HIGH-DIMENSIONAL INFERENCE FOR SPATIAL ERROR MODELS

By

Liqian Cai

A DISSERTATION

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

Statistics—Doctor of Philosophy

2016

.

ABSTRACT

HIGH-DIMENSIONAL INFERENCE FOR SPATIAL ERROR MODELS

By

Liqian Cai

In the literature of econometric theory and application, issues relating to urban, real estate, agricultural, and environmental economics, etc., where the data are collected spatially from cross-sectional units, are common and in these circumstances, the spatial relation among the sampling sites can not be ignored. Spatial autocorrelation is thus introduced to model the correlation among values of a single variable strictly attributable to their relatively close locational positions on a two-dimensional surface, which extends autocorrelation in time series to spatial dimensions.

With the growth of computer capabilities, databases are becoming progressively larger and more complex, making traditional statistical methods less effective or sometimes even unsuitable. Data from high-frequency economic transactions, detailed macroeconomic data collected by a multitude of sources with varying data quality and varying sampling frequencies, and data on large economic and social networks are just a few examples of the content of enormous databases that are now subject to thorough examination.

This dissertation discusses applicable (high-dimensional) variable selection and estimation methods and corresponding theories focusing on a spatial error model where the spatial autocorrelation comes from the disturbances across cross-sectional units, in a regression context.

In the first part, we propose a generalized Lasso-type of estimator for the spatial error model, where the disturbance terms are autocorrelated across cross-sectional units. We

further prove the estimation consistency and selection sign consistency of the parameter estimator under both the low dimensional setting when the dimension of the parameter p is fixed and smaller than the sample size n, as well as the high dimensional setting when p is greater than and can be growing with n. The number of non-zero components of the parameter in both settings are considered relatively smaller than the number of observations (sparsity).

In the second part, we continue to investigate post-model selection estimators that apply estimation to the model selected by first-step variable selection. We show that by separating the model selection and estimation process, the post-model selection estimator can perform at least as well as the simultaneous variable selection and estimation method in terms of the rate of convergence. The convergence rate of the estimation error in both the ℓ_2 and *sup* norms are studied. Moreover, under perfect model selection, that is, when the selection process is able to correctly identify the significant covariates of the true model with probability goes to 1, the oracle convergence rate can be reached.

In the last part, a sketch of the future work on high-dimensional analysis on mixed regressive, spatial autoregressive model, where the response unit depends not only on the explanatory variables but also on the response from its neighboring units, is described.

To my beloved parents and my foundation, Hengyi Cai and Hehua Zhong, for their endless love, support and encouragement.

ACKNOWLEDGMENTS

First and foremost, I would like to express my sincerest gratitude to my major dissertation advisor Dr. Tapabrata Maiti, for his patient guidance, invaluable assistance, constructive comments and immense knowledge. His continuous support and encouragement helped me through the research and writing process of this dissertation.

Special thanks to Dr. Roger Calantone, for providing me with a precious experience working on real life business data and learning about practical methodologies and also for serving as one of my committee members. Special thanks to Dr. Arnab Bhattarcharjee, for the countless discussions and useful suggestions on my work.

Thanks goes to Dr. Chae Young Lim and Dr. Ping-Shou Zhong, for serving as members of my guidance committee and providing me with constructive advices for my dissertation.

Thanks goes to the entire faculty and staff members in the Department of Statistics and Probability who have taught me and helped me during my study at Michigan State University. Thanks goes to the graduate school, the College of Natural Science and the Department of Statistics and Probability who provided me the Dissertation Continuation Fellowship, Dissertation Completion Fellowship and traveling fellowships for working on my research and attending academic conferences.

I would also like to thank my academic family members at Department of Statistics and Probability for all the time we have had in the past years. Knowing all of you make even the toughest days enjoyable.

Last but not the least, I would like to thank my beloved parents, Hengyi Cai and Hehua Zhong, for giving birth to me, loving me and supporting me to become who I want to be at every stage of life.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

Chapter 1

Introduction

1.1 Spatial Econometric Models

Spatial econometrics is a subfield of econometrics which deals with the spatial correlation among geographical units. The units, depending on the nature of the problem, can refer to zip codes, regions, cities, states and so on. Applied work relating to transportation, house pricing, agriculture growth, etc relies heavily on sampled data that is collected from different locations. The geographical units do not need to be confined as the concrete physical locations in space, they can also be used to explain the abstract interaction between economic agents, and a good illustration will be the connection through social networks.

What separates spatial econometrics from traditional econometrics is that sampled data with a location component involves two issues that traditional econometrics has mostly ignored, spatial dependence between the observations and spatial heterogeneity in the modeling relationships. Spatial dependence comes from the fact that observation from location i is dependent on the values of observation j, where i, j can be any sample location. And spatial heterogeneity refers to the the situation when we expect a different regression relationship for every location in sample space. From the Gauss-Markov assumption used in regression, the explanatory variables are fixed in repeated sampling and a constant variance exists for data from different locations. Thus these two issues violate the Gauss-Markov assumptions and alternative modeling procedures are needed here to address the issues.

One way to deal with the spatial effect is to impose the spatial structure onto the non-spatial regression model. Starting with the standard linear regression model, which takes the form

$$Y = X\beta + \varepsilon,$$

where Y is a $n \times 1$ vector consisting of observations of the dependent variable in each sample, X represents an $n \times p$ matrix of explanatory variables, β is the corresponding parameter vector of interest and ε is the disturbance vector with independently and identically distributed error terms. The general linear regression model is commonly estimated by ordinary least squares estimator. However, when spatial interaction effects exist, a spatial econometrics model can be constructed by adding different combinations of interaction effects to the linear model. Typically, we think the association of an observation at a specific location with observations at other locations come from three sources: the endogenous interaction effects among the dependent variable (Y), the exogenous interaction effects among the independent variables (X), and the interaction effects among the error terms (ε). A full model containing all types of spatial effects can be presented as

$$Y = \delta W_1 Y + X\beta + W_2 X\theta + u,$$

$$u = \rho W_3 u + \varepsilon$$

where W_1Y, W_2X, and W_3u denote the spatial interaction effects among the dependent variable, independent variable and error terms, respectively. Here W_1, W_2, W_3 are $n \times n$ spatial weight matrices used to describe the spatial arrangement of the geographical units in the sample and they may or may not be identical. δ and ρ denote the spatial autoregressive

parameters and very often the parameters are assumed to be within the range of -1 and 1, just like in a time-series model. The parameter describes the clustering (positive ρ) or dissimilarity (negative ρ) in space of certain random variable. Of the two types of spatial autocorrelation, positive autocorrelation is by far the more intuitive. Negative spatial autocorrelation implies a checker- board pattern of values and does not always have a meaningful substantive interpretation. Even though the structure of the spatial econometrics models mentioned above look very alike the ones used in time series, it is important to be aware that spatial econometrics is not a straightforward extension of time series to two dimensions. In time series, the focus is on the dependence among observations over time and each observation is only correlated with the observations from the past, while in spatial econometrics, more attention is paid on the spatial dependence among observations across space, which is multi-dimensional and there is no nature ordering for the data arrangement, so spatial econometrics methodologies can not be a direct transposition from time-series.

The spatial interaction effects among different locations, wherever the source, are all quantified by the spatial weight matrix. The spatial weight matrix, usually denote W, is a nonnegative matrix of known constants. The elements in the matrix are decided mainly from two sources of information. The first one is the latitude and longitude of the location in Cartesian space. The coordinates of the location provide the distance of any two locations. And based on the fundamental theorem of regional science, observations that are nearer will reflect a greater degree of spatial dependence than those more distant from each other. This suggests use of functions of distance between location i and j as element w_{ij} in the matrix. The other source of information is the contiguity, which represents the relative position in space of one observation to other observations. From the nature of the size and shape of the observation units, we can determine the definition of neighbors, and neighboring units

should exhibit a higher degree of spatial dependence than units located further apart. The neighbors in the spatial matrix can be differed by elements 1 and 0. And the diagonal elements are always set to zero, assuming no spatial unit can be viewed as its own neighbor.

The full model mentioned above incorporates all interaction effects, yet in real applications, models that contain fewer sources of interaction effects can be obtained by imposing restrictions on one or more of the parameters. And theoreticians are mainly focusing on some of the mostly used models. To start with, models with only interaction effects among the error terms are called linear regression model with spatial autoregressive disturbance, also known as, spatial error models. The model is specified as

$$Y = X\beta + u, \quad u = \rho M u + \varepsilon,$$

where ε has zero mean and variance $\sigma^2 I$. In this regression model, the disturbances $u_i's$ in u are following a spatial autoregressive process, and are correlated to each other across units. Another popular type of model with only endogenous interaction effects on the dependent variable is called mixed regressive, spatial autoregressive model

$$Y = \delta W Y + X\beta + \varepsilon,$$

where $\varepsilon_i's$ are independently identically distributed with zero mean and variance σ^2. This model differs from a usual Spatial Autoregressive Process in the presence of exogenous regressors X as explanatory variables in the model. There has also been a growing interest in models containing more than just one spatial interaction effect. A lot of econometric problems use the model which combines endogenous interaction effects and interaction effects

among the error terms

$$Y = \delta WY + X\beta + u, \quad u = \rho Mu + \varepsilon,$$

where $\varepsilon_i's$ are independently identically distributed with zero mean and variance σ^2. The estimation methods for the spatial econometric models, which have been considered in the existing literature, are mainly the (Quasi-) Maximum Likelihood method, Instrumental Variable method, Generalized Method of Moments or Bayesian Markov Chain Monte Carlo methods. The (Quasi-) Maximum Likelihood method assumes the error term ε to be normal and permits the actual distribution to be different from normal distribution. It has good finite sample properties with one order spatial lag. However, it is not computationally attractive for large sample size problems because of the estimation complexities of spatial autoregressive parameters. The Monte Carlo methods come to use for the computational challenge. The Instrumental Variable method and Generalized Method of Moments are feasible for higher spatial lag models and they do not assume the normality of error term ε. Plus, the Generalized Methods of Moments is also computationally feasible and asymptotically consistent under an explicit set of conditions.

1.2 Variable Selection in High-dimensional Setting

In statistical research, we are continuously dealing with the problem of building a model using a collection of potentially relevant predictors for the purpose of forecasting a response of interest. And variable selection serves a fundamental role in identifying the relevant predictors that truly makes a contribution to the response. The main goals of variable selection

are to simplify the prediction models in order to make them easier to interpret, to shorten the training times, to enhance generalization by reducing overfitting, as well as hopefully to construct an improved estimation method. Nowadays, with the development of scientific research and advanced technology, the collection of vast quantities of data becomes possible and increasingly easy. Sometimes the dimension of the attributes collected becomes so large, maybe even larger than the sample size, then it becomes a high dimensional statistical problem. Examples of high-dimensional data can be frequently seen in high-frequency economic transactions, genomics, high-resolution images, among others. The good thing is, when dealing with high-dimensional data problem, we make the assumption that the final regression function lies in a low dimensional manifold, and the regression parameter vector is sparse with many of the components being zero, which is not only reasonable but also makes high dimensional statistics inference possible.

When the attention focuses on identifying the significant predictors, criteria are needed to select a manageable subset model. In the linear model context, the earliest developments of variable selection were based on attempts to minimize the mean squared error of the prediction with different adjustments depending on the goal of modeling. One of the most familiar criteria is Mallow's C_p criterion, where $C_p = \frac{SSE_p}{MSE_{full}} - (n - 2p)$, here C_p considers the ratio of SSE for p variable model to MSE for full model, then penalizes for the number of variables. Two other popular criteria, motivated from very different points of view, are AIC (a.k.a Akaike Information Criterion) and BIC (a.k.a Bayesian Information Criterion). Letting $\log L$ denote the maximum log likelihood of the candidate model with k-dimension parameters, AIC selects the model which minimizes $2k - 2\log L$, whereas BIC selects the model which minimizes $k \log n - 2\log L$. Traditional variable selection criteria is a specific form of penalized likelihood, providing a unified framework for comparison. However, in the

high dimensional setup when the dimensionality becomes comparable to or even larger than the sample size, computational and inferential challenges are significant.

There has been evolving amount of literature working on techniques that are capable of reducing the high dimensionality of the variable as well as producing optimal estimators. Approaches to cope with high dimensionality is usually the penalized regression methods. Consider the linear regression model

$$Y = X\beta + \varepsilon.$$

Suppose we have n observations indexed by i, and for each observation, one response variable y_i, along with p features $\{x_{i1}, \cdots, x_{ip}\}$ are observed. Typically, this type of linear relation between Y and X can be easily solved by least squares estimators. However, it becomes unfeasible when the dimension of the features p becomes larger than the sample size n. Penalized regression methods can now be used, which penalize the model fitting with various regularization terms to encourage model sparsity, by minimizing an objective function Q with a generalized form

$$Q(\beta) = \frac{1}{n} L(\beta|X, Y) + P_\lambda(\beta),$$

which consists of a loss function L and a penalty term P. The penalty term P_λ is indexed by a regularization parameter λ that controls the level of penalty of the objective function Q. Typically, the penalty function P has the following properties: it is symmetric about the origin, $P(0) = 0$, and P is nondecreasing in $||\beta||$.One of the most popular has been the Least Absolute Shrinkage and Selection Operator, as known as, the Lasso method, first proposed

by Tibshirani (1996). The Lasso estimator is estimated by minimizing the objective function

$$Q_{Lasso}(\beta) = \frac{1}{n}||Y - X\beta||_2^2 + \lambda||\beta||_1,$$

hoping to simultaneously select variables and estimate the associated regression coefficients. Since the L_1 norm of the estimator is controlled by the penalty term, a portion of values in $\hat{\beta}$ will be reduced to exact zero while minimizing the squared loss. Ever since the development of Lasso estimator, much progress has been made in understanding its statistical properties. The development of Lasso remedies the disadvantages of an earlier regularization method, by Hoerl and Kennard (1970), known as the ridge regression. They proposed the objective function as follows

$$Q_{Ridge}(\beta) = \frac{1}{n}||Y - X\beta||_2^2 + \lambda||\beta||_2^2.$$

Ridge regression heavily penalizes large coefficients, leading to biased estimates when some of the coefficients are large. And also, ridge regression does not produce sparse solutions and thus fails to improve the interpretability of the model.

1.3 Overview

There have already been rich literature working on high-dimensional variable selection in the linear regression set up, but still not much has been talked about for spatial econometric models. High-dimensional problems also arise from time to time in economics, for example, survey data which contains hundreds or thousands of variables may only have a few that actually relate to the response of interest; house price data which contains all cross-sectional effects of geographic neighbors may only have significant relation with a few

neighbors nearby. Recently, Belloni and his group have worked on a series of papers focusing on high dimensional variable selection methods for sparse econometrics models, with application focused on instrumental variable. However, they do not consider the possible spatial interaction effects that might be involved in the model, and there are not much references to high-dimensional variable selection for spatial econometric models.

So we fill in the gap here. In Chapter 2, we will introduce a generalized variable selection and estimation method for a regression model with spatial autoregressive error, the basic spatial econometric model. Additionally, we will develop the parameter consistency and model selection consistency for the estimator in both the low dimensional setting when the dimension of the parameter p is fixed and smaller than the sample size n as well as the high dimensional setting when p is greater than and can be growing with n.

In Chapter 3, we continue with the same model we talked about in Chapter 2 and investigate post-model selection estimators that apply least squares estimation to the model selected by first-step penalized estimation. We manage to show that by separating the model selection and estimation process, the post-model selection estimator can perform just as well as the simultaneous variable selection and estimation method in terms of the rate of convergence. And it can strictly outperform the simultaneous variable selection and estimation estimator when the selection process is able to correctly identify the significant covariates of the true model with large probability.

In Chapter 4, we will extend the work to mixed regressive, spatial autoregressive model, where the endogenous interaction effects among the dependent variable are considered, and related future work is discussed.

Chapter 2

Variable Selection With Spatial Autoregressive Errors

2.1 Introduction

In the literature of econometric theory and application, issues relating to urban, real estate, agricultural, and environmental economics, etc., where the data are collected spatially from cross-sectional units are common and in these circumstances, the spatial relation among the sampling sites can not be ignored. Thus in 1973, Cliff and Ord first put forth a spatial autoregressive model (also known as SAR) to model the spatial autocorrelation of the disturbances across cross-sectional units in a regression context. This model extends autocorrelation in time series to spatial dimensions and is a variant of the model suggested in Whittle (1954). In this Spatial model, the disturbance term corresponding to a cross-sectional unit is modeled as a weighted average of disturbances corresponding to other cross-sectional units, plus an innovation. To be more precise, the disturbance u_n is written as

$$u_n = \rho M_n u_n + \varepsilon_n.$$

And the regression model with SAR disturbance u_n is specified as

$$Y_n = X_n\beta + u_n.$$

The subscript n indicates the sample size. The term $M_n u_n$ is often referred as "spatial lag". Typically the innovations ε_n are assumed to be i.i.d with mean 0 and variance σ^2 and the parameter of interest is ρ, σ^2 and β. For now, we assume the $n \times n$ spatial weight matrix M_n is known. Contrary to time-series models which are associated with uni-directional time flow, the spatial data can be viewed as multi-directional, with each location correlating with all the other locations nearby in every direction. Because of this particular characteristic of spatial processes, a simple transposition of time-series methodologies can not be applied.

Since the introduction of the spatial autoregressive model, several methods have been developed for estimating the regression coefficients for spatial models with spatial autoregressive error. To summarize, the most widely known methods with theoretical basis are the (Quasi-) Maximum Likelihood method (Ord, 1975, Smirnov and Anselin, 2001), and the methods of moments (Kelejian and Prucha, 1999). The Quasi-Maximum Likelihood method allows for the case when the actual likelihood function differs to some extent from the normal distribution assumed. One obstacle of the ML method in practice is its huge computational burden, since the maximization of the log-likelihood involves a nonlinear optimization that requires the evaluation of the determinant of dimension $n \times n$, where n is the size of the data set for each value of the autoregressive parameter ρ used. Common computational approaches to this problem is the use of eigenvalues of the spatial weights matrix, as done in Ord (1975), but the computation of the eigenvalues quickly becomes numerically unstable for more than 1000 observations, or the solution can be the use of Monte Carlo estimation

to approximate as well as bound the determinant proposed by Barry and Pace (1999). In Smirnov and Anselin (2001), a new method for evaluating the Jacobian term which is based on the characteristic polynomial of the spatial weights matrix M_n is introduced and this algorithm can approach linear computational complexity. However, the simulation method remains an approximation and cannot yield the theoretical properties. Compared with the computational difficulty of ML method, Kelejian and Prucha (1999) proposed an alternative estimator for the spatial autoregressive parameter ρ and variance parameter σ^2 in the spatial autoregressive error model based on a generalized moments approach which is computationally simple irrespective of sample size. Further, the conditions needed do not involve the assumption of normality. This estimator of ρ is also proved to be consistent, thus can be treated as a nuisance parameter and the asymptotic properties of the regression parameter β solved based on the estimated ρ can retain all the good properties of the OLS for the model where ρ is assumed to be known.

In the regression context, often we find ourselves in the face of need to identify the important factors in order to explain certain phenomena. Current days, high dimensional statistical problems arise from diverse fields of scientific research and technological development. Here, high dimensional data refers to the general case of growing dimensionality and ultra-high dimensional which specifies the case where the dimensionality grows at a non-polynomial rate as the sample size increases. Example of high-dimensional data includes but not limited to: high-resolution images, microarray or proteomics data, high-frequency financial data and gene data (Fan and Lv, 2010). And as a result of the wide availability of inexpensive global positioning systems and other advances in technology, the collection of vast quantities of data with geo-referenced sample locations becomes possible and the models for spatially correlated data become increasingly important. Sometimes the number

of attributes collected becomes so large, maybe even larger than the sample size, and this makes it impossible to conduct the standard estimation method that we discussed earlier. The good thing is, we believe that among all the information we collect, many of them do not have significant impact on the subject variable we are interested in, thus the p-dimensional regression parameters are assumed to be sparse with majority of the components being zero. Recently, this kind of high-dimensional variable selection problem has drawn great attention and many mechanisms for linear regression models have been discovered. Among all, one of the most popular has been the least absolute shrinkage and selection operator, a.k.a., the Lasso method first introduced by Tibshirani (1996) and much progress has been made in understanding the statistical properties ever since. For example, in Knight and Fu (2000), the problem of the asymptotic distribution of Lasso-type estimators is studied in the low-dimensional setting where the dimension of regression p is smaller than the sample size n and is fixed. Later in Zhao and Yu (2006), an almost necessary and sufficient Irrepresentable Condition for Lasso is constructed to select the true model consistently both in the fixed p setting and in the large p setting when p can grow as the sample size n gets larger. Other results concerning the asymptotic properties of the Lasso can be found in the Meinshausen and B$\ddot{u}$hlmann (2006), Bickel et al (2009) and B$\ddot{u}$hlmann and van de Geer (2011), among others.

Econometrics are also in the need of tools to deal with huge amounts of data as computers are getting more involved in the middle of economic transactions. Lasso and its penalized regression estimators can be computed quite efficiently and are providing good predictions in practice (Varian, 2014). Recently, Belloni and his group introduce high dimensional variable selection methods for sparse econometrics models, with application focused on instrumental variable. The series of papers include Belloni and Chernozhukov (2011), Belloni,

Chernozhukov and Wang (2011), Belloni, Chen, Chernozhukov and Hansen (2012), and Belloni and Chernozhukov (2013). However, all these variable selection methods assume that the error in a regression model is independent, which is not the case in the spatial autoregressive model context. We argue a variable selection method under the independent error assumption, e.g. standard Lasso, may not perform well for spatially dependent data.

The literature regarding the theoretical results on asymptotic properties of the Lasso estimator with spatial autoregressive errors is very limited. So we fill the gap here. We combine the spirit of the Lasso method and the generalized method of moments for spatial autoregressive error models, and develop a generalized Lasso estimator which performs the variable selection and estimation simultaneously for the regression parameter β in a two-stage process. Also, we use the consistency property of the estimator for model parameter ρ and the fact that it is a nuisance parameter (Kelejian and Prucha, 1999), to prove that the asymptotic properties of the Lasso estimator of β remain valid even when the model parameter ρ is replaced by its moment estimator. Both the parameter consistency and model sign consistency of the estimator are addressed. Here, parameter consistency refers to the asymptotic property that as the sample size n goes to infinity, the resulting sequence of estimates converges in probability to the true parameter value, that is,

$$\hat{\beta}^n \to \beta, \quad as \quad n \to \infty$$

And an estimate is model sign consistent if and only if the probability that the sign of each component of the estimator equals to that of the true parameter converges to one, that is,

there exists $\lambda_n = f(n)$, a function of n and independent of Y_n or X_n such that

$$\lim_{n\to\infty} P(\hat{\beta}^n(\lambda_n) =_s \beta^n) = 1.$$

The rest of the chapter is organized as follows. Section 2 below describes the procedure to construct a generalized Lasso estimator in order to select and estimate the nonzero components of the regression parameter in a spatial autoregressive error setup. Section 3 discusses the asymptotic properties of the estimator which includes the parameter consistency and model sign consistency in the set up where the dimension of parameter p is fixed and smaller than the sample size n. Section 4 tackles the asymptotic properties of the estimator as in section 3 in the high dimension setting when p can be growing with n. Section 5 provides the simulation studies of the performance of the estimator for different choices of parameter ρ in the spatial autoregressive error model with proper selection of the penalty parameter λ_n, which will be defined later. Section 6 illustrates data example of Aveiro-Ílhavo urban housing market in Portugal where the proposed method can be applied. Additionally, all the proofs of lemmas and theorems in detail are relegated to the section 7.

2.2 A generalized moments LASSO (GLASSO) estimator

In this section, we propose a two-stage estimation procedure which combines GMM and LASSO estimation at the same time. We will focus on the simple spatial model where the

error term is assumed to be spatially autoregressive:

$$\begin{aligned} Y_n &= X_n\beta + u_n, \\ u_n &= \rho M_n u_n + \varepsilon_n, \end{aligned} \tag{2.1}$$

where Y_n is the $n \times 1$ vector of observations on the dependent variable, X_n is the $n \times p$ matrix of observations on the explanatory variables, β is the $p \times 1$ vector of unknown model parameters, and u_n is the vector of spatial autoregressive errors with spatial autoregressive parameter ρ, a scalar parameter, M_n is a spatial weighting matrix, a $n \times n$ matrix of known constants, and ε_n is an $n \times 1$ vector of idiosyncratic errors.

For generality, we permit the elements of M_n and ε_n to depend on n. We make several standard assumptions as follows:

Assumption 1. *For all n, the idiosyncratic errors $\varepsilon_1, \varepsilon_2, \cdots, \varepsilon_n$ are independently and identically distributed with zero mean and positive bounded variance σ^2. Additionally, we assume $E(\varepsilon_1^4) < \infty$.*

Assumption 2. *M_n is an exogenous $n \times n$ matrix. All diagonal elements of M_n are zero, $|\rho| < 1$ and the matrix $I - \rho M_n$ is nonsingular for all $|\rho| < 1$.*

M_n is a spatial weights matrix whose elements defines the relationship between different units. In a cross-sectional setting, if the ith and j th units are not related, we can set $m_{ij} = m_{ji} = 0$ where m_{ij} is the (i, j)th element of M_n. Often, M_n is set as a contiguity (adjacency) matrix, in which case the non-zero off-diagonal elements are symmetric and have unit value. In other cases, the elements may reflect economic or geographic distances between the units, in which case they are non-negative and symmetric; M_n can still be asymmetric if it is considered in row-standardized form. In other modeling contexts, for example Bailey et

al. (2016), the matrix can be symmetric but the elements can assume values $\{-1, 0, 1\}$. In yet other contexts, the weights can be asymmetric and without any sign or other restrictions, beyond the conditions in Assumption 2; see, for example, Bhattacharjee et al. (2016). In Assumption 2, $|\rho| < 1$ is a stability (spatial granularity) condition, and the invertibility of the matrix $I - \rho M_n$ is to ensure identification in reduced form, that is, the error vector u_n is uniquely defined in terms of the idiosyncratic error vector ε_n, as $(I - \rho M_n)^{-1}\varepsilon_n$. These assumptions are standard; see for example, Kelejian and Prucha (1999) and Lee (2004).

The first step of our estimation procedure is to obtain a generalized moments estimator of ρ. The estimation process follows the same method as Kelejian and Prucha (1999), and we outline this below for convenience of exposition. Let $\tilde{u}_n$ be a predictor for u_n. Further, let $\bar{u}_n = M_n u_n$ and $\bar{\bar{u}}_n = M_n M_n u_n$, and correspondingly, $\tilde{\bar{u}}_n = M_n \tilde{u}_n$, and $\tilde{\bar{\bar{u}}}_n = M_n M_n \tilde{u}_n$. Similarly, let $\bar{\varepsilon}_n = M_n \varepsilon_n$. Then, under Assumptions 1 and 2:

$$E[\frac{1}{n}\varepsilon_n'\varepsilon_n] = \sigma^2 \qquad E[\frac{1}{n}\bar{\varepsilon}_n'\bar{\varepsilon}_n] = \sigma^2 n^{-1} Tr(M_n' M_n) \qquad E[\frac{1}{n}\bar{\varepsilon}_n'\varepsilon_n] = 0 \tag{2.2}$$

The spatial autoregressive parameter ρ is included in the above moments equations through the expression $\epsilon_n = u_n - \rho\bar{u}_n$. Thus the equations can be used to obtain a generalized moments estimator for ρ. From Equation (2.1), and Equation (2.2), we obtain

$$\Gamma_n[\rho, \rho^2, \sigma^2]' - \gamma_n = 0. \tag{2.3}$$

Here

$$\Gamma_n = \begin{bmatrix} \frac{2}{n}E(u_n'\bar{u}_n) & \frac{-1}{n}E(\bar{u}_n'\bar{u}_n) & 1 \\ \frac{2}{n}E(\bar{\bar{u}}_n'\bar{u}_n) & \frac{-1}{n}E(\bar{\bar{u}}_n'\bar{\bar{u}}_n) & \frac{1}{n}Tr(M_n'M_n) \\ \frac{1}{n}E(u_n'\bar{\bar{u}}_n + \bar{u}_n'\bar{u}_n) & \frac{-1}{n}E(\bar{u}_n'\bar{\bar{u}}_n) & 0 \end{bmatrix},$$

$$\gamma_n = \begin{bmatrix} \frac{1}{n}E(u_n'u_n) \\ \frac{1}{n}E(\bar{u}_n'\bar{u}_n) \\ \frac{1}{n}E(u_n'\bar{u}_n) \end{bmatrix}$$

Now if we consider the sample moments based on $\tilde{u}_n$, and use these to replace the moments of u_n shown above, similar to Equation (2.3), we get the equation :

$$G_n[\rho, \rho^2, \sigma^2]' - g_n = \nu_n(\rho, \sigma^2), \tag{2.4}$$

where

$$G_n = \begin{bmatrix} \frac{2}{n}\tilde{u}_n'\tilde{\bar{u}}_n & \frac{-1}{n}\tilde{\bar{u}}_n'\tilde{\bar{u}}_n & 1 \\ \frac{2}{n}\tilde{\bar{\bar{u}}}_n'\tilde{\bar{u}}_n & \frac{-1}{n}\tilde{\bar{\bar{u}}}_n'\tilde{\bar{\bar{u}}}_n & \frac{1}{n}Tr(M_n'M_n) \\ \frac{1}{n}(\tilde{u}_n'\tilde{\bar{\bar{u}}}_n + \tilde{\bar{u}}_n'\tilde{\bar{u}}_n) & \frac{-1}{n}\tilde{\bar{u}}_n'\tilde{\bar{\bar{u}}}_n & 0 \end{bmatrix}$$

$$g_n = \begin{bmatrix} \frac{1}{n}\tilde{u}_n'\tilde{u}_n \\ \frac{1}{n}\tilde{\bar{u}}_n'\tilde{\bar{u}}_n) \\ \frac{1}{n}\tilde{u}_n'\tilde{\bar{\bar{u}}}_n \end{bmatrix}$$

The 3×1 vector $\nu_n(\rho, \sigma^2)$ can be viewed as a vector of residuals, and the GMM estimator for ρ and σ^2 can be defined as the nonlinear least squares estimator, $\hat{\rho}_n$ and $\hat{\sigma}_n^2$, which minimizes

the norm of the residual vector. Specifically,

$$(\hat{\rho}_n, \hat{\sigma}_n^2) = \arg\min_{\rho,\sigma^2} \left[G_n[\rho, \rho^2, \sigma^2]' - g_n \right]' \left[G_n[\rho, \rho^2, \sigma^2]' - g_n \right]. \tag{2.5}$$

Several additional assumptions are required to obtain the asymptotic properties of the GMM estimator.

Assumption 3. *The row and column sums of M_n and $(I - \rho M_n)^{-1}$ are bounded uniformly in absolute value. Note that the bound for $(I - \rho M_n)^{-1}$ may depend on ρ.*

Assumption 4. *Let $\tilde{u}_{i,n}$ denote the i th element of $\tilde{u}_n$, we assume that there exist (finite dimensional) random vectors d_{in} and Δ_n such that $|\tilde{u}_{i,n} - u_{i,n}| \leqslant \|d_{in}\|\|\Delta_n\|$ with $n^{-1}\sum_{i=1}^n \|d_{in}\|^{2+\delta} = O_p(1)$ for some $\delta > 0$ and $n^{\frac{1}{2}}\|\Delta_n\| = O_p(1)$.*

Assumption 5. *The smallest eigenvalue of $\Gamma_n'\Gamma_n$ is bounded away from zero, that is,*

$$\lambda_{min}(\Gamma_n'\Gamma n) \geqslant \lambda_* > 0,$$

where λ_* may depend on ρ and σ^2.

For a discussion of these assumptions, we refer to Kelejian and Prucha (1999). Given Assumption 1 to 5, the nonlinear least squares estimators $\hat{\rho}_n$ and $\hat{\sigma}_n^2$ defined in Equation (2.5) are consistent estimator of ρ and σ^2, that is, $\hat{\rho}_n \to_p \rho$ and $\hat{\sigma}_n^2 \to_p \sigma^2$ as $n \to \infty$ (Kelejian and Prucha, 1999). Let us now focus on the context of a spatial regression model whose errors are autoregressive. It is easy to see that, if ρ were known, we could rewrite model (2.1) as

$$(I - \rho M_n)Y_n = (I - \rho M_n)X_n\beta + \varepsilon_n.$$

Then, LASSO variable selection and estimation of β can be conducted using the L_1 penalized

least squares criterion

$$(Y_n - X_n\beta)'\Sigma_n(\rho)(Y_n - X_n\beta) + \lambda_n \sum_{j=1}^{p} |\beta_j|,$$

where $\Sigma_n(\rho) = (I - \rho M_n)'(I - \rho M_n)$; for a given penalty λ_n, we denote this estimator as $\hat{\beta}_L(\rho)$. Of course, in practical applications ρ is typically unknown, and thus the direct LASSO estimator defined above is infeasible. In this case, we may replace ρ by the generalized moments estimator $\hat{\rho}_n$, and propose a feasible generalized moments LASSO (GLASSO) estimator $\hat{\beta}_L(\hat{\rho}_n)$ for model (2.1) in the second step of the estimation process. To be specific,

$$\hat{\beta}_L(\hat{\rho}_n) = \arg\min_{\beta}(Y_n - X_n\beta)'\Sigma_n(\hat{\rho}_n)(Y - X_n\beta) + \lambda_n \sum_{j=1}^{p} |\beta_j|. \tag{2.6}$$

The above function can be numerically optimized using the package "glmnet" in R developed by Friedman et al. (2010). The glmnet algorithms use cyclical coordinate descent, which optimizes the objective function over each parameter successively while keeping others fixed, with the cycles repeating until convergence. The tuning parameter λ_n is chosen by cross-validation with a certain lower bound inferred from the theoretical results discussed below.

2.3 Asymptotic Properties for fixed p and q

In this section, we consider the asymptotic behavior of the generalized moments LASSO estimator (2.6) under the setting when p (the dimension of all candidate covariates) and q (the dimension of covariates with non-zero coefficients) are both finite and fixed and smaller than the sample size n; that is, $q << p < n$. We show that under the classical setting

mentioned above, our proposed GLASSO estimator achieves consistency in terms of both parameter estimation and model selection.

2.3.1 Parameter Consistency

In the following theorem, we show that when the tuning parameter λ_n grows at a rate slower than n, the proposed estimator $\hat{\beta}_L(\hat{\rho}_n)$ achieves the parameter consistency and if we add more control on the growth rate of λ_n, the asymptotic normality of the estimator can also be derived. We need one more regularity condition:

Assumption 6. *The elements of X_n are nonstochastic and uniformly bounded in absolute value. The matrix $C(\rho) = \lim_{n\to\infty} \frac{1}{n} X_n' \Sigma(\rho) X_n$ is finite and nonsingular for all $|\rho| < 1$ and $\frac{1}{n} \max_{1\leqslant i\leqslant n} z_i z_i' \to 0$, where z_i is the ith row of the matrix $(I - \rho M_n)X_n$.*

The assumption is justified since the parametrization of the linear model $(I-\rho M_n)Y_n = (I-\rho M_n)X_n\beta + \varepsilon_n$, which is a transformation of (2.1) is unique if the matrix $C_n = \frac{1}{n} X_n' \Sigma(\rho) X_n$ is nonsingular for all n, and we further assume that $C(\rho)$ is nonsingular. The nonstochastic design matrix assumption can be relaxed and is assumed here for explanation simplicity. In fact, the results in this section can also hold quite generally for random designs. If X_n is a random design matrix, the asymptotic results still apply as long as the probability of the set for Assumption 6 to hold is 1. Similar extension can be seen in Zhao and Yu (2006), see also Bühlmann and van de Geer (2011).

For all the X_n matrices we are considering here, along with those we will consider in later chapters, in general, the elements of the design matrix are uncorrelated with those of the disturbance vector u_n. However, even if the design matrix is correlated with the disturbance, an instrumental matrix that is independent of the disturbance, denote H_n for example, can be used to instrument X_n. By regressing each column of X_n on H_n, and replace the endogenous

covariates with the predicted value from the regression, the problem is transformed back to a high-dimensional spatial error model with uncorrelated design matrix and disturbance.

Theorem 2.1. *Under Assumptions 1 to 6, if* $\lambda_n/n \to 0$*, the generalized moments LASSO estimator* $\hat{\beta}_L(\hat{\rho}_n)$ *is a consistent estimator for* β*. That is,* $\hat{\beta}_L(\hat{\rho}_n) \to_p \beta, \quad as \quad n \to \infty$*. If we assume further that* $\lambda_n/\sqrt{n} \to \lambda_0 \geqslant 0$*, then*

$$\sqrt{n}(\hat{\beta}_L(\hat{\rho}_n) - \beta) \longrightarrow_D \arg\min(V(w)),$$

where $V(w) = -2w'U + w'C(\rho)w + \lambda_0 \sum_{j=1}^{p}[w_j sgn(\beta_j)I(\beta_j \neq 0) + |w_j|I(\beta_j = 0)]$*, and* $U \sim N(0, \sigma^2 C(\rho))$.

The above theorem establishes parameter consistency of the GLASSO estimator in the setting where both the dimension of all covariates p and the dimension of non-zero covariates q are fixed and smaller than the sample size n. Further, if we control the rate of convergence of the penalty parameter λ_n in a specific way, the estimator achieves asymptotic normality towards the minimizer of a function $V(w)$. In the function $V(w)$, w is a $p \times 1$ vector, U is a $p \times 1$ random vector with normal distribution, and $C(\rho)$, defined in Assumption 6, involves the spatial parameter ρ and spatial weight matrix M_n. Specifically, if the tuning parameter λ_n grows to infinity at a slower rate than the square root of n, we have a nice result. Compared with the asymptotic properties of the naive LASSO estimator in the linear model setting, here we have spatial correlation. We find that the spatial autoregressive parameter ρ is involved in the asymptotic distribution of the GLASSO estimator and controls the convergence rate; if $\rho = 0$, the asymptotic distribution reduces to the same as that for the naive LASSO.

2.3.2 Sign Consistency

Above, we have shown parameter consistency of our generalized moments LASSO (GLASSO) estimator $\hat{\beta}_L(\hat{\rho}_n)$. However, a consistent estimator does not necessarily consistently select the correct model. Here, we may have a large number of irrelevant predictors, even in the low dimensional settings, and our primary goal is to correctly identify those which are relevant so that the final model will not only fit well but also be easily interpretable. So another property we desire is the model selection consistency of the estimation, which requires that

$$P(\{i : \hat{\beta}_i \neq 0\} = \{i : \beta_i \neq 0\}) \to 1, \quad as \quad n \to \infty.$$

Thus, we follow the idea of Zhao and Yu (2006) and achieve the result through sign consistency of the estimator, in which case,

$$sign(\hat{\beta}_L(\hat{\rho}_n)) = sign(\beta),$$

where $sign(\cdot)$ maps positive entry to 1, negative entry to -1 and zero to zero. We denote the above sign consistency condition as

$$\hat{\beta}_L(\hat{\rho}_n) =_s \beta.$$

Note that sign consistency is stronger than model selection consistency, in the sense that, if our estimator is sign consistent, then the model selection consistency condition is automatically satisfied. Further, sign consistency avoids the undesirable situation that the model is estimated only with zeros matched but reversed signs for some of the relevant covariates.

Notation : Assume $\beta = (\beta_1, \cdots, \beta_q, \beta_{q+1}, \cdots, \beta_p)'$ where $\beta_j \neq 0$ for $j = 1, \cdots, q$ and $\beta_j = 0$ for $j = q+1, \cdots, p$. Let $\beta(1) = (\beta_1, \cdots, \beta_q)'$ and $\beta(2) = (\beta_{q+1}, \cdots, \beta_p)'$, and for any p-column matrix Z, write $Z(1)$ and $Z(2)$ as the first q and final $p-q$ columns of Z respectively. Define $C^n(\rho) = \frac{1}{n}[(I - \rho M_n)X_n]'[(I - \rho M_n)X_n]$. By setting $C^n_{11}(\rho) = \frac{1}{n}[(I - \rho M_n)X_n](1)'[(I - \rho M_n)X_n](1)$, $C^n_{22}(\rho) = \frac{1}{n}[(I - \rho M_n)X_n](2)'[(I - \rho M_n)X_n](2)$, $C^n_{12}(\rho) = \frac{1}{n}[(I - \rho M_n)X_n](1)'[(I - \rho M_n)X_n](2)$, and $C^n_{21}(\rho) = \frac{1}{n}[(I - \rho M_n)X_n](2)'[(I - \rho M_n)X_n](1)$, we can express $C^n(\rho)$ as follows:

$$C^n(\rho) = \begin{pmatrix} C^n_{11}(\rho) & C^n_{12}(\rho) \\ C^n_{21}(\rho) & C^n_{22}(\rho) \end{pmatrix}.$$

For the same reason as Assumption 6, here we assume that C^n_{11} is invertible based on the uniqueness of the parametrization of the first relevant q covariates. Since $\hat{\rho}_n$ is a consistent estimator of ρ, the invertibility of $C^n_{11}(\hat{\rho}_n)$ is inherited from that of $C^n_{11}(\rho)$ when the sample size is large enough. This can hold even for the high dimension case when $p > n$ and Assumption 6 does not hold. In the rest of the paper, we will use the notation C^n to denote $C^n(\hat{\rho}_n)$ unless specified otherwise.

The following proposition places a lower bound on the probability of LASSO picking the true model which quantitatively relates to the probability of LASSO selecting the correct model. This is a modification of the Proposition 1 in Zhao and Yu (2006).

Proposition 2.1. *Assume that* $|C^n_{21}(C^n_{11})^{-1} sign(\beta(1))| \leqslant 1 - \eta$ *holds with a constant* $\eta > 0$, *where the inequality holds element-wise. Then,*

$$P(\hat{\beta}_L(\hat{\rho}_n; \lambda) =_s \beta) \geqslant P(A_n \cap B_n)$$

for

$$A_n = \{|(C_{11}^n)^{-1}W^n(1)| < \sqrt{n}(|\beta(1)| - \frac{\lambda_n}{2n}|(C_{11}^n)^{-1}sign(\beta(1))|)\}$$

$$B_n = \{|C_{21}^n(C_{11}^n)^{-1}W^n(1) - W^n(2)| \leqslant \frac{\lambda_n}{2\sqrt{n}}\eta\}$$

where

$$W^n(1) = \frac{1}{\sqrt{n}}[(I - \rho M_n')^{-1}\Sigma(\hat{\rho}_n)](1)'\varepsilon_n$$

and

$$W^n(2) = \frac{1}{\sqrt{n}}[(I - \rho M_n')^{-1}\Sigma(\hat{\rho}_n)X_n](2)'\varepsilon_n$$

In order to prove Proposition 2.1 and the following Theorem 2.2, we need the following lemma which is a direct consequence of the Karush-Kuhn-Tucker conditions:

Lemma 2.1. $\hat{\beta}^n(\lambda) = (\hat{\beta}_1^n, \cdots, \hat{\beta}_j^n, \cdots)$ *are the LASSO estimates defined by*

$$\hat{\beta}^n(\lambda) = \arg\min_{\beta} ||Y_n - X_n\beta||_2^2 + \lambda||\beta||_1$$

if and only if

$$\frac{d||Y_n - X_n\beta||_2^2}{d\beta_j}|_{\beta_j=\hat{\beta}_j^n} = -\lambda sign(\hat{\beta}_j^n) \quad for \quad j \quad such\ that \quad \hat{\beta}_j^n \neq 0$$

$$\left|\frac{d||Y_n - X_n\beta||_2^2}{d\beta_j}\right|_{\beta_j=\hat{\beta}_j^n} \leqslant \lambda \quad for \quad j \quad such\ that \quad \hat{\beta}_j^n = 0.$$

In our context, the generalized moments LASSO estimator $\hat{\beta}_L(\hat{\rho}_n)$ is defined to minimize $(Y_n - X_n\beta)'\Sigma(\hat{\rho}_n)(Y_n - X_n\beta) + \lambda_n||\phi||_1$ for some λ_n for all ϕ, where $\hat{\rho}_n$ is the GMM estimator

of the parameter ρ in (2.1). Hence, applying Lemma 2.1 in our case, we have

$$\left|\frac{d||(I-\hat{\rho}_n M_n)Y_n-(I-\hat{\rho}_n M_n)X_n\beta||_2^2}{d\beta_j}\right|_{\beta_j=\hat{\beta}_{Lj}(\hat{\rho}_n)} = -\lambda_n sign(\hat{\beta}_{Lj}(\hat{\rho}_n)) \quad ,$$
$$for \quad j \quad \text{such that} \quad \hat{\beta}_{Lj}(\hat{\rho}_n)\neq 0,$$

$$\left|\frac{d||(I-\hat{\rho}_n M_n)Y_n-(I-\hat{\rho}_n M_n)X_n\beta||_2^2}{d\beta_j}\right|_{\beta_j=\hat{\beta}_{Lj}(\hat{\rho}_n)} \leqslant \lambda_n \quad ,$$
$$for \quad j \quad \text{such that} \quad \hat{\beta}_{Lj}(\hat{\rho}_n)= 0,$$

where $\hat{\beta}_{Lj}(\hat{\rho}_n)$ is the jth element of the estimator $\hat{\beta}_L(\hat{\rho}_n)$. With this result, we are now able to prove the Proposition 2.1. The proof follows Zhao and Yu (2006) with appropriate adjustments to our case.

Recall that in this section, we are only focusing on the classical setting where q, p,and β are all fixed as $n \to \infty$. Under the above conditions and assumptions, we have the following result about sign consistency of our proposed GLASSO estimator $\hat{\beta}_L(\hat{\rho}_n)$.

Theorem 2.2. *For fixed q, p, and β, under Assumptions 1-6, the generalized moments LASSO estimator is sign consistent if the condition $|C_{21}^n(C_{11}^n)^{-1}sign(\beta(1))| \leqslant 1-\eta$ holds. That is, for every λ_n that satisfies $\lambda_n/n \to 0$ and $\lambda_n/n^{\frac{1+c}{2}} \geqslant r$ for any $r>0$ with $0 \leqslant c < 1$, we have*

$$P(\hat{\beta}_L(\hat{\rho}_n) =_s \beta^n) = 1 - o\left(s(\rho)e^{\frac{-n^c}{s^2(\rho)}}\right).$$

From the above result, it is clear that the convergence rate for the estimation method to choose the correct model is a bounded function of the spatial parameter ρ times the

exponential of a function of n and s . Here, $s(\rho)$ is the bound for the diagonal elements of $C_{11}^{-1}\sigma^2$ and $C_{22} - C_{21}C_{11}^{-1}C_{12}\sigma^2$. Because of the spatial structure added to the linear model, the convergence rate is affected. While the convergence rate in the i.i.d case is related only to n, now this depends also on ρ. One remark here is that, for Theorem 2.2, the effect of the spatial correlation to the estimator in the form of a function of ρ can instead be applied to the penalty parameter λ_n as a lower bound. In this way, additional information can be used for the choice of λ_n besides cross-validation.

2.4 Asymptotics for large p and q

In the previous section, we proved parameter consistency and sign consistency, as well as the asymptotic distribution, of our generalized moments LASSO estimator $\hat{\beta}_L(\hat{\rho}_n)$ as $n \to \infty$ under the classical setting where p, q, and β are all fixed, and p and q are smaller than n. The setting is simplified in the sense that it is natural to assume the regularity conditions as stated in Assumption 6:

$$C = \lim_{n\to\infty} \frac{1}{n} X_n' \Sigma(\rho) X_n$$

where C is finite and nonsingular.

However, in practice, there are many situations where large p and thus q are needed; it can either be larger than the sample size n or increase at some rate as n. In the large p and q case, we allow the dimension of the designs C^n grow and model parameter β change as n grows, that is, $p = p_n$ and $q = q_n < n$ and $\beta = \beta_n$. Consequently, the assumptions and regularity conditions in the previous sections are not appropriate since C^n may no longer converge and $\beta = \beta_n$ may change as n grows. Towards this situation, we first prove an oracle inequality for the generalized moments LASSO when the design is non-random; this in turn

will imply consistency as well. Then, in the second part of this section, we also prove that with high probability we can correctly select the model in the case that $p > n$.

2.4.1 Parameter Consistency

In this section, we prove that, with an appropriate choice of λ_n , the generalized moments LASSO estimator $\hat{\beta}_L(\hat{\rho}_n)$ obeys the following oracle inequality with a probability that can be made arbitrarily close to unity. That is, for large enough n, the condition

$$\frac{||(I - \hat{\rho}_n M_n)X_n(\hat{\beta} - \beta)||_2^2}{n} + \lambda_n||\hat{\beta} - \beta||_1 \leqslant \frac{4\lambda_n^2 s_0}{\phi_0^2}$$

is satisfied with an arbitrarily large probability. The inequality provides a bound for $||\hat{\beta}-\beta||_1$, and thus the estimator is consistent if the bound converges to zero. Here, β is the true value of the unknown parameter, $\lambda_n = O\left(\frac{\log p}{n}\right)$, we denote the GLASSO estimator by $\hat{\beta}$ for notational simplicity, s_0 is the cardinality of the set of nonzero components of β, and S_0 and ϕ_0 are constants depending on the design matrix X_n.

By the definition of the generalized moments estimator:

$$\hat{\beta} := \arg\min_{\phi} \left\{(Y_n - X_n\phi)'\Sigma(\hat{\rho}_n)(Y_n - X_n\phi) + \lambda_n||\phi||_1\right\}.$$

Since $\hat{\beta}$ provides the minima of this penalized objective function, we have the inequality below with change of scale of λ_n :

$$\frac{||(I - \hat{\rho}_n M_n)(Y_n - X_n\hat{\beta})||_2^2}{n} + \lambda_n||\hat{\beta}||_1 \leqslant \frac{||(I - \hat{\rho}_n M_n)(Y_n - X_n\beta)_2^2}{n} + \lambda_n||\beta||_1$$

Rearranging terms and using the triangle inequality, we obtain our Basic Inequality:

$$\frac{||(I-\hat{\rho}_n M_n)X_n(\hat{\beta}-\beta)||_2^2}{n}+\lambda_n||\hat{\beta}||_1 \leqslant \frac{2\epsilon_n'(I-\rho M_n')^{-1}\Sigma(\hat{\rho}_n)X_n(\hat{\beta}-\beta)}{n}+\lambda_n||\beta||_1. \quad (2.7)$$

Note that the first term on the RHS of the Basic Inequality (2.7) can be easily bounded in terms of the L_1-norm of parameters involved:

$$2\left|\varepsilon_n'(I-\rho M_n')^{-1}\Sigma(\hat{\rho}_n)X_n(\hat{\beta}-\beta)\right| \leqslant \left(\max_{1\leqslant j\leqslant p} 2|\varepsilon_n' T^{(j)}|\right)||\hat{\beta}-\beta||_1$$

where $T^{(j)}$ is the jth column of the matrix $T=(I-\rho M_n')^{-1}\Sigma(\hat{\rho}_n)X_n$.

Next, we introduce the set

$$\Im := \left\{\max_{1\leqslant j\leqslant p} 2|\epsilon_n' T^{(j)}|/n \leqslant \lambda_0\right\}$$

where we arbitrarily assume that $2\lambda_0 \leqslant \lambda_n$ to make sure that on $\Im$ we can get rid of the random part of the problem. Now, we have the following result.

Proposition 2.2. *Suppose Assumptions 1-5 hold, and further assume all the elements of X_n are nonstochastic and uniformly bounded in absolute value, then for all $t>0$, if we define*

$$\lambda_0 = 2\sigma(\exp[t^2/2]+1)\sqrt{\frac{log2p}{n}},$$

we have

$$P(\Im) \geqslant 1-K\exp[-t^2/2].$$

for some positive constant K.

Proof is given in the last section. Since we are in a situation where p is growing with

n, and possibly $p > n$, we generally consider the fact that only a few, say s_0, of the β_j are non-zero. To quantify the sparsity of the true β^0, we denote

$$S_0 := \{j : \ \beta_j^0 \neq 0\},$$

so that $s_0 = |S_0|$. In the literature, S_0 is called the active set, and s_0 the sparsity index of β^0.

Before we state the final oracle inequality, using $\lambda_n \geqslant 2\lambda_0$ and the Basic Inequality (2.7), we have on $\Im$,

$$2||(I - \hat{\rho}_n M_n)X_n(\hat{\beta} - \beta)||_2^2/n + 2\lambda_n||\hat{\beta}||_1 \leqslant \lambda_n||\hat{\beta} - \beta^0||_1 + 2\lambda_n||\beta||_1.$$

Since

$$||\hat{\beta}||_1 = ||\hat{\beta}_{S_0}||_1 + ||\hat{\beta}_{S_0^c}||_1 \geqslant ||\beta_{S_0}||_1 - ||\hat{\beta}_{S_0} - \beta_{S_0}||_1 + ||\hat{\beta}_{S_0^c}||_1,$$

and also,

$$||\hat{\beta} - \beta^0||_1 = ||\hat{\beta}_{S_0} - \beta_{S_0}||_1 + ||\hat{\beta}_{S_0^c}||_1.$$

Therefore,

$$2||(I - \hat{\rho}_n M_n)X_n(\hat{\beta} - \beta)||_2^2/n + \lambda_n||\hat{\beta}_{S_0^c}||_1 \leqslant 3\lambda_n||\hat{\beta}_{S_0} - \beta_{S_0}||_1. \tag{2.8}$$

Here, λ_n is some regularization parameter satisfying the relationship with λ_0 defined in Proposition 2.2. From Assumption 1, we have $0 < \sigma^2 < b$, hence $\lambda_n = 2\sqrt{b}(\exp[t^2/2] + 1)\sqrt{\frac{\log 2p}{n}}$ is a possible choice.

In order to prove the oracle inequality mentioned at the beginning of this section, we

need one more condition on the design matrix corresponding with the consistent estimator of ρ to make our proof go through and it is similar to the "compatibility condition" given in Bü hlmann and van de Geer (2001) with only minor changes. Since we know from inequality (2.8), on $\Im$,

$$||\hat{\beta}_{S_0^c}||_1 \leqslant 3||\hat{\beta}_{S_0} - \beta_{S_0}||_1,$$

we will only require the condition restricted on β_{S_0}. Thus, the compatibility condition in our case is stated as follows.

Condition 1. *Condition 1 is said to be satisfied for the set S_0, if for some constant $\phi_0 > 0$, and for all β satisfying $||\beta_{S_0^c}||_1 \leqslant 3||\beta_{S_0}||_1$, it holds that*

$$||\beta_{S_0}||_1^2 \leqslant (\beta' X_n' \Sigma(\hat{\rho}_n) X_n \beta) s_0/(n\phi_0^2).$$

Note that, when we solve for the LASSO estimator in the second step of our estimation process, $\hat{\rho}_n$ is considered a known parameter. Finally, we obtain parameter consistency as follows.

Theorem 2.3. *Suppose Condition 1 holds for S_0, for some $t > 0$, and let the regularization parameter $\lambda_n \geqslant 2\lambda_0$, then on $\Im$, we have*

$$\frac{||(I - \hat{\rho}_n M_n) X_n(\hat{\beta} - \beta)||_2^2}{n} + \lambda_n ||\hat{\beta} - \beta||_1 \leqslant \frac{4\lambda_n^2 s_0}{\phi_0^2}.$$

The result also means that with probability at least $1 - K\exp[-t^2/2]$, we have

$$\frac{||(I - \hat{\rho}_n M_n) X_n(\hat{\beta} - \beta)||_2^2}{n} + \lambda_n ||\hat{\beta} - \beta||_1 \leqslant \frac{4\lambda_n^2 s_0}{\phi_0^2}.$$

As discussed above, the above result tells states that with high probability, the L_1 norm of the difference between the estimator and the true value of the parameter of interest is bounded by a function of λ_n and s_0 (same as the dimension of non-zero parameters q_n). Further, the consistency of the estimator is achieved when the bound converges to 0 as n goes to infinity, and p_n and q_n in this case need to satisfy:

$$\frac{q_n^2 \log 2p_n}{n} \rightarrow \ 0.$$

2.4.2 Sign Consistency

We have already proved sign consistency which infers the model selection consistency of our generalized moments LASSO estimator with a condition similar to the Strong Irrepresentable Condition in Zhao and Yu (2006). Now, we extend the result to sign consistency of the estimator in the high dimensional case when p and q are large and growing with n, following the previous arguments but with an additional assumption:

Assumption 7. *There exists* $0 \leqslant c_1 < c_2 \leqslant 1$ *and* $K_1, K_2, K_3, K_4 > 0$ *so that the following holds:*

$$\frac{1}{n}(X_n' \Sigma(\rho))_{ii} \leqslant K_1, \ \ \forall i,$$

$$\alpha' C_{11}^n(\rho)\alpha \geqslant K_2, \ \ \forall ||\alpha||_2^2 = 1, \tag{2.9}$$

$$q_n = O(n^{2c_1}),$$

$$n^{\frac{1-c_2}{2}} \min_{i=1,\cdots,q} |\beta_i^n| \geqslant K_3.$$

Under the above assumptions, we can have the following result.

Theorem 2.4. *Under Assumptions 1-5 and 7, if the condition* $|(C_{21}^n)(C_{11}^n)^{-1}sign(\beta(1))| \leqslant 1-\eta$ *holds for some* $\eta > 0$, *then for* $p_n = o(n^{2(c_2-c_1)})$, *and* $\forall \lambda_n$ *that satisfies* $\frac{\lambda_n}{\sqrt{n}} = O(n^{\frac{c_2-c_1}{2}})$, *we have*

$$P(\hat{\beta}_L(\hat{\rho}_n;\lambda_n) =_s \beta) \geqslant 1 - O\left(r(\rho)n^{2c_2-2}\right) - o(1) \to 1, \quad as \ \ n \to \infty.$$

Here, we denote $r(\rho)$ as a function of the spatial parameter ρ which controls the maximum of the absolute value of the element in the matrix $(C_{11}^n(\rho))^{-1}$. The term $r(\rho)$ controlling the convergence rate of the estimator to correctly select the true model in our spatial autoregressive errors setting differs from that in the traditional independent data linear regression setting.

2.5 Simulation Studies

In this section we study the finite sample performance of the generalized moments LASSO estimator (GLASSO) $\hat{\beta}_L(\hat{\rho}_n)$ in both the low-dimensional setting and the high-dimensional setting and compare these with the traditional LASSO estimator $\hat{\beta}_L$ as well as the ordinary least squares estimator $\hat{\beta}_{OLS}$ (when applicable in the low dimensional case); both the $\hat{\beta}_L$ and the $\hat{\beta}_{OLS}$ ignore spatial dependence in the data. For this purpose, we conduct a two-part Monte Carlo study. Throughout, we set the distribution of ϵ to be normal, and without loss of generality, $N(0,1)$. This is because the estimators for ρ defined earlier do not depend on σ^2. We consider 6 choices of ρ, covering the range from -1 to 1, together with 5 choices of the sample size n, and thus we have a total of 30 cases for our simulation study. For each

case, the results are summarized over 200 Monte Carlo replications.

The detail design of the study is as follows. The weight matrix M_n is defined as an idealized weighting matrix M_n following Kelejian and Prucha (1999), which means, M_n was selected such that each element of u_{ni} is directly related to the one immediately before and after it. We assume the above relationship to be circular, so that u_n is related to $u_{n,1}$ and $u_{n,n-1}$, for instance. For simplicity, we specify M_n such that all the non-zero elements of M_n are equal and that the respective rows sum to 1.

Our main object of interest lies in the ability of the generalized moments LASSO estimator to consistently choose the correct parameter, and the simulation result shows the mean (over 200 replicates) for the value of Correctly, Falsely, and Sign-correctly identified component of the parameter for our GLASSO, traditional LASSO and OLS (only in the low dimensional case), respectively. Note that Huang et al. (2010) demonstrate the selection consistency of using group LASSO for variable selection with a certain lower bound for the penalty parameter λ_n. For the analysis, we also set a proper chosen lower bound for the cross-validation selection of λ_n, which satisfies the conditions implied by our theoretical results.

In the low-dimensional set up, the dimension of the parameter β is chosen as $p = 50$ with the first $q = 5$ non-zero components independently generated from a uniform distribution over the interval $(-2, 5)$ and the rest are zero coefficients. The covariates X_i's are IID from a 50-dimensional Gaussian distribution with each component having mean zero and variance 1. The pairwise correlation is set to be $cor(x_{ij}, x_{ik}) = 0.5^{|j-k|}$. The results for the low-dimensional setup are shown in Table 2.1 and 2.2. In each of the tables, the reported figures are the means of the statistics from 200 repetitions; NZ represents the correctly selected components, IZ represents the incorrectly selected components and SC represents the the

number of correctly selected components with correct sign.

For the high-dimensional set up, we set the dimension of the parameter $p = 1000$ but the true number of components that are significant is only $q = 20$. M_n is specified the same as it is in the low-dimensional setting. The first 20 non-zero components are also generated independently from a uniform distribution over the interval $(-2, 5)$. The design matrix X_n likewise is the same as that of the low-dimensional design with only change of dimensions. Traditional OLS becomes impossible so we only compare the performance of the generalized LASSO and the traditional LASSO. Note here, in the traditional LASSO approach, we ignore the autocorrelation of the error u_n and treat the errors as IID. The estimator of LASSO is achieved by using the package "glmnet" in R and the penalty parameter λ_n is chosen by 10-fold cross-validation. Another issue that distinguishes our method from the traditional LASSO is the use of a lower bound for λ_n, which justifies the consistency of our approach. The results are recorded in Tables 2.3 and 2.4.

Table 2.1: Means of NZ, IZ, SC for SEM when $p < n$ with positive ρ

n		ρ=0.25			ρ=0.5			ρ=0.75		
		NZ (5)	IZ(45)	SC(5)	NZ (5)	IZ(45)	SC(5)	NZ (5)	IZ(45)	SC(5)
100	GLASSO	5	19.85	5	5	19.87	5	5	19.71	5
	LASSO	5	14.90	5	5	19.24	5	5	24.64	5
	OLS	5	2.93	5	5	3.43	5	4.91	4.72	4.91
200	GLASSO	5	13.10	5	5	11.96	5	5	9.90	5
	LASSO	5	14.88	5	5	21.03	5	5	25.33	5
	OLS	5	3.46	5	5	4.51	5	5	6.55	5
400	GLASSO	5	6.61	5	5	6.33	5	5	6.7	5
	LASSO	5	15.73	5	5	20.85	5	5	26.29	5
	OLS	5	3.39	5	5	5.08	5	5	7.3	5
600	GLASS	5	1.66	5	5	1.59	5	5	1.76	5
	LASSO	5	15.78	5	5	20.37	5	5	26.10	5
	OLS	5	3.74	5	5	5.23	5	5	7.23	5
800	GLASS	5	0.23	5	5	0.25	5	5	0.35	5
	LASSO	5	15.11	5	5	21.62	5	5	26.18	5
	OLS	5	3.76	5	5	5.44	5	5	7.46	5

From the simulation results, we can see clearly that in all of the cases, all the methods considered above can (almost) consistently select the non-zero components of the regression

Table 2.2: Means of NZ, IZ, SC for SEM when $p < n$ with negative ρ

n		ρ=-0.25			ρ=-0.5			ρ=-0.75		
		NZ (5)	IZ(45)	SC(5)	NZ (5)	IZ(45)	SC(5)	NZ (5)	IZ(45)	SC(5)
100	GLASS	5	16.2	5	5	14.20	5	5	11.08	5
	LASSO	5	8.38	5	5	5.32	5	5	4.07	5
	OLS	4.99	1.9	4.99	4.93	1.63	4.93	4.64	1.1	5
200	GLASSO	5	11.87	5	5	11.97	5	5	10.46	5
	LASSO	5	7.3	5	5	4.56	5	5	2.75	5
	OLS	5	1.6	5	5	0.93	5	5	0.6	5
400	GLASSO	5	2.83	5	5	2.90	5	5	2.53	5
	LASSO	5	6.79	5	5	3.82	5	5	2.54	5
	OLS	5	1.44	5	5	0.66	5	5	0.29	5
600	GLASSO	5	0.33	5	5	0.37	5	5	0.23	5
	LASSO	5	6.58	5	5	4.06	5	5	2.21	5
	OLS	5	1.27	5	5	0.55	5	5	0.25	5
800	GLASSO	5	0.01	5	5	0.02	5	5	0.03	5
	LASSO	5	6.21	5	5	3.64	5	5	2.22	5
	OLS	5	1.22	5	5	0.47	5	5	0.22	5

Table 2.3: Means of NZ, IZ, SC for SEM when $p > n$ with positive ρ

n		ρ=0.25			ρ=0.5			ρ=0.75		
		NZ(20)	IZ(980)	SC(20)	NZ(20)	IZ(980)	SC(20)	NZ(20)	IZ(980)	SC(20)
100	GL	15.5	77.4	15.4	15.6	79.7	15.6	15.2	80.7	15.2
	L	16.4	105.6	16.4	16.4	113.3	16.4	15.9	117.5	15.9
200	GL	19.6	96.5	19.6	19.5	104.5	19.5	19.3	122.2	19.3
	L	19.7	116.7	19.7	19.7	168.1	19.7	19.6	254.2	19.6
400	GL	20	93.1	20	19.9	116.3	19.9	19.7	158	19.7
	L	20	129.7	20	20	194.9	20	19.9	277.3	19.9
600	GL	20	48.5	20	20	69.1	20	19.9	127.0	19.9
	L	20	140.3	20	20	226.3	20	20	325.7	20
800	GL	20	14.9	20	20	23.7	20	20	61.8	20
	L	20	150.9	20	20	258.2	20	20	374.6	20

parameter when the sample size n gets larger. What distinguishes the methods truly is their ability to identify the irrelevant components and set these to zero. From Tables 2.1 and 2.2, in the low-dimensional case, it is clear that the traditional LASSO is not suitable for dependent data and OLS works reasonably well for all choices of n. However, even though our generalized moments LASSO estimator (GLASSO) falsely selects more zero components in small sample sizes, the results get much better with increasing data and performs better than the OLS when n exceeds 400. These results are consistent for all choices of the autoregressive parameter ρ.

Table 2.4: Means of NZ, IZ, SC for SEM when $p > n$ with negative ρ

n		ρ=-0.25			ρ=-0.5			ρ=-0.75		
		NZ(20)	IZ(980)	SC(20)	NZ(20)	IZ(980)	SC(20)	NZ(20)	IZ(980)	SC(20)
100	GL	15.09	69.58	15.07	14.35	70.90	14.34	13.93	67.04	13.89
	L	15.98	88.43	15.95	14.92	82.18	14.88	13.64	60.09	13.61
200	GL	19.62	72.61	19.62	19.56	71.12	19.56	19.50	68.42	19.50
	L	19.65	71.69	19.65	19.46	58.9	19.46	19.05	48.67	19.05
400	GL	19.97	43.43	19.97	19.99	41.87	19.99	20	40.36	20
	L	19.99	54.26	19.99	19.95	35.72	19.95	19.76	23.83	19.76
600	GL	20	14.10	20	20	15.27	20	20	15.07	20
	L	20	46.13	20	20	28.38	20	19.95	16.71	19.95
800	GL	20	3.07	20	20	4.12	20	20	4.2	20
	L	20	42.16	20	20	24.28	20	19.99	14.14	19.99

In the high-dimensional setting, since OLS becomes unavailable, we only compare the performance of the traditional LASSO and our two-stage GLASSO estimator. Still, for different choices of ρ, even though both methods can mostly select the non-zero regression coefficients correctly, the traditional LASSO performs poorly relative to the generalized moments LASSO estimator in correctly identifying the zero elements. The most interesting observation is the way the LASSO over-selects in the presence of even a little bit of spatial dependence. This inability is not a finite sample bias: if anything, the problem worsens with sample size. There is also an important asymmetry between positive and negative dependence, which has to do with challenging inferences in negative autocorrelation situations. In summary, the LASSO loses its selection ability when the errors are not independent.

2.6 Application to a hedonic housing price model

In this section, we illustrate the proposed two-step GLASSO by application to housing market data for the municipalities of Aveiro and Ílhavo and the adjoining peri-urban and rural area in central Portugal (Figure 2.1); see Bhattacharjee et al. (2016) for the data and for further information.

Figure 2.1: **The Aveiro-Ílhavo Housing Market**

The dataset was provided by the firm Janela Digital S.A, which owns the largest portal in Portugal for real estate advertisement, and contains $n = 12,467$ observations (houses on sale) sampled from 76 different locations within the above housing market over the period October 2000 and March 2010. Our interest here lies mainly in estimating the spatially varying implicit price of living space that is modeled by the living space elasticity of house price. We estimate this elasticity by regressing the logarithm of house price per square meter of living area on the logarithm of square meters of living space. This is an example of a hedonic house price model; see Bhattacharjee at al. (2016) for further discussion.

Potentially, several other regressors relating to the attributes of the house, as well as the characteristics of the neighborhood, also affect house prices and hence should be included as controls. However, the effect of these hedonic characteristics on the spatially varying estimates of living space elasticity is not substantial, after spatial dependence is adequately modeled. Hence, for this illustrative example, we abstract from the full estimation of a hedonic house price model, and focus on spatial dependence.

We model the spatial aspects quite fully. This is done in three ways. First, we allow full spatial heterogeneity by allowing the shadow price of living space (β_{ii}) to vary across the $L = 76$ locations. In addition, we allow for L location specific fixed effects (α_i) to account for neighborhood level unobserved heterogeneity. Second, we model spatial spillovers in house price shocks by spatial autoregressive errors, where the spatial weights matrix (M_n) is a row-standardized version of inverse geographical distance weights. That is, we first construct a weights matrix where, corresponding to two houses in different locations, the off diagonal elements are reciprocal of the Euclidean distance between the locations; if the houses are in the same location, the corresponding spatial weight is the reciprocal of half the distance of that location to its nearest neighbor location. This weights matrix is then row-standardized by dividing each element by the sum of all entries in its row, and this transformed matrix then constitutes our spatial weights matrix M_n. Third, and most importantly in the context of this work, we allow spillovers of the quality of housing stock from neighboring locations to affect housing price in an index location. The most popular way to accommodate such spillovers in exogenous covariates is the spatial Durbin model (LeSage and Pace, 2009):

$$\begin{aligned} Y_n &= X_n\beta + W_nX_n\gamma + u_n, \\ u_n &= \rho M_n u_n + \epsilon_n. \end{aligned}$$

Here, in addition to the (direct) effect of the covariates and the spatial autoregressive errors, there is also the effect coming arising from covariate values in the neighborhood, and captured through a spatial lag term (W_nX_n) with corresponding effect γ. The above spatial Durbin model can have the structural interpretation of capturing the true spillovers in the effect of characteristics in the neighborhood, but may also sometimes be seen reflection in the reduced

form of omitted or inappropriately modeled spatial dependence (LeSage and Pace, 2009). Whatever the mechanism, the spatial Durbin model is an important workhorse model in contemporary spatial econometrics.

Typically, the spatial weights matrix W_n is assumed known *a priori*, and usually taken to be the same as M_n. However, mis-measured spatial weights can have serious implications on the inferences drawn, and a current branch of the literature focuses on inferences on the spatial weights themselves; see, for example, Bhattacharjee et al. (2016). Here, we use the GLASSO for identifying the neighbors that matter and for estimating the implied weights matrix γW_n, which has $L(L-1)$ elements. This allows spillovers and their strength to vary over the spatial domain, which is natural in the current context of hedonic pricing.

In a typical application, this would imply adding covariates for all locations on the right hand side of the regression model and then use LASSO based model selection to estimate both the spatially varying slope (β) and spillovers from other locations (γW_n). In the context of our application, the estimation of a three dimensional functional surface of the spatial varying effect of living space can be tailored to the regression of a linear combination of the effect of living space over nearby locations, besides the effect of living space on each specific location. Thus, the generalized moments LASSO variable selection and estimation method proposed is useful when we select neighboring locations whose living space have an effect on the index location and to estimate how large the effect is; in the process, we can build a parsimonious model by eliminating those locations that are irrelevant for housing prices at each index location.

Further, there is some simplification because of replications in our data. For every house in an index location, houses in any specific other location is expected to be exchangeable, and hence what matters is not the living spaces of these houses, but their average at each

location. To be specific, the chosen linear model can be described as:

$$Y_{ij} = \alpha_i + x_{ij}\beta_{ii} + \sum_{k \neq i} \bar{x}_{k.}\beta_{ik} + u_{ij}, \;\; i = 1, 2, \cdots, L, \;\; j = 1, 2, \cdots, n_i, \;\; n = \sum_i n_i.$$

Here Y_{ij} is the logarithm of house price per square meter of living space for the jth replication in the ith location, while X_{ij} represents the logarithm of living space of the corresponding house, and the average of the logarithm of living space at each of other locations k is denoted as $\bar{x}_{k.}$. Further, u_{ij} is a spatial autoregressive error with spatial weight matrix defined based on the distance between the locations (M_n).

Under the model, for each replication j in the location i, the logarithm of house price can be modeled as the linear combination of its corresponding logarithm of living space, along with the average of sampled logarithms of living space at other locations, plus an error term. The response variable y_{ij} is arranged by locations into a column vector of dimension $n = 12467$. We are interested in selection and inferences on the effect of the average of logarithm of living space at location k on each of the sample of logarithm of house price at location i, which is denoted β_{ik}, together with location specific fixed effects (α_i) and spatially varying living space elasticities of house price (β_{ii}). This makes our parameter of interest to be a $p = 76 \times 77$ dimension vector. Even though in this dataset, it is not the high-dimensional setting we defined earlier which requires $p > n$, it is high dimension in the sense that p is considerably large. In addition to estimating the spatially varying implicit price of living space, we wish to identify those locations with "living space" effect on each other, so hence we implement the two step method for variable selection and estimation for the proposed model. We compare these results with the traditional LASSO method.

By conducting variable selection and estimation for the proposed model at each location,

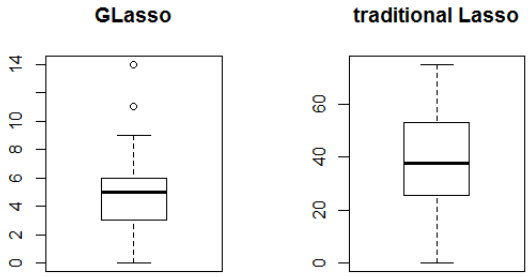

Figure 2.2: **Identified neighbors with spillover effect of living space**

the effect of living space from locations in the neighborhood is thus identified. Part of the results are shown in Table 2.5 as illustration and we summarize the number of identified neighbors through box plot (Figure 2.2).

Table 2.5: Numerical summary of the significant effect of living size from neighbors

Location	No. of neighbors (GLASSO)	No. of neighbors (LASSO)
1	5	56
2	5	59
3	4	63
4	6	56
5	5	26
6	7	25
7	4	8
8	3	43
9	4	25
10	1	11

Table 2.5 illustrates the differences in number of selected neighbors with spillover effects of living space effect identified by the generalized moments LASSO estimator and the traditional LASSO method. Coinciding with the simulation results, the traditional LASSO estimator tends to over-select irrelevant variables compared to the generalized moments LASSO estimator and thus weaken selection power. The summary of box plots in Figure 2.2 further supports the above conclusion using the overall distribution of the number of

(a) Spatial effect for location 1 (b) Spatial effect for location 6

Figure 2.3: **Computation results**

neighbors identified. While the GLASSO selects parsimonious models with a median of 5 neighbors, the traditional LASSO selects enormously large models with a median of about 40 neighbors.

The network graphs in Figure 2.3 use two example locations to illustrate the relationship between these specific locations (Locations 1 and 6 in our case) and the identified locations with significant spillover effects. The magnitudes of the locations with large spatial effects are also shown. Each point is located at its own location defined by coordinates and the distance between locations represent the relative distance between points. Take Location 1 for example. The house price in this location is effected by the living space at five other surrounding locations and Location 2 has the largest effect of these 5 locations.We can see that identified locations are not completely random but somewhat conditioned by distances on the spatial domain. At the same time, one can clearly see the advantages of allowing the patterns of spillovers to be different in different locations. Comparing the results with Figure 2.1, we see that our method can successfully identify the spillover locations in the nearby area.

2.7 Proofs

PROOF of Theorem 2.1. Define a random function of ρ and ϕ,

$$Z_n(\phi, \rho) = \frac{1}{n}(Y_n - X_n\phi)'\Sigma(\rho)(Y_n - X_n\phi) + \frac{\lambda_n}{n}\sum_{j=1}^{p}|\phi_j|.$$

By the definition of LASSO estimator, for any fixed ρ, $Z_n(\phi, \rho)$ is minimized at $\phi = \hat{\beta}_L(\rho)$. However, we not have the true value of ρ, but instead, we use the GMM estimator $\hat{\rho}_n$ as a substitute. Then the function $Z_n(\phi, \hat{\rho}_n)$ is minimized at the generalized moments LASSO estimator $\phi = \hat{\beta}_L(\hat{\rho}_n)$. Furthermore, denote by β the true value of the unknown parameter, and let

$$Z(\phi, \rho) = (\beta - \phi)'C(\rho)(\beta - \phi) + \sigma^2.$$

Then, it is easy to see that for any given ρ, $Z(\beta, \rho)$ is minimized at $\phi = \beta$.

For each $\phi \in \mathcal{R}^p$,

$$\begin{aligned} Z_n(\phi, \hat{\rho}_n) &= \frac{1}{n}(Y_n - X_n\phi)'\Sigma(\hat{\rho}_n)(Y_n - X_n\phi) + \frac{\lambda_n}{n}\sum_{j=1}^{p}|\phi_j| \\ &= \Phi_1 - \Phi_2 + \Phi_2 + \Phi_3 \end{aligned}$$

where

$$\Phi_1 = \frac{1}{n}(Y_n - X_n\phi)'\Sigma(\hat{\rho}_n)(Y_n - X_n\phi)$$

$$\Phi_2 = \frac{1}{n}(Y_n - X_n\phi)'\Sigma(\rho)(Y_n - X_n\phi)$$

$$\Phi_3 = \frac{\lambda_n}{n}\sum_{j=1}^{p}|\phi_j|$$

Since $\frac{\lambda_n}{n} \to 0$, we have $\Phi_3 \to 0$. Also,

$$\begin{aligned}
\Phi_2 &= \frac{1}{n}[(I-\rho M_n)X_n(\beta-\phi)+\varepsilon_n]'[(I-\rho M_n)X_n(\beta-\phi)+\varepsilon_n] \\
&= \frac{1}{n}(\beta-\phi)'X_n'\Sigma(\rho)X_n(\beta-\phi)+\frac{1}{n}\varepsilon_n'(I-\rho M_n)X_n(\beta-\phi)+ \\
&\quad \frac{1}{n}(\beta-\phi)'X_n'(I-\rho M_n)'\varepsilon_n+\frac{1}{n}\varepsilon_n'\varepsilon_n \\
\to_p &\quad (\beta-\phi)'C(\rho)(\beta-\phi)+\sigma^2 \\
&= Z(\phi,\rho),
\end{aligned}$$

by Assumption 6 and the weak law of large numbers.

Moreover, since $\hat{\rho}_n$ is a consistent estimator of ρ,

$$\begin{aligned}
\Phi_1 - \Phi_2 &= \frac{1}{n}(Y_n - X_n\phi)'[\Sigma(\hat{\rho}_n) - \Sigma(\rho)](Y_n - X_n\phi) \\
&= \frac{1}{n}(Y_n - X_n\phi)'[(\rho - \hat{\rho}_n)(M_n + M_n') + (\hat{\rho}_n^2 - \rho^2)M_n'M_n](Y_n - X_n\phi) \\
&= \frac{1}{n}[(\beta - \phi)'X_n' + \varepsilon_n'(I - \rho M_n')^{-1}][(\rho - \hat{\rho}_n)(M_n + M_n') + (\hat{\rho}_n^2 - \rho^2)M_n'M_n] \\
&\quad [X_n(\beta - \phi) + (I - \rho M_n)^{-1}\varepsilon_n] \\
&= \frac{1}{n}(\rho - \hat{\rho}_n)(\beta - \phi)'X_n'(M_n + M_n')X_n(\beta - \phi) \\
&\quad + \frac{1}{n}(\hat{\rho}_n^2 - \rho^2)(\beta - \phi)'X_n'(M_n'M_n)X_n(\beta - \phi) \\
&\quad + \frac{1}{n}(\rho - \hat{\rho}_n)\varepsilon_n'(I - \rho M_n')^{-1}(M_n + M_n')X_n(\beta - \phi) \\
&\quad + \frac{1}{n}(\hat{\rho}_n^2 - \rho^2)\varepsilon_n'(I - \rho M_n')^{-1}(M_n'M_n)X_n(\beta - \phi) \\
&\quad + \frac{1}{n}(\rho - \hat{\rho}_n)(\beta - \phi)'X_n'(M_n + M_n')(I - \rho M_n)^{-1}\varepsilon_n \\
&\quad + \frac{1}{n}(\hat{\rho}_n^2 - \rho^2)(\beta - \phi)'X_n'(M_n'M_n)(I - \rho M_n)^{-1}\varepsilon_n \\
&\quad + \frac{1}{n}(\rho - \hat{\rho}_n)\varepsilon_n'(I - \rho M_n')^{-1}(M_n + M_n')(I - \rho M_n)^{-1}\varepsilon_n \\
&\quad + \frac{1}{n}(\hat{\rho}_n^2 - \rho^2)\varepsilon_n'(I - \rho M_n')^{-1}(M_n'M_n)(I - \rho M_n)^{-1}\varepsilon_n \\
\to_p \quad & 0
\end{aligned}$$

Therefore, $Z_n(\phi, \hat{\rho}_n) - Z(\phi, \rho) \to_p 0$ for any $\phi \in \mathcal{R}^p$. Combined with the fact that $Z_n(\phi, \hat{\rho}_n)$ is a convex function of ϕ, we have

$$\sup_{\phi \in \mathcal{K}} |Z_n(\phi, \hat{\rho}_n) - Z(\phi, \rho)| \to_p 0$$

for any compact set $\mathcal{K}$ and $\hat{\beta}_L(\hat{\rho}_n) \in O_p(1)$ by applying the convexity lemma in Pollard

(1991). From the above result we have

$$\arg\min(Z_n(\phi, \hat{\rho}_n)) \to_p \arg\min(Z(\phi, \rho))$$

which implies that

$$\hat{\beta}_L(\hat{\rho}_n) \to_p \beta.$$

For asymptotic normality of the estimator, we need λ_n to grow slowly, and further assume that $\lambda_n = O(\sqrt{n})$. From the above proof, we already know that

$$nZ_n(\phi, \hat{\rho}_n) = (Y_n - X_n\phi)'\Sigma(\hat{\rho}_n)(Y_n - X_n\phi) + \lambda_n \sum_{j=1}^{p} |\phi_j|$$

is minimized at $\phi = \hat{\beta}_L(\hat{\rho}_n)$. Now define $w = \sqrt{n}(\phi - \beta)$. Then $nZ_n(\phi, \hat{\rho}_n)$ can be treated as a function of w and

$$\begin{aligned} nZ_n(\phi, \hat{\rho}_n) &= \left[Y_n - X_n\left(\frac{w}{\sqrt{n}} + \beta\right)\right]' \Sigma(\hat{\rho}_n) \left[Y_n - X_n\left(\frac{w}{\sqrt{n}} + \beta\right)\right] + \lambda_n \sum_{j=1}^{p} \left|\frac{w_j}{\sqrt{n}} + \beta_j\right| \\ &= \tilde{V}_n(w) \end{aligned}$$

is minimized at $\sqrt{n}\left(\hat{\beta}_L(\hat{\rho}_n) - \beta\right)$. The same is true for

$$V_n(w) = \tilde{V}_n(w) - (Y_n - X_n\beta)'\Sigma(\hat{\rho}_n)(Y_n - X_n\beta) - \lambda_n \sum_{j=1}^{p} |\beta_j|.$$

It follows that

$$\lambda_n \sum_{j=1}^{p} \left[\left| \frac{w_j}{\sqrt{n}} + \beta_j \right| - |\beta_j| \right] \to \lambda_0 \sum_{j=1}^{p} [w_j sgn(\beta_j) I(\beta_j \neq 0) + |w_j| I(\beta_j = 0)].$$

Also, define

$$\begin{aligned} \Omega_n(w) &= (Y_n - X_n \frac{w}{\sqrt{n}} - X_n\beta)' \Sigma(\hat{\rho}_n)(Y_n - X_n \frac{w}{\sqrt{n}} - X_n\beta) - \\ &\quad (Y_n - X_n\beta)' \Sigma(\hat{\rho}_n)(Y_n - X_n\beta) \\ &= \Omega_n(w) - \Omega_1(w) + \Omega_1(w), \end{aligned}$$

where

$$\Omega_1(w) = (Y_n - X_n \frac{w}{\sqrt{n}} - X_n\beta)' \Sigma(\rho)(Y_n - X_n \frac{w}{\sqrt{n}} - X_n\beta) - \varepsilon_n' \varepsilon_n.$$

Easy to see that

$$\begin{aligned} \Omega_1(w) &= \left[\varepsilon_n - (I - \rho M_n) X_n \frac{w}{\sqrt{n}} \right]' \left[\varepsilon_n - (I - \rho M_n) X_n \frac{w}{\sqrt{n}} \right] - \varepsilon_n' \varepsilon_n \\ &= -2 \frac{1}{\sqrt{n}} w' X_n' (I - \rho M_n)' \varepsilon_n + \frac{1}{n} w' X_n' (I - \rho M_n)' (I - \rho M_n) X_n w \\ \to_D & \quad -2w'U + w'C(\rho)w, \end{aligned}$$

where $U \sim N(0, \sigma^2 C(\rho))$. Also

$$\begin{aligned}
\Omega_n(w) - \Omega_1(w) &= \frac{-2}{\sqrt{n}}\varepsilon_n'(I - \rho M_n')^{-1}\Sigma(\hat{\rho}_n)X_n w + \frac{1}{n}w' X_n'\Sigma(\hat{\rho}_n)X_n w \\
&\quad + \frac{2}{\sqrt{n}}\varepsilon_n'(I - \rho M_n)X_n w - \frac{1}{n}w' X_n'\Sigma(\rho)X_n w \\
&= \frac{2}{\sqrt{n}}\varepsilon_n'(I - \rho M_n')^{-1}[(\hat{\rho}_n - \rho)(M_n' + M_n) - (\hat{\rho}_n^2 - \rho^2)M_n' M_n]X_n w \\
&\quad - \frac{1}{n}w' X_n'[(\hat{\rho}_n - \rho)(M_n' + M_n) - (\hat{\rho}_n^2 - \rho^2)M_n' M_n]X_n w \\
\to_p \quad & 0
\end{aligned}$$

where we use the consistency of $\hat{\rho}_n$ in the proof above. Thus $V_n(w) \to_D V(w)$, and combined with the fact that V_n is convex and V has a unique minimum, it follows from Geyer (1996) that

$$\arg\min(V_n) = \sqrt{n}\left[\hat{\beta}_L(\hat{\rho}_n) - \beta\right] \to_D \arg\min(V(w)).$$

PROOF of Proposition 2.1. By the definition of estimator in the second estimation step,

$$\hat{\beta}_L(\hat{\rho}_n) = \arg\min_{\phi}[(Y_n - X_n\phi)\Sigma(\hat{\rho}_n)(Y_n - X_n\phi)] + \lambda_n||\phi||_1,$$

where the estimator is the minimizer of the penalized least square when the true spatial parameter ρ is replaced by its consistent estimator $\hat{\rho}_n$. Let $\varphi = \phi - \beta$, which is equivalent to $\frac{w}{\sqrt{n}}$ in the proof of Theorem 2.1. The following proof is similar to that of the proof of Theorem 2.1. Define

$$\begin{aligned}
D_n(\varphi) &= [(Y_n - X_n(\varphi + \beta))'\Sigma(\hat{\rho}_n)(Y_n - X_n(\varphi + \beta))] + \lambda_n||\varphi + \beta||_1 \\
&\quad -(Y_n - X_n\beta)'\Sigma(\hat{\rho}_n)(Y_n - X_n\beta)
\end{aligned}$$

Then

$$\begin{aligned}\hat{\varphi} &= \hat{\beta}_L(\hat{\rho}_n) - \beta \\ &= \arg\min_{\varphi} D_n(\varphi).\end{aligned}$$

Separate $D_n(\varphi)$ into two parts, $D_{n1}(\varphi)$ and $D_{n2}(\varphi)$. Let

$$\begin{aligned}D_{n1}(\varphi) &= [(Y_n - X_n(\varphi+\beta))'\Sigma(\hat{\rho}_n)(Y_n - X_n(\varphi+\beta))] - (Y_n - X_n\beta)'\Sigma(\hat{\rho}_n)(Y_n - X_n\beta) \\ &= [(I-\hat{\rho}_n M_n)((I-\rho M_n)^{-1}\varepsilon_n - X_n\varphi)]'[(I-\hat{\rho}_n M_n)((I-\rho M_n)^{-1}\varepsilon_n - X_n\varphi)] \\ &\quad -\varepsilon_n'(I-\rho M_n')^{-1}(I-\hat{\rho}_n M_n')(I-\hat{\rho}_n M_n)(I-\rho M_n)^{-1}\varepsilon_n \\ &= -2\varphi' X_n'(I-\hat{\rho}_n M_n)'(I-\hat{\rho}_n M_n)(I-\rho M_n)^{-1}\varepsilon_n \\ &\quad +\varphi' X_n'(I-\hat{\rho}_n M_n')(I-\hat{\rho} M_n)X_n\varphi \\ &= -2(\sqrt{n}\varphi)'W^n + (\sqrt{n}\varphi)'C^n(\hat{\rho}_n)(\sqrt{n}\varphi)\end{aligned}$$

where

$$W^n = W^n(\hat{\rho}_n) = X_n'\Sigma(\hat{\rho}_n)(I-\rho M_n)^{-1}\varepsilon_n/\sqrt{n},$$

Differentiate $D_n(\varphi)$ w.r.t. φ, we have

$$\frac{dD_{n1}(\varphi)}{d\varphi} = -2\sqrt{n}W^n + 2nC^n(\hat{\rho}_n)\varphi.$$

Note here that both $\hat{\varphi}(1)$ and $W^n(1)$ are vectors of dimension $p \times 1$. Let $\hat{\varphi}(1)$, $W^n(1)$ and $\hat{\varphi}(2)$, $W^n(2)$ denote the first q and last $p-q$ entries of $\hat{\varphi}$ and W^n respectively. Then by

definition:

$$\{sign(\hat{\beta}_{Lj}(\hat{\rho}_n)) = sign(\beta_j), for\ \ j = 1, 2, \cdots, q.\} \supseteq \{sign(\beta(1))\hat{\varphi}(1) > -|\beta(1)|\}.$$

Hence if there exists $\hat{\varphi}$ such that

$$C_{11}^n(\hat{\rho}_n)(\sqrt{n}\hat{\varphi}(1)) - W^n(1) = -\frac{\lambda_n}{2\sqrt{n}} sign(\beta(1)),$$

$$|\hat{\varphi}(1)| < |\beta(1)|,$$

$$-\frac{\lambda_n}{2\sqrt{n}}\mathbf{1} \leqslant C_{21}^n(\hat{\rho}_n)(\sqrt{n}\hat{\varphi}(1)) - W^n(2) \leqslant \frac{\lambda_n}{2\sqrt{n}}\mathbf{1},$$

then by Lemma 2.1 and the uniqueness of LASSO solution, $sign(\hat{\beta}_L(\hat{\rho}_n)(1)) = sign(\beta(1))$ and $\hat{\beta}_L(\hat{\rho}_n)(2) = \beta(2) = 0$.

And the existence of such $\hat{\varphi}$ is implied by

$$|(C_{11}^n(\hat{\rho}_n))^{-1}W^n(1)| < \sqrt{n}(|\beta(1)| - \frac{\lambda_n}{2n}|(C_{11}^n(\hat{\rho}_n))^{-1}sign(\beta(1)|), \tag{2.10}$$

$$|C_{21}^n(\hat{\rho}_n)(C_{11}^n(\hat{\rho}_n))^{-1}W^n(1) - W^n(2)| \leqslant \frac{\lambda_n}{2\sqrt{n}}\left(\mathbf{1} - |C_{21}^n(\hat{\rho}_n)\,(C_{11}^n(\hat{\rho}_n))^{-1}\,sign(\beta(1))|\right) \tag{2.11}$$

here (2.10) coincides with A_n and (2.11) contains B_n. The result for Proposition 2.1 follows.

PROOF of Theorem 2.2. From Proposition 2.1, we have

$$P(\hat{\beta}_L(\hat{\rho}_n; \lambda) =_s \beta) \geqslant P(A_n \cap B_n).$$

Thus,

$$\begin{aligned} P(A_n \cap B_n) &\geqslant 1 - P(A_n^c) - P(B_n^c) \\ &\geqslant 1 - \sum_{i=1}^{q} P(|z_i^n| \geqslant \sqrt{n}(|\beta_i^n| - \frac{\lambda_n}{2n} b_i^n) - \sum_{i=1}^{p-q} P(|\zeta_i^n| > \frac{\lambda_n}{2\sqrt{n}} \eta_i). \end{aligned}$$

where $z^n = (z_1^n, \cdots, z_q^n)' = (C_{11}^n)^{-1} W^n(1)$, $\zeta^n = (\zeta_1^n, \cdots, \zeta_{p-q}^n)' = C_{21}^n (C_{11}^n)^{-1} W^n(1) - W^n(2)$ and $b^n = (b_1^n, \cdots, b_q^n)' = (C_{11}^n)^{-1} sign(\beta(1))$.

Since $\hat{\rho}_n$ is a consistent estimator of ρ, similar to the proof of Theorem 2.1, and under the regularity conditions in Assumption 6, we have

$$(C_{11}^n)^{-1} W^n(1) \to_D N(0, C_{11}^{-1}(\rho)\sigma^2)$$

This is because

$$\begin{aligned} C^n &= \frac{1}{n} X_n' \Sigma(\hat{\rho}_n) X_n \\ &= \frac{1}{n} X_n' \Sigma(\hat{\rho}_n) X_n - \frac{1}{n} X_n' \Sigma(\rho) X_n + \frac{1}{n} X_n' \Sigma(\rho) X_n \\ &= \frac{1}{n} X_n' [(\hat{\rho}_n^2 - \rho^2) M_n' M_n - (\hat{\rho}_n - \rho)(M_n' + M_n)] X_n + \frac{1}{n} X_n' \Sigma(\rho) X_n \\ \to_p & \quad C \end{aligned}$$

The final step follows from Assumption 3 and 6, together with the consistency of $\hat{\rho}_n$. Thus,

$(C_{11}^n(\hat{\rho}_n))^{-1} \to_p (C_{11}(\rho))^{-1}$. Similarily,

$$\begin{aligned} & X_n'\Sigma(\hat{\rho}_n)(I-\rho M_n)^{-1}\varepsilon_n/\sqrt{n} \\ = \; & X_n'[(\hat{\rho}_n^2-\rho^2)M_n'M_n-(\hat{\rho}_n-\rho)(M_n'+M_n)](I-\rho M_n)^{-1}\varepsilon_n/\sqrt{n} \\ & +X_n'(I-\rho M_n')\varepsilon_n/\sqrt{n} \\ = \; & o_p(1)O_p(1)+X_n'(I-\rho M_n')\varepsilon_n/\sqrt{n} \end{aligned}$$

Since $X_n'(I-\rho M_n')\epsilon_n/\sqrt{n} \to_d N(0,\sigma^2 C(\rho))$, we have

$$W_n = X_n'\Sigma(\hat{\rho}_n)(I-\rho M_n)^{-1}\epsilon_n/\sqrt{n} \to_d N(0,\sigma^2 C(\rho))$$

Thus $W^n(1) \to_D N(0,\sigma^2 C_{11}(\rho))$. Applying Slutsky's theorem, we have

$$z^n = (C_{11}^n)^{-1}W^n(1) \to_D N(0,(C_{11}(\rho))^{-1}\sigma^2).$$

Making use of the above result, combined with the fact that

$$C_{21}^n(C_{11}^n)^{-1}W^n(1) - W^n(2) = (C_{21}^n(C_{11}^n)^{-1}, -I_{p-q})W_n$$

we have

$$\zeta^n = C_{21}^n(C_{11}^n)^{-1}W^n(1) - W^n(2) \to_d N(0, C_{22}(\rho) - C_{21}(\rho)C_{11}(\rho)^{-1}C_{12}(\rho)\sigma^2).$$

Hence all z_i^n's and ζ_i^n's converge in distribution to Gaussian random variables with mean

0 and finite variance bounded by $s^2(\rho)$ for some constant function $s(\rho)$. For $t > 0$, the Gaussian distribution has its tail probability bounded by

$$1 - \Phi(t) < t^{-1} e^{-\frac{1}{2}t^2}$$

Since $\lambda_n/n \to 0$ and $\lambda_n/n^{\frac{1+c}{2}} \geqslant r$ with $0 \leqslant c < 1$, we have

$$\begin{aligned} & \sum_{i=1}^{q} P(|z_i^n| \geqslant \sqrt{n}(|\beta_i| - \frac{\lambda_n}{2n} b_i^n) \\ \leqslant \; & (1 + o(1)) \sum_{i=1}^{q} 2 \left(1 - \Phi \left(\frac{1}{s(\rho)} n^{\frac{1}{2}} |\beta_i^n| \, (1 + o(1)) \right) \right) \\ = \; & o\left(s(\rho) e^{\frac{-n^c}{s^2(\rho)}} \right) \end{aligned}$$

and

$$\sum_{i=1}^{p-q} P\left(|\zeta_i^n| \geqslant \frac{\lambda_n}{2\sqrt{n}} \eta_i \right) = (1 + o(1)) \sum_{i=1}^{p-q} 2 \left(1 - \Phi \left(\frac{1}{s} \frac{\lambda_n}{2\sqrt{n}} \eta_i \right) \right) = o\left(s(\rho) e^{\frac{-n^c}{s^2(\rho)}} \right).$$

Theorem 2.2 follows.

PROOF of Proposition 2.2.

$$
\begin{aligned}
& P(\max_{1\leqslant j\leqslant p} 2|\varepsilon_n' T^{(j)}|/n > \lambda_0) \\
= \; & P(\max_{1\leqslant j\leqslant p} \left| \frac{\varepsilon_n'(I-\rho M_n)X_n^{(j)}}{n} + \frac{\varepsilon_n'(I-\rho M_n')^{-1}\Sigma(\hat{\rho}_n)X_n^{(j)}}{n} - \frac{\varepsilon_n'(I-\rho M_n)X_n^{(j)}}{n} \right| \\
& > \frac{\lambda_0}{2}) \\
\leqslant \; & P(\max_{1\leqslant j\leqslant p} \left| \frac{\varepsilon_n'(I-\rho M_n)X_n^{(j)}}{n} \right| \\
& + \max_{1\leqslant j\leqslant p} \left| \frac{\varepsilon_n'(I-\rho M_n')^{-1}\Sigma(\hat{\rho}_n)X_n^{(j)} - \varepsilon_n'(I-\rho M_n')^{-1}\Sigma(\rho)X_n^{(j)}}{n} \right| > \frac{\lambda_0}{2})
\end{aligned}
$$

Let $r = \sigma\sqrt{\frac{\log 2p}{n}}$, and denote

$$
A = \max_{1\leqslant j\leqslant p} \left| \frac{\varepsilon_n'(I-\rho M_n')^{-1}\Sigma(\hat{\rho}_n)X_n^{(j)} - \varepsilon_n'(I-\rho M_n')^{-1}\Sigma(\rho)X_n^{(j)}}{n} \right|
$$

Then define

$$
A1 = \max_{1\leqslant j\leqslant p} \left| \frac{\varepsilon_n'(I-\rho M_n')^{-1}(\hat{\rho}_n^2-\rho^2)M_n' M_n X_n^{(j)}}{n} \right| \quad \text{and}
$$

$$
A2 = \max_{1\leqslant j\leqslant p} \left| \frac{\varepsilon_n'(I-\rho M_n')^{-1}(\hat{\rho}_n-\rho)(M_n'+M_n)X_n^{(j)}}{n} \right|;
$$

therefore,

$$
\begin{aligned}
P(A > r) &\leqslant P(A1 + A2 > r) \\
&\leqslant P\left(A1 > \frac{r}{2}\right) + P\left(A2 > \frac{r}{2}\right).
\end{aligned}
$$

Since $\hat{\rho}_n$ is a consistent estimator of ρ, that is, $\hat{\rho}_n \to_p \rho$, we have

$\forall t > 0$, defining $c = \frac{1}{2}\exp(-\frac{t^2}{2})$, when n is large enough,

$$P(|\hat{\rho}_n - \rho| > c) < c$$

and

$$P(|\hat{\rho}_n^2 - \rho^2| > c) < c.$$

Then, it is easy to see that

$$\begin{aligned}
P(A1 > \frac{r}{2}) &= P(\frac{\left\{\max_{1\leqslant j\leqslant p} |\varepsilon_n'(I-\rho M_n')^{-1}M_n'M_nX_n^{(j)}|\right\}|\hat{\rho}_n^2-\rho^2|}{n} > \frac{r}{2}) \\
&= P(\frac{\left\{\max_{1\leqslant j\leqslant p} |\varepsilon_n'(I-\rho M_n')^{-1}M_n'M_nX_n^{(j)}|\right\}|\hat{\rho}_n^2-\rho^2|}{n} > \frac{r}{2} \\
&\quad \bigcap |\hat{\rho}_n^2 - \rho^2| > c) \\
&\quad + P(\frac{\left\{\max_{1\leqslant j\leqslant p} |\varepsilon_n'(I-\rho M_n')^{-1}M_n'M_nX_n^{(j)}|\right\}|\hat{\rho}_n^2-\rho^2|}{n} > \frac{r}{2} \\
&\quad \bigcap |\hat{\rho}_n^2 - \rho^2| \leqslant c) \\
&\leqslant c + P(\max_{1\leqslant j\leqslant p} |\varepsilon_n'(I-\rho M_n')^{-1}M_n'M_nX_n^{(j)}| > \frac{rn}{2c})
\end{aligned}$$

and

$$\begin{aligned}
P(A2 > \frac{r}{2}) &= P(\frac{\left\{\max_{1\leqslant j\leqslant p} |\varepsilon_n'(I-\rho M_n')^{-1}(M_n'+M_n)X_n^{(j)}|\right\}|\hat{\rho}_n-\rho|}{n} > \frac{r}{2}) \\
&= P(\frac{\left\{\max_{1\leqslant j\leqslant p} |\varepsilon_n'(I-\rho M_n')^{-1}(M_n'+M_n)X_n^{(j)}|\right\}|\hat{\rho}_n-\rho|}{n} > \frac{r}{2} \\
&\quad \bigcap |\hat{\rho}_n-\rho| > c) \\
&\quad + P(\frac{\left\{\max_{1\leqslant j\leqslant p} |\varepsilon_n'(I-\rho M_n')^{-1}(M_n'+M_n)X_n^{(j)}|\right\}|\hat{\rho}_n-\rho|}{n} > \frac{r}{2} \\
&\quad \bigcap |\hat{\rho}_n-\rho| \leqslant c) \\
&\leqslant c + P(\max_{1\leqslant j\leqslant p} |\varepsilon_n'(I-\rho M_n')^{-1}(M_n'+M_n)X_n^{(j)}| > \frac{rn}{2c})
\end{aligned}$$

Next, we need the tail probability of

$$\max_{1\leqslant j\leqslant p} |\varepsilon_n'(I-\rho M_n')^{-1}M_n'M_nX_n^{(j)}|$$

and

$$\max_{1\leqslant j\leqslant p} |\varepsilon_n'(I-\rho M_n')^{-1}(M_n'+M_n)X_n^{(j)}|.$$

However, note that in our case, we do not assume Gaussian distribution for the error ϵ_n, instead, we only have zero mean and finite second moment assumption (Assumption 1). Thus, we use the moment inequality derived from the Nemirovski's inequality:

$$E\left(\max_{1\leqslant j\leqslant p} |\varepsilon_n'U^{(j)}|\right)^2 \leqslant 8\log(2p)\sum_{i=1}^{n}\left(\max_{1\leqslant j\leqslant p}|U_i^{(j)}|\right)^2 E\varepsilon_i^2$$

for any design matrix U, with $U^{(j)}$ as its jth column. Based on the assumption, the row

and column sums of M_n and $(I-\rho M_n)^{-1}$ are bounded uniformly in absolute value and each element of X_n are non-stochastic and uniformly bounded in absolute value. Also, we know that, if A_n and B_n are matrices that are conformable for multiplication with row and column sums uniformly bounded in absolute value, then the row and column sums of A_nB_n are also uniformly bounded in absolute value. Further, this result follows to 3 or more matrices. Thus, the row and column sums of $I-\rho M_n$, $(I-\rho M_n')^{-1}M_n'M_n$ and $(I-\rho M_n')^{-1}(M_n'+M_n)$ are all bounded uniformly in absolute value. So every element in matrices $(I-\rho M_n)X_n^{(j)}$, $(I-\rho M_n')^{-1}M_n'M_nX_n^{(j)}$ and $(I-\rho M_n')^{-1}(M_n'+M_n)X_n^{(j)}$ are bounded, and denote the common bound for all of them as κ_B.

Then, we have

$$\begin{aligned} P\left(A1 > \frac{r}{2}\right) &\leqslant c + \frac{E[\max_{1\leqslant j\leqslant p}|\varepsilon_n'(I-\rho M_n')^{-1}M_n'M_nX_n^{(j)}|]^2}{(rn/2c)^2} \\ &\leqslant c + \frac{8(2c)^2\log(2p)\sigma^2\kappa_B}{nr^2}, \end{aligned}$$

and similarly,

$$\begin{aligned} P\left(A2 > \frac{r}{2}\right) &\leqslant c + \frac{E[\max_{1\leqslant j\leqslant p}|\varepsilon_n'(I-\rho M_n')^{-1}(M_n'+M_n)X_n^{(j)}|]^2}{(rn/2c)^2} \\ &\leqslant c + \frac{8(2c)^2\log(2p)\sigma^2\kappa_B}{nr^2}. \end{aligned}$$

As a result,

$$\begin{aligned} P(A > r) &\leqslant P\left(A1 > \frac{r}{2}\right) + P\left(A2 > \frac{r}{2}\right) \\ &\leqslant 2c + \frac{(2c)^2\log(2p)\sigma^2\kappa_{B0}}{nr^2} \end{aligned}$$

Substituting the above probability bounds, we have

$$
\begin{aligned}
& P\left(\max_{1\leqslant j\leqslant p} 2|\varepsilon_n' T^{(j)}|/n > \lambda_0\right) \\
\leqslant\ & P\left(\max_{1\leqslant j\leqslant p}\left|\frac{\varepsilon_n'(I-\rho M_n)X_n^{(j)}}{n}\right| + A > \frac{\lambda_0}{2}\right) \\
\leqslant\ & P\left(\max_{1\leqslant j\leqslant p}\left|\frac{\varepsilon_n'(I-\rho M_n)X_n^{(j)}}{n}\right| + A > \frac{\lambda_0}{2}\bigcap A > r\right) \\
& + P\left(\max_{1\leqslant j\leqslant p}\left|\frac{\varepsilon_n'(I-\rho M_n)X_n^{(j)}}{n}\right| + A > \frac{\lambda_0}{2}\bigcap A \leqslant r\right) \\
\leqslant\ & 2c + \frac{(2c)^2\log(2p)\sigma^2\kappa_{B0}}{nr^2} + P\left(\max_{1\leqslant j\leqslant p}\left|\varepsilon_n'(I-\rho M_n)X_n^{(j)}\right| > n\left(\frac{\lambda_0}{2}-r\right)\right) \\
\leqslant\ & 2c + \frac{(2c)^2\log(2p)\sigma^2\kappa_{B0}}{nr^2} + \frac{E\left(\max_{1\leqslant j\leqslant p}|\varepsilon_n'(I-\rho M_n)X_n^{(j)}|\right)^2}{n^2(\frac{\lambda_0}{2}-r)^2} \\
\leqslant\ & 2c + \frac{(2c)^2\log(2p)\sigma^2\kappa_{B0}}{nr^2} + \frac{\log(2p)\sigma^2\kappa_{B0}}{n(\frac{\lambda_0}{2}-r)^2} \\
\leqslant\ & \exp[-t^2/2] + \kappa_{B0}\exp[-t^2] + \kappa_{B0}\exp[-t^2/2] \\
\leqslant\ & K\exp[-t^2/2].
\end{aligned}
$$

This then implies the proof of the result:

$$
\begin{aligned}
P(\Im) &= 1 - P\left(\max_{1\leqslant j\leqslant p} 2|\varepsilon_n' T^{(j)}|/n > \lambda_0\right) \\
&\geqslant 1 - K\exp[-t^2/2].
\end{aligned}
$$

PROOF of Theorem 2.3. On the set $\Im$, with $\lambda_n \geqslant 2\lambda_0$,

$$\begin{aligned}
& 2\frac{||(I-\hat{\rho}_n M_n)X_n(\hat{\beta}-\beta)||_2^2}{n} + \lambda_n||\hat{\beta}-\beta||_1 \\
= \; & 2\frac{||(I-\hat{\rho}_n M_n)X_n(\hat{\beta}-\beta)||_2^2}{n} + \lambda_n||\hat{\beta}_{S_0}-\beta_{S_0}||_1 + \lambda_n||\hat{\beta}_{S_0^c}||_1 \\
\leqslant \; & 4\lambda_n||\hat{\beta}_{S_0}-\beta_{S_0}||_1 \\
\leqslant \; & 4\lambda_n\sqrt{s_0}||(I-\hat{\rho}_n M_n)X_n(\hat{\beta}-\beta)||_2/(\sqrt{n}\phi_0) \\
\leqslant \; & ||(I-\hat{\rho}_n M_n)X_n(\hat{\beta}-\beta)||_2^2/n + 4\lambda_n^2 s_0/\phi_0^2,
\end{aligned}$$

where the final inequality follows from the fact that

$$4uv \leqslant u^2 + 4v^2.$$

Further, combining the oracle inequality with the Proposition regarding the set $\Im$, the result follows.

PROOF of Theorem 2.4. Using the result of Proposition 1 and the line of proof of Theorem 2, we have

$$\begin{aligned}
P(A_n \cap B_n) &\geqslant 1 - P(A_n^c) - P(B_n^c) \\
&\geqslant 1 - \sum_{i=1}^{q} P\left(|z_i^n| \geqslant \sqrt{n}(|\beta_i| - \frac{\lambda_n}{2n}b_i^n\right) - \sum_{i=1}^{p-q} P\left(|\zeta_i^n| > \frac{\lambda_n}{2\sqrt{n}}\eta_i\right).
\end{aligned}$$

where $z^n = (z_1^n, \cdots, z_q^n)' = (C_{11}^n)^{-1}W^n(1)$, $\zeta^n = (\zeta_1^n, \cdots, \zeta_{p-q}^n)' = C_{21}^n(C_{11}^n)^{-1}W^n(1) - W^n(2)$ and $b^n = (b_1^n, \cdots, b_q^n)' = (C_{11}^n)^{-1}sign(\beta(1))$.

Replace all the $\hat{\rho}_n$ in the notations above with the true parameter value ρ, and denote these as C_0^n, W_0^n, z_0^n, ζ_0^n, and b_0^n for simple notation. Then each element in the first term on the

right hand side of the above inequality is:

$$\begin{aligned}
& P(|z_i^n| \geqslant \sqrt{n}\left(|\beta_i| - \frac{\lambda_n}{2n} b_i^n\right) \\
= \; & P\left(|z_i^n| \geqslant \sqrt{n}(|\beta_i| - \frac{\lambda_n}{2n} b_i^n), |z_{0i}^n - z_i^n| > \delta, |b_{0i}^n - b_i^n| > \delta\right) \\
& + P\left(|z_i^n| \geqslant \sqrt{n}(|\beta_i| - \frac{\lambda_n}{2n} b_i^n), |z_{0i}^n - z_i^n| \leqslant \delta, |b_{0i}^n - b_i^n| \leqslant \delta\right) \\
& + P\left(|z_i^n| \geqslant \sqrt{n}(|\beta_i| - \frac{\lambda_n}{2n} b_i^n), |z_{0i}^n - z_i^n| > \delta, |b_{0i}^n - b_i^n| \leqslant \delta\right) \\
& + P\left(|z_i^n| \geqslant \sqrt{n}(|\beta_i| - \frac{\lambda_n}{2n} b_i^n), |z_{0i}^n - z_i^n| \leqslant \delta, |b_{0i}^n - b_i^n| > \delta\right) \\
= \; & A_1 + A_2 + A_3 + A_4
\end{aligned}$$

for any $\delta > 0$.

Since $C^n - C_0^n \to_p 0, W^n - W_0^n \to_p 0$, then $z^n - z_0^n = o_p(1), \zeta^n - \zeta_0^n = o_p(1)$ and $b^n - b_0^n = o_p(1)$.

Note that here we cannot use $C = \lim_{n\to\infty} \frac{1}{n} X_n' \Sigma(\rho) X_n$ as defined in Assumption 6, since this may not be nonsingular or maybe not even convergent in the high-dimensional context.

Thus, $A_1 + A_3 + A_4 < 3\delta$, and

$$\begin{aligned}
A_2 \; &= \; P\left(|z_i^n| \geqslant \sqrt{n}(|\beta_i^n| - \frac{\lambda_n}{2n} b_i^n), |z_{0i}^n - z_i^n| \leqslant \delta, |b_{0i}^n - b_i^n| \leqslant \delta\right) \\
&\leqslant \; P\left(|z_{0i}| \geqslant \sqrt{n}(|\beta_i^n| - \frac{\lambda_n}{2n}(b_{i0}^n + \delta)) - \delta\right).
\end{aligned}$$

Now if we write $z_0^n = H_A' \varepsilon_n$, where $H_A' = (h_1^a, \cdots, h_q^a)' = (C_{11}^0)^{-1} \frac{1}{\sqrt{n}} \left[(I - \rho M_n) X_n\right](1)'$, then

$$H_A' H_A = (C_{11}^0)^{-1} n^{-1} [(I - \rho M_n) X_n](1)' [(I - \rho M_n) X_n](1) (C_{11}^0)^{-1} = (C_{11}^0)^{-1}.$$

Therefore, $z_{0i}^n = (h_i^a)'\varepsilon_n$ with

$$||h_i^a||_2^2 \leqslant \frac{1}{K_2} \ \forall i = 1, \cdots, q. \tag{2.12}$$

Similarly,

$$\begin{aligned} & P\left(|\zeta_i^n| > \frac{\lambda_n}{2\sqrt{n}}\eta_i\right) \\ = \ & P\left(|\zeta_i^n| > \frac{\lambda_n}{2\sqrt{n}}\eta_i, |\zeta_i^n - \zeta_{0i}| > \delta) + P(|\zeta_i^n| > \frac{\lambda_n}{2\sqrt{n}}\eta_i, |\zeta_i^n - \zeta_{0i}| \leqslant \delta\right) \\ \leqslant \ & \delta + P\left(|\zeta_{0i}| > \frac{\lambda_n}{2\sqrt{n}}\eta_i - \delta\right). \end{aligned}$$

If we write $\zeta_0^n = H_B'\varepsilon_n$ where
$H_B' = (h_1^b, \cdots, h_{p-q}^b)' = C_{21}^0(C_{11}^0)^{-1}n^{-\frac{1}{2}}[(I-\rho M_n)X_n](1)' - n^{-\frac{1}{2}}[(I-\rho M_n)X_n](2)'$, then

$$\begin{aligned} & H_B'H_B \\ = \ & \frac{1}{n}[(I-\rho M_n)X_n](2)'\{I - [(I-\rho M_n)X_n](1) \\ & \{[(I-\rho M_n)X_n](1)'[(I-\rho M_n)X_n](1)\}^{-1}[(I-\rho M_n)X_n](1)'\}[(I-\rho M_n)X_n](2). \end{aligned}$$

Since $I - [(I-\rho M_n)X_n](1)\{[(I-\rho M_n)X_n](1)'[(I-\rho M_n)X_n](1)\}^{-1}[(I-\rho M_n)X_n](1)'$ has eigenvalues between 0 and 1, therefore $\zeta_{0i}^n = (h_i^b)'\varepsilon_n$ with

$$||h_i^b||_2^2 \leqslant K_1 \ \forall i = 1, \cdots, q. \tag{2.13}$$

Also note that,

$$\left|\frac{\lambda_n}{n}b_0^n\right| = \frac{\lambda_n}{n}\left|(C_{11}^0)^{-1}sign(\beta(1))\right| \leqslant \frac{\lambda_n}{nK_2}\left\|sign(\beta(1))\right\|_2 = \frac{\lambda_n}{nK_2}\sqrt{q} \tag{2.14}$$

Now given (2.12) and (2.13), it can be shown that $E(\varepsilon_i^n)^4 < \infty$ in Assumption 1 implies $E(z_i^n)^4 < \infty$ and $E(\zeta_i^n)^4 < \infty$. In fact, given any constant n-dimensional vector α,

$$E(\alpha'\varepsilon^n)^{2k} \leqslant (2k-1)!\left\|\alpha\right\|_2{}^2 E(\varepsilon_i^n)^{2k}.$$

For i.i.d. errors with bounded 4th moments, we have their tail probability bounded by

$$P(z_{i0}^n > t) = O(t^{-4})$$

Therefore, for $\lambda_n/\sqrt{n} = O(n^{\frac{c_2-c_1}{2}})$, using (2.14), if we make δ arbitrary small, we have

$$\begin{aligned}
&\sum_{i=1}^{q} P\left(|z_i^n| \geqslant \sqrt{n}\left(|\beta_i| - \frac{\lambda_n}{2n}b_i^n\right)\right) \\
\leqslant\ & q(3\delta + O(\sqrt{n}(|\beta_i| - \frac{\lambda_n}{2n}(b_{i0}^n + \delta)) - \delta)^{-4}) \\
=\ & qO\left(r(\rho)n^{-2c_2+2c_1-2}\right) \\
=\ & O\left(r(\rho)n^{-2+2c_2}\right),
\end{aligned}$$

where $r(\rho)$ is the bound for the absolute value of the elements in the matrix $(C_{11}^n(\rho))^{-1}$.

Likewise,

$$\begin{aligned}
& \sum_{i=1}^{p-q} P\left(|\zeta_i^n| > \frac{\lambda_n}{2\sqrt{n}}\eta_i\right) \\
\leqslant\ & \delta + (p-q)O\left(\frac{n^2}{\lambda_n^4}\right) \\
=\ & O\left(\frac{pn^2}{\lambda_n^4}\right) \\
=\ & o(1)
\end{aligned}$$

Adding these two terms, Theorem 2.4 follows.

Chapter 3

Post-model Selection Estimation for Regression Models with Spatial Autoregressive Error

3.1 Introduction

From the previous chapter, we have mentioned that there exists an extensive literature working on spatial econometric models where the data are collected spatially from cross-sectional units in one time period, and the spatial relation among the sampling sites can not be ignored. Among all, the spatial autoregressive model, which was first introduced by Cliff and Ord (Cliff and Ord, 1973, 1981) as a variant of the model suggested in Whittle (Whittle, 1954), is one of the most widely referenced models of spatial autocorrelation. In a regression context, if the spatial influence comes only through the error terms, we can model the disturbance term for one cross-sectional unit as a weighted average of disturbances corresponding to other cross-sectional units, plus an innovation. The weighted average involves a scalar parameter, denote ρ, and a spatial weight matrix whose elements describe the spatial interactions. And the innovations are typically assumed to be independent, identically distributed with mean zero and standard deviation σ^2. The parameter of interest in this case

will be ρ, σ^2 and the vector of regression coefficients β.

In a high-dimensional set up, which is easily encountered these days, traditional methods for regression models with spatial autoregressive errors can not be directly applied. One of the most common approaches to variable selection and estimation in high dimension models has been the least absolute shrinkage and selection operator, the ℓ_1 penalized Lasso estimator first introduced by Tibshirani (Tibshirani, 1996). It has been proved as a fundamental result that the Lasso type ℓ_1 penalized estimator obtains both the parameter consistency (Knight and Fu, 2000, B$\ddot{u}$hlmann and van de Geer, 2011) and model selection consistency (Zhao and Yu, 2006). And the ℓ_1 penalized least squares estimators achieve the ℓ_2 error convergence at the rate of $\sqrt{s \log p/n}$, which adds a penalty $\sqrt{\log p}$ to the oracle rate $\sqrt{s/n}$ of convergence when the true model is known. Here, n is the sample size, p is the total number of parameters and s is the number of parameters with non-zero coefficients (Bickle, Ritov, Tsybakov, 2009, Zhang, Huang, 2008).

Based on previous literature, Belloni and Chernozhukov (Belloni, Chernozhukov, 2013) proposed a two-step procedure, which applies ordinary least squares to the model selected by first-step Lasso estimator. They show that the post-model selection estimation performs at least as well as the Lasso in terms of the rate of convergence, even when Lasso did a unsatisfactory job in eliminating insignificant parameters in the variable selection step and it can be strictly better when Lasso can perfectly select the true regression model. We want to derive similar results for the spatial model in the high dimensional setting. In Chapter 2, we combine the idea of generalized moments estimator and ℓ_1 penalized estimator, and develop a generalized two-stage Lasso estimator as a first step model selection. It turns out that the variables selected in the first step is able to contain the true parameter set and the difference between the selection and true set $\hat{m}$ is at the same order with s. Then the least

squares estimator in the second step can achieve a ℓ_2 error rate as well as the estimator from the first step. Further if the first step can perfectly select the true model, that is, the difference $\hat{m}$ goes to zero in probability, then the two step estimator is able to attain the oracle rate of $\sqrt{s/n}$. Similar result for the sup norm estimation error rate is also derived. This is non-trivial work since the literature on high-dimensional models has focused mostly on ℓ_1 and ℓ_2 estimation errors, ℓ_∞ bounds on the estimation error was established for linear regression model (Lounici, 2008, van de Geer, 2014), but has not been mentioned in the high-dimensional spatial models.

The rest of the chapter is organized as follows. Section 2 discusses the properties of the model selection and estimation by first step model selection applying the generalized Lasso method described in Chapter 2. Section 3 proves the ℓ_2 error as well as the *sup* norm error convergence rate for the least squares post-model selection estimation. Section 4 provides the simulation studies of the performance of the post-model selection estimator compared with simultaneous variable selection and estimation estimator empirically. Section 5 applies the proposed method to a small Columbus crime dataset as an illustration. All the proofs are relegated to the Section 6.

3.2 Model Estimation and Selection properties

Following the estimation procedures described in the previous chapter, we can see that the objective of the model selection steps is to recover the true support T with $card(T) = s$, of the parameter vector β and the properties of the post-model selection estimators depend crucially on both the estimation and model selection properties of the generalized Lasso. In this section, we will develop the estimation properties of the generalized Lasso in the form of

ℓ_2 norm, followed by selection properties given by the selection support $\hat{T}$ of the generalized Lasso estimator $\hat{\beta}_L(\hat{\rho}_n)$(we will denote $\hat{\beta}$ for easy notation).

Assumption 8. *Assume the elements of X_n are uniformly bounded in absolute value, further assume that* $\max_{i,j} |x_{ij}| = O(\frac{1}{\sqrt{s}})$.

Assumption 8 does nothing but controls the magnitude of the components of the design matrix, and it can be achieved by normalizing the design matrix. For the analysis of the estimator, we need to use the following restricted eigenvalue condition on the Gram matrix, similar statements can be found in Bickel, Ritov, and Tsybakow (2009):

Condition (**RE**). *For a given $\bar{c} > 0$, there exists a constant $\kappa(\bar{c})$ such that*

$$\kappa(\bar{c}) = \inf_{||\delta_{T^c}||_1 \leqslant \bar{c}||\delta_T||_1, \delta \neq 0} \frac{\delta' X_n' X_n \delta}{n||\delta||_2^2} > 0$$

For the analysis of post-model selection estimators, we also need the following restricted sparse eigenvalue condition on the empirical Gram matrix, we can find the same condition in Belloni and Chernozhukov (2013):

Condition (**RSE**). *For any given $m < n - s$, there exists finite positive constants τ_x and ω_x, such that*

$$\tau_x = \inf_{||\delta_{T^c}||_0 \leqslant m, \delta \neq 0} \frac{\delta' X_n' X_n \delta}{n||\delta||_2^2} > 0,$$

$$\omega_x = \sup_{||\delta_{T^c}||_0 \leqslant m, \delta \neq 0} \frac{\delta' X_n' X_n \delta}{n||\delta||_2^2} < \infty$$

An extended condition can be derived for Condition (RE) in order to cater to the spatial autoregressive error.

Lemma 3.1. *Suppose Condition (RE) holds for $\bar{c} = \frac{c+1}{c-1}$, for some $c > 1$, and that $\Sigma(\rho) = (I - \rho M_n)'(I - \rho M_n)$, then for the generalized moments estimator $\hat{\rho}_n$, when n is large enough,*

we have

$$\kappa(\bar{c})\mu_{\min}(\Sigma(\rho)) = \inf_{||\delta_{T^c}||_1 \leqslant \bar{c}||\delta_T||_1, \delta \neq 0} \frac{\delta' X_n' \Sigma(\hat{\rho}_n) X_n \delta}{n||\delta||_2^2},$$

where $\mu_{\min}(\Sigma(\rho))$ *is the minimum eigenvalue of the matrix* $\Sigma(\rho)$.

From the Lemma 3.1 we can see that the effect of spatial autoregressive error in the model will add a coefficient depending on the eigenvalues of $\Sigma(\rho)$. And similar extension can be derived for Condition (RSE) as well. The following theorem constructs the main estimation properties of the generalized Lasso estimator $\hat{\beta}$.

Theorem 3.1. *Suppose that Assumption 1-5 and 8 holds, and Condition (RE) hold for* $\bar{c} = \frac{c+1}{c-1}$, *for some* $c > 1$. *Choose the regularization parameter* $\lambda_n \geqslant 2cb(\exp(t^2/2) + 1)\sqrt{\frac{\log 2p}{n}}$, *then with probability at least* $1 - K\exp\{-t^2/2\}$, $t > 0$, *we have*

$$||\hat{\beta} - \beta||_2 \leqslant \frac{(1 + \frac{1}{c})\sqrt{s}\lambda_n}{\kappa(\bar{c})\mu_{\min}(\Sigma(\rho))}$$

The bound for ℓ_2 norm of the parameter estimation error is derived when the disturbance ε_n is distribution free, and the convergence rate of the ℓ_2 error is $\sqrt{\frac{s\log p}{n}}$. A lower bound for the regularization λ_n is required in order to control the randomness brought by ε_n. In practice, we would like to select the regularization parameter to be close to the lower bound since too much penalization will have a negative effect on the selection capability. In fact, the estimation error converges to zero in ℓ_2 norm when $\sqrt{s}\lambda_n \to 0$.

In the preceding paragraphs, we will discuss the model selection properties of the generalized lasso estimator and provide the bounds on the false positive selected variables.

Theorem 3.2. *(1) If the coefficients are well separated from 0, that is,*

$$\min_{j \in T} |\beta_{0j}| > \zeta + t, \quad forsome \ \ t \geqslant 0, \ \ \zeta = \max_{j=1,\cdots,p} |\hat{\beta}_j - \beta_{0j}|,$$

then under the conditions used in Theorem 3.1, the true model is a subset of the selected model, $T := support(\beta_0) \subseteq \hat{T} := support(\hat{\beta}_n)$*, with high probability.*

(2) Suppose Assumption 1-5, Assumption 8 and Condition (RE), (RSE) holds, and choose the regularization parameter λ_n *same as in Theorem 3.1, then with probability at least* $1 - K\exp\{-t^2/2\}$*,* $t > 0$*, we can have upper bounds on the number of noise variables* $\hat{m} = card(\hat{T} - T)$*, that is,*

$$\hat{m} \lesssim s.$$

From Theorem 3.2, we obtain that the selected support set of the estimator from ℓ_1 penalization $\hat{T}$ contains the true model T with a high probability converges to 1 and also, the number of noise variables selected $\hat{m}$ is bounded by a value in the same order as the cardinality s. These are the two variable selection properties we need in order to proceed to the properties of estimators derived from post-model selection estimation. In fact, many of the known variable selection properties meet these requirements, but we are focusing on the ℓ_1 penalization for better illustration.

3.3 Post-model estimation properties

In this section, we will present a general result on the performance of a post-model selection estimator $\hat{\beta}_p$ with the model selected from previous steps, in terms of both the ℓ_2 norm and *sup* norm error convergence rate. We will show that the estimator $\hat{\beta}_p$ can perform at least as well as the estimates provided by ℓ_1 penalization estimation and even strictly outperforms $\hat{\beta}$ if certain properties are achieved by the first step model selection.

We define the least-square post-model selection selection estimator as

$$\hat{\beta}_p = \arg\min_{\beta} \frac{1}{n}||(I - \hat{\rho}M_n)(Y_n - X_n\beta)||_2^2, \quad where \;\; \beta_j = 0, \;\; j \in \hat{T}^c. \tag{3.1}$$

The following lemma provides an upper bound for a stochastic term involving the disturbances ε_n, and it is a crucial part for the derivation of ℓ_2 convergence rate theorem.

Lemma 3.2. *Suppose Assumption 1-5, 8 and Condition (RSE), $\hat{\rho}_n$ is a generalized moments estimator for ρ, for $m = 1, 2, \cdots, n - s$, define*

$$e_n(m, \eta) = \frac{2\sigma\sqrt{\omega_x \mu_{min}(\Sigma(\rho))}}{\sqrt{n}}\left(\sqrt{\log \binom{p}{m}} + \sqrt{(m+s)\log(3D)} + \sqrt{(m+s) + \log\frac{1}{\eta}}\right)$$

for any $\eta \in (0, 1)$ and some constant D. Then for all m,

$$\sup_{||\delta_{T^c}||_0 \leqslant m, ||\delta||_2 > 0} \left|\frac{\varepsilon_n'(I - \rho M_n')^{-1}\Sigma(\hat{\rho}_n)X_n\delta}{n||\delta||_2}\right| \leqslant e_n(m, \eta),$$

with probability at least $1 - \eta\exp^{-s}(1 - 1/e)$.

Now with the lemma established above, with a high probability, the ℓ_2 error bounds of the post-model selection estimator $\hat{\beta}_p$ is obtained thereby.

Theorem 3.3. *Let $\hat{\beta}$ be any estimator from a model selector and let $\hat{T} = supp(\hat{\beta})$. Suppose $\hat{\beta}$ satisfies $T \subseteq \hat{T}$, the true support and $\hat{m} = card(\hat{T} - T) \lesssim s$. Now if we let $\hat{\beta}_p$ be the post-model selection estimator defined in* (3.1)*, with assumption 1-6 and Condition (RSE), we have for any $\eta \in (0, 1)$ with probability at least $1 - \eta\exp^{-s}(1 - 1/e)$ that*

$$||\hat{\beta}_p - \beta_0||_2 \leqslant \frac{2}{\tau_x \mu_{min}(\Sigma(\rho))} e_n(\hat{m}, \eta).$$

This theorem establishes a bound for the ℓ_2 error bounds of the post-model selection. The convergence rate of this bound implies that the least square post-model selection estimator performs just as well as the ℓ_1 penalization. And in fact, based on the summarization corollary below, the performance of the post-model selection estimator becomes strictly better than generalized Lasso when the number of noise variables selected $\hat{m}$ goes to 0 with probability to 1. And if the model selection step manages to perfectly select the true model, that is, $\hat{T} = T$, with high probability, the ℓ_2 estimation error can achieve the $\sqrt{\frac{s}{n}}$ oracle rate of convergence.

Corollary 3.1. *Let $\hat{\beta}_p$ be the post-model selection estimator defined in* (3.1)*, then under the conditions from Theorem 3.3, we can get directly*

$$||\hat{\beta}_p - \beta_0||_2 \lesssim \begin{cases} \sqrt{\frac{s \log p}{n}}, & \text{in general} \\ \sqrt{\frac{o(1) s \log p}{n}}, & \text{if } T \subseteq \hat{T}, \hat{m} = o(s), \text{w.p.} \to 1 \\ \sqrt{\frac{s}{n}}, & \text{if } T = \hat{T}, \text{w.p.} \to 1 \end{cases}$$

Now that we have achieve the estimation ℓ_2 error bound, we now proceed to construct the convergence rate of the estimation error in the form of *sup* norm. At this point, we need one more assumption of the design matrix.

Assumption 9. *Define $\Psi = \frac{X_n' \Sigma(\hat{\rho}_n) X_n}{n}$, with out loss of generalization, assume $\Psi_{i,i} = 1$, and $\max_{i \neq j} |\Psi_{i,j}| \leqslant \frac{c}{s}$ for some constant c.*

Similar assumption has been made in Donoho, Elad and Temlyakov (2006), where the authors require that the value of $\max_{i \neq j} |\Psi_{i,j}|$ to be sufficiently small. In fact, this can also

be derived from Assumption 8. With these, now we are able to proceed to the result of convergence rate of the estimation error in *sup* norm.

Lemma 3.3. *Let Assumption 1-5, 8-9 be satisfied. And for any* $\eta \in (0,1)$*, take*

$$h_n(m,\eta) := \sqrt{\frac{2}{n}}\sigma(\sqrt{m\log p} + \sqrt{\log(m+s)} + \sqrt{\log(\frac{1}{\eta})},$$

then with a probability at least 1-$2\eta p/(p-1)$*,*

$$\max_{i\in\tilde{T}, card(\tilde{T}-T)\leqslant m} \frac{1}{n}|\sum_{j=1}^{n} T_{i,j}\varepsilon_j| \leqslant h_n(m,\eta), \quad for \ \ any \ \ m \leqslant n-s,$$

where $T_{i,j}$ *is the element in the matrix* $T = X_n'\Sigma(\hat{\rho}_n)(I-\rho M_n)^{-1}$*.*

Theorem 3.4. *Let* $\hat{T}$ *be the support of any first step model selector, and let* $h_n(m,\eta)$ *be the same function defined in Lemma 3.3, then assume Assumption 1-5, 8-9 and Condition (RSE) holds, then we have with high probability,*

$$||\hat{\beta}_p - \beta_0||_\infty \leqslant (1 + \frac{c(\hat{m}+s)}{s\mu_{min}(\Sigma(\rho))\tau_x})h_n(\hat{m},\eta).$$

Theorem 3.4 can be used to derive the rates of convergence of the estimation error in *sup* norm of the post model selection estimator and the result is summarized as follows:

Corollary 3.2. *Let* $\hat{\beta}_p$ *be the post-model selection estimator defined in* (3.1)*, then under the*

conditions from Theorem 3.4, we will get

$$||\hat{\beta}_p - \beta_0||_\infty \lesssim \begin{cases} \sqrt{\frac{s \log p}{n}}, & \textit{in general} \\ \sqrt{\frac{o(s) \log p}{n}}, & \textit{if } T \subseteq \hat{T}, \hat{m} = o(s), \textit{w.p.} \to 1 \\ \sqrt{\frac{\log s}{n}}, & \textit{if } T = \hat{T}, \textit{w.p.} \to 1 \end{cases}$$

3.4 Simulation Studies

In this section, we will use computational simulations to show how the post-model selection estimator $\hat{\beta}_p$ defined earlier outperform the simultaneous variable selection and estimation $\hat{\beta}$ in terms of the ℓ_2 error rate. The distribution of ε_n in the Monte Carlo study is always set to be normal, and without loss of generality, $N(0, 1)$. This is because the estimators for ρ defined earlier do not depend on σ^2. The weight matrix M_n is defined as an idealized $n \times n$ weighting matrix in a "circular world" following Kelejian and Prucha (1999), and specifies M_n such that each element of u_{ni} is directly related to the 5 elements immediately before and after it. For simplicity, we specify M_n such that all the non-zero elements of M_n are equal and that the respective rows sum to 1. In the $n \times p$ design matrix X_n, the covariates X_i's are i.i.d. from a p dimensional Gaussian distribution with each component having mean zero and variance 1. And the pairwise correlation is set to be $cor(x_{ij}, x_{ik}) = 0.5^{|j-k|}$, for $1 \leqslant j, k \leqslant p$. The first $q = 20$ non-zero components of the p dimensional parameter of interest β_0 are generated independently from a uniform distribution over the interval $(-2, 5)$. We consider 4 different choices of ρ, along with $4 \times 4 = 16$ combinations of (n, p), and this will give us 64 model settings in total. For each case, the results are summarized over 100

Monte Carlo replications. At the end of calculations in each setting, we record and compare the Relative Estimation Error, which is defined by $||\beta - \beta_0||_2/||\beta_0||_2$, of the estimate β, for the post-model selection estimator $\hat{\beta}_p$, the generalized lasso estimator $\hat{\beta}$, and the oracle estimates for the spatial error model with only the true non zero parameters. We denote them as REE_p, REE_g, and REE_o, respectively. Besides, the number in the parenthesis record the sum of estimation variance for the 20 significant parameters.

The ℓ_1 penalization computation used in finding the generalized lasso estimator and involved in getting the post-model selection is achieved by using the "glmnet" in the R package developed by Friedman et. al. (2010), and the penalty level λ is chosen by cross validation controlled by a data-driven choice of lower bound. The idea is that, based on the proof, λ is chosen to dominate the randomness brought by ε_n, that is,

$$\lambda_n \geqslant 2c \max_{1 \leqslant j \leqslant p} |\varepsilon_n' T^{(j)}|/n, \quad with\ probability\ at\ least\ 1-\alpha,$$

where $T^{(j)}$ is the jth column of the matrix $T = (I - \rho M_n')^{-1}\Sigma(\hat{\rho}_n)X_n$, probability 1-$\alpha$ needs to be close to 1 and c is a constant greater than 1. Therefore, the lower bound for the penalization is proposed to be

$$\lambda = c'\hat{\sigma}Q(1-\alpha|X, \hat{\rho}_n), \quad for\ some\ fixed\ c' > c > 1,$$

here $Q(1-\alpha|X, \hat{\rho}_n)$ is the maximum $(1-\alpha)$ quantile of $|z_n' T^{(j)}|/n$, where z_n is a $n \times 1$ standard normal vector and $\hat{\sigma}$ is the estimate of σ. Table 3.1 to 3.4 display the superiority of the post-model selection over the simultaneous variable selection and estimation method via ℓ_1 penalization and they also show the post-model selection estimator reach the same order

of estimation error rate as the oracle estimator as the sample size is large enough. Figure 3.1 to 3.4 compare the coverage rate of the 20 significant variables in each pair of n, p scenario at each value of ρ. Here, the coverage is defined so that the true value of each variable is located within the 95% confidence interval constructed from the estimates.

Table 3.1: Means of REE for $\hat{\beta}_p$, $\hat{\beta}$ and $\hat{\beta}_{oracle}$ of 100 data sets repetition for $\rho = 0.3$

		p=500	p=800	p=1000	p=1200
n=225	REE_p	0.084(0.159)	0.101(0.189)	0.119(0.213)	0.123(0.224)
	REE_g	0.212	0.216	0.230	0.226
	REE_o	0.031(0.101)	0.029(0.099)	0.031(0.102)	0.029(0.100)
n=400	REE_p	0.032(0.058)	0.037(0.064)	0.041(0.066)	0.043(0.072)
	REE_g	0.137	0.137	0.137	0.137
	REE_o	0.021(0.055)	0.022(0.056)	0.021(0.056)	0.021(0.057)
n=625	REE_p	0.022(0.034)	0.022(0.035)	0.023(0.035)	0.024(0.035)
	REE_g	0.116	0.114	0.112	0.110
	REE_o	0.017(0.035)	0.017(0.035)	0.017(0.035)	0.017(0.035)
n=900	REE_p	0.016(0.023)	0.016(0.023)	0.017(0.023)	0.016(0.023)
	REE_g	0.100	0.096	0.096	0.093
	REE_o	0.014(0.024)	0.014(0.024)	0.015(0.024)	0.014(0.024)

Table 3.2: Means of REE for $\hat{\beta}_p$, $\hat{\beta}$ and $\hat{\beta}_{oracle}$ of 100 data sets repetition for $\rho = 0.75$

		p=500	p=800	p=1000	p=1200
n=225	REE_p	0.101(0.221)	0.119(0.261)	0.135(0.299)	0.144(0.290)
	REE_g	0.232	0.221	0.236	0.268
	REE_o	0.035(0.134)	0.034(0.129)	0.032(0.134)	0.031(0.126)
n=400	REE_p	0.059(0.142)	0.085(0.176)	0.083(0.167)	0.075(0.173)
	REE_g	0.179	0.166	0.152	0.137
	REE_o	0.028(0.120)	0.028(0.114)	0.028(0.106)	0.027(0.108)
n=625	REE_p	0.047(0.173)	0.073(0.209)	0.077(0.219)	0.081(0.230)
	REE_g	0.218	0.173	0.246	0.222
	REE_o	0.033(0.197)	0.033(0.187)	0.032(0.194)	0.031(0.191)
n=900	REE_p	0.044(0.173)	0.041(0.177)	0.053(0.215)	0.047(0.169)
	REE_g	0.192	0.145	0.138	0.200
	REE_o	0.031(0.190)	0.031(0.189)	0.031(0.207)	0.030(0.182)

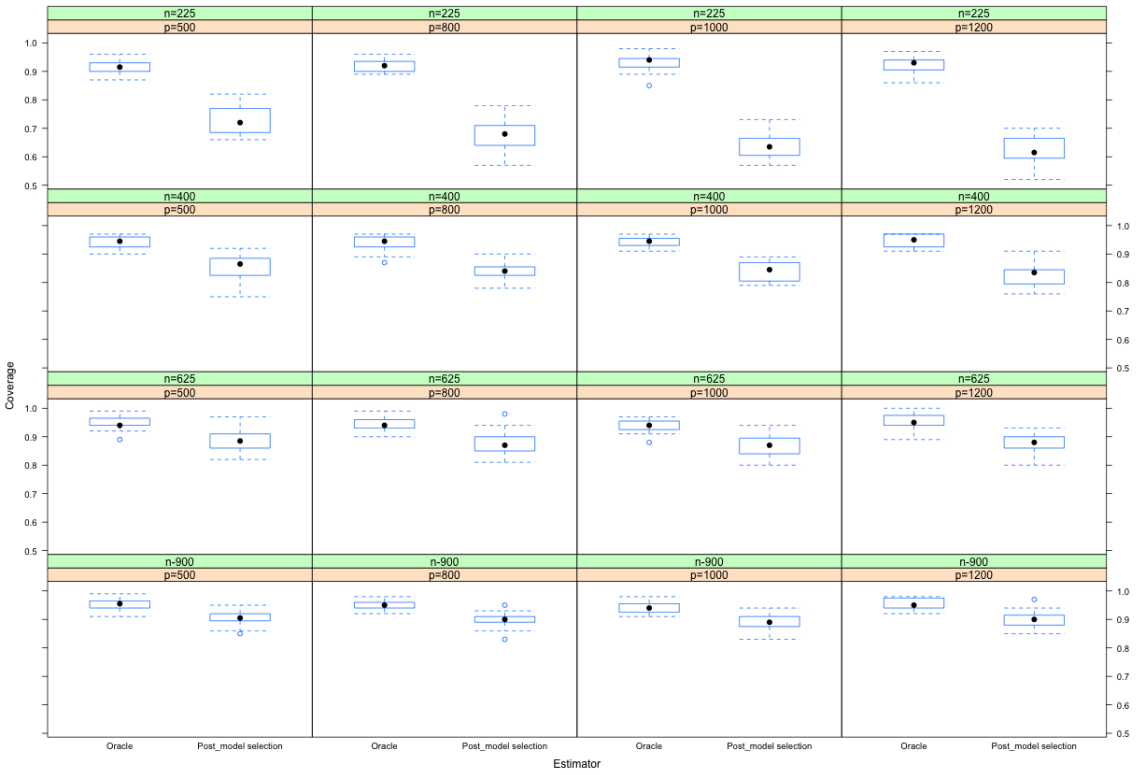

Figure 3.1: **Coverage rate of post-model selection and oracle estimators for** $\rho = 0.3$

Figure 3.2: **Coverage rate of post-model selection and oracle estimators for** $\rho = 0.75$

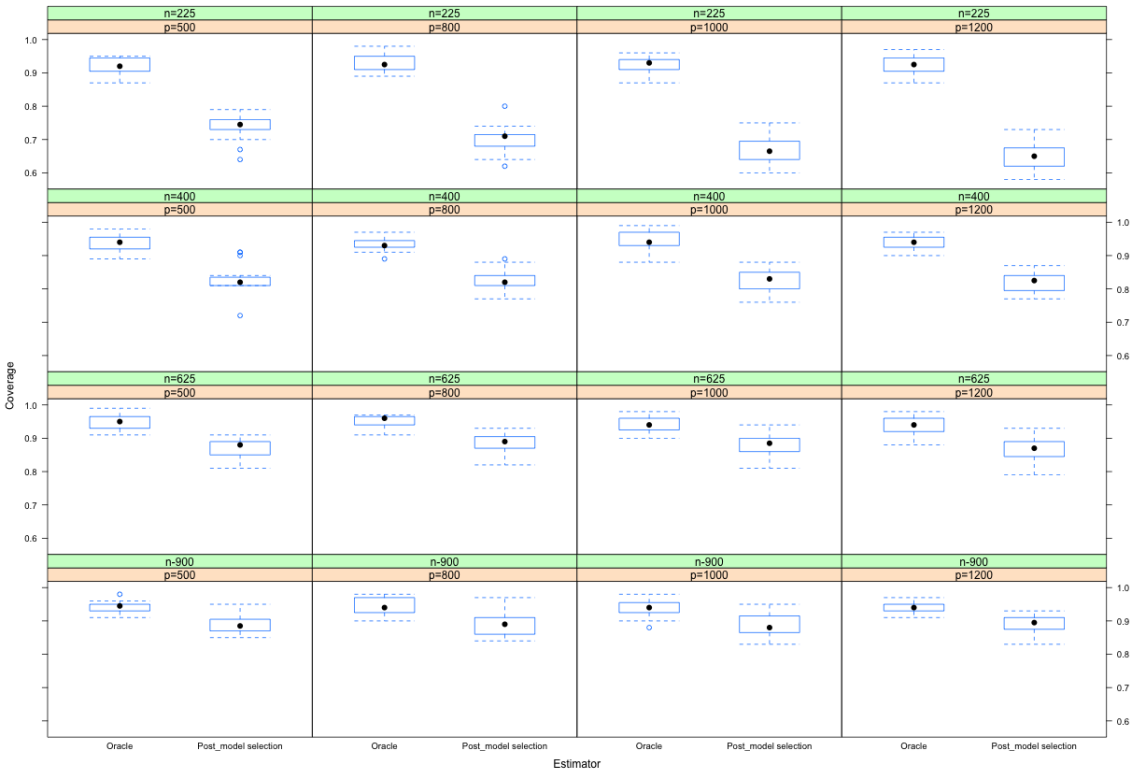

Figure 3.3: **Coverage rate of post-model selection and oracle estimators for** $\rho = -0.3$

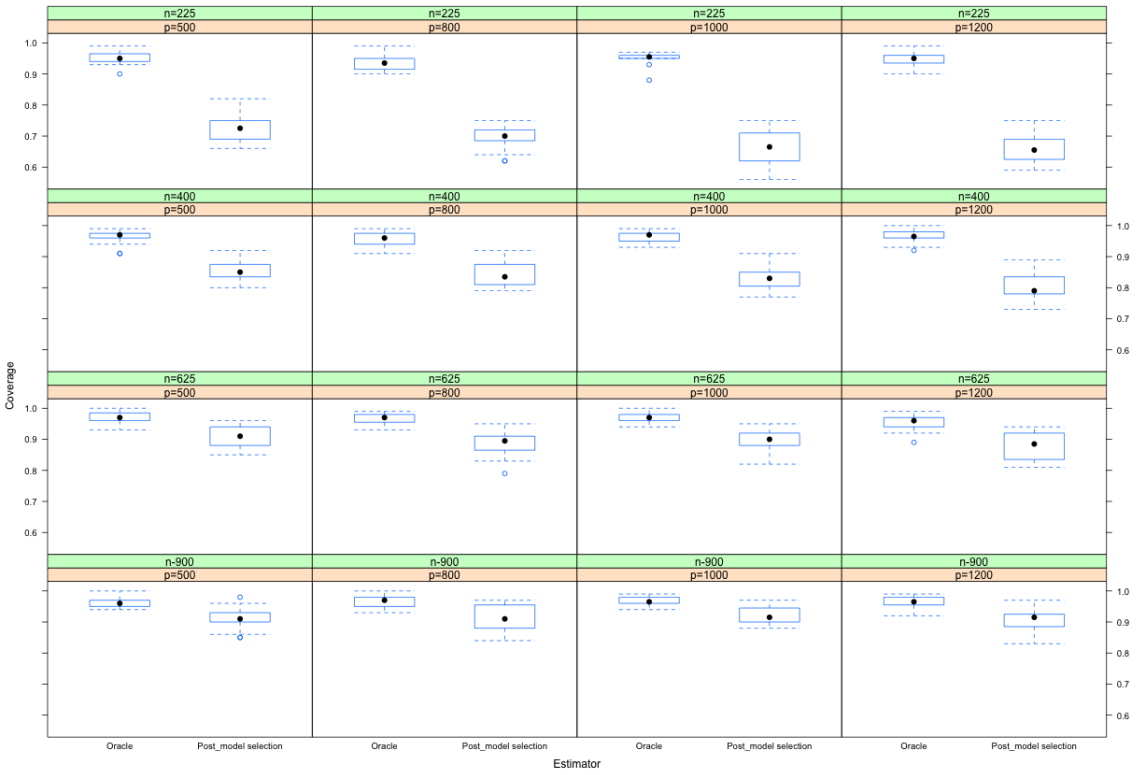

Figure 3.4: **Coverage rate of post-model selection and oracle estimators for** $\rho = -0.75$

Table 3.3: Means of REE for $\hat{\beta}_p$, $\hat{\beta}$ and $\hat{\beta}_{oracle}$ of 100 data sets repetition for $\rho = -0.3$

		p=500	p=800	p=1000	p=1200
n=225	REE_p	0.082(0.168)	0.091(0.189)	0.106(0.217)	0.120(0.241)
	REE_g	0.221	0.228	0.258	0.258
	REE_o	0.028(0.086)	0.028(0.087)	0.029(0.085)	0.028(0.083)
n=400	REE_p	0.034(0.054)	0.039(0.061)	0.042(0.065)	0.042(0.063)
	REE_g	0.137	0.148	0.144	0.148
	REE_o	0.020(0.045)	0.019(0.045)	0.019(0.045)	0.020(0.045)
n=625	REE_p	0.020(0.029)	0.020(0.029)	0.022(0.030)	0.023(0.030)
	REE_g	0.105	0.108	0.110	0.108
	REE_o	0.015(0.028)	0.015(0.028)	0.015(0.028)	0.015(0.028)
n=900	REE_p	0.015(0.019)	0.016(0.019)	0.015(0.019)	0.016(0.019)
	REE_g	0.088	0.087	0.088	0.090
	REE_o	0.012(0.019)	0.013(0.019)	0.012(0.019)	0.012(0.019)

Table 3.4: Means of REE for $\hat{\beta}_p$, $\hat{\beta}$ and $\hat{\beta}_{oracle}$ of 100 data sets repetition for $\rho = -0.75$

		p=500	p=800	p=1000	p=1200
n=225	REE_p	0.092(0.197)	0.102(0.220)	0.114(0.253)	0.115(0.252)
	REE_g	0.215	0.217	0.235	0.235
	REE_o	0.026(0.093)	0.027(0.092)	0.026(0.094)	0.026(0.092)
n=400	REE_p	0.037(0.062)	0.039(0.070)	0.045(0.077)	0.051(0.082)
	REE_g	0.126	0.130	0.134	0.142
	REE_o	0.018(0.048)	0.018(0.048)	0.017(0.048)	0.019(0.048)
n=625	REE_p	0.021(0.033)	0.024(0.034)	0.024(0.034)	0.025(0.034)
	REE_g	0.099	0.100	0.101	0.099
	REE_o	0.014(0.030)	0.015(0.030)	0.014(0.029)	0.015(0.029)
n=900	REE_p	0.015(0.020)	0.015(0.021)	0.015(0.021)	0.016(0.021)
	REE_g	0.080	0.078	0.079	0.079
	REE_o	0.012(0.020)	0.011(0.020)	0.011(0.020)	0.011(0.020)

3.5 Real Data Example

In this section, we are going to apply the proposed method to work on a small real life data example as illustration. The data we chose is a built in sample data set in the R package "spdep", which can also be found in Anselin (1988) book. It includes 49 samples describing the Columbus crime including the necessary spatial information. This data set we use originally is not high-dimensional, but we intentionally pick this one since the low-dimensional nature of the data set can be used as a criterion to check capability of variable

selection.

The "classic" Columbus crime regression is to predict the variable Crime, the residential burglaries and auto thefts per 1000 households with variables HOVAL, the house value and INC, income amount. The Lagrange Multiplier Test Statistics for spatial dependence is significant for spatial error models and an initial analysis with the sample data returns an estimate of $\hat{\beta}_{HOVAL} = -1.17$, $\hat{\beta}_{INC} = -0.30$, with standard error 0.348 and 0.095, respectively. Both the house value and house income have a negative impact on the burglaries and auto thefts rate. To test the power of variable selection for spatial error models, we manually input 500 spurious covariates that are uncorrelated with the response, so that the sample dataset becomes high-dimensional with 49×502 dimension. Now a generalized ℓ_1 penalty variable selection method is applied to the sample dataset and the penalization parameter λ is chosen by cross-validation with a data-driven lower bound. The variable selection method shows great competence by correctly selecting the only two authentic variables, with a estimate of HOVAL and INC -0.4592141and -0.1476459, respectively. A post-model selection estimation is conducted with the selected two variables and it turns out returning a much accurate prediction, with parameter estimates $\hat{\beta}_{HOVAL} = -0.99$, $\hat{\beta}_{INC} = -0.31$, and their standard error 0.348 and 0.095, the same with the oracle case.

3.6 Proofs

PROOF of Lemma 3.1. Note that the generalized moments estimator $\hat{\rho}_n$ is a consistent estimator for the spatial autoregressive parameter ρ, then

$$\begin{aligned}
\delta' X_n' \Sigma(\hat{\rho}_n) X_n \delta &= \delta' X_n' \Sigma(\rho) X_n \delta + \delta' X_n' \Sigma(\hat{\rho}_n) X_n \delta - \delta' X_n' \Sigma(\rho) X_n \delta \\
&= \delta' X_n' \Sigma(\rho) X_n \delta + \delta' X_n' [(\rho - \hat{\rho}_n)(M_n' + M_n) + (\hat{\rho}_n^2 - \rho^2) M_n' M_n] X_n \delta \\
&= \Delta_1 + \Delta_2
\end{aligned}$$

where

$$\Delta_1 = \delta' X_n' \Sigma(\rho) X_n \delta,$$

$$\Delta_2 = \delta' X_n' [(\rho - \hat{\rho}_n)(M_n' + M_n) + (\hat{\rho}_n^2 - \rho^2) M_n' M_n] X_n \delta.$$

Now if we look at the term Δ_2, since $\hat{\rho}_n - \rho \to_p 0$, $\hat{\rho}_n^2 - \rho^2 \to_p 0$ and the boundedness condition of X_n and M_n, when n becomes large enough,

$$\frac{\Delta_2}{n||\delta||_2^2} \to 0$$

And thus,

$$\inf_{||\delta_{T^c}||_1 \leqslant \bar{c}||\delta_T||_1, \delta \neq 0} \frac{\delta' X_n' \Sigma(\hat{\rho}_n) X_n \delta}{n||\delta||_2^2} = \inf_{||\delta_{T^c}||_1 \leqslant \bar{c}||\delta_T||_1, \delta \neq 0} \frac{\Delta_1}{n||\delta||_2^2}$$

Already known that $\Sigma(\rho) = (I-\rho M_n)'(I-\rho M_n)$ is a symmetric and positive definite matrix, so there exists a unique decomposition of $\Sigma(\rho)$,

$$\Sigma(\rho) = Q'UQ,$$

where

$$U = diag(\mu_1, \cdots, \mu_n)$$

is the diagonal matrix composed of the eigenvalues of $\Sigma(\rho)$, and Q is an orthogonal matrix. Based on these,

$$\begin{aligned} \frac{\Delta_1}{n||\delta||_2^2} &= \frac{\delta' X_n' Q' U Q X_n \delta}{n||\delta||_2^2} \\ &= \frac{\Sigma_{i=1}^n \mu_i (QX_n\delta)_i^2}{n||\delta||_2^2} \\ &\geqslant \mu_{\min} \frac{\delta' X_n' X_n \delta}{n||\delta||_2^2} \end{aligned}$$

Combine with Condition (RE), we yield result of Lemma 3.1.

PROOF of Theorem 3.1. By definition, the generalized Lasso estimator $\hat{\beta}$ can be expressed as

$$\hat{\beta} = \arg\min_{\beta} \frac{1}{n}(Y_n - X_n\beta)'\Sigma(\hat{\rho}_n)(Y_b - X_n\beta) + \lambda_n||\beta||_1,$$

where $\Sigma(\hat{\rho}_n) = (I - \hat{\rho}_n M_n)'(I - \hat{\rho}_n M_n)$.

Thus, if we denote β_0 as the true value for parameter β, intuitively,

$$\frac{1}{n}(Y_n - X_n\hat{\beta})'\Sigma(\hat{\rho}_n)(Y_n - X_n\hat{\beta}) + \lambda_n||\hat{\beta}||_1 \leqslant \frac{1}{n}(Y_n - X_n\beta_0)'\Sigma(\hat{\rho}_n)(Y_n - X_n\beta_0) + \lambda_n||\beta_0||_1. \tag{3.2}$$

Since,

$$\begin{aligned}
& \frac{1}{n}(Y_n - X_n\hat{\beta})'\Sigma(\hat{\rho}_n)(Y_n - X_n\hat{\beta}) - \frac{1}{n}(Y_n - X_n\beta_0)'\Sigma(\hat{\rho}_n)(Y_n - X_n\beta_0) \\
= & \frac{1}{n}[X_n(\beta_0 - \hat{\beta}) + (I - \rho M_n)^{-1}\varepsilon_n]'\Sigma(\hat{\rho}_n)[X_n(\beta_0 - \hat{\beta}) + (I - \rho M_n)^{-1}\varepsilon_n] \\
& -\frac{1}{n}[(I - \rho M_n)^{-1}\varepsilon_n]'\Sigma(\hat{\rho}_n)[(I - \rho M_n)^{-1}\varepsilon_n] \\
= & \frac{1}{n}[X_n(\beta_0 - \hat{\beta})]'\Sigma(\hat{\rho}_n)[X_n(\beta_0 - \hat{\beta})] + 2\frac{1}{n}\varepsilon_n'(I - \rho M_n')^{-1}\Sigma(\hat{\rho}_n)X_n(\beta_0 - \hat{\beta}) \\
\geqslant & \frac{1}{n}||(I - \hat{\rho}_n M_n)X_n(\beta_0 - \hat{\beta})||_2^2 - 2\frac{1}{n}(\max_{1\leqslant j\leqslant p}|\varepsilon_n'T^{(j)}|)||\hat{\beta} - \beta_0||_1
\end{aligned}$$

where $T^{(j)}$ is the jth column of the matrix $T = (I - \rho M_n')^{-1}\Sigma(\hat{\rho})X_n$. Then according to the result in previous chapter, the set

$$\Im := \left\{\max_{1\leqslant j\leqslant p} 2|\epsilon_n'T^{(j)}|/n \leqslant \lambda_0\right\}$$

in which the random part can be get rid of, has a probability at least $1 - K\exp[-t^2/2]$, where $\lambda_0 = 2\sigma(\exp[t^2/2] + 1)\sqrt{\frac{\log 2p}{n}}$. Now assume an arbitrarily constant $c > 1$, so that $\lambda_n \geqslant c\lambda_0$, then on the set $\Im$,

$$-2\frac{1}{n}(\max_{1\leqslant j\leqslant p}|\varepsilon_n'T^{(j)}|)||\hat{\beta} - \beta_0||_1 \geqslant -\frac{\lambda_n}{c}||\hat{\beta} - \beta_0||_1$$

Now bring the result back to (3.2), on the set $\Im$,

$$\frac{1}{n}(Y_n - X_n\hat{\beta})'\Sigma(\hat{\rho}_n)(Y_n - X_n\hat{\beta}) + \lambda_n||\hat{\beta}||_1 \leqslant \frac{\lambda_n}{c}||\hat{\beta} - \beta_0||_1 + \lambda_n||\beta_0||_1. \tag{3.3}$$

With simple transformation,

$$||\hat{\beta}||_1 = ||\hat{\beta}_T||_1 + ||\hat{\beta}_{T^c}||_1 \geqslant ||\beta_{0T}||_1 - ||\hat{\beta}_T - \beta_{0T}||_1 + ||\hat{\beta}_{T^c}||_1,$$

and

$$||\hat{\beta} - \beta_0||_1 = ||\hat{\beta}_T - \beta_{0T}||_1 + ||\hat{\beta}_{T^c}||_1,$$

we can have the relationship of the difference between the generalized Lasso estimator and the true value of the parameter on the support and non-support set,

$$\frac{c}{n}(Y_n - X_n\hat{\beta})'\Sigma(\hat{\rho}_n)(Y_n - X_n\hat{\beta}) + (c-1)\lambda_n||\hat{\beta}_{T^c}||_1 \leqslant (c+1)\lambda_n||\hat{\beta}_T - \beta_{0T}||_1$$

Denote $\phi = \hat{\beta} - \beta_0$ for notation simplicity, since the first term on the left hand side is positive, thus on $\Im$,

$$||\phi_{T^c}||_1 \leqslant \frac{c+1}{c-1}||\phi_T||_1$$

Thus ϕ belongs to the restricted set in condition $RE(\bar{c})$, where $\bar{c} = \frac{c-1}{c+1}$, and we can have

$$\begin{aligned}\frac{1}{n}(Y_n - X_n\hat{\beta})'\Sigma(\hat{\rho}_n)(Y_n - X_n\hat{\beta}) &= \frac{1}{n}||(I - \hat{\rho}M_n)X_n\phi||_2^2 \\ &\geqslant \kappa(\bar{c})\mu_{\min}(\Sigma(\rho))||\phi||_2^2.\end{aligned}$$

Based on (3.3),

$$\kappa(\bar{c})\mu_{\min}(\Sigma(\rho))||\phi||_2^2 - \frac{\lambda_n}{c}(||\phi_T||_1 + ||\phi_{T^c}||_1) \leqslant \lambda_n(||\phi_T||_1 - ||\phi_{T^c}||_1),$$

$$\begin{aligned}
\kappa(\bar{c})\mu_{\min}(\Sigma(\rho))||\phi||_2^2 &\leqslant (1+\frac{1}{c})\lambda_n||\phi_T||_1 - (1-\frac{1}{c})\lambda_n||\phi_{T^c}||_1 \\
&\leqslant (1+\frac{1}{c})\lambda_n||\phi_T||_1 \\
&\leqslant \sqrt{s}(1+\frac{1}{c})\lambda_n||\phi_T||_2.
\end{aligned}$$

The last inequality makes use of Cauchy-Schwarz inequality, and $\sqrt{s}$ is the price to pay when you replace the ℓ_1 with ℓ_2 norm. And thus we finish the proof with bound of the ℓ_2 norm of the estimator,

$$||\hat{\beta} - \beta_0||_2 \leqslant \frac{(1+\frac{1}{c})\sqrt{s}\lambda_n}{\kappa(\bar{c})\mu_{\min}(\Sigma(\rho))}.$$

PROOF of Theorem 3.2. (1) Based on the assumption of the magnitude of β_0, if $T \nsubseteq \hat{T}$, then $\exists k \in \{1, 2, \cdots, p\}$, such that $\beta_{0k} \neq 0$, but $\hat{\beta}_k = 0$. Thus for any $l \in \{1, 2, \cdots, p\}$,

$$||\hat{\beta}_l - \beta_{0l}||_2 \geqslant |\beta_{0k}| \geqslant \min_{j\in T}|\beta_{j0}| > \max_{j=1,2,\cdots,p}|\hat{\beta}_j - \beta_{0j}|.$$

A contradiction occurs, and thus $T \subseteq \hat{T}$.

(2) Recall the definition of the generalized Lasso estimator defined in (2.6), and make use of

the Karush-Kuhn-Tucker condition, for $\forall j \in \hat{T}$,

$$\begin{aligned} & \frac{d||\frac{1}{n}(I-\hat{\rho}_n M_n)Y_n - (I-\hat{\rho}_n M_n)X_n\beta||_2^2}{d\beta_i}|_{\beta_j=\hat{\beta}_j} \\ = \; & [-2\frac{1}{n}(Y_n - X_n\beta)'\Sigma(\hat{\rho}_n)X_n]_j|_{\beta_j=\hat{\beta}_j} \\ = \; & -\lambda_n sign(\hat{\beta}_j) \end{aligned}$$

Thus, $\forall j \in \hat{T}$,

$$|\frac{2}{n}(Y_n - X_n\hat{\beta})'\Sigma(\hat{\beta}_n)X_n|_j = \lambda_n.$$

Since we are only looking at the support set $\hat{T}$ of the estimated parameter $\hat{\beta}$,

$$\begin{aligned} \sqrt{|\hat{T}|}\lambda_n &= ||(\frac{2}{n}(Y_n - X_n\hat{\beta})'\Sigma(\hat{\beta}_n)X_n)_{\hat{T}}||_2 \\ &= 2||(\frac{1}{n}(Y_n - X_n\beta_0 + X_n\beta_0 - X_n\hat{\beta})'\Sigma(\hat{\rho}_n)X_n)_{\hat{T}}||_2 \\ &\leqslant 2||(\frac{1}{n}\varepsilon_n'(I-\rho M_n')^{-1}\Sigma(\hat{\rho})X_n)_{\hat{T}}||_2 + 2||(\frac{1}{n}(\beta_0-\hat{\beta})'X_n'\Sigma(\hat{\rho})X_n)_{\hat{T}}||_2 \\ &\leqslant \frac{\lambda_n}{c}\sqrt{|\hat{T}|} + 2\Delta, \end{aligned}$$

here $\Delta = ||(\frac{1}{n}(\beta_0-\hat{\beta})'X_n'\Sigma(\hat{\rho})X_n)_{\hat{T}}||_2$. The last inequality is satisfied by $\varepsilon_n's$ from the set $\Im$, which we know from Theorem 3.1, has a probability at least $1 - K\exp\{-t^2/2\}$. On the

other hand, using Holder inequality, and the extension of Condition (RSE),

$$\begin{aligned}
\Delta &= ||(\frac{1}{n}(\beta_0-\hat{\beta})'X_n'\Sigma(\hat{\rho})X_n)_{\hat{T}}||_2 \\
&\leqslant \sup_{||\delta_{T^c}||_0\leqslant\hat{m},||\delta||\leqslant 1}|\delta'\frac{1}{n}X_n'\Sigma(\hat{\rho}_n)X_n(\hat{\beta}-\beta_0)| \\
&\leqslant \sup_{||\delta_{T^c}||_0\leqslant\hat{m},||\delta||\leqslant 1}||\delta'\frac{1}{\sqrt{n}}X_n'(I-\hat{\rho}_nM_n)'||_2||\frac{1}{\sqrt{n}}(I-\hat{\rho}_n)X_n(\hat{\beta}-\beta_0)||_2 \\
&\leqslant \omega_x(\hat{m})\mu_{\max}(\Sigma(\rho))||\hat{\beta}-\beta_0||_2
\end{aligned}$$

To summarize,

$$(1-\frac{1}{c})\lambda_n\sqrt{|\hat{T}|}\leqslant 2\omega_x(\hat{m})\mu_{\max}(\Sigma(\rho))||\hat{\beta}-\beta_0||_2,$$

combined with the order of ℓ_2 norm of the difference between the generalized Lasso estimator $\hat{\beta}$ and β_0, easily to get

$$\hat{m}\lesssim s.$$

PROOF of Lemma 3.2. For each nonnegative integer $m\leqslant n-s$, and consider each set $\tilde{T}\subset\{1,2,\cdots,p\}$, with $card(\tilde{T}-T)\leqslant m$, define a class of functions

$$\mathcal{G}_{\tilde{T}}=\{f_\delta,\delta\in\mathcal{R}^p,support(\delta)\subseteq\tilde{T},||\delta||_1=1\},$$

where $f_\delta=\varepsilon_iD_i'\delta$, with $D_i'=[(I-\rho M_n')^{-1}\Sigma(\hat{\rho}_n)X_n]_{ith\ \ row}$. Further define the set

$$\mathcal{F}_m=\{\mathcal{G}_{\tilde{T}}:\tilde{T}\subseteq\{1,2,\cdots,p\},withcard(\tilde{T}-T)\leqslant m\}.$$

combining all possible choices of $\tilde{T}$. With the definition of $e_n(m,\eta)$, it follows directly,

$$P(\sup_{f\in\mathcal{F}_m}|\frac{1}{n}\sum_{i=1}^{n}f_\delta(\varepsilon_i)|\geqslant e_n(m,\eta))\leqslant\binom{p}{m}\max_{card(\tilde{T}-T)\leqslant m}P(\sup_{f\in\mathcal{G}_{\tilde{T}}}|\frac{1}{n}\sum_{i=1}^{n}f_\delta(\varepsilon_i)|\geqslant e_n(m,\eta)). \tag{3.4}$$

Now consider any two functions $f,g\in\mathcal{G}_{\tilde{T}}$, recall that $E\frac{1}{\sqrt{n}}\sum_{i=1}^{n}f_\delta(\varepsilon_i)=\frac{1}{\sqrt{n}}\sum_{i=1}^{n}g_\delta(\varepsilon_i)=0$, let

$$\gamma(f,g):=\sqrt{E[\frac{1}{\sqrt{n}}\sum_{i=1}^{n}f_\delta(\varepsilon_i)-\frac{1}{\sqrt{n}}\sum_{i=1}^{n}g_\delta(\varepsilon_i)]^2},$$

which can be seen as a "natural semimetric". Also, the covering number of $\mathcal{G}_{\tilde{T}}$ with respect to γ obeys

$$N(t,\mathcal{G}_{\tilde{T}},\gamma)\leqslant(\frac{3R}{t})^{m+s},\ \ for\ \ each\ \ 0<t<\sigma\sqrt{\omega_x\mu_{max}(\Sigma(\rho))}. \tag{3.5}$$

and $\sigma^2(\mathcal{G}_{\tilde{T}}):=\sup_{f\in\mathcal{G}_{\tilde{T}}}E[\frac{1}{\sqrt{n}}\sum_{i=1}^{n}f_\delta(\varepsilon_i)]^2=\sigma^2\omega_x\mu_{max}(\Sigma(\rho))$. Now in the following part we will detail the procedure to get (3.5).

For two p-dimensional vector δ, and $\tilde{\delta}\in\mathbb{R}^p$, consider two functions $f_\delta(\varepsilon_i)$ and $f_{\tilde{\delta}}(\varepsilon_i)$ in $\mathcal{G}_{\tilde{T}}$, for a given $\tilde{T}$ which satisfies $\tilde{T}\subset\{1,2,\cdots,p\}$ with $card(\tilde{T}-T)\leqslant m$. Then

$$\begin{aligned}
&\sqrt{E\frac{1}{n}\sum_{i=1}^{n}(f_\delta-f_{\tilde{\delta}})^2}\\
=\ &\sqrt{\frac{1}{n}\sigma^2||(I-\rho M_n')^{-1}\Sigma(\hat{\rho}_n)X_n(\delta-\tilde{\delta})||_2^2}\\
\leqslant\ &\sigma\sqrt{\omega_x\mu_{max}(\Sigma(\rho))}||\delta-\tilde{\delta}||_2,
\end{aligned}$$

when n is large enough. So

$$N(t, \mathcal{G}_{\tilde{T}}, \gamma) \leqslant N(t/R, B(0,1), ||\cdot||_2) \leqslant (\frac{3R}{t})^{m+s},$$

where $R = \sigma\sqrt{\omega_x \mu_{max}(\Sigma(\rho))}$ for any $0 < t < R$.

By Proposition A.2.7 of van der Vaart and Wellner, let $\bar{\Phi}(z) = \int_z^\infty \phi(x)dx \leqslant z^{-1}\phi(z)$ be the tail probability of a standard normal variable, then there exists a universal constant D such that,

$$P(\sup_{f\in\mathcal{G}_{\tilde{T}}} |\frac{1}{n}\sum_{i=1}^n f_\delta(\varepsilon_i)| \geqslant e_n(m,\eta)) \leqslant (\frac{3DR\sqrt{n}e_n(m,\eta)}{\sqrt{m+s}\sigma^2\omega_x\mu_{max}(\Sigma(\rho))})^{m+s}\bar{\Phi}(\frac{\sqrt{n}e_n(m,\eta)}{\sigma\sqrt{\omega_x\mu_{max}(\Sigma(\rho))}}).$$

By the definition of $e_n(m,\eta)$, and denote $E = \sqrt{n}e_n(m,\eta)$ for notation simplicity,

$$\begin{aligned}
& (\frac{3DR\sqrt{n}e_n(m,\eta)}{\sqrt{m+s}\sigma^2\omega_x\mu_{max}(\Sigma(\rho))})^{m+s}\bar{\Phi}(\frac{\sqrt{n}e_n(m,\eta)}{\sigma\sqrt{\omega_x\mu_{max}(\Sigma(\rho))}}) \\
\leqslant\ & \exp\{-\frac{E^2}{2\sigma^2\omega_x\mu_{max}(\Sigma(\rho))} + (m+s)\log\frac{E}{\sqrt{m+s}\sigma\sqrt{\omega_x\mu_{max}(\Sigma(\rho))}} + (m+s)\log 3D\} \\
=\ & \exp\{-\frac{m+s}{2}(\frac{E}{\sqrt{m+s}\sigma\sqrt{\omega_x\mu_{max}(\Sigma(\rho))}})^2 + (m+s)\log\frac{E}{\sqrt{m+s}\sigma\sqrt{\omega_x\mu_{max}(\Sigma(\rho))}} \\
& +(m+s)\log 3D\}.
\end{aligned}$$

Take $D \geqslant e/3$, so $\frac{E}{\sqrt{m+s}\sigma\sqrt{\omega_x\mu_{max}(\Sigma(\rho))}} \geqslant \sqrt{2}$, and combine with the fact that $\log x \leqslant \frac{x^2}{4}$,

if $x \geqslant \sqrt{2}$, then the inequalities continues as follows:

$$\begin{aligned}
& (\frac{3DR\sqrt{n}e_n(m,\eta)}{\sqrt{m+s}\sigma^2\omega_x\mu_{max}(\Sigma(\rho))})^{m+s}\bar{\Phi}(\frac{\sqrt{n}e_n(m,\eta)}{\sigma\sqrt{\omega_x\mu_{max}(\Sigma(\rho))}}) \\
\leqslant & \exp\{-\frac{m+s}{4}(\frac{E}{\sqrt{m+s}\sigma\sqrt{\omega_x\mu_{max}(\Sigma(\rho))}})^2 + (m+s)\log 3D\} \\
= & \exp\{-\frac{E^2}{4\sigma^2\omega_x\mu_{max}(\Sigma(\rho))} + (m+s)\log 3D\} \\
\leqslant & \exp\{-\log\binom{p}{m} - (m+s) - \log(\frac{1}{\eta})\}
\end{aligned}$$

thus the above probability is bounded by $\eta e^{-m-s}/\binom{p}{m}$. From (3.4),

$$P(\sup_{f\in\mathcal{F}_m}|\frac{1}{n}\sum_{i=1}^{n}f_\delta(\varepsilon_i)| \geqslant e_n(m,\eta)) \leqslant \eta e^{-m-s}$$

And thus

$$\begin{aligned}
P(\sup_{f\in\mathcal{F}_m}|\frac{1}{n}\sum_{i=1}^{n}f_\delta(\varepsilon_i)| \geqslant e_n(m,\eta), \exists m \leqslant n-s) & \leqslant \sum_{m=0}^{n}\eta e^{-m-s} \\
& \leqslant \eta e^{-s}/(1-1/e).
\end{aligned}$$

And Lemma 3.2 therefore is proved.

PROOF of Theorem 3.3. By the definition of $\hat{\beta}_p$,

$$\frac{||(I-\hat{\rho}_nM_n)(Y_n-X_n\hat{\beta}_p)||_2^2}{n} \leqslant \frac{||(I-\hat{\rho}_nM_n)(Y_n-X_n\beta_0)||_2^2}{n},$$

thus,

$$\begin{aligned}
& \frac{||(I-\hat{\rho}_n M_n)(Y_n - X_n\hat{\beta}_p)||_2^2}{n} - \frac{||(I-\hat{\rho}_n M_n)(Y_n - X_n\beta_0)||_2^2}{n} \\
= & \frac{(\hat{\beta}_p-\beta_0)' X_n' \Sigma(\hat{\rho}_n) X_n(\hat{\beta}_p-\beta_0)}{n} - \frac{2\varepsilon_n'(I-\rho M_n')^{-1}\Sigma(\hat{\rho}_n)X_n(\hat{\beta}_p-\beta_0)}{n} \\
\leqslant & \ 0
\end{aligned}$$

Suppose Assumption 1-5, and 8 holds, and combined with the fact that $\hat{m} = card(\hat{T} - T)$, then using the result of Lemma 3.2, we will get

$$\begin{aligned}
\frac{(\hat{\beta}_p-\beta_0)' X_n' \Sigma(\hat{\rho}_n) X_n(\hat{\beta}_p-\beta_0)}{n} & \leqslant \left|\frac{2\varepsilon_n'(I-\rho M_n')^{-1}\Sigma(\hat{\rho}_n)X_n(\hat{\beta}_p-\beta_0)}{n}\right| \\
& \leqslant 2e_n(\hat{m},\eta)||(\hat{\beta}_p-\beta_0)||_2,
\end{aligned}$$

with probability $1-\eta\exp^{-s}(1-1/e)$.

Since $\hat{m} \leqslant n-s$, from the extension of Condition (RSE), we obtain

$$\frac{(\hat{\beta}_p-\beta_0)' X_n' \Sigma(\hat{\rho}_n) X_n(\hat{\beta}_p-\beta_0)}{n} \geqslant \tau_x \mu_{min}(\Sigma(\rho))||\hat{\beta}_p-\beta_0||_2^2,$$

remind here $\Sigma(\rho) = (I-\rho M_n)'(I-\rho M_n)$. And combine the above results, with probability $1-\eta\exp^{-s}(1-1/e)$, the ℓ_2 norm of the post-model selection estimation error has an upper bound,

$$||\hat{\beta}_p-\beta_0||_2 \leqslant 2e_n(\hat{m},\eta)/\tau_x\mu_{min}(\Sigma(\rho)),$$

which proves the Theorem.

PROOF of Lemma 3.3. For any fixed $m \in \{0,1,\cdots,n-s\}$, easy to see $||T\varepsilon_n||_\infty =$

$\max_{i\in\tilde{T},card(\tilde{T}-T)\leqslant m}\frac{1}{n}|\sum_{j=1}^{n}T_{i,j}\varepsilon_j|$. Then

$$\begin{aligned}
& P(||T\varepsilon_n||_\infty > h_n(m,\eta)) \\
= \; & P(\max_{i\in\tilde{T},card(\tilde{T}-T)\leqslant m}\frac{1}{n}|\sum_{j=1}^{n}T_{i,j}\varepsilon_j| > h_n(m,\eta)) \\
\leqslant \; & (m+s)\max_i P(\frac{1}{\sqrt{n}}|\sum_{j=1}^{n}T_{i,j}\varepsilon_j| > \sqrt{h}_n(m,\eta))
\end{aligned}$$

Since the ε_j is i.i.d with $N(0,\sigma^2)$, the linear combination for each i, $\frac{1}{\sqrt{n}}\sum_{j=1}^{n}T_{i,j}\varepsilon_j$ is also normally distributed with mean 0 and variance $\frac{1}{n}\sum_{j=1}^{n}T_{i,j}^2$. And the variance is indeed a multiplication of the ith element of the diagonal of the matrix TT'. Consider the fact that $\frac{1}{n}TT'-\Psi\to_p 0$, and that all the diagonal element in Ψ is equal to 1. Thus for any $\eta\in(0,1)$

$$\begin{aligned}
& \max_i P(\frac{1}{\sqrt{n}}|\sum_{j=1}^{n}T_{i,j}\varepsilon_j| > \sqrt{h}_n(m,\eta)) \\
\leqslant \; & 2\exp\{-\frac{nh_n^2(m,\eta)}{2\sigma^2}\}
\end{aligned}$$

And if we bring in the definition of $h_n(m,\eta)$

$$\begin{aligned}
& P(\max_{i\in\tilde{T},card(\tilde{T}-T)\leqslant m}\frac{1}{n}|\sum_{j=1}^{n}T_{i,j}\varepsilon_j| > h_n(m,\eta), for\ \ any\ \ m\leqslant n-s) \\
\leqslant \; & 2\Sigma_{m=0}^{n-s}\eta p^{-m} \\
\leqslant \; & 2\eta p/(p-1).
\end{aligned}$$

PROOF of Theorem 3.4. Still use the definition of the post-model selection estimator $\hat{\beta}_p$,

$$\frac{||(I-\hat{\rho}_nM_n)(Y_n-X_n\hat{\beta}_p)||_2^2}{n}\leqslant\frac{||(I-\hat{\rho}_nM_n)(Y_n-X_n\beta_0)||_2^2}{n},$$

and therefore, easy to get

$$\frac{(\hat{\beta}_p - \beta_0)' X_n' \Sigma(\hat{\rho}_n) X_n (\hat{\beta}_p - \beta_0)}{n} - \frac{2\varepsilon_n' (I - \rho M_n')^{-1} \Sigma(\hat{\rho}_n) X_n (\hat{\beta}_p - \beta_0)}{n} \leqslant 0.$$

Since $||(\hat{\beta}_p - \beta_0)_{T^c}||_0 = \hat{m} \subset \{0, \cdots, n-s\}$, then combine with Condition (RSE), and Lemma 3.3,

$$\mu_{min}(\Sigma(\rho))\tau_x ||\hat{\beta}_p - \beta_0||_2^2 \leqslant h_n(\hat{m}, \eta)||\hat{\beta}_p - \beta_0||_1.$$

The cardinality of the support of $\hat{\beta}_p - \beta_0$ is bounded by $m + s$, so

$$||\hat{\beta}_p - \beta_0||_1^2 \leqslant (m+s)||\hat{\beta}_p - \beta_0||_2^2,$$

with high probability. Combine the two results, we can have

$$||\hat{\beta}_p - \beta_0||_1 \leqslant \frac{h_n(m, \eta)(m+s)}{\mu_{min}(\Sigma(\rho))\tau_x}.$$

On the other hand, since the subset $\hat{T}$ elementsof $\hat{\beta}_p$ is the least squares estimator for the linear model with response vector Y_n and covariate matrix $X_{\hat{T}}$, then

$$\frac{1}{n} X_{\hat{T}}' \Sigma(\hat{\rho}_n)(Y_n - X_{\hat{T}} \hat{\beta}_{ps}) = 0,$$

and easy to see the equation above is the same as

$$\frac{1}{n} X_n' \Sigma(\hat{\rho}_n)(Y_n - X_n \hat{\beta}_p) = 0.$$

Keep in mind the definition of Ψ, we have,

$$\begin{aligned} ||\Psi(\hat{\beta}_p - \beta_0)||_\infty &= ||\frac{1}{n}X_n'\Sigma(\hat{\rho}_n)(X_n\hat{\beta}_p - Y_n) - \frac{1}{n}X_n'\Sigma(\hat{\rho}_n)(X_n\beta_0 - Y_n)||_\infty \\ &\leqslant ||\frac{1}{n}X_n'\Sigma(\hat{\rho}_n)(X_n'\hat{\beta}_p - Y_n)||_\infty + ||\frac{1}{n}X_n'\Sigma(\hat{\rho}_n)(X_n\beta_0 - Y_n)||_\infty \\ &\leqslant ||\frac{1}{n}X_n'\Sigma(\hat{\rho}_n)(X_n\beta_0 - Y_n)||_\infty \\ &\leqslant h_n(\hat{m}, \eta), \end{aligned}$$

from the result of Lemma 3.3.

Thus, with high probability ,for $1 \leqslant j \leqslant p$,

$$(\Psi(\hat{\beta}_p - \beta_0))_j = (\hat{\beta}_{pj} - \beta_{0j}) + \sum_{i \neq j} \Psi_{i,j}(\hat{\beta}_{pi} - \beta_{0i}),$$

then

$$\begin{aligned} |(\Psi(\hat{\beta}_p - \beta_0))_j - (\hat{\beta}_{pj} - \beta_{0j})| &\leqslant \frac{c}{s}\sum_{i \neq j}|\hat{\beta}_{pi} - \beta_{0i}| \\ |\hat{\beta}_{pj} - \beta_{0j}| &\leqslant |(\Psi(\hat{\beta}_p - \beta_0))_j - (\hat{\beta}_{pj} - \beta_{0j})| + \frac{c}{s}\sum_{i \neq j}|\hat{\beta}_{pi} - \beta_{0i}|, \end{aligned}$$

and we will have

$$\begin{aligned} ||\hat{\beta}_p - \beta_0||_\infty &\leqslant ||\Psi(\hat{\beta}_p - \beta_0)||_\infty + \frac{c}{s}||\hat{\beta}_p - \beta_0||_1 \\ &\leqslant (1 + \frac{c(\hat{m} + s)}{s\mu_{min}(\Sigma(\rho))\tau_x})h_n(\hat{m}, \eta). \end{aligned}$$

And this proves the result of Theorem 3.4.

Chapter 4

Future work

4.1 An extension to Mixed Regressive, Spatial Autoregressive Models

Tobler's first law of geography encapsulates this situation: "everything is related to everything else, but near things are more related than distant things." One way of approach is through spatial interaction. According to Anselin and Bera (1998), high or low values for a random variable tend to cluster in space or locations tend to be surrounded by neighbors with very dissimilar values. The spatial interactions generally come from three resources: the endogenous interaction effects among the dependent variables, the exogenous interaction effects among the independent variables and the interaction effects among the error terms. To capture the spatial dependence, the general approach in a spatial econometrics is to impose structures on a model. In an empirical economic problem, if the spatial influence only comes from the error terms, econometricians will prefer to use a regression model with spatial autoregressive errors, that is, a spatial error model as we have discussed in the previous chapters. Compared with others, the spatial error model is conceptually simpler in the sense that the only problems involved are heteroskedasticity and non-linearity in the spatial parameter ρ.

Another popular type of model, which has also been heavily discussed in the literature,

considers the endogenous interaction effects on the dependent variable in a regression context, and this type of model is called Mixed Regressive, Spatial Autoregressive Model:

$$Y_n = \delta W_n Y_n + X_n \beta + \varepsilon_n, \tag{4.1}$$

where n is the total number of spatial cross-sectional units, Y_n is an n-dimensional vector of response, X_n is an $n \times p$ matrix of constant regressors, W_n is the spatial weights matrix, similarly defined to matrix M_n in the Spatial Error Model, and ε_n is an n-dimensional i.i.d. disturbances with zero mean and finite variance σ^2. The $W_n Y_n$ in (4.1) is called a "spatial lag" and the spatial autoregressive parameter ρ represents the spatial effect due to the influence of neighboring units. The main interest in estimation of the model is, in general, the parameters δ, β and σ^2. The interpretation of the model is that the response of a unit depends not only on the explanatory variables but also on the response of its neighboring units. Therefore, the spatial lag model is widely used in spatial econometrics, social sciences, agricultures, and health (Bertrand, Luttmer, and Mullainathan, 2000, Topa, 2001).

Clearly we would not want to run ordinary least squares (OLS) on this model, since the presence of Y_n on both sides of the equation means that there exists a correlation between regressors and disturbances, and the estimates will thus be biased and inconsistent. The estimating methods, which have been widely discussed in the literature for mixed regressive, spatial autoregressive models, are mainly the (quasi-) maximum likelihood estimator (Lee, 2004), the two-stage-least-square (2SLS) or Instrumental Variable method (Kelejian and Prucha, 1998, Lee, 2002) and the generalized method of moments (Lee, 2007). The instrumental variables (IV) are usually generated from exogenous regressors X_n and the spatial weights matrix W_n of the model, and most of them are computationally simple.

However, they are inefficient relative to the ML estimator, when the disturbances are normally distributed so that the likelihood function is correctly specified. Also, as the IVs are functions of the spatial weights matrices and exogenous variables, the 2SLS method would not be applicable to the (pure) Spatial Autoregressive Process when there are no exogenous variables relevant in the model. The generalized method of moments (GMM) approach combines the IV estimation with a generalization of the method of moments (MOM) in Kelejian and Prucha (1999) that has been discussed for the estimation of Spatial Error Model. Of all, the most popular and traditional estimation method is the (quasi-) maximum likelihood estimator under the assumption that the error term ε_n is normally distributed, and the quasi-maximum likelihood estimator allows for the case when the true distribution of error is different from normal.

We want to continue the idea of exploring variable selection and estimation methods in the high-dimensional setup, where the data contains larger number of parameters than the sample size but most of them are excessive, for spatial econometric models and extend the theoretical discussions to a mixed regressive, spatial autoregressive model, where the response variable is spatially correlated with units in neighbors in a regression context.

Consider the mixed regressive, spatial autoregressive model defined in (4.1) in a high-dimensional setting, and let $S_n(\delta) = I_n - \delta W_n$. We advocate an ℓ_1 penalized likelihood estimator which is defined as,

$$\hat{\beta} = \arg\max_{\beta \in \mathbb{R}^p} \{\hat{l}_n(\beta) - \lambda_n ||\beta||_1\}, \tag{4.2}$$

where

$$\hat{l}_n(\beta) = -\frac{n}{2}\ln(2\pi) - \frac{n}{2}\ln\sigma^2 + \ln|S_n(\delta)| - \frac{1}{2\sigma^2}(Y_nS_n(\delta) - X_n\beta)'(Y_nS_n(\delta) - X_n\beta).$$

is the log-likelihood of the model parameters given the sample. In reality these days, there are increasing number of datasets containing spatial interaction that comes from the response variable, and for many of them, large amounts of irrelevant parameters exist because of easy data collection. Because of this, it will be extremely helpful to achieve asymptotic consistency as well as theoretical inference results to justify for the penalized estimator defined in (4.2).

4.2 Future Work

So far, spatial literature has not paid much importance on model selection in a high dimensional context, and neither has the model selection literature accounted for spatial dependence in any substantial way. The combination of these two concepts can be widely extended to more complex spatial models. For example, besides the spatial error model and mixed regressive, spatial autoregressive model we discussed in the dissertation, when only one source of spatial interaction is considered, we can also develop similar approach for the spatial autoregressive model with autoregressive disturbances, where both interactions from the neighbors are considered. Also, the spatial models we mentioned only contain one spatial lag term (the ρM_n or δW_n part). Spatial models with higher-order, which incorporate two or more spatial lags, have also been discussed in the literature (Lee and Liu, 2010) and how to develop efficient statistical approaches for them in the high-dimensional set up is worth studying.

In the existing literature of spatial econometrics, the spatial weight matrix plays an irreplaceable role in describing the interactions between cross-sectional units. However, it has been pointed out by Manski (1993) that the literature of spatial autoregressive model family fail to specify how the spatial weight matrix should change when the sample size changes. Even though the increase of sample size will inevitably affect the magnitude of the spatial autocorrelation, and there exist large literature working on the asymptotic properties of estimators in the low-dimensional set up, study of how the spatial weight matrix will change is largely neglected. More work is needed for the specification of spatial weight matrix to allow for a changing sample size context.

Besides, the spatial autoregressive model family in the literature always treat the spatial weight matrix as priori knowledge and the spatial weights are typically defined as functions of some pre-defined measure of distance. The choice may or may not be consistent with reality and incorrect specification may result in consequences. Some literature has already noticed the issue (Bhattacharjee, Jensen-Butler, 2013, Bailey et al, 2014), but most of the methods are developed for panel data and do not take into consideration of the high-dimensionality of the predictor variables. We anticipate the analysis of spatial weight matrix for data collected from one time period in a high-dimensional set up could lead us to a new understanding of spatial interaction and the method can also be extended to a wider research field.

BIBLIOGRAPHY

BIBLIOGRAPHY

[1] Anselin, L. and Bera, A. K. (1998). Spatial Dependence in Linear Regression Models with an Introduction to Spatial Econometrics. *Handbook of Applied Economic Statistics*, 237–289.

[2] Bailey, N., Holly, S. and Pesaran, M. H. (2016). A Two-Stage Approach to Spatio-Temporal Analysis with Strong and Weak Cross-Sectional Dependence. *Journal of Applied Econometrics*, **31**, 249–280.

[3] Barry, R. P. and Pace, R. K. (1999). Monte Carlo Estimates of the Log Determinant of Large Sparse Matrices. *Linear Algebra and its Applications*, **289** , 41–54.

[4] Belloni, A., Chen, D., Chernozhukov, V. and Hansen, C. (2012). Sparse Models and Methods for Optimal Instruments With an Application to Eminent Domain. *Econometrica*, **80** , 2369–2429.

[5] Belloni, A. and Chernozhukov, V. (2011). High Dimensional Sparse Econometric Models: An Introduction. arXiv:1106.5242v2.

[6] Belloni, A. and Chernozhukov, V. (2013). Least Squares After Model Selection in High-dimensional Sparse Models. *Bernoulli*, **19** , 521–547.

[7] Belloni, A., Chernozhukov, V. and Wang, L. (2011). Square-root Lasso: Pivotal Recovery of Sparse Signals via Conic Programming. *Biometrika*, **98** , 791–806.

[8] Bertrand, M., Luttmer, E.F.P., and Mullainathan, S. (2000). Network Effects and Welfare Cultures. *Quarterly Journal of Economics*, **115** , 1019–1055.

[9] Bhattacharjee, A., Castro, E., Maiti, T. and Marques, J. (2016). Endogenous Spatial Regression and Delineation of Submarkets: A New Framework with Application to Housing Markets. *Journal of Applied Econometrics*, **31**, 32–57.

[10] Bhattacharjee, A. and Jensen-Butler, C. (2013). Estimation of the Spatial Weights Matrix under Structural Constraints. *Regional Science and Urban Economics*, **43**, 617–634.

[11] Bickel, P. J., Ritov, Y. and Tsybakov, A. B. (2009). Simultaneous Analysis of Lasso and Dantzig Selector. *The Annals of Statistics*, **37**, 1705–1732.

[12] B*ü*hlmann, P. and van de Geer, S. (2011). Statistics for High-Dimensional Data. *Springer*.

[13] Cliff, A. D. and Ord, J. K. (1973). Spatial Autocorrelation. *London: Pion*.

[14] Cliff, A. D. and Ord, J. K. (1981). Spatial Process: Models and Applications. *London: Pion*.

[15] Donoho, D. L., Elad, M. and Temlyakov, V. (2006). Stable Recovery of Sparse Overcomplete Representations in the Presence of Noise. textitIEEE Transactions On Information Theory, **52** , 6–18.

[16] Fan, J. and Lv, J. (2010). A Selective Overview of Variable Selection in High Dimensional Feature Space. *Statistica Sinica*, **20** , 101–148.

[17] Friedman, J. Hastie, T. and Tibshirani, R. (2010). Regularization Paths For Generalized Linear Models via Coordinate Descent. *Journal of Statistical Software*, **33**, 1–22.

[18] Fu, W. and Knight, K. (2000). Asymptotics for Lasso-type Estimators. *The Annals of Statistics*, **28** , 1356–1378.

[19] Geyer, C.J. (1996). On the asymptotics of convex stochastic optimization. *Unpublished manuscript.*

[20] Hoerl, A. E. and Kennard, R. W (1970). Ridge Regression: Biased Estimation for Nonorthogonal Problems. *Technometrics*, **12**, 69–82.

[21] Huang, J., and Horowitz, J. L. and Wei, F. (2010). Variable Selection in Nonparametric Additive Models. *The Annals of Statistics*, **38**, 2282–2313.

[22] Kelejian, H. H. and Prucha, I. R. (1998). A Generalized Moments Estimator for the Autoregressive Parameter in a Spatial Model. *International Economics Review*, **40** , 509–533.

[23] Kelejian, H. H. and Prucha, I. R. (1999). A Generalized Spatial Two-stage Least Squares Procedure for Estimating a Spatial Autoregressive Model with Autoregressive Disturbance. *Journal of Real Estate Finance and Economics*, **17** , 99–121.

[24] Lee, L-F. (2002). Consistency and Efficiency of Least Squares Estimation for Mixed Regressive, Spatial Autoregressive Models. *Econometric Theory* , **18**, 252–277.

[25] Lee, L-F. (2004). Asymptotic Distributions of Quasi-maximum Likelihood Estimators for Spatial Autoregressive Models. *Econometrica* , **72**, 1899–1925.

[26] Lee, L-F. (2007). GMM and 2SLS Estimation of Mixed Regressive, Spatial Autoregressive Models. *Journal of Econometrics* , **137**, 489–514.

[27] Lee, L-F. and Liu, X. (2010). Efficient GMM Estimation of High Order Spatial Autoregressive Models with Autoregressive Disturbances. *Econometric Theory* , **26**, 187–230.

[28] LeSage, J. and Pace, R. (2009). *Introduction to Spatial Econometrics*. Boca Raton, FL: CRC Press.

[29] Lounici, K. (2008). Sup-norm Convergence Rate and Sign Concentration Property of Lasso and Dantzig Estimators. *Electronic Journal of Statistics*, **2** , 90–102.

[30] Manski, C. F. (1993). Identification of Endogenous Social Effects: The Reflection Problem. *The Review of Economic Studies*, **60** , 531–542.

[31] Meinshausen, N. and Bühlmann, P. (2006). High Dimensional Graphs and Variable Selection with the Lasso. *The Annals of Statistics*, **34** , 1436–1462.

[32] Ord, J. K. (1975). Estimation Methods for Models of Spatial Interaction. *Journal of American Statistical Association*, **70** , 120–126.

[33] Pollard, D. (1991). Asymptotic for Least Absolute Deviation Regression Estimators. *Econometric Theory*, **7**, 186–199.

[34] Smirnov, O. and Anselin, L. (2001). Fast Maximum Likelihood Estimation of Very Large Spatial Autoregressive Models: A Characteristic Polynomial Approach. *Computational Statistics and Data Analysis*, **35** , 301–319.

[35] Tibshirani, R. (1996). Regression Shrinkage and Selection via the LASSO. *Journal of the Royal Statistical Society. Series B*, **58** , 267–288.

[36] Topa, G. (2001). Social Interactions, Local Spillovers and Unemployment. *The Review of Economic Studies*, **68** , 261–295.

[37] van de Geer, S. A. (2014). Statistical Theory for High-dimensional models. *Lecture Notes*.

[38] Varian, H. R. (2014). Big Data: New Tricks for Econometrics. *Journal of Economic Perspectives*, **28** , 3–28.

[39] Whittle, P. (1954). On stationary Processes in the Plane. *Biometrica* **41**, 434–449.

[40] Zhang, C. H. and Huang, J. (2008). The Sparsity and Bias of the Lasso Selection in High-dimensional Linear Regression. *The Annals of Statistics*, **36**, 1567–1594.

[41] Zhao, P. and Yu, B. (2006). On Model Selection Consistency of Lasso. *Journal of Machine Learning Research*, **7** , 2541–2563.