
141
579
THS

ESIS
2
2009

LIBRARY
Michigan State
University

This is to certify that the
dissertation entitled

SOCIAL LEARNING AND PARAMETER UNCERTAINTY IN IRREVERSIBLE INVESTMENTS

AND

PARTIAL MAXIMUM LIKELIHOOD ESTIMATION OF A SPATIAL PROBIT MODEL

presented by

HONGLIN WANG

has been accepted towards fulfillment
of the requirements for the

Ph.D.	degree in	Agricultural Economics and Economics

Major Professor's Signature

May 11, 2009

Date

MSU is an Affirmative Action/Equal Opportunity Employer

PLACE IN RETURN BOX to remove this checkout from your record.
TO AVOID FINES return on or before date due.
MAY BE RECALLED with earlier due date if requested.

DATE DUE	DATE DUE	DATE DUE

5/08 K:/Proj/Acc&Pres/CIRC/DateDue.indd

SOCIAL LEARNING AND PARAMETER UNCERTAINTY IN IRREVERSIBLE INVESTMENTS AND PARTIAL MAXIMUM LIKELIHOOD ESTIMATION OF A SPATIAL PROBIT MODEL

BY

HONGLIN WANG

A DISSERTATION

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

DOCTOR OF PHILOSOPHY

Agricultural Economics and Economics

2009

ABSTACT

SOCIAL LEARNING AND PARAMETER UNCERTAINTY IN IRREVERSIBLE INVESTMENTS AND PARTIAL MAXIMUM LIKELIHOOD ESTIMATION OF A SPATIAL PROBIT MODEL

BY

HONGLIN WANG

The dissertation is composed of two essays.

The first paper discusses the social leaning and parameter uncertainty in irreversible investments. The adoption of new technology usually involves irreversible investments where the future payoff is uncertain. In addition, investors often have to contend with a limited understanding of the technology itself, which can be modeled as uncertainty regarding the parameters of the stochastic process describing the future payoff. It is hypothesize that social learning (having previous adopters in the farmer's social network) increases the probability of the farmer adopting the new technology. This is posited based on theory: social learning would reduce parameter uncertainty, and thus the overall level of risk facing the farmer-investor, and thus induce investment. The paper tests this hypothesis using Chinese farm household data on adoption of greenhouses. The latter are of the "intermediate technology" type, made of clay walls, a plastic-sheet roof, and a straw mat roll-out awning for cold nights. The empirical findings of this paper support the hypothesis. It is also found that market volatility discourages adoption.

The second paper analyzes a spatial Probit model for cross sectional dependent data

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

ACKNOWLEDGMENT

I owe my gratitude to all those people who have made this dissertation possible and because of whom my graduate experience has been one that I will cherish forever.

My deepest gratitude is to my major advisor in agricultural economics, Dr. Thomas Reardon, and my major advisor in economics, Dr. Jeffrey Wooldridge. I would like to sincerely thank Dr. Reardon for his guidance, understanding, and strong support whenever and for whatever I need. He encouraged me to pursue dual degree program, which provides me a wider view of the long-term career path. His extensive international experience, enthusiasms on global development, encouragement and his intelligent visions on human society and culture exhibit me an exciting world for further exploring. I have been amazingly fortunate to have Dr. Wooldridge as my major professor in economics. His wisdom, effective guidance and great teaching, patience and constructive comments help me enrich but focus my ideas at different stages. I am grateful to him for holding me to a high research standard. Inspired by Dr. Wooldridge, learning econometrics becomes a very enjoyable part of my life. I hope one day I could become a great teacher like Dr. Wooldridge to my students.

I would like to extend my special thanks to my committee members, Dr. Songqing Jin, Dr. Emma Iglesias and Dr. John Giles. They always spent a lot of time to discuss with me and give very helpful advice. I would also express my sincere thanks to Dr. Fan Yu. His insightful critics, guidance, and careful reviews on my paper help me overcome difficulties

and finish my research. My special thanks also go to Dr. Robert Myers and Dr. Zhengfei Guan, who always offer me their help when I have difficulties in my research.

Grateful and sincere thanks also go to Dr. Jikun Huang, Dr. Scott Rozelle and Dr. Linxiu Zhang. I am indebted to them for their guidance on designing survey questionnaire, support on field surveys and the trip in China, and comments on the research paper. They have been advising me since my pursuit of M.S. in China, and I am thankful for their continuous encouragement, support and inspiration.

I am grateful to my graduate colleagues Jinxia Wang, Chengfang Liu, Xiaoxia Dong, Zijun Wang, Haiqing Zhang and Ruijian Chen in China, who worked very hard in assisting surveys, data collection, validation and data entry. I am grateful for their hard work, sharing the thoughts, and friendships. I would like express gratitude to my graduate fellows Wei Zhang, Yanyan Liu, Zhiying Xu, Feng Song, Fang Xie, Feng Wu, Lili Gao, Wolfgang Pejuan, Kirimi Sindi, Vandana Yadav, Ricardo Hernandez, Kang-Hung Chang and Panutat Satchachai in both department of agricultural economics, and department of economics, for the valuable discussion, 'happy hour', and care from them. I highly value such friendships. They make my staying in the MSU a pleasant and unforgettable experience.

Finally, and most importantly, I would like thank my wife Qing Xiang. None of this would have been possible without the love and patience of my wife. My wife and my parents, Ruixia Chen and Jinyu Wang, have been a constant source of love, concern, encouragement and strength all these years.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

Chapter 1: Social Learning and Parameter Uncertainty in Irreversible Investments: Evidence from Greenhouse Adoption in Northern China

1.1 Introduction

Risk and uncertainty have been important themes in the agricultural technology adoption literature since the 1970s. They were included in studies of green revolution technology adoption to explain lagged or partial adoption or even disadoption. Examples include Roumasset (1976) and Feder (1980). This can be seen as part of a wider strand of literature on the economics of risk and uncertainty, and their constraining effects on investment (Newbery and Stiglitz, 1981).

Distinctions in two dimensions in particular that interest us here have been drawn from the initial foundation of inclusion of risk and uncertainty in agricultural technology adoption analysis. The first dimension is the modeling of various forms of "information capital" as part of the vector of capital assets in the adoption function. The earliest forms modeled were public information in the form of farmers' education and access to extension services. Then, and of most interest to us here, came the introduction of personal experience with a technology ("learning by doing") and observation of neighbors' experience with the technology ("learning from neighbors"). These were introduced for example in Besley and Case (1994) and Foster and Rosenzweig (1995).

The modeling of "learning from neighbors" has been further refined in recent papers that model "social learning," such as: (1) Conley and Udry (2001) in their modeling of Ghana farmers' adoption of fertilizer in pineapple production, conditioned by their incomplete information and communication networks with neighbors; (2) Bandiera and

Rasul (2006) in their modeling of Mozambique farmers' adoption of sunflowers, conditioned by their social network (neighbors and friends who have adopted); and (3) Munshi (2004) in his modeling of Indian farmers' adoption of HYV of rice and wheat, conditioned by their neighbors' experiences but differentiated over rice and wheat areas due to the influence of heterogeneous population. This body of work has demonstrated the effects of social learning on technology adoption. In most cases the social learning's effect on adoption is interpreted as increasing the capacity of the farmer to adopt as well as reducing the farmer's uncertainty and perception of risk in adoption.

The second dimension is the modeling of irreversible investments in capital embodying technology, such as tube wells, greenhouses, and so on. This distinction - between reversible investments such as adoption of an annual crop, a hybrid seed, fertilizer, or a new planting technique - and irreversible investments where the salvage value of the asset is negligible or the asset cannot be transferred or sold, is important in the analysis of risk and uncertainty in technology adoption.

Because of incomplete information with respect to the performance, reliability, and appropriateness of agricultural equipment, irreversibility entails substantial risk for the investor (Dixit and Pindyck, 1994, and Sunding and Zilberman, 2000). McDonald and Siegel (1986) and Dixit and Pindyck (1994) show that the ability to delay an irreversible investment can be considered as a real option; a higher level of uncertainty regarding future benefits raises the option value and causes the investment decision to deviate from the classical NPV rule. Specifically, investors may rationally delay investment to gain additional information, reduce the level of uncertainty, and increase discounted expected payoffs. This has been modeled in two strands of literature.

On the one hand, delayed investment to gain additional information in the face of uncertainty has been studied in the economics literature, inspired by McDonald and Seigel and Dixit and Pindyck. Examples include Olmstead and Rhode (1993), Zilberman et al. (2004), Hassett and Metcalf (1995), and Nelson and Amegbeto (1998), inter alia. These studies have tended to assume that all parameters of the dynamic process are known to agents, and the only uncertainty in the model comes from the future value of the dynamic process.

On the other hand, investment under parameter uncertainty has been examined in the finance literature. Merton (1980) shows that while the variance of the return can be estimated precisely from continuous observations on a finite interval, the estimator of mean return does not converge unless the length of the interval becomes large. Gennotte (1986) studies portfolio choice under incomplete information about the stock return process. He uses tools of nonlinear filtering from Lipster and Shiryaev (1978) to derive the optimal drift estimator as agents continuously observe the returns. Brennan (1998) and Xia (2001) construct similar models to examine how learning about unknown parameters and unknown predictability affects portfolio choice. More recently, Abasov (2005) modeled irreversible investment under parameter uncertainty, and Huang and Liu (2007) modeled learning from discrete noisy signals about the true drift in their study of periodic news on portfolio selection. Note that much of the finance literature is primarily theoretical, with few empirical applications and none in the domain of investment in agriculture capital as an embodiment of agricultural technology adoption.

The present paper aims at a particular, and a particularly important, gap left by the two dimensions discussed. That is, while the literature on social learning and technology

adoption has modeled the effect of social learning as a means of reducing uncertainty, that literature has not treated the issue of irreversibility of the investment per se, and thus has not modeled the effect of social learning in a real options context. Moreover, while the literature on irreversible investment and uncertainty has indeed modeled investment in a real options framework, it has not examined uncertainty-reduction measures taken by adopters, in particular, social learning.

There is thus a gap in the literature, both theoretical and empirical, where an analysis of irreversible investment under parameter uncertainty models the effect of social learning. The contribution of the present paper is to address that gap.

We address the gap empirically by modeling greenhouse investments with primary data from Shandong province in China. The data are multi-year, observing the characteristics, including their social network of prior adopters, of the adopters the year before their adoption, and thus, new to this literature, we capture causality of social learning and adoption.

We address it theoretically, by presenting a new model to the literature of these links. Following McDonald and Siegel (1986), we assume that a farmer is considering an investment project, whose value follows a geometric Brownian motion. Departing from the standard framework, we assume that the true drift of the Brownian motion is unobservable to the farmer (we call this parameter uncertainty). In essence, the farmer is imperfectly informed as to the expected rate of return of his investment. He must make an inference about the true expected return based on his information and, at the same time, determine the optimal timing for investing in the project. The farmer can learn about the unknown parameter in two ways. First, he extracts information on the true drift from a continuous

observation of past realized returns on the project value. This captures the process of continuous learning from public information about the project. Second, he obtains discrete noisy signals of the true drift. This represents the process of social learning from early adopters in his social network, who might possess information about the project that the public do not have. In our model, parameter uncertainty adds to the overall risk that the farmer faces; this raises the threshold project value needed to induce the farmer to invest. In contrast, social learning reduces parameter uncertainty, which decreases the overall level of uncertainty and reduces the investment threshold, thereby increasing the likelihood of adoption. In our model, social learning also causes the farmer's belief about the expected return to converge to the average belief of his social network; the higher the average belief, the higher is the investment threshold, and the less likely the farmer will adopt the technology.

The rest of the paper is organized as follows: In Section 2, we present the theoretical model. In Section 3, we provide background information about the greenhouse technology in northern China. In Section 4, we outline our sample selection and summarize the data. In Section 5, we explain our empirical methodology. In Section 6, we present the empirical findings using linear probability models. We conclude in Section 7.

1.2 The Theoretical Model Framework

In this section, we use a real options model to articulate the effect of parameter uncertainty and social learning on technology adoption. We begin with a model of continuous learning, which is essentially that of Abasov (2005). Specifically, a farmer is considering whether to pay a sunk cost of I for an agricultural technology, whose value V evolves according to:

$$dV_t = V_t (\mu d_t + \sigma dZ_t)$$

where Z is a Brownian motion.

Motivated by Merton (1980), we assume that the farmer can observe V continuously and knows its volatility σ; however, he only knows that the drift μ is a normal random variable with mean m_0 and variance γ_0 in the beginning. According to Lipster and Shiryaev (1978), the conditional mean of the drift given the farmer's information set, $m_t = E\left(\mu \mid \mathbf{F}_t^V \right)$, follows:

$$dm_t = \frac{\gamma t}{\sigma} dZ_t^{'}$$

where $\gamma_t = E\left[(\mu - m_t)^2 \mid \mathbf{F}_t^V \right]$ is the conditional variance of the drift, satisfying:

$$d\gamma_t = -\frac{\gamma_t^2}{\sigma^2} dt \qquad (1.1)$$

and $Z^{'}$ is a new Brownian motion related to the original Brownian motion through:

$$dZ_t^{'} = dZ_t^{'} + \frac{m_t - \mu}{\sigma} dt$$

We can solve equation 1.1 for γ_t:

$$\gamma_t = \frac{\gamma_0 \sigma^2}{\gamma_0 t + \sigma^2}.$$

This result shows that continuous learning decreases the conditional variance of the unknown parameter. Thus the longer the farmer observes the value process, the less uncertain he is about the drift. This is consistent with Merton (1980)'s results: the uncertainty of the drift is not related to the number of observations, but is rather related to the length of the observation period. However, the conditional mean of the drift can fluctuate up or down, depending on new observations of the Brownian motion Z^t.

According to Gennotte (1986), the farmer's decision can be separated into two problems: the inference of the unknown parameter given $\{Z_s'\}_{0 \le s \le t}$, and the optimal stopping decision based on the current state variables (m_t, γ_t, V_t) and the dynamics of (m, γ, V). Putting everything together, we can characterize the farmer's problem using observable processes:

$$\begin{aligned} J(m_0, \gamma_0, V_0) &= \max_{\tau \in \mathbf{F}^V} E\left[e^{-\rho \tau} (V_\tau - I) \right], \\ s.t.\ dV_t &= V_t \left(m_t dt + \sigma dZ_t' \right), \\ dm_t &= \frac{\gamma_t}{\sigma} dZ_t', \\ d\gamma_t &= -\frac{\gamma_t^2}{\sigma^2} dt. \end{aligned} \tag{1.2}$$

Here, ρ is the farmer's discount rate, and τ has to be an $\mathbf{F}^{V'}$-stopping time, reflecting that the farmer must make a decision based on his information set. The stopping rule takes

the form of:

$$\tau = \inf\left\{t \geq 0 \ : \ V_t \geq V^*(m_t, \gamma_t)\right\},$$

where $V^*(m,\gamma)$ is the trigger value of investing, which depends on the state variables.[1]

Abasov (2005) derives the Hamilton-Jacobi-Bellman equation for the optimal stopping problem (1.2) and transforms it into a linear complementarity problem, which he solves with the finite difference method. His numerical results demonstrate that the trigger value of investing, $V^*(m_0,\gamma_0)$, obtained as a part of the solution, increases with γ_0. This result is sensible given that the trigger value in the McDonald and Siegel (1986) and Dixit and Pindyck (1994) model increases with σ, and parameter uncertainty contributes to the total uncertainty in our model. In addition, Abasov shows that V^* increases with m_0; this is also consistent with the traditional real options model without parameter uncertainty.

In developed countries, there are public economic forecasts and newsletters informing investors. Therefore, agents can make inferences based on past realized returns. However, in rural China, information is more likely to come from local private sources. Similar to Huang and Liu (2007), we allow farmers to obtain direct signals of the drift from early adopters in their social networks. These signals are noisy, reflecting the fact that even early adopters are unlikely to learn everything about the technology from their own experience. Different from Huang and Liu (2007), we assume that the signals are costless. However, the number of signals to which a farmer has access is limited by the scope of his social

[1] Since γ is a deterministic function of t, we can equivalently formulate the problem in terms of state variables (m,t).

network, which we take as exogenous. For simplicity, we also assume that these signals are received at time 0, just as the farmer begins to consider his adoption decision. Since discrete signals are much more effective than continuous learning in changing the farmer's belief, it seems reasonable to assume that he would seek out these signals at the very beginning of his decision-making process. This implies that discrete updating affects the farmer's optimal stopping problem only insofar as it changes his initial belief; discrete updating plays no role in the dynamics of the conditional mean and conditional volatility.

Let signal i be given by:

$$\mu_i = \mu + \varepsilon_i, \tag{1.3}$$

where $\varepsilon_i \sim N\left(0, \sigma_\varepsilon^2\right)$ is independently and identically distributed. After receiving n such signals, it can be shown that the conditional mean and variance of the drift are given by:

$$m_0' = \frac{\sigma_\varepsilon^2}{n\gamma_0 + \sigma_\varepsilon^2} m_0 + \frac{n\gamma_0}{n\gamma_0 + \sigma_\varepsilon^2} \bar{\mu}, \tag{1.4}$$

$$\gamma_0' = \frac{\gamma_0 \sigma_\varepsilon^2}{n\gamma_0 + \sigma_\varepsilon^2}, \tag{1.5}$$

where $\bar{\mu} = \frac{1}{n} \Sigma_{i=1}^{n} \mu_i$. Equation (1.4) shows that the conditional variance is decreasing in the number of signals, which can be taken as the scope of social learning. Therefore, social learning reduces parameter uncertainty. Using Abasov's numerical results, this implies that social learning decreases the trigger value for adoption, making it more likely that the farmer would adopt the technology.

Considering the conditional mean equation (1.3), we find that as the number of signals increases, m_0' tends to move away from m_0 and approach $\bar{\mu}$. This indicates that social learning causes the farmer's belief about the drift to converge to the average belief in the farmer's social network. The net effect depends on the relation between m_0 and $\bar{\mu}$. If $m_0 > \bar{\mu}$, the farmer is initially too optimistic; social learning causes him to lower his expectation about the project's return. This, in turn, lowers the trigger value and facilitates adoption. If the farmer is, on average, unbiased in his initial belief, then social learning is unlikely to change the probability of adoption through its effect on the conditional mean return.

If we generalize this model to allow the dynamics of social learning to enter the farmer's decision making, then we can write down the following optimal stopping problem, where we combine continuous filtering with discrete updating:

$$
\begin{aligned}
J(m_0,\gamma_0,V_0) &= \max_{\tau\in\mathbf{F}^V\vee\mathbf{F}^N} E\left[e^{-\rho\tau}(V_\tau - I)\right], \\
s.t.\ dV_t &= V_t\left(m_t dt + \sigma dZ_t'\right), \\
dm_t &= \frac{\gamma_t}{\sigma} dZ_t' + \frac{\gamma_{t-}}{\gamma_{t-}+\sigma_\varepsilon^2}(\mu(t) - m_{t-})dN_t, \\
d\gamma_t &= -\frac{\gamma_t^2}{\sigma^2}dt - \frac{\gamma_{t-}^2}{\gamma_{t-}+\sigma_\varepsilon^2}dN_t.
\end{aligned}
\tag{1.6}
$$

Here, $\mu(t)$ refers to the independently and identically distributed noisy signals described in equation (1.5), and N_t is a counting process that counts the number of signals that the farmer has received up to time t. It can be periodic and deterministic as in Huang

and Liu (2007), or stochastic, as in the case of a Poisson process with arrival rate λ, which describes social interaction as a random phenomenon. In all cases, however, the first part of the dynamic equations for (m, γ) captures the effect of continuous updating as the farmer learns from the past history of V. The second part represents a jump in the conditional mean and variance when the farmer receives a noisy signal of the drift. Because γ and N are deterministically related through the conditional variance relation, we have suppressed the dependence of the value function on N. Similarly, we can write the trigger value as $V^*(m_t, \gamma_t)$, with the understanding that the effect of N_t is already reflected in the conditional variance γ_t.

Generally, the optimal stopping problem (1.6) must be solved numerically. The adoption decision is related to the amount of social learning that the farmer has experienced. According to the above model, this is measured by N_t. As the conditional variance equation shows, a larger N (more social learning) always reduces γ. We conjecture that the trigger value is increasing in γ, regardless of whether farmers are cognizant or ignorant of future social learning.[2] This implies that social learning can lower the trigger level for adoption.

Summarizing the various models, the classical real options analysis of McDonald and Siegel (1986) predicts that the trigger value for investment increases with the uncertainty of the project value. We show that this result also extends to parameter uncertainty. Building from recent work on social learning and technology diffusion (such as Bandiera and Rasul, 2006), we argue that social learning can facilitate adoption by reducing parameter

[2] One can conceive of cases in which knowledge of the social learning dynamics can actually delay adoption. For example, if the farmer knows that parameter uncertainty will be fully resolved tomorrow, he is unlikely to invest today.

uncertainty. In rural China, where public extension information is not easily accessible to small farmers, information from social learning could play an important role in their adoption decisions. The rest of our paper is dedicated to testing this hypothesis.

1.3 Greenhouse Intermediate-Technology in Northern China

Before economic reforms, China gave first priority to the development of heavy industry. In agriculture, China emphasized the importance of self-sufficiency for grains - the "iron rice bowl policy." After the "household responsibility system" reform started in 1981, the shortage of grain supply was relieved by a significant increase in grain production. This made it possible for China to diversify into horticulture and livestock husbandry. Meanwhile, rapid income growth in the 1980s and 1990s created an increasing demand for high-value horticultural products. However, poor infrastructure and high energy costs prevented the transportation of perishable products from southern China to northern China, and affordable fresh vegetables were still unavailable in the 1990s to consumers during the winter season in northern China.

The huge demand for cheap fresh vegetables led to the development and widespread diffusion of an affordable greenhouse technology for northern Chinese farmers. Rather than the modern, expensive type made of steel frame, plastic or glass walls and ceilings, and requiring energy-using heating and cooling mechanisms (promoted in the 1970s in China but saw very little adoption because of the cost, Wan 2000), the greenhouse adopted in the 1990s in northern China was of the "intermediate technology" type, made of simple clay walls, bamboo frame, a plastic-sheet roof, and a straw mat roll-out awning for cold nights. The sun warms the interior, with the greenhouse built with an orientation to maximize sunlight capture. These greenhouses changed not only the food consumption pattern for hundreds of millions of consumers, but also the face of farming in northern China. These greenhouses helped to transform China from a modest global player to the

volume leader in horticulture - growing one third of the fruits and vegetables on the planet by 2003. By 2004, China grew 47 percent of the vegetable volume in the world (Weinberger and Lumpkin 2005). The vegetable greenhouse area in China reached 150,000 hectares in 2004 (Chinese Agriculture Yearbook 2006), and at least half a million farmers were by that year using the intermediate-technology greenhouse.

Greenhouse yields exceed open-field cropping: for example, the tomato yield is 200 tons/hectare/year in the greenhouse, versus 40 tons in an open field. Several factors, including labor intensive production, contribute to this high yield. For example, the popular greenhouse size in Shandong province is only about 60 meters long and 10 meters wide, but it usually employs two full-time workers. Greenhouse production usually lasts more than eight months, because the temperature inside the greenhouse is high enough during the winter months to sustain production. Moreover, high quality crop varieties and intensive use of organic fertilizers are common in greenhouse production. Nutrient replacement is important due to the intensive and continuous use of the land under the greenhouse.

The intermediate-technology greenhouse is far cheaper than a modern type, but is still a major investment for the very small farmers of Shandong. The construction cost of intermediate-technology greenhouses is roughly four dollars per square meter, much cheaper than modern greenhouses of glass or plastic which cost about 80 dollars per square meter to construct. Yet even four dollars per square meter is a large investment for very small farmers. For example, if a greenhouse is 60 meters long and 10 meters wide, the construction cost would be about $2,400, while the average Chinese farmer earned less than $500 in 2005. Moreover, the labor time involved in building the greenhouse is substantial: the farmer spends months creating the main component - the rear-wall of the

greenhouse, which is usually made of pounded clay bricks.

Moreover, the investment is "irreversible," in the sense of Bertola and Caballero (1994), as the structure can only be used in immediate production, and has little to no salvage value and cannot be sold or transferred. The bricks cannot be reused or sold; if the farmer decides to demolish the greenhouse (as it cannot be transferred or sold as it is not movable), the bricks would be broken into dirt clods, and the old straw awning and old bamboo beams worth little in salvage.

1.4 Data Description

1.4.1 Sample Selection

Our survey area is in Shandong province, the leading horticulture province in China. It has seven percent of China's cropland, but 12 percent of China's horticultural land in 2004. The latter share has been steadily rising over time. The number of greenhouses and the level of commercialization as well as yields in Shandong are higher than in the rest of China.

In Shandong, we conducted two coordinated community and household level surveys in 2005 and 2006, respectively. The first one, the Shandong village survey, provided a representative sample of tomato and cucumber growing villages in Shandong. During the first step of the survey, we created sampling frames of county-level tomato and cucumber production in order to select five sample counties per crop. Specifically, with knowledge of county production of each crop, we ranked counties by the output per capita of that crop. For each crop in our sample, one high production county was randomly selected from the counties in the top quintile; the other high production county was randomly selected from the second quintile. The two medium production counties were randomly chosen from the third and fourth quintiles, respectively. After eliminating five percent of the counties with the lowest production, the low production county was randomly chosen from the lowest quintile. In the end, there were two counties in the high production set, two counties in the medium production set, and one county in the low production set.

After the sample counties were chosen, a similar process was used to select sample townships and villages. For each crop, the survey teams visited a total of ten townships.

Moreover, for each crop (among the five counties and ten townships), we interviewed respondents in 35 villages (22 in high production counties, 10 in medium production counties, and 3 in low production counties). Since we collected area data on all villages, townships, and counties in the sample, we were able to construct area-based weights in order to create point estimates of our variables that are provincially representative.

Having selected the villages, the enumeration team visited each community and undertook data collection. Specifically, the enumerator conducted a two-hour interview with three village leaders for the village survey. In each village, we divided all households into two groups. For the cucumber sample, they are non-cucumber households and cucumber households. We randomly sampled seven cucumber farmers and three non-cucumber farmers. As a result, we obtained 350 households from cucumber growing villages.[3] With knowledge of the distribution of cucumber farmers and non-cucumber farmers, plus the distribution of greenhouse adopters in each village, we calculated the weights to adjust for selection bias. Following this procedure, we also obtained 350 households from tomato growing villages.

After data cleaning, we collected 638 valid household observations. Among this sample, 204 (64 percent) out of 317 households from tomato growing villages were found to have adopted greenhouses, while 158 (49 percent) out of 321 households from cucumber growing villages were found to have adopted greenhouses. That a higher share of tomato growers adopted greenhouses is apparently due to the fact that in cucumber production, a shading shed is a substitute for a greenhouse, while in tomato production there is no

[3] The reason why we did not directly stratify on greenhouse use is that our survey is part of a large horticulture production survey, which required stratified sampling of cucumber/tomato and non-cucumber/tomato households.

substitute for a greenhouse, and the options are only growing in the open field or in a greenhouse.

Shandong farmers did not adopt greenhouses all at once, but rather, in a process typical of diffusion of new technology, over years. The greenhouse diffusion process can be roughly divided into three stages: early stage, take-off stage, and slow-down stage. Figure 1.1 shows that the diffusion process is relatively slow in the early stage before 1990; only a few farmers adopted the technology. Between 1990 and 1995, many more farmers adopted. The diffusion process reached its peak between 1996 and 2000, after which the trend began to slow down. This diffusion curve is similar to the "s-curve" observed by Griliches (1957) for the adoption of hybrid maize in the US, and subsequently documented in many other settings.

1.4.2 Social Learning

We are interested in the effect of social learning on farmers' adoption of greenhouses. Our theoretical model predicts that social learning helps to reduce parameter uncertainty, thus facilitating adoption. Empirically, however, social learning could be one of many factors affecting adoption. For example, farmers may have other options such as off-farm jobs. Alternatively, farmers may be credit-constrained because greenhouse adoption is a major investment. To disentangle the effect of social learning from other determinants, we need to find appropriate empirical proxies for social learning and control for other factors that might influence farmers' decisions.

Social learning is a key variable in our study. We measure social learning in a way similar to the approach of Bandiera and Rasul (2006). Specifically, we asked the farmers

who adopted, "How many people do you know who adopted greenhouses *before you adopted* in your village?" We asked the non-adopters how many adopters they knew at the time of the survey. We control for year with year dummy variables. We then asked, "How many of these people are your relatives and friends?" (We did not include neighbors as a separate category because Chinese farmers usually consider neighbors among friends.) The answer to the second question is taken as our empirical proxy for social learning. Differing from Bandiera and Rasul (2006) (who asked about the social network at the time of the survey, not before adoption), we obtained the size of the farmer's social network of adopters *before* his adoption, so that we can infer causality.

There are several reasons why our measure of social network of adopters is an appropriate measure of social learning before adoption. First, the number of earlier adopters among relatives and friends is likely to be positively correlated with the number of different sources of information on greenhouse adoption that the farmer accessed before adoption, which corresponds to the number of discrete signals in our theoretical model. Second, village membership, kinship, and friends are the defining elements of a farmer's social network, or a group of people with whom the farmer has close contact, and from whom information can be most easily obtained. By concentrating on the number of earlier adopters among relatives and friends, we also mitigate the concern for *ex post* social network formation. While this is obvious for kin adopters, we noticed during our survey that Shandong farmers tended to define friendship based on long-term relation, such as classmates, neighbors, and people who served with them in the army. Typically, they consider a friend someone from whom they can borrow money in case of illness; they would not consider passing acquaintances as friends. Third, we found that farmers were

easily able to remember the number of adopters they knew before they adopted; we surmise that this is because a greenhouse is a big investment for local farmers and hence easily observable.

The first two rows of Table 1.1 provide the means and standard errors of our social learning measures by adoption status. In the last column, tests of equality of the means are provided to examine whether the differences between adopters and non-adopters are significant. The first row indicates that, on average, adopters know about 6.9 earlier adopters among relatives and friends in their own village, while non-adopters only know about 4.7 earlier adopters in their social network. The result of the *t*-test shows that this difference is significant. This implies that there is more social learning for adopters than for non-adopters. When we extend the scope of the social network to include earlier adopters among relatives and friends in nearby villages (the second row), the findings are similar.

1.4.3 Other Household Characteristics

Table 1.1 presents other household characteristics by adoption status. There are several salient points.

(1) Demographics differ between adopters and non-adopters. The family size of adopters is significant larger than that of non-adopters, while the amount of farm labor is significant smaller for adopters than for non-adopters. This is because adopters have more dependent family members (either young children or old parents) than non-adopters. For such households, greenhouse adoption could be a good choice because it allows the adults to work close to home, so that they can care for dependent family members. Non-adopters are, on average, substantially older than non-adopters - a point consistent with younger

farmers having more young children and old parents to care for.

(2) Off-farm employment and income are significantly larger for non-adopters than for adopters, which suggests that greenhouse labor and off-farm jobs are substitutes.

(3) There is no significant difference in education between adopters and non-adopters in our sample. This suggests that education is not the main determinant of greenhouse adoption when the main source of information for the technology is social learning.

(4) The farm size of adopters is larger than that of non-adopters, which indicates that farmers with more land are more dependent on agricultural income, and farmers with less land are more likely to favor off-farm jobs.

(5) Irrigation is of course important to greenhouse farming, and 89 percent of the adopters have access to irrigation. However, 80 percent of the non-adopters also have access to irrigation, showing that there is not much variation in irrigation access among farmers in this well-irrigated region.

(6) Adopters have greater land tenure security than non-adopters. This is a sensible result given the long-term nature of greenhouse investment. We proxy land tenure security by the number of land reallocations undertaken by village leaders every few years to ensure relative land distribution equality in the village.

(7) Adopters and non-adopters have no significant difference in grain land share, which suggests that both groups have a similar agricultural production pattern except that adopters use greenhouses to produce vegetables and non-adopters produce vegetables in the open field.

(8) The presence of a credit constraint would in theory undermine an important

investment such as greenhouses, all else equal. However, it is difficult to measure a credit constraint facing a farmer, as this is equivalent to examining whether a farmer can borrow as much as he would like at the going market interest rate (Banerjee and Duflo, 2002). Since we are focusing on greenhouse adoption rather than testing whether the farmer has invested in a greenhouse of optimal scale, we only need to know whether a farmer is capable of building a greenhouse by borrowing money or using his savings. Therefore, we observed the house value as a proxy for household wealth. We also collected the household's credit history (maximum borrowing and maximum lending) before adoption as an indicator of how much credit/savings is available. Our data shows that non-adopters are significantly wealthier than are adopters before the latter's adoption; non-adopters have a mean house value of 8,773 yuan vs. 4,294 yuan for adopters. Similarly, non-adopters have significantly greater credit/savings than adopters. The maximum borrowing is 1,352 yuan for non-adopters vs. 925 yuan for adopters, and the maximum lending is 862 yuan for non-adopters vs. 368 yuan for adopters. Given that non-adopters are both wealthier and have more access to credit, credit constraints are unlikely to play an important role in greenhouse adoption in Shandong.

1.5 Empirical Methodology

In this section, we illustrate the connection between our theoretical model and the empirical framework. According to our real option model of greenhouse adoption, the farmer decides to adopt or to wait based on a comparison between the current value of the technology and the trigger value. Therefore, we can define the farmer's adoption status at time t as:

$$\begin{aligned} Y_t &= 1 \text{ (adopt)}, \text{if } Y_t^* = V_t - V_t^* > 0, \\ Y_t &= 0 \text{ (non - adopt)}, \text{if } Y_t^* = V_t - V_t^* \le 0, \end{aligned} \tag{1.7}$$

where V_t is the discounted expected value of all future cash flow from greenhouse vegetable production, and V_t^* is the trigger value.

McDonald and Siegel (1986)'s model, in which the drift μ is known, shows the trigger value V^* as a function of the parameters (ρ, μ, I, σ). However, the drift μ is unknown in our model. Thus, the trigger value also depends on the conditional mean and variance of the drift, (m_t, γ_t). According to the dynamics of (m, γ) in equation (1.6), we can substitute (m_t, γ_t) with functions of $\left(m_0, \gamma_0, Z_t', N_t, \sigma, \sigma_\varepsilon, \bar{\mu}, t\right)$.[4] Therefore, we can express the trigger value V_t^* as:

$$V_t^* = g\left(\rho, I, \sigma, m_0, \gamma_0, Z_t', N_t, \sigma_\varepsilon, \bar{\mu}, t\right) \tag{1.8}$$

---

[4] This is only a simplified representation; strictly speaking, the solution of (m_t, γ_t) according to equation (1.6) depends on the paths of Z' and N, as well as the history of the signals up to time t.

Following similar reasoning, the current project value V_t can be written as a function of the same group of variables. Therefore, we can express $Y_t^* = V_t - V_t^*$ as:

$$Y_t^* = h\left(\rho, I, \sigma, m0, \gamma 0, Z_t^{'}, N_t, \sigma_\varepsilon, \bar{\mu}, t\right). \tag{1.9}$$

To motivate the empirical proxies for the variables in equation (1.9), we first note that $Z_t^{'}$ represents the stochastic change in the project value. A good proxy for $Z_t^{'}$ is the observed profitability of greenhouse production in the current period. We proxy that profitability by the ratio of the output price to the input price. Because historical data are not available on vegetable prices in Shandong, we use the ratio of the vegetable price index and the input price index at the national level as a proxy for the profitability of greenhouse production over the years. For the investment cost *I*, we use the greenhouse construction cost (real value) for each adopter. For non-adopters, we use the average construction cost for adopters in their village or nearby villages as the proxy.

Continuing with the interpretation of equation (1.9), σ is the volatility of the project value, which we measure as the standard deviation of the national vegetable price index over the three years prior to the farmer's adoption. $\bar{\mu}$ represents the average signal received by the farmer from his social network, the proxy for which is the vegetable price index growth rate over the three years preceding the farmer's adoption. This is a reasonable assumption if the expected return of the project is close to the average return in the economy. The time t in our model is equated with the amount of time the agent spent in continuous learning. We use the number of years that the farmer had been aware of the technology before adoption to represent the continuous learning effect. As noted above, N_t is the key variable in our study. We measure it by the number of earlier adopters in a

farmer's social network, which includes relatives and friends in his own village and nearby villages.

Besides these theoretically motivated variables, there may be other factors that affect greenhouse adoption in practice, such as land tenure security, off-farm employment, and household wealth. These factors were discussed in the preceding section. In addition, we do not have compelling empirical proxies for farmers' discount factor ρ, their initial values of the conditional mean and variance (m_0, γ_0) before any learning had taken place, and the standard deviation of their signals σ_ε. These parameters, however, are likely correlated with household characteristics such as age, family size, and education, which we include in our empirical analysis to capture potential omitted factors.

Our theoretical model is based on observables; with knowledge of these observables, the model predicts adoption with certainty. In reality, however, we do not observe all information relevant for determining adoption. Therefore, our empirical model must allow for the presence of unobserved determinants.

In brief, our empirical model can be written as:

$$Y_i^* = f(X_i, Z_i, N_i, D_1, D_2) + e_i, \tag{1.10}$$

where i denotes a household, Y_i^* is the adoption criterion in year t according to equation (1.7), and X_i are household characteristics before adoption (year $t-1$), which include age, education of household head, family size, farm size, off-farm employment and income, family labor, irrigation conditions, family wealth, years of awareness of the technology, and greenhouse construction costs. Z_i are institutional and market variables at

$t-1$, which include the number of land reallocations, the ratio of the output price index to the input price index, the volatility of the vegetable price index, and the average growth rate of the vegetable price index. N_i is the number of earlier adopters in the farmer's social network at $t-1$. D_1 and D_2 are, respectively, year and county dummies that control for heterogeneity in farmers' adoption across different years and counties. Finally, e_i represents the effect of unobservable determinants of adoption. According to equation (5.4), the probability of adoption is:

$$P\left(Y_i^* > 0\right) = P(e_i > -f(X_i, Z_i, N_i, D_1, D_2)). \tag{1.11}$$

In our empirical analysis, we estimate a linear probability model (LPM), which specifies the above probability as a linear function of the explanatory variables. LPM has its strengths and weaknesses. (1) It is a linear model, which offers convenience in model estimation. For example, OLS provides consistent and even unbiased estimators and ease in dealing with heteroskedasticity using heteroskedasticity-robust standard errors and *t*-statistics. (2) However, the coefficients in the linear model measure the effect of the explanatory variables on the response probability. Unless the range of the explanatory variables is severely restricted, the LPM cannot be a good description of the population response probability. The hope is that the linear specification approximates the response probability for common values of the covariates; fortunately, this often turns out to be the case (Wooldridge 2002). (3) The LPM model allows us to use year dummies to control for heterogeneities over time, which is important to this empirical study given the structure of our data set (in which different farmers adopted greenhouse in different years). Therefore, even with some weaknesses, LPM often provides good estimates of the partial effects on

the response probability near the center of the distribution of the explanatory variables.

1.6 Empirical Results

1.6.1 Identification Strategy

In this section we focus on the potential endogeneity of the social learning effect and our identification strategy. The endogeneity problem is one of the most formidable problems in empirical studies. In order to find an appropriate identification strategy for this study, it is crucial to understand the reasons why we could face the problem.

Manski (1993) uses the reflection problem to describe the tendency for people in the same social network to behave in similar ways. He identifies two possibilities: (1) an endogenous effect, wherein the propensity of an individual to behave in certain ways varies with the prevalence of the behavior in the group; (2) a correlated effect, wherein common environment and personal characteristics produce similar behavior.

In this paper, we attempt to show that farmers' adoption decision is influenced by social learning. Therefore, we need to empirically distinguish the social learning effect from the endogenous effect and the correlated effect.

In our context, the endogenous effect is essentially the social pressure problem. Psychologists often use social pressure as a way of explaining herd behavior. For greenhouse adoption, adopters are usually the minority in most villages. From this observation one can infer that it would be rare for farmers to choose greenhouse adoption because of social pressure.

In our context, the correlated effect poses a more serious challenge. An endogeneity problem could arise from the simultaneous determination of adoption and network formation: for example, a farmer could know more adopters because he adopted the

greenhouse. In other words, the adoption could affect social learning instead of social learning affecting adoption (endogeneity from simultaneous determination). To mitigate this problem, we collected household and institutional information for the year before the adoption for adopters. For non-adopters, we collected the information in the year before the survey occurred (2005).

Moreover, farmers who are entrepreneurial in spirit are likely to know more people (hence more adopters). At the same time, they are more likely to try out new things (thus more likely to adopt). Therefore, a farmer's adoption could be explained by his personality, rather than by learning from others in his social network. Thus, a key problem is how to identify social learning from unobservable error terms such as similar personalities in the social network. We need to find at least one instrumental variable which is (1) correlated with social learning after we control for other factors, but that is (2) not correlated with the error terms. We can test the first condition. We cannot test the second condition directly because the error terms are not observable.

Fortunately, we have an appropriate instrument in this study: the walking time from the farm to a farmer's neighborhood. More specifically, we ask farmers the following question in the field survey: "How many minutes does it take to walk by your 20 closest neighbors?" The logic of this question is that social learning could be negatively correlated to the walking time. For example, if a farmer lives in a mountainous area, it could take two hours or even more to walk by his 20 closest neighbors. On the contrary, it only takes 10 minutes for farmers to walk by his 20 closest neighbors if people live closely. We surmise that farmers in the second case are more likely to have access to social learning. We test this hypothesis with data after controlling for other factors: we find that walking time is

significantly negatively correlated with social learning (first row of Table 1.3 for both social learning measures). This result demonstrates that the walking time variable satisfies the first condition for a valid instrument.

For an analysis of whether this instrument meets the second condition (lack of correlation with the error term in the adoption equation), the following three-step discussion provides further justification for the validity of the instrument.

First, we use a heuristic explanation to justify the instrument. In rural China, it is not unusual for a family to live in the same place for decades. A well-functioning real estate market does not exist in rural China for several reasons: (1) a farmer could own his house, but not the land on which his house is built because all land is owned by the village collective; (2) it is illegal to buy a house in a village if the buyer is not a member of the village; (3) it is also illegal for a household to buy an additional house from another villager because Chinese law forbids any household to occupy two pieces of land for housing in a village; (4) if a farmer wants to change his house location, either he has to obtain a new piece of land from the village collective under very strict conditions due to land scarcity in Shandong, or he can find another household in the village that is willing to give up its housing land, which is very rare. In addition, in both cases the farmer has to give up his old housing land. Based on these observations, it appears very difficult, if not impossible, for a household to change its location. In other words, the farmer's housing location in rural China can be considered as fixed in most cases. From this we infer that the walking time to the neighborhood is fixed and exogenous to greenhouse adoption.

Second, we constructed interaction terms between the IV (distance to neighborhood) and year dummies. We used the Hansen-J over-identification test to examine the validity of

the IV given that we believe the other instruments (the interaction terms) to be truly exogenous. The C-statistic from the Hansen-J test (the last row of Table 1.4) indicates that the distance to neighborhood variable passes the validity test in both social learning measurements. We must be cautious by not over-emphasizing this result, as the power of the Hansen-J test depends on the exogeneity of the other instruments. However, this is the best test we can do to check the validity of an instrumental variable.

Finally, we tabulate the distance to neighborhood by household characteristics such as education, age, and wealth. These simple but reliable summary statistics can tell us whether the distance to neighborhood is correlated with typical household characteristics. If the distance to neighborhood is truly exogenous due to the fixed housing location in rural China, we would not expect to see a significant correlation with household characteristics. Indeed, the results in Table 1.5 indicate that the distance to neighborhood does not show any robust correlation with the education and age of the household head, or the real value of the house. These findings lend support to our working hypothesis that that the distance to neighborhood is exogenous to greenhouse adoption.

As a result of these discussions, we are fairly confident that the IV (distance to neighborhood) is exogenous to greenhouse adoption, and therefore it allows us to obtain consistent estimators given that social learning is shown to be endogenous by the Durbin-Wu-Hausman Test (last row of Table 1.2).

1.6.2 Linear Probability Model

Table 1.2 presents the estimation results for the linear probability model estimated by 2SLS with cluster-robust standard errors using distance to neighborhood as the instrument.

The first two columns report the results using a measure of social learning within the farmer's own village; the next two columns report the results using a measure of social learning that also includes the farmer's nearby villages. Generally speaking, the two sets of results are very similar, suggesting that village boundaries are not crucial to how social learning affects greenhouse adoption.

We will focus on the first two columns for a detailed discussion of our results. The first row confirms the key result for our study: social learning has a significantly positive impact on greenhouse adoption. Specifically, one more adopter in a farmer's social network increases the probability of his adoption by 1.9 percent after controlling for other factors. In other words, if there are currently 10 earlier adopters in the farmer's social network, his adoption probability in the next year will increase by about 19 percent. Given that the greenhouse adoption rate is still low in rural China, this amount of increasing probability is economically significant.

The third row shows how adoption is affected by the conditional mean return to the greenhouse technology. From our theoretical model, we know that the farmer's belief about the mean return will converge to the average belief of his social network as a result of social learning. Because we cannot observe farmers' expectations, we use the vegetable price index (national level) growth rate before adoption to approximate the average belief of project return in the social network. The coefficient is not significant; however, the sign is consistent with the prediction of our theoretical model, namely, higher expected return results in a higher trigger value for investment and a lower probability for adoption. It is also possible that the price index growth rate is acting as a proxy for farmers' outside opportunities; however, we have already included off-farm income in our regression

specification.

We use the market volatility of vegetable prices before adoption to represent the uncertainty in the stochastic project value in our theoretical model. Our result indicates that this source of uncertainty discourages adoption. This finding is consistent with theory, which predicts that the option value of waiting to invest is larger when the future investment value is more uncertain.

We use the number of years that the farmer had been aware of the technology before adoption to represent the continuous learning effect. However, it is not significant according to our estimation. It could be that farmers in rural China simply did not have continuous access to information about the greenhouse technology and its returns. It is also possible that the main source of information about the greenhouse technology is discrete social learning.

Our proxy for the current profitability of the greenhouse technology is the ratio of the output price index to the input price index: the higher is the stochastic project value, the higher is the probability of adoption. Our result confirms this prediction.

Among the included household characteristics, only the age of the family head is statistically significant. However, the effects of most household characteristics are consistent with our discussion in section 1.4.3. The R^2 of this regression is 0.83, which suggests that we have included most of the factors that could affect the adoption decision. It also reinforces the idea that our irreversible investment model is an appropriate choice for describing the greenhouse adoption behavior.

In Table 1.4, the interaction terms between the distance to neighborhood and the year dummies are included as extra instruments in the regression. The results are very similar to

the results in Table 1.2, which suggests that the results are robust. Moreover, the extra instruments allow us to use the Hansen-J test to test the validity of the IV (distance to neighborhood).

1.7 Conclusion

In technology adoption with irreversible investment, agents commonly face two sources of uncertainty. First, the future value of the investment is uncertain. Second, agents have incomplete information regarding the parameters of the process describing the future investment value. In this paper, we model social learning as a way of reducing parameter uncertainty, thus facilitating technology adoption with irreversible investment. We use household-level data from intermediate-technology greenhouse adoption in northern China to test the predictions, with the following main results.

(1) Social learning has a significantly positive impact on greenhouse adoption. Ten more adopters in the farmer's social network increase the probability of adoption by 19 percent, which is an economically significant effect.

(2) The empirical data confirms what we know from the conventional theory of irreversible investment: higher uncertainty about the future investment value results in less adoption.

(3) Social learning could also affect technology adoption through its influence on the farmer's belief about the expected return on the technology. The empirical data offers some support for this hypothesis.

Our paper also provides an answer to the following question: how could small farmers in developing countries deal with the risk from irreversible investment and incomplete information? Our results suggest that social learning can be an effective solution. Therefore, the policy implication from this paper is clear: when small farmers face technology adoptions such as investing in tube wells or machinery, helping several farmers adopt

successfully may be the best way to induce more adoption in their village.

Figure 1.1 Greenhouse Diffusion Curve at the Household Level

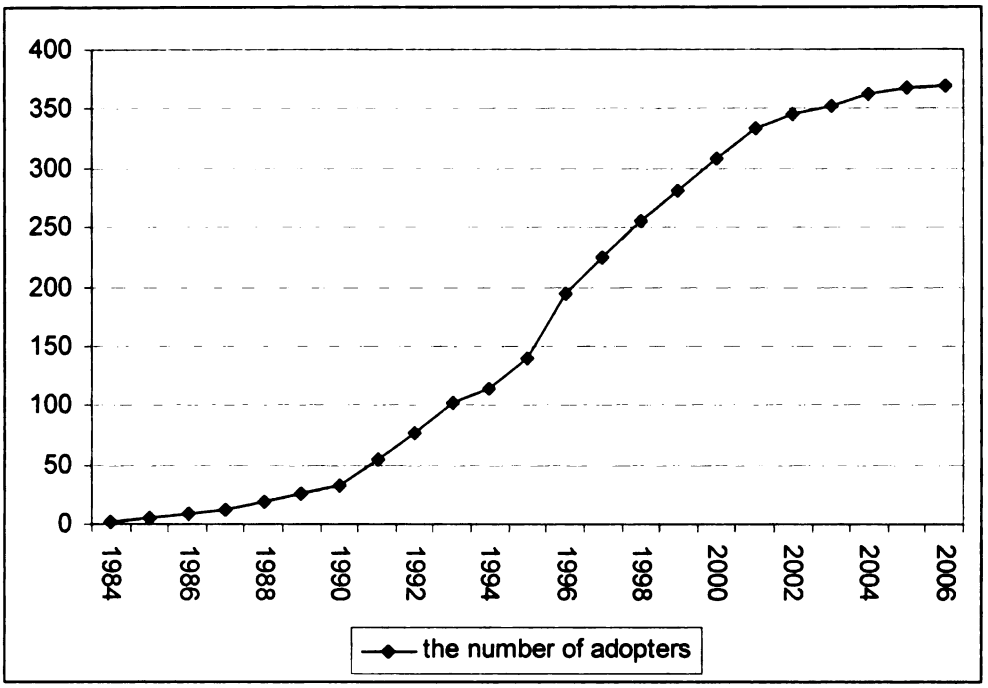


Table 1.1 Descriptive Statistics: Household Level Data

This table contains the basic household characteristics used in our study. The mean value for each variable is presented with the associated standard error in parentheses. For adopters, all variables are measured in the year before adoption. For non-adopters, all variables are measured in the year before the survey. *** denotes significance at one-percent, ** five-percent, and * ten-percent level.

Basic characteristics	Non-adopter	Adopter	Test of equality of the means (p-value)
Social learning within village	4.7	6.9	0.027**
	(0.7)	(0.67)	
Social learning within village and nearby villages	5.8	8.45	0.018**
	(0.8)	(0.76)	
Family size	3.7	3.9	0.016**
	(0.07)	(0.06)	
Farm labor	2.92	2.46	0.01***
	(0.07)	(0.043)	
Off-farm employment	0.8	0.24	0.01***
	(0.054)	(0.022)	
Age of family head	46.4	35	0.01***
	(0.6)	(0.46)	
Education of family head	7.0	7.24	0.25
	(0.17)	(0.14)	
Off-farm income (yuan)	8420	1643	0.01***
	(649)	(182)	
Farm size (mu)	5.6	6.01	0.09*
	(0.19)	(0.16)	
Irrigation ratio	0.80	0.89	0.01***
	(0.019)	(0.013)	
Major land reallocations since 1980	1.44	0.79	0.01***
	(0.067)	(0.05)	
Minor land reallocations since 1980	4.29	3.19	0.01***
	(0.26)	(0.19)	
House value (yuan)	8773	4294	0.01***
	(539)	(413)	
Grain Land Share (percent)	0.579	0.577	0.92
	(0.282)	(0.252)	
Maximum lend	862	368	0.01***
	(104)	(66)	
Maximum borrow	1352	925	0.01**
	(146)	(102)	

Table 1.2 Greenhouse Adoption and Social Learning: LPM Estimated by 2SLS

This table contains a 2SLS estimation of the linear probability model for farmers' adoption decision. The instrumental variable for social learning is distance to neighborhood (measured by the walking time to the 20 closest neighbors). The dependent variable is 1 for adopters and 0 for non-adopters. *** denotes significance at one-percent, ** five-percent, and * ten-percent level.

Explanatory variables	Coefficient	Robust std error	Coefficient	Robust std error
Social Learning				
Social learning within village	0.019	0.01**		
Social learning within village and nearby villages			0.017	0.009**
Conditional mean of market return	-0.41	0.34	-0.37	0.31
Market volatility	-0.0017	0.0006**	-0.0017	0.0006***
Years of awareness of the technology	-0.009	0.0073	-0.01	0.008
Output price/input price	0.83	0.21***	0.86	0.22***
Household Characteristics				
Family size	0.020	0.017	0.021	0.016
Age of family head	-0.0034	0.0017**	-0.0037	0.0015**
Education of family head	0.0012	0.0044	-0.0003	0.004
Off-farm income	-0.0068	0.0065	-0.007	0.006
Farm size	0.006	0.006	0.0075	0.0054
Irrigation ratio	0.058	0.045	0.060	0.038
House value	-0.0017	0.0032	-0.0022	0.0031
Greenhouse construction cost	0.0073	0.015	0.006	0.14
Times of major reallocations	-0.017	0.03	-0.025	0.031
Times of minor reallocations	-0.001	0.008	0.0005	0.008
Grain share	0.158	0.095	0.144	0.089
Dummies and constant terms				
Crop dummy	0.0043	0.038	0.061	0.044
County dummies	Yes		Yes	
Year dummies	Yes		Yes	
Constant terms	-0.687	0.26**	-0.69	0.26**
Observations	626		626	
Adjusted R-squared	0.83		0.84	
Durbin–Wu–Hausman Test for Endogeneity	*p*-value	0.014	*p*-value	0.013

Table 1.3 Greenhouse Adoption and Social Learning: First Stage 2SLS Results

This table contains the first stage results of a 2SLS estimation of the linear probability model for farmers' adoption decision. The dependent variable is social learning within village or social learning within village or nearby villages. *** denotes significance at one percent, ** five percent, and * ten percent level.

	Social learning within village		Social learning within village and nearby villages	
Explanatory variables	Coefficient	Robust std error	Coefficient	Robust std error
Walking time to 20 closest neighbors	-0.088	0.044**	-0.093	0.046**
Conditional mean of market return	1.02	2.535	-1.375	2.585
Market volatility	0.005	0.012	0.007	0.015
Years of awareness of the technology	0.657	0.30**	0.771	0.304**
Output price/input price	14.82	16.2	14.09	16.94
Household Characteristics				
Family size	0.132	0.743	0.086	0.784
Age of family head	-0.062	0.080	-0.046	0.089
Education of family head	-0.132	0.286	-0.052	0.308
Off-farm income	-0.187	0.338	-0.185	0.346
Farm size	-0.030	0.358	-0.107	0.351
Irrigation ratio	1.729	4.076	1.700	3.656
House value	-0.035	0.143	-0.010	0.143
Greenhouse construction cost	-1.169	0.605*	-1.168	0.580**
Times of major reallocations	1.416	0.974	1.946	0.970**
Times of minor reallocations	-0.610	0.521	-0.727	0.526
Grain share	0.847	4.289		
Dummies and constant terms				
Crop dummy	-1.898	2.092	-3.11	3.05
County dummies	Yes		Yes	
Year dummies	Yes		Yes	
Constant terms	-5.996	19.69	-6.21	20.62
Observations	626		626	
Adjusted R-squared	0.267		0.293	

Table 1.4 Greenhouse Adoption and Social Learning: LPM with Interaction Terms

This table contains a 2SLS estimation of the linear probability model for farmers' adoption decision. The instrumental variables for social learning include distance to neighborhood (measured by the walking time to the 20 closest neighbors) and its interaction with year dummies. The dependent variable is 1 for adopters and 0 for non-adopters. *** denotes significance at one-percent, ** five-percent, and * ten-percent level.

Explanatory variables	Coefficient	Robust std error	Coefficient	Robust std error
Social Learning				
Social learning within village	0.019	0.009**		
Social learning within village and nearby villages			0.018	0.008**
Conditional mean of market return	-0.415	0.357	-0.37	0.33
Market volatility	-0.0017	0.0006**	-0.0017	0.0005***
Years of awareness of the technology	-0.0094	0.0073	-0.01	0.0073
Output price/input price	0.82	0.211***	0.85	0.22***
Household Characteristics				
Family size	0.020	0.017	0.021	0.016
Age of family head	-0.0034	0.0017**	-0.0037	0.0015**
Education of family head	0.0013	0.0044	-0.0003	0.004
Off-farm income	-0.0068	0.0067	-0.007	0.006
Farm size	0.006	0.006	0.0076	0.0057
Irrigation ratio	0.057	0.047	0.059	0.039
House value	-0.0017	0.0032	-0.0022	0.0031
Greenhouse construction cost	0.0082	0.015	0.007	0.13
Times of major reallocations	-0.018	0.03	-0.026	0.032
Times of minor reallocations	-0.0004	0.0082	0.0009	0.008
Grain share	0.157	0.097	0.143	0.09
Dummies and constant terms				
Crop dummy	0.0044	0.039	0.064	0.044
County dummies	Yes		Yes	
Year dummies	Yes		Yes	
Interaction terms	Yes		Yes	
Constant terms	-0.681	0.26**	-0.684	0.27**
Observations	626		626	
Adjusted R-squared	0.82		0.83	
Over-Identification Hansen J Test: C-Statistics	*p*-value	0.20	*p*-value	0.21

Table 1.5 Distance to Neighborhood and Characteristics of Household

This table summarizes the walking time to the 20 closest neighbors for households categorized by their education, wealth, and age levels.

Education of family head (school year)	Distance to 20 closest neighbors (minute)	Real value of House (10,000 Yuan)	Distance to 20 closest neighbors (minute)	Age of head of household (year)	Distance to 20 closest neighbors (minute)
0	14	<0.2	14	<20	18
1	21	0.2~0.5	25	20~25	16
2	13	0.5~1	16	25~30	14
3	15	1~2	16	30~35	17
4	21	2~3	16	35~40	16
5	19	3~4	16	40~45	16
6	17	4~5	16	45~50	15
7	13	5~6	13	50~55	17
8	15	6~7	15	55~60	16
9	16	7~8	17	>60	15
10	15	8~9	13		
>11	13	9~10	16		
		>10	16		

BIBLIOGRAPY

Abasov, T. M. (2005): Dynamic learning effect in corporate finance and risk management. Ph.D. Dissertation. University of California, Irvine.

Banerjee, A., and Duflo, E. (2002). Do firms want to borrow more? Testing credit constraints using a directed lending Program. MIT Department of Economics, Working Paper No. 02-25.

Bradiera, O., and Rasul, I. (2006). Social network and technology adoption in Northern Mozambique. Economic Journal: 116, 869-902.

Bertola, G., and Caballero, R. (1994). Irreversibility and aggregate investment. Review of Economic Studies: 61, 223-246.

Besley, T., and Case, A. (1994). Diffusion as a learning process. Evidence from HYV cotton. mimeo, Princeton University.

Brennan, M. J. (1998). The role of learning in dynamic portfolio decisions. European Economic Review: 1, 295-306.

Chinese Agricultural Yearbook (2006). Chinese Agricultural Press.

Conley, T., and Udry, C. (2001). Learning about a new technology: pineapple in Ghana. American Journal of Agricultural Economics: 83, 668-673.

Dixit, A. K., and Pindyck, R. S. (1994). Investment under Uncertainty. Princeton University Press.

Feder, G. (1980). Farm size, risk aversion and the adoption of new technology under uncertainty. Oxford Economic Papers, New Series: 32, 2, 263-283.

Foster, A., and Rosenzweig, M. (1995). Learning by doing and learning from others: human capital and technical change in agriculture. Journal of Political Economy: 103, 1176-1209.

Gennotte, G. (1986). Optimal portfolio choice under incomplete information. Journal of Finance: 41, 733-746.

Griliches, Z. (1957). Hybrid corn: an exploration in the economics of technological change. Econometrica: 25, 501-522.

Hassett, K. A., and Metcalf, G. E. (1995). Energy tax credits and residential conservation investment. NBER Working Paper No. W4020.

Huang, L., and Liu, H. (2007). Rational inattention and portfolio selection. Journal of Finance: 62, 1999-2040.

Liptser, R., and Shiryaev, A. (2001). Statistics of random processes. Springer-Verlag, Berlin.

Manski, C. F. (1993). Identification of social effects: reflection problem. Review of Economic Studies: 60, 531-542.

McDonald, R., and Siegel, D. (1986). The value of waiting to invest. Quarterly Journal of Economics: 101, 707-728.

Merton, R. C. (1980). On estimating the expected return on the market. Journal of Financial Economics: 8, 323-361.

Munshi, K. (2004). Social learning in a heterogeneous population: social learning in the Indian green revolution. Journal of Development Economics: 73, 185-213.

Nelson, A. W., and Amegbeto, K. (1998). Option values to conservation and agricultural price policy: application to terrace construction in Kenya. American Journal of Agricultural Economics: 80, 409-418.

Newbery, D. and J. Stiglitz (1981). The theory of commodity price stabilization. Oxford: Clarendon Press.

Olmstead, A. L., and Rhode, P (1993). Induced innovation in American agriculture: a reconsideration. Journal of Political Economy: 101, 100-118.

Roumasset, J. (1976). Rice and risk: decision making among low income farmers. Amsterdam: North Holland.

Sunding, D., and Zilberman, D. (2000). Research and technology adoption in a changing agricultural sector. Draft for the Handbook of Agricultural Economics.

Wan, X. (2000). The Chinese protection agriculture outlook and trend. Agricultural Machinery: 2000, 4-6 (in Chinese).

Weinberger, K., and Lumpkin, T. (2005). Horticulture for poverty alleviation: the unfunded revolution. AVRDC Working Paper 15.

Wooldridge, J. (2002). Econometric analysis of cross section and panel data. MIT Press, Cambridge.

Xia, Y. (2001). Learning about predictability: the effects of parameter uncertainty on

dynamic asset allocation. Journal of Finance: 56, 205-246.

Zilberman, D., Sunding, D., Howitt, R., Dinar, A., and MacDougall, R. (1994). Water for California agriculture: lessons from the drought and new water market reform. Choices: 4, 25-28.

Chapter 2: Partial Maximum Likelihood Estimation of a Spatial Probit Model

2.1 Introduction

Most econometrics techniques on cross-section data are based on the assumption of independence of observations. However, economic activities become more and more correlated over space with modern communication and transportation improvements. On the other hand, technological advances in communications and the geographic information system (GIS) make spatial data more available than before. Spatial correlations among observations received more and more attentions in regional, real estate, agricultural, environmental and industrial organizations economics (Lee 2004).

Econometricians began to pay more attention on spatial dependence problems in the last two decades and some important advances have been done in both theoretical and empirical studies[5]. Spatial dependence not only means lack of independence between observations, but also a spatial structure underlying these spatial correlations (Anselin and Florax 1995). There are two ways to capture spatial dependence by imposing structures on a model: one is in the domain of geostatistics where the spatial index is continuous (Conley 1999), the other is that spatial sites form a countable lattice (Lee 2004). Among the lattice models, there are also two types of spatial dependence models according to spatial correlation between variables or error terms: the spatial autoregressive dependent variable model (SAR) and the spatial autoregressive error model (SAE). In most applications of

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

spatial models, the dependent variables are continuous (Conley 1999; Lee 2004; Kelejian and Prucha, 1999, 2001; among others), and only few applications address the spatial dependence with discrete choice dependent variables (exceptions include: Case 1991; McMillen 1995; Pinkse and Slade, 1998; Lesage 2000; Beron and Vijerberg 2003). This paper is designed to address this gap and we are concerned about the SAE model with discrete choice dependent variables.

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

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

Dependence is the other aspect of this problem. General forms of dependence are rarely allowed for in cross-sectional data, although routinely allowed for in time-series data (Conley 1999). For example, some scholars discussed discrete choice models with dependence in time-series data: Robinson (1982) relaxed Amemiya (1973) assumptions of

independence in Tobit model, and proved that the MLE with dependent observations is strongly consistent and asymptotically normal under some regularity conditions. Poirier and Rudd (1988) discussed the Probit model with dependence in time-series data, and developed generalized conditional moment (GCM) estimators which are computational attractive and relatively more efficient.

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

Therefore, how to deal with dependence in space in estimation is the key to spatial econometricians. Inspired by works about dependence in time-series data, Conley (1999) uses metrics of economic distance to characterize dependence among agents, and shows that the GMM estimator is consistent and asymptotically normal under some assumptions similar to time-series data. He also provides how to get consistent covariance matrix estimator by an approach similar to Newey-West (1987). Pinkse and Slade (1998) use GMM in the discrete choice setting with the SAE model, and show that the GMM estimator remains consistent and asymptotically normal under some regularity conditions. Although Pinkse and Slade (1998) generated generalized residuals from the MLE as the basis of the GMM estimators, they do not take advantage of information from spatial correlations among observations, and hence the GMM estimator is much less efficient than full ML

estimators. Lee (2004) examines carefully the asymptotic properties of MLE and quasi-MLE for the linear spatial autoregressive model (SAR), and he shows that the rate of convergence of those estimators may depend on some general features of the spatial weights matrix of the model. If each units are influenced by only a few neighboring units, the estimators may have $\sqrt{n}$ -rate of convergence and asymptotic normality; otherwise, it may have lower rate of convergence and estimators could be inconsistent.

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

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

2.2 Discrete Choice Models with Spatial Dependence

2.2.1 Probit Model without Dependence

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

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

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

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

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

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

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

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

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

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

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

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

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

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

where ϕ is the probability density function (pdf) of the standard normal distribution. However, the simple closed-form expressions for the MLE are not available because the cdf of the normal distribution has no close-form solution. So the MLE must be solved by using numerical algorithms[6]. In general, we can prove that the conditional MLE is consistent and the most efficient estimator given some regularity conditions[7] such as correctly specifying a parametric model, an identified β and a log-likelihood function that is continuous in β.

2.2.2 A Probit Model with Spatial Error Correlation

Consider the Probit model with spatial error correlation (SAE), where the underlying

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

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

linear latent variable model is

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

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

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

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

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

so that the variance-covariance matrix for the model is

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

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

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

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

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

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

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

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

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

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

(above, (16) and (17)

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

2.2.3 Probit Models with Other Forms of Spatial Correlation

Generally, there is no reason to think that spatial correlation is properly modeled by (9). Other forms are possible. For example, one might assume that, outside of a certain geographic radius from a given observation i, ε_i is uncorrelated with shocks to the

outlying regions. So, for example, we might assume a constant correlation with any unit within a given radius -- similar to a random effects structure for unbalanced panel data. Alternatively, we may prefer more of a moving average structure, such as

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

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

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

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

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

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

2.3 Using Partial MLEs to Estimate General Spatial Probit Models

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

To get an estimator that is computationally feasible, Pinkse and Slade (1998) proposed using generalized method of moments (GMM) using information on the marginal distributions of the binary responses. In particular, the generalized residuals from the marginal probit log likelihood are used to construct moment conditions for the GMM method. Pinske and Slade show that, under conditions very similar to those in this paper, the GMM estimator is consistent and asymptotically normal. The consistent variance-covariance matrix can also be obtained theoretically without a covariance stationary assumption, although Pinske and Slade (1998) do not discuss estimation of the asymptotic variance. Therefore, the GMM estimator is almost practically useful, but it is fundamentally based on the marginal probit models. Thus, while a GMM estimator can be obtained that is efficient given the information on the marginal likelihood, the method throws out much useful information. We describe a simplified version of this approach in section 2.3.1, which, in effect, uses a heteroskedastic probit model to estimate the β_j along with any spatial autocorrelation parameter.

Using only the marginal distribution of Y_i, conditional on the covariates and weights, likely results in serious loss of information for estimating both β and the spatial autocorrelation parameters. Our key contribution in this paper is to explore the use of partial maximum likelihood where we group small numbers of nearby observations and obtain the joint distribution of those observations. Naturally, these distributions are determined by the fully specified spatial autocorrelation model -- just as we must obtain the implied variance to apply marginal probit methods. Once the covariances between observations are found as a function of the weights and λ , we can use that information in multivariate probit estimation. Section 2.3.2 covers the case of where we describe a bivariate probit approach, with heteroskedasticity and covariance implied by the particular spatial autocorrelation model. Using a single covariance in addition to the variance seems likely to improve efficiency of estimation.

2.3.1 Univariate Probit Partial MLE

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

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

where $\sigma_i^2(\lambda) = Var(\varepsilon_i \mid X, W) = Var(\varepsilon_i \mid W)$ is a function of all weights, W, and the spatial correlation parameters λ. As is well-known in time series contexts – for example,

Poirier and Ruud (1988) or Robinson (1982) – using probit while ignoring the time series correlation leads to consistent estimation under standard regularity conditions, provided the data are weakly dependent. Thus, it is not surprising that pooled probit that accounts for the heteroskedasticity in the marginal distribution is generally consistent for spatially correlated data, too -- provided, of course, we limit the amount of spatial correlation.

The log likelihood can be written generically as

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

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

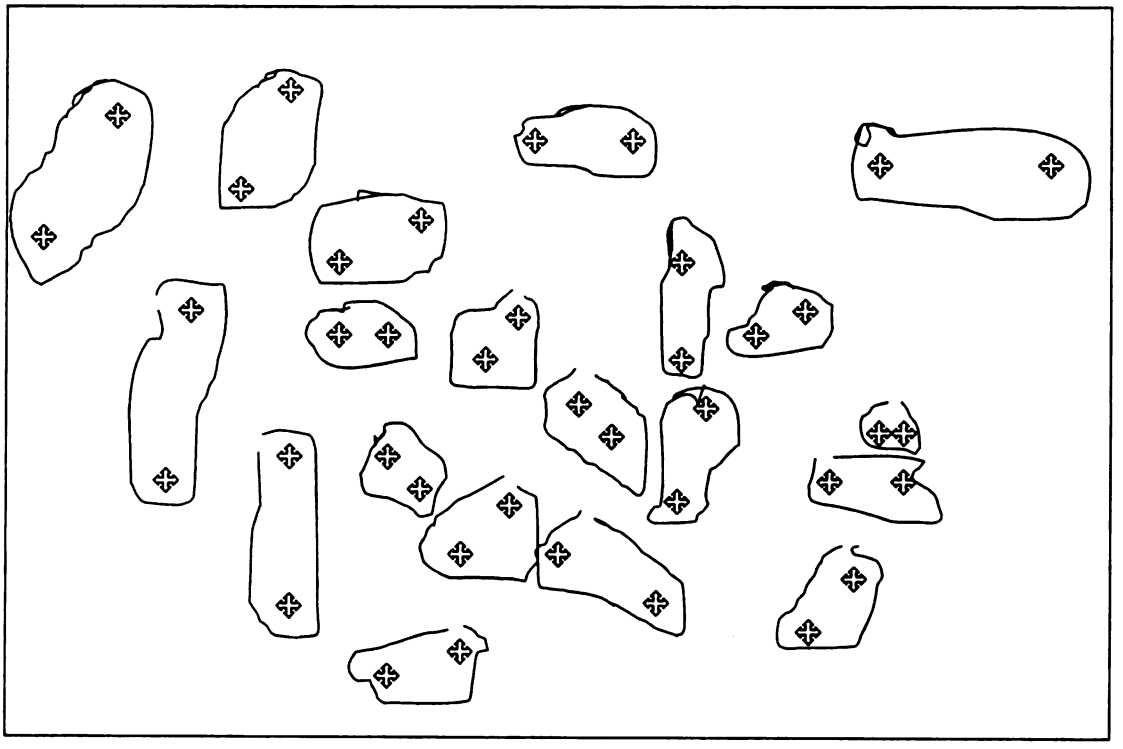
2.3.2 Bivariate Probit Partial MLE

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

For illustration, assume a sample includes *2n* observations, and we divide the *2n* observations into *n* pairwise groups according to the spatial Euclidean distance between them (Figure 2.1). In other words, each group includes two observations, with the idea being that the internal correlation between the two observations is more important than external correlations with observations in other groups. Within a group, the two

observations follow a conditional bivariate normal distribution because error terms are assumed to have a joint normal distribution.

Figure 2.1 N pairwise groups of 2n observations based on Euclidean Distance



In group g , we have

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

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

Rewrite the above equations in matrix form as

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

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

where X_{g1} and X_{g2} are $1 \times K$ vectors of regressors and β is a $K \times 1$ vector. ε_{g1} and ε_{g2}

are scalars. In group g, observation A and observation B are not only correlated with each other, but also correlated with other observations over space. Therefore, the variances and covariance between ε_{g1} and ε_{g2} not only depend on the weight within group, but also weights with other observations out of the group, and of course the parameters, λ. See, for example equation (20).

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

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

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

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

$$Y_g = \begin{Bmatrix} 1 & \text{if } Y_g^* > 0, \\ 0 & \text{if } Y_g^* \le 0 \end{Bmatrix}. \tag{28}$$

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

$$P(Y_{g1}=1,Y_{g2}=1\mid X_g)=P(X_{g1}\beta+\varepsilon_{g1}>0,X_{g2}\beta+\varepsilon_{g2}>0\mid X_g) \tag{29}$$

$$=P(\varepsilon_{g1}<X_{g1}\beta,\varepsilon_{g2}<X_{g2}\beta\mid X_g)=\Phi_2(\frac{X_{g1}\beta}{\sqrt{\Omega_{g11}}},\frac{X_{g2}\beta}{\sqrt{\Omega_{g22}}},\rho_g), \tag{30}$$

$$\rho_g=\frac{Cov(\varepsilon_{g1},\varepsilon_{g2})}{\sqrt{Var(\varepsilon_{g1})}\sqrt{Var(\varepsilon_{g2})}}=\frac{\Omega_{g12}}{\sqrt{\Omega_{g11}\Omega_{g22}}}, \tag{31}$$

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

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

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

where

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

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

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

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

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

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

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

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

Therefore

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

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

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

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

$P(Y_{g1} = 1 \mid Y_{g2} = 1, X_g)$.

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

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

Therefore,

$$P(Y_{g1} = 1 \mid Y_{g2} = 1, X_g) = E[P(Y_{g1} = 1 \mid X_g, \varepsilon_{g2}) \mid Y_{g2} = 1, X_g] \tag{40}$$

$$= E\left[\Phi\left(\frac{X_{g1}\beta + \delta_{g1}\varepsilon_{g2}}{\sqrt{Var(e_{g1})}}\right) \mid Y_{g2} = 1, X_g\right] \tag{41}$$

$$= \frac{1}{\Phi\left(\frac{X_{g2}\beta}{\sqrt{Var(\varepsilon_{g2})}}\right)} \int_{-X_{g2}\beta}^{\infty} \Phi\left(\frac{X_{g1}\beta + \delta_{g1}\varepsilon_{g2}}{\sqrt{Var(e_{g1})}}\right) \phi\left(\frac{\varepsilon_{g2}}{\sqrt{Var(\varepsilon_{g2})}}\right) d\varepsilon_{g2} \tag{42}$$

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

Similarly, we can get

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

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

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

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

$$\times\Phi(\frac{X_{g2}\beta}{\sqrt{Var(\varepsilon_{g2})}}) \tag{44}$$

$$=\int_{-X_{g2}\beta}^{\infty}\Phi(\frac{X_{g1}\beta+\delta_{g1}\varepsilon_{g2}}{\sqrt{Var(e_{g1})}})\phi(\frac{\varepsilon_{g2}}{\sqrt{Var(\varepsilon_{g2})}})d\varepsilon_{g2}, \tag{45}$$

and similarly we can obtain finally

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

$$P(Y_{g1}=1,Y_{g2}=0\mid X_g)=\int_{-\infty}^{X_{g2}\beta}\Phi(\frac{X_{g1}\beta+\delta_{g1}\varepsilon_{g2}}{\sqrt{Var(e_{g1})}})\phi(\frac{\varepsilon_{g2}}{\sqrt{Var(\varepsilon_{g2})}})d\varepsilon_{g2} \tag{46}$$

$$P(Y_{g1}=0,Y_{g2}=0\mid X_g) \tag{47}$$

$$=[1-\Phi(\frac{X_{g2}\beta}{\sqrt{Var(\varepsilon_{g2})}})]-\int_{-\infty}^{X_{g2}\beta}\Phi(\frac{X_{g1}\beta+\delta_{g1}\varepsilon_{g2}}{\sqrt{Var(e_{g\theta 1})}})\phi(\frac{\varepsilon_{g2}}{\sqrt{Var(\varepsilon_{g2})}})d\varepsilon_{g2}. \tag{48}$$

2.4 Partial Maximum Likelihood Estimation

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

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

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

We can apply a similar idea to the spatial Probit model: if we have the bivariate normal densities $\phi_{2g}(Y_{g1}, Y_{g2} \mid X_g, \theta)$ correctly specified for each group, we could get a consistent estimator by partial ML. However, there are several differences between panel data and spatial dependent data: first, the panel data model assumes that the cross section

dimension (N) is sufficiently large relative to the time dimension (T), but in spatial data we do not have this assumption. Second, in the panel data model, we view the cross section observations as independent, while in the spatial data model, even though we divided the sample into n groups, however, we are definitely not assuming independence among groups. Observations in different groups are still correlated, but the correlations are assumed to decay as distances become further away. Third, as we discussed before, dependence in space is more complicated than dependence in time, and we need to assume that the correlations between groups die out quickly enough as distance goes further away. In short, we need to examine carefully how the weak law of large numbers (WLLN) and central limit theorem (CLT) can be applied in the spatial dependent case. We will discuss these issues and provide proofs in the following sections.

First, we can write the partial log likelihood function as

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

and for the sake of brevity, we define

$$P_g(1,1) \equiv \log P_g(Y_{g1}=1, Y_{g2}=1 \mid X_g); \quad P_g(1,0) \equiv \log P_g(Y_{g1}=1, Y_{g2}=0 \mid X_g); \tag{51}$$

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

Therefore, we can rewrite the partial log likelihood function as

$$\begin{aligned} L = \sum_{g=1}^{n} \{ & Y_{g1}Y_{g2}P_g(1,1) + Y_{g1}(1-Y_{g2})P_g(1,0) \\ & + (1-Y_{g1})Y_{g2}P_g(0,1) + (1-Y_{g1})(1-Y_{g2})P_g(0,0)\}. \end{aligned} \tag{53}$$

2.4.1 Consistency of Bivariate Probit Estimation

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

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

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

i.e, in other words,

$$\hat{\theta} = \arg\max_{\theta \in \Theta} Q_n(\theta) \tag{55}$$

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

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

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

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

The following Theorem 1 states the main consistency result. We define

$$S(\theta) \equiv \frac{\partial Q_n}{\partial \theta}(\theta) \quad \text{and} \quad \lim_{n\to\infty} E[S_n(\theta)] = S(\theta).$$

THEOREM 1. *If (i) θ_0 is the interior of a compact set Θ, which is the closure of a concave set, (ii) Q attains a unique maximum over the compact set Θ at θ_0, (iii) Q is continuous on Θ, (iv) the density of observations in any region whose area exceeds a fixed minimum is bounded, (v) as $n\to\infty$,*

$$\sup_g \left(\left\| \frac{1}{\Pr(Y_{g1}=1, Y_{g2}=1|X_g)} + \frac{1}{\Pr(Y_{g1}=1, Y_{g2}=0|X_g)} + \frac{1}{\Pr(Y_{g1}=0, Y_{g2}=1|X_g)} + \frac{1}{\Pr(Y_{g1}=0, Y_{g2}=0|X_g)} \right\| \right) < \infty,$$

(vi) as $n \to \infty$, $\sup_g (\|X_g\| + \|Y_g\|) = O(1)$, (vii) $\sup_{ngj} |Cov(Y_{gi}, Y_{ji})| \le \alpha(d_{gj}), i = 1,2$

where d_{gj} denotes the distance between group g and j, and $\alpha(d) \to 0$ as $d \to \infty$, and

(viii) $\lim_{n\to\infty} E[Q_n(\theta)]$ exists, (ix) $\sup_g \|W_g\| < \infty$, then $\hat{\theta} - \theta_0 = o_p(1)$

Proof: Given in Appendix 1.

Condition (i) is a standard assumption from set theory. Condition (ii) is the identification condition for MLE. Condition (iii) assumes that the function Q is continuous

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

2.2.2 Asymptotic Normality

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

so that the CLT argument can be applied to this derived sequence (Davidson 1994).

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

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

$$\sqrt{n}(\hat{\theta} - \theta_0) \to N[0, A(\theta_0)^{-1} B(\theta_0) A(\theta_0)^{-1}] \quad \text{where} \quad S_n(\theta_0) \equiv \frac{\partial Q_n}{\partial \theta}(\theta_0) \quad \text{and}$$

$$H(\theta_0) = \frac{\partial^2 Q_n}{\partial \theta \partial \theta^T}(\theta_0).$$

Proof: Given in Appendix 1.

Condition (i) is stronger than condition (vii) in Theorem 1, and it is also stronger than the usual condition in time series data because spatial dependent data has more dimension correlations than time series data. It shows that how dependence decays when distance between groups gets further away, and the dependence decays at the rate fast enough. Condition (ii) just repeats the assumption in the Bernstein's blocking method, the two non-opposing directions just exclude sampling area grows at two parallel directions, which

does not make much sense in spatial dependent case. Conditions in (iii) are natural conditions about matrices, which are implied by the previous assumptions. Matrices are semidefinite if some extreme situations happen such as Pr (Y_{g1} = 1, Y_{g2} =1| X_g) = 0, which are assumed to be excluded in the previous assumptions.

2.4.3 Estimation of Variance-covariance Matrices

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

covariance estimators. However, it is difficult to get closed form expressions for $B_n(\hat{\theta})$ in practice, and hence we follow an alternative approach proposed by Conley (1999).

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

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

$$\alpha_{k,l}(n) \equiv \sup\{|P(A \cap B) - P(A)P(B)|\},\ A \in \Xi_{\Lambda_1}, B \in \Xi_{\Lambda_2} \quad \text{and}$$
$$|\Lambda_1| \le k,\ |\Lambda_2| \le l,\ \Upsilon(\Lambda_1, \Lambda_2) \ge n.$$

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

$$R_S(\theta) = \left\{ \begin{array}{l} S(\theta) \text{ if } W_S^* = 1, \\ 0 \quad \text{if } W_S^* = 0. \end{array} \right\}$$

Then

THEOREM 3. *If (i)* Λ_τ *grows uniformly in two non-opposing directions as* $\tau \to \infty$, *(ii)* $B(\theta_0) \equiv \lim_{n\to\infty} E[S_n(\theta_0) S_n^T(\theta_0)]$ *and* $A(\theta_0) \equiv \lim_{n\to\infty} -E[H(\theta_0)]$ *are uniformly positive definite matrices, (iii)* Y_{gi}, Y_{ji} *as defined in Theorem 1* , $i = 1,2$ *and*

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

$$\hat{B}_\tau - B(\theta_0) = o_p(1) \text{ as } \tau \to \infty$$

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

$$\hat{B}_T = n_T^{-1} \sum_{j=0}^{L_M} \sum_{k=0}^{L_P} \sum_{m=j+1}^{M} \sum_{p=k+1}^{P} K_{MP}(j,k) \left(\begin{array}{c} R_{m,p}\left(\hat{\theta}\right) R_{m-j,p-k}\left(\hat{\theta}\right)^T + \\ R_{m-j,p-k}\left(\hat{\theta}\right) R_{m,p}\left(\hat{\theta}\right)^T \end{array} \right)$$
$$- n_T^{-1} \sum_{m=1}^{M} \sum_{p=1}^{P} R_{m,p}\left(\hat{\theta}\right) R_{m,p}\left(\hat{\theta}\right)^T .$$

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

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

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

2.5 Simulation Study

In the previous section, we have proved that the partial maximum likelihood estimator (PMLE) based on the bivariate normal distribution is consistent and asymptotically normal. Moreover, one of the most attractive properties of our new PMLE is that we can get a more efficient estimator compared to the GMM estimator, and the approach is much less computational demanding when compared to full information methods. In order to learn about the gains in efficiency that we obtain in the context of a Bivariate Spatial Probit model when using PML versus GMM, we conduct in this Section a simulation study to show the efficiency gains of PML.

2.5.1 Simulation Design and Results

Instead of comparing our PMLE to the GMM estimator of Pinkse and Slade (1998) directly, we choose to compare the PMLE to the heteroskedastic Probit estimator (HPE) because of two reasons: First, the HPE uses similar information with the GMM estimator because both methods use generalized residuals from the Probit estimation to construct the moment conditions, which means that both methods use the information from the heterogeneities of the diagonal terms of the variance-covariance matrix, while our PMLE uses both diagonal and off-diagonal correlations information between two closest neighbors. Second, the STATA [8] source codes for bivariate probit estimation and heteroskedastic Probit estimation are available online, and we can easily add the spatial parts into these existing source codes to compare PML estimators with Heteroskedastic

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

Probit Estimators.

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

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

Therefore, it seems reasonable to do the following. Let R be the weighting matrix which can be generated in STATA[9] according to the distance between observations

$$Y_i^* = X_{i1}\beta_1 + X_{i2}\beta_2 + X_{i3}\beta_3 + \varepsilon_i \tag{58}$$
$$\varepsilon = \lambda R u, \tag{59}$$

where $u \sim Normal(0, I_n)$. The weighting matrix R is standardized so that the diagonal elements are ones, and then the elements of R shrink as distance is increasing. The reason we do this is because it is easier to determine $Var(\varepsilon_i)$ and $Cov(\varepsilon_i, \varepsilon_j)$ to apply the HP

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

and the bivariate probit estimators. In this way, we still allow general correlation across groups, and we are able to compare the efficiency gains from only using the diagonal information (the HP approach) to using both diagonal and off-diagonal information (bivariate probit), and we do not require to know the exact formula for the elements in Ω_g (given in equation (57)) to reach the same goal.

Therefore, we generate the dataset according to equations (58) and (59), which allows spatial correlation between any two observations, and we set the true parameter values for β_1, β_2 and β_3 equal to 1, 1 and 1 respectively. Since our main focus in this study is on the estimation of the spatial parameter λ, we also set different λ true values for each simulated sample: $\lambda = 0.2; 0.4; 0.6$ and 0.8, to test for the performance of the two estimation methods (PML and HP). These values for λ are in the range of the estimated value in the empirical application of Pinkse and Slade (1998). In this setting and with 1000 replications, we consider a sample size of $N = 1000$ observations (where the sample size is divided into 500 pairwise groups). Finally, we also simulate samples of sizes 500 and 1500 (with 250 and 750 pairwise groups respectively) to check the performance of the two methods in different samples sizes. The simulation results are reported in Tables 2.1 (for the spatial parameter λ) and Table 2.2 (for the β_1, β_2 and β_3) in Appendix 2.

From Table 2.2, we can observe that both the HPE and the PMLE of β_1, β_2 and β_3 converge to true parameter values across the different parameter values as sample size increasing. Also the PML estimator has much less bias than the HPE. Moreover, as expected, PML always provides smaller standard errors than the HP estimation method and bias and standard errors decrease in general when sample size increases.

Furthermore, it is in Table 2.1 where we can observe the largest advantages of using PML versus HP. We can see that the PMLE is much better than the HPE in terms of estimating the spatial parameter λ . The PMLE is always much closer to the true parameter values and with small standard errors across different sample sizes and parameter values (as expected from our theoretical results), while the HPE is much further away from true parameter values and it is has a much larger standard deviation over the different sample sizes, even though HPE also shows the trend to converge to the true values in general as the sample increases. The HPE has always much larger standard deviation than the PMLE, showing clearly the gains in efficiency of PML versus HPE/GMM as predicted by our theory. Since both the HPE and the GMM estimator use generalized residuals from Probit estimation to construct the moment conditions, we conjecture that the GMM estimator is subject to similar inefficiency problems in estimating the spatial coefficient. Also, as it is expected, the bias of the PMLE decreases when N increases.

In summary, from the simulation results of Tables 2.1 and Table 2.2, we see how the PMLE outperforms clearly the HPE (i.e., the GMM estimator of Pinkse and Slade (1998)), specially when estimating the spatial parameter λ , which implies that the PMLE is much more robust and efficient in the context of the spatial probit model. The simulation results provide clear evidence of the gains in efficiency that can be obtained by PML versus GMM, as predicted by our theoretical results in the previous section.

2.6 Conclusions

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

APPENDIX I

A.1 Proofs to Theorems

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

To prove consistency, the proof includes three parts:

(i) Q has a unique maximum at θ_0.

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

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

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

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

$$\frac{\partial \theta_n}{\partial \theta}(\hat{\theta}) = 0 = \frac{\partial \theta_n}{\partial \theta}(\theta_0) + \frac{\partial^2 Q_n}{\partial \theta \partial \theta^T}(\theta^*)(\hat{\theta} - \theta_0) \tag{60}$$

$$\Rightarrow (\hat{\theta} - \theta_0) = -[\frac{\partial^2 Q_n}{\partial \theta \partial \theta^T}(\theta^*)]^{-1} \frac{\partial Q_n}{\partial \theta}(\theta_0) \tag{61}$$

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

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

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

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

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

where

$$\frac{\partial^2 P_g(1,1)}{\partial\theta\partial\theta^T} = \frac{-1}{[\Pr(Y_{g1}=1,Y_{g2}=1\mid X_g)]^2}[\frac{\partial \Pr(Y_{g1}=1,Y_{g2}=1\mid X_g)}{\partial\theta}]^2 + \frac{1}{\Pr(Y_{g1}=1,Y_{g2}=1\mid X_g)}\frac{\partial^2[\Pr(Y_{g1}=1,Y_{g2}=1\mid X_g)]}{\partial\theta\partial\theta^T}, \tag{64}$$

and all other terms behave similar.

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

$$\frac{1}{n}\sum_{g=1}^{n}[Y_{g1}Y_{g2}\frac{\partial^2 P_g(1,1)}{\partial\theta\partial\theta^T}(\theta^*)]$$
$$= \frac{1}{n}\sum_{g=1}^{n} Y_{g1}Y_{g2}\{\frac{-1}{[\Pr(Y_{g1}=1,Y_{g2}=1\mid X_g)]^2}[\frac{\partial \Pr(Y_{g1}=1,Y_{g2}=1\mid X_g)}{\partial\theta}(\theta^*)]^2 \tag{65}$$
$$+\frac{1}{\Pr(Y_{g1}=1,Y_{g2}=1\mid X_g)}\frac{\partial^2[\Pr(Y_{g1}=1,Y_{g2}=1\mid X_g)]}{\partial\theta\partial\theta^T}(\theta^*)\}.$$

Look at the first term of the above equation given by

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

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

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

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

In order to prove

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

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

$$\varpi^T\{\frac{1}{n}\sum_{g=1}^{n} K_{g11}[\frac{\partial \Pr(Y_{g1}=1,Y_{g2}=1\mid X_g)}{\partial\theta}(\hat{\theta})]^2$$

$$-\frac{1}{n}\sum_{g=1}^{n} K_{g11}[\frac{\partial \Pr(Y_{g1}=1,Y_{g2}=1\mid X_g)}{\partial\theta}(\theta_0)]^2\}$$

$$=\frac{1}{n}\sum_{g=1}^{n}\overline{K_{g11}}\{[\frac{\partial \Pr(Y_{g1}=1,Y_{g2}=1\mid X_g)}{\partial\theta}(\hat{\theta})]^2-\frac{\partial \Pr(Y_{g1}=1,Y_{g2}=1\mid X_g)}{\partial\theta}(\theta_0)]^2\}$$

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

(above, equation (69), (70), (71))

From the proof of Theorem 1, we know that $\sup_g\left\|\dfrac{\partial \Pr(Y_{g1}=1,Y_{g2}=1|X_g)}{\partial\theta}\right\|<\infty$.

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

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

$$=o_p(1)$$

$$\Rightarrow \varpi^T\{\frac{1}{n}\sum_{g=1}^{n}K_{g11}[\frac{\partial \Pr(Y_{g1}=1,Y_{g2}=1\mid X_g)}{\partial\theta}(\hat{\theta})]^2-\frac{1}{n}\sum_{g=1}^{n}K_{g11}[\frac{\partial \Pr(Y_{g1}=1,Y_{g2}=1\mid X_g)}{\partial\theta}(\theta_0)]^2\}$$

$$=o_p(1)$$

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

(above, (72), (73), (74).

By definition,

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

and therefore,

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

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

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

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

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

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

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

If we define

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

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

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

Therefore, it remains to show the asymptotic normality of the score term: $\frac{\partial Q_n}{\partial\theta}(\theta_0)$.

For the sake of brevity, redefine the score as: $S_n(\theta_0) \equiv \frac{\partial Q_n}{\partial\theta}(\theta_0)$. Then

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

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

dependent processes. First, define $T_{na_n} \equiv \Pi_{j=}^{a_n}(1+i\gamma D_{n,j})$, where $i^2=-1$, and $D_{n,j}(j=1,2...a_n)$ is an array of random variables on the probability triple (Ω, F, P). γ is a real number. McLeish's (1974) central limit theorem for dependent processes requires the following four conditions

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

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

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

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

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

indices corresponding to the observations in area j . By assumption a number $C > 0$ exists such that $Max_j(\#\Lambda_{nj}) < Cb_n$. Define $D_{n,j} \equiv n^{-\frac{1}{2}} \Sigma_{t\in\Lambda_{nj}} A_{nt}$, and hence we can write $Y_{0n} = \Sigma_{j=1}^{a_n} D_{nj}$.

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

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

Since by assumption

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

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

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

Since $B(\theta_0)$ is positive definite, $B(\theta_0)^{-\frac{1}{2}}$ is bounded for sufficiently large n, and we have that $\sup_g \|Y_{gn}\| < \infty$ by assumption (vi) in Theorem 1. We have also proved that

$\sup_g \left\| \frac{\partial P_g(1,1)}{\partial \theta} \right\| < \infty$ in Lemma 2. Therefore, we are able to prove that $\sup \|A_{nt}\| < \infty$.

Then $Cb_n \times n^{-\frac{1}{2}} \sup \|A_{nt}\| = O_p(Cb_n \times n^{-\frac{1}{2}}) = o_p(1)$ by construction of b_n.

Hence we can get that $\underset{j \le a_n}{Max} |D_{n,j}| = o_p(1)$.

Second, let us discuss condition (i): $\{T_{na_n}\}$ is uniformly integrable. Following Davidson (1994), if a random variable is integrable, the contribution to the integer of extreme random variable values must be negligible. In other words, if $E|T_{na_n}| < \infty, E(|T_{na_n}| 1|T_{na_n}| > K) \to 0$, as $K \to \infty$, it is equivalent to say $P[\sup_{n>N} |T_{na_n}| > K] = 0$, for some $K > 0$ as $n \to \infty$. Here we follow the proof of Lemma 10 in Pinkse and Slade (1998). We have that

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

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

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

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

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

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

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

(92)

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

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

(93)

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

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

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

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

by construction of Y_{0n}, since $E(Y_{0n}^2) = 1$. It remains to show that

$\sum_{i \neq j}^{a_n} E(D_{ni} D_{nj}) = o(1)$. This condition is proved in Lemmas 5-7[10]. *Q.E.D.*

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

A.2 Technical Lemmas

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

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

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

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

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

It is easy to see that

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

where $\gamma_{ng1}=[P_g(1,1)-P_g(1,0)-P_g(0,1)+P_g(0,0)], \gamma_{ng2}=[P_g(1,0)-P_g(0,0)]$, and

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

Note that here

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

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

Therefore, we can write

$$Sup_{ngj} |Var[Q_n(\theta)]| =$$

$$\frac{1}{n^2}\sum_{g=1}^{n}\sum_{j=1}^{n}\{4\,\mathrm{cov}(Y_{g1}Y_{g2}, Y_{j1}Y_{j2}) + 8\,\mathrm{cov}(Y_{g1}Y_{g2}, Y_{j1}) + 8\,\mathrm{cov}(Y_{g1}Y_{g2}, Y_{j2}) \tag{98}$$
$$+ 4\,\mathrm{cov}(Y_{g1}, Y_{j1}) + 8\,\mathrm{cov}(Y_{g1}, Y_{j2}) + 4\,\mathrm{cov}(Y_{g2}, Y_{j2}).$$

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

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

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

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

Following Pinkse and Slade (1998), we also use the Bernstein's (1927) blocking method to prove this as follows. We split the region in which observations are located up to

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

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

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

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

(102)

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

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

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

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

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

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

By the mean value theorem

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

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

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

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

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

and note that $P_g(Y_{g1}=1,Y_{g2}=1 \mid X_g) = \Phi 2(\frac{X_{g1}\beta}{\sqrt{\Omega_{g11}}}, \frac{X_{g2}\beta}{\sqrt{\Omega_{g22}}}, \rho_g \mid X_g)$, where Φ_2 is the bivariate normal distribution function. Also

$$\frac{\partial P_g(1,1)}{\partial \theta^T} = \frac{\partial[\log \Phi 2(\frac{X_{g1}\beta}{\sqrt{\Omega_{g11}}}, \frac{X_{g2}\beta}{\sqrt{\Omega_{g22}}}, \rho)]}{\partial \theta^T}. \tag{108}$$

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

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

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

$$\frac{\partial P_g(1,1)}{\partial \beta^T} = \frac{\partial[\log \Phi 2(\frac{X_{g1}\beta}{\sqrt{\Omega_{g11}}}, \frac{X_{g2}\beta}{\sqrt{\Omega_{g22}}}, \rho)]}{\partial \beta^T} = \frac{\frac{s_{g1}X_{g1}}{\sqrt{\Omega_{g11}}} + \frac{s_{g2}X_{g2}}{\sqrt{\Omega_{g22}}}}{\Phi 2(\frac{X_{g1}\beta}{\sqrt{\Omega_{g11}}}, \frac{X_{g2}\beta}{\sqrt{\Omega_{g22}}}, \rho_g)}, \tag{110}$$

with

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

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

By assumption (v)

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

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

Therefore,

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

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

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

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

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

provided that $\sup_g \|W_g\| < \infty$.

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

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

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

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

LEMMA 3 *Under the assumptions in Theorem 2,*

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

Proof: From Lemma 2, we know that

$$\frac{\partial \Pr(Y_{g1}=1,Y_{g2}=1\mid X_g)}{\partial \beta^T}$$

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

$$\Rightarrow\frac{\partial^2 \Pr(Y_{g1}=1,Y_{g2}=1\mid X_g)}{\partial\beta\partial\beta^T}$$

$$=\frac{X_{g1}\phi(\frac{X_{g1}\beta}{\sqrt{\Omega_{g11}}})\{X_{g1}\frac{X_{g1}\beta}{\sqrt{\Omega_{g11}}}\Phi[\frac{\frac{X_{g2}\beta}{\sqrt{\Omega_{g22}}}-\rho\frac{X_{g1}\beta}{\sqrt{\Omega_{g11}}}}{\sqrt{1-\rho_g^2}})]+\phi[\frac{\frac{X_{g2}\beta}{\sqrt{\Omega_{g22}}}-\rho\frac{X_{g1}\beta}{\sqrt{\Omega_{g11}}}}{\sqrt{1-\rho_g^2}})]\frac{\frac{X_{g2}}{\sqrt{\Omega_{g22}}}-\rho\frac{X_{g1}}{\sqrt{\Omega_{g11}}}}{\sqrt{1-\rho_g^2}})}{\sqrt{\Omega_{g11}}}$$

$$+\frac{X_{g2}\phi(\frac{X_{g2}\beta}{\sqrt{\Omega_{g22}}})\{X_{g2}\frac{X_{g2}\beta}{\sqrt{\Omega_{g22}}}\Phi[\frac{\frac{X_{g1}\beta}{\sqrt{\Omega_{g11}}}-\rho\frac{X_{g2}\beta}{\sqrt{\Omega_{g22}}}}{\sqrt{1-\rho_g^2}})]+\phi[\frac{\frac{X_{g1}\beta}{\sqrt{\Omega_{g11}}}-\rho\frac{X_{g2}\beta}{\sqrt{\Omega_{g22}}}}{\sqrt{1-\rho_g^2}})]\frac{\frac{X_{g1}}{\sqrt{\Omega_{g11}}}-\rho\frac{X_{g2}}{\sqrt{\Omega_{g22}}}}{\sqrt{1-\rho_g^2}})}{\sqrt{\Omega_{g22}}}$$

(above, (118) and (119))

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

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

$$\frac{\partial \Pr(Y_{g1}=1,Y_{g2}=1\mid X_g)}{\partial\lambda}$$

$$= \frac{\phi_2(\frac{\varepsilon_{g1}}{\sqrt{\Omega_{g11}}},\frac{\varepsilon_{g2}}{\sqrt{\Omega_{g22}}},\rho_g)}{\Phi_2(\frac{X_{g1}\beta}{\sqrt{\Omega_{g11}}},\frac{X_{g2}\beta}{\sqrt{\Omega_{g22}}},\rho_g)} \times \frac{\partial\phi_2(\frac{\varepsilon_{g1}}{\sqrt{\Omega_{g11}}},\frac{\varepsilon_{g2}}{\sqrt{\Omega_{g22}}},\rho_g)}{\partial\lambda}, \tag{121}$$

$$\Rightarrow \frac{\partial^2 \Pr(Y_{g1}=1,Y_{g2}=1\mid X_g)}{\partial\lambda^2}$$

$$= \frac{(\frac{\partial\phi_2}{\partial\lambda})^2(\Phi_2-\phi_2\phi_2)}{(\Phi_2)^2} + \frac{\phi_2}{\Phi_2}\frac{\partial^2\phi_2}{\partial\lambda^2}. \tag{122}$$

It is easy to see that the first term of the above equation is bounded from previous results (i.e. $\sup_g \left\| \frac{\partial\phi_2}{\partial\lambda} \right\| < \infty$) and the second term can be also proved bounded since $\frac{\partial^2\phi_2}{\partial\lambda^2}$ can be proved to be bounded given that $\sup_g \|W_g\| < \infty$ after some algebra. Hence

$$\sup_g \left\| \frac{\partial^2 \Pr(Y_{g1}=1,Y_{g2}=1|X_g)}{\partial\theta\partial\theta^T} \right\| < \infty. \quad Q.E.D.$$

LEMMA 4 *Under the assumptions in Theorem 2,* $ET_{na_n} - 1 = o(1)$, where

$$T_{na_n} \equiv \Pi_{j=1}^{a_n}(1 + i\gamma D_{n,j}).$$

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

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

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

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

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

We know that

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

(above, (125) and (126)).

First, we look at the term $E[D_{nj} T_{Rnj}]$. Since $T_{Rnj} \equiv \Pi_{k\notin\Xi_{nj-1}}(1 + i\gamma D_{n,k})$, that means the group is not adjacent to group j. By Bernstein's method, we split the region

in such a way that the distance between group j and non-adjacent group is at least $b_n^{\frac{1}{2}}$.

Hence, $Max \mid E[D_{nj}T_{Rnj}] \mid = Max \mid \mathrm{cov}(D_{nj}, T_{Rnj}) = \alpha(\sqrt{b_n})$ provided $E(D_{nj}) = 0$ and by assumption (vi) in Theorem 1. By construction of a_n and b_n, $\alpha(\sqrt{b_n})a_n = o(1)$, and hence we obtain $Max \mid E[D_{nj}T_{Rnj}] \mid = o(a_n^{-1})$.

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

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

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

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

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

Proof: We know that

$\Sigma_{i\neq j}^{a_n} E(D_{ni}D_{nj}) = \Sigma_{i=1}^{a_n}\Sigma_{j=1}^{a_n} E(D_{ni}D_{nj}) - \Sigma_{i=j}^{a_n} E(D_{ni}D_{nj}) = o(1)$ if we can

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

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

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

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

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

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

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

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

$$\leq Max_{i \neq j} C_1 n^{-1} \sum_{s \in \Lambda ni, t \in \Lambda nj} \alpha(d_{ts}) \tag{131}$$

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

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

$$Max_{i\notin j}C_1 n^{-1} \sum_{s\in\Lambda ni, t\in\Lambda nj} \alpha(d_{ts}) \le C_3 Max_{i\notin j} n^{-1}\sqrt{b_n} \sum_{d=0}^{C_4\sqrt{b_n}} d^2\alpha(d), \quad (132)$$

where $C_3 = C_1C_2, C_4 > 0$.

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

$$C_3 Max_{i\notin j} n^{-1}\sqrt{b_n} \sum_{d=0}^{C_4\sqrt{b_n}} d^2\alpha(d) = o(n^{-1}b_n). \quad (133)$$

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

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

Proof:

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

$$Max \sum_{l=2}^{\sqrt{a_n}} \sum_{j \in \Xi_{nil}} |E(D_{ni}D_{nj})| \leq C_5 Max \sum_{l=2}^{\sqrt{a_n}} \#\Xi_{nil} n^{-1} \times \#\Lambda_{ni} \times \#\Lambda_{nj} \alpha(\sqrt{b_n}(l-1)) \quad (134)$$

$$\leq C_6 n^{-1} b_n^2 l \sum_{l=1}^{\sqrt{a_n}} \alpha(\sqrt{b_n} l) = o(n^{-1} b_n l \sum_{l=1}^{\sqrt{a_n}} \alpha(l) = o(n^{-1} b_n) \quad (135)$$

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

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

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

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

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

We have that

$$\sum_{i=1}^{a_n} \sum_{j=1}^{a_n} Cov(D_{n,i}^2, D_{n,j}^2) = \sum_{i=1}^{a_n} \sum_{j=1}^{a_n} \{[D_{n,i}^2 - E(D_{n,i}^2)][D_{n,j}^2 - E(D_{n,j}^2)]\} \quad (138)$$

$$\leq C_7 \sum_{l=0}^{C_8\sqrt{a_n}} (l+1)\alpha(\sqrt{b_n} l) Max E(D_{ni}^4), \quad (139)$$

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

$$MaxE(D_{ni}^4) \le n^{-2} Max \sum_{t1,t2,t3,t4 \in \Lambda_{nj}} |E[A_{nt1}, A_{nt2}, A_{nt3}, A_{nt4}] \tag{140}$$

$$\le C9 n^{-2} Max_j \sum_{t1,t2,t3,t4 \in \Lambda_{nj}} \{\alpha(d_{t1,t2}) + ... + \alpha(d_{t3,t4})\} \tag{141}$$

$$\le C10 n^{-2} Max_j \sum_{t1,t2 \in \Lambda_{nj}} \{\alpha(d_{t1,t2})\} \tag{142}$$

$$\le C11 n^{-2} b_n^2 Max_j \sum_{t1 \in \Lambda_{nj}} \sum_{l=0}^{c_{12}\sqrt{b_n}} l\alpha(l) = O(n^{-2} b_n^3), \tag{143}$$

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

$$C7 \sum_{l=0}^{C_8\sqrt{a_n}} (l+1)\alpha(\sqrt{b_n} l) MaxE(D_{ni}^4) = O(n^{-2} b_n^3 a_n) = o(1), \tag{144}$$

because $a_n b_n = n$ and $n^{-1} b_n^2 \to 0$ as $n \to \infty$. *Q.E.D.*

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

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

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

Theorem 2, we have already proved that

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

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

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

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

and we can rewrite it as

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

For the sake of brevity, we redefine

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

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

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

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

Therefore,

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

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

As before, we just need to provide the proof for one of these terms, and the same logic applies to other terms. We consider the most complicated term and the rest follow the same argument

$$\begin{aligned} & n^{-1} \sum_{g=1}^{n} \sum_{j=1}^{n} [\psi_{ng1}\psi_{nj1} Cov(Y_{g1}Y_{g2}, Y_{j1}Y_{j2})] \\ &= n^{-1} \sum_{g=1}^{n} \sum_{j=1}^{n} \psi_{ng1}\psi_{nj1} [E(Y_{g1}Y_{g2}Y_{j1}Y_{j2}) - E(Y_{g1}Y_{g2})E(Y_{j1}Y_{j} \end{aligned} \tag{153}$$

$$E(Y_{g1}Y_{g2}Y_{j1}Y_{j2}) = \Pr(Y_{g1} = 1, Y_{g2} = 1, Y_{j1} = 1, Y_{j2} = 1 \mid X_g) \tag{154}$$

$$= \Phi_4(y_{g1}, y_{g2}, y_{j1}, y_{j2}, \rho_{12}, \rho_{13}, \rho_{14}, \rho_{23}, \rho_{24}, \rho_{34}) \tag{155}$$

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

$$E(Y_{g1}Y_{g2}) = \Pr(Y_{g1}=1, Y_{g2}=1 \mid X_g) = \Phi_2(y_{g1}, y_{g2}, \rho_{12}), \quad (156)$$
$$E(Y_{j1}Y_{j2}) = \Pr(Y_{j1}=1, Y_{j2}=1 \mid X_g) = \Phi_2(y_{j1}, y_{j2}, \rho_{34}), \quad (157)$$

and therefore,

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

so we can write the first term as

$$B_n(\theta_0) = n^{-1} \sum_{g=1}^{n} \sum_{j=1}^{n} \psi_{ng1}\psi_{nj1}[E(Y_{g1}Y_{g2}Y_{j1}Y_{j2}) - E(Y_{g1}Y_{g2})E(Y_{j1}Y_{j2})] \quad (160)$$

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

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

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

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

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

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

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

and the same argument applies to

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

so that

$$n^{-1}(\hat{\theta}-\theta_0)\sum_{g=1}^{n}\sum_{j=1}^{n}\psi_{ng1}\psi_{nj1}$$

$$\{\frac{\partial\Phi 4(y_{g1},y_{g2},y_{j1},y_{j2},\rho_{12}(\theta^*),\rho_{13}(\theta^*),\rho_{14}(\theta^*),\rho_{23}(\theta^*\hat{\theta}),\rho_{24}(\theta^*\hat{\theta}),\rho_{34}(\theta^*)}{\partial\theta}$$

$$-\frac{\partial\Phi 4(y_{g1},y_{g2},y_{j1},y_{j2},\rho_{12}(\theta^*),0,0,0,0,\rho_{34}(\theta^*)}{\partial\theta}\}\to 0 \qquad (167)$$

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

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

APPENDIX II

TABLE 2.1: Simulation Results of Different Estimators of lambda in the Context of the Bivariate Spatial Probit Model*

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

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

TABLE 2.2: Simulation Results of Different Estimators of betas in the Context of the Bivariate Spatial Probit Model*

			$\beta_1 = 1$		$\beta_2 = 1$		$\beta_3 = 1$	
			HPE	PMLE	HPE	PMLE	HPE	PMLE
$\lambda = 0.2$	$N = 500$	mean	5.322	2.618	5.333	2.619	5.329	2.623
		(s.d.)	(8.844)	(0.839)	(8.872)	(0.855)	(8.863)	(0.870)
	$N = 1000$	mean	5.308	2.616	5.296	2.616	5.289	2.618
		(s.d)	(7.612)	(0.560)	(7.570)	(0.560)	(7.568)	(0.564)
	$N = 1500$	mean	5.247	2.604	5.239	2.602	5.235	2.604
		(s.d.)	(6.624)	(0.540)	(6.606)	(0.536)	(6.613)	(0.543)
$\lambda = 0.4$	$N = 500$	mean	3.610	1.329	3.614	1.329	3.608	1.328
		(s.d.)	(5.305)	(0.362)	(5.311)	(0.365)	(5.290)	(0.366)
	$N = 1000$	mean	3.600	1.318	3.593	1.316	3.588	1.315
		(s.d.)	(4.192)	(0.355)	(4.177)	(0.355)	(4.178)	(0.353)
	$N = 1500$	mean	3.456	1.281	3.441	1.281	3.438	1.278
		(s.d.)	(3.818)	(0.342)	(3.793)	(0.343)	(3.798)	(0.339)
$\lambda = 0.6$	$N = 500$	mean	2.898	0.972	2.876	0.966	2.885	0.969
		(s.d.)	(3.761)	(0.271)	(3.723)	(0.268)	(3.735)	(0.271)
	$N = 1000$	mean	2.669	0.981	2.669	0.979	2.657	0.978
		(s.d.)	(2.951)	(0.261)	(2.953)	(0.261)	(2.916)	(0.259)
	$N = 1500$	mean	2.508	1.016	2.499	1.015	2.501	1.016
		(s.d.)	(2.726)	(0.250)	(2.706)	(0.250)	(2.708)	(0.253)
$\lambda = 0.8$	$N = 500$	mean	2.246	0.805	2.237	0.801	2.249	0.802
		(s.d.)	(2.810)	(0.373)	(2.803)	(0.373)	(2.841)	(0.392)
	$N = 1000$	mean	2.098	0.843	2.096	0.843	2.082	0.843
		(s.d.)	(2.281)	(0.349)	(2.279)	(0.349)	(2.246)	(0.340)
	$N = 1500$	mean	2.086	0.884	2.096	0.886	2.094	0.886
		(s.d.)	(2.059)	(0.316)	(2.071)	(0.314)	(2.073)	(0.318)

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

BIBLIOGRAPHY

Andrews, D. W. K. (1991). Heteroskedasticity and autocorrelation consistent covariance matrix estimation. Econometrica: 59, 3, 817-858.

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

Anselin, L. (1988). Spatial econometrics: methods and models. Kluwer Academic Publishers.

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

Anselin, L. Florax, R.J.G.M, and Rey, J.S. (2004). Econometrics for spatial models: recent advances, in Advances in spatial econometrics. Springer-Verlag, Berlin, Germany,1-28.

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

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

Case, A.C. (1991). Spatial patterns in household demand. Econometrica: 59, 953-965.

Case, A.C. (1992). Neighborhood influence and technology change. Regional Science and Urban Economics 22, 491-508.

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

Davidson, J. (1994). Stochastic limit theory. Oxford: Oxford University Press.

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

Gourieroux, C. (2000). Econometrics of qualitative dependent variables. Cambridge University Press.

Greene, W.H. (2003). Econometrics analysis. 4th Edition, Prentice-Hall, Upper Saddle River, N.J.

Harvey, A. (1976). Estimating regression models with multiplicative heteroscedasticity. Econometrica : 44, 461-465.

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

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

Kotz, S. Balakrishnan, N. and Johnson, N. (2000). Continuous multivariate distributions, 2nd Edition. Wiley Series in Probability and Statistics.

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

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

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

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

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

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

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

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

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

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

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

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

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

White, H. (2001). Asymptotic theory for econometricians, 2nd Edition. Orlando, FL. Academic Press.

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

MICHIGAN STATE UNIVERSITY LIBRARIES
3 1293 03062 6224