conditional mean is correctly specified. Let $\mathrm{E}(y_i|\mathbf{x}_i, \mathbf{D}_N) = m_i\left(\mathbf{x}_i, \mathbf{D}_N; \theta_0\right)$ be a correctly specified mean function along with a LEF density $f_i\left(y_i|\mathbf{x}_i;\theta\right)$. For example, in a nonlinear regression model, the objective function for the nonlinear weighted least squares (NWLS) estimator is $\frac{1}{N}\sum_{i=1}^{N}\left\{\left[y_i - m_i\left(\mathbf{x}_i;\theta\right)\right]^2/v_i\right\}$, where v_i is the variance of the error term (usually based on the LEF density), while for the partial maximum likelihood estimation the objective function is $\frac{1}{N}\sum_{i=1}^{N}\log f_i\left(y_i|\mathbf{x}_i, \mathbf{D}_N;\theta\right)$, with $\mathrm{E}(y_i|\mathbf{x}_i, \mathbf{D}_N) = m_i\left(\mathbf{x}_i, \mathbf{D}_N; \theta_0\right)$.

For the M-estimator to be consistent, we need a uniform law of large numbers (ULLN). To derive the ULLN, we need the following assumptions.

Assumption 1: The pairwise distances d_{ij} are finite and lower bounded by some $\varepsilon > 0$. I employ increasing domain asymptotics. That is, the sample size grows as the sampling region expands.

Assumption 2: $\mathbf{x}_i, y_i$ are uniformly bounded variables.

Assumption 3: $\mathbf{\Theta}$ is a compact subset on $\mathcal{R}^p$.

Assumption 4 (**Pointwise Convergence**)**:** For each $\theta \in \mathbf{\Theta}$, $Q_N\left(\theta\right) - \bar{Q}_N\left(\theta\right) = op\left(1\right)$, where $\bar{Q}_N\left(\theta\right) = \mathrm{E}\left(Q_N\left(\theta\right)\right) = \frac{1}{N}\sum_{i=1}^{N}\mathrm{E}\left(q_i\left(\theta\right)\right)$, and $\lim_{N\to\infty}\bar{Q}_N\left(\theta\right) = \ddot{Q}$.

Assumption 5: $Q_N\left(\theta\right)$ is stochastically equicontinuous. Let ($\mathbf{\Theta}$,v) be a metric space. Let $B\left(\theta', \delta\right)$ be the open ball $\left\{\theta \in \mathbf{\Theta} : v\left(\theta, \theta'\right) < \delta\right\}$. $Q_N\left(\theta\right)$ is stochastically equicontinuous in the sense that, for every $\varepsilon > 0$,

$$\limsup_{N\to\infty} P\left(\sup_{\theta,\theta'\in\mathbf{\Theta}, v\left(\theta,\theta'\right)}\left|Q_N\left(\theta\right) - Q_N\left(\theta'\right)\right| > \varepsilon\right) \to 0 \quad \text{as} \quad \delta \to 0\ .$$

Assumption 6: $\bar{Q}_N\left(\theta\right)$ is also stochastically equicontinuous on $\mathbf{\Theta}$. The definition is similar to the statement above for $Q_N\left(\theta\right)$ to be stachostically equicontinuous.

Assumption 7: $\ddot{Q}$ attains unique global minimization at $\theta_0 \in \mathbf{\Theta}$.

Assumption 8: No perfect multicollinearity in $\mathbf{x}_i$. For exponential and logistic regression functions, the objective function is a function of a linear function of independent variables. Multicollinearity should be ruled out in order to identify the model.

Assumption 4 provides a pointwise law of large numbers. **Assumption 7** and **8** provide the identification conditions that make sure the model has unique solution to the minimization problem.

Proposition 6 *Under* ***Assumptions 1-8,*** *the M-estimator in 2.2 is consistent, that is, $\hat{\theta}_N \rightarrow^{a.s.} \theta_0$ as $N \rightarrow \infty$. See proof in Chapter 4.*

Following the above proposition, we can apply the ULLN to different nonlinear estimators. The above M-estimator can be expressed utilizing groups of observations (group notation). That is, we can divide the observations into groups according to geographical properties or other economic or social relationships. Then we can write the objective function as

$$Q_G\left(\mathbf{w}_G, \mathbf{D}_G; \theta\right) = \frac{1}{G}\sum_{i=1}^{G} q_g\left(\mathbf{w}_g, \mathbf{D}_G; \theta\right). \tag{2.3}$$

Let $q_g(\theta) \equiv q_g\left(\mathbf{w}_g, \mathbf{D}_G; \theta\right)$, which is a real valued function of the gth group of observations, $g = 1, 2, ..., G$. $\mathbf{w}_g$ contains the observations of the gth group. $\mathbf{D}_G$ represents the lattice with group information other than just locations. Q_G denotes the objective function which indicate that the total number of groups is G, although the total number of observations is still N. Note that when the group size is equal to one, the group notation is the same as the individual notation. I will use the group notation in the rest of this paper.The above **Assumptions 1-8** for the group notation are basically the same as the individual notation (equation (1) and equation (2)) except that the subscript i changes into g. Let L be the number of observations in each group, then the total number of groups $G = N/L$.

$$\hat{\boldsymbol{\theta}}_G = \arg\min_{\theta \in \boldsymbol{\Theta}} Q_G\left(\mathbf{w}_G, \mathbf{D}_G; \theta\right) \tag{2.4}$$

Proposition 7 *Follow Proposition 1, the groupwise M-estimator in 2.3 is consistent, that is, $\hat{\theta}_G \rightarrow^{a.s.} \theta_0$ as $G \rightarrow \infty$. See proof in Chapter 4.*

2.2.2 Two-step Estimation

In some situations, we have a preliminary estimator. For example, from a partial QMLE estimator, we can get a preliminary consistent estimator. After that we can get an estimated working covariance matrix as weight to get a more efficient estimator. A two-step M-estimator $\hat{\theta}_G$ of θ_0 solves the problem

$$\hat{\theta}_G = \arg\min_{\theta \in \boldsymbol{\Theta}} Q_G\left(\mathbf{w}_G, \mathbf{D}_G; \theta, \hat{\gamma}\right), \tag{2.5}$$

$$Q_G\left(\mathbf{w}_G, \mathbf{D}_G; \theta, \hat{\gamma}\right) = \frac{1}{G}\sum_{g=1}^{G} q_g\left(\mathbf{w}_g, \mathbf{D}_G; \theta, \hat{\gamma}\right),$$

where $\hat{\gamma}$ is a preliminary estimator based on the sample $\left\{\mathbf{w}_g : g = 1, 2, ..., G\right\}$ which exhibits spatial correlations. $p\lim \hat{\gamma} = \gamma^*$, where γ^* is some element in the parameter space $\boldsymbol{\Gamma}$. I will discuss a specific example of the two-step M-estimator in the next section, the PMLE and GEE method.

Assumption 9: $\bar{Q}_G\left(\theta, \gamma^*\right)$ attains unique global minimization at $\theta_0 \in \boldsymbol{\Theta}$, where $\bar{Q}_G\left(\theta, \gamma^*\right) = \frac{1}{G}\sum_{g=1}^{G} \mathrm{E}\left[q_g\left(\mathbf{w}_g, \mathbf{D}_G; \theta, \gamma^*\right)\right]$; That is $\bar{Q}_G\left(\theta_0, \gamma^*\right) < \bar{Q}_G\left(\theta, \gamma^*\right)$, for all $\theta \in \boldsymbol{\Theta}, \theta \neq \theta_0$.

Assumption 9 provides necessary identification condition for the two-step M-estimator. If $q_g\left(\mathbf{w}_i, \theta_0; \gamma\right)$ is stachastically equicontinuous over $\mathbf{\Theta} \times \mathbf{\Gamma}$, then a ULLN applies. Along with identification, this result can be shown to imply consistency of $\hat{\theta}_G$ for θ_0. The consistency argument is the same as Proposition 1.

2.3 Two-step Estimator: QMLE and GEE

Due to the complexity of the joint distribution of spatial random processes, econometricians have developed a variety of ways to reduce the computational burden. One way is to specify the partial conditional distribution, and maximize the summand of log likelihoods for each observation. The parameters can be consistently estimated if the partial log likelihood function satisfies the assumptions for consistency of M-estimation. A consistent variance estimator should be provided for valid inference[5]. Moreover, one can divide data into different groups, and specify the marginal distribution for each group to get a more efficient estimator (Wang, Iglesias, & Wooldridge 2012). However, this approach requires correctly specified marginal distributions, and when we increase the group size, the joint distribution for each group of variables becomes more and more difficult to compute. In this section, I propose a two-step estimator in a QMLE framework. The first step is a pooled QMLE procedure and the second step is a GEE procedure. Only the correct conditional mean and a density function in the LEF need to be specified.

The partial (pooled) QMLE method requires correctly specified conditional mean functions, $\mathrm{E}(y_i|\mathbf{x}_i) = m_i\left(\mathbf{x}_i; \theta_0\right), i = 1, 2, ..., N$, along with a LEF density $f_i\left(y_i|\mathbf{x}_i; \theta\right)$. Then

[5] 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 extent of how much one ignores it in the estimation.

one proceeds with minimizing the sum of log individual likelihoods ignoring any spatial dependence. Note that, the true log likelihood cannot be written by the sum of the individual likelihoods. PQMLE is an approximation of the true MLE. However, under certain conditions, PQMLE delivers consistent estimators with computation ease.

The the partial quasi-log likelihood is

$$L_N(\theta) = \frac{1}{N}\sum_{i=1}^{N} \log f_i(y_i|\mathbf{x}_i, \mathbf{D}_N; \theta). \tag{2.6}$$

The partial QMLE is found by solving the score function,

$$S_N(\check{\theta}) = \sum_{i=1}^{N} \mathbf{s}_i(\check{\theta}) = \mathbf{0}. \tag{2.7}$$

One characterization of QMLE in LEF is that the individual score function has the following form:

$$\mathbf{s}_i(\theta) = \nabla m_i(\mathbf{x}_i, \mathbf{D}_N; \theta)' [y_i - m_i(\mathbf{x}_i, \mathbf{D}_N; \theta)] / v(m_i(\mathbf{x}_i, \mathbf{D}_N; \theta)), \tag{2.8}$$

where $\nabla m_i(\mathbf{x}_i, \mathbf{D}_N; \theta)$ is the $1 \times P$ gradient of the mean function and $v(m_i(\mathbf{x}_i, \mathbf{D}_N; \theta))$ is the variance function associated with the chosen LEF density. For Bernoulli,

$$v(m_i(\mathbf{x}_i, \mathbf{D}_N; \theta)) = m_i(\mathbf{x}_i, \mathbf{D}_N; \theta)(1 - m_i(\mathbf{x}_i, \mathbf{D}_N; \theta)),$$

and for Poisson distribution,

$$v(m_i(\mathbf{x}_i, \mathbf{D}_N; \theta)) = m_i(\mathbf{x}_i, \mathbf{D}_N; \theta).$$

$\mathrm{E}(\mathbf{s}_i(\theta)|\mathbf{x}_i, \mathbf{D}_N) = \mathbf{0}$ if $\mathrm{E}(y_i|\mathbf{x}_i, \mathbf{D}_N) = m_i(\mathbf{x}_i, \mathbf{D}_N; \theta_0)$, which implies Fisher consistency.

The above gives a consistent estimator by **Proposition 1**. However, this estimator is not likely to be the most efficient estimator among the estimators that are based on the same distributional assumptions, because it ignores the spatial correlations between observations. If we can use some or all pairwise correlation information, we can possibly improve the estimation efficiency. A common way to make use of the pairwise information is to divide observations into groups, and use spatial correlations within groups while ignore correlations between groups. In empirical studies, there exist some natural groups of data, e.g., the technology spillover effects within a certain state. Suppose we know the groupwise distribution but not the full distribution of the whole data, we can get a consistent estimator by using only groupwise information.

Let g be the number of groups, $g = 1, 2, ..., G$. The group size is the number of observations divided by the number of groups, and it can vary from $1, 2, ...,$ to N. 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 partial QMLE 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 pairwise information. If the group size is not equal to one or N, the estimation is actually a partial QMLE. By "partial", I mean that I do not use full information, only the information within groups. Suppose we divide data into G groups and assume that there are the same number of observations in each group. Let $L = N/G$, which is fixed. The group numbers and the sample size both increase as the sampling domain increases. That is, we get more observations by increasing the space where we obtain a sample. For group g, $\mathbf{X}_g$ is an $L \times K$ matrix and $\mathbf{y}_g$ is an $L \times 1$ vector, where g denotes the gth group, $g = 1, 2, ..., G$. Let $\mathbf{X}$ denote the $N \times P$ covariate matrix. We can write the assumptions in terms of group notation. Those assumptions are basically same, the difference is the notation. Thus I will not readdress the assumptions again.

Assumption 10: Conditional mean is correctly specified for each group. $\mathrm{E}\left(\mathbf{y}_g|\mathbf{X}_g\right) = \mathbf{m}_g\left(\mathbf{X}_g, \mathbf{D}_G; \theta\right) \equiv \mathbf{m}_g, g = 1, 2, ..., G$. I will use $\mathbf{m}_g$ for short and $\mathbf{m}_g\left(\cdot\right)$ to emphasize certain parameters. When $G = N$, we get $\mathrm{E}(y_i|\mathbf{x}_i) = m_i\left(\mathbf{x}_i, \mathbf{D}_N; \theta_0\right)$.

The QMLE estimator is given by setting the score function equal to $\mathbf{0}$.

$$\sum_{g=1}^{G} \mathbf{s}_g\left(\hat{\theta}\right) = \mathbf{0}, \tag{2.9}$$

where $\mathbf{s}_g\left(\theta\right)$ is the score function for each group. And the group score $\mathbf{s}_g\left(\theta\right)$ has the following form,

$$\mathbf{s}_g\left(\theta\right) = \nabla\mathbf{m}_g'\mathbf{W}_g^{-1}\left(\mathbf{y}_g - \mathbf{m}_g\right). \tag{2.10}$$

where $\nabla\mathbf{m}_g$ is the $L \times P$ gradient of the group mean function and $\mathbf{W}_g$ is the LEF variance covariance matrix for group g. Notice that $\mathbf{W}_g$ is not a diagonal matrix that only contain the variances of each individual, but also contains the covariances of pairwise individuals. It is because of this property that we can improve efficiency by doing a so called generalized estimating equations (GEE) approach. The GEE approach was first extended to correlated data by Zeger and Liang (1986). In the spatial data context, I propose that the generalized estimating equations (GEE) for the mean parameters, which is given by

$$\sum_{g=1}^{G} \nabla\mathbf{m}_g'\mathbf{W}_g^{-1}\left(\mathbf{y}_g - \mathbf{m}_g\right) = \mathbf{0}. \tag{2.11}$$

In order to use GEE, we need to get a consistent estimator for $\mathbf{W}_g$ which depends on the pairwise distances and a spatial dependence parameter. Suppose $\hat{\mathbf{W}}_g$ is a consistent estimator for $\mathbf{W}_g$. GEE is a "pseudo" weighted multivariate nonlinear least squares (MWNLS), because GEE only use the groupwise information. The GEE estimator is given by:

$$\hat{\theta}_{GEE} = \arg\min_{\theta} \sum_{g=1}^{G} \left(\mathbf{y}_g - \mathbf{m}_g\right)' \hat{\mathbf{W}}_g^{-1} \left(\mathbf{y}_g - \mathbf{m}_g\right). \tag{2.12}$$

As a special case, pooled QMLE is the same as the nonlinear weighted least squares estimator (NWLS):

$$\tilde{\theta}_{NWLS} = \arg\min_{\theta} \sum_{i=1}^{N} \left[y_i - m_i\left(\mathbf{x}_i, \mathbf{D}_N; \theta\right)\right]^2 / v\left(m_i\left(\mathbf{x}_i, \mathbf{D}_N; \theta\right)\right). \tag{2.13}$$

The following demonstrates how to find a consistent estimator for $\mathbf{W}_g$. We can write

$$\mathbf{W}_g = \mathbf{V}_g(\mathbf{X}_g; \theta)^{1/2} \mathbf{R}_g\left(\rho, \mathbf{D}_G\right) \mathbf{V}_g(\mathbf{X}_g; \theta)^{1/2}.$$

The diagonal elements of $\mathbf{W}_g$ correspond to the variances of dependent variables drawn from a density in LEF. The off-diagonal elements are the covariances that depend on the spatial parameter and distances.

$$\mathbf{V}_g(\mathbf{X}_g, \mathbf{D}_G; \theta) = \begin{pmatrix} v_1 & \cdots & & 0 \\ 0 & v_2 & & \vdots \\ \vdots & & \ddots & 0 \\ 0 & ... & 0 & v_L \end{pmatrix}, \tag{2.14}$$

where the lth element on the diagonal is $\mathbf{v}_l = \mathrm{Var}(\mathbf{y}_{gl} | \mathbf{X}_{gl})$ 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 $\mathbf{X}_g$.

Let $\mathbf{R}_N\left(\rho, \mathbf{D}_N\right)$ be the $N \times N$ correlation matrix for the whole sample, and let $\mathbf{R}_g\left(\rho, \mathbf{D}_G\right)$

be the $L \times L$ correlation matrix for the group g. A common assumption of the ijth element of $\mathbf{R}_N(\rho, \mathbf{D}_N)$ is that

$$\mathbf{R}_{ij} = 1 - \gamma\left(d_{ij}, \rho\right), \gamma\left(d_{ij}, h\right) = \begin{cases} 0 & \text{if } d_{ij} = 0, \\ c + b\left[1 - \exp\left(-d_{ij}/\rho\right)\right] & \text{otherwise,} \end{cases} \tag{2.15}$$

where the vector of spatial parameters $h = (c, b, \rho), c \geq 0, b \geq 0, \rho \geq 0$, and $c + b \leq 2$.[6]

Set $b = c = 1$ without loss of generality. Then

$$\mathbf{R}_{ij} = \begin{cases} 1 & \text{if } d_{ij} = 0, \\ \exp\left(-d_{ij}/\rho\right) & \text{otherwise.} \end{cases} \tag{2.16}$$

Another example would be

$$\mathbf{R}_{ij} = \begin{cases} 1 & \text{if } d_{ij} = 0, \\ \rho/d_{ij} & \text{otherwise.} \end{cases} \tag{2.17}$$

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

Let $\check{\theta}$ be the partial QMLE estimator. $\hat{u}_i = y_i - m_i\left(x_i; \check{\theta}\right)$, for $i = 1, 2, ..., N$, is the QMLE residual. $\check{v}_i = v\left(m_i\left(\mathbf{x}_i, \mathbf{D}_N; \check{\theta}\right)\right)$ is the fitted variance of individual i associated with the chosen LEF density. Let $\hat{r}_i = \check{u}_i/\check{v}_i$ be the standardized residual. Let $\hat{\mathbf{r}} = (\hat{r}_1, \hat{r}_2, ..., \hat{r}_N)'$. Then $\hat{\mathbf{r}}\hat{\mathbf{r}}'$ is the sample correlation matrix. We can use a method in Prentice (1988) to find a consistent estimator for ρ. Let $\mathbf{e}$ be a vector containing $N(N-1)/2$ different elements of

[6] See Cressie (1993) p.61 for more examples.

the lower (or upper) triangle of $\hat{\mathbf{r}}\hat{\mathbf{r}}'$, excluding the diagonal. Let $\mathbf{z}$ be the vector containing the elements in $\mathbf{R}$ corresponding to the same entries of elements in $\hat{\mathbf{r}}\hat{\mathbf{r}}'$. We can get the parameter estimator $\hat{\rho}$ by solving:

$$\hat{\rho} = \arg\min(\mathbf{e}-\mathbf{z})'\boldsymbol{\Xi}^{-1}(\mathbf{e}-\mathbf{z}), \tag{2.18}$$

where $\boldsymbol{\Xi}$ is the working correlation matrix for $\mathbf{e}$, the sample correlation vector. Ξ is a diagonal matrix with $\left(\xi_{21}, \xi_{31}, ..., \xi_{n1}, \xi_{32}, \xi_{42}, ..., \xi_{N2}, ..., \xi_{N,N-1}\right)$ as the elements on the diagonal, which are the corresponding variances of element in $\mathbf{e}$. If the variance covariance matrix $\mathbf{W}$ is correctly specified, a model-based consistent variance estimator of $\hat{\theta}$ is $\left(\sum_{g=1}^{G} \nabla\hat{\mathbf{m}}_g' \hat{\mathbf{W}}_g^{-1} \nabla\hat{\mathbf{m}}_g\right)^{-1}$, where $\nabla\hat{\mathbf{m}}_g = \nabla\hat{\mathbf{m}}_g\left(\mathbf{X}_g; \hat{\theta}_{GEE}\right)$. As an alternative, the variance estimator of $\hat{\theta}$ that is robust to misspecification of the variance covariance matrix is

$$\left(\sum_{g=1}^{G} \nabla\hat{\mathbf{m}}_g' \hat{\mathbf{W}}_g^{-1} \nabla\hat{\mathbf{m}}_g\right)^{-1} \left(\sum_{g=1}^{G}\sum_{h=1}^{G} \nabla\hat{\mathbf{m}}_g' \hat{\mathbf{W}}_g^{-1} k\left(g,h\right) \hat{\mathbf{u}}_g \hat{\mathbf{u}}_h' \hat{\mathbf{W}}_h^{-1} \nabla\hat{\mathbf{m}}_h\right)$$
$$\left(\sum_{g=1}^{G} \nabla\hat{\mathbf{m}}_g' \hat{\mathbf{W}}_g^{-1} \nabla\hat{\mathbf{m}}_g\right)^{-1}, \tag{2.19}$$

where $k\left(g,h\right)$ is the kernel function depending on the distance between group g and h.

The GEE approach is summarized as follows:

First, find the partial QMLE estimator for the mean parameters and obtain the residuals from PQMLE.

Second, use the first step estimator to get a fitted variance covariance matrix according to the LEF density, and obtain the spatial correlation parameter using the standardized residuals from the first step. After obtaining a working matrix, undertake a multivariate weighted least squares procedure. This gives the GEE estimator.

Finally, we can get a consistent variance estimator for the mean parameter that is robust to heteroskedasticity and spatial correlation. Further, we can obtain the average partial effect (APE).

2.4 Asymptotic Distribution for PQMLE and GEE

A central limit theorem is needed to develop the asymptotic distributions for M-estimators. Bolthausen (1982) provides central limit theorem (CLT) for strictly stationary processes. However, in economic applications there are a lot of nonstationary spatial processes in the sense that they are heterogenous; that is, the joint distribution of dependent variables varies with locations. There are also spatial processes which may have asymptotically unbounded moments. However, in this paper, I will only discuss uniformly bounded random variables.

From the above equation, we can take derivatives with respect to the parameter, and get the scores for each group. Let $\mathbf{s}_g(\theta)$ denote the $P \times 1$ vector of score for $q_g(\theta)$. The group score $\mathbf{s}_g(\theta)$ and Hessian $\mathbf{H}_g(\theta)$ have the following forms,

$$\mathbf{s}_g(\theta) = \nabla \mathbf{m}_g'(\theta) \mathbf{W}_g^{-1} \left[\mathbf{y}_g - \mathbf{m}_g(\theta)\right], \tag{2.20}$$

$$\mathrm{E}\left(\mathbf{s}_g(\theta) | \mathbf{x}_g, \mathbf{D}_G\right) = \mathbf{0}.$$

By the assumption of correctly specified conditional mean, the above condition implies Fisher consistency for QMLE for linear exponential family.

While

$$\begin{aligned}\mathbf{H}_g(\theta) &= \partial \mathbf{s}_g(\theta)/\partial\theta' && (2.21)\\ &= -\nabla_\theta \mathbf{m}_g' \mathbf{W}_g^{-1} \nabla \mathbf{m}_g + \partial \nabla_\theta^2 \mathbf{m}_g' \mathbf{W}_g^{-1}(\mathbf{y}_g - \mathbf{m}_g),\end{aligned}$$

where $\nabla_\theta \mathbf{m}_g \equiv \partial \mathbf{m}_g/\partial\theta$ is the $L \times P$ gradient of the group mean function, $\partial\nabla_\theta^2 \mathbf{m}_g \equiv \partial^2 \mathbf{m}_g/\partial\theta\partial\theta'$ is the $L \times P$ jacobian of the group mean function and $\mathbf{W}_g$ is the LEF variance covariance matrix for group g. Taking the expected value of the score function over the distributions of $\mathbf{w}$ gives

$$\begin{aligned}\mathrm{E}\left[\mathbf{H}_g(\theta_0)\right] &= \mathrm{E}\left\{\mathrm{E}\left[\mathbf{H}_g(\theta_0)|\mathbf{w}_g, \mathbf{D}_G\right]\right\} && (2.22)\\ &= \mathrm{E}\left\{-\nabla \mathbf{m}_g' \mathbf{W}_g^{-1} \nabla \mathbf{m}_g + \mathrm{E}\left[\partial \nabla_\theta^2 \mathbf{m}_g' \mathbf{W}_g^{-1}(\mathbf{y}_g - \mathbf{m}_g)|\mathbf{w}_g, \mathbf{D}_G\right]\right\}\\ &= \mathrm{E}\left\{\left(-\nabla \mathbf{m}_g' \mathbf{W}_g^{-1} \nabla \mathbf{m}_g\right) + \partial \nabla_\theta^2 \mathbf{m}_g' \mathbf{W}_g^{-1}\left[\mathrm{E}\left(\mathbf{y}_g|\mathbf{w}_g, \mathbf{D}_G\right) - \mathbf{m}_g\right]\right\}\\ &= \mathrm{E}\left(-\nabla \mathbf{m}_g' \mathbf{W}_g^{-1} \nabla \mathbf{m}_g\right) + \partial \nabla_\theta^2 \mathbf{m}_g' \mathbf{W}_g^{-1}\left[\mathbf{m}_g - \mathbf{m}_g\right]\\ &= \mathrm{E}\left(-\nabla \mathbf{m}_g' \mathbf{W}_g^{-1} \nabla \mathbf{m}_g\right)\end{aligned}$$

Notice that $\mathbf{W}_g$ is not a diagonal matrix that only contain the variances of each individual, but also contains the covariances of pairwise individuals. It is because of this property that we can improve efficiency by doing a so called generalized estimating equations (GEE) approach. The GEE approach was first extended to correlated data by Zeger and Liang (1986). In the spatial data context, I propose that the generalized estimating equations (GEE) for the mean parameters are given by

$$\frac{1}{G}\sum_{g=1}^{G} \mathbf{s}_g\left(\hat{\theta}\right) = \mathbf{0}, \tag{2.23}$$

$$\frac{1}{G}\sum_{g=1}^{G}\nabla\mathbf{m}_g'\left(\hat{\theta}\right)\mathbf{W}_g^{-1}\left[\mathbf{y}_g-\mathbf{m}_g\left(\hat{\theta}\right)\right]=\mathbf{0}. \tag{2.24}$$

Because each score is a function of spatial processes, they are correlated with each other. The score function for the total sample is $S_G\left(\hat{\theta}\right)=\frac{1}{G}\sum_{g=1}^{G}\mathbf{s}_g\left(\hat{\theta}\right)=0$. The score function can be expanded about θ_0 in a mean-value expansion:

$$S_G\left(\hat{\theta}\right)=\frac{1}{G}\sum_{g=1}^{G}s_g\left(\theta_0\right)+\frac{1}{G}\sum_{g=1}^{G}H_g\left(\ddot{\theta}\right)\left(\hat{\theta}-\theta_0\right). \tag{2.25}$$

where $\ddot{\theta}\in\mathbf{\Theta}$ is between $\hat{\theta}$ and θ_0. $H_g\left(\ddot{\theta}\right)$ is the $P\times P$ Hessian of the objective function $q_g\left(\theta\right)$.

$$\sqrt{G}\left(\hat{\theta}-\theta_0\right)=\left[-\frac{1}{G}\sum_{g=1}^{G}\mathbf{H}_g\left(\ddot{\theta}\right)\right]^{-1}\frac{1}{\sqrt{G}}\sum_{g=1}^{G}s_g\left(\theta_0\right). \tag{2.26}$$

Assumption 11 (**Uniform $L_{2+\delta}$ integrability**): The elements in the scores are uniformly bounded and have a limit expectation equal to zero. That is,

$$\lim_{k\to\infty}\sup_{g}\mathrm{E}\left[\left|s_{gl}\right|^{2+\delta}\mathbf{1}\left(s_{gl}>k\right)\right]=0, \tag{2.27}$$

where $\mathbf{1}\left(\cdot\right)$ is the indicator function and s_g is the group score matrix, which is $P\times L$. s_{gl} is an element in the score matrix, and k is a constant.

Assumption 12: The second moment of the score function is positive and uniformly bounded.

$$0<\lim_{G\to\infty}\frac{1}{G}\mathrm{Var}\left(\sum_{g=1}^{G}s_g\left(\theta_0\right)\right)<\infty. \tag{2.28}$$

Proposition 8 *Under **Assumptions 1-12**,* $\sqrt{G}\left(\check{\theta}-\theta_0\right)\Rightarrow\mathrm{N}\left(\mathbf{0},\mathbf{A}_0^{-1}\mathbf{B}_0\mathbf{A}_0^{-1}\right)$, *where* $\mathbf{A}_0=\lim_{G\to\infty}\left\{-\frac{1}{G}\sum_{g=1}^{G}\mathrm{E}\left[\mathbf{H}_g\left(\theta_0\right)\right]\right\}$, *and* $\mathbf{B}_0=\lim_{G\to\infty}\mathrm{Var}\left[\frac{1}{\sqrt{G}}\sum_{g=1}^{G}s_g\left(\theta_0\right)\right]$. *See proof in Chapter 4.*

$\mathrm{E}\left[\mathbf{H}_g\left(\theta_0\right)\right]$ has already been given in Equation (22). $\mathrm{Var}\left[\frac{1}{\sqrt{G}}\sum_{g=1}^{G} s_g\left(\theta_0\right)\right]$ is given as follows,

$$\begin{aligned}
\mathrm{Var}\left[\frac{1}{\sqrt{G}}\sum_{g=1}^{G} s_g\left(\theta_0\right)\right] &= \mathrm{Var}\left\{\frac{1}{\sqrt{G}}\sum_{g=1}^{G} \nabla\mathbf{m}_g'\left(\theta_0\right)\mathbf{W}_g^{-1}\left[\mathbf{y}_g-\mathbf{m}_g\left(\theta_0\right)\right]\right\} \qquad (2.29)\\
&= \mathrm{Var}\left[\frac{1}{\sqrt{G}}\sum_{g=1}^{G} \nabla\mathbf{m}_g'\left(\theta_0\right)\mathbf{W}_g^{-1}\mathbf{u}_g\right]\\
&= \frac{1}{G}\sum_{g=1}^{G}\mathrm{E}\left[\nabla\mathbf{m}_g'\left(\theta_0\right)\mathbf{W}_g^{-1}\mathbf{u}_g\mathbf{u}_g'\mathbf{W}_g^{-1}\nabla\mathbf{m}_g\left(\theta_0\right)\right]\\
&\quad +\frac{1}{G}\sum_{g=1}^{G}\sum_{g\neq h}^{G}\mathrm{E}\left[\nabla\mathbf{m}_g'\left(\theta_0\right)\mathbf{W}_g^{-1}\mathbf{u}_g\mathbf{u}_h\mathbf{W}_h^{-1}\nabla\mathbf{m}_h\left(\theta_0\right)\right].
\end{aligned}$$

If we have a first step estimator, say $\hat{\gamma}$, $p\lim\hat{\gamma}=\gamma^*$. By linear expansion, the score should be written as

$$\begin{aligned}
S_G\left(\hat{\theta}\right) &= \frac{1}{G}\sum_{g=1}^{G}\mathbf{s}_g\left(\theta_0,\hat{\gamma}\right)+op\left(1\right) \qquad (2.30)\\
&= \frac{1}{G}\sum_{g=1}^{G}\mathbf{s}_g\left(\theta_0,\gamma^*\right)+\mathbf{F}_0\left(\hat{\gamma}-\gamma^*\right)+op\left(1\right),
\end{aligned}$$

where

$$\mathbf{s}_g\left(\theta_0,\hat{\gamma}\right)=\nabla\mathbf{m}_g'\left(\theta\right)\hat{\mathbf{W}}_g^{-1}\left[\mathbf{y}_g-\mathbf{m}_g\left(\theta\right)\right]=\mathbf{0}.$$

where $\hat{\mathbf{W}}_g^{-1}$ is a function of $\hat{\gamma}$, $\mathbf{F}_0$ is a $P\times J$ matrix. J is the dimension of γ. $\mathbf{F}_0=\lim_{g\to\infty}\frac{1}{G}\sum_{g=1}^{G}\mathrm{E}\left[\nabla_\gamma\mathbf{s}_g\left(\mathbf{w}_g,\theta_0;\gamma^*\right)\right]$. We can see $\nabla\mathbf{m}_g'\left(\theta\right)$ and $\mathbf{m}_g\left(\theta\right)$ do not rely on γ. When we take derivatives with respect to γ, it only matters with $\mathbf{W}_g^{-1}$. $\nabla_\gamma\mathbf{s}_g\left(\mathbf{w}_g,\theta_0;\gamma^*\right)$ is a linear combination of elements of $\left[\mathbf{y}_g-\mathbf{m}_g\left(\theta\right)\right]$. Since $\mathrm{E}[\left(\mathbf{y}_g-\mathbf{m}_g\right)\mid\mathbf{w},\mathbf{D}\ ]=0$,

$\mathrm{E}[\nabla_\gamma \mathbf{s}_g \left(\mathbf{w}_g, \theta_0; \gamma^*\right) | \mathbf{w}, \mathbf{D}] = 0$. By law of iterated expectations, $\mathrm{E}\left[\nabla_\gamma \mathbf{s}_g \left(\mathbf{w}_g, \theta_0; \gamma^*\right)\right] = \mathbf{0}$. Thus $\mathbf{F}_0 = \mathbf{0}$. Then we can write the score function as

$$S_G\left(\hat{\theta}\right) = \frac{1}{G}\sum_{g=1}^{G} \mathbf{s}_g\left(\theta_0, \gamma^*\right) + op\left(1\right), \tag{2.31}$$

and

$$S_G\left(\hat{\theta}\right) = \frac{1}{G}\sum_{g=1}^{G} s_g\left(\theta_0, \gamma^*\right) + \frac{1}{G}\sum_{g=1}^{G} H_g\left(\ddot{\theta}, \gamma^*\right)\left(\hat{\theta} - \theta_0\right) \tag{2.32}$$

where $\ddot{\theta} \in \mathbf{\Theta}$ is between $\hat{\theta}$ and θ_0. $H_g\left(\ddot{\theta}\right)$ is the $P \times P$ Hessian of the objective function $q_g\left(\theta\right)$.

$$\sqrt{G}\left(\hat{\theta} - \theta_0\right) = \left[-\frac{1}{G}\sum_{g=1}^{G} \mathbf{H}_g\left(\ddot{\theta}, \gamma^*\right)\right]^{-1} \frac{1}{\sqrt{G}}\sum_{g=1}^{G} s_g\left(\theta_0, \gamma^*\right). \tag{2.33}$$

Thus the first step estimation will not affect the asymptotic distribution of the second step. γ^* is a fixed number in the second step score function.

Proposition 9 *Under **Assumptions 1-12,*** $\sqrt{G}\left(\hat{\theta} - \theta_0\right) \Rightarrow \mathrm{N}\left(\mathbf{0}, \mathbf{A}_0^{-1}\mathbf{B}_0\mathbf{A}_0^{-1}\right)$, *where* $\mathbf{A}_0 = \lim_{G\to\infty}\left\{-\frac{1}{G}\sum_{g=1}^{G} \mathrm{E}\left[\mathbf{H}_g\left(\theta_0, \gamma^*\right)\right]\right\}$, *and* $\mathbf{B}_0 = \lim_{G\to\infty} \mathrm{Var}\left[\frac{1}{\sqrt{G}}\sum_{g=1}^{G} s_g\left(\theta_0, \gamma^*\right)\right]$. *See proof in Chapter 4.*

2.5 Variance-Covariance Matrix Estimator

2.5.1 Parametric Variance-Covariance Estimator

Assumption 13: $\mathrm{Var}\left(\mathbf{u}_g\right) = \mathrm{Var}\left(\mathbf{u}_g | \mathbf{x}_g, \mathbf{D}_G\right) = \mathbf{W}_g$;

$$\mathrm{Cov}\left(\mathbf{u}_g,\mathbf{u}_h\right)=\mathrm{Cov}\left(\mathbf{u}_g,\mathbf{u}_h|\mathbf{x}_g,\mathbf{x}_h,\mathbf{D}_G\right)=\mathbf{C}_{gh}.$$

$$\begin{aligned}\mathrm{Var}\left[\frac{1}{\sqrt{G}}\sum_{g=1}^{G}\mathbf{s}_g\left(\theta_0\right)\right] &= \frac{1}{G}\sum_{g=1}^{G}\mathrm{E}\left[\nabla\mathbf{m}_g'\left(\theta_0\right)\mathbf{W}_g^{-1}\nabla\mathbf{m}_g\left(\theta_0\right)\right] \\ &\quad+\frac{1}{G}\sum_{g=1}^{G}\sum_{g\neq h}^{G}\mathrm{E}\left[\nabla\mathbf{m}_g'\left(\theta_0\right)\mathbf{W}_g^{-1}\mathbf{C}_{gh}\mathbf{W}_h^{-1}\nabla\mathbf{m}_h\left(\theta_0\right)\right].\end{aligned}\tag{2.34}$$

Under **Assumption 13,** the asymptotic variance estimator for $\hat{\theta}_G$ can be estimated by

$$\begin{aligned}\widehat{\mathrm{Avar}_1}\left(\hat{\theta}\right) &= \frac{1}{G}\hat{\mathbf{A}}^{-1}\hat{\mathbf{B}}_1\hat{\mathbf{A}}^{-1} \\ &= \left(\sum_{g=1}^{G}\nabla\hat{\mathbf{m}}_g'\hat{\mathbf{W}}_g^{-1}\nabla\hat{\mathbf{m}}_g\right)^{-1}\left(\sum_{g=1}^{G}\nabla\hat{\mathbf{m}}_g'\hat{\mathbf{W}}_g^{-1}\hat{\mathbf{C}}_{gh}\hat{\mathbf{W}}_h^{-1}\nabla\hat{\mathbf{m}}_h'\right) \\ &\quad \left(\sum_{g=1}^{G}\nabla\hat{\mathbf{m}}_g'\hat{\mathbf{W}}_g^{-1}\nabla\hat{\mathbf{m}}_g\right)^{-1},\end{aligned}\tag{2.35}$$

where

$$\hat{\mathbf{A}}=\frac{1}{G}\sum_{g=1}^{G}\nabla\hat{\mathbf{m}}_g'\hat{\mathbf{W}}_g^{-1}\nabla\hat{\mathbf{m}}_g,$$

and

$$\hat{\mathbf{B}}_1=\frac{1}{G}\sum_{g=1}^{G}\sum_{h=1}^{G}\nabla\hat{\mathbf{m}}_g'\hat{\mathbf{W}}_g^{-1}\hat{\mathbf{C}}_{gh}\hat{\mathbf{W}}_h^{-1}\nabla\hat{\mathbf{m}}_h'.$$

2.5.2 Nonparametric Variance-Covariance Estimator

Assumption 13 is not always obtained. And most of times it can not be easily known. Since $\mathrm{Avar}\sqrt{G}\left(\hat{\theta}_G-\theta_0\right)=\mathbf{A}_0^{-1}\mathbf{B}_0\mathbf{A}_0^{-1}$, we can consistently estimate $\mathrm{Avar}\sqrt{G}\left(\hat{\theta}_G-\theta_0\right)$ by $\hat{\mathbf{A}}^{-1}\hat{\mathbf{B}}_2\hat{\mathbf{A}}^{-1}$. The asymptotic standard errors are obtained from the matrix $\widehat{\mathrm{Avar}_2}\left(\hat{\theta}_G\right)=$

$\hat{\mathbf{A}}^{-1}\hat{\mathbf{B}}_2\hat{\mathbf{A}}^{-1}/G$. A robust nonparametric variance covariance estimator is given as

$$\begin{aligned}\widehat{\text{Avar}_2}\left(\hat{\theta}\right) &= \frac{1}{G}\hat{\mathbf{A}}^{-1}\hat{\mathbf{B}}_2\hat{\mathbf{A}}^{-1} \qquad (2.36)\\ &= \left(\sum_{g=1}^{G}\nabla\hat{\mathbf{m}}_g'\hat{\mathbf{W}}_g^{-1}\nabla\hat{\mathbf{m}}_g\right)^{-1}\left(\sum_{g=1}^{G}\sum_{h=1}^{G}k(d_{gh})\nabla\hat{\mathbf{m}}_g'\hat{\mathbf{W}}_g^{-1}\hat{\mathbf{u}}_g\hat{\mathbf{u}}_h'\hat{\mathbf{W}}_h^{-1}\nabla\hat{\mathbf{m}}_h'\right)\\ &\quad\left(\nabla\hat{\mathbf{m}}_g'\hat{\mathbf{W}}_g^{-1}\nabla\hat{\mathbf{m}}_g\right),\end{aligned}$$

where

$$\hat{\mathbf{A}} = \frac{1}{G}\sum_{g=1}^{G}\nabla\hat{\mathbf{m}}_g'\hat{\mathbf{W}}_g^{-1}\nabla\hat{\mathbf{m}}_g,$$

and

$$\hat{\mathbf{B}}_2 = \frac{1}{G}\sum_{g=1}^{G}\sum_{h=1}^{G}k(d_{gh})\nabla\hat{\mathbf{m}}_g'\hat{\mathbf{W}}_g^{-1}\hat{\mathbf{u}}_g\hat{\mathbf{u}}_h'\hat{\mathbf{W}}_h^{-1}\nabla\hat{\mathbf{m}}_h'.$$

Proposition 10 *Under **Assumptions 1-13**, $\widehat{\text{Avar}_2}\left(\hat{\theta}_G\right)_{robust} \to \frac{1}{G}\mathbf{A}_0^{-1}\mathbf{B}_0\mathbf{A}_0^{-1}$. See proof in Chapter 4.*

Although we do not need to make adjustment to the two-step QMLE estimator in this paper, it is worth to mention that in lot of cases $\mathbf{F}_0 \neq 0$, and we need to make adjustment to the asymptotic variances.

$$\sqrt{G}\left(\hat{\theta} - \theta_0\right) = \mathbf{A}_0\frac{1}{\sqrt{G}}\sum_{g=1}^{G}\left[-\mathbf{g}_g\left(\theta_0;\gamma^*\right)\right] + op\left(1\right),$$

where

$$\mathbf{g}_g\left(\theta_0;\gamma^*\right) \equiv s_g\left(\theta_0,\gamma^*\right) + \mathbf{F}_0\mathbf{r}_g\left(\gamma^*\right).$$

Let

$$\mathbf{D}_0 \equiv \lim_{G\to\infty}\frac{1}{G}\text{E}\left[\sum_{g=1}^{G}\mathbf{g}_g\left(\theta_0;\gamma^*\right)\sum_{h=1}^{G}\mathbf{g}_h\left(\theta_0;\gamma^*\right)'\right],$$

the asyptotic distribution of $\hat{\theta}$ can be written as

$$\sqrt{G}\left(\hat{\theta}-\theta_0\right) \Rightarrow^d \mathrm{N}\left(0, \mathbf{A}_0^{-1}\mathbf{D}_0\mathbf{A}_0^{-1}\right).$$

A robust estimator after adjustment is given as

$$\begin{aligned}
\widehat{\mathrm{Avar}_3}\left(\hat{\theta}\right) &= \frac{1}{G}\hat{\mathbf{A}}^{-1}\hat{\mathbf{D}}\hat{\mathbf{A}}^{-1} && (2.37)\\
&= \left(\sum_{g=1}^{G}\nabla\hat{\mathbf{m}}_g'\hat{\mathbf{W}}_g^{-1}\nabla\hat{\mathbf{m}}_g\right)^{-1}\\
&\quad \left(\sum_{g=1}^{G}\sum_{h=1}^{G}k(d_{gh})\left(\hat{\mathbf{s}}_g\hat{\mathbf{s}}_h'+\hat{\mathbf{s}}_g\hat{\mathbf{r}}_h'\hat{\mathbf{F}}'+\hat{\mathbf{F}}\hat{\mathbf{r}}_g\hat{\mathbf{s}}_h'+\hat{\mathbf{F}}\hat{\mathbf{r}}_g\hat{\mathbf{r}}_h'\hat{\mathbf{F}}'\right)\right)\\
&\quad \left(\sum_{g=1}^{G}\nabla\hat{\mathbf{m}}_g'\hat{\mathbf{W}}_g^{-1}\nabla\hat{\mathbf{m}}_g\right)^{-1},
\end{aligned}$$

where

$$\hat{\mathbf{A}} = \frac{1}{G}\sum_{g=1}^{G}\nabla\hat{\mathbf{m}}_g'\hat{\mathbf{W}}_g^{-1}\nabla\hat{\mathbf{m}}_g,$$

and

$$\hat{\mathbf{D}}=\frac{1}{G}\sum_{g=1}^{G}\sum_{h=1}^{G}k(d_{gh})\left(\hat{\mathbf{s}}_g\hat{\mathbf{s}}_h'+\hat{\mathbf{s}}_g\hat{\mathbf{r}}_h'\hat{\mathbf{F}}'+\hat{\mathbf{F}}\hat{\mathbf{r}}_g\hat{\mathbf{s}}_h'+\hat{\mathbf{F}}\hat{\mathbf{r}}_g\hat{\mathbf{r}}_h'\hat{\mathbf{F}}'\right). \qquad (2.38)$$

2.6 Two Examples: Spatial Probit Model and Poisson Regression Model

The setup of nonlinear models with spatial data could be tricky. We need to incorporate the spatial correlated term in an appropriate way. In this section, I will use two nonlinear models to demonstrate how we can incorporate the spatial correlated term and use a GEE procedure. This could vary with different models. The first example is a Probit model, and the second one is a count data model.

2.6.1 Example 1. A Probit Model with Spatial Correlation in the Latent Error

The Probit model is one of the popular binary response models. The dependent variable has conditional Bernoulli distribution. The dependent variable y 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 ot the latent variable y^* determines the outcome of y.

Assume the Probit model is

$$y_i = 1\left[y_i^* \geq 0\right], \tag{2.39}$$

$$y_i^* = \mathbf{x}_i \beta + e_i, \tag{2.40}$$

The latent error e has a standard multivariate normal distribution, but the covariances depend on pairwise distances $\mathbf{D}_N$, which is different from the usual multivariate normal distribution.[7]

$$\text{corr}\left(e_i, e_j\right) = f\left(d_{ij,}\rho\right), \tag{2.41}$$

where f $(\cdot)$is a function increases in ρ and decreases in d_{ij}.

We do not observe y_i^*; we only observe y_i. Let $\Phi\left(\cdot\right)$ be the standard normal cumulative density function (CDF), and ϕ be the standard normal probability density function

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

(PDF). Assume that the mean function $m_i(\mathbf{x}_i;\beta) \equiv \mathrm{E}(y_i|\mathbf{x}_i, \mathbf{D}_N) = \Phi(\mathbf{x}_i\beta)$ is correctly specified. Because of the nonlinearity of y_i and non-observability of the latent variable y_i^*, $\mathrm{Cov}\left(y_i, y_j|\mathbf{x}_i, \mathbf{x}_j, \mathbf{D}_N\right)$ is hard to discover without more information on the multivariate distribution. In order to proceed with GEE, we need to specify a working matrix, which is possibly misspecified.

The partial QMLE delivers a consistent first-step estimator for the mean parameters as in **Proposition 1**. Using the Bernoulli density function, the log likelihood function for each observation is:

$$l_i(\beta) = y_i \log\Phi(\mathbf{x}_i\beta) + (1-y_i)\log\left[1-\Phi(\mathbf{x}_i\beta)\right]. \tag{2.42}$$

The partial QMLE solves:

$$\begin{aligned} \breve{\beta} &= \arg\max_{\theta\in\Theta} L_N(\beta) \qquad (2.43) \\ L_N(\beta) &= \sum_{i=1}^{N} l_i(\beta) = \sum_{i=1}^{N} y_i \log\Phi(\mathbf{x}_i\beta) + \sum_{i=1}^{N}(1-y_i)\log\left[1-\Phi(\mathbf{x}_i\beta)\right]. \end{aligned}$$

The score of the likelihood function for each individual is

$$\mathbf{s}_i \equiv \frac{\phi(\mathbf{x}_i\beta)\,\mathbf{x}_i'\left[y_i - \Phi(\mathbf{x}_i\beta)\right]}{\Phi(\mathbf{x}_i\beta)\left[1-\Phi(\mathbf{x}_i\beta)\right]}. \tag{2.44}$$

The expected Hessian[8] for each observation is

$$\mathrm{E}(\mathbf{H}_i|\mathbf{x}_i, \mathbf{D}_N) = -\frac{\phi^2(\mathbf{x}_i\beta)\,\mathbf{x}_i'\mathbf{x}_i}{\Phi(\mathbf{x}_i\beta)\left\{1-\Phi(\mathbf{x}_i\beta)\right\}}. \tag{2.45}$$

[8] Unexpected Hessian can be derived. Because the score has the special form that contains $y_i - \Phi(\mathbf{x}_i\beta)$, by taking the expectations conditional on $\mathbf{x}_i$ and $\mathbf{D}_N$, we can get a cleaner expression.

Let $\mathbf{A}_N$ be the sum of negative expected Hessians $\mathbf{A}_N = \sum_{i=1}^{N} \frac{\phi^2(\mathbf{x}_i\beta)\mathbf{x}_i'\mathbf{x}_i}{\Phi(\mathbf{x}_i\beta)\{1-\Phi(\mathbf{x}_i\beta)\}}$, and $\mathbf{A}_0 = \mathrm{E}(\mathbf{H}_i)$ on the distribution of $\mathbf{x}$. Let $\breve{u}_i = y_i - \Phi\left(\mathbf{x}_i\breve{\beta}\right), i = 1, 2, ..., N$ be the residuals. At this stage, a robust variance and covariance estimator for $\breve{\beta}$ can be computed as follows:

$$\begin{aligned}\widehat{\mathrm{Var}}\left(\breve{\beta}\right) &= \left(\sum_{i=1}^{N} \frac{\phi^2\left(\mathbf{x}_i\breve{\beta}\right)\mathbf{x}_i'\mathbf{x}_i}{\Phi\left(\mathbf{x}_i\breve{\beta}\right)\left\{1-\Phi\left(\mathbf{x}_i\breve{\beta}\right)\right\}}\right)^{-1} \\ &\quad \left(\sum_{i=1}^{N}\sum_{j\neq i}^{N} k\left(d_{ij}\right) \frac{\phi\left(\mathbf{x}_i\breve{\beta}\right)\phi\left(\mathbf{x}_j\breve{\beta}\right)\mathbf{x}_i'\breve{u}_i\breve{u}_j\mathbf{x}_j}{\Phi\left(\mathbf{x}_i\beta\right)\left[1-\Phi\left(\mathbf{x}_i\beta\right)\right]}\right) \\ &\quad \left(\sum_{i=1}^{N} \frac{\phi^2\left(\mathbf{x}_i\breve{\beta}\right)\mathbf{x}_i'\mathbf{x}_i}{\Phi\left(\mathbf{x}_i\breve{\beta}\right)\left\{1-\Phi\left(\mathbf{x}_i\breve{\beta}\right)\right\}}\right)^{-1},\end{aligned} \tag{2.46}$$

where $k\left(d_{ij}\right)$ is the kernel weight function that depend on pairwise distances. This partial QMLE and its robust variance covariance estimator provides a legitimate way of the estimation of the spatial Probit model.

The next is to find out how how the two-step estimator GEE works. The second step is to construct the weighting matrix using the first-step estimators and residuals. As the data can be divided into groups, the working matrix can be the weight for a specific group. If the group size equals two, the working matrix is a two by two matrix. We can write the working variance covariance matrix as $\hat{\mathbf{W}}_g = \hat{\mathbf{V}}_g{}^{1/2}\hat{\mathbf{R}}_g\left(\hat{\rho}, \mathbf{D}_G\right)\hat{\mathbf{V}}_g^{1/2}$. An estimator for the working variance matrix for each group is

$$\hat{\mathbf{V}}_g = \begin{pmatrix} \check{v}_1 & 0 & 0 & \cdots & \cdots & 0 \\ 0 & \check{v}_2 & 0 & & & 0 \\ 0 & 0 & \ddots & \ddots & & \vdots \\ \vdots & & \ddots & \check{v}_l & \ddots & \vdots \\ \vdots & & & \ddots & \ddots & 0 \\ 0 & 0 & \cdots & \cdots & 0 & \check{v}_L \end{pmatrix}, \tag{2.47}$$

where

$$\check{v}_l = \Phi\left(\mathbf{x}_l\check{\beta}\right)\left[1-\Phi\left(\mathbf{x}_l\check{\beta}\right)\right], \quad l = 1, ..., L. \tag{2.48}$$

Next we will find an estimator for the working correlation matrix for $y_i - \Phi\left(\mathbf{x}_i\beta\right)$. Suppose the structure of the true correlation matrix $\mathbf{R}$ is $\mathbf{R}_{ij} = \mathbf{C}_{ij}\left(d_{ij}, \lambda\right)$,where $\mathbf{C}_{ij}\left(d_{ij}, \lambda\right)$ is a function that increases in λ and decreases in d_{ij}. Note that λ is the spatial correlation parameter for the dependent variables, while ρ is for the latent error. This two parameters are generally different. Let $\hat{r}_i = \check{u}_i/\sqrt{\check{v}_i}$, 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 $\check{u}_j/\sqrt{\check{v}_j}$. Let $\hat{\mathbf{R}} \equiv \mathbf{R}\left(\mathbf{D}_G, \hat{\lambda}\right)$ and $\hat{\mathbf{R}}_g$ stand for the correlation matrix $\mathbf{R}_g\left(\mathbf{D}_g, \hat{\lambda}\right)$ for the gth group. The function $\mathbf{C}_{ij}\left(d_{ij}, \lambda\right)$ is unknown, but we can choose a correlation function to approximate it and use it in the estimation. For example, say $\mathbf{C}\left(d_{ij}, \rho\right) = \frac{\lambda}{d_{ij}}$ or $\exp\left(-\frac{dij}{\lambda}\right)$. By only using the correlations within groups, an estimator of λ is $\hat{\lambda} = \arg\min \sum_{g=1}^{G}\sum_{i=1}^{L}\sum_{j\neq i}^{L}\left[\hat{r}_i\hat{r}_j - C_{ij}\left(d_{ij}, \rho\right)\right]^2$ for $i < j$.

The second step GEE estimator for β is

$$\hat{\beta} = \arg\min_{\beta} \sum_{g=1}^{G} \left(\mathbf{y}_g - \Phi\left(\mathbf{x}_g\beta\right)\right)' \hat{\mathbf{W}}_g^{-1} \left(\mathbf{y}_g - \Phi\left(\mathbf{x}_g\beta\right)\right). \tag{2.49}$$

If one believes the working correlation matrix is correctly specified, the non-robust variance estimator for $\hat{\beta}$ is

$$\widehat{\mathrm{Var}}_1\left(\hat{\beta}\right) = \left(\sum_{g=1}^{G} \hat{\mathbf{D}}_g' \hat{\mathbf{W}}_g^{-1} \hat{\mathbf{D}}_g\right)^{-1} \tag{2.50}$$

where $\hat{\mathbf{D}}_g = \partial\Phi\left(\mathbf{x}_g\hat{\beta}\right)/\partial\hat{\beta} = \phi\left(\mathbf{x}_g\hat{\beta}\right)\mathbf{x}_g'$. $\hat{\beta}$ is consistent even for misspecified spatial correlation structure. The robust variance estimator to misspecification of spatial correlation is:

$$\begin{aligned}\widehat{\mathrm{Var}}_R\left(\hat{\beta}\right) = & \left(\sum_{g=1}^{G} \hat{\mathbf{D}}_g' \hat{\mathbf{W}}_g^{-1} \hat{\mathbf{D}}_g\right)^{-1} \\ & \left(\sum_{g=1}^{G}\sum_{h=1}^{G} k(d_{gh}) \hat{\mathbf{D}}_g' \hat{\mathbf{W}}_g^{-1} \hat{\mathbf{u}}_g \hat{\mathbf{u}}_h' \hat{\mathbf{W}}_h^{-1} \hat{\mathbf{D}}_h\right) \\ & \left(\sum_{g=1}^{G} \hat{\mathbf{D}}_g' \hat{\mathbf{W}}_g^{-1} \hat{\mathbf{D}}_g\right)^{-1}\end{aligned} \tag{2.51}$$

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

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, Iglesias and Wooldridge (2012).

2.6.2 Example 2. A Poisson Model with a Multiplicative Spatial Error

A count variable is a variable that takes on nonnegative integer values. Many variables that we would like to explain in terms of covariates come as counts, such as the number of times someone is arrested during a given year, and the number of patents applied for by a firm during a 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). A count data is usually characterized by a density in LEF and a population mean. Now let's use a specific example to demonstrate a count data model with a spatial error term. The conditional Poisson density is specified as

$$f\left(y|\mathbf{x},\mathbf{D}\right)=\exp\left[-\mu\right]\mu^{y}/y!, \tag{2.52}$$

where $y! = 1\cdot 2\cdot \ldots \cdot (y-1)\cdot y$ and $0! = 1$. μ is the conditional mean of y. For a given sample, specify the conditional mean as the exponential form:

$$\mathrm{E}\left(y_i|\mathbf{x}_i,\mathbf{D}_N\right)=\exp\left(\mathbf{x}_i\beta\right). \tag{2.53}$$

Assume that the conditional mean function is correctly specified and model the spatial correlation in the conditional mean function:

$$\mathrm{E}\left(y_i|\mathbf{x}_i,v_i,\mathbf{D}_N\right)=v_i\exp\left(\mathbf{x}_i\beta_0\right), \tag{2.54}$$

where v_i is the multiplicative spatial error term. Let $\mathbf{v}$ equal $\left(v_{1,}v_2, ..., v_N\right)'$. The count data model can be written in a conditional mean form:

$$y_i = v_i \exp\left(\mathbf{x}_i\beta_0\right) + \delta_i, \tag{2.55}$$

and

$$\mathrm{E}\left(\delta_i|\mathbf{x}_i, v_i, \mathbf{D}_N\right) = 0.$$

A count data model with a multiplicative spatial error can be characterized by the following assumptions:

(1) $\{(\mathbf{x}_i, v_i, \delta_i)\}$ is a mixing sequence on the sampling space, with mixing coefficient α or ϕ.

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

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

(4) v_i is independent of $\mathbf{x}_i$, $\mathrm{E}(v_i) = 1$, $\mathrm{Var}\left(v_i\right) = \tau^2$, and $\mathrm{Cov}\left(v_i, v_j\right) = \tau^2 \cdot c\left(d_{ij}\right)$, where $c\left(d_{ij}\right)$ is the spatial correlation depending on the distance between observation i and j.

Note that, we only specify the conditional mean, instead of the distribution. Even if the data do not follow the Poisson distribution, the quasi-MLE approach will give you a consistent estimator if you use the Poisson density function. Moreover, y even need not to be a count number. Under the above assumptions we can integrate out v_i by using the law of iterated expectations.

$$\begin{aligned}
\mathrm{E}\left(y_i|\mathbf{x}_i,\mathbf{D}_N\right) &= \mathrm{E}\left(\mathrm{E}\left(y_i|\mathbf{x}_i,v_i,\mathbf{D}_N\right)|\mathbf{x}_i,\mathbf{D}_N\right) \qquad (2.56)\\
&= \mathrm{E}\left(v_i\exp\left(\mathbf{x}_i\beta_0\right)|\mathbf{x}_i,\mathbf{D}_N\right)\\
&= \exp\left(\mathbf{x}_i\beta_0\right)\mathrm{E}\left(v_i|\mathbf{x}_i,\mathbf{D}_N\right)\\
&= \exp\left(\mathbf{x}_i\beta_0\right)\mathrm{E}\left(v_i\right)\\
&= \exp\left(\mathbf{x}_i\beta_0\right)
\end{aligned}$$

And we can calculate the variances and covariances of y conditional on $\mathbf{x}$:

$$\begin{aligned}
\mathrm{Var}\left(y_i|\mathbf{x}_i,\mathbf{D}_N\right) &= \mathrm{E}\left[\mathrm{Var}\left(\left(y_i|\mathbf{x}_i,v_i,\mathbf{D}_N\right)|\mathbf{x}_i,\mathbf{D}_N\right)\right] \qquad (2.57)\\
&\quad +\mathrm{Var}\left[\mathrm{E}\left(\left(y_i|\mathbf{x}_i,v_i,\mathbf{D}_N\right)|\mathbf{x}_i,\mathbf{D}_N\right)\right]\\
&= \mathrm{E}\left[v_i\exp\left(\mathbf{x}_i\beta_0\right)|\mathbf{x}_i,\mathbf{D}_N\right]+\mathrm{Var}\left[v_i\exp\left(\mathbf{x}_i\beta_0\right)|\mathbf{x}_i,\mathbf{D}_N\right]\\
&= \exp\left(\mathbf{x}_i\beta_0\right)\mathrm{E}\left(v_i|\mathbf{x}_i,\mathbf{D}_N\right)+\exp\left(2\mathbf{x}_i\beta_0\right)\mathrm{Var}\left(v_i|\mathbf{x}_i,\mathbf{D}_N\right)\\
&= \exp\left(\mathbf{x}_i\beta_0\right)+\exp\left(2\mathbf{x}_i\beta_0\right)\cdot\tau^2
\end{aligned}$$

$$\begin{aligned}
\mathrm{Cov}\left(y_i, y_j|\mathbf{x}_i, \mathbf{x}_j, \mathbf{D}_N\right) &= \mathrm{E}\left[\mathrm{Cov}\left(\left(y_i, y_j|\mathbf{x}_i, \mathbf{x}_j, v_i, v_j, \mathbf{D}_N\right)|\mathbf{x}_i, \mathbf{x}_j, \mathbf{D}_N\right)\right] \qquad (2.58)\\
&\quad +\mathrm{Cov}\{\left[\mathrm{E}\left(y_i|\mathbf{x}_i, v_i, \mathbf{D}_N\right), E\left(y_j|\mathbf{x}_i, v_i, \mathbf{D}_N\right)\right]|\mathbf{x}_i, \mathbf{x}_j, \mathbf{D}_N\}\\
&= 0 + \mathrm{Cov}\left(v_i \exp\left(\mathbf{x}_i\beta_0\right), v_j \exp\left(\mathbf{x}_j\beta_0\right)|\mathbf{x}_i, \mathbf{x}_j, \mathbf{D}_N\right)\\
&= \exp\left(\mathbf{x}_i\beta_0\right)\exp\left(\mathbf{x}_j\beta_0\right)\mathrm{Cov}\left[v_i, v_j|\mathbf{x}_i, \mathbf{x}_j, \mathbf{D}_N\right]\\
&= \exp\left(\mathbf{x}_i\beta_0\right)\exp\left(\mathbf{x}_j\beta_0\right)\mathrm{Cov}\left(v_i, v_j\right)\\
&= \exp\left(\mathbf{x}_i\beta_0\right)\exp\left(\mathbf{x}_j\beta_0\right)\cdot\tau^2\cdot c\left(d_{ij}\right)
\end{aligned}$$

The model can also be expressed as

$$\begin{aligned}
\mathrm{E}\left(y_i|\mathbf{x}_i, e_i, \mathbf{D}_N\right) &= \exp\left(\mathbf{x}_i\beta + e_i\right) \qquad (2.59)\\
&= \exp\left(e_i\right)\exp\left(\mathbf{x}_i\beta_0\right)
\end{aligned}$$

We can see that $v_i = \exp(e_i)$. We can model the spatial correlation between e_i and e_j, for $i, j = 1, 2, ..., N$. Let vector $\mathbf{e}$ denote $\left(e_{1,}e_2, ..., e_N\right)'$. In this paper, only positive correlations are considered for convenience. Negative correlations[9] are possible and one can extend this method to it. For convenience, I assume $\mathbf{e}$ follows a multivariate normal distribution, $\mathrm{N}(\mu, \mathbf{\Omega})$. Also I use this in the simulaion later. Then $\mathbf{v}$ will follow a multivariate lognormal distribution. One could use other multivariate distributions for $\mathbf{e}$, and $\mathbf{v}$ will follow the corresponding multivariate distribution. Or one could use a multivariate distribution for v_i directly as long as the first two moments assumptions for v_i are satisfied.

[9] Bloom, schankerman and Reenen's working paper (2012) in NBER identifies negative product market rivalry effect.

Although $\mathbf{y}$ follow a multivariate Poisson distribution, the above equation shows that β_0 can be consistently estimated by the partial Poisson maximum likelihood estimation. The density function for y_i is

$$f_i\left(y_i|\mathbf{x}_i, \mathbf{D}_N\right) = \exp\left[-\exp\left(\mathbf{x}_i\beta\right)\right]\left[\exp\left(\mathbf{x}_i\beta\right)\right]^{y_i}/y_i!, \tag{2.60}$$

where $y_i! = 1 \cdot 2 \cdot ... \cdot (y_i - 1) \cdot y_i$ and $0! = 1$.

The log likelihood for each observation is

$$l_i\left(\beta\right) = -\exp\left(\mathbf{x}_i\beta\right) + y_i\mathbf{x}_i\beta - \log\left(y_i!\right), \tag{2.61}$$

and the score is

$$\mathbf{s}_i\left(\beta\right) \equiv \frac{\partial L_i\left(\beta\right)}{\partial\beta} = \mathbf{x}_i'\left[y_i - \exp\left(\mathbf{x}_i\beta\right)\right] = \mathbf{x}_i'u_i, \tag{2.62}$$

and the Hessian is

$$\mathbf{H}_i\left(\beta\right) = -\exp\left(\mathbf{x}_i\beta\right)\mathbf{x}_i'\mathbf{x}_i. \tag{2.63}$$

Let $\mathbf{A}_i \equiv -\mathrm{E}\left(\mathbf{H}_i\left(\beta\right)|\mathbf{x}_i, \mathbf{D}_N\right) = \exp\left(\mathbf{x}_i\beta\right)\mathbf{x}_i'\mathbf{x}_i$, and $\mathbf{A}_0 \equiv \mathrm{E}\left[\mathbf{A}_i\right]$ over the distribution of $\mathbf{x}$ and the spatial space $\mathbf{D}$. Let $\mathbf{B}_0 \equiv \mathrm{E}\left(\mathbf{s}_i\left(\beta_0\right)\mathbf{s}_i\left(\beta_0\right)'\right) = \mathrm{E}\left(u_i^2\mathbf{x}_i'\mathbf{x}_i\right)$.

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

$$\check{\beta} = \arg\max_{\theta\in\Theta} L\left(\beta\right) = \sum_{i=1}^{N} l_i\left(\beta\right) = \sum_{i=1}^{N} y_i\mathbf{x}_i\beta - \sum_{i=1}^{N}\exp\left(\mathbf{x}_i\beta\right) - \sum_{i=1}^{N}\log\left(y_i!\right). \tag{2.64}$$

A robust partial QMLE variance covariance estimator is

$$\left[\sum_{i=1}^{N}\exp\left(\mathbf{x}_i\breve{\beta}\right)\mathbf{x}_i'\mathbf{x}_i\right]^{-1}\sum_{i=1}^{N}\sum_{j=1}^{N}k\left(d_{ij}\right)\mathbf{x}_i'\hat{u}_i\hat{u}_j\mathbf{x}_j\left[\sum_{i=1}^{N}\exp\left(\mathbf{x}_i\breve{\beta}\right)\mathbf{x}_i'\mathbf{x}_i\right]^{-1}, \qquad (2.65)$$

where $k\left(i,j\right)$ is a kernel function depending on the distance between observations i and j.

The partial QMLE does not make any use of the pairwise correlations. If we can use them, we may improve efficiency. The GEE approach is:

$$\hat{\beta}=\arg\min_{\beta}\sum_{g=1}^{G}\left[\mathbf{y}_g-\exp\left(\mathbf{X}_g\beta\right)\right]'\hat{\mathbf{W}}_g^{-1}\left[\mathbf{y}_g-\exp\left(\mathbf{X}_g\beta\right)\right]. \qquad (2.66)$$

where $\hat{\mathbf{W}}_g$ is an estimated weighting matrix. Again, like the Probit model, $\hat{\mathbf{W}}_g=\hat{\mathbf{V}}_g^{1/2}\hat{\mathbf{R}}_g\hat{\mathbf{V}}_g^{1/2}$, where $\hat{\mathbf{V}}$ is a diagonal matrix with the estimated variances on the diagonal, and $\hat{\mathbf{R}}$ is the estimated working correlation matrix. The most efficient weighting matrix is the true covariance matrix of $\mathbf{y}-\exp\left(\mathbf{X}\beta\right)$.

There are two pivotal parameters that we do not know in the estimation, τ^2 and ρ. We could use the partial Poisson QMLE residuals to estimate the parameters. Let $\breve{\beta}$ be the partial Poisson regression estimator. Let $\breve{u}_i^2=\left[y_i-\exp\left(\mathbf{x}_i\breve{\beta}\right)\right]^2$ be the squared residual. τ^2 can be estimated in the following way: $\hat{\tau}^2$ equals the coefficient by regressing $\breve{u}_i^2-\exp\left(\mathbf{x}_i\breve{\beta}\right)$ on $\exp\left(2\mathbf{x}_i\breve{\beta}\right)$. The situation to estimate ρ is a little bit complicated. First, we do not know how σ_{ij} depends on ρ and d_{ij}. If we use the wrong structure, we probably will get a wrong estimator for ρ. Suppose the correlation structure is $c\left(d_{ij}\right)=\frac{\rho}{d_{ij}}$, then an estimator for ρ is: $\hat{\rho}=$ coefficient by regressing $\frac{\breve{u}_i\breve{u}_j}{\exp\left(\mathbf{x}_i\breve{\beta}\right)\exp\left(\mathbf{x}_j\breve{\beta}\right)}$ on $\frac{\hat{\tau}^2}{d_{ij}}$. However, this estimator sometimes suffers from the negative products of the Poisson regression residuals, and the estimated parameter $\hat{\rho}$ is biased downward. $\hat{\mathbf{W}}_g$ is obtained by plugging $\hat{\tau}^2$ and $\hat{\rho}$ back in the variance covariance matrix. By minimizing the above equation, we can get the GEE estimator.

If **Assumption 13** holds, a non-robust variance estimator for GEE is

$$\widehat{\mathrm{Var}}\left(\hat{\beta}\right)=\left(\sum_{g=1}^{G}\hat{\mathbf{D}}_g'\hat{\mathbf{W}}_g^{-1}\hat{\mathbf{D}}_g\right)^{-1}, \tag{2.67}$$

where $\hat{\mathbf{D}}_g=\partial\exp\left(\mathbf{X}_g\beta\right)/\partial\beta=\exp\left(\mathbf{X}_g\hat{\beta}\right)\mathbf{X}_g'$.

$\hat{\beta}$ is still consistent even for a misspecified spatial correlation structure when **Assumption 13** does not hold. The robust variance estimator to misspecification of spatial correlation is:

$$\begin{aligned}\widehat{\mathrm{Var}}\left(\hat{\beta}\right) &= \left(\sum_{g=1}^{G}\hat{\mathbf{D}}_g'\hat{\mathbf{W}}_g^{-1}\hat{\mathbf{D}}_g\right)^{-1} \\ &\quad \left(\sum_{g=1}^{G}\sum_{h=1}^{G}k(d_{gh})\hat{\mathbf{D}}_g'\hat{\mathbf{W}}_g^{-1}\hat{\mathbf{u}}_g\hat{\mathbf{u}}_h'\hat{\mathbf{W}}_h^{-1}\hat{\mathbf{D}}_h\right) \\ &\quad \left(\sum_{g=1}^{G}\hat{\mathbf{D}}_g'\hat{\mathbf{W}}_g^{-1}\hat{\mathbf{D}}_g\right)^{-1}\end{aligned} \tag{2.68}$$

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

2.7 Monte Carlo Simulations

In this section, I want to do the Monte Carlo simulation to investigate the efficiency gain of the GEE approach compared to the pooled QMLE. The simulation mechanism is described as follows:

(1) Each individual resides on a intersection of the square lattice. Thus the pairwise

distance can be calculated using the coordinates of each observation.

(2) According to pairwise distances, generate the pairwise correlations. In the simulation, there are two cases of pairwise correlations. One is that each observation is correlated with all other observations; the other is that only observations in the same group are correlated with each other. That is, assuming groupwise independence. The group are demonstrated in the graph in the next subsection.

(3) After generated correlated data, the first step is to find the pooled QMLE estimator. The second step is divide observations into groups and only use within group information to estimate the correlation parameters. In the case of groupwise dependence, although each pair of the data are correlated, we do not use pairwise correlations between observations in different groups. In the case of groupwise dependence, this method sounds natural. After estimation of the spatial correlation parameters, estimate the mean parameters again using the GEE procedure.

2.7.1 Sampling Space

Graph 1 demonstrates the case when the sample size is 400. Thus, I create a 20 × 20 square lattice. Each observation resides on the intersections of the lattice. The locations for the data are $\{(r, s) : r, s = 1, 2, ...20\}$. The distance d_{ij} between location i and j is 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 $\sqrt{(a_i - b_i)^2 + (a_j - b_j)^2}$. Then the spatial correlation is based on a given parameter ρ and d_{ij}. The assumed correlation for the spatial correlated error term is ρ/d_{ij} for $i, j = 1, 2, ..., N$. For the Probit model this correlation is between the latent error, and for the count data model this correlation is between the multiplicative error. The data are divided into groups of 4, 16. And you can use more observations in one group, say, 25 and

100. But this increase the computation burden and 25 and 100 are too big for the sample size 400. In Graph 1, the left upper corner of the lattice demonstrates the case that the group size is four; the left lower corner of the lattice demonstrates the case that the group size is 16; and the right lower part of the lattice demonstrates the case that the group size is 100. The idea of this two-step method is to use a small number of pairwise correlations and conveniently get more efficient estimators. Thus, I use 4 or 16 as the number of observations in each group. Notice, for convenience, I make the pairwise distances in different groups the same.

2.7.2 Spatial Probit Data

For the Probit model, we can not easily find the variances and covariances for the dependent variables conditional on the covariates. The corrlated latent errors result in correlated binary response varibles. However, the correlations in latent error usually do not reflect the exact correlations in the binary dependent variables. The correlations are much smaller in the binary response variable because of the nonlinear tranformation. In the following simulation, let the sample size be 400 and replication be 500. Consider the following data generating process:

1. $x_i = [1, x_{i1}], \qquad x_{i1} \sim \mathrm{N}\left(1, 1\right); \qquad \beta = [-1, 1]'$.

2. $y_i^* = x_i\beta + e_i, \qquad e \sim \mathrm{MVN}(0, \Omega), \qquad \mathrm{Var}\left(e_i\right) = 1, \qquad \mathrm{Cov}\left(e_i, e_j\right) = \frac{\rho}{d_{ij}}; \qquad \rho = 0, 0.1, 0.2, 0.3, 0.4$.

3. $y_i = 1$ if $y_i^* \geq 0, \qquad y_i = 0$ if $y_i^* < 0$.

4. Use $\mathrm{Cov}\left(y_i, y_j | \mathbf{x}_i, \mathbf{x}_j, \mathbf{D}_N\right) = \frac{\lambda}{d_{ij}}$ as the correlation between the dependent variables. This form is arbitrary. Based on the information, this is not likely to be true and λ is unknown.

The simulation results are in Table 1-2. $\hat{\lambda}$ is the estimator for λ. $\hat{\lambda}$ is calculated by the minimum distance estimator

$$\min_{\lambda} \sum_{g=1}^{G} sum_{i=1}^{L} \sum_{j \neq i}^{L} \left[\frac{\check{u}_i \check{u}_j}{\sqrt{\Phi\left(\mathbf{x}_i \check{\beta}\right)\left[1-\Phi\left(\mathbf{x}_i \check{\beta}\right)\right]}\sqrt{\Phi\left(\mathbf{x}_j \check{\beta}\right)\left[1-\Phi\left(\mathbf{x}_j \check{\beta}\right)\right]}} - \frac{\lambda}{d_{ij}} \right],$$

for i, j in the same group g, and L is the number of individuals in a group. Because λ is unknown, we can not calculate the bias of $\hat{\lambda}$. $\check{\beta}$ is the first-step partial QMLE estimator for β. $\hat{\beta}_1$ is the two-step weighted nonlinear least squred estimator (WNLS) that only uses the variances as weight in the second step. $\hat{\beta}_4$ is the two-step GEE estimator that uses a 4×4 variance covariance matrix as weight for each group. $\hat{\beta}_{16}$ is the two-step GEE estimator that uses a 16×16 variance covariance matrix as weight for each group. Note that when $L = 4$, $\hat{\beta}_{16}$ is calculated by using $\hat{\lambda}$ which is calculated using group size equal 4. Similarly for $\hat{\beta}_4$ when $L = 16$. The following two tables show two cases of the simulation: (1) N=400, L=4. (2) N=400, L=16.

Table 2.1: Binary Data N=400, L=4

N=400,	L=4,	T=500,	$\beta = 1$			
ρ		$\hat{\lambda}$	$\breve{\beta}$	$\hat{\beta}_1$	$\hat{\beta}_4$	$\hat{\beta}_{16}$
0	estimate	0.0032	1.0162	1.0162	1.0155	1.0169
	s.d	0.0419	0.1111	0.1011	0.1144	0.1118
0.1	estimate	0.0281	1.0139	1.0139	1.0145	1.1053
	s.d	0.0501	0.1166	0.1166	0.1165	0.1172
0.2	estimate	0.0616	1.0241	1.0241	1.0243	1.0261
	s.d	0.0575	0.1103	0.1103	0.1111	0.1136
0.3	estimate	0.0903	1.0242	1.0233	1.0253	1.0251
	s.d	0.0558	0.1138	0.1147	0.1142	0.1157
0.4	estimate	0.1281	1.0444	1.0434	1.0484	1.0515
	s.d	0.0745	0.1128	0.1148	0.1185	0.1162

Table 2.2: Binary Data N=400, L=16

N=400,	L=16,	T=500,	$\beta = 1$			
ρ		$\hat{\lambda}$	$\breve{\beta}$	$\hat{\beta}_1$	$\hat{\beta}_4$	$\hat{\beta}_{16}$
0	estimate	-0.0033	1.0042	1.0043	1.0037	1.0050
	s.d	0.0267	0.1082	0.1081	0.1103	0.1081
0.1	estimate	0.0238	1.0165	1.0165	1.0164	1.0166
	s.d	0.0338	0.1174	0.1174	0.1170	0.1171
0.2	estimate	0.0504	1.0295	1.0280	1.0300	1.0303
	s.d	0.0407	0.1118	0.1154	0.1120	0.1125
0.3	estimate	0.0822	1.0336	1.0336	1.0346	1.0366
	s.d	0.0539	0.1121	0.1121	0.1135	0.1135
0.4	estimate	0.1100	1.0483	1.0483	1.0494	1.0516
	s.d	0.0601	0.1277	0.1277	0.1283	0.1298

From the above results, $\hat{\lambda}$ is much smaller than ρ, which implies that a high spatial correlation in the latent error term may only cause a tiny correlation in the binary response variables. The PQMLE $\breve{\beta}$ delivers almost the same efficiency as the other three two-step estimators. Since the correlations between the binary dependent variables are low, a two-step estimator may not be necessary. One can use other data generating process to generate highly correlated binary dependent variables to examine the two-step method.

2.7.3 Spatial Count Data

In the count data case, the variances and covariances of the count dependent variable can be written in closed forms. The spatial correlation in the underlying spatial error term is then transformed to spatial correlation in the response variable term. That is, by knowing the correlations in the spatial error term, we know the correlations in the count dependent variable. This avoids the situation that we can not compare the estimated spatial correlation parameter and the true parameter.

Remember the assumption of conditional mean function for the count data is $\mathrm{E}\left(y_i|\mathbf{x}_i, v_i,\ \mathbf{D}_N\right) = \exp\left(\mathbf{x}_i\beta + e_i\right) = v_i\exp\left(\mathbf{x}_i\beta_0\right).$ Consider the following case of spatial count data generating process:

1. $e_i, i = 1, 2, ..., N$ follows a multivariate normal distribution with $\mathrm{E}(e_i) = -\frac{1}{2}$ and $\mathrm{Var}\left(e_i\right) = \sigma^2 = 1$; $\mathrm{Cov}\big(e_i, e_j\big) = \frac{\rho}{d_{ij}}$, and $\rho = 0, 0.1, 0.2, 0.3, 0.4, 0.5, 0.6$. Therefore, v_i follows a multivariate lognormal distribution with $\mathrm{E}(v_i) = 1$, $\mathrm{Var}\left(v_i\right) = e - 1$ and $\mathrm{Cov}\big(v_i, v_j\big) = \exp\left(\frac{\rho}{d_{ij}}\right) - 1$.

2. $x_i = [1, x_{i1}], \qquad x_{i1} \sim \text{Uniform}\,(0, 1)$

3. $\beta = [1, -1]'$

4. $m_i = \exp\left(\mathbf{x}_i\beta{+}e_i\right)$

5. $y_i \sim Poisson\,(m_i)$

The parameter ρ represents the correlation in the underlying error term. The true correlations for y_i and y_j conditional on $\mathbf{x}_i, \mathbf{x}_j, d_{ij}$ is

$$\begin{aligned} & \text{Corr}\left(y_i, y_j | \mathbf{x}_i, \mathbf{x}_j, d_{ij}\right) \qquad (2.69) \\ = & \frac{\exp\left(\mathbf{x}_i \beta_0\right) \exp\left(\mathbf{x}_j \beta_0\right) \left[\exp\left(\rho / d_{ij}\right) - 1\right]}{\sqrt{\exp\left(\mathbf{x}_i \beta_0\right) + \exp\left(2\mathbf{x}_i \beta_0\right) \cdot \tau^2} \sqrt{\exp\left(\mathbf{x}_j \beta_0\right) + \exp\left(2\mathbf{x}_j \beta_0\right) \cdot \tau^2}}. \end{aligned}$$

We can calculate the sample correlations in the simulated count data based on the above expression. For each replication t, let $u_i = y_i - \exp\left(\mathbf{x}_i \beta_0\right)$, and the sample correlation of $\mathbf{y}$ for each t is:

$$\begin{aligned} & \widehat{corr}_t\left(y_i, y_j | \mathbf{x}_i, \mathbf{x}_j, d_{ij}\right) \\ = & \frac{1}{N_{d^*}} \sum_i^N \sum_{j>i}^N \frac{u_i u_j}{\sqrt{\exp\left(\mathbf{x}_i \beta_0\right) + \exp\left(2\mathbf{x}_i \beta_0\right) \cdot \tau^2} \sqrt{\exp\left(\mathbf{x}_j \beta_0\right) + \exp\left(2\mathbf{x}_j \beta_0\right) \cdot \tau^2}}, \end{aligned}$$

where N_{d^*} is the number of distinct pairs of observations whose distance is equal to a certain distance $d_{ij} = d^*$, and $\text{Var}\left(\mathbf{y}\right)$ is the sample variance of count variable at time t. Table 2.3 shows the sample correlations in y over the replications. The correlations are calculated separately for pairs of $y_i^\prime s$ that are 1, 2, 5 apart from each other. As pairwise distances increase, the correlations decrease. There is almost no correlation between two observations when distance is 5. Moreover, the spatial parameter ρ can be as high as 0.6, and the sample correlation of y can be as high as more than 0.3.

Table 2.3: Correlations in Simulated Count Data

T=2000	N=400			N=1600		
d_{ij}	1	2	5	1	2	5
N_{ij}	760	720	1688	3170	3040	8028
ρ	$\rho_{y,d=1}$	$\rho_{y,d=2}$	$\rho_{y,d=5}$	$\rho_{y,d=1}$	$\rho_{y,d=2}$	$\rho_{y,d=5}$
0	-0.001	-0.001	-0.000	0.000	0.000	-0.000
0.1	0.043	0.022	0.009	0.043	0.021	0.008
0.2	0.089	0.043	0.017	0.089	0.043	0.015
0.3	0.139	0.068	0.026	0.139	0.064	0.023
0.4	0.194	0.091	0.036	0.194	0.088	0.032
0.5	0.250	0.115	0.046	0.253	0.112	0.040
0.6	0.312	0.139	0.056	0.315	0.136	0.048

As long as we can find consistent estimators for τ^2 and ρ, we can do the second step estimation. As in 2.57 and 2.58, the conditional variances and covariances of count dependent variable can be written in a closed form according to which the spatial parameters can be estimated. Using the information above, τ^2 can be estimated by $\hat{\tau}^2$ as the coefficient by regressing $\breve{u}_i^2 - \exp\left(\mathbf{x}_i\breve{\beta}\right)$ on $\exp\left(2\mathbf{x}_i\breve{\beta}\right)$. Obviously $\hat{\tau}^2$ does not depend on distances. For simplicity ρ is estimated by$\frac{1}{N_0}\sum_{i=1}^{N}\sum_{j\neq i}^{N}\log[\frac{\breve{u}_i\breve{u}_j}{\exp(\mathbf{x}_i\breve{\beta})\exp\left(\mathbf{x}_j\breve{\beta}\right)}+1]$ for those pairs whose pairwise distance is one, where N_0 is the number of pairs whose distance is one.

Table 2.4 shows the simulation results of the case in which: N=1600, L=4 and L=16. Replication is 2000. $\breve{\beta}$ is the one-step QMLE estimator; $\hat{\beta}_1$, $\hat{\beta}_4$ and $\hat{\beta}_{16}$ are three two-step estimators which uses different variance and covariance matrices of the count variable . $\hat{\beta}_1$ only uses the estimated variances as weights; $\hat{\beta}_4$ uses the covariance matrix with group size equal to four, and $\hat{\beta}_{16}$ uses the covariance matrix with group size equal to sixteen. We can see that all the three two-step estimators $\hat{\beta}_1$, $\hat{\beta}_4$ and $\hat{\beta}_{16}$ are more efficient than the one-step estimator $\breve{\beta}$. The two GEE estimators $\hat{\beta}_4$ and $\hat{\beta}_{16}$ has more improvement when ρ grows larger. When sample size increases, efficiency gains. $\hat{\rho}$ and $\hat{\tau}^2$ are both slightly biased downwards.

Table 2.4: Count Data N=1600, L=4 and 16

N=1600			$\sigma^2 = 1, \quad \beta = -1$			L=4	L=16
ρ		$\hat{\rho}$	$\hat{\sigma}^2$	$\check{\beta}$	$\hat{\beta}_1$	$\hat{\beta}_4$	$\hat{\beta}_{16}$
0	estimate	−0.002	0.984	-1.000	-0.998	-0.998	-0.998
	bias	−0.002	0.016	0.000	0.002	0.002	0.002
	s.d.	(0.043)	(0.156)	(0.139)	(0.135)	(0.135)	(0.135)
0.1	estimate	0.095	0.980	-1.000	-0.998	-0.998	-0.998
	bias	−0.005	−0.020	0.000	0.002	0.002	0.002
	s.d.	(0.048)	(0.158)	(0.137)	(0.134)	(0.134)	(0.134)
0.2	estimate	0.191	0.970	−1.001	-0.998	-0.998	-0.999
	bias	-0.008	-0.030	-0.001	0.002	0.002	0.001
	s.d.	(0.055)	(0.160)	(0.131)	(0.128)	(0.126)	(0.126)
0.3	estimate	0.287	0.969	-0.998	-0.996	-0.995	-0.997
	bias	−0.013	-0.031	0.002	0.004	0.005	0.003
	s.d.	0.063	(0.161)	(0.138)	(0.135)	(0.132)	(0.131)
0.4	estimate	0.384	0.967	−1.000	-0.998	-0.998	-0.999
	bias	−0.016	-0.033	0.000	0.002	0.002	0.001
	s.d.	(0.070)	(0.172)	(0.136)	(0.134)	(0.128)	(0.126)
0.5	estimate	0.480	0.974	-0.998	-0.997	-0.996	-0.998
	bias	−0.020	−0.026	0.002	0.003	0.004	0.002
	s.d.	(0.076)	(0.196)	(0.134)	(0.130)	(0.120)	(0.118)

2.8 Conclusions

The spatial correlation in nonlinear models makes it more difficult to get efficient estimators. For a Probit model, because of the nonlinearity, it is hard to find the real form of spatial correlations between two dependent variables. Thus the estimated variance covariance matrix is likely to be misspecified. In the spatial count data model case, since we can actually write down the variances and covariances of the dependent variables based on certain assumptions, we can use a multivariate nonlinear weighted least squares (MNWLS) to improve the efficiency. Accounting for spatial correlation will improve efficiency in the count data example. The further study will focus on how to get a better estimator for the spatial correlation parameter, how to get a good approximation of the correlation structure and how to incorporate spatial correlation in other nonlinear models, such as multi-catogary binary choice, fractional response, two-part model and so on.

Chapter 3

Conditions for the Numerical Equality of the OLS, GLS and Amemiya-Cragg Estimators

3.1 Introduction

Conditions under which the ordinary least squares (OLS) and generalized least squares (GLS) estimators are equal are well known. This paper extends these results in two ways. First, we give conditions under which GLS based on one assumed error variance matrix equals GLS based on a different assumed variance matrix. Second, we give conditions under which GLS equals the GMM estimator of Amemiya (1983) and Cragg (1983).

3.2 Equivalence of OLS and GLS

Consider the linear regression model

$$y = X\beta + \varepsilon \tag{3.1}$$

where y is $T \times 1$ and X is $T \times K$. Let Σ be a $T \times T$, positive definite, assumed or estimated variance matrix of ε. We consider the ordinary least squares (OLS) estimator $\hat{\beta}$ and the generalized least squares (GLS) estimator $\tilde{\beta}$ defined as follows:

$$\hat{\beta} = (X'X)^{-1}X'y, \tilde{\beta} = (X'\Sigma^{-1}X)^{-1}X'\Sigma^{-1}y. \tag{3.2}$$

The question is under what conditions on X and Σ it is the case that $\hat{\beta} = \tilde{\beta}$. Since these will be conditions for numerical equality of the estimators, they will not depend on assumptions about ε, and in particular they will not depend on whether Σ does or does not actually equal the variance matrix of ε. We assume only that Σ, $X'X$ and $X'\Sigma^{-1}X$ are positive definite.

This is an old and classic problem. The basic equivalence results were first given by Zyskind (1962, 1967), Rao (1967) and Kruskal (1968). These results were summarized in textbook fashion by Amemiya (1985). Since then the number of equivalent conditions for equality of the two estimators has grown. The survey by Puntanen and Styan (1989) lists 20 such conditions. See also Baltagi (1989) and McAleer (1992) for more discussion and applications of these results.

A related but different questions is under what conditions on Σ we have OLS = GLS for all X. McElroy (1967) showed that, if X contains an intercept, a necessary and sufficient condition is that $\sum$ have the "equicorrelated" form (all diagonal elements equal, and all off-diagonal elements equal). Balestra (1970) extended this result to the case that a subset of the regressors must have a certain form and then OLS = GLS for all possible values of the remaining regressors. We do not seek to extend these results in this paper.

Because of the numerous different equivalent sets of conditions for the equality of OLS and GLS, we will focus on the conditions given in Theorem 6.1.1 of Amemiya (1985, p. 182).

The following Theorem is a minor extension of that result.

THEOREM 1. Suppose that Σ, $X'X$ and $X'\Sigma^{-1}X$ are positive definite. Then the following statements are equivalent.

(A) $(X'X)^{-1}X'\Sigma X(X'X)^{-1} = \left(X'\Sigma^{-1}X\right)^{-1}$

(B) $\Sigma X = XB$ for some nonsingular B

(C) $(X'X)^{-1}X' = \left(X'\Sigma^{-1}X\right)^{-1} X'\Sigma^{-1}$

(D) $X = HA$ for some nonsingular A, where the columns of H are K eigenvectors of Σ

(E) $Z'\Sigma X = 0$ for any Z such that $Z'X = 0$

(F) $\Sigma = X\Gamma X' + Q\Theta Q' + c^2 I$ for some Γ, some Θ, some c, and some Q such that $Q'X = 0$

(F') $\Sigma = X\Psi X' + R\Phi R'$ for some Ψ, some Φ, and some R such that $R'X = 0$

Conditions (A)-(F) are, apart from a few changes in notation, as given by Amemiya, p. 182.. Condition (F') is new. Amemiya shows the equivalence of (A)-(E) and refers to Rao (though not to the correct paper by Rao) for a proof of the equivalence of (F). It is trivial that (F) implies (E) but the proof in Rao (1967, p. 364) that (E) implies (F) is not transparent (or even complete; it just says that "it is easy to verify ..."). In Chapter 4 we show that (F') is equivalent to (F) and provide a straightforward proof that (D) implies (F').

3.3 Equivalence of Two Different GLS Estimators

We now consider two different GLS estimators:

$$\tilde{\beta} = (X'\Sigma^{-1}X)^{-1}X'\Sigma^{-1}y, \ddot{\beta} = (X'\Omega^{-1}X)^{-1}X'\Omega^{-1}y. \tag{3.3}$$

We will provide conditions (on X, Σ and Ω) such that these two estimators are equal. (Obviously the results of the previous section are a special case corresponding to $\Omega = I$.) The context that we have in mind is that Σ is the correct variance matrix of ε and Ω is an incorrect or approximate variance matrix. In this case $\ddot{\beta}$ could be considered to be a pseudo-GLS estimate. However, our conditions are just conditions for equality of the two estimators and so which (if either) is based on the correct variance matrix of ε is irrelevant. Note also that any of the conditions given below must still hold if we reverse the roles of Σ and Ω, which is why there are two versions of results (A), (B), etc.

The phrase "pseudo-GLS" has been used in the literature with a different meaning, namely, GLS using the original error variance matrix but after some of the regressors have been partialled out. See, e.g., Fiebig, Bartels and Krämer (1996) or Gross and Puntanen (2000). This section does not apply to that topic, because we assume that both Σ and Ω are nonsingular. Matthew (1983) gives conditions for the two different GLS estimators to be equal for all X in a certain class. This is the same sort of question that was addressed for equality of OLS and GLS by McElroy (1967) and Balestra (1970). Our Theorem below is different because it applies for a given X.

THEOREM 2. Suppose that $\Sigma, \Omega, X\prime X, X'\Sigma^{-1}X$ and $X'\Omega^{-1}X$ are positive definite. Then the following statements are equivalent.

(A1) $\left(X'\Sigma^{-1}X\right)^{-1} X'\Sigma^{-1}\Omega\Sigma^{-1}X \left(X'\Sigma^{-1}X\right)^{-1} = \left(X'\Omega^{-1}X\right)^{-1}$

(A2) $\left(X'\Omega^{-1}X\right)^{-1}X'\Omega^{-1}\Sigma\Omega^{-1}X\left(X'\Omega^{-1}X\right)^{-1}=\left(X'\Sigma^{-1}X\right)^{-1}$

(B1) $\Omega\Sigma^{-1}X=XB$ for some nonsingular B

(B2) $\Sigma\Omega^{-1}X=XB$ for some nonsingular B

(C) $\left(X'\Sigma^{-1}X\right)^{-1}X'\Sigma^{-1}=\left(X'\Omega^{-1}X\right)^{-1}X'\Omega^{-1}$

(D1) $\Sigma^{-1/2}X=HA$ for some nonsingular A, where the columns of H are K eigenvectors of

$$\Sigma^{-1/2}\Omega\Sigma^{-1/2}$$

(D2) $\Omega^{-1/2}X=HA$ for some nonsingular A, where the columns of H are K eigenvectors of

$$\Omega^{-1/2}\Sigma\Omega^{-1/2}$$

(E1) $Z'\Omega\Sigma^{-1}X=0$ for any Z such that $Z'X=0$

(E2) $Z'\Sigma\Omega^{-1}X=0$ for any Z such that $Z'X=0$

(F1) $\Omega=X\Gamma X'+\Sigma Q\Theta Q'\Sigma+c^2\Sigma$ for some Γ, some Θ, some c, and some Q such that

$$Q'X=0$$

(F2) $\Sigma=X\Gamma X'+\Omega Q\Theta Q'\Omega+c^2\Omega$ for some Γ, some Θ, some c, and some Q such that

$$Q'X=0$$

(F1') $\Omega=X\Psi X'+\Sigma R\Phi R'\Sigma$ for some Ψ, some Φ, and some R such that $R'X=0$

(F2') $\Sigma=X\Psi X'+\Omega R\Phi R'\Omega$ for some Ψ, some Φ, and some R such that $R'X=0$

The proof of these results is given in Chapter 4. Our proof is essentially an exercise in translation of Amemiya's conditions. An alternative that we do not pursue would be to use the results of Gourieroux and Monfort (1980).

The results in Theorem 2 have some applications that are similar to existing applications of Theorem 1 in the literature.

Example 1. Random coefficients. Consider the random coefficient model

$$y = X\beta_* + u, \beta_* = \beta + v \text{ so } y = X\beta + \varepsilon \text{ where } \varepsilon = u + Xv. \tag{3.4}$$

Suppose that u and v are uncorrelated and $V(v) = \Gamma$. If $V(u) = \sigma^2 I$, then $V(\varepsilon) \equiv \Sigma = X\Gamma X\prime + \sigma^2 I$. Therefore condition (F) of Theorem 1 applies and GLS = OLS, as has been pointed out by Rao (1967), Amemiya(1985) and Baltagi(1989), among others. However, now suppose that $V(u) = \Omega$. Then $\Sigma = X\Gamma X\prime + \Omega$. Therefore condition (F2) of Theorem 2 holds, and GLS based on Σ equals GLS based on Ω, which so far as we know is a previously unknown result.

Example 2. SUR with the same regressors in every equation. Consider a set of G seemingly unrelated regression equations, with T observations per equation and a common regressor matrix X of dimension $T \times K$. Using standard notation, write the stacked system of equations as

$$y_* = X_*\beta_* + \varepsilon_* \tag{3.5}$$

where y_* is $GT \times 1$, $X_* = (I_G \otimes X)$ is $GT \times KG$, etc. Suppose that $V(\varepsilon_*) = \Sigma \otimes I_T \equiv \Sigma_*$.

It is well known that GLS using Σ_* is the same as OLS, and Baltagi (1989) proves this using condition (B) of Theorem 1: $\Sigma_* X_*$ =$X_* B$ where $B = \Sigma \otimes I_K$. Now suppose that Ω_*=$\Omega \otimes I_T$ is another possible variance matrix. Then GLS based on Σ_* is the same as GLS based on Ω_*. This is probably obvious, since both must equal OLS, but it can also be proved using Theorem 2. Define $\Psi = \Omega\Sigma^{-1}$ and $\Psi_* = \Psi \otimes I_T = \Omega_*\Sigma_*^{-1}$. Then

$$\Omega_*\Sigma_*^{-1}X_* = \Psi_*X_* = \Psi \otimes X = X_*B \tag{3.6}$$

where $B = \Psi \otimes I_K$. So condition (B1) of Theorem 2 holds and the two GLS estimators are the same.

3.4 Equivalence of GLS and the Amemiya-Cragg Estimator

We now consider the two estimators:

$$\tilde{\beta} = (X'\Sigma^{-1}X)^{-1}X'\Sigma^{-1}y \text{ and } \breve{\beta} = [X'H(H\prime\Sigma H)^{-1}H\prime X]^{-1}X\prime H(H\prime\Sigma H)^{-1}H\prime y. \tag{3.7}$$

The first of these is the GLS estimator while the second is the estimator of Amemiya (1983) and Cragg (1983). We will provide conditions (on X,H and Σ) such that these two estimators are equal. If Σ is the correct error variance matrix, then the GLS estimator is efficient relative to the Amemiya-Cragg estimator, and the Amemiya-Cragg estimator is efficient relative to OLS if X is contained in H. However, again, our results are just algrebraic results for the numerical equivalence of the two estimators.

THEOREM 3. Suppose that Σ, $X'\Sigma^{-1}X$ and $H\prime\Sigma H$ are positive definite and that $H\prime X$ has full column rank. Then the following statements are equivalent.

(A) $X\prime H(H\prime\Sigma H)^{-1}H\prime X = X\prime\Sigma^{-1}X$

(B) $\Sigma^{-1}X = HB$ for some B [or, $X = \Sigma HB$ for some B]

(C) $[X\prime H(H\prime\Sigma H)^{-1}H\prime X]^{-1}X\prime H(H\prime\Sigma H)^{-1}H\prime = (X'\Sigma^{-1}X)^{-1}X'\Sigma^{-1}$

(D) $X = SA$ where the columns of S are the eigenvectors that correspond to the non-zero

eigenvalues of $\Sigma H(H\prime\Sigma H)^{-1}H\prime$

(E) $Z\prime\Sigma^{-1}X = 0$ for any Z such that $Z\prime H = 0$ [or, $Z\prime X = 0$ for any Z such that $Z\prime\Sigma H = 0$]

(F') $\Sigma = X\Gamma X\prime + Q\Theta Q\prime$ for some nonsingular Γ, some Θ, some Q and some B (of dimension $L \times K$, where H is $T \times L$) such that $X\prime HB$ is nonsingular and $Q\prime HB = 0$

The proofs of these results are given in Chapter 4.

Chapter 4

Proofs of Propositions and Theorems

4.1 Proofs of Propositions in Chapter 1

This sections provides proofs of propositions in Chapter 1.

Proof of Proposition 1

(1) Consistency.

$$\begin{aligned}
\hat{\beta}_{OLS} &= (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y} \\
&= \beta + (N^{-1}\sum_{i=1}^{N}\mathbf{x}_i'\mathbf{x}_i)^{-1}(N^{-1}\sum_{i=1}^{N}\mathbf{x}_i'u_i) \\
&= \beta + [p\lim(N^{-1}\sum_{i=1}^{N}\mathbf{x}_i'\mathbf{x}_i)^{-1} + op\,(1)][p\lim(N^{-1}\sum_{i=1}^{N}\mathbf{x}_i'u_i) + op\,(1)].
\end{aligned}$$

Under **Assumption 2,** since $\{(\mathbf{x}_i, u_i)\}$ is a mixing sequence, $\mathbf{x}_i'u_i$ and $\mathbf{x}_i'\mathbf{x}_i$ are also mixing. By law of large numbers for mixing sequences (Arbia 2006, page 70), we have the following two equations.

$$N^{-1}\sum_{i=1}^{N}\mathbf{x}_i'u_i \xrightarrow{p} \mathrm{E}\left(N^{-1}\sum_{i=1}^{N}\mathbf{x}_i'u_i\right),$$

$$N^{-1}\sum_{i=1}^{N}\mathbf{x}_i'\mathbf{x}_i \xrightarrow{p} \mathrm{E}\left(N^{-1}\sum_{i=1}^{N}\mathbf{x}_i'\mathbf{x}_i\right).$$

Under **Assumption 3** and **4**,

$$\begin{aligned} \mathrm{E}\left(N^{-1}\sum_{i=1}^{N}\mathbf{x}_i^{'}u_i\right) &= N^{-1}\sum_{i=1}^{N}\mathrm{E}\left(\mathbf{x}_i^{'}u_i\right)=\mathbf{0}, \\ \mathrm{E}\left(N^{-1}\sum_{i=1}^{N}\mathbf{x}_i^{'}\mathbf{x}_i\right) &\rightarrow \mathbf{A}_1. \end{aligned}$$

And

$$p\lim\left(N^{-1}\sum_{i=1}^{N}\mathbf{x}_i^{'}\mathbf{x}_i\right)^{-1}=\mathbf{A}_1^{-1}.$$

Thus

$$\hat{\beta}_{OLS}\overset{p}{\rightarrow}\beta+\mathbf{A}_1^{-1}\cdot\mathbf{0}=\beta.$$

(2) Asymptotic Normality.

$$\sqrt{N}\left(\hat{\beta}_{OLS}-\beta\right)=\left(N^{-1}\sum_{i=1}^{N}\mathbf{x}_i^{'}\mathbf{x}_i\right)^{-1}\left(\frac{1}{\sqrt{N}}\sum_{i=1}^{N}\mathbf{x}_i^{'}u_i\right).$$

Under **Assumption 3**,

$$\mathrm{E}\left(\frac{1}{N}\sum_{i=1}^{N}\mathbf{x}_i^{'}u_i\right)=\frac{1}{N}\sum_{i=1}^{N}\mathrm{E}\left(\mathbf{x}_i^{'}u_i\right)=\mathbf{0}.$$

Under **Assumption 5**,

$$\mathrm{Var}\left(\frac{1}{\sqrt{N}}\sum_{i=1}^{N}\mathbf{x}_i^{'}u_i\right)=\mathbf{B}_N\overset{p}{\rightarrow}\mathbf{B}.$$

Let $Z_N\equiv\frac{\frac{1}{N}\sum_{i=1}^{N}\mathbf{x}_i^{'}u_i-\mathrm{E}\left(\frac{1}{N}\sum_{i=1}^{N}\mathbf{x}_i^{'}u_i\right)}{\sqrt{\mathrm{Var}\left(\frac{1}{N}\sum_{i=1}^{N}\mathbf{x}_i^{'}u_i\right)}}$. Since $\mathbf{x}_i^{'}u_i$ is mixing, by the central limit theorem of Wooldridge and White (1988) for mixing sequences, $Z_N\overset{d}{\rightarrow}\mathrm{N}(0,1)$. Thus

$\frac{1}{\sqrt{N}}\sum_{i=1}^{N}\mathbf{x}_i^{\prime}u_i \xrightarrow{d} \mathrm{N}\left(\mathbf{0},\mathbf{B}\right)$. Therefore,

$$\sqrt{N}\left(\hat{\beta}_{OLS}-\beta\right)\to^{d}\mathrm{N}\left(\mathbf{0},\mathbf{A}_1^{-1}\mathbf{B}\mathbf{A}_1^{-1}\right).$$

Proof of Proposition 2

(1) Consistency.

$\hat{\Omega}_N \equiv \Omega_N\left(\hat{\lambda}\right)$is a consistent estimator in the sense that $\hat{\lambda}\xrightarrow{p}\lambda$. Thus $N^{-1}\mathbf{X}'\hat{\Omega}_N^{-1}\mathbf{X}\xrightarrow{p} N^{-1}\mathbf{X}'\Omega_N^{-1}\mathbf{X}$, and $N^{-1}\mathbf{X}'\hat{\Omega}_N^{-1}\mathbf{u}\xrightarrow{p} N^{-1}\mathbf{X}'\Omega_N^{-1}\mathbf{u}$.

$$\begin{aligned}\hat{\beta}_{FGLS} &= (\mathbf{X}'\hat{\Omega}_N^{-1}\mathbf{X})^{-1}\left(\mathbf{X}'\hat{\Omega}_N^{-1}\mathbf{Y}\right)\\ &= \beta+\left(N^{-1}\mathbf{X}'\hat{\Omega}_N^{-1}\mathbf{X}\right)^{-1}\left(N^{-1}\mathbf{X}'\hat{\Omega}_N^{-1}\mathbf{u}\right)\\ &\xrightarrow{p}\beta+\left(N^{-1}\mathbf{X}'\Omega_N^{-1}\mathbf{X}\right)^{-1}\left(N^{-1}\mathbf{X}'\Omega_N^{-1}\mathbf{u}\right)\\ &= \beta+\left[p\lim\left(N^{-1}\mathbf{X}'\Omega_N^{-1}\mathbf{X}\right)^{-1}+op\left(1\right)\right]\left[p\lim\left(N^{-1}\mathbf{X}'\Omega_N^{-1}\mathbf{u}\right)+op\left(1\right)\right].\end{aligned}$$

For any symmetric positive definite matrix Ω_N,

$$N^{-1}\mathbf{X}'\Omega_N^{-1}\mathbf{X}\xrightarrow{p}\mathrm{E}\left(N^{-1}\mathbf{X}'\Omega_N^{-1}\mathbf{X}\right)\xrightarrow{p}\mathbf{B},$$

$$p\lim\left(N^{-1}\mathbf{X}'\Omega_N^{-1}\mathbf{X}\right)^{-1}=\mathbf{B}^{-1}.$$

Under **Assumption 6**,

$$N^{-1}\mathbf{X}'\Omega_N^{-1}\mathbf{u}\xrightarrow{p}\mathrm{E}\left(N^{-1}\mathbf{X}'\Omega_N^{-1}\mathbf{u}\right)=\mathbf{0}.$$

Thus

$$\hat{\beta}_{FGLS} = \beta + \left[\mathbf{B}^{-1} + op\left(1\right)\right] \cdot op\left(1\right) \xrightarrow{p} \beta.$$

(2) Asymptotic Normality.

$$\begin{aligned}\sqrt{N}\left(\hat{\beta}_{FGLS} - \beta\right) &= \left(N^{-1}\mathbf{X}'\hat{\Omega}_N^{-1}\mathbf{X}\right)^{-1}\left(\frac{1}{\sqrt{N}}\mathbf{X}'\hat{\Omega}_N^{-1}\mathbf{u}\right) \\ &= \left(N^{-1}\mathbf{X}'\Omega_N^{-1}\mathbf{X}\right)^{-1}\left(\frac{1}{\sqrt{N}}\mathbf{X}'\Omega_N^{-1}\mathbf{u}\right) + op\left(1\right).\end{aligned}$$

Under **Assumption 6,** $\mathrm{E}\left(N^{-1}\mathbf{X}'\Omega_N^{-1}\mathbf{u}\right) = \mathbf{0}$.

Under **Assumption 7,**

$$\mathrm{Var}\left(\frac{1}{\sqrt{N}}\mathbf{X}'\Omega_N^{-1}\mathbf{u}\right) = \mathrm{E}\left(\frac{1}{N}\mathbf{X}'\boldsymbol{\Omega}_N^{-1}\mathbf{X}\right) \xrightarrow{p} \mathbf{B}_2.$$

Let $P_N \equiv \frac{N^{-1}\mathbf{X}'\Omega_N^{-1}\mathbf{u} - \mathrm{E}\left(N^{-1}\mathbf{X}'\Omega_N^{-1}\mathbf{u}\right)}{\sqrt{\mathrm{Var}\left(N^{-1}\mathbf{X}'\Omega_N^{-1}\mathbf{u}\right)}} = \frac{N^{-1}\mathbf{X}'\Omega_N^{-1}\mathbf{u} - \mathrm{E}\left(N^{-1}\mathbf{X}'\Omega_N^{-1}\mathbf{u}\right)}{\sqrt{N^{-1}\mathrm{Var}\left(\frac{1}{\sqrt{N}}\mathbf{X}'\Omega_N^{-1}\mathbf{u}\right)}}$, by the central limit theorem of Wooldridge and White (1988) for mixing sequences, $P_N \xrightarrow{d} \mathrm{N}\left(0, 1\right)$. Thus,

$$\frac{1}{\sqrt{N}}\mathbf{X}'\Omega_N^{-1}\mathbf{u} \rightarrow \mathrm{N}\left(\mathbf{0}, \mathbf{B}_2\right).$$

$$\sqrt{N}\left(\hat{\beta}_{FGLS} - \beta\right) \rightarrow \mathrm{N}\left(\mathbf{0}, \mathbf{B}_2^{-1}\mathbf{B}_2\mathbf{B}_2^{-1}\right) = \mathrm{N}\left(\mathbf{0}, \mathbf{B}_2^{-1}\right).$$

Proof of Proposition 3

In order to prove $\hat{\beta}_{FPGLS}$ is consistent and asymptotically normal, we need to prove that first $\hat{\beta}_{PGLS}$ is consistent and asymptotically normal, and that $\hat{\beta}_{FPGLS}$ and $\hat{\beta}_{PGLS}$ are asymptotically equivalent.

(1) Consistency.

$$\begin{aligned}
\hat{\beta}_{PGLS} &= \left(\frac{1}{G}\sum_{g=1}^{G}\mathbf{X}_g'\boldsymbol{\Lambda}_g^{-1}\mathbf{X}_g\right)^{-1}\left(\frac{1}{G}\sum_{g=1}^{G}\mathbf{X}_g'\boldsymbol{\Lambda}_g^{-1}\mathbf{y}_g\right) \\
&= \beta + \left(\frac{1}{G}\sum_{g=1}^{G}\mathbf{X}_g'\boldsymbol{\Lambda}_g^{-1}\mathbf{X}_g\right)^{-1}\left(\frac{1}{G}\sum_{g=1}^{G}\mathbf{X}_g'\boldsymbol{\Lambda}_g^{-1}\mathbf{u}_g\right)
\end{aligned}$$

Under **Assumption P3-P5**, since $\{(\mathbf{x}_i, u_i)\}$ is a mixing sequence, $\mathbf{X}_g'\boldsymbol{\Lambda}_g^{-1}\mathbf{X}_g$ and $\mathbf{X}_g'\boldsymbol{\Lambda}_g^{-1}\mathbf{u}_g$ are also two mixing sequences. By law of large numbers for mixing sequences (Theorem 3.57 in White), we have the following two equations.

$$\begin{aligned}
\frac{1}{G}\sum_{g=1}^{G}\mathbf{X}_g'\boldsymbol{\Lambda}_g^{-1}\mathbf{X}_g \xrightarrow{p} \mathrm{E}\left[\frac{1}{G}\sum_{g=1}^{G}\left(\mathbf{X}_g'\boldsymbol{\Lambda}_g^{-1}\mathbf{X}_g\right)\right] &= \mathbf{Q}_g \xrightarrow{p} \mathbf{Q} \\
\frac{1}{G}\sum_{g=1}^{G}\mathbf{X}_g'\boldsymbol{\Lambda}_g^{-1}\mathbf{u}_g \xrightarrow{p} \mathrm{E}\left(\mathbf{X}_g'\boldsymbol{\Lambda}_g^{-1}\mathbf{u}_g\right) &= \mathbf{0}.
\end{aligned}$$

Thus

$$\begin{aligned}
\hat{\beta}_{PGLS} &= \beta + \left(\frac{1}{G}\sum_{g=1}^{G}\mathbf{X}_g'\boldsymbol{\Lambda}_g^{-1}\mathbf{X}_g\right)^{-1}\left(\frac{1}{G}\sum_{g=1}^{G}\mathbf{X}_g'\boldsymbol{\Lambda}_g^{-1}\mathbf{u}_g\right) \\
&= \beta + \left[\mathbf{Q}^{-1} + op\left(1\right)\right]\cdot op\left(1\right) \\
&\to \beta
\end{aligned}$$

(2) Asymptotic Normality.

$$\sqrt{G}\left(\hat{\beta}_{PGLS} - \beta\right) = \left(\frac{1}{G}\sum_{g=1}^{G}\mathbf{X}_g'\boldsymbol{\Lambda}_g^{-1}\mathbf{X}_g\right)^{-1}\left(\frac{1}{\sqrt{G}}\sum_{g=1}^{G}\mathbf{X}_g'\boldsymbol{\Lambda}_g^{-1}\mathbf{u}_g\right),$$

By central limit theorem for mixing sequences (Wooldridge and White 1988),

$$\frac{1}{\sqrt{G}}\sum_{g=1}^{G}\mathbf{X}_g'\mathbf{\Lambda}_g^{-1}\mathbf{u}_g \to^d \mathrm{N}\left(\mathbf{0},\mathbf{S}\right),$$

$$\left(\frac{1}{G}\sum_{g=1}^{G}\mathbf{X}_g'\mathbf{\Lambda}_g^{-1}\mathbf{X}_g\right)^{-1} = \mathbf{Q}^{-1}+op\left(1\right).$$

Thus

$$\sqrt{G}\left(\hat{\beta}_{PGLS}-\beta\right) \xrightarrow{d} \mathrm{N}\left(0,\mathbf{Q}^{-1}\mathbf{S}\mathbf{Q}^{-1}\right).$$

(3) $\hat{\beta}_{FPGLS}$ and $\hat{\beta}_{PGLS}$ are asymptotically equivalent.

Let $\hat{\mathbf{\Lambda}} \equiv \mathbf{\Lambda}\left(\hat{\lambda}\right)$. Write down the formulas for the two estimators as follows.

$$\begin{aligned}\hat{\beta}_{FPGLS} &= \left(\sum_{g=1}^{G}\mathbf{X}_g'\hat{\mathbf{\Lambda}}_g^{-1}\mathbf{X}_g\right)^{-1}\left(\sum_{g=1}^{G}\mathbf{X}_g'\hat{\mathbf{\Lambda}}_g^{-1}\mathbf{y}_g\right),\\ \hat{\beta}_{PGLS} &= \left(\sum_{g=1}^{G}\mathbf{X}_g'\mathbf{\Lambda}_g^{-1}\mathbf{X}_g\right)^{-1}\left(\sum_{g=1}^{G}\mathbf{X}_g'\mathbf{\Lambda}_g^{-1}\mathbf{y}_g\right).\end{aligned}$$

In order to prove the asymptotic equivalence of $\hat{\beta}_{FPGLS}$ and $\hat{\beta}_{PGLS}$, we need to prove the following two conditions hold. The procedure follows Schmidt (1971).

$$\begin{aligned}\left(\frac{1}{G}\sum_{g=1}^{G}\mathbf{X}_g\hat{\mathbf{\Lambda}}_g^{-1}\mathbf{X}_g\right)^{-1} - \left(\frac{1}{G}\sum_{g=1}^{G}\mathbf{X}_g'\mathbf{\Lambda}_g^{-1}\mathbf{X}_g\right) &\xrightarrow{p} op\left(1\right),\\ \left(\frac{1}{\sqrt{G}}\sum_{g=1}^{G}\mathbf{X}_g'\hat{\mathbf{\Lambda}}_g^{-1}\mathbf{u}_g\right) - \left(\frac{1}{\sqrt{G}}\sum_{g=1}^{G}\mathbf{X}_g'\mathbf{\Lambda}_g^{-1}\mathbf{u}_g\right) &\xrightarrow{p} op\left(1\right).\end{aligned}$$

Lemma 1 $\frac{1}{G}\sum_{g=1}^{G}\mathbf{X}_g'\hat{\mathbf{\Lambda}}_g^{-1}\mathbf{X}_g - \frac{1}{G}\sum_{g=1}^{G}\mathbf{X}_g'\mathbf{\Lambda}_g^{-1}\mathbf{X}_g = op\left(1\right)$

Proof:

We have consistent estimator $\hat{\lambda} \to \lambda$, $p\lim\left(\hat{\lambda}\right) = \lambda$. $\hat{\mathbf{\Lambda}}_g^{-1}$ is a continuous function of $\hat{\lambda}$, thus $\hat{\mathbf{\Lambda}}_g^{-1} \xrightarrow{p} \Lambda_g^{-1}$.

$$\mathbf{X}_g'\hat{\mathbf{\Lambda}}_g^{-1}\mathbf{X}_g \xrightarrow{p} \mathbf{X}_g'\mathbf{\Lambda}_g^{-1}\mathbf{X}_g$$

$$\mathbf{X}_g'\hat{\mathbf{\Lambda}}_g^{-1}\mathbf{X}_g - \mathbf{Q}_g \xrightarrow{p} \mathbf{X}_g'\mathbf{\Lambda}_g^{-1}\mathbf{X}_g - \mathbf{Q}_g$$

$$\frac{1}{G}\sum_{g=1}^{G}\mathbf{X}_g'\hat{\mathbf{\Lambda}}_g^{-1}\mathbf{X}_g - \frac{1}{G}\sum_{g=1}^{G}\mathbf{Q}_g \xrightarrow{p} \frac{1}{G}\sum_{g=1}^{G}\mathbf{X}_g'\mathbf{\Lambda}_g^{-1}\mathbf{X}_g - \frac{1}{G}\sum_{g=1}^{G}\mathbf{Q}_g$$

$$\begin{aligned}
\left(\frac{1}{G}\sum_{g=1}^{G}\mathbf{X}_g'\hat{\mathbf{\Lambda}}_g^{-1}\mathbf{X}_g - \frac{1}{G}\sum_{g=1}^{G}\mathbf{Q}_g\right) - \left(\frac{1}{G}\sum_{g=1}^{G}\mathbf{X}_g'\mathbf{\Lambda}_g^{-1}\mathbf{X}_g - \frac{1}{G}\sum_{g=1}^{G}\mathbf{Q}_g\right) &= op\left(1\right) \\
\left(\frac{1}{G}\sum_{g=1}^{G}\mathbf{X}_g'\hat{\mathbf{\Lambda}}_g^{-1}\mathbf{X}_g - \frac{1}{G}\sum_{g=1}^{G}\mathbf{Q}_g\right) - op\left(1\right) &= op\left(1\right) \\
\left(\frac{1}{G}\sum_{g=1}^{G}\mathbf{X}_g'\hat{\mathbf{\Lambda}}_g^{-1}\mathbf{X}_g - \frac{1}{G}\sum_{g=1}^{G}\mathbf{Q}_g\right) &= op\left(1\right)
\end{aligned}$$

Thus the first equation holds.

Lemma 2 $\frac{1}{\sqrt{G}}\sum_{g=1}^{G}\mathbf{X}_g'\hat{\mathbf{\Lambda}}_g^{-1}\mathbf{u}_g - \frac{1}{\sqrt{G}}\sum_{g=1}^{G}\mathrm{E}\left(\mathbf{X}_g'\mathbf{\Lambda}_g^{-1}\mathbf{u}_g\right) = op\left(1\right).$

Similar to the argument in Lemma 1, we can get

$$\begin{aligned}
\frac{1}{G}\sum_{g=1}^{G}\mathbf{X}_g'\hat{\mathbf{\Lambda}}_g^{-1}\mathbf{u}_g - \frac{1}{G}\sum_{g=1}^{G}\mathrm{E}\left(\mathbf{X}_g'\mathbf{\Lambda}_g^{-1}\mathbf{u}_g\right) &= op\left(1\right), \\
\frac{1}{G}\sum_{g=1}^{G}\mathbf{X}_g'\hat{\mathbf{\Lambda}}_g^{-1}\mathbf{u}_g - \mathbf{0} &= op\left(1\right), \\
\frac{1}{G}\sum_{g=1}^{G}\mathbf{X}_g'\hat{\mathbf{\Lambda}}_g^{-1}\mathbf{u}_g &\to {}^{p}\mathbf{0},
\end{aligned}$$

$$\begin{aligned}\hat{\mathbf{S}}_G &= \widehat{\mathrm{Var}}\left(\frac{1}{\sqrt{G}}\sum_{g=1}^{G}\mathbf{X}_g'\hat{\mathbf{\Lambda}}_g^{-1}\mathbf{u}_g\right)\\ &= \frac{1}{G}\sum_{g=1}^{G}\sum_{h=1}^{G}\mathrm{E}\left(\mathbf{X}_g'\hat{\mathbf{\Lambda}}_g^{-1}\mathbf{u}_g\mathbf{u}_h'\hat{\mathbf{\Lambda}}_h^{-1}\mathbf{X}_h\right)\end{aligned}$$

$$\begin{aligned}\mathrm{E}\left(\mathbf{X}_g'\hat{\mathbf{\Lambda}}_g^{-1}\mathbf{u}_g\mathbf{u}_h'\hat{\mathbf{\Lambda}}_h^{-1}\mathbf{X}_h\right) &= \mathrm{E}\left\{\mathbf{X}_g'\left[\mathbf{\Lambda}_g^{-1}+op\left(1\right)\right]\mathbf{u}_g\mathbf{u}_h'\left[\mathbf{\Lambda}_h^{-1}+op\left(1\right)\right]\mathbf{X}_h\right\}\\ &= \mathrm{E}\left(\mathbf{X}_g'\mathbf{\Lambda}_g^{-1}\mathbf{u}_g\mathbf{u}_h'\mathbf{\Lambda}_h^{-1}\mathbf{X}_h\right)+\mathrm{E}\left(\mathbf{X}_g'\cdot op\left(1\right)\cdot\mathbf{u}_g\mathbf{u}_h'\mathbf{\Lambda}_h^{-1}\mathbf{X}_h\right)\\ &\quad +\mathrm{E}\left\{\mathbf{X}_g'\mathbf{\Lambda}_g^{-1}\mathbf{u}_g\mathbf{u}_h'\cdot op\left(1\right)\cdot\mathbf{X}_h\right\}\\ &\quad +\mathrm{E}\left\{\mathbf{X}_g'\cdot op\left(1\right)\cdot\mathbf{u}_g\mathbf{u}_h'\cdot op\left(1\right)\cdot\mathbf{X}_h\right\}\end{aligned}$$

$$\begin{aligned}\lim_{G\to\infty}\hat{\mathbf{S}}_G &= \lim_{G\to\infty}\frac{1}{G}\sum_{g=1}^{G}\sum_{h=1}^{G}\mathrm{E}\left(\mathbf{X}_g'\hat{\mathbf{\Lambda}}_g^{-1}\mathbf{u}_g\mathbf{u}_h'\hat{\mathbf{\Lambda}}_h^{-1}\mathbf{X}_h\right)\\ &= \lim_{G\to\infty}\frac{1}{G}\sum_{g=1}^{G}\sum_{h=1}^{G}\mathrm{E}\left(\mathbf{X}_g'\mathbf{\Lambda}_g^{-1}\mathbf{u}_g\mathbf{u}_h'\mathbf{\Lambda}_h^{-1}\mathbf{X}_h\right)+\mathbf{0}\\ &= \lim_{G\to\infty}\mathbf{S}_G=\mathbf{S}\end{aligned}$$

In addition, by the central limit theorem for mixing sequence in Wooldridge and White (1988), we get the asymptotic distribution for $\frac{1}{\sqrt{G}}\sum_{g=1}^{G}\mathbf{X}_g'\hat{\mathbf{\Lambda}}_g^{-1}\mathbf{u}_g$,

$$\frac{1}{\sqrt{G}}\sum_{g=1}^{G}\mathbf{X}_g'\hat{\mathbf{\Lambda}}_g^{-1}\mathbf{u}_g\to^{d}\mathrm{N}\left(\mathbf{0},\mathbf{S}\right),$$

which is the same as $\frac{1}{\sqrt{G}}\sum_{g=1}^{G}\mathbf{X}_g'\mathbf{\Lambda}_g^{-1}\mathbf{u}_g$. Thus the second equation holds:

$$\frac{1}{\sqrt{G}}\sum_{g=1}^{G}\mathbf{X}_g'\hat{\mathbf{\Lambda}}_g^{-1}\mathbf{u}_g-\frac{1}{\sqrt{G}}\sum_{g=1}^{G}\mathbf{X}_g'\mathbf{\Lambda}_g^{-1}\mathbf{u}_g=op\left(1\right).$$

Finally we can write:

$$\begin{aligned}
\sqrt{G}\left(\hat{\beta}_{FPGLS}-\beta\right) &= \left(\frac{1}{G}\sum_{g=1}^{G}\mathbf{X}_g'\hat{\mathbf{\Lambda}}_g^{-1}\mathbf{X}_g\right)^{-1}\left(\frac{1}{\sqrt{G}}\sum_{g=1}^{G}\mathbf{X}_g'\hat{\mathbf{\Lambda}}_g^{-1}\mathbf{u}_g\right) \\
&= \left[\left(\frac{1}{G}\sum_{g=1}^{G}\mathbf{X}_g'\mathbf{\Lambda}_g^{-1}\mathbf{X}_g\right)^{-1}+op(1)\right] \\
&\quad \left[\left(\frac{1}{\sqrt{G}}\sum_{g=1}^{G}\mathbf{X}_g'\mathbf{\Lambda}_g^{-1}\mathbf{u}_g\right)+op(1)\right] \\
&= \left(\frac{1}{G}\sum_{g=1}^{G}\mathbf{X}_g'\mathbf{\Lambda}_g^{-1}\mathbf{X}_g\right)^{-1}\left(\frac{1}{\sqrt{G}}\sum_{g=1}^{G}\mathbf{X}_g'\mathbf{\Lambda}_g^{-1}\mathbf{u}_g\right)+op(1) \\
&= \sqrt{G}\left(\hat{\beta}_{PGLS}-\beta\right)+op(1)
\end{aligned}$$

$$\sqrt{G}\left(\hat{\beta}_{FPGLS}-\hat{\beta}_{PGLS}\right)=op(1)$$

Therefore, $\hat{\beta}_{FPGLS}$ and $\hat{\beta}_{PGLS}$ are asymptotically equivalent.

Proof of Proposition 4

As stated in Section 4, the HAC estimator that is robust to groupwise spatial correlation and misspecification as in equation (1.26) is

$\widehat{\text{Avar}}\left(\hat{\beta}_{FPGLS}\right)_{rob}=\left[\sum_{g=1}^{G}\left(\mathbf{X}_g'\hat{\mathbf{\Lambda}}_g^{-1}\mathbf{X}_g\right)\right]^{-1}$
$\left[\sum_{g=1}^{G}\sum_{h=1}^{G}p\left(d_{gh}\right)\mathbf{X}_g'\hat{\mathbf{\Lambda}}_g^{-1}\tilde{\mathbf{u}}_g\tilde{\mathbf{u}}_h'\hat{\mathbf{\Lambda}}_h^{-1}\mathbf{X}_h\right]\left[\sum_{g=1}^{G}\left(\mathbf{X}_g'\hat{\mathbf{\Lambda}}_g^{-1}\mathbf{X}_g\right)\right]^{-1}.$

Define

$\hat{\mathbf{S}}\equiv\frac{1}{G}\sum_{g=1}^{G}\sum_{h=1}^{G}p\left(d_{gh}\right)\mathbf{X}_g'\hat{\mathbf{\Lambda}}_g^{-1}\tilde{\mathbf{u}}_g\tilde{\mathbf{u}}_h'\hat{\mathbf{\Lambda}}_g^{-1}\mathbf{X}_h$

$= \frac{1}{G}\sum_{g=1}^{G}\mathbf{X}_g'\hat{\mathbf{\Lambda}}_g^{-1}\tilde{\mathbf{u}}_g\tilde{\mathbf{u}}_g^{'}\hat{\mathbf{\Lambda}}_g^{-1}\mathbf{X}_g + \frac{1}{G}\sum_{g=1}^{G}\sum_{h\neq g}^{G} p\left(d_{gh}\right)\mathbf{X}_g'\hat{\mathbf{\Lambda}}_g^{-1}\tilde{\mathbf{u}}_g\tilde{\mathbf{u}}_h^{'}\hat{\mathbf{\Lambda}}_h^{-1}\mathbf{X}_h$

Define the following expressions:

$$\mathbf{S}^p = \frac{1}{G}\sum_{g=1}^{G}\sum_{h=1}^{G} p\left(d_{gh}\right)\mathbf{X}_g'\mathbf{\Lambda}_g^{-1}\mathbf{u}_g\mathbf{u}_h^{'}\mathbf{\Lambda}_h^{-1}\mathbf{X}_h \tag{4.1}$$

$$\mathbf{S}_0^p = \mathrm{E}\left[\frac{1}{G}\sum_{g=1}^{G}\sum_{h=1}^{G} p\left(d_{gh}\right)\mathbf{X}_g'\mathbf{\Lambda}_g^{-1}\mathbf{u}_g\mathbf{u}_h^{'}\mathbf{\Lambda}_h^{-1}\mathbf{X}_h\right] \tag{4.2}$$

$$\mathbf{S}_0 = \mathrm{E}\left[\frac{1}{G}\sum_{g=1}^{G}\sum_{h=1}^{G}\mathbf{X}_g'\mathbf{\Lambda}_g^{-1}\mathbf{u}_g\mathbf{u}_h^{'}\mathbf{\Lambda}_h^{-1}\mathbf{X}_h\right] \tag{4.3}$$

$$\left|\hat{\mathbf{S}} - \mathbf{S}\right| \leq \left|\hat{\mathbf{S}} - \mathbf{S}^p\right| + \left|\mathbf{S}^p - \mathbf{S}_0^p\right| + \left|\mathbf{S}_0^p - \mathbf{S}_0\right| + \left|\mathbf{S}_0 - \mathbf{S}\right| \tag{4.4}$$

In order to prove Proposition 4, we need to prove each of the above four terms on the right-hand side to be $op\left(1\right)$. Since $p\lim\left(\hat{\lambda}\right) = \lambda, \hat{\mathbf{\Lambda}}_g \to \mathbf{\Lambda}_g$, as $G \to \infty$. By **Assumption P6(b)**, $\left|\mathbf{S}_0 - \mathbf{S}\right| = op\left(1\right)$. Rewrite the other three terms as follows:

$$\left|\hat{\mathbf{S}} - \mathbf{S}^p\right| = \frac{1}{G}\sum_{g=1}^{G}\sum_{h=1}^{G} p\left(d_{gh}\right)\mathbf{X}_g'\mathbf{\Lambda}_g\left(\lambda\right)^{-1}\left(\tilde{\mathbf{u}}_g\tilde{\mathbf{u}}_h^{'} - \mathbf{u}_g\mathbf{u}_h^{'}\right)\mathbf{\Lambda}_h(\lambda)^{-\mathbf{1}}\mathbf{X}_h + op\left(1\right). \tag{4.5}$$

$$\begin{aligned}\left|\mathbf{S}^p - \mathbf{S}_0^p\right| &= \frac{1}{G}\sum_{g=1}^{G}\sum_{h=1}^{G} p\left(d_{gh}\right) \\ &\cdot\left\{\mathbf{X}_g'\mathbf{\Lambda}_g\left(\lambda\right)^{-1}\mathbf{u}_g\mathbf{u}_h^{'}\mathbf{\Lambda}_h(\lambda)^{-\mathbf{1}}\mathbf{X}_h - \mathrm{E}\left[\mathbf{X}_g'\mathbf{\Lambda}_g\left(\lambda\right)^{-1}\mathbf{u}_g\mathbf{u}_h^{'}\mathbf{\Lambda}_h(\lambda)^{-\mathbf{1}}\mathbf{X}_h\right]\right\}\end{aligned} \tag{4.6}$$

$$\left|\mathbf{S}_0^p - \mathbf{S}_0\right| = \mathrm{E}\left\{\frac{1}{G}\sum_{g=1}^{G}\sum_{h=1}^{G}\left[1 - p\left(d_{gh}\right)\right]\mathbf{X}_g'\mathbf{\Lambda}_g\left(\lambda\right)^{-1}\mathbf{u}_g\mathbf{u}_h^{'}\mathbf{\Lambda}_h(\lambda)^{-\mathbf{1}}\mathbf{X}_h\right\}. \tag{4.7}$$

First, prove the first term is $op\left(1\right).$

$$\begin{aligned}\tilde{\mathbf{u}}_g\tilde{\mathbf{u}}_h^{'} &= \mathbf{u}_g\mathbf{u}_h^{'} - \mathbf{u}_g\left(\hat{\beta}_{FPGLS}-\beta\right)^{'}\mathbf{X}_h^{\prime} - \mathbf{X}_g\left(\hat{\beta}_{FPGLS}-\beta\right)\mathbf{u}_h^{'} \\ &\quad +\mathbf{X}_g\left(\hat{\beta}_{FPGLS}-\beta\right)\left(\hat{\beta}_{FPGLS}-\beta\right)^{'}\mathbf{X}_h^{\prime}.\end{aligned} \tag{4.8}$$

It suffices to show that the averages of the last three terms converge in probability to zero. The average of the vec of the first term can be written as $\frac{1}{G}\sum_{g=1}^{G}\sum_{h=1}^{G}\left(\mathbf{X}_h\otimes\mathbf{u}_g\right)$ vec $\left(\hat{\beta}_{FPGLS}-\beta\right)$, which is $op\left(1\right)$ because $p\lim\hat{\beta}_{FPGLS}-\beta=\mathbf{0}$ and $\frac{1}{G}\sum_{g=1}^{G}\sum_{h=1}^{G}$ $\left(\mathbf{X}_h\otimes\mathbf{u}_g\right)\rightarrow^p\mathbf{0}$. The third term is the transposition of the second. The average of the last term can be written as $\frac{1}{G}\sum_{g=1}^{G}\sum_{h=1}^{G}\left(\mathbf{X}_h\otimes\mathbf{X}_g\right)vec[(\hat{\beta}_{FPGLS}-\beta)(\hat{\beta}_{FPGLS}-\beta)^{\prime}]$. $vec[(\hat{\beta}_{FPGLS}-\beta)(\hat{\beta}_{FPGLS}-\beta)^{\prime}]=op\left(1\right)$. Assuming $\mathbf{X}_h$ and $\mathbf{X}_g$ have finite second moment, each element of $\frac{1}{G}\Sigma_{g=1}^{G}\Sigma_{h=1}^{G}(\mathrm{X}_h\otimes\mathrm{X}_g)$ is $Op\left(1\right).$ Since $Op\left(1\right)\cdot op\left(1\right)=op\left(1\right)$, the last term is $op\left(1\right)$. Thus $\tilde{\mathbf{u}}_g\tilde{\mathbf{u}}_h^{'}-\mathbf{u}_g\mathbf{u}_h^{'}=op\left(1\right).$ Further assuming each element in $\mathbf{\Lambda}_h$ is finite, each element in the first term is $op\left(1\right).$

The rest terms can be proved to be $op\left(1\right)$ by similar arguments.

Proof of Proposition 5

A sufficient condition is L_G is stochastically equicontinuous according to Newey (1991).

$$\begin{aligned}L_G &= \sum_{g=1}^{G}l_g\left(\beta,\lambda\right) \\ &= \sum_{g=1}^{G}l_g\left(\beta_0,\lambda_0\right)+\sum_{g=1}^{G}s_g\left(\ddot{\beta},\ddot{\lambda}\right)\begin{pmatrix}\beta-\beta_0\\ \lambda-\lambda_0\end{pmatrix},\end{aligned}$$

where $\ddot{\beta}$ is between β_0 and β, and $\ddot{\lambda}$ is between λ_0 and λ.

$$\begin{aligned}
& \sup \left\| L_G(\theta) - L_G(\theta_0) \right\| \\
= & \sup \left\| \sum_{g=1}^{G} l_g(\beta_0, \lambda_0) + \sum_{g=1}^{G} s_g\left(\ddot{\beta}, \ddot{\lambda}\right) \begin{pmatrix} \beta - \beta_0 \\ \lambda - \lambda_0 \end{pmatrix} - \sum_{g=1}^{G} l_g(\beta_0, \lambda_0) \right\| \\
= & \sup \left\| \sum_{g=1}^{G} s_g\left(\ddot{\beta}, \ddot{\lambda}\right) \begin{pmatrix} \beta - \beta_0 \\ \lambda - \lambda_0 \end{pmatrix} \right\| \\
= & \sup \left\| \sum_{g=1}^{G} s_g\left(\ddot{\beta}, \ddot{\lambda}\right) \right\| \left\| \begin{pmatrix} \beta - \beta_0 \\ \lambda - \lambda_0 \end{pmatrix} \right\| \\
< & \infty.
\end{aligned}$$

Thus $L_G(\theta)$ is stochastically equicontinuous. The PQMLE estimator is consistent.

The $(k+p) \times 1$ score of the log likelihood of group g is simply

$$\mathbf{s}_g(\beta, \lambda) \equiv \begin{pmatrix} \nabla_\beta l_g(\beta, \lambda)' \\ \nabla_\lambda l_g(\beta, \lambda)' \end{pmatrix},$$

where

$$\frac{\partial L(\beta; \lambda)}{\partial \beta} = \nabla_\beta l_g(\beta, \lambda)' = \mathbf{X}_g' \Lambda_g^{-1}(\lambda) \left(\mathbf{y}_g - \mathbf{X}_g \beta\right),$$

and

$$\begin{aligned}
\frac{\partial L(\beta; \lambda)}{\partial \lambda} &= \nabla_\lambda l_g(\beta, \lambda)' \\
&= -\frac{1}{2} \nabla_\lambda \Lambda_g(\lambda)' \, vec\left[\Lambda_g^{-1}(\lambda)\right] \\
&\quad + \frac{1}{2} \nabla_\lambda \Lambda_g(\lambda)' \operatorname{vec}\left\{\boldsymbol{\Lambda}_g^{-1}(\lambda)\left[\mathbf{u}_g \mathbf{u}_g' - \Lambda_g(\lambda)\right] \Lambda_g^{-1}(\lambda)\right\} \\
&= \frac{1}{2} \nabla_\lambda \Lambda_g(\lambda)' \left[\boldsymbol{\Lambda}_g^{-1}(\lambda) \otimes \Lambda_g^{-1}(\lambda)\right] \operatorname{vec}\left[\mathbf{u}_g \mathbf{u}_g' - \Lambda_g(\lambda)\right]
\end{aligned}$$

Set

$$\sum_{g=1}^{G} \mathbf{s}_g\left(\beta, \lambda\right) = \mathbf{0}.$$

$$\hat{\beta}_{PQMLE} = \left(\sum_{g=1}^{G} \mathbf{X}_g' \mathbf{\Lambda}_g^{-1}\left(\hat{\lambda}\right)\mathbf{X}_g\right)^{-1} \sum_{g=1}^{G} \mathbf{X}_g' \mathbf{\Lambda}_g^{-1}\left(\hat{\lambda}\right)\mathbf{y}_g,$$

$\hat{\lambda}_{PQMLE}$ may or may not have a closed functional form.

The groupwise QMLE is normally distributed asymptotically.

$$\sqrt{G}\left(\hat{\theta} - \theta_0\right) \Rightarrow^d \mathrm{N}\left(0, \mathbf{A}^{-1}\mathbf{B}\mathbf{A}^{-1}\right)$$

$$\begin{aligned} S_G\left(\hat{\theta}\right) &= \sum_{g=1}^{G} s_g\left(\hat{\theta}\right) = \mathbf{0} \\ &= \sum_{g=1}^{G} s_g\left(\theta_0\right) + \sum_{g=1}^{G} H_g\left(\ddot{\theta}\right)\left(\hat{\theta} - \theta_0\right) \end{aligned}$$

$$\begin{aligned} \left(\hat{\theta} - \theta_0\right) &= \left[-\sum_{g=1}^{G} \mathbf{H}_g\left(\ddot{\theta}\right)\right]^{-1} \sum_{g=1}^{G} s_g\left(\theta_0\right) \\ \sqrt{G}\left(\hat{\theta} - \theta_0\right) &= \left[-\frac{1}{G}\sum_{g=1}^{G} \mathbf{H}_g\left(\ddot{\theta}\right)\right]^{-1} \frac{1}{\sqrt{G}} \sum_{g=1}^{G} \mathbf{s}_g\left(\theta_0\right) \end{aligned}$$

By ergodic theorem for mixing fields,

$$\begin{aligned} -\frac{1}{G}\sum_{g=1}^{G} \mathbf{H}_g\left(\ddot{\theta}\right) &\xrightarrow{p} \mathbf{A} \\ \mathbf{A} &= \mathrm{E}\left[\mathbf{H}_g\left(\theta_0\right)\right] \end{aligned}$$

By central limit theorem in Jenish and Prucha (2007),

$$\frac{1}{\sqrt{G}}\sum_{g=1}^{G}\mathbf{s}_g(\theta_0) \overset{d}{\Rightarrow} \mathrm{N}(0,\mathbf{B})$$
$$\mathbf{B} = \lim_{G\to\infty}\frac{1}{G}\mathrm{Var}\left[\sum_{g=1}^{G}s_g(\theta_0)\right].$$

Thus we have

$$\sqrt{G}\left(\hat{\theta}-\theta_0\right) \overset{d}{\Rightarrow} \mathbf{N}\left(0,\mathbf{A}^{-1}\mathbf{B}\mathbf{A}^{-1}\right).$$

4.2 Proofs of Propositions in Chapter 2

This section provides proofs of propositions in Chapter 2.

Proof of Proposition 6

The proof follows **Theorem 2** in Jenish and Prucha (2009). A sufficient condition for consistency estimators is that the objective function satisfies the uniform law of large numbers (ULLN). To prove the ULLN, we need a pointwise law of large numbers (LLN). From **Assumption 4**, we have a pointwise LLN, and we get $Q_N(\theta) - \bar{Q}_N(\theta) \to^{a.s.} 0$, where $\bar{Q}_N(\theta) = \mathrm{E}(Q_N(\theta))$ .Since $Q_N(\theta)$ is the sample average of q_i, it is also stochastically equicontinuous. Then we have $\sup_\theta\left|Q_N(\theta) - \ddot{Q}(\theta)\right| \to^{a.s.} 0$ as $N \to \infty$, which is the uniform law of large numbers.

Proof of Proposition 7

Proof of Proposition 7 is similar to the proof of Proposition 6. Instead of writing the objective function in the sum of individuals, we write it as the sum of groups. The pointwise LLN can be written as $Q_G(\theta) - \bar{Q}_G(\theta) \rightarrow^{a.s.} \mathbf{0}$, where $\bar{Q}_G(\theta) = \mathrm{E}(Q_G(\theta))$. In addition with the stochatically equicontinuity condition, $\sup_\theta \left| Q_G(\theta) - \ddot{Q}(\theta) \right| \rightarrow^{a.s.} 0$ as $G \rightarrow \infty$.

Proof of Proposition 8

$$
\begin{aligned}
S_G\left(\check{\theta}\right) &= \sum_{g=1}^{G} s_g\left(\check{\theta}\right) = 0 \\
&= \sum_{g=1}^{G} s_g(\theta_0) + \sum_{g=1}^{G} H_g\left(\ddot{\theta}\right)\left(\hat{\theta} - \theta_0\right)
\end{aligned}
$$

$$
\begin{aligned}
\left(\check{\theta} - \theta_0\right) &= \left[-\sum_{g=1}^{G} \mathbf{H}_g\left(\ddot{\theta}\right)\right]^{-1} \sum_{g=1}^{G} s_g(\theta_0) \\
\sqrt{G}\left(\check{\theta} - \theta_0\right) &= \left[-\frac{1}{G}\sum_{g=1}^{G} \mathbf{H}_g\left(\ddot{\theta}\right)\right]^{-1} \frac{1}{\sqrt{G}}\sum_{g=1}^{G} s_g(\theta_0)
\end{aligned}
$$

By uniform law of large numbers for mixing fields,

$$
\begin{aligned}
-\frac{1}{G}\sum_{g=1}^{G} \mathbf{H}_g\left(\ddot{\theta}\right) &\rightarrow {}^{p}\mathbf{A}_0 \\
\mathbf{A}_0 &= \mathrm{E}\left[\mathbf{H}_g(\theta_0)\right]
\end{aligned}
$$

By Central Limit Theorem in Jenish and Prucha (2009),

$$
\begin{aligned}
\frac{1}{\sqrt{G}}\sum_{g=1}^{G} s_g(\theta_0) &\Rightarrow {}^{d}N(0, \mathbf{B}_0) \\
\mathbf{B}_0 &= \lim_{G\rightarrow\infty} \frac{1}{G} \mathrm{Var}\left[\sum_{g=1}^{G} s_g(\theta_0)\right]
\end{aligned}
$$

Thus we have

$$\sqrt{G}\left(\breve{\theta}-\theta_0\right) \Rightarrow^d N\left(0, \mathbf{A}_0^{-1}\mathbf{B}_0\mathbf{A}_0^{-1}\right).$$

Proof of Proposition 9

$$\sqrt{G}\left(\hat{\theta}_G - \theta_0\right) = \left[-\frac{1}{G}\sum_{g=1}^{G}\mathbf{H}_g\left(\ddot{\theta};\hat{\gamma}\right)\right]^{-1}\sum_{g=1}^{G}s_g\left(\theta_0;\hat{\gamma}\right),$$

$$\sqrt{G}\left(\hat{\theta}_G - \theta_0\right) = \mathbf{A}_0\left(-\frac{1}{\sqrt{G}}\sum_{g=1}^{G}\mathbf{s}_g\left(\theta_0;\hat{\gamma}\right)\right) + op\left(1\right),$$

where $\mathbf{A}_0 = \lim_{G\to\infty}\left\{-\frac{1}{G}\sum_{g=1}^{G}\mathrm{E}\left[\mathbf{H}_g\left(\theta_0,\gamma^*\right)\right]\right\}$ by law of large numbers for mixing sequences.

$$\frac{1}{\sqrt{G}}\sum_{g=1}^{G}s_g\left(\theta_0;\hat{\gamma}\right) = \frac{1}{\sqrt{G}}\sum_{i=1}^{G}s_g\left(\theta_0;\gamma^*\right) + \mathbf{F}_0\sqrt{G}\left(\hat{\gamma}-\gamma^*\right) + op\left(1\right),$$

where $\mathbf{F}_0$ is a $P \times J$ matrix, $\mathbf{F}_0 = \lim_{i\to\infty}\left\{\frac{1}{G}\sum_{g=1}^{G}\mathrm{E}\left[\nabla_\gamma s_g\left(\mathbf{w}_i,\theta_0;\gamma^*\right)\right]\right\}$.

Assume $\sqrt{G}\left(\hat{\gamma}-\gamma^*\right)$ can be written in the following form:

$$\sqrt{G}\left(\hat{\gamma}-\gamma^*\right) = \frac{1}{\sqrt{G}}\sum_{g=1}^{G}\mathbf{r}_g\left(\gamma^*\right) + op\left(1\right),$$

where $\mathbf{r}_g\left(\gamma^*\right)$ is a $J \times 1$ vector with $\mathbf{E}\left[\mathbf{r}_g\left(\gamma^*\right)\right] = \mathbf{0}$. Then

$$\sqrt{G}\left(\hat{\theta}-\theta_0\right) = \mathbf{A}_0\frac{1}{\sqrt{G}}\sum_{g=1}^{G}\left[-\mathbf{e}_g\left(\theta_0;\gamma^*\right)\right] + op\left(1\right)$$

$$\mathbf{e}_g\left(\theta_0;\gamma^*\right) \equiv \mathbf{s}_g\left(\theta_0;\gamma^*\right) + \mathbf{F}_0\mathbf{r}_g\left(\gamma^*\right)$$

$$\mathbf{D}_0 \equiv \lim_{G\to\infty}\frac{1}{G}\mathrm{E}\left[\sum_{g=1}^{G}\mathbf{e}_g\left(\theta_0;\gamma^*\right)\sum_{h=1}^{G}\mathbf{e}_h\left(\theta_0;\gamma^*\right)'\right]$$

$$\text{Avar}\sqrt{G}\left(\hat{\theta}-\theta_0\right)=\mathbf{A}_0^{-1}\mathbf{D}_0\mathbf{A}_0^{-1}$$

In the case of QMLE, we can see $\nabla\mathbf{m}_g'(\theta)$ and $\mathbf{m}_g(\theta)$ do not rely on γ. When we take derivatives with respect to γ, it only matters with $\mathbf{W}_g^{-1}$. $\nabla_\gamma\mathbf{s}_g\left(\mathbf{w}_g,\theta_0;\gamma^*\right)$ is a linear combination of elements of $\left[\mathbf{y}_g-\mathbf{m}_g(\theta)\right]$. Since the $\text{E}\left[\left(\mathbf{y}_g-\mathbf{m}_g\right)|\mathbf{w},\mathbf{D}\right]=\mathbf{0}$, $\text{E}\left[\nabla_\gamma\mathbf{s}_g\left(\mathbf{w}_g,\theta_0;\gamma^*\right)|\mathbf{w},\mathbf{D}\right]$ $=\mathbf{0}$. By law of iterated expectations, $\text{E}\left[\nabla_\gamma\mathbf{s}_g\left(\mathbf{w}_g,\theta_0;\gamma^*\right)\right]=\mathbf{0}$. Thus $\mathbf{F}_0=\mathbf{0}$.

$$\text{Avar}\sqrt{G}\left(\hat{\theta}-\theta_0\right)=\mathbf{A}_0^{-1}\mathbf{B}_0\mathbf{A}_0^{-1}$$

and $\mathbf{B}_0=\lim_{G\to\infty}\text{Var}\left[\frac{1}{\sqrt{G}}\sum_{g=1}^{G}s_g\left(\theta_0,\gamma^*\right)\right]$

Proof of Proposition 10

Let $\nabla\hat{\mathbf{m}}_g=\nabla\mathbf{m}_g\left(\hat{\theta}\right)$, $\quad\nabla\hat{\mathbf{m}}_g'=\nabla\mathbf{m}_g'\left(\hat{\theta}\right)$. The proof consists two parts: (1) $\hat{\mathbf{A}}\to\mathbf{A}_0$; (2)$\hat{\mathbf{B}}_2\to\mathbf{B}_0$.

Part (1) prove that $\hat{\mathbf{A}}\to\mathbf{A}_0$.

$\hat{\mathbf{A}}=\frac{1}{G}\sum_{g=1}^{G}\nabla\hat{\mathbf{m}}_g'\hat{\mathbf{W}}_g^{-1}\nabla\hat{\mathbf{m}}_g\to\lim_{G\to\infty}\frac{1}{G}\sum_{g=1}^{G}\text{E}\left(\nabla\mathbf{m}_g'\mathbf{W}_g^{-1}\nabla\mathbf{m}_g\right)=\mathbf{A}_0$

Part (2) prove that $\hat{\mathbf{B}}_2\to\mathbf{B}_0$.

$$\hat{\mathbf{B}}_2=\frac{1}{G}\sum_{g=1}^{G}\sum_{h=1}^{G}k(d_{gh})\nabla\hat{\mathbf{m}}_g'\hat{\mathbf{W}}_g^{-1}\hat{\mathbf{u}}_g\hat{\mathbf{u}}_h'\hat{\mathbf{W}}_h^{-1}\nabla\hat{\mathbf{m}}_h,$$

and

$$\begin{aligned}\mathbf{B}_0 &= \lim_{G\to\infty}\frac{1}{G}\sum_{g=1}^{G}\text{E}\left[\nabla\mathbf{m}_g'\mathbf{W}_g^{-1}\mathbf{u}_g\mathbf{u}_g'\mathbf{W}_g^{-1}\nabla\mathbf{m}_g\right]\\ &\quad+\lim_{G\to\infty}\frac{1}{G}\sum_{g=1}^{G}\sum_{g\neq h}^{G}\text{E}\left[\nabla\mathbf{m}_g'\mathbf{W}_g^{-1}\mathbf{u}_g\mathbf{u}_h\mathbf{W}_h^{-1}\nabla\mathbf{m}_h\right].\end{aligned}$$

Define

$$Z_g = \nabla \mathbf{m}_g \mathbf{W}_g^{-1} \mathbf{u}_g$$

$$\hat{Z}_g = \nabla \hat{\mathbf{m}}_g' \hat{\mathbf{W}}_g^{-1} \hat{\mathbf{u}}_g$$

For a given G,

$$\begin{aligned}
\hat{\mathbf{B}}_2 &= \frac{1}{G} \sum_{g=1}^{G} \nabla \hat{\mathbf{m}}_g' \hat{\mathbf{W}}_g^{-1} \hat{\mathbf{u}}_g \hat{\mathbf{u}}_g' \hat{\mathbf{W}}_g^{-1} \nabla \hat{\mathbf{m}}_g + \frac{1}{G} \sum_{g=1}^{G} \sum_{g \neq h}^{G} k(d_{gh}) \nabla \hat{\mathbf{m}}_h \hat{\mathbf{W}}_g^{-1} \hat{\mathbf{u}}_g \hat{\mathbf{u}}_h' \hat{\mathbf{W}}_h^{-1} \nabla \hat{\mathbf{m}}_h \\
&= \frac{1}{G} \sum_{g=1}^{G} \hat{Z}_g' \hat{Z}_g + \frac{1}{G} \sum_{g=1}^{G} \sum_{g \neq h}^{G} k(d_{gh}) \hat{Z}_g' \hat{Z}_h,
\end{aligned}$$

$$\begin{aligned}
\mathbf{B}_0 &= \frac{1}{G} \sum_{g=1}^{G} \mathrm{E}\left[\nabla \mathbf{m}_g' \mathbf{W}_g^{-1} \mathbf{u}_g \mathbf{u}_g' \mathbf{W}_g^{-1} \nabla \mathbf{m}_g \right] + \frac{1}{G} \sum_{g=1}^{G} \sum_{g \neq h}^{G} \mathrm{E}\left[\nabla \mathbf{m}_g' \mathbf{W}_g^{-1} \mathbf{u}_g \mathbf{u}_h' \mathbf{W}_h^{-1} \nabla \mathbf{m}_h \right] \\
&= \frac{1}{G} \sum_{g=1}^{G} \mathrm{E}\left[Z_g' Z_g \right] + \frac{1}{G} \sum_{g=1}^{G} \sum_{g \neq h}^{G} \mathrm{E}\left[Z_g' Z_h \right].
\end{aligned}$$

Define

$$\begin{aligned}
\mathbf{B}_0^k &= \frac{1}{G} \sum_{g=1}^{G} \mathrm{E}\left[\nabla \mathbf{m}_g' \mathbf{W}_g^{-1} \mathbf{u}_g \mathbf{u}_g' \mathbf{W}_g^{-1} \nabla \mathbf{m}_g \right] \\
&\quad + \frac{1}{G} \sum_{g=1}^{G} \sum_{g \neq h}^{G} k(d_{gh}) \mathrm{E}\left[\nabla \mathbf{m}_g' \mathbf{W}_g^{-1} \mathbf{u}_g \mathbf{u}_h' \mathbf{W}_h^{-1} \nabla \mathbf{m}_h \right] \\
&= \frac{1}{G} \sum_{g=1}^{G} \mathrm{E}\left(Z_g' Z_g \right) + \frac{1}{G} \sum_{g=1}^{G} \sum_{g \neq h}^{G} k(d_{gh}) \mathrm{E}\left(Z_g' Z_h \right),
\end{aligned}$$

$$
\begin{aligned}
\mathbf{B}^{k} &= \frac{1}{G}\sum_{g=1}^{G}\left[\nabla\mathbf{m}_g'\mathbf{W}_g^{-1}\mathbf{u}_g\mathbf{u}_g'\mathbf{W}_g^{-1}\nabla\mathbf{m}_g\right]+\frac{1}{G}\sum_{g=1}^{G}\sum_{g\neq h}^{G}k(d_{gh})\left[\nabla\mathbf{m}_g'\mathbf{W}_g^{-1}\mathbf{u}_g\mathbf{u}_h'\mathbf{W}_h^{-1}\nabla\mathbf{m}_h\right]\\
&= \frac{1}{G}\sum_{g=1}^{G}Z_g^{'}Z_g+\frac{1}{G}\sum_{g=1}^{G}\sum_{g\neq h}^{G}k(d_{gh})Z_g^{'}Z_h.
\end{aligned}
$$

Next,

$$
\begin{aligned}
\left|\hat{\mathbf{B}}_2-\mathbf{B}_0\right| &= \left|\hat{\mathbf{B}}_2-\mathbf{B^k}+\mathbf{B^k}-\mathbf{B_0^k}+\mathbf{B_0^k}-\mathbf{B}_0\right|\\
&\leq \left|\hat{\mathbf{B}}_2-\mathbf{B^k}\right|+\left|\mathbf{B^k}-\mathbf{B_0^k}\right|+\left|\mathbf{B_0^k}-\mathbf{B}_0\right|\\
&= \left|\hat{\mathbf{B}}_2-\left[\frac{1}{G}\sum_{g=1}^{G}Z_g^{'}Z_g+\frac{1}{G}\sum_{g=1}^{G}\sum_{g\neq h}^{G}k(d_{gh})Z_g^{'}Z_h\right]\right|\\
&\quad+\left|\frac{1}{G}\sum_{g=1}^{G}\left[Z_g^{'}Z_g-\mathrm{E}\left(Z_g^{'}Z_g\right)\right]+\frac{1}{G}\sum_{g=1}^{G}\sum_{g\neq h}^{G}k(d_{gh})\left[Z_g^{'}Z_h-\mathrm{E}\left(Z_g^{'}Z_h\right)\right]\right|\\
&\quad+\left|\frac{1}{G}\sum_{g=1}^{G}\mathrm{E}\left(Z_g^{'}Z_g\right)+\frac{1}{G}\sum_{g=1}^{G}\sum_{g\neq h}^{G}k(d_{gh})\mathrm{E}\left(Z_g^{'}Z_h\right)-\mathbf{B}_0\right|\\
&= \left|\hat{\mathbf{B}}_2-\left[\frac{1}{G}\sum_{g=1}^{G}Z_g^{'}Z_g+\frac{1}{G}\sum_{g=1}^{G}\sum_{g\neq h}^{G}k(d_{gh})Z_g^{'}Z_h\right]\right|\\
&\quad+\left|\frac{1}{G}\sum_{g=1}^{G}\left[Z_g^{'}Z_g-\mathrm{E}\left(Z_g^{'}Z_g\right)\right]+\frac{1}{G}\sum_{g=1}^{G}\sum_{g\neq h}^{G}k(d_{gh})\left[Z_g^{'}Z_h-\mathrm{E}\left(Z_g^{'}Z_h\right)\right]\right|\\
&\quad+\left|\frac{1}{G}\sum_{g=1}^{G}\sum_{g\neq h}^{G}k(d_{gh})\mathrm{E}\left(Z_g^{'}Z_h\right)-\frac{1}{G}\sum_{g=1}^{G}\sum_{g\neq h}^{G}\mathrm{E}\left[Z_g^{'}Z_h\right]\right|\\
&\leq \left|\hat{\mathbf{B}}_2-\left[\frac{1}{G}\sum_{g=1}^{G}Z_g^{'}Z_g+\frac{1}{G}\sum_{g=1}^{G}\sum_{g\neq h}^{G}k(d_{gh})Z_g^{'}Z_h\right]\right| \qquad (4.9)\\
&\quad+\left|\frac{1}{G}\sum_{g=1}^{G}\left[Z_g^{'}Z_g-\mathrm{E}\left(Z_g^{'}Z_g\right)\right]+\frac{1}{G}\sum_{g=1}^{G}\sum_{g\neq h}^{G}k(d_{gh})\left[Z_g^{'}Z_h-\mathrm{E}\left(Z_g^{'}Z_h\right)\right]\right|\\
&\quad+\frac{1}{G}\sum_{g=1}^{G}\sum_{g\neq h}^{G}\left|k(d_{gh})-1\right|\left|\mathrm{E}\left(Z_g^{'}Z_h\right)\right|
\end{aligned}
$$

Next, prove each of the right hand side term goes to zero. Define $p_{gh} = Z_g^{\prime} Z_h - \mathrm{E}\left(Z_g^{\prime} Z_h\right)$. Use this device to make the mean value expansion go to zero and this will complete the proof.

4.3 Proofs of Theorems in Chapter 3

This section provides proofs of theorems in Chapter 3.

Proof of Theorem 1

Amemiya (1985, pp. 182-183) showed the equivalence of conditions (A), (B), (C), (D) and (E). It is trivial that (F) implies (E) and that (F') implies (F). We still need to establish that one of (A), (B), (C), (D) or (E) implies (F').

Proof that (D) implies (F'): Condition (D) says that $X = F_1 A$ for some nonsingular A, where the eigenvectors of Σ are $F = [F_1, F_2]$ and the eigenvalues are $\Lambda = \begin{bmatrix} \Lambda_1 & 0 \\ 0 & \Lambda_2 \end{bmatrix}$, a diagonal matrix. We can choose F such that $F\prime F = I$ and therefore $F_1\prime F_2 = 0$. Then the spectral representation of Σ says

$$\Sigma = F\Lambda F\prime = F_1\Lambda_1 F_1\prime + F_2\Lambda_2 F_2\prime.$$

Since $F_1 = XA^{-1}$,

$$\Sigma = X(A^{-1}\Lambda_1 A\prime^{-1})X\prime + F_2\Lambda_2 F_2\prime.$$

Finally, $X\prime F_2 = A\prime F_1\prime F_2 = 0$. So (F') holds.

Because the result is somewhat counterintuitive, we also give a proof that (F) implies (F').

Proof that (F) implies (F'): Suppose that (F) holds so $\Sigma = X\Gamma X\prime + Q\Theta Q\prime + c^2 I$ with $Q\prime X = 0$. Now define $P_X = I - X(X\prime X)^{-1}X\prime$ and $M_X = I - P_X$. Since Q is in the null space of X, $Q = M_X B$ for some B. Also, there exists a matrix A, of dimension $T \times (T - K)$, such that

$M_X = AA', A'A = I_{T-K}$ and $A'X = 0$. (See, e.g., Theil (1971, pp. 203-209).) So we can write

$$\begin{aligned}\Sigma &= X\Gamma X' + Q\Theta Q' + c^2(P_X + M_X)\\ &= X[\Gamma + c^2(X'X)^{-1}]X' + M_X B\Theta B' M_X + c^2 M_X\\ &= X[\Gamma + c^2(X'X)^{-1}]X' + A[c^2 I + A'B\Phi B'A]A'\end{aligned}$$

Since $A'X = 0$, condition (F') holds.

Proof of Theorem 2

The model is $y = X\beta + \varepsilon$, and $\tilde{\beta} = (X'\Sigma^{-1}X)^{-1}X'\Sigma^{-1}y$ is GLS assuming that the variance matrix of ε is Σ. Now define $X_* = \Omega^{-1/2}X$, $y_* = \Omega^{-1/2}y$ and $\Sigma_* = \Omega^{-1/2}\Sigma\Omega^{-1/2}$ (so $\Sigma_*^{-1} = \Omega^{1/2}\Sigma^{-1}\Omega^{1/2}$). Then $\ddot{\beta} = (X'\Omega^{-1}X)^{-1}X'\Omega^{-1}y$ is OLS of y_* on X_*, and $\tilde{\beta} = (X_*'\Sigma_*^{-1}X_*)^{-1}X_*'\Sigma_*^{-1}y_*$ is GLS of y_* on X_* using the error variance matrix Σ_*. So Amemiya's results apply if we simply replace his X by X_*, and Σ by Σ_*.

(A) $(X_*'X_*)^{-1}X_*'\Sigma_* X_*(X_*'X_*)^{-1} = (X_*'\Sigma_*^{-1}X_*)^{-1}$

$$\left(X'\Omega^{-1/2}\Omega^{-1/2}X\right)^{-1} X'\Omega^{-1/2}\cdot\Omega^{-1/2}\Sigma\Omega^{-1/2}\cdot\Omega^{-1/2}X\left(X'\Omega^{-1/2}\Omega^{-1/2}X\right)^{-1}$$
$$= \left(X'\Omega^{-1/2}\cdot\Omega^{1/2}\Sigma^{-1}\Omega^{1/2}\cdot\Omega^{-1/2}X\right)^{-1}$$

$$\left(X'\Omega^{-1}X\right)^{-1}X'\Omega^{-1}\Sigma\Omega^{-1}X\left(X'\Omega^{-1}X\right)^{-1} = \left(X'\Sigma^{-1}X\right)^{-1}$$

which is condition (A2) of Theorem 2.

(B) $\Sigma_* X_* = X_* B$ for some nonsingular B

$\Omega^{-1/2}\Sigma\Omega^{-1/2}\cdot\Omega^{-1/2}X = \Omega^{-1/2}XB$ for some nonsingular B

$\Sigma\Omega^{-1/2}X = XB$ for some nonsingular B

which is condition (B2) of theorem 2.

(C) $(X_*'X_*)^{-1}X_*' = (X_*'\Sigma_*^{-1}X_*)^{-1}X_*'\Sigma_*^{-1}$

$$\left(X'\Omega^{-1}X\right)^{-1}X'\Omega^{-1/2} = X'\Omega^{-1/2}\cdot\Omega^{1/2}\Sigma^{-1}\Omega^{1/2}\cdot\Omega^{-1/2}X$$

$$\left(X'\Omega^{-1}X\right)^{-1}X'\Omega^{-1} = \left(X'\Sigma^{-1}X\right)^{-1}X'\Sigma^{-1}$$

which is condition (C) of Theorem 2.

(D) $X_* = HA$ for some nonsingular A, where the columns of H are Keigenvectors of Σ_*.

$\Omega^{-1/2}X = HA$ for some nonsingular A, where the columns of H are Keigenvectors of $\Omega^{-1/2}\Sigma\Omega^{-1/2}$.

which is condition (D2) of Theorem 2.

(E) $X_*'\Sigma_*\,Z = 0$ for any Z such that $Z'X_* = 0$.

$X'\Omega^{-1/2}\cdot\Omega^{-1/2}\Sigma\Omega^{-1/2}\cdot Z = 0$ for any Z such that $Z'\Omega^{-1/2}X = 0$.

$X'\Omega^{-1}\Sigma\Omega^{-1/2}\cdot Z = 0$ for any Z such that $Z'\Omega^{-1/2}X = 0$.

Now define $Z_* = \Omega^{-1/2}Z$. Then this condition becomes $X'\Omega^{-1}\Sigma Z_*$ for any Z_* such that $Z_*'X = 0$, which is condition (E1) of Theorem 2.

(F) $\Sigma_* = X_*\Gamma X_*' + Q\Theta Q' + c^2I$ with $Q'X_* = 0$.

$\Omega^{-1/2}\Sigma\Omega^{-1/2} = \Omega^{-1/2}X\Gamma X'\Omega^{-1/2} + Q\Theta Q' + c^2I$ with $Q'\Omega^{-1/2}X = 0$.

$\Sigma = X\Gamma X' + \Omega^{1/2}Q\Theta Q'\Omega^{1/2} + c^2\Omega$ with $Q'\Omega^{-1/2}X = 0$.

Now let $Q_* = \Omega^{-1/2}Q$. Then we obtain

$\Sigma = X\Gamma X' + \Omega Q_*\Theta Q_*'\Omega + c^2\Omega$ with $Q_*'X = 0$

which is condition (F2) of Theorem 2.

(F') The proof is essentially the same as for condition (F).

Proof of Theorem 3

Proof that (A) and (C) are equivalent: Define

$$\Delta = \left[X'H\left(H'\Sigma H\right)^{-1}H'X\right]^{-1}X'H\left(H'\Sigma H\right)^{-1}H' - (X'\Sigma^{-1}X)^{-1}X'\Sigma^{-1}$$

then $\left[X'H\left(H'\Sigma H\right)^{-1}H'X\right]^{-1}-(X'\Sigma^{-1}X)^{-1}=\Delta\Sigma\Delta'$. So $\Delta=0$ (Which is condition (C)) is equivalent to $\left[X'H\left(H'\Sigma H\right)^{-1}H'X\right]^{-1}=(X'\Sigma^{-1}X)^{-1}$, which is condition (A).

Proof that (A) and (B) are equivalent:

$$X'\Sigma^{-1}X-X'H\left(H'\Sigma H\right)^{-1}H'X=X'\Sigma^{-1/2}[I-P_{[\Sigma^{1/2}H]}]\Sigma^{-1/2}X$$

and this equals zero if and only if $\Sigma^{-1/2}X$ is in the column space of $\Sigma^{1/2}H$, that is , $\Sigma^{-1/2}X=\Sigma^{1/2}\ HB$ for some B, or $\Sigma^{-1}X=HB$ or $X=\Sigma HB$, which is condition (B).

Proof that (B) and (E) are equivalent: The proof that (B) implies (E) is trivial. To show that (E) implies (B), suppose that $Z'H=0$. Then $Z=M_HS$ where $M_H=I-P_H=I-H(H'H)^{-1}H'$. Then (E) says that $Z'\Sigma^{-1}X=0$, or $S'M_H\Sigma^{-1}X=0$. This is true for any S, so it must be true that $M_H\Sigma^{-1}X=0$, that is $\Sigma^{-1}X$ is in the column space of H, or $\Sigma^{-1}X=HB$ for some B. This is condition (B).

Proof that (B) and (D) are equivalent: Note that

$$\Sigma H\left(H'\Sigma H\right)^{-1}H'\cdot\Sigma H=\Sigma H\bullet I$$

That is, the column of ΣH are eigenvectors of $\Sigma H\left(H'\Sigma H\right)^{-1}H'$, and the eigenvalues equal one. So, if (B) holds, $X=\Sigma HB$ and X is a linear combination of these eigenvectors, so (D) holds. Conversely, if (D) holds, then $X=(\Sigma H)A$ and (B) holds.

Proof that (F') implies (B): Suppose (F') holds, so that

$$\Sigma=X\Gamma X'+Q\Theta Q'$$

and Q satisfies $Q'HB=0$ for some B such that $X'HB$ is nonsingular. Then, for that B,

$$\Sigma HB=X\Gamma X'HB+Q\Theta Q'HB=X\Gamma X'HB$$

and so

$$X=\Sigma H[B(X'HB)^{-1}\Gamma^{-1}]$$

So condition (B) holds.

Proof that (B) implies (F'): Suppose (B) holds so that $X = \Sigma HB$. Let $\Gamma = (B\prime H\prime \Sigma HB)^{-1}$,

where the inverse must exist because X has rank K so HB must have rank K. Then

$\Sigma = X\Gamma X\prime + C$

where

$C = \Sigma - \Sigma HB(B\prime H\prime \Sigma HB)^{-1}B\prime H\prime \Sigma$

$= \Sigma^{1/2}[I - P_{[\Sigma^{1/2}HB]})]\Sigma^{1/2}$

$= QQ\prime$

where $Q = \Sigma^{1/2}[I - P_{[\Sigma^{1/2}HB]}]$.

Then $Q\prime HB = [I - P_{[\Sigma^{1/2}HB]}]\Sigma^{1/2}HB = 0$. So (F') holds.

BIBLIOGRAPHY

BIBLIOGRAPHY

[1] Amemiya, T. (1983), “Partially Generalized Least Squares and Two-Stage Least Squares Estimators,” Journal of Econometrics, 23, 275-283.

[2] Amemiya, T. (1985), Advanced Econometrics, Harvard University Press, Cambridge, MA.

[3] Anderson, T.W. (1971), The Statistical Analysis of Time Series, John Wiley and Sons, New York.

[4] Anselin, L. (2010), “Thirty Years of Spatial Econometrics,” Papers in Regional Science, 89, 3–25.

[5] Arbia, G. (2006), Spatial Econometrics: Statistical Foundations and Applications to Regional Convergence . Springer.

[6] Baltagi, B.H. (1989), “Applications of a Necessary and Sufficient Condition for OLS to be BLUE,” Statistics and Probability Letters, 8, 457-461.

[7] Balestra, P. (1970), “On the Efficiency of Ordinary Least Squares in Regression Analysis,” Journal of the American Statistical Association, 65, 1330-1337.

[8] Bester, A.C., Conley, T.G., Hansen, C.B. and Vogelsang, T.J. (2008), “Fixed-b Asymptotics for Spatially Dependent Robust Nonparametric Covariance Matrix Estimators.”

[9] Bloom,N., Schankerman, M. and VanReenen, J. (2012), “Identifying technology spillovers and product market rivalry,” NBER working paper.

[10] Bollerslev, T. and Wooldridge, J.M. (1992), “Quasi-Maximum Likelihood Estimation and Inference in Dynamic Models with Time-varying Covariances,” Econometric Reviews, 11(2), 143-172.

[11] Bolthausen, E. (1982), “On the Central Limit Theorem for Stationary Mixing Random Fields,” The Annals of Probability, Vol. 10, No. 4, 1047-1050.

[12] Brown, J.R., Ivkovic, Z., Smith, P.A. and Wwisbenner, S. (2008), “Neighbors Matter: Causal Community Effects and Stock Market Participation,” The Journal of Finance, 63: 1509–1531.

[13] Christakos, G. (1987), “On the Problem of Permissible Covariance and Variogram Models,” Water Resources Research, 20, 251-265.

[14] Cragg, J. (1983), “More Efficient Estimation in the Presence of Heteroscedasticity of Unknown Form,” Econometrica, 51, 751-763.

[15] Conley,T.G. (1999), “GMM Estimation with Cross Sectional Dependence,” Journal of Econometrics, Volume 92, Issue 1, 1-45.

[16] Cressie, N.A.C. (1993), Statistics for Spatial Data, Wiley Series in Probability and Mathematical Statistics: Applied Probability and Statistics, New York: John Wiley & Sons Inc., a Wiley-Interscience Publication.

[17] Du, J. and Zhang, H. (2007), “Covariance Tapering in Spatial Statistics,” Working paper.

[18] Dubin, R.A. (1988), “Estimation of Regression Coefficients in the Presence of Spatially Autocorrelated Error Terms,” The Review of Economics and Statistics , Vol. 70, No. 3, 466-474.

[19] Fiebig, D.G., Bartels, R. and Krämer, W. (1996), “The Frisch-Waugh Theorem and Generalized Least Squares,” Econometric Reviews, 15, 431-443.

[20] Gneiting, T. (2002), “Compactly Supported Correlation Functions,” Journal of Multivariate Analysis, 83, 493-508.

[21] Gourieroux, C. and Monfort A. (1980), “Sufficient Linear Structures: Econometric Applications,” Econometrica, 48, 1083-1097.

[22] Gourieroux, C., Monfort, A. and Trognon, A. (1984), “Pseudo Maximum Likelihood Methods: Applications to Poisson Models,” Econometrica , Vol. 52, No. 3, pp. 701-720

[23] Gross, J. and Puntanen, S. (2000), “Remark on Pseudo-Generalized Least Squares,” Econometric Reviews, 19, 131-134.

[24] Kelejian, H.H. and Prucha I.R. (2007), “HAC estimation in a spatial framework,” Journal of Econometrics, Volume 140, Issue 1, September 2007, Pages 131-154

[25] Kruskal, G. (1968), “When Are Gauss-Markov and Least Squares Estimators Identical? A Coordinate-Free Approach,” Annals of Mathematical Statistics, 39, 70-75.

[26] 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: 86–98.

[27] Kaufman, C., Schervish, M. and Nychka, D. (2008), “Covariance tapering for likelihood-based estimation in large spatial datasets,” J. Amer. Statist. Assoc. 103 1545–1555.

[28] Lee, L.F.(2004), “Asymptotic Distributions of Quasi-Maximum Likelihood Estimators for Spatial Autoregressive Models,” Econometrica, Vol. 72, 1899-1925.

[29] Matthew, T. (1983), “Linear Estimation with an Incorrect Dispersion Matrix in Linear Models with a Common Linear Part,” Journal of the American Statistical Association, 78, 468-471.

[30] McAleer, M. (1992), “Efficient Estimation: The Rao-Zyskind Condition, Kruskal’s Theorem and Ordinary Least Squares,” Economic Record, 68, 65-72.

[31] McElroy, F.W. (1967), “A Necessary and Sufficient Condition that Ordinary Least Squares Estimators Be Best Linear Unbiased,” Journal of the American Statistical Association, 63, 1302-1304.

[32] Newey, W.K. and West K.D., 1987. “A simple, positive semi-definite, heteroskedasticity and autocorrelation consistent covariance matrix,” Econometrica, Vol.55, No.3, 703-708.

[33] Prentice, R.L. (1988), “Correlated binary regression with covariates specific to each binary observation,” Biometrics AA, 1033-48.

[34] Puntanen, S. and Styan, G.P.H. “The Equality of the Ordinary Least Squares Estimator and the Best Linear Unbiased Estimator,” The American Statistician, 43, 153-161.

[35] Rao, C.R. (1967), “Least Squares Theory Using an Estimated Dispersion Matrix and Its Application to Measurement of Signals,” Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability, vol. 1, 355-372, University of California Press, Berkeley, CA.

[36] Theil, H. (1971), Principles of Econometrics, John Wiley and Sons, New York.

[37] Wang, H., Iglesias, E. and Wooldrige, J.M. (2012). “Partial Maximum Likelihood Estimation of Spatial Probit Models,” Journal of Econonometrics.

[38] Wooldridge, J.M. (2011), Econometric Analysis of Cross Section and Panel Data. MIT.

[39] Zeger, S.L., Liang, K.Y., and Albert, P.S. (1988), “Models for Longitudinal Data: A Generalized Estimating Equation Approach,” Biometrics, 44, 1049- 1060.

[40] Zyskind, G. (1962), “On Conditions for Equality of Best and Simple Linear Least Squares Estimators (abstract),” Annals of Mathematical Statistics, 33, 1502-1503.

[41] Zyskind, G. (1967), “On Canonical Forms, Non-Negative Covariance Matrices and Best and Simple Least Squares Linear Estimators in Linear Models,” Annals of Mathematical Statistics, 38, 1092-1109.