12-31-2021

Parameter estimation and inference of spatial autoregressive model by stochastic gradient descent

Gan Luan
New Jersey Institute of Technology, gl238@njit.edu

Copyright Warning & Restrictions

The copyright law of the United States (Title 17, United States Code) governs the making of photocopies or other reproductions of copyrighted material.

Under certain conditions specified in the law, libraries and archives are authorized to furnish a photocopy or other reproduction. One of these specified conditions is that the photocopy or reproduction is not to be "used for any purpose other than private study, scholarship, or research." If a, user makes a request for, or later uses, a photocopy or reproduction for purposes in excess of "fair use" that user may be liable for copyright infringement,

This institution reserves the right to refuse to accept a copying order if, in its judgment, fulfillment of the order would involve violation of copyright law.

Please Note: The author retains the copyright while the New Jersey Institute of Technology reserves the right to distribute this thesis or dissertation

Printing note: If you do not wish to print this page, then select "Pages from: first page # to: last page #" on the print dialog screen

ABSTRACT

PARAMETER ESTIMATION AND INFERENCE OF SPATIAL AUTOREGRESSIVE MODEL BY STOCHASTIC GRADIENT DESCENT

by
Gan Luan

Stochastic gradient descent (SGD) is a popular iterative method for model parameter estimation in large-scale data and online learning settings since it goes through the data in only one pass. While SGD has been well studied for independent data, its application to spatially-correlated data largely remains unexplored. This dissertation develops SGD-based parameter estimation and statistical inference algorithms for the spatial autoregressive (SAR) model, a common model for spatial lattice data.

This research contains three parts. **(I)** The first part concerns SGD estimation and inference for the SAR mean regression model. A new SGD algorithm based on maximum likelihood estimator (MLE) is proposed to accommodate the spatial correlation in the SAR model. Also, a statistical inference algorithm is proposed based on the online bootstrap resampling procedure (Fang et al., 2018). The asymptotic properties are then developed for the estimators and the finite sample properties for the estimators are investigated by simulations. The SGD-based parameter estimation procedures are shown to be more than 40 times faster than MLE for the settings examined. The SGD estimators for all parameters are close to the true values. The empirical coverages of confidence intervals (CIs) are at the nominal levels for the coefficients of the covariates but not for the spatial parameter. Two methods are proposed to improve the empirical coverage of CI for the spatial parameter. **(II)** The second part is regarding the SAR quantile regression mode. SGD algorithms based on one-stage quantile regression (1SQR) and two-stage quantile regression (2SQR) are developed for parameter estimation and statistical inference. Simulation results show that SGD estimator based on 2SQR is unbiased while that based on 1SQR

is biased. Also, the empirical coverages of CIs constructed using SGD based on 2SQR are all at the nominal levels. **(III)** In the last part, this research analyzes a real dataset on charges for medical services provided by physicians and healthcare professionals. Both SAR mean regression and quantile regression models are fitted to study the effect of location and other characteristics of medical facilities on medical prices. Modeling results show that the spatial correlation parameter is significantly different from 0 (95% CI is (-0.27, -0.23) for the mean regression), suggesting spatial correlation of medical charges. Also the models find that charges depend on the total number of services provided yearly, gender of the provider, facility type, and whether the provider is in a metropolitan area.

PARAMETER ESTIMATION AND INFERENCE OF SPATIAL AUTOREGRESSIVE MODEL BY STOCHASTIC GRADIENT DESCENT

by
Gan Luan

A Dissertation
Submitted to the Faculty of
New Jersey Institute of Technology and
Rutgers, The State University of New Jersey – Newark
in Partial Fulfillment of the Requirements for the Degree of
Doctor of Philosophy in Mathematical Sciences

Department of Mathematical Sciences
Department of Mathematics and Computer Science, Rutgers-Newark

December 2021

APPROVAL PAGE

PARAMETER ESTIMATION AND INFERENCE OF SPATIAL AUTOREGRESSIVE MODEL BY STOCHASTIC GRADIENT DESCENT

Gan Luan

Dr. Ji Meng Loh, Dissertation Advisor — Date
Associate Professor of Mathematical Sciences, NJIT

Dr. Sunil Dhar , Committee Member — Date
Professor of Mathematical Sciences, NJIT

Dr. Wenge Guo , Committee Member — Date
Associate Professor of Mathematical Sciences, NJIT

Dr. Sundarraman Subramanian, Committee Member — Date
Associate Professor of Mathematical Sciences, NJIT

Dr. Yixin Fang, Committee Member — Date
Director of Global Medical Affairs Statistics, AbbVie, Chicago, Illinois

BIOGRAPHICAL SKETCH

Author: Gan Luan

Degree: Doctor of Philosophy

Date: December 2021

Undergraduate and Graduate Education:

- Doctor of Philosophy in Mathematical Sciences,
  New Jersey Institute of Technology, Newark, New Jersey, 2021
- Doctor of Philosophy in Biomedical Sciences,
  Rutgers University, Newark, New Jersey, 2018
- Bachelor of Science in Biotechnology,
  University of Science of Technology of China, Hefei, Anhui, China, 2012

Major: Mathematical Sciences

Presentations:

Gan Luan, Ji Meng Loh "Parameter Estimation and Inference of Spatial Autoregressive Model by Stochastic Gradient Descent," *International Chinese Statistical Association Applied Statistics Symposium*, Houston, TX, USA, December 14, 2020 (held virtually due to COVID19).

Gan Luan, Ji Meng Loh "Application of Stochastic Gradient Descent in Parameter Estimation for Models with Spatial Correlation," *Joint Statistical Meetings*, Philadelphia, PA, USA, August 4, 2020 (held virtually due to COVID19).

To my beloved Shenshen, Mom, Dad, and Grandma

ACKNOWLEDGMENT

There are many people I would like to thank for their help in completing the present dissertation. I would like start by thanking my advisor Prof. Ji Meng Loh. I first met Prof. Loh when taking the course of Probability Distribution taught by him. I was highly impressed by how good he is at teaching. He always can explain difficult concepts in a simple way. I benefit a lot from this when I was first introduced to the area of spatial statistics. He provided all the guidance and support I need for doing the research. I am so grateful that Prof. Loh never blamed me for any mistakes I made and always helped me to learn from them and try to avoid them next time. He is the kind of mentor who always puts the student's need on the first place. I am interested in working in industry and did two summer internships during my PhD study. This cannot be possible without Prof. Loh' support.

I want to say thank you to my committee, Prof. Sunil Dhar, Prof. Wenge Guo, Prof. Sundarraman Subramanian, and Dr. Yixin Fang. They provided a lot of wonderful suggestions for my research and encouraged me a lot. They are always there whenever I have problems with my research. Their optimism and persistence on research motivate me all the time. Besides research, they also helped me a lot with my career development and my daily life.

I would like to thank National Institute of Health for sponsoring my research with grant NIH R15AG061651-01. And I want to thank the Math Department for providing the opportunity of teaching assistance.

I would like to thank professors in the Math Department and the Computer Science Department, Prof. Antai Wang, Prof. Cyrill Muratov, Prof. Guiling Wang, Prof. Pantelis Monogioudis and many others. Courses taught by them prepared me well for my research. Also I want to thank the administrative staff in the Math Department, Michelle Llado-Wrzos, Rey Sentina, Alison Boldero, Prof. Shahriar

Afkhami, and Prof. Michael Siegel. They are always there to help for all kinds of questions I have during my graduate studies.

I want to express my gratitude to my friends. We had a lot of fun hanging out together and they made my life as a PhD student much easier. I would like to specially mention Lei Chen, Feng Liu, Beibei Li, and Atefeh Javidialsaadi here. Lei and Feng encouraged me a lot to purse this PhD degree in Mathematics after I finished my PhD degree in Microbiology. Beibei, Atefeh, and I spent a lot time taking course, preparing for exams, doing projects, searching for jobs together. Their help and encouragement motivate me a lot when I am facing obstacles.

Last but not the least, I would like to say thank you to my family. I feel so blessed to meet my fiancee Shenshen during this hard time of pandemic. She brings sunshine to my life and I cannot imagine how I can survive this pandemic and PhD study without her love and support. Also for my family members back in China, my grandma, mom, and dad, it would be impossible for me to study in the US without their sacrifice and love. And of course I want to mention our lovely dog Amur. Walking her and playing with her is always a great leisure time for me.

TABLE OF CONTENTS

TABLE OF CONTENTS
(Continued)

Chapter **Page**

TABLE OF CONTENTS
(Continued)

Chapter **Page**

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1

OVERVIEW

1.1 Spatial Data and Stochastic Gradient Descent

This section briefly introduces two main topics of this dissertation: spatial data and stochastic gradient descent (SGD).

1.1.1 Spatial data

Many studies make observations of one or more variables at multiple sites. If the location information of these sites are also observed and attached to the data, then the resulting data is called spatial data. The specific type of spatial data studied in this dissertation is the lattice data. For lattice data, the domain under study is fixed and discrete and the attributes of interest can be observed at a number of fixed locations. These locations can be points or regions, but usually census track, states, zip codes etc. [9, 39].

Lattice data has been analyzed and studied in many area, such as economic, environmental and geographical research [9]. For example, Haider and Miller studied the effects of transportation infrastructure and location on the residential real estate values [17]. Permai et al. modelled the average expenditure of Papua providence in Indonesia and considered spatial correlations [44]. Trzpiot and Orwat-Acedańska investigated the healthy life years in European countries with a spatial quantile regression model [53]. Kanaroglou et al. estimated sulfur dioxide air pollution concentrations with a spatial autoregressive model [19].

One concrete example of lattice data is shown in Figure 1.1. It is a map with number of confirmed COVID-19 cases in each state of the United States as of October, 27, 2021 (data source: `https://coronavirus.1point3acres.com/`). Clustering appears in this map and this suggests the existence of spatial correlation. Also, it is

reasonable to assume that the number of confirmed cases in New Jersey is correlated with that of New York since they are neighbors and communication between these two states are in a high degree. For more general data, one can use Moran's I to study the strength of spatial dependence [40, 35]. Ignoring spatial correlation when it is present can significantly affect the modeling of the data.

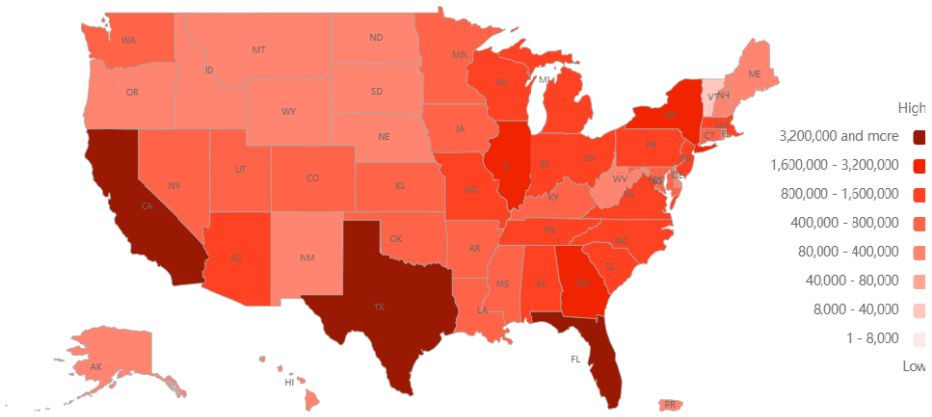


Figure 1.1 Map of confirmed COVID19 cases in the United States as of October, 27, 2021.

1.1.2 Stochastic gradient descent

An important parameter estimation method in statistics is estimating by minimizing a target function (loss function). Optimization is a common problem in machine learning [14]. Stochastic gradient descent (SGD) is an recursive algorithm for optimization and parameter estimation. It was first proposed by Robin & Monro [47] and then studied by many others (for example, [56, 48, 45, 42]). Unlike many other optimization algorithms that require the availability of all the data, SGD only uses one data point at each iteration. Let $F(\theta) = \mathbb{E}[f(\theta, z)]$ be the function to be minimized; where, z is an observation of the random variable Z, f the loss function, θ the unknown parameter, and the expectation is with respect to (w.r.t.) the random variable Z. Let θ_0 be the minimizer of $F(\theta)$ over parameter space Θ; that is $\theta_0 = \text{argmin}_{\theta\in\Theta} F(\theta)$. Let

$\{z_1, z_2, \ldots, z_n\}$ be independent observations of random variable Z, and $\hat{\theta}_0$ an initial value . If f is differentiable, then the estimate for θ can be updated as:

$$\hat{\theta}_k = \hat{\theta}_{k-1} - \gamma_k \nabla f(\hat{\theta}_{k-1}, z_k), \;\; k = 1, \ldots, n, \tag{1.1}$$

where, γ_k is often called the learning rate. One common way to set γ_k is $\gamma_k = \gamma_1 k^{-\alpha}$, with $\gamma_1 > 0$ and $\alpha \in (0.5, 1)$. Also, Ruppert and Polyak & Juditsky suggested that the convergence of SGD estimates can be accelerated by taking the mean of the estimates ([48, 45]):

$$\bar{\theta}_k = \frac{1}{k} \sum_{i=1}^{k} \hat{\theta}_i. \tag{1.2}$$

The averaged estimate also can be recursively updated by $\bar{\theta}_k = ((k-1)\bar{\theta}_{k-1} + \hat{\theta}_k)/k$.

We can use SGD for parameter estimation of the linear regression model:

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}, \;\; \boldsymbol{\epsilon} \sim N(\boldsymbol{0}, \sigma^2 \boldsymbol{I}) \tag{1.3}$$

where, $\boldsymbol{y}_{n\times 1}$ is the response variable; $\boldsymbol{X}_{n\times p}$ are covariates, $\boldsymbol{\beta}$ and σ^2 unknown parameters. The log-likelihood for $\boldsymbol{y}$ (omitting constant term and constant coefficient) is:

$$\ell(\boldsymbol{\beta}|\boldsymbol{y}) = \sum_{i=1}^{n} -(y_i - \boldsymbol{\beta}^T \boldsymbol{x}_i)^2 = \sum_{i=1}^{n} f_i, \tag{1.4}$$

where, $f_i = -(y_i - \boldsymbol{\beta}^T \boldsymbol{x}_i)^2$ is the log-likelihood for i-th data point. The derivative of f_i w.r.t. $\boldsymbol{\beta}$ is:

$$\nabla f_{\boldsymbol{\beta},i} = 2\boldsymbol{x}_i(y_i - \boldsymbol{\beta}^T \boldsymbol{x}_i). \tag{1.5}$$

Let $F = \mathbb{E}(f_i)$ and let

$$\boldsymbol{\beta}_0 = \operatorname{argmin}(-F) = \operatorname{argmin} \mathbb{E}(y_i - \boldsymbol{\beta}^T \boldsymbol{x}_i)^2 \tag{1.6}$$

Given an initial estimate $\hat{\boldsymbol{\beta}}_0$, the SGD algorithm for updating $\hat{\boldsymbol{\beta}}_k$ is:

$$\hat{\boldsymbol{\beta}}_k = \hat{\boldsymbol{\beta}}_{k-1} + \gamma_k(y_k - \hat{\boldsymbol{\beta}}_{k-1}^T \boldsymbol{x}_k)\boldsymbol{x}_k. \tag{1.7}$$

Note that '+' is used in (1.7) rather than '−' like in (1.1) because we are trying to maximize the log-likelihood. Also, the constant '2' in (1.5) is not carried over to the updating algorithm (1.7), since we can adjust the initial value for the learning rate, γ_1. Here, in linear regression, we can easily write out the log-likelihood for i-th data point since data are assumed to be independent. Also, $\ell(\boldsymbol{\beta}|\boldsymbol{y})$, the log-likelihood for $\boldsymbol{y}$, can be treated as the sample estimation of nF. The maximum likelihood estimator (MLE) is the value that maximizes $\ell(\boldsymbol{\beta}|\boldsymbol{y})$, while SGD converges to the value that maximize nF. This is the connection between the estimators obtained by SGD and MLE.

SGD has several advantages over other algorithms that require the availability of the whole dataset. First, it does not require storage of all the data, since the data are only used once and are not revisited. This can reduce storage need. Also, SGD is very useful in stream learning (or online learning), where we observe the data one by one [55]. With SGD, we can estimate the parameters based on available data and update the estimate whenever new data arrive. When updating the estimate, we only need the current data point and do not need to revisit the previous data. Finally, SGD is fast and easy to scale up. Many optimization algorithms involve matrix calculations, which become very computational intensive for large sample size. Thus, these algorithms are difficult to scale up. On contrast, SGD can easily scale up since it only uses one data point for each recursive step.

The asymptotic properties of SGD estimates have been well studied. However, there are only a few studies look at the inference of SGD estimates. Chen et al. (2016) suggested two methods for estimating variance and constructing confidence intervals for SGD estimators: the plug-in method and the batch-mean method [5]. The plug-in

method requires the computation of a Hessian matrix and its inverse, which can be computational intensive. The batch-mean method has a relatively slower convergence rate compared to the plug-in method. Also, it tends to underestimate the variance due to the correlation between batch means. Li et al. (2017) proposed a statistical inference method similar to the batch-mean method and tried to reduce correlation between batch means by discarding some intermediate estimates [37]. However, their method only works for M-estimation based on SGD with a fixed learning rate. Su & Zhu (2018) proposed the statistical inference procedure HiGrad, short for hierarchical incremental gradient descent [51]. They used a hierarchical tree structure and updated SGD estimates along the tree. Their method can provide confidence intervals for predictions but not for vanilla SGD estimators.

Fang et al. (2018) proposed an online bootstrap resampling procedure to estimate the variance and construct confidence intervals for SGD estimators [13]. This method is simple and can be applied to a general class of models. For this method, to construct confidence interval (CI), besides the SGD estimates, a number of perturbed estimates are also obtained. Let $\hat{\theta}^*$ denote the perturbed estimate, and with $\hat{\theta}_0^* \equiv \hat{\theta}_0$, the perturbed SGD estimate can be updated as:

$$\hat{\theta}_k^* = \hat{\theta}_{k-1}^* - \gamma_k W_k \nabla f(\hat{\theta}_{k-1}^*, z_k),\ k = 1, \ldots, n \tag{1.8a}$$

$$\bar{\theta}_k^* = \frac{1}{k} \sum_{i=1}^{k} \hat{\theta}_i^*, \tag{1.8b}$$

where, W_k is the perturbation variable, and $W_k \overset{iid}{\sim} W,\ k = 1, \ldots, n,\ W > 0,\ E(W) = Var(W) = 1$. Fang et.al (2018) show that $\{\bar{\theta}_n^* - \bar{\theta}_n\}$ has the same asymptotic distribution as $\{\bar{\theta}_n - \theta_0\}$ [13]. In practice, we obtain $\bar{\theta}_k^{*,b}$ by sequentially updating perturbed SGD estimates for each sample, $b = 1, \ldots, B$.

$$\hat{\theta}_k^{*,b} = \hat{\theta}_{k-1}^{*,b} - \gamma_k W_{k,b} \nabla f(\hat{\theta}_{k-1}^{*,b}, z_k),\ k = 1, \ldots, n \tag{1.9a}$$

$$\bar{\theta}_k^{*,b} = \frac{1}{k}\sum_{i=1}^{k}\hat{\theta}_i^{*,b}, \tag{1.9b}$$

where, $W_{k,b} \overset{iid}{\sim} W,\ k = 1, \ldots, n$. We can approximate the sample distribution of $\{\bar{\theta}_n - \theta_0\}$ using the empirical distribution of $\{\bar{\theta}_n^{*,b} - \bar{\theta}_n\}, b = 1, \ldots, B$. They proposed two procedures for constructing $(1-\alpha)$ confidence intervals of θ_0. One is based on the upper and lower $\alpha/2$ quantiles of $\{\bar{\theta}_n^* - \bar{\theta}_n\}$. Assume we want to construct confidence interval for j-th element of θ_0. Let L and U be the empirical $\alpha/2$ and $1-\alpha/2$ quantile of $\bar{\theta}_n^{(j)*,1} - \bar{\theta}_n^{(j)}, \bar{\theta}_n^{(j)*,2} - \bar{\theta}_n^{(j)}, \ldots, \bar{\theta}_n^{(j)*,B} - \bar{\theta}_n^{(j)}$, then

$$\begin{aligned} P(L \leq \bar{\theta}_n^{(j)} - \theta_0^{(j)} \leq U) &= 1-\alpha \\ \text{i.e., } P(\bar{\theta}_n^{(j)} - U \leq \theta_0^{(j)} \leq \bar{\theta}_n^{(j)} - L) &= 1-\alpha \end{aligned} \tag{1.10}$$

Thus, the $(1-\alpha)$ CI for $\theta_0^{(j)}$ is $[\bar{\theta}_n^{(j)} - U, \bar{\theta}_n^{(j)} - L]$. The other confidence interval construction method is based on the sample variance of $\{\bar{\theta}_n^* - \bar{\theta}_n\}$. Let $S^{(j)}$ be the sample variance of $\bar{\theta}_n^{(j)*,1} - \bar{\theta}_n^{(j)}, \bar{\theta}_n^{(j)*,2} - \bar{\theta}_n^{(j)}, \ldots, \bar{\theta}_n^{(j)*,B} - \bar{\theta}_n^{(j)}$, then the confidence interval for $\theta_0^{(j)}$ can be constructed as $[\bar{\theta}_n - Z_{\alpha/2}\sqrt{S^{(j)}}, \bar{\theta}_n + Z_{\alpha/2}\sqrt{S^{(j)}}]$; here, $Z_{\alpha/2}$ is the $\alpha/2$ percentile of the standard normal distribution. An advantage of this online bootstrap resamping method for confidence interval construction is that it is recursive and only uses one data point at each step. Data that have been used are not revisited again. This algorithm retains all the nice properties of the SGD algorithm.

1.2 Dissertation Outline

Large datasets with spatial correlation are common nowadays in the big data era [36, 57]. As discussed above, SGD is a scaleable way for parameter estimation when analyzing large datasets. We consider SGD for parameter estimation of spatial models and incorporate a perturbation method for inference. The structure of this dissertation is listed below.

Chapter 2 discusses applying SGD for mean regression in spatial autoregressive (SAR) models. We first introduce the SAR mean regression model and common methods for parameter estimation. The maximum likelihood estimator (MLE) is an unbiased and efficient estimator but suffers from heavy computation burden. We circumvent this by using SGD for parameter estimation based on the likelihood. Different from linear mean regression model, data in SAR model are correlated. We modify the SGD procedures and the corresponding online bootstrapping procedure for confidence interval construction to accommodate the data correlation. We derive the asymptotic properties of the SGD estimator and study the finite sample properties using simulations. Simulations show that the estimates are close to true value. CIs for the regressor coefficients achieve nominal level. However, the empirical coverage of CIs does not reach the desired level for spatial parameter. We propose new ways to improve the coverage. In addition to the MLE based SGD procedure, we study the two-stage least square based SGD procedure. Simulations show that the estimates are close to true value and CIs all achieve nominal level. Besides the SAR model, we also consider the SGD estimation and CI construction for the spatial regression model with autoregressive disturbance. Finally, an R package is created to automate the parameter estimation and CIs construction procedure.

In Chapter 3 we consider SGD for quantile regression in the SAR model for modelling quantiles of response variables. Quantile regression can provide a more detailed analysis of the distribution and is more robust to outliers and less restrictive on error distributions. This chapter first introduces the SAR quantile regression model and discusses the available parameter estimation methods. Then it investigates the possibilities and advantages of applying the SGD procedure based on each of these estimation methods. After that it applies SGD based on the two selected estimation methods, ie. one-stage quantile regression and two-stage quantile regression methods.

Lastly finite sample properties of these two SGD estimation methods are investigated and compared with not applying SGD using simulations.

In Chapter 4, we analyze the Physician and Other Supplier Public Use File (PUF) data set. This dataset contains information on services and procedures provided to Medicare beneficiaries by physicians and other healthcare professionals at medical facilities. Effect of locations as well as other characteristics of the medical facilities on medical service charges are of interest. We use both mean regression and quantile regression SAR models to study the effect.

Chapter 5 summarizes this dissertation and provides directions for future study.

CHAPTER 2

ESTIMATION AND INFERENCE FOR THE SAR MEAN REGRESSION MODEL USING SGD

2.1 Introduction

2.1.1 Spatial autoregressive model

The spatial autoregressive (SAR) and conditional autoregressive (CAR) models are often used to model spatial lattice data [11]. SAR model assumes that the response variable at a given location depends on the response variable at neighboring locations, whereas the CAR model models the conditional distribution of the response given the neighboring values. In this dissertation, we focus on the SAR model. The was first introduced by Cliff and Ord [7] and then studied by many researchers e.g., [8, 21, 32, 46, 38]. The SAR model considers effect of covariates and the spatial correlation on the response variable. The general form for SAR model is:

$$\boldsymbol{y} = \rho \boldsymbol{W} \boldsymbol{y} + \boldsymbol{X} \boldsymbol{\beta} + \boldsymbol{\epsilon}, \tag{2.1}$$

where, $\boldsymbol{y}$ is the $n \times 1$ response variable vector, n the total number of data points, $\boldsymbol{X}$ the $n \times p$ covariate matrix, p the dimension of covariate, $\boldsymbol{\beta}$ the parameter for effects of covariate, $\boldsymbol{\epsilon}$ the random error term with $\mathbb{E}(\boldsymbol{\epsilon}) = \boldsymbol{0}$, ρ the autoregressive parameter, and $\boldsymbol{W}$ the $n \times n$ neighborhood matrix. Usually ρ is restricted to be between -1 and 1. The term $\rho \boldsymbol{W} Y$ is the spatial lag term and controls the effects of neighborhood units. In this SAR model, we assume ϵ_i and ϵ_j are independent for $i \neq j$. We further assume $\boldsymbol{\epsilon} \sim N(\boldsymbol{0}, \sigma^2 \boldsymbol{I})$, where, $\boldsymbol{I}$ is the $n \times n$ identity matrix. The unknown parameters are $\boldsymbol{\beta}, \rho$ and σ^2, of which $\boldsymbol{\beta}$ and ρ are usually of more interest in data analysis.

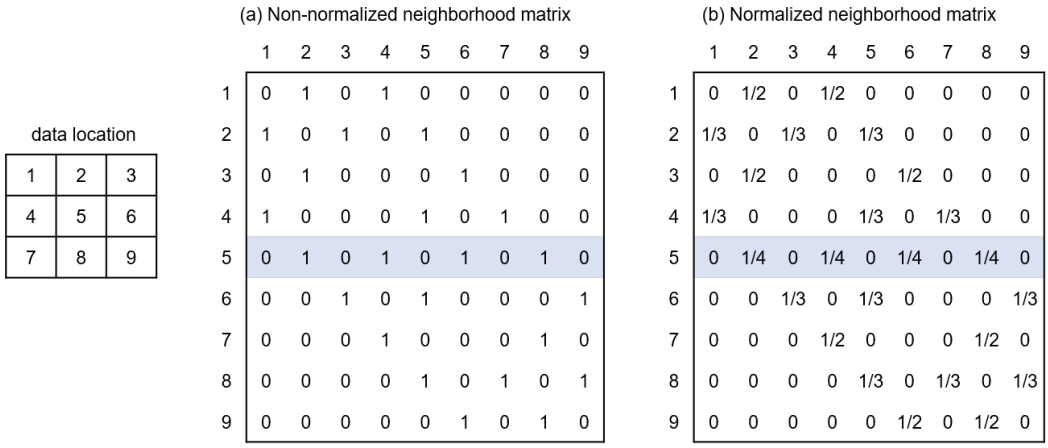
The neighborhood matrix $\boldsymbol{W}$ specifies the neighbors of each data point and the range of the correlation. The element $w_{i,j}$ is non-zero if and only if data points i and j are neighbors and the value of $w_{i,j}$ represents how strong the correlation is between these two data points. There are several common ways to specify $\boldsymbol{W}$. It can be the determined by whether two data points share borders, or by the whether the distance of two data points are within a certain threshold, or as the reciprocal of the distance of two data points. Usually the diagonal elements of $\boldsymbol{W}$ are zero, since the term $\rho\boldsymbol{WY}$ represents the effect of other spatial data points. Also, $\boldsymbol{W}$ is often row normalized, i.e., the sum of each row of $\boldsymbol{W}$ is 1. After row normalization, the elements of $\boldsymbol{W}$ are between 0 and 1 and $\boldsymbol{WY}$ can be interpreted as the weighted average of neighboring values. Another reason for row normalization concerns the invertibility of the matrix $\boldsymbol{A} = \boldsymbol{I} - \rho\boldsymbol{W}$. If $\boldsymbol{A}$ is invertible, we can rewrite the SAR model in (2.1) as:

$$\boldsymbol{y} = \boldsymbol{A}^{-1}(\boldsymbol{X\beta} + \boldsymbol{\epsilon}). \tag{2.2}$$

Let λ_i be the eigenvalues of $\boldsymbol{W}$, and $|\lambda_i| \leq 1$ if $\boldsymbol{W}$ is row-normalized. The determinant of $\boldsymbol{A}$, $|\boldsymbol{A}| = \prod_i(1 - \rho\lambda_i)$, is greater than 0 if ρ is between -1 and 1. Thus, a row-normalized $\boldsymbol{W}$ can guarantee that $\boldsymbol{A}$ is invertible for ρ between -1 and 1. Figure 2.1 shows a simple example of a neighborhood matrix. The left panel of this figure shows a dataset of 9 data points arranged on a 3 by 3 grid. Two data points are neighbors if and only if they share a border. The middle panel shows an adjacency matrix $\boldsymbol{C}$ before row normalization. Here, $C_{i,j} = 1$ if and only if data points i and j are neighbors. The right panel shows the neighborhood matrix after row normalization. Based on this neighborhood matrix, we can write the model for y_5 as:

$$y_5 = \rho\frac{y_2 + y_4 + y_6 + y_8}{4} + \boldsymbol{\beta}^T\boldsymbol{x}_5 + \epsilon_5. \tag{2.3}$$

From Equation (2.3), we can see that the value of y_5 is both affected by the value of its neighbors and the value of its covariates.



	1	2	3	4	5	6	7	8	9
1	0	1	0	1	0	0	0	0	0
2	1	0	1	0	1	0	0	0	0
3	0	1	0	0	0	1	0	0	0
4	1	0	0	0	1	0	1	0	0
5	0	1	0	1	0	1	0	1	0
6	0	0	1	0	1	0	0	0	1
7	0	0	0	1	0	0	0	1	0
8	0	0	0	0	1	0	1	0	1
9	0	0	0	0	0	1	0	1	0

	1	2	3	4	5	6	7	8	9
1	0	1/2	0	1/2	0	0	0	0	0
2	1/3	0	1/3	0	1/3	0	0	0	0
3	0	1/2	0	0	0	1/2	0	0	0
4	1/3	0	0	0	1/3	0	1/3	0	0
5	0	1/4	0	1/4	0	1/4	0	1/4	0
6	0	0	1/3	0	1/3	0	0	0	1/3
7	0	0	0	1/2	0	0	0	1/2	0
8	0	0	0	0	1/3	0	1/3	0	1/3
9	0	0	0	0	0	1/2	0	1/2	0

Figure 2.1 Simple example of neighborhood matrix.

2.1.2 Estimation methods for SAR model

Various parametric estimators for the SAR model have been studied. The ordinary least square (OLS) estimator treats the SAR model the same as the regular linear regression model. It defines $\tilde{\boldsymbol{X}} = [\boldsymbol{W}\boldsymbol{y} \;\; \boldsymbol{X}]$ and unknown parameter $\boldsymbol{\gamma} = [\rho, \boldsymbol{\beta}^T]^T$ and the OLS estimators is given by:

$$\hat{\gamma}_{OLS} = (\tilde{\boldsymbol{X}}^T \tilde{\boldsymbol{X}})^{-1} \tilde{\boldsymbol{X}}^T \boldsymbol{y} \tag{2.4}$$

This OLS estimators is easy to calculate but is shown to be inconsistent in general and consistent only for some special settings [31]. Kyriacou et al. proposed an indirect inference method to correct the bias in the OLS estimators [29] for the pure SAR model (the model without the $\boldsymbol{X}\boldsymbol{\beta}$ term in Equation (2.1)).

The two-stage least square (2SLS) estimator is another method for parameter estimation of the SAR model [30, 22]. Let $\boldsymbol{Z}$ denote the instrumental variables (IV) that are exogenous for $\boldsymbol{y}$. Common choices for Z are $\boldsymbol{W}\boldsymbol{X}, \boldsymbol{W}^2\boldsymbol{X}, \cdots$. For the

first stage, $\boldsymbol{Wy}$ is regressed on $\boldsymbol{X}$ and $\boldsymbol{Z}$ using OLS. Then in the second stage, $\boldsymbol{y}$ is regressed on $\widehat{\boldsymbol{Wy}}$ and $\boldsymbol{X}$ using OLS. $\widehat{\boldsymbol{Wy}}$ is the predicted value for $\boldsymbol{Wy}$ in the first stage. The estimated coefficient for $\widehat{\boldsymbol{Wy}}$ and $\boldsymbol{X}$ are the estimates for ρ and $\boldsymbol{\beta}$ respectively. The 2SLS estimator is consistent in general, but it is less efficient than maximum likelihood (MLE). Also, it can not be used when all the coefficients of covariate $\boldsymbol{X}$ are not significant [32].

The generalized method of moments (GMM) is also used for fitting the SAR model. Let $\boldsymbol{Q}$ denote the instrumental variable matrix, $\boldsymbol{\gamma} = [\rho, \boldsymbol{\beta}^T]^T$, and $\boldsymbol{\epsilon} = (\boldsymbol{I} - \rho\boldsymbol{W})\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}$. The moment function is defined as $g(\boldsymbol{\gamma}) = \boldsymbol{Q}\boldsymbol{\epsilon}$. Note that $\mathbb{E}[g(\boldsymbol{\gamma}_0)] = \boldsymbol{0}$ for the true parameter values $\boldsymbol{\gamma}_0$. Parameters are estimated by solving $g(\boldsymbol{\gamma}) = \boldsymbol{0}$. The GMM estimator relies on the choice of instrumental variables and is less efficient than MLE in general [34]. MLE is consistent but is computational intensively [43, 33, 20]. We describe the MLE in more detail below.

2.1.3 MLE for SAR model

To derive MLE for the SAR model, we rewrite the SAR model as:

$$\boldsymbol{\epsilon} = \boldsymbol{A}\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}. \tag{2.5}$$

Then, given $\boldsymbol{\epsilon} \sim N(\boldsymbol{0}, \sigma^2\boldsymbol{I})$, the likelihood of the SAR is:

$$\begin{aligned} L(\boldsymbol{\theta}|\boldsymbol{y}) &= L(\boldsymbol{\theta}|\boldsymbol{\epsilon})|\frac{d\boldsymbol{\epsilon}}{d\boldsymbol{y}}| = (2\pi\sigma^2)^{-n/2}\exp(-\frac{\boldsymbol{\epsilon}^T\boldsymbol{\epsilon}}{2\sigma^2})|\boldsymbol{A}| \\ &= (2\pi\sigma^2)^{-n/2}\exp(-\frac{(\boldsymbol{A}\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})^T(\boldsymbol{A}\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})}{2\sigma^2})|\boldsymbol{A}|. \end{aligned} \tag{2.6}$$

Here, $\boldsymbol{\theta} = [\boldsymbol{\beta}^T, \sigma^2, \rho]^T$, and $|\boldsymbol{A}|$ represents the determinant of matrix $\boldsymbol{A}$. The log-likelihood of SAR model (omitting constants) is:

$$\ell(\boldsymbol{\theta}|\boldsymbol{y}) = -\frac{\ln(\sigma^2)}{2}n - \frac{(\boldsymbol{A}\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})^T(\boldsymbol{A}\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})}{2\sigma^2} + \ln|\boldsymbol{A}|. \tag{2.7}$$

We can obtain the maximum likelihood estimator (MLE) of $\boldsymbol{\beta}$ and σ^2 by setting $\frac{\partial \ell(\boldsymbol{\theta}|\boldsymbol{y})}{\partial \boldsymbol{\beta}}$ and $\frac{\partial \ell(\boldsymbol{\theta}|\boldsymbol{y})}{\partial \boldsymbol{\sigma^2}}$ equal to zero, respectively. The MLE of these two parameters as a function of ρ are then

$$\hat{\boldsymbol{\beta}} = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{A}\boldsymbol{y}, \quad \hat{\sigma^2} = (\boldsymbol{A}\boldsymbol{y})^T(\boldsymbol{I}-\boldsymbol{M})^T(\boldsymbol{I}-\boldsymbol{M})(\boldsymbol{A}\boldsymbol{y})/n, \tag{2.8}$$

where, $\boldsymbol{M} = \boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T$. We can get the profile log-likelihood for ρ by plugging these estimates into (2.7):

$$\ell(\rho|\boldsymbol{y}) = \ln|\boldsymbol{A}| - \frac{n}{2}\ln(\boldsymbol{A}\boldsymbol{y})^T(\boldsymbol{I}-\boldsymbol{M})^T(\boldsymbol{I}-\boldsymbol{M})(\boldsymbol{A}\boldsymbol{y}). \tag{2.9}$$

Thus, the MLE of ρ is the ρ value that maximizes $\ell(\rho|\boldsymbol{y})$ (same as minimizing $-\ell(\rho|\boldsymbol{y})$) subject to $\rho \in (-1, 1)$:

$$\hat{\rho} = \underset{|\rho|<1}{\operatorname{argmin}}[-\ell_\rho(\boldsymbol{y})] = \underset{|\rho|<1}{\operatorname{argmin}}[-\ln|\boldsymbol{A}| + \frac{n}{2}\ln(\boldsymbol{A}\boldsymbol{y})^T(\boldsymbol{I}-\boldsymbol{M})^T(\boldsymbol{I}-\boldsymbol{M})(\boldsymbol{A}\boldsymbol{y})] \tag{2.10}$$

The MLE of ρ can be plugged into (2.8) to get the final estimate for $\boldsymbol{\beta}$ and σ^2.

The existence and uniqueness of solutions to (2.10) under some regularity conditions has been established [18]. However, the solution cannot be given in closed form for most cases due to the $\ln|\boldsymbol{A}|$ term in (2.10) [18]. Thus, numerical methods have to be used to find the MLE of ρ based on profile likelihood [21, 20]. These involve calculating $|\boldsymbol{A}|$ multiple times, which is computational intensive if A is large. Thus, this method is difficult to scale up.

Ord [43] has proposed a way to avoid evaluating $|\boldsymbol{A}|$ multiple times by using the relation:

$$\ln|\boldsymbol{A}| = \ln|(\boldsymbol{I}_n - \rho\boldsymbol{W})| = \sum_{i=1}^{n}\ln(1-\rho\lambda_i), \tag{2.11}$$

where, λ_i are the eigenvalues of $\boldsymbol{W}$. Since $\boldsymbol{W}$ is known and fixed, we only need to calculate its eigenvalues once and $\ln|\boldsymbol{A}|$ can be calculated very easily once the eigenvalues of $\boldsymbol{W}$ are known. This method can be more efficient than

directly evaluating $|\boldsymbol{A}|$ multiple times. However, calculation of eigenvalues is also computational intensive for large matrices. Also, Kelejian and Prucha [21] pointed out that for a general large matrix without any special structure, the eigenvalues may not be calculated correctly using current computation technology [1]. Thus, it is desirable an estimation method based on maximum likelihood that can scale well. Stochastic gradient descent is an optimization algorithm that can serve this purpose.

2.1.4 Outline

This chapter is organized as follows. In Section 2.2, we discuss the difficulities in applying SGD directly to the SAR model. Section 2.3 develops the SGD algorithm by writing the derivative of the overall log-likelihood as a sum of derivative of log-likelihood for each data point and also introduces the perturbation method for CIs construction [13]. Section 2.4 studies the asymptotic properties of the SGD estimators and Section 2.5 examines the finite sample properties with simulations. Section 2.6 proposes some methods to increase the empirical coverage of the constructed CI for ρ. Section 2.7 briefly describes the SGD algorithm for spatial model with autoregressive disturbance. Section 2.8 develops the 2SLS based SGD algorithm and studies its finite sample properties with simulations. Finally, we summarize and discuss the results in Section 2.9.

2.2 Difficulties in Applying SGD Directly

To develop the SGD algorithm for estimating parameters for SAR model, we start by trying following the same SGD algorithm used for parameter estimation of linear

[1] They concluded this by calculating eigenvalues for matrices with all real eigenvalues. The absolute value of imaginary parts of some of calculated eigenvalues are more than 0.5. We also calculated the eigenvalues of some matrices with all real eigenvalues. The imaginary parts of the calculated eigenvalues are very close to 0. This might due to the improvement of computational technology.

regression model. As discussed in Section 1.1.2, the log-likelihood for SAR model is:

$$\ell(\boldsymbol{\theta}|\boldsymbol{y}) = -\frac{\ln(\sigma^2)}{2}n - \frac{(\boldsymbol{A}\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})^T(\boldsymbol{A}\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})}{2\sigma^2} + \ln|\boldsymbol{A}| \tag{2.7}$$

Here, $\boldsymbol{\theta} = [\boldsymbol{\beta}^T, \sigma^2, \rho]^T$. If we can write $\ell(\boldsymbol{\theta})$ into the form of $\sum_i f_i(\boldsymbol{\theta})$, where f_i represent the log-likelihood of i-th data point, then we can apply SGD for parameter estimation by minimizing $R(\boldsymbol{\theta}) = \mathbb{E}(-f_i(\boldsymbol{\theta}))$. However, it is not easy to do so, because of the $\ln|\boldsymbol{A}|$ term. Though we can write $\ln|\boldsymbol{A}|$ in the summation form as in (2.11), we do not have a one-to-one correspondence between one data point and a specific eigenvalue of $\boldsymbol{W}$. This $\ln|\boldsymbol{A}|$ term appears in log-likelihood because data are correlated in SAR model. The fact that data are correlated, not independent, brings more challenges in applying SGD with SAR model.

We circumvent this problem by writing the derivative of the log-likelihood as a sum of one unit for each data point. As shown in Equation (1.1), it is the derivative not the target function that is used for updating parameters in each iterative steps of SGD. Thus, if we have the derivative of the likelihood for each data point, we can use it in each iterative step. The expression for derivative w.r.t. $\boldsymbol{\theta}$ is complicated. We work on the derivative w.r.t. $\boldsymbol{\beta}, \rho$ and σ^2 respectively:

$$\nabla\ell_{\boldsymbol{\beta}}(\boldsymbol{y}) = \sum_i \nabla\ell_{\boldsymbol{\beta},i}, \quad \nabla\ell_{\rho}(\boldsymbol{y}) = \sum_i \nabla\ell_{\rho,i}, \quad \nabla\ell_{\sigma^2}(\boldsymbol{y}) = \sum_i \nabla\ell_{\sigma^2,i} \tag{2.12}$$

$\nabla\ell_{\boldsymbol{\beta},i}, \nabla\ell_{\rho,i}$ and $\nabla\ell_{\sigma^2,i}$ are derivative of log-likelihood for i-th data point and can be used for updating SGD estimates.

2.3 Applying SGD for the SAR Model

2.3.1 Derivative of $\ell(\boldsymbol{\theta}|\boldsymbol{y})$ with respect to $\boldsymbol{\beta}$

Taking derivative $\ell(\boldsymbol{\theta}|\boldsymbol{y})$ w.r.t. $\boldsymbol{\beta}$, we get:

$$\nabla\ell_{\boldsymbol{\beta}} = \frac{1}{\sigma^2}(\boldsymbol{X}^T\boldsymbol{A}\boldsymbol{y} - \boldsymbol{X}^T\boldsymbol{X}\boldsymbol{\beta}). \tag{2.13}$$

The terms in the right hand side of (2.13) can be written in summation form:

$$\boldsymbol{X}^T\boldsymbol{X}\boldsymbol{\beta} = \sum_{i=1}^{n} \boldsymbol{x}_i\boldsymbol{x}_i^T\boldsymbol{\beta} \tag{2.14a}$$

$$\begin{aligned}\boldsymbol{X}^T\boldsymbol{A}\boldsymbol{y} &= \sum_{i=1}^{n} \boldsymbol{x}_i^T(\boldsymbol{A}_i\boldsymbol{y}) = \sum_{i=1} \boldsymbol{x}_i^T((\boldsymbol{I}-\rho\boldsymbol{W})_i\boldsymbol{y}) \\ &= \sum_{i=1} \boldsymbol{x}_i^T(y_i - \rho\bar{y}_i), \bar{y}_i = w_i\boldsymbol{y},\end{aligned} \tag{2.14b}$$

where, w_i is the i-th row of $\boldsymbol{W}$. Clearly $\boldsymbol{x}_i\boldsymbol{x}_i^T\boldsymbol{\beta}$ only depends on the i-th data point, and $\bar{y}_i = w_i\boldsymbol{y}$ is the weighted average of the neighbors of the i-th data point. The term $\boldsymbol{x}_i(y_i - \rho\bar{y}_i)$ involves not only the i-th data point but also its neighbors. Note that usually neighborhood matrix $\boldsymbol{W}$ is a sparse matrix and most elements of vector w_i are 0 [20] - the j-th element of w_i is non-zero if and only if data points i and j are neighbors. Though the expression for $\bar{y}_i$ involves $\boldsymbol{y}$, it is only involves data point i and the data around it, i.e., neighbors of the i-th data point. It is reasonable to treat the i-th data point and its neighbors as a whole unit, and we can write the derivative of the overall log-likelihood as a sum of the derivative of log-likelihood for each data unit as following:

$$\nabla\ell_{\boldsymbol{\beta},i} = \frac{1}{\sigma^2}\boldsymbol{x}_i(y_i - \rho\bar{y}_i - \boldsymbol{x}_i^T\boldsymbol{\beta}) \tag{2.15a}$$

$$\nabla\ell_{\boldsymbol{\beta}} = \sum_i \nabla\ell_{\boldsymbol{\beta},i} \tag{2.15b}$$

We use $\nabla\ell_{\boldsymbol{\beta},i}$ for our SGD algorithm in the SAR model. We can compare the derivative for the SAR model in (2.15a) with that for the linear model in (1.5). If we ignore the constant coefficient ($\frac{1}{\sigma^2}$ in (2.15a) and 2 in (1.5)), the only difference between (2.15a) and (1.5) is that y_i in (1.5) is replaced by $y_i - \rho\bar{y}_i$ in (2.15a). This is consistent with the nature of the SAR model, which is that the response variable is affected both by the neighbors and the covariates. If the effect of neighbors is

subtracted, the left over part, $y_i - \rho\bar{y}_i$, is equivalent to y_i in linear model. Also, this suggests that if ρ is known, the SGD algorithm for estimating $\boldsymbol{\beta}$ in SAR model is simplified to that of the linear model.

2.3.2 Derivative of $\ell(\boldsymbol{\theta}|\boldsymbol{y})$ with respect to ρ

Taking derivative of $\ell(\boldsymbol{\theta}|\boldsymbol{y})$ w.r.t. ρ, we get:

$$\nabla\ell_\rho = -tr(\boldsymbol{A}^{-1}\boldsymbol{W}) + \frac{(\boldsymbol{A}\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})^T\boldsymbol{W}\boldsymbol{y}}{\sigma^2}. \tag{2.16}$$

The $-tr(\boldsymbol{A}^{-1}\boldsymbol{W})$ term is the derivative of $\ln(\boldsymbol{A})$ w.r.t. ρ:

$$\frac{d(\ln(|\boldsymbol{A}|))}{d\rho} = tr(\boldsymbol{A}^{-1}\frac{d\boldsymbol{A}}{d\rho}) = tr(\boldsymbol{A}^{-1}\frac{d(\boldsymbol{I} - \rho\boldsymbol{W})}{d\rho}) = -tr(\boldsymbol{A}^{-1}\boldsymbol{W}). \tag{2.17}$$

We then write $\nabla\ell_\rho$ as a sum of the derivative of log-likelihood for each data point. The term $(\boldsymbol{A}\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})^T\boldsymbol{W}\boldsymbol{y}$ on the right hand side of (2.16) can be written as:

$$(\boldsymbol{A}\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})^T\boldsymbol{W}\boldsymbol{y} = \sum_{i=1}^{n}[(y_i - \rho\bar{y}_i - \boldsymbol{x}_i^T\boldsymbol{\beta})\bar{y}_i] \tag{2.18}$$

As discussed above, if we treat y_i and the mean of its neighbors $\bar{y}_i$ as a single unit, we can associate the summand in (2.18) with the i-th data point. The first term on the right hand of (2.16) can be written as:

$$\begin{aligned} tr(\boldsymbol{A}^{-1}\boldsymbol{W}) &= tr(\boldsymbol{A}^{-1}\frac{1}{\rho}(\boldsymbol{I} - \boldsymbol{A})), \text{ since } \boldsymbol{A} = \boldsymbol{I} - \rho\boldsymbol{W} \\ &= tr((\frac{1}{\rho}(\boldsymbol{A}^{-1} - \boldsymbol{I}) = \frac{1}{\rho}tr(\boldsymbol{A}^{-1})) - \frac{n}{\rho} \\ &= \frac{1}{\rho}\sum_{i}^{n}[(\boldsymbol{A}^{-1})_{ii} - 1]. \end{aligned} \tag{2.19}$$

Here, $(\boldsymbol{A}^{-1})_{ii}$ is the i-th diagonal element of $\boldsymbol{A}^{-1}$. If we want to use (2.19) to write $\nabla\ell_\rho$ as summation of individual terms, we need to prove that there is a

one-to-one correspondence between $(\boldsymbol{A}^{-1})_{ii}$ and the i-th data point. This is shown by **Proposition 1** stated below (see Appendix for proof).

Proposition 1. *A one-to-one correspondence exists between each diagonal element of $\boldsymbol{A}^{-1}$ and each data point of the SAR model.*

Thus, we can write the derivative of the overall log-likelihood w.r.t. ρ as a sum of derivative log-likelihoods w.r.t. ρ for each data as following:

$$\nabla \ell_{\rho,i} = -\frac{1}{\rho}((\boldsymbol{A}^{-1})_{ii} - 1) + \frac{1}{\sigma^2}(y_i - \rho\bar{y}_i - \boldsymbol{x}_i^T\boldsymbol{\beta})\bar{y}_i \tag{2.20a}$$

$$\nabla \ell_{\rho} = \sum_i \nabla \ell_{\rho,i} \tag{2.20b}$$

We can use $\nabla \ell_{\rho,i}$ for our SGD algorithm in SAR model. For each recursive step, ρ is updated, and as a result $\boldsymbol{A}$ is also updated. In this SGD algorithm we need to calculate the inverse of $\boldsymbol{A}$ and get the i-th diagonal element for each recursive step. Since calculation of matrix inverses is computationally heavy, it can affect the scalability of the SGD algorithm. Fortunately we can avoid this with the following **Proposition 2**.

Proposition 2. $\boldsymbol{A}^{-1} = (\boldsymbol{I} - \rho\boldsymbol{W})^{-1} = \sum_{k=0}^{\infty}(\rho\boldsymbol{W})^k$, *given $\boldsymbol{W}$ is row normalized and* $\rho \in (-1, 1)$.

Proof. $\boldsymbol{W}$ is row normalized, thus, $\|\boldsymbol{W}\|_\infty \leq 1$. Since $|\rho| < 1$, $\|\rho\boldsymbol{W}\|_\infty < 1$. Then we can apply Lemma 2.3.3 on P74 of [15]. □

In practice, if ρ is bounded away from 1, one can truncate the sum to, say, K=30 terms and this results in a negligible error [20]. We can the use $\sum_{k=0}^{K}\rho^k(\boldsymbol{W})_{ii}^k$ to approximate the i-th diagonal element of $\boldsymbol{A}^{-1}$. Since $\boldsymbol{W}$ is known and fixed, we only need to calculate $\boldsymbol{W}^k$ once and save all its diagonal elements for $k = 1, \ldots, K$. Then whenever we need to calculate the i-th diagonal element of $\boldsymbol{A}^{-1}$, we can just

get the i-th diagonal element of $\boldsymbol{W}^k$, multiply with ρ^k and sum for $k = 1, \ldots, K$. Thus, for each recursive step of SGD, to calculate $(\boldsymbol{A}^{-1})_{ii}$ we only need to perform some scalar multiplication and summation, which is much faster than calculating the inverse of a matrix.

2.3.3 Derivative of $\ell(\boldsymbol{\theta}|\boldsymbol{y})$ with respect to σ^2

Sometimes the variance σ^2 is not of great interest. For example, in linear regression, if we are only interested in estimating $\boldsymbol{\beta}$, we can use SGD to only estimate $\boldsymbol{\beta}$ and avoid estimating the variance σ^2, as discussed in Section 2.1. However, for the SAR model σ^2 appears in the derivative of log-likelihood w.r.t. ρ, $\nabla \ell_{\rho,i}$ as shown in (2.20a). Thus, σ^2 can not be omitted and an estimate of σ^2 is needed for estimating ρ. Also, the estimation of $\boldsymbol{\beta}$ requires the estimation of ρ, and vice versa. Thus, we need to estimate all three parameters, $\boldsymbol{\beta}, \rho, \sigma^2$ together even if we are only interested in some of them.

Taking derivative of $\ell(\boldsymbol{\theta}|\boldsymbol{y})$ w.r.t. σ^2, we get:

$$\nabla \ell_{\sigma^2} = -\frac{n}{2\sigma^2} + \frac{1}{2(\sigma^2)^2}(\boldsymbol{A}\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})^T(\boldsymbol{A}\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}) \tag{2.21}$$

The term of $(\boldsymbol{A}\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})^T(\boldsymbol{A}\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})$ can be written as

$$(\boldsymbol{A}\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})^T(\boldsymbol{A}\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}) = \sum_{i=1}^{n}(y_i - \rho\bar{y}_i - \boldsymbol{x}_i^T\boldsymbol{\beta})^2. \tag{2.22}$$

Similar to what was discussed before, if we treat y_i and its neighbors as a single unit, the summand in right hand side of (2.22) only involves the i-th data unit. We can write the derivative of the overall log-likelihood w.r.t. σ^2 as a sum of derivative of log-likelihoods w.r.t. σ^2 for each data unit:

$$\nabla \ell_{\sigma^2,i} = -\frac{1}{2\sigma^2} + \frac{1}{2(\sigma^2)^2}(y_i - \rho\bar{y}_i - \boldsymbol{x}_i^T\boldsymbol{\beta})^2, \tag{2.23a}$$

$$\nabla \ell_{\sigma^2} = \sum_i \ell_{\sigma^2,i}. \tag{2.23b}$$

$\nabla \ell_{\sigma^2,i}$ can be used for our SGD algorithm for the SAR model.

2.3.4 SGD algorithm

With all preliminary steps discussed above, we can now implement the SGD algorithm for the SAR model. Given arbitrary initial values $\hat{\boldsymbol{\beta}}_0, \hat{\rho}_0$ and $\hat{\sigma^2}_0$, parameter estimates $\hat{\boldsymbol{\beta}}_k, \hat{\boldsymbol{\rho}}_k$ and $\hat{\sigma^2}_k$ can be updated as:

$$
\begin{aligned}
\hat{\boldsymbol{\beta}}_k &= \hat{\boldsymbol{\beta}}_{k-1} + \gamma_k \nabla \ell_{\boldsymbol{\beta},k}(\hat{\boldsymbol{\beta}}_{k-1}, \hat{\rho}_{k-1}, \hat{\sigma^2}_{k-1}) \\
&= \hat{\boldsymbol{\beta}}_{k-1} + \gamma_k \frac{1}{\hat{\sigma^2}_{k-1}} \boldsymbol{x}_k (y_k - \hat{\rho}_{k-1} \bar{y}_k - \boldsymbol{x}_k^T \hat{\boldsymbol{\beta}}_{k-1}) \\
\hat{\rho}_k &= \hat{\rho}_{k-1} + \gamma_k \nabla \ell_{\rho,k}(\hat{\boldsymbol{\beta}}_{k-1}, \hat{\rho}_{k-1}, \hat{\sigma^2}_{k-1}) \\
&= \hat{\rho}_{k-1} + \gamma_k [-\frac{1}{\hat{\rho}_{k-1}}((\boldsymbol{A}(\hat{\rho}_{k-1})^{-1})_{kk} - 1) + \frac{1}{\hat{\sigma^2}_{k-1}}(y_k - \hat{\rho}_{k-1}\bar{y}_k - \boldsymbol{x}_k^T \hat{\boldsymbol{\beta}}_{k-1})\bar{y}_k]
\end{aligned}
\tag{2.24}
$$

$$
\begin{aligned}
\hat{\sigma^2}_k &= \hat{\sigma^2_0}_{k-1} + \gamma_k \nabla \ell_{\sigma_0^2,k}(\hat{\boldsymbol{\beta}}_{k-1}, \hat{\rho}_{k-1}, \hat{\sigma^2}_{k-1}) \\
&= \hat{\sigma^2}_{k-1} + \gamma_k [-\frac{1}{2\hat{\sigma^2}_{k-1}} + \frac{1}{2(\hat{\sigma^2}_{k-1})^2}(y_k - \hat{\rho}_{k-1}\bar{y}_k - \boldsymbol{x}_k^T \hat{\boldsymbol{\beta}}_{k-1})^2] \\
\bar{\boldsymbol{\beta}}_k &= \frac{1}{k}\sum_{i=1}^{k} \hat{\boldsymbol{\beta}}_i, \quad \bar{\rho}_k = \frac{1}{k}\sum_{i=1}^{k} \hat{\rho}_i, \quad \bar{\sigma^2}_k = \frac{1}{k}\sum_{i=1}^{k} \hat{\sigma^2}_i
\end{aligned}
$$

where, $\gamma_k = \gamma_1 k^{-\alpha}, \alpha \in (0.5, 1)$, $\bar{y}_k = w_k \boldsymbol{y}$ and w_i is the i-th row of neighborhood matrix $\boldsymbol{W}$. The term $(\boldsymbol{A}(\hat{\rho}_{k-1})^{-1})_{kk}$ means that we use $\hat{\rho}_{k-1}$ to calculate $\boldsymbol{A}$, take its inverse, and extract the k-th diagonal element (This is only to illustrate the meaning of this term. For real application we do not directly calculate $\boldsymbol{A}^{-1}$. We use the approximation method discussed above). Another point worth mentioning is that when updating $\hat{\rho}_k$ and $\hat{\sigma^2}_k$, $\hat{\boldsymbol{\beta}}_{k-1}$ is used rather than $\hat{\boldsymbol{\beta}}_k$. What we applied here is called simultaneous updating. There are also algorithms do not apply simultaneous updating [16]. These algorithms would use $\hat{\boldsymbol{\beta}}_k$ not $\hat{\boldsymbol{\beta}}_{k-1}$ for calculating $\hat{\rho}_k$ and $\hat{\boldsymbol{\beta}}_k$ not

$\hat{\boldsymbol{\beta}}_{k-1}$, $\hat{\rho}_k$ not $\hat{\rho}_{k-1}$ for calculating $\hat{\sigma^2}_k$. There are different convergence properties for these two different ways. We will use simultaneous updating for this project.

This algorithm (2.24) as written does not consider the constrain that $\rho \in (-1, 1)$ and $\sigma^2 > 0$. When true value of ρ is close to 1 or -1, the estimate of ρ could end up outside the range for ρ. Also, it is possible for the estimate of σ^2 to be negative when true value of σ^2 is close to 0. To incorporate these two constrains, we introduce two more parameters η and ϕ. The relation between ρ, σ^2 and η, ϕ are $\rho = \sin\eta$ and $\sigma^2 = e^{\phi}$. Instead of directly updating ρ and σ^2, η and ϕ are updated in each recursive step and then estimates of ρ and σ^2 are calculated based on the relation above. In this way, we can guarantee that ρ is in the range of $(-1, 1)$ and σ^2 is always positive. To update η and ϕ in each recursive step, we need to calculate the derivative of log-likelihood for each data w.r.t. η and ϕ. They can be easily calculated by chain rule as shown below:

$$\nabla\ell_{\eta,i} = \nabla\ell_{\rho,i}\cos\eta, \quad \nabla\ell_{\phi,i} = \nabla\ell_{\sigma^2,i}e^{\phi}. \tag{2.25}$$

Thus, we got our updated SGD algorithm for parameter estimation of the SAR model. Given arbitrary initial values $\hat{\boldsymbol{\beta}}_0, \hat{\rho}_0$ and $\hat{\sigma^2}_0$, we first calculate the initial value for $\hat{\eta}_0$ and $\hat{\phi}_0$ by $\hat{\eta}_0 = \arcsin\hat{\rho}_0$ and $\hat{\phi}_0 = \ln\hat{\sigma^2}_0$. The parameter estimates can be updated as:

$$\begin{aligned}
\hat{\boldsymbol{\beta}}_k &= \hat{\boldsymbol{\beta}}_{k-1} + \gamma_k\nabla\ell_{\boldsymbol{\beta},k}(\hat{\boldsymbol{\beta}}_{k-1}, \hat{\rho}_{k-1}, \hat{\sigma^2}_{k-1}) \\
\hat{\eta}_k &= \hat{\eta}_{k-1} + \gamma_k\nabla\ell_{\rho,k}(\hat{\boldsymbol{\beta}}_{k-1}, \hat{\rho}_{k-1}, \hat{\sigma^2}_{k-1})\cos\hat{\eta}_{k-1} \\
\hat{\phi}_k &= \hat{\phi}_{k-1} + \gamma_k\nabla\ell_{\sigma_0^2,k}(\hat{\boldsymbol{\beta}}_{k-1}, \hat{\rho}_{k-1}, \hat{\sigma^2}_{k-1})e^{\hat{\phi}_{k-1}} \\
\hat{\rho}_k &= \sin\hat{\eta}_k, \quad \hat{\sigma^2}_k = e^{\hat{\phi}_k} \\
\bar{\boldsymbol{\beta}}_k &= \frac{1}{k}\sum_{i=1}^{k}\hat{\boldsymbol{\beta}}_i, \quad \bar{\rho}_k = \frac{1}{k}\sum_{i=1}^{k}\hat{\rho}_i, \quad \bar{\hat{\sigma^2}}_k = \frac{1}{k}\sum_{i=1}^{k}\hat{\sigma^2}_i
\end{aligned} \tag{2.26}$$

There are two ways for calculating the mean to obtain the final estimate for ρ and σ^2. One is to direct taking the mean of $\hat{\rho}$ and $\hat{\sigma^2}$, respectively as shown in (2.26). The other method is to first calculate $\bar{\eta}_n$ and $\bar{\phi}_n$ as $\bar{\eta}_n = 1/n \sum_{i=1}^n \hat{\eta}_i$ and $\bar{\phi}_n = 1/n \sum_{i=1}^n \hat{\phi}_i$, and set $\bar{\rho}_n = \sin \bar{\eta}_n$ and $\bar{\sigma^2}_n = e^{\bar{\phi}_n}$. Our simulation results (not shown) suggest that the former method works better.

For confidence interval construction, we applied the online bootstrap resampling method proposed by Fang et al. [13, 12]. Given arbitrary initial values $\hat{\boldsymbol{\beta}}_0, \hat{\rho}_0$ and $\hat{\sigma^2}_0$, let $\hat{\boldsymbol{\beta}}_0^{*,b} \equiv \hat{\boldsymbol{\beta}}_0, \hat{\rho}_0^{*,b} \equiv \hat{\rho}_0$ and $\hat{\sigma^2}_0^{*,b} \equiv \hat{\sigma^2}_0$, for $b = 1, 2, \ldots, B$. Similar to the SGD estimates, we use η and ϕ for the constrains of ρ and σ^2 with $\hat{\eta}_0^{*,b} = \arcsin \hat{\rho}_0^{*,b}, \hat{\phi}_0^{*,b} = \ln \hat{\sigma^2}_0^{*,b}$. These perturbed estimate can be updated as:

$$
\begin{aligned}
\hat{\boldsymbol{\beta}}_k^{*,b} &= \hat{\boldsymbol{\beta}}_{k-1}^{*,b} + \gamma_k W_k^{*,b} \nabla \ell_{\boldsymbol{\beta},k}(\hat{\boldsymbol{\beta}}_{k-1}^{*,b}, \hat{\rho}_{k-1}^{*,b}, \hat{\sigma^2}_{k-1}^{*,b}) \\
\hat{\eta}_k^{*,b} &= \hat{\eta}_{k-1}^{*,b} + \gamma_k W_k^{*,b} \nabla \ell_{\rho,k}(\hat{\boldsymbol{\beta}}_{k-1}^{*,b}, \hat{\rho}_{k-1}^{*,b}, \hat{\sigma^2}_{k-1}^{*,b}) \cos \hat{\eta}_{k-1}^{*,b} \\
\hat{\phi}_k^{*,b} &= \hat{\phi}_{k-1}^{*,b} + \gamma_k W_k^{*,b} \nabla \ell_{\sigma_0^2,k}(\hat{\boldsymbol{\beta}}_{k-1}^{*,b}, \hat{\rho}_{k-1}^{*,b}, \hat{\sigma^2}_{k-1}^{*,b}) e^{\hat{\phi}_{k-1}^{*,b}} \\
\hat{\rho}_k^{*,b} &= \sin \hat{\eta}_k^{*,b}, \quad \hat{\sigma^2}_k^{*,b} = e^{\hat{\phi}_k^{*,b}} \\
\bar{\boldsymbol{\beta}}_k^{*,b} &= \frac{1}{k} \sum_{i=1}^k \hat{\boldsymbol{\beta}}_i^{*,b}, \quad \bar{\rho}_k^{*,b} = \frac{1}{k} \sum_{i=1}^k \hat{\rho}_i^{*,b}, \quad \bar{\sigma^2}_k^{*,b} = \frac{1}{k} \sum_{i=1}^k \hat{\sigma^2}_i^{*,b}
\end{aligned}
\tag{2.27}
$$

Here, $W_k^{*,b} \overset{iid}{\sim} W, k = 1, 2, \cdots, n, W > 0, E(W) = Var(W) = 1$. With the perturbed estimates, we can construct confidence intervals using the methods discussed in Section 1.1.2.

2.4 Theoretical Properties

In this section, we describes the theoretical properties of the SGD estimates and the perturbed estimates. Let $\boldsymbol{\theta} = [\boldsymbol{\beta}^T, \sigma^2, \rho]^T$ and $\ell(\boldsymbol{\theta})$ denote the log-likelihood of the SAR model. We have discussed that it is not easy to write out $\ell_i(\boldsymbol{\theta})$, which is the contribution of the i-th data point to the log-likelihood. Thus, there is no such loss

function, defined as $R(\boldsymbol{\theta}) = -\mathbb{E}[\ell_i(\boldsymbol{\theta})]$, to minimize. However, this loss function, $R(\boldsymbol{\theta})$ is usually needed for developing asymptotic properties. In this dissertation, for SAR model, the estimates are updated based on $\nabla\ell_i(\boldsymbol{\theta})$, which is the contribution of i-th data unit to the derivative of log-likelihood. However, $\mathbb{E}[\nabla\ell_i(\boldsymbol{\theta})|\boldsymbol{Y}]$ usually depends on i (Though we have shown that when $\boldsymbol{\theta} = \boldsymbol{\theta}_0$, the true value, $\mathbb{E}[\nabla\ell_i(\boldsymbol{\theta})|\boldsymbol{Y} = \mathbf{0}$ and does not depend on i. See Appendix B.2). Thus, $\mathbb{E}[\nabla\ell_i(\boldsymbol{\theta})|\boldsymbol{Y}]$ cannot be used to develop asymptotic properties either.

In this section, we develops the asymptotic properties with a setting slightly different from the one discussed above in Section 2.3.4. Let $Z_1, Z_2, \cdots$ be i.i.d. samples for Z, which follows the SAR model shown in Equation (2.1). Each i.i.d sample represents all K data points in the SAR model. Let $\ell(\boldsymbol{\theta})$ be the log-likelihood for this one dataset and SGD is then used to estimate $\boldsymbol{\theta}$ by minimizing $L(\boldsymbol{\theta}) = \mathbb{E}[-\ell(\boldsymbol{\theta})]$. One dataset is used as one data point for this SGD procedure. Given a initial value for $\boldsymbol{\theta}$ as $\hat{\boldsymbol{\theta}}_0$ and the SGD estimates and perturbed estimates are updated as:

$$\begin{aligned} \hat{\boldsymbol{\theta}}_k &= \hat{\boldsymbol{\theta}}_{k-1} + \gamma_k \nabla\ell(\hat{\boldsymbol{\theta}}_{k-1}, Z_k) \\ \hat{\boldsymbol{\theta}}^*_k &= \hat{\boldsymbol{\theta}}^*_{k-1} + \gamma_k W_k \nabla\ell(\hat{\boldsymbol{\theta}}^*_{k-1}, Z_k) \end{aligned} \tag{2.28}$$

Also, we have the following Assumptions.

- **A1**. Neighborhood matrix $\boldsymbol{W}$ is row-normalized and symmetric.
- **A2**. Exist $a > 0$ such that $\sigma^2 \in [a, \infty)$; exist $b > 0$, such that $\|\boldsymbol{\beta}\| \in [0, b]$; exist ρ_{min}, ρ_{max} with $0 < \rho_{min} \leq \rho_{max} < 1$ and $|\rho| \in [\rho_{min}, \rho_{max}]$.
- **A3**. Assumptions 1-8 from [33]. They are required for the existence and uniqueness of MLE for the SAR model.
- **A4**. The learning rates are chosen as $\gamma_k = \gamma_1 k^{-\alpha}$ and $\alpha \in (0.5, 1)$.
- **A5**. The perturbation variables, $W_1, W_2, \ldots$, are non-negative i.i.d. random variables satisfying that $\mathbb{E}(W) = \mathrm{Var}(W) = 1$.

Assumption A4-A5 are from [13]. Let $\boldsymbol{S}(\boldsymbol{\theta}) = \nabla^2 L(\boldsymbol{\theta})$ and $\boldsymbol{S}_0 = \boldsymbol{S}(\boldsymbol{\theta}_0)$, $\boldsymbol{V}_0 = \mathbb{E}\{[\nabla\ell(\boldsymbol{\theta}_0, Z)][\nabla\ell(\boldsymbol{\theta}_0, Z)]^T]\}$ and we get the following theorem (See Appendix A.1 for proof).

Theorem 1. *Given Assumption A1-A4, then*

$$\sqrt{n}(\bar{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0) \implies N(\mathbf{0}, \boldsymbol{S}_0^{-1}\boldsymbol{V}_0\boldsymbol{S}_0^{-1}), \text{ in distribution, as } n \to \infty$$

Also, we can derive the following two theorems, which are essentially Theorems 2 and 3 of [13].

Theorem 2. *Given Assumption A1-A4, and the perturbed variables, W_1, W_2, $\cdots$, are non-negative i.i.d. random variables with $\mathbb{E}(W) = 1$, then we have,*

$$\sqrt{n}\left(\bar{\boldsymbol{\theta}}_n^* - \boldsymbol{\theta}_0\right) = -\frac{1}{\sqrt{n}}\boldsymbol{S}_0^{-1}\sum_{i=1}^{n} W_i \nabla\ell\left(\boldsymbol{\theta}_0; Z_i\right) + o_p(1) \tag{2.29}$$

By **Theorem** 2, let $W \equiv 1$, we can derive the the following representation for $\bar{\theta}_n$,

$$\sqrt{n}\left(\bar{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0\right) = -\frac{1}{\sqrt{n}}\boldsymbol{S}_0^{-1}\sum_{i=1}^{n} \nabla\ell\left(\boldsymbol{\theta}_0; Z_i\right) + o_p(1) \tag{2.30}$$

Consider the difference between Equations (2.29) and (2.30), we have

$$\sqrt{n}\left(\bar{\boldsymbol{\theta}}_n^* - \boldsymbol{\theta}_n\right) = -\frac{1}{\sqrt{n}}\boldsymbol{S}_0^{-1}\sum_{i=1}^{n} \left(W_i - 1\right)\nabla\ell\left(\boldsymbol{\theta}_0; Z_i\right) + o_p(1). \tag{2.31}$$

Let $\mathbb{P}^*$ denote the conditional probability and expectation given the data. Starting from Equation (2.31), we derive the following theorem.

Theorem 3. *If Assumptions A1 to A5 hold, then we have*

$$\sup_{v\in\mathcal{R}^p}\left|\mathbb{P}^*\left(\sqrt{n}\left(\bar{\boldsymbol{\theta}}_n^* - \bar{\boldsymbol{\theta}}_n\right) \le v\right) - \mathbb{P}^*\left(\sqrt{n}\left(\bar{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0\right) \le v\right)\right| \to 0, \text{ in probability.} \tag{2.32}$$

Proof for **Theorem 2** and **3** are straightforward, since we have verified that our assumptions imply assumptions listed in [13] as shown in Appendix A.1 when proving **Theorem 1**.

2.5 Simulation Studies

2.5.1 Simulation settings

We use simulations to study the finite sample properties of the SGD algorithms proposed for SAR model parameter estimation. Data samples are located in a regular grid as shown in Figure 2.2. Each small square represents one data location and we have 81 data values in this 9 by 9 grid. We consider three different structures in our simulation studies. The first is the '4-neighbors' structure, where two data points are neighbors if and only if they share a common border. Consider the highlighted yellow data point. In this '4-neighbors' structure the four data containing the thick blue arrows are its neighbors. The second is the '8-neighbor' structure. In this structure all the 8 other data points inside the red 3 by 3 block around the yellow highlighted data point are its neighbors. Similarly, for the '24-neighbor' structure, all the other 24 data points inside the green 5 by 5 block around the yellow highlighted data point are its neighbors. These 4, 8, or 24 are the number of neighbors for a majority of the data points and for data points located on the edge of the grid the numbers of neighbors are less. For example, with '4-neighbors' structure, data labelled 1 only has 2 neighbors, data 2 and data 10. To construct the neighborhood matrix for n data points, we first generate a n by n matrix $\boldsymbol{C}$, with:

$$c_{i,j} = \begin{cases} 1, & \text{if data points } i \text{ and } j \text{ are neighbors} \\ 0, & \text{else} \end{cases} \tag{2.33}$$

The diagonal element of $\boldsymbol{C}$ are set to be 0. Then the neighborhood matrix $\boldsymbol{W}$ are calculated by row normalization of $\boldsymbol{C}$ as shown below.

$$w_{ij} = c_{ij}/s_i, \quad s_i = \sum_j c_{ij} \tag{2.34}$$

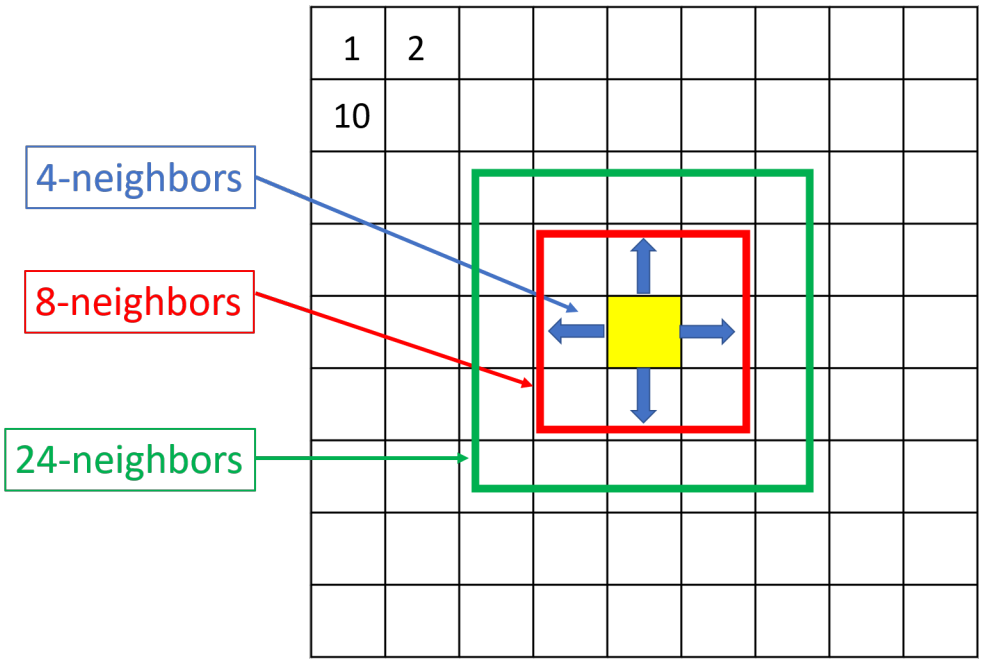


Figure 2.2 Three neighborhood structures.

For simulations, we consider the covariates matrix $\boldsymbol{X} = [\boldsymbol{1}, \boldsymbol{X}_1, \boldsymbol{X}_2]_{n\times 3}$ and each element of $\boldsymbol{X}_1, \boldsymbol{X}_2$ are generated independently from the uniform $(-1, 1)$ distribution. The true values for $\boldsymbol{\beta}$ is $\boldsymbol{\beta}_0 = [\beta_0, \beta_1, \beta_2]^T = [0.5, 0.5, -0.5]^T$, while the true value for σ^2 is set to be $\sigma_0^2 = 1$. We try several different values for parameter ρ and denote its true value as ρ_0. Simulation data are generated as follows: given $\boldsymbol{W}$ and ρ_0, we first calculate $\boldsymbol{A}^{-1}$ as $(\boldsymbol{I} - \rho_0\boldsymbol{W})^{-1}$. The inverse is calculated using **Proposition 2**. We truncate the sum for the first 81 terms, i.e., $\boldsymbol{A}_0^{-1} \approx \sum_{k=0}^{80}(\rho_0\boldsymbol{W})^k$ (We used 81 terms instead of 30 terms as discussed in Section 2.3 to reduce the approximation error when $|\rho|$ is close to 1). Then we sample $\boldsymbol{\epsilon}$ from the $N(\boldsymbol{0}, \sigma_0^2\boldsymbol{I})$ distribution. The response variable $\boldsymbol{y}$ is calculated as $\boldsymbol{A}_0^{-1}(\boldsymbol{X}\boldsymbol{\beta}_0 + \boldsymbol{\epsilon})$. Based on the SAR model shown in (2.1) we have $\boldsymbol{y} \sim N(\boldsymbol{A}_0^{-1}\boldsymbol{X}\boldsymbol{\beta}, \sigma_0^2\boldsymbol{A}_0^{-1}(\boldsymbol{A}_0^{-1})^T)$. We do not directly sample $\boldsymbol{y}$

from this normal distribution because directly sampling from a multivariate normal distribution is very time consuming. Since the elements of $\boldsymbol{\epsilon}$ are i.i.d., we can easily sample $\boldsymbol{\epsilon}$ by making n draws from the univariate normal distribution $N(0, \sigma_0^2)$. This is much faster than directly sampling $\boldsymbol{y}$ from a multivariate normal distribution.

For the learning rate, $\gamma_k = \gamma_1 k^{-\alpha}$ we use $\alpha = 2/3$. The value of γ_1 has to be carefully selected. Although in theory, the value of γ_1 does not affect the convergence of SGD estimates [45, 13], however, in practice the sample size is finite and the value of γ_1 can have a great impact. If γ_1 is too small, the convergence rate is slow and the estimate may not converge. If γ_1 is too large, the convergence rate can also be slow and sometimes the estimate may even diverge. There is not a standardized guideline for deciding the value for γ_1 to use. For our simulation study, we try several values and select the one that works best. For each of these simulation studies, we use 500 independent replicates. For each of these, the initial estimates of parameters are randomly generated. For this simulation, we discard the SGD estimates from the first 20% recursive steps since they are usually not stable.

2.5.2 Comparison of SGD estimates and MLE

First we compare the performance of SGD with maximizing the profile-likelihood (MLE method). For this study, we consider the '4-neighbors' neighborhood structure. The true value of ρ is chosen to be 0.3. As mentioned in Section 2.1, MLE of ρ cannot be expressed in closed form and a numerical method is needed for estimating it. We used golden section search [23, 1] to maximize (2.10). We need to directly evaluate the determinant of $\boldsymbol{A}$ multiple times or calculate the eigenvalues of $\boldsymbol{W}$ once and use (2.11) with different ρ values. We used a sample size of 22,500, the largest $\boldsymbol{W}$ we can use with our computation clusters available (Details about the clusters used can be found at `https://ist.njit.edu/high-performance-computing-hpc-clusters`). This corresponds to a 150 by 150 regular grid. Eigenvalue calculation of $\boldsymbol{W}$ for 22,500

data points takes about 16.5 hours. On contrast the time required for calculating $\boldsymbol{W}^k$ for $k = 0, \ldots, 80$ is only about 10 minutes. This suggests that SGD can really speed up the parameter estimation and the difference will be greater with larger sample size.

We perform 500 independent runs. For each run, we generate the data and perform parameter estimation with SGD or profile maximum likelihood on the same data. With eigenvalues of $\boldsymbol{W}$ and $\boldsymbol{W}^k$ for $k = 0, \ldots, 80$ calculated and stored, the time for each run is about 5 seconds for SGD and about 200 seconds for MLE. We calculate the mean and standard deviation (SD) of these 500 SGD estimates (see Table 2.1). For $\boldsymbol{\beta}$, we only show the results for β_1 since the results of β_1 and β_2 are very similar. The true values are given in parentheses beside the parameter name. The result suggests that both the SGD and ML estimates are close to the true value, with MLE closer to the true value and a smaller standard error. The estimates for ρ are also shown in Figure 2.3. Here, each pair of black and red dots with the same x coordinate represent the SGD and MLE estimates of ρ from the same simulated dataset, respectively. In most cases we see that the SGD and ML estimates and MLE are very close. However, there are some instances when the parameters are poorly estimated by SGD. The reason for this could be that the sample size is not large enough for convergence. In summary, this simulation suggests that our SGD algorithm works for SAR model at least in this setting. For the sample size of $22,500$, SGD estimates are very close to the MLE and SGD estimates and can be calculated much faster than MLE.

2.5.3 Effect of ignoring spatial correlation

Sometimes researchers ignore the spatial correlation and fit the regular linear regression (LR) model (shown in Equation (1.3)) for spatial correlated data. We study the effect of ignoring spatial correlation with simulations in this section. We use

Table 2.1 Comparison of SGD Estimate and MLE

method	parameter	mean	SD
SGD	β_1 (0.5)	0.500	0.015
	σ^2 (1.0)	1.003	0.011
	ρ (0.3)	0.298	0.010
MLE	β_1 (0.5)	0.500	0.012
	σ^2 (1.0)	0.999	0.010
	ρ (0.3)	0.300	0.009

'4-neighbors' neighborhood structure, set $\rho_0 = 0.3$, and generate 90,000 data points (in a 300 by 300 regular grid space) using the SAR model. We use Equations (17) and (18) in Fang, et al., 2018 [13] to obtain SGD estimates and perturbed estimates under LR model (ignoring spatial correlation), and compare them to estimates obtained using Equations (2.26) and (2.27) based on the correctly specified SAR model. To construct confidence intervals, we use $B = 200$ perturbed estimates.

Simulation results shown in Table 2.2 [2] suggest that the estimates from the SAR model are unbiased and the constructed CIs are at the nominal level. However, estimates from the LR model are biased and the constructed CIs below the nominal level. Thus, accounting for spatial correlation when data is spatially correlated is necessary for correct parameter estimation. Also, comparing with Table 2.1, where sample size is 25,000, we find that empirical standard deviation of estimates decrease as sample size increase. Also, the empirical coverage for CI for ρ is close to but not at the nominal level. A possible reason for this is that the online bootstrapping based inference method is designed for independent data and SAR data are dependent. In

[2]Here, the confidence interval is constructed for each individual parameter separately. It is not family wise confidence interval for several parameters. Same for tables below.

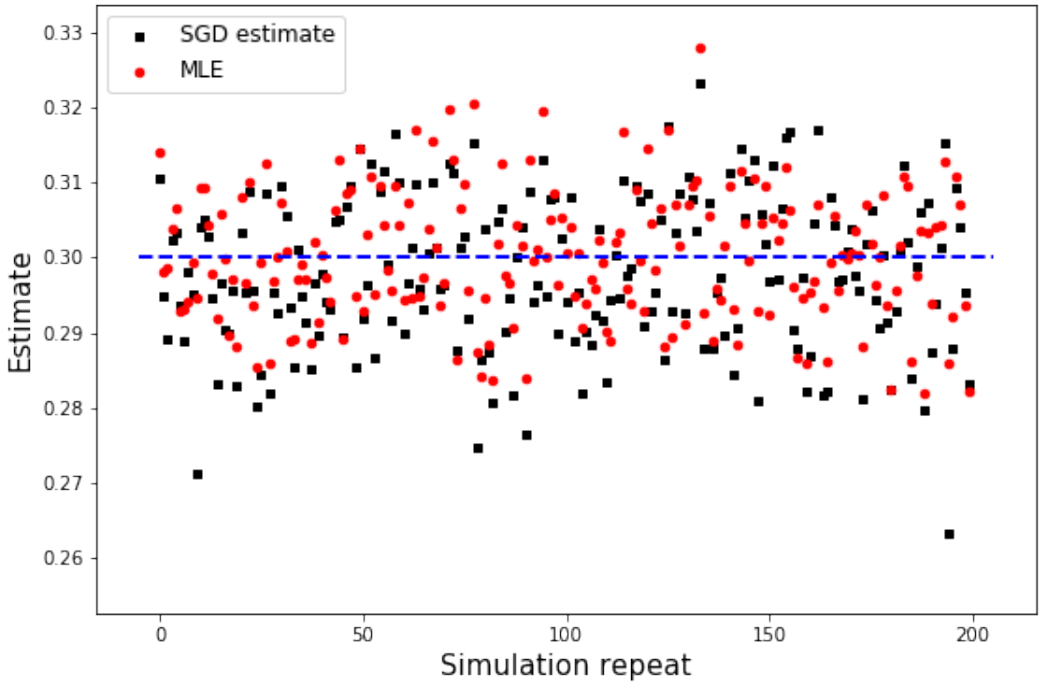


Figure 2.3 Comparison of SGD estimate and MLE of ρ.

Section 2.6, we propose some modifications to improve the empirical coverage of CIs for ρ.

2.5.4 Robustness of SGD algorithm

We study the robustness of SGD algorithm by varying the neighborhood structures, ρ_0, and order of data used in the algorithm.

Effect of neighborhood structures First, we study the effect of the neighborhood structure on the performance of SGD estimates. Three kinds of neighborhood structure: '4-neighbors', '8-neighbors', and '24-neighbors' are compared. Clearly, with the same spatial parameter, as the average number of neighbors for each data get larger, the spatial correlation between data gets larger also. ρ_0 is set to be 0.3 and 90,000 data points are generated based on each of these three neighborhood structures.

Table 2.2 Effect of Ignoring Spatial Correlation

model	parameter	mean	SD	coverage
SAR	β_1 (0.5)	0.500	0.006	0.958
	β_2 (-0.5)	-0.500	0.006	0.956
	ρ (0.3)	0.299	0.005	0.850
LR	β_1 (0.5)	0.512	0.006	0.594
	β_2 (-0.5)	-0.512	0.006	0.546

Table 2.3 shows the mean (SD) of the 500 estimates and the empirical coverage of the CIs. Comparing the result of the '4-neighbors' structure here with that from Table 2.1, we find that estimation improves with sample size. Also, the neighborhood structure has little effect on estimates of $\boldsymbol{\beta}$ and σ^2, but has a large effect on the estimate of ρ. As spatial correlation increases, both standard deviation and bias increase. Due to the limitation of computer memory, we cannot try sample sizes larger than 90,000. However, we suspect that estimate of ρ will improve with sample size like the case with the '4-neighbors' structure.

The coverage of CIs for $\boldsymbol{\beta}$ are close to 0.95 for all three neighborhood structures. Confidence interval coverage for σ^2 is close to 0.95 for the first two neighborhood structures but much lower for '24-neighbors' neighborhood structure. The confidence interval for ρ is not close to 0.95 for any of these three structures and the coverage is far below 0.95 for '24-neighborhood' structure. Thus, this result suggests that neighborhood structure has little effect on the CI coverage of $\boldsymbol{\beta}$; has some effect on that of σ^2, and has a big impact for ρ.

Effect of ρ_0 Besides the neighborhood structure, another factor that affects data correlation in the SAR model is the spatial parameter ρ. We want to investigate how the SGD algorithm performs with different ρ values. We consider ρ_0 equals to 0.2, 0.3,

Table 2.3 Effect of Neighborhood Structures

neigh struc	parameter	mean	SD	coverage
4-neighbors	β_1 (0.5)	0.500	0.006	0.958
	σ^2 (1.0)	1.001	0.006	0.934
	ρ (0.3)	0.299	0.005	0.850
8-neighbors	β_1 (0.5)	0.500	0.005	0.985
	σ^2 (1.0)	1.001	0.006	0.935
	ρ (0.3)	0.299	0.006	0.865
24-neighbors	β_1 (0.5)	0.500	0.006	0.940
	σ^2 (1.0)	1.001	0.006	0.900
	ρ (0.3)	0.296	0.036	0.580

0.7, 0.8, and -0.3. We use the '4-neighbors' neighborhood structure here. Different learning rates are used for the different ρ_0 values. The final learning rate for each ρ_0 is selected to be the best out of several learning rates tested for that ρ_0 value. We calculate the SGD estimate and construct CIs for each run and calculate the empirical coverage. To save space, we only show the results for β_1 and ρ here (Table 2.4)[3]. Wall suggested that higher $|\rho|$ does not necessarily mean higher spatial correlation [54]. To compare the spatial correlation of the data generated with different ρ_0 values, we calculated the Moran's I index for each dataset. The expected value of Moran's I is $-\frac{1}{n-1}$ and the value of that usually lies between -1 and 1. Values significantly below or above $-\frac{1}{n-1}$ indicate negative or positive spatial correlation, respectively. The last column of Table 2.4 shows the mean value of Moran's I calculated from each simulated dataset.

[3]The summary about estimates for $\boldsymbol{\beta}$ and ρ with $\rho_0 = 0.3$ are different from that shown in 2.3. This is due to randomness in data generation and estimates initialization in different simulation studies. This is also the reason for 'inconsistencies' among other tables in this chapter.

As shown in Table 2.4, Moran's I for positive ρ_0 are positive and that for negative ρ_0 are negative. Also, for positive ρ_0, as ρ_0 increasing, Moran's I also increases. This suggests that larger $|\rho_0|$ corresponds with larger spatial correlation. The means of the β_1, ρ estimates are close to their true value for all ρ_0 with standard errors roughly the same across all ρ_0. The confidence interval coverage for β_1 are all close to 0.95. There is a sightly increasing trend for the confidence interval coverage of ρ as ρ_0 increases from 0.2 to 0.8. Though ρ and $\boldsymbol{W}$ both affect the spatial correlation of the data, their effects on the CI coverage of our algorithm are different. For $\boldsymbol{W}$, as the averaged number of neighbors increasing, the empirical SD of ρ estimates increase and empirical coverages of CIs for ρ decrease. ρ_0 does not have much effect on the empirical SD of the ρ estimates and the empirical coverage of CIs increase as ρ_0 increasing from 0.2 to 0.8.

Table 2.4 Effect of ρ_0

ρ_0	parameter	mean	SD	coverage	Moran's I
0.2	β_1 (0.5)	0.500	0.006	0.980	0.1019
	ρ (0.2)	0.199	0.006	0.790	
0.3	β_1 (0.5)	0.500	0.006	0.940	0.1558
	ρ (0.3)	0.300	0.005	0.870	
0.7	β_1 (0.5)	0.500	0.006	0.960	0.4378
	ρ (0.7)	0.699	0.003	0.865	
0.8	β_1 (0.5)	0.501	0.008	0.935	0.5479
	ρ (0.8)	0.799	0.005	0.920	
-0.3	β_1 (0.5)	0.500	0.006	0.955	-0.1557
	ρ (-0.3)	-0.300	0.005	0.875	

Effect of data order For regular linear regression, the sequence of data used in SGD recursive step does not matter, since the data are uncorrelated. However, the sequence of data used in SGD for the SAR model might be important, since the data are dependent. We investigate the effect of sequence of data on SGD estimation.

For this study, we use the '4-neighbors' neighborhood structure, and $\rho_0 = 0.3$. We study four different data sequences. We label the data 1 to n by rows and from left to right in each row according their location in the regular grid space (See Figure 2.4 for an example with $n = 25$). The first data order we use is 1 to n in orders.

One possible reason we think might be responsible for the bad performance of the algorithm is that data are used multiple times in several recursive steps. We can illustrate this by taking data 13 in Figure 2.4 as an example. Data is used once when updating SGD estimate based on itself, data 13. It is also used as neighbors when updating estimate based on data 8, data 12, data 14, data 18. A piece of information used multiple times may affect statistical inference [52]. To study this, we separate our data into two parts, part A and part B in the following way: first put data 1 in part A and then put its neighbors into part B; then go to next data not in part A or part B and put it in part A and its neighbors into part B and so on. In setting 2, when running SGD algorithm, we first use data in part A and then those in part B. In this way, before the data in part B are used, all the data in part A are only used once. Since learning rate is decreasing, data used at beginning tends to have a larger impact on the final estimate. Using this 2nd data order might improve the performance of the CI construction algorithm.

For the regular data order used (from data 1 to data n), data used in consecutive recursive steps are highly correlated. Clearly correlation between data 1 and data 2 are higher than that between data 1 and data 20. We investigate whether using less correlated data in consecutive SGD recursive steps can improve performance. For settings 3 and 4, we use a randomized order of data for SGD. The difference between

these two settings is that same random order is used for all runs in setting 3, but each run has a different random order in setting 4. For each data realization, we apply SGD with these four different data orders. Table 2.5 shows the result. We find that the performance is very similar, suggesting that data order may not crucial in our SGD algorithm for SAR model.

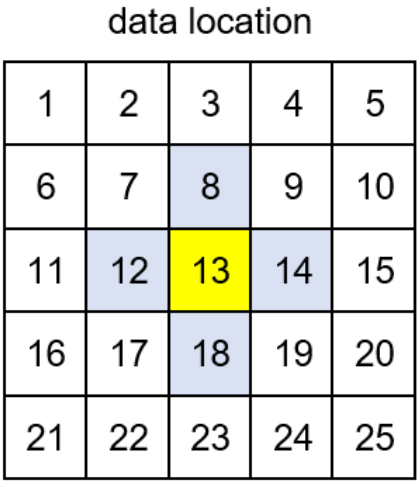


Figure 2.4 Illustration of data orders.

2.6 New Confidence Interval Construction Algorithm

This section summarizes the attempts to modify the confidence interval construction algorithm to improve the empirical coverage of CI for ρ.

2.6.1 Fisher transformation

Let $\theta_0, \bar{\theta}_n, \bar{\theta}_n^*$ be true value, the SGD estimate and, the perturbed SGD estimate, respectively. As discussed above, there are two ways to construct the confidence intervals for $\bar{\theta}_n$ based on $\bar{\theta}_n^*$. One is to construct the confidence interval based on the quantile of $\bar{\theta}_n^*$. This is based on assumption that the asymptotic distribution of $\sqrt{n}(\bar{\theta}^* - \bar{\theta}_n)$ is the same as that of $\sqrt{n}(\bar{\theta}_n - \theta_0)$. One way to generate confidence interval

Table 2.5 Effect of Data Orders

Settings	parameter	mean	SD	coverage
1	β_1 (0.5)	0.500	0.006	0.958
	ρ (0.3)	0.299	0.005	0.850
2	β_1 (0.5)	0.500	0.006	0.956
	ρ (0.3)	0.300	0.005	0.870
3	β_1 (0.5)	0.500	0.006	0.944
	ρ (0.3)	0.300	0.005	0.856
4	β_1 (0.5)	0.500	0.007	0.942
	ρ (0.3)	0.300	0.005	0.828

is to use the standard deviation of $\bar{\theta}_n^*$. This is based on the additional assumption that $\bar{\theta}_n^*$ is asymptotic normal. For ρ, its range is between -1 and 1, not $\mathbb{R}$, which is the domain for normal distribution. The asymptotic distribution of $\bar{\rho}_n^*$ cannot be normal due to this range constraint. For the algorithms described above (Equations (2.26) and 2.27), we let $\rho = \sin(\eta)$, but still the range for η is between $-\pi/2$ to $\pi/2$, not $\mathbb{R}$. Then we try Fisher transformation:

$$\begin{aligned} \eta = \frac{1}{2}\ln\frac{1+\rho}{1-\rho} \implies \rho = \frac{e^{2\eta}-1}{e^{2\eta}+1}, \\ \frac{\partial\rho}{\partial\eta} = \frac{4e^{2\eta}}{(e^{2\eta+1})^2}. \end{aligned} \tag{2.35}$$

The range for η is $\mathbb{R}$. We develop the SGD estimation and confidence interval construction algorithm by modifying Equations (2.26) and (2.27) according. Two methods are used to construct confidence intervals for ρ:

- **Method 1**. Get perturbed estimates $\bar{\rho}_n^* = \frac{e^{2\bar{\eta}_n^*}-1}{e^{2\bar{\eta}_n^*}+1}$ and then get the CIs for $\bar{\rho}_n$ using the SD of $\bar{\rho}_n^*$.

- **Method 2**. Get the CIs for $\bar{\eta}_n$ using the SD of $\bar{\eta}_n^*$ and then get the CIs for $\bar{\rho}_n$ by the relationship that $\rho = \frac{e^{2\eta}-1}{e^{2\eta}+1}$.

For this simulation, we use the '4-neighbors' neighborhood structure and $\rho_0 = 0.3$. The result is summarized in Table 2.6. The estimate and empirical coverage of CIs for β_1 and σ^2 for this Fisher transformation method are similar to that of the original algorithm (see the corresponding part in Tables 2.3 and 2.4). For ρ, neither Method 1 or Method 2 improves the empirical coverage of CIs compared with Tables 2.3 and 2.4, and Method 2 performs even worse than the the original algorithm in terms of bias of estimate and empirical coverage of CIs. This simulation result suggests that Fisher transformation does not help in improving the empirical coverage of confidence interval for ρ. And this can be seen from the histogram of the perturbed estimates from one of the simulation replicates with '4-neighbors' neighborhood matrix and $\rho_0 = 0.3$ with the original SGD (Figure 2.5). The probability for the perturbed estimates to go below 0.25 or above 0.35 is very low. In summary, the reason for low coverage of CI for ρ is not due to the range constraint of ρ.

Table 2.6 Effect of Fisher Transformation

Setting	parameter	mean	SD	coverage
1	β_1 (0.5)	0.500	0.005	0.948
	σ^2 (1.0)	1.001	0.005	0.960
	ρ (0.3)	0.299	0.005	0.898
2	β_1 (0.5)	0.500	0.007	0.948
	σ^2 (1.0)	1.001	0.006	0.960
	ρ (0.3)	0.295	0.004	0.668

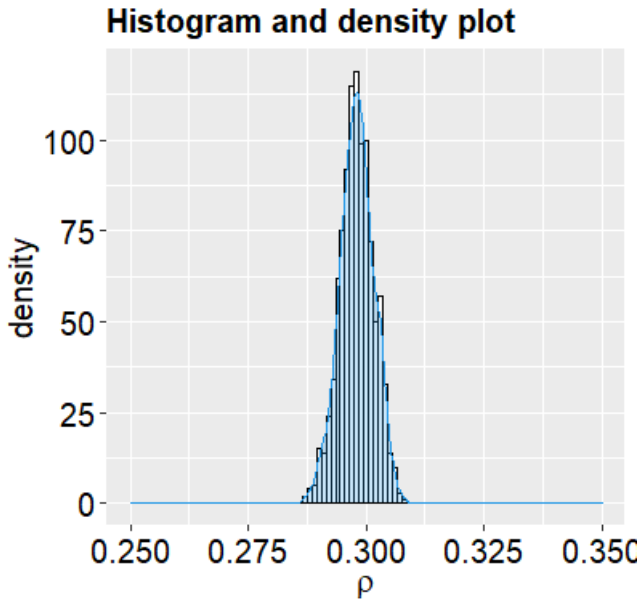


Figure 2.5 Histogram and density plot of perturbed estimate for ρ.

2.6.2 Increase ranges of confidence intervals

To further investigate the confidence interval construction algorithm, we plot out 25 of the constructed confidence intervals for ρ in Figure 2.6 obtained from a simulation study with '4-neighbors' neighborhood structure and $\rho_0 = 0.3$. Each red dot represents the SGD estimate for each run and each black vertical line the range of its corresponding confidence interval. The true value of ρ is indicated by the blue dash line. Confidence intervals that not covering the true values are boxed. Though these confidence intervals do not cover the true value, one of their ends is very close to the true value. They can cover the true value, if their ranges increase a little bit.

Also, we compare the standard deviation obtained from the perturbed estimates of each independent replicates, denoted as s_i, with the standard deviation of estimates from all of the independent replicates, the empirical standard deviation, denoted as s. The empirical standard deviation s can be treated as the true standard deviation of the estimates. To construct confidence interval at desired coverage level, we need s_i to be close to s. Figure 2.7 is the plot of s_i and s from a simulation

with '4-neighbors' neighborhood structure and $\rho_0 = 0.3$. The result shows that for β_1, most of the standard deviations obtained from the perturbed estimates of each independent replicate are higher than the standard deviation of the estimates. The relation is reversed for ρ. Table 2.7 lists the empirical coverage of confidence intervals constructed using s_i and s as $\hat{\theta} \pm Z_{\alpha/2} * sd$ ($\hat{\theta}$ the SGD estimate). For ρ, using s as the standard deviation increases the empirical coverage of CIs to the desired level. As a control, for β_1, using s as the standard deviation decreases the empirical coverage of CIs, but it is still around the desired level. Motivated by this result, we then study whether we can increase the range of the confidence intervals so that the empirical coverage is near the desired level.

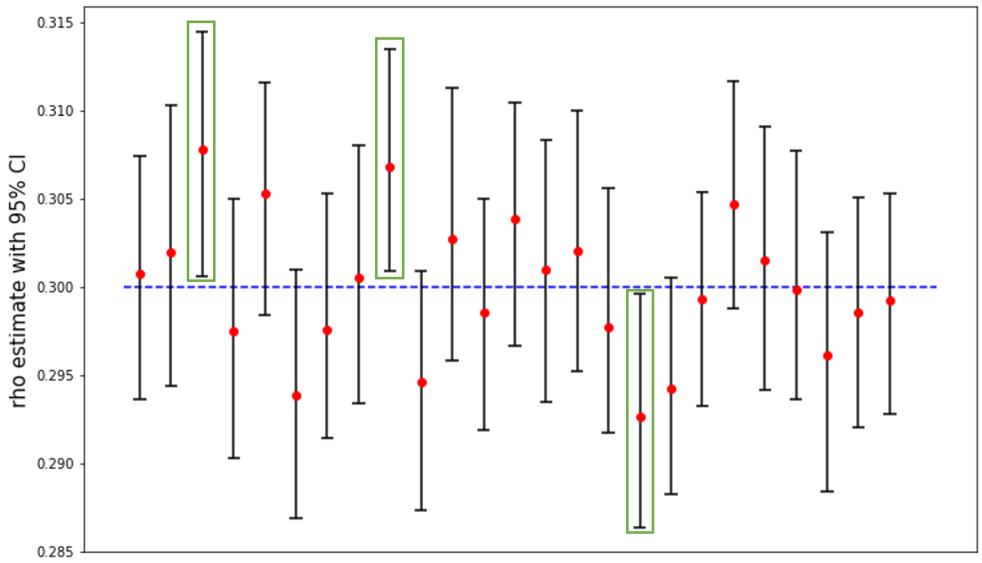


Figure 2.6 Plot of confidence interval of ρ.

Using different perturbation variables for β and ρ The range of confidence intervals for ρ are determined by the sample variance of $\bar{\rho}_n^{*,1} - \bar{\rho}_n, \bar{\rho}_n^{*,2} - \bar{\rho}_n, \ldots, \bar{\rho}_n^{*,B} - \bar{\rho}_n$. We can increase the sample variance to increase the range of confidence intervals. These perturbed estimates are generated by introducing the perturbation random variable $W_k^{*,b}$ as shown in (2.26). As suggested by [13] , $W_k^{*,b} \overset{iid}{\sim} W, W > 0, E(W) =$

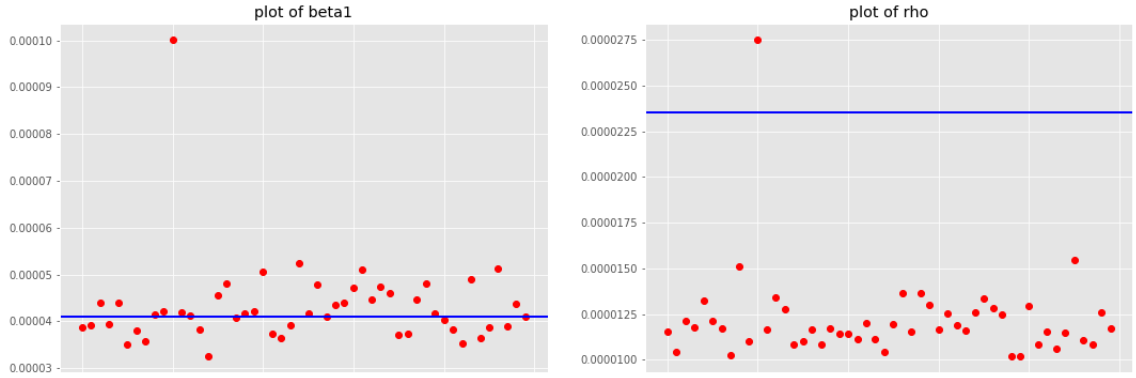


Figure 2.7 Comparison of empirical sd and sd from perturbed estimates for β_1 and ρ. Each red dot represents one sd from perturbed estimates of one independent replicate. The blue line represents the empirical sd.

Table 2.7 Comparison of Empirical Coverage of CIs with SD of Perturbed Estimates and Empirical SD

	β_1	ρ
s_i	0.956	0.852
s	0.946	0.948

$Var(W) = 1$. To increase the sample variance of $\bar{\rho}_n^{*,1} - \bar{\rho}_n, \bar{\rho}_n^{*,2} - \bar{\rho}_n, \ldots, \bar{\rho}_n^{*,B} - \bar{\rho}_n$, we can still keep the expectation of W to be 1 but increase the variance of W to be greater than 1. We try this idea by simulation studies with $Var(W)$ be several values greater than 1. We use a Gamma distribution instead of the $exp(1)$ distribution for W, choosing the shape parameter and scale parameter so that $E(W) = 1, Var(W) = a$ and $a > 1$. The results (not shown) suggest that larger $Var(W)$ applied leads to a larger confidence interval coverage not only for ρ but also for $\boldsymbol{\beta}$. Since confidence interval coverage for $\boldsymbol{\beta}$ is at the desired level, this large $Var(W)$ causes CI coverage for $\boldsymbol{\beta}$ to be beyond the desired level.

To solve the problem that empirical CI coverage for $\boldsymbol{\beta}$ also increases with $Var(W)$, we use different perturbed variable $W_k^{*,b}$ for $\boldsymbol{\beta}, \sigma^2$ and for ρ. This new

algorithm is shown in (2.36).

$$
\begin{aligned}
\hat{\boldsymbol{\beta}}_k^{*,b} &= \hat{\boldsymbol{\beta}}_{k-1}^{*,b} + \gamma_k W_k^{*,b} \nabla \ell_{\boldsymbol{\beta},k}(\hat{\boldsymbol{\beta}}_{k-1}^{*,b}, \hat{\rho}_{k-1}^{*,b}, \hat{\sigma^2}_{k-1}^{*,b}) \\
\hat{\eta}_k^{*,b} &= \hat{\eta}_{k-1}^{*,b} + \gamma_k \widetilde{W}_k^{*,b} \nabla \ell_{\rho,k}(\hat{\boldsymbol{\beta}}_{k-1}^{*,b}, \hat{\rho}_{k-1}^{*,b}, \hat{\sigma^2}_{k-1}^{*,b}) \cos \hat{\eta}_{k-1}^{*,b} \\
\hat{\phi}_k^{*,b} &= \hat{\phi}_{k-1}^{*,b} + \gamma_k W_k^{*,b} \nabla \ell_{\sigma_0^2,k}(\hat{\boldsymbol{\beta}}_{k-1}^{*,b}, \hat{\rho}_{k-1}^{*,b}, \hat{\sigma^2}_{k-1}^{*,b}) e^{\hat{\phi}_{k-1}^{*,b}} \\
\hat{\rho}_k^{*,b} &= \sin \hat{\eta}_k^{*,b}, \quad \hat{\sigma^2}_k^{*,b} = e^{\hat{\phi}_k^{*,b}} \\
\bar{\boldsymbol{\beta}}_k^{*,b} &= \frac{1}{k} \sum_{i=1}^{k} \hat{\boldsymbol{\beta}}_i^{*,b}, \quad \bar{\rho}_k^{*,b} = \frac{1}{k} \sum_{i=1}^{k} \hat{\rho}_i^{*,b}, \quad \bar{\hat{\sigma^2}}_k^{*,b} = \frac{1}{k} \sum_{i=1}^{k} \hat{\sigma^2}_i^{*,b}
\end{aligned}
\tag{2.36}
$$

Here, $W_k^{*,b} \overset{iid}{\sim} W, W > 0, E(W) = Var(W) = 1$, $\widetilde{W}_k^{*,b} \overset{iid}{\sim} \widetilde{W}, \widetilde{W} > 0, E(\widetilde{W}) = 1, Var(\widetilde{W}) \geq 1$. We study this with a simulation study, using an $exp(1)$ distribution for W, a Gamma distribution for $\widetilde{W}$ with W, $\widetilde{W}$ independent. The results (not shown) suggest that a larger $Var(\widetilde{W})$ leads to a larger confidence interval coverage not only for ρ but also for $\boldsymbol{\beta}$ though $Var(W)$ is 1. The reason for this could be that by using two perturbed parameter $(W_k^{*,b}, \widetilde{W}_k^{*,b})$ more variability is introduced, affecting not only the perturbed estimates of ρ but also those of $\boldsymbol{\beta}$. To reduce the variability, we still use different perturbed parameters for $\boldsymbol{\beta}$ and for ρ but with a new way to generate them. For recursive step k and perturbed estimate with index b, we first generate the perturbed parameter $W_k^{*,b}$ for $\boldsymbol{\beta}$ and σ^2. Then the perturbed parameter used for ρ, $\widetilde{W}_k^{*,b}$ is given by

$$
\widetilde{W}_k^{*,b} = C W_k^{*,b} - (C-1), C \geq 1. \tag{2.37}
$$

And like above, $W_k^{*,b} \overset{iid}{\sim} W, W > 0, E(W) = Var(W) = 1$ and it is easy to verify that $E(\widetilde{W}_k^{*,b}) = 1$, $Var(\widetilde{W}_k^{*,b}) = C^2$. Here, the correlation between $W_k^{*,b}$ and $\widetilde{W}_k^{*,b}$ is 1, thus, we reduced the extra variability caused by using two different perturbed parameters.

We use simulations to study this new algorithm. We try different C values, i.e., different variance for $\widetilde{W}_k^{*,b}$, with all three neighborhood structures and $\rho_0 = 0.3$.

For each neighborhood structure, we generate the data and apply CIs construction algorithm with several C values. Results shown in Table 2.8 are for the simulation with '4-neighbors' neighborhood structure. The results suggest that increasing C does increase the CI coverage of ρ but not that of $\boldsymbol{\beta}$. With a carefully tuned value of C, we can make the coverage of ρ close to the desired level. The C value required for making the coverage of ρ to be close to the desired level is different for different neighborhood structures. We find the best C value needed by hyperparameter tuning for this simulation study.

Table 2.8 CIs Coverage for Algorithm with Two Sets of Perturbed Parameters

C	CI coverage	
	β_1	ρ
1.0	0.941	0.856
1.15	0.941	0.906
1.2	0.942	0.922
1.25	0.942	0.940
1.3	0.942	0.952

In order to determine the value for C to use in a real application, we study the asymptotic distribution of the SGD estimates. For a given and fixed data point index, i, let $\nabla\ell_i$ be the contribution from the i-th data unit to the derivative of the likelihood, $\boldsymbol{\theta}_{0[q]}$ the true parameter value, $\boldsymbol{S}_{0,i} = -\mathbb{E}[\nabla^2\ell_i(\boldsymbol{\theta}_0)]$ and $\boldsymbol{V}_{0,i} = \mathbb{E}[\nabla\ell_i(\boldsymbol{\theta}_0)\nabla\ell_i(\boldsymbol{\theta}_0)^T]$. The i.i.d. $\nabla\ell_{i,1}, \nabla\ell_{i,2}, \cdots, \nabla\ell_{i,n}$ from independent datasets are used for parameter estimation with SGD. Based on **Theorem 1**, the variance of the estimate $\bar{\boldsymbol{\theta}}_n$ is $\boldsymbol{M}_i = \boldsymbol{S}_{0,i}^{-1}\boldsymbol{V}_{0,i}\boldsymbol{S}_{0,i}^{-1}/n$. The variance for $\bar{\rho}_n$ is $\boldsymbol{M}_{i,[q,q]}$. If $\nabla\ell_i$ is the derivative from a genuine likelihood, then $\boldsymbol{S}_{0,i} = \boldsymbol{V}_{0,i}$ and $\boldsymbol{S}_{0,i}^{-1}\boldsymbol{V}_{0,i}\boldsymbol{S}_{0,i}^{-1}/n = \boldsymbol{S}_{0,i}^{-1}/n$ (let $\boldsymbol{G}_i = \boldsymbol{S}_{0,i}^{-1}/n$). In this case, the variance for $\hat{\rho}$ is $\boldsymbol{S}_{0,i,[q,q]}^{-1}$. In Appendix C we derive the explicit

expressions of $\boldsymbol{S}_{0,i}$ and $\boldsymbol{V}_{0,i}$, and also show that all elements $\boldsymbol{S}_{0,i}$ and $\boldsymbol{V}_{0,i}$ are equal except that $\boldsymbol{S}_{0,,i,[q,q]} \neq \boldsymbol{V}_{0,i,[q,q]}$.

To further investigate the difference between $\boldsymbol{S}_{0,,i,[q,q]}$ and $\boldsymbol{V}_{0,i,[q,q]}$, we generate simulation data with '4-neighbors' neighborhood structure and $\rho_0 = 0.3$. For each sample with $N = 90,000$ data points, we calculate $\boldsymbol{M}_1, \boldsymbol{M}_2, \cdots, \boldsymbol{M}_N$ and $\boldsymbol{G}_1, \boldsymbol{G}_2, \cdots, \boldsymbol{G}_N$ [4]. Then we calculate $\bar{\boldsymbol{M}} = \frac{1}{N}\sum_i \boldsymbol{M}_i$ and $\bar{\boldsymbol{G}} = \frac{1}{N}\sum_i \boldsymbol{G}_i$. The ratio of $\bar{\boldsymbol{G}}_{[q,q]}/\bar{\boldsymbol{M}}_{[q,q]}$ is about 2. And $\sqrt{2}$ is close to the C value needed in Table 2.8 to make the empirical coverage of CIs of ρ to close to the nominal level. Hence, we believe that $\sqrt{\bar{\boldsymbol{G}}_{[q,q]}/\bar{\boldsymbol{M}}_{[q,q]}}$ could be used the value of C in Equation (2.37) to improve the empirical coverage of confidence intervals. This will be part of future work.

Using randomized learning rates For simulation studies above, we use a decaying learning rate in the format of $\gamma_k = \gamma_1 k^{-\alpha}, k = 1, 2, \cdots, n, \gamma_1 > 0, \alpha \in (0.5, 1)$. A decaying learning rate helps to guarantee that the SGD increments will converge to zero and convergence of the SGD estimate [41]. We try two methods to modify the learning rate to increase the coverage of confidence intervals. The first method is to shuffle all the learning rates used, $\gamma_1, \gamma_1 \cdot 2^{-\alpha}, \cdots, \gamma_1 \cdot n^{-\alpha}$ and then use them according to the order after shuffling. In this way the learning rate is not necessarily decreasing and this could cause the final SGD estimate to fluctuate around a point but not converge to it. This fluctuation can increase the sample variance of perturbed estimates and therefore increase the coverage of confidence intervals. The other method we try is to use a constant learning rate for all recursive steps, $\gamma_k \equiv \gamma_1, \gamma_1 > 0, k = 1, 2, \cdots, n$.

We use simulations to study these two methods. For this simulation, we use the '4-neighbors' neighborhood structure and $\rho_0 = 0.3$. For the first method, we

[4] We note that for a given data index i, $\boldsymbol{M}_{i,[q,q]}$ and $\boldsymbol{G}_{i,[q,q]}$ do not depend on random samples generated

use the same randomized order of learning rates for all simulation runs as well as different randomized order of learning rates for different simulation runs. Simulation results (summarized in Table 2.9) show that means of the SGD estimates from both methods are close to the true value. Only the method of using randomized learning rates increases the standard error of SGD estimates comparing with the algorithm using decaying learning rate (for example see setting 1 in Table 2.5). Also, the coverage of CI for both $\boldsymbol{\beta}$ and ρ from the first method are around 0.95. For the second method, the coverage of CI does not improve compared to using a decaying learning rate. In summary, though using randomized learning rate can improve CI coverage, it increases variance of the final SGD estimate. Also, proof of convergence for this method could be challenging. This will be a focus for future work.

Table 2.9 Effect of Learning Rates

learning rate	parameter	mean	SD	coverage
randomized	β_1 (0.5)	0.501	0.027	0.976
(same for all runs)	ρ (0.3)	0.298	0.015	0.960
randomized	β_1 (0.5)	0.501	0.028	0.974
(diff. for each run)	ρ (0.3)	0.297	0.016	0.944
constant	β_1 (0.5)	0.500	0.006	0.958
	ρ (0.3)	0.299	0.005	0.822

2.7 SAR Model with Autoregressive Disturbance

Besides SAR model (2.1) described above, another spatial model is spatial model with autoregressive disturbance, also called the spatial error model. This model is useful when spatial correlation exists but adding the spatial lag term as in Equation (2.1)

does not provide significant improvement [11]. The model is given as:

$$\begin{aligned} \boldsymbol{y} &= \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{U} \\ \boldsymbol{U} &= \rho \boldsymbol{W}\boldsymbol{U} + \boldsymbol{\epsilon}, \end{aligned} \tag{2.38}$$

The meaning of these notations are the same as described for the SAR model. As before, we only consider the case that $\boldsymbol{\epsilon} \sim N(\boldsymbol{0}, \sigma^2 \boldsymbol{I})$.

We develop the SGD procedure for the spatial error model based on MLE. First, we work out the log-likelihood for this model. Given Equation (2.38) we have:

$$\boldsymbol{\epsilon} = \boldsymbol{A}(\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta}). \tag{2.39}$$

Here, $\boldsymbol{A} = \boldsymbol{I} - \rho \boldsymbol{W}$. Then the likelihood is:

$$\begin{aligned} L(\boldsymbol{\theta}|\boldsymbol{y}) = L(\boldsymbol{\theta}|\boldsymbol{\epsilon})|\frac{d\boldsymbol{\epsilon}}{d\boldsymbol{y}}| &= (2\pi\sigma^2)^{-n/2} \exp(-\frac{\boldsymbol{\epsilon}^T\boldsymbol{\epsilon}}{2\sigma^2})|\boldsymbol{A}| \\ &= (2\pi\sigma^2)^{-n/2} \exp(-\frac{(\boldsymbol{A}\boldsymbol{y} - \boldsymbol{A}\boldsymbol{X}\boldsymbol{\beta})^T(\boldsymbol{A}\boldsymbol{y} - \boldsymbol{A}\boldsymbol{X}\boldsymbol{\beta})}{2\sigma^2})|\boldsymbol{A}|. \end{aligned} \tag{2.40}$$

The log-likelihood (omitting constant) is:

$$\ell(\boldsymbol{\theta}|\boldsymbol{y}) = -\frac{\ln(\sigma^2)}{2}n - \frac{(\boldsymbol{A}\boldsymbol{y} - \boldsymbol{A}\boldsymbol{X}\boldsymbol{\beta})^T(\boldsymbol{A}\boldsymbol{y} - \boldsymbol{A}\boldsymbol{X}\boldsymbol{\beta})}{2\sigma^2} + \ln|\boldsymbol{A}|. \tag{2.41}$$

Clearly, due to the existence of the term $\ln|\boldsymbol{A}|$, this log-likelihood can not be written as summation of contribution from each individual data point. We follow the procedure for the SAR model, by writing the derivative as the contribution from each individual data unit.

- **for $\boldsymbol{\beta}$**

$$\begin{aligned} \nabla l(\boldsymbol{\beta}) &= \frac{1}{\sigma^2}(\boldsymbol{X}^T\boldsymbol{A}^T\boldsymbol{A}\boldsymbol{y} - \boldsymbol{X}^T\boldsymbol{A}^T\boldsymbol{X}\boldsymbol{\beta}) \\ &= \sum_i \frac{1}{\sigma^2}(\boldsymbol{x}_i - \rho\bar{\boldsymbol{x}}_i)(y_i - \rho\bar{y}_i - (\boldsymbol{x}_i - \rho\bar{\boldsymbol{x}}_i)^T\boldsymbol{\beta}) \\ &= \sum_i \nabla l_i(\boldsymbol{\beta}). \end{aligned} \tag{2.42}$$

Here, $\boldsymbol{x}_i, y_i$ are the i-th data point, $\bar{\boldsymbol{x}}_i = \boldsymbol{X}^T \boldsymbol{W}_i^T$, $\boldsymbol{W}_i$ is the i-th row of $\boldsymbol{W}$; i.e., $\bar{\boldsymbol{x}}_i$ is the mean $\boldsymbol{x}$ of neighbors of i-th data point.

- **for** σ^2

$$\begin{aligned}\nabla l(\sigma^2) &= -\frac{n}{2\sigma^2} + \frac{1}{2(\sigma^2)^2}(\boldsymbol{A}\boldsymbol{y} - \boldsymbol{A}\boldsymbol{X}\boldsymbol{\beta})^T(\boldsymbol{A}\boldsymbol{y} - \boldsymbol{A}\boldsymbol{X}\boldsymbol{\beta}) \\ &= \sum_i -\frac{1}{2\sigma^2} + \frac{1}{2(\sigma^2)^2}(y_i - \rho\bar{y}_i - (\boldsymbol{x}_i - \rho\bar{\boldsymbol{x}}_i)^T\boldsymbol{\beta})^2 \\ &= \sum_i \nabla l_i(\rho).\end{aligned} \tag{2.43}$$

- **for** ρ

$$\begin{aligned}\nabla l(\rho) &= -tr(\boldsymbol{A}^{-1}\boldsymbol{W}) + \frac{(\boldsymbol{A}\boldsymbol{y} - \boldsymbol{A}\boldsymbol{X}\boldsymbol{\beta})^T\boldsymbol{W}(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})}{\sigma^2} \\ &= \sum_i -\frac{1}{\rho}(\boldsymbol{A}_{ii}^{-1} - 1) + \frac{1}{\sigma^2}(y_i - \rho\bar{y}_i - (\boldsymbol{x}_i - \rho\bar{\boldsymbol{x}}_i)^T\boldsymbol{\beta})(\bar{y}_i - \bar{\boldsymbol{x}}_i\boldsymbol{\beta}) \\ &= \sum_i \nabla l_i(\rho).\end{aligned} \tag{2.44}$$

Also, to incorporate the constrain that $\rho \in (-1, 1)$, we let $\rho = \sin\eta$, and incorporate the constrain $\sigma^2 = e^{\phi}$. We follow exactly as in Equations (2.26) and (2.27) for updating SGD estimates and perturbed estimates for CIs construction.

We then study the performance of the above algorithm with simulations. The simulation setting is the same as described in Section 2.5.1, except that here for spatial error model, the response variable $\boldsymbol{y}$ is generated as $\boldsymbol{X}\boldsymbol{\beta}_0 + \boldsymbol{A}_0^{-1}\boldsymbol{\epsilon}$. We tried all three neighbor structures described above. The summary of the estimates and empirical coverages of CIs are shown in Table 2.10 [5]. Clearly for all three neighborhood structures studied, the proposed algorithm can provide correct estimates and the empirical coverages of CIs are at the nominal level for β_1 and σ^2. However, the empirical coverages for CIs of ρ are lower than the nominal level. Also, like the SAR

[5] As in SAR model, we only show the results for β_1, and that for β_2 are very similar to β_1

model, when the spatial correlation is high (24-neighbors), the performance is worse. Future work is needed to find ways to improve the performance of confidence intervals.

Table 2.10 Simulation Results for the Spatial Error Model

neigh struc	parameter	mean	SD	coverage
4-neighbors	β_1 (0.5)	0.500	0.006	0.932
	σ^2 (1.0)	1.00	0.006	0.936
	ρ (0.3)	0.300	0.005	0.824
8-neighbors	β_1 (0.5)	0.500	0.006	0.968
	σ^2 (1.0)	1.000	0.005	0.952
	ρ (0.3)	0.300	0.007	0.824
24-neighbors	β_1 (0.5)	0.500	0.007	0.944
	σ^2 (1.0)	1.000	0.006	0.924
	ρ (0.3)	0.308	0.023	0.724

2.8 SGD Based on Two-Stage Least Square

For the SGD algorithm discussed above, the correlation between data points used in each iterative step could be the reason that the empirical coverage of CI for ρ is not at the nominal level. As discussed in Section 2.1.2, besides MLE, two-stage least square (2SLS) is also a unbiased estimation method. Also, with the help of instrumental variables, the regressors are exogenous in both stages. We can develop the SGD algorithm based on two-stage least squares. We apply SGD on both stages and generate perturbed estimates on on second stage to construct CIs. We illustrate the detailed algorithm below.

For the first stage, we fit the model $\boldsymbol{W}\boldsymbol{y} = \boldsymbol{X}\tilde{\boldsymbol{\beta}} + \boldsymbol{Z}\tilde{\boldsymbol{\zeta}}$ using SGD for a giving initial value $\hat{\tilde{\boldsymbol{\beta}}}_0$ and $\hat{\tilde{\boldsymbol{\zeta}}}_0$ as:

$$\begin{aligned}\hat{\tilde{\boldsymbol{\beta}}}_k &= \hat{\tilde{\boldsymbol{\beta}}}_{k-1} + \gamma_k(\bar{y}_k - \boldsymbol{x}_k^T\hat{\tilde{\boldsymbol{\beta}}}_{k-1} - \boldsymbol{z}_k^T\hat{\tilde{\boldsymbol{\zeta}}}_{k-1})\boldsymbol{x}_k \\ \hat{\tilde{\boldsymbol{\zeta}}}_k &= \hat{\tilde{\boldsymbol{\zeta}}}_{k-1} + \gamma_k(\bar{y}_k - \boldsymbol{x}_k^T\hat{\tilde{\boldsymbol{\beta}}}_{k-1} - \boldsymbol{z}_k^T\hat{\tilde{\boldsymbol{\zeta}}}_{k-1})\boldsymbol{z}_k\end{aligned} \tag{2.45}$$

The final estimate of $\tilde{\boldsymbol{\beta}}$ and $\tilde{\boldsymbol{\zeta}}$ are given by $\hat{\tilde{\boldsymbol{\beta}}}_n = 1/n\sum_i^n \hat{\tilde{\boldsymbol{\beta}}}_i$ and $\hat{\tilde{\boldsymbol{\zeta}}}_n = 1/n\sum_i^n \hat{\tilde{\boldsymbol{\zeta}}}_i$. We then calculate the fitted value of $\boldsymbol{W}\boldsymbol{y}$ as:

$$\widehat{\boldsymbol{W}\boldsymbol{y}} = \boldsymbol{X}\hat{\tilde{\boldsymbol{\beta}}}_n + \boldsymbol{Z}\hat{\tilde{\boldsymbol{\zeta}}}_n \tag{2.46}$$

For the second stage, we fit the model $\boldsymbol{y} = \rho\widehat{\boldsymbol{W}\boldsymbol{y}} + \boldsymbol{X}\boldsymbol{\beta}$ with SGD based parameter estimation and generate perturbed estimates for CI constructions. Given initial value for $\boldsymbol{\beta}$ and ρ of $\hat{\boldsymbol{\beta}}_0$ and $\hat{\rho}_0$, respectively. Also, let $\rho = \sin\eta$ and $\hat{\eta}_0 = \arcsin\hat{\rho}_0$. The SGD estimates and perturbed estimates are updated as follows:

$$\begin{aligned}\hat{\boldsymbol{\beta}}_k &= \hat{\boldsymbol{\beta}}_{k-1} + \gamma_k(y_k - \hat{\rho}_{k-1}\hat{\tilde{y}}_k - \hat{\boldsymbol{\beta}}_{k-1}^T\boldsymbol{x}_k)\boldsymbol{x}_k \\ \hat{\eta}_k &= \hat{\eta}_{k-1} + \gamma_k(y_k - \hat{\rho}_{k-1}\hat{\tilde{y}}_k - \hat{\boldsymbol{\beta}}_{k-1}^T\boldsymbol{x}_k)\hat{\tilde{y}}_k\cos\hat{\eta}_{k-1} \\ \hat{\rho}_k &= \sin\hat{\eta}_k \\ \hat{\boldsymbol{\beta}}_k^* &= \hat{\boldsymbol{\beta}}_{k-1}^* + \gamma_k(y_k - \hat{\rho}_{k-1}^*\hat{\tilde{y}}_k - \hat{\boldsymbol{\beta}}_{k-1}^{*T}\boldsymbol{x}_k)\boldsymbol{x}_k \\ \hat{\eta}_k^* &= \hat{\eta}_{k-1}^* + \gamma_k(y_k - \hat{\rho}_{k-1}^*\hat{\tilde{y}}_k - \hat{\boldsymbol{\beta}}_{k-1}^{*T}\boldsymbol{x}_k)\hat{\tilde{y}}_k\cos\hat{\eta}_{k-1}^* \\ \hat{\rho}_k^* &= \sin\hat{\eta}_k^*.\end{aligned} \tag{2.47}$$

Here, $\hat{\tilde{y}}_k$ is the k-th element of $\widehat{\boldsymbol{W}\boldsymbol{y}}$. The CIs are constructed as described in Section 1.1.2. Please note that for this SGD estimation method, we do not estimate σ^2. As mentioned in Section 2.3.3, σ^2 is usually not of great interest. Here, we do not need to know the estimate of σ^2 to estimate $\boldsymbol{\beta}$ and ρ. Thus, it is not estimated for this algorithm.

We study the final sample propriety of this SGD estimate and CIs with simulations. Samples are generated the same as described in Section 2.5.1 for

'4-neighbor' neighborhood structure and $\rho_0 = 0.3$. The estimates are all close to true values (Table 2.11) and the empirical coverage of CIs are all at the nominal levels. Compared with the result in Table 2.3, the empirical SD of ρ estimates is higher than that of MLE based SGD. This is consistent with that MLE is more efficient than 2SLS [33].

Table 2.11 Simulation Results for SGD Based on Two-Stage Least Square

parameter	mean	SD	coverage
$\beta_1(0.5)$	0.499	0.006	0.950
$\beta_2(0.5)$	-0.499	0.006	0.960
ρ (0.3)	0.308	0.019	0.965

2.9 Summary and Discussion

This chapter develops the SGD-based parameter estimation and confidence intervals construction algorithm for the SAR mean regression model. The algorithm is developed based on the MLE and modification on the algorithm are made to accommodate the correlation between data points in the SAR model. The asymptotic properties of the estimates and perturbed estimates are studied. We then use simulations to study the performance of the proposed algorithms. The estimates are unbiased for various spatial parameter values, neighborhood structures, and orders of data used. The empirical coverages of CIs constructed for $\boldsymbol{\beta}$ are at the nominal level, while those for ρ are below the nominal level. We propose two ways to improve the empirical coverage for ρ. One is to use different perturbation variables for $\boldsymbol{\beta}$ and for ρ. The other is to use randomized learning rates instead of decaying learning ones. Both methods can increase the empirical coverage of CIs for ρ.

As discussed in the Section 2.4, the setting we use to develop the theoretical properties is different from the algorithm developed for the SAR model. To generalize

from the setting discussed for theoretical study to the SGD procedure used for the SAR model, we can make further assumptions about the neighborhood structure. Assume we have a very large grid and we are only looking at a proportion of the data (N data points). Also, assume all the data points have the same neighborhood structure. For example, one scenario is that each data point has two neighbors, the data points left and right to itself. Under this setting, expectation of derivative of log-likelihood from each individual data should equal to $1/N$ of expectation of the overall log-likelihood. Since these two derivatives only differ by a constant, we might get the convergence of the regular SGD for SAR from the convergence of the SGD in above setting. However, there are still difference between this setting and the SGD procedure we used. For this setting, data used in each iterative step are independent while in our SGD procedure, data in each iterative step are not. And this correlation could be responsible for the lower empirical coverage of the CIs constructed based on perturbed estimates. Further investigation on this is needed.

Simulation results show that empirical coverage of confidence interval for ρ is below the desired level. Analysis in Section 2.6.2 suggests that the width of the confidence interval might be responsible for this low empirical coverage. Two methods are proposed to increase the width of the confidence intervals. For the one using randomized learning rates, the empirical standard deviation of the estimates also increase. Clearly this method is less efficient compared to using decaying learning rates. Also, it is challenging for theoretical justification of both methods. Further investigations on improving the coverage of CIs are required.

We also studied parameter estimation and CIs construction with 2SLS based SGD. For the simulation setting studied, the estimates are all close to true value and empirical coverages of CIs are all at nominal levels. Compared with MLE based SGD algorithm, the disadvantage of this method is that it uses each data point twice. This means we cannot discard the data point after the first use. Also, for estimating ρ, this

method is less efficient. More simulations for various neighborhood structures and various ρ_0 values are needed to further investigate the robustness of this algorithm. Also, we need to study this asymptotic properties of this algorithm.

CHAPTER 3

ESTIMATION AND INFERENCE FOR THE SAR QUANTILE REGRESSION MODEL USING SGD

3.1 Introduction

Quantile regression was first introduced by Koenker and Basset (1978) [25], as an extension of the median regression. Compared with mean regression, quantile regression provides more detailed information about the model distribution, is more robust to outliers, and less restrictive on error distributions [10]. The linear quantile regression is defined as below:

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta}_\tau + \boldsymbol{\epsilon}. \tag{3.1}$$

Here, $\tau \in (0, 1)$ is the quantile of interest, $\boldsymbol{y}$ a $N \times 1$ vector for response variable, $\boldsymbol{X}$ the $N \times p$ regressor matrix. $\boldsymbol{\beta}_\tau$ is the unknown parameter vector for quantile τ and $\boldsymbol{\epsilon}$ is the error. Unlike the classic mean regression model Equation (1.3), there are no restrictions on the error term in quantile regression. Also, the coefficient $\boldsymbol{\beta}_\tau$ can depend on quantile τ. This provides more flexibility on the regression model. The parameter $\boldsymbol{\beta}_\tau$ can be estimated by a minimization problem defined as:

$$\boldsymbol{\beta}_\tau = \underset{\boldsymbol{\beta}}{\operatorname{argmin}} \sum_i \lambda_\tau(y_i - \boldsymbol{\beta}^T \boldsymbol{x}_i). \tag{3.2}$$

Here, $y_i, \boldsymbol{x}_i$ are for the i-th data point, $\lambda_\tau(u) = u(\tau - I(u < 0))$, often referred to as the check function[1], and $I(\cdot)$ the indicator function. This minimization problem can be solved by linear programming (see [4] for more details).

Spatial autoregressive quantile (SARQ) regression, like spatial autoregressive mean regression, allows the spatial correlation between data points. There are two

[1]Check function is usually denoted as ρ in literature. Here, we use λ since ρ is reserved for spatial parameter.

definitions of spatial quantile regression [28]. Here, we consider the one proposed by Kostov in 2009 [27], defined as follows:

$$\boldsymbol{y} = \rho_\tau \boldsymbol{W}\boldsymbol{y} + \boldsymbol{X}\boldsymbol{\beta}_\tau + \boldsymbol{\epsilon}. \tag{3.3}$$

Here, $\boldsymbol{y}$ is the $N \times 1$ vector of response, $\boldsymbol{W}$ the $N \times N$ neighborhood matrix (see Section 2.1 for more details), $\boldsymbol{X}$ the $N \times p$ regressor matrix, ρ_τ, $\boldsymbol{\beta}_\tau$ unknown parameters, and $\boldsymbol{\epsilon}$ the error. The quantity $\rho_\tau \boldsymbol{W}\boldsymbol{Y}$ is the spatial lag term. Similar to the SAR mean regression model, SARQ model assumes that the response variable at a certain location depends not only on its covariates, also the value of the response variable of its neighboring data points. Unlike the SAR mean model, in the SARQ model, coefficients $\boldsymbol{\beta}_\tau$ and spatial parameter ρ_τ can depend on the quantile τ. This allows for varying degree of spatial dependence at different quantiles of the response distribution, for example, spatial dependence may exist in some portion of the distribution but not elsewhere, with constant $\boldsymbol{\beta}$ and ρ as a special case.

Several methods have been proposed for estimating the model parameters. One way is to treat the endogenous term $\boldsymbol{W}\boldsymbol{Y}$ as an exogenous term like $\boldsymbol{X}$. Then the parameters are estimated via:

$$(\boldsymbol{\beta}_\tau, \rho_\tau) = \underset{(\boldsymbol{\beta},\rho)}{\operatorname{argmin}} \sum_i \lambda_\tau(y_i - \rho\bar{y}_i - \boldsymbol{\beta}^T \boldsymbol{x}_i). \tag{3.4}$$

Here y_i and $\boldsymbol{x}_i$ are from the i-th data point, and $\bar{y}_i$ is the i-th element of $\boldsymbol{W}\boldsymbol{y}$, which is the weighted average of the response variables for the neighbors of the i-th data point. Like for SAR mean regression, this estimator is biased in general case as it ignores the endogeneity of $\boldsymbol{W}\boldsymbol{y}$. We refer this estimator as the one-stage quantile regression (1SQR) estimator.

Kostov proposed the instrumental quantile regression estimation to deal with the endogeneity issue [27]. This can be implemented in two ways. Kim and Muller proposed the two stage quantile regression estimator with the first stage based on

quantile regression [24]. For parameters estimated in this method, first instrumental variables $\boldsymbol{Z}_{[N\times p']}$ are generated. $\boldsymbol{Z}$ is usually set as $\boldsymbol{WX}, \boldsymbol{W^2X}, \cdots$. In the first stage, a quantile regression is fitted:

$$\boldsymbol{WY} = \boldsymbol{X}\tilde{\boldsymbol{\beta}} + \boldsymbol{Z}\tilde{\boldsymbol{\zeta}}. \tag{3.5}$$

The fitted value of $\boldsymbol{WY}$, $\widehat{\boldsymbol{WY}} = \boldsymbol{X}\hat{\tilde{\boldsymbol{\beta}}} + \boldsymbol{Z}\hat{\tilde{\boldsymbol{\zeta}}}$, is then used for the second stage quantile regression:

$$\boldsymbol{Y} = \rho\widehat{\boldsymbol{WY}} + \boldsymbol{X}\boldsymbol{\beta}. \tag{3.6}$$

We refer to this method as 2SQR.

Another way to implement the instrumental quantile regression estimation is proposed by Chernozhukov and Hansen (referred as **CH** below) [6]. This estimation method also uses the instrumental variables $\boldsymbol{Z}$. First $\hat{\boldsymbol{\beta}}(\rho, \tau)$ and $\hat{\boldsymbol{\zeta}}(\rho, \tau)$ depending on ρ are estimated using the model:

$$\boldsymbol{Y} = \rho\boldsymbol{WY} + \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{Z}\boldsymbol{\zeta}. \tag{3.7}$$

Then ρ is estimated by:

$$\hat{\rho}_\tau = \underset{\rho}{\operatorname{argmin}} \sqrt{\hat{\boldsymbol{\zeta}}(\rho, \tau)^T \hat{\boldsymbol{\zeta}}(\rho, \tau)}. \tag{3.8}$$

Then $\hat{\rho}_\tau$ is plugged back to $\hat{\boldsymbol{\beta}}(\rho, \tau)$ to get the estimate of $\boldsymbol{\beta}_\tau$. In practice, this methods can be applied by defining a grid of values for ρ, $\{\ \rho_1, \rho_2, \cdots, \rho_K\}$, and selecting the value ρ_k that minimizes $\sqrt{\hat{\boldsymbol{\zeta}}(\rho, \tau)^T \hat{\boldsymbol{\zeta}}(\rho, \tau)}$ for a given quantile τ. Then we can get the estimate for $\boldsymbol{\beta}$ by plugging ρ_k into $\hat{\boldsymbol{\beta}}(\rho, \tau)$. A drawback of this method is that it requires a predefined grid of ρ values. This can be difficult in some scenarios. Also, as described above, this method requires fitting multiple models which limits the scalability of the method. Hence, we do not use this method with the SGD algorithm (See more discussion on this in Section 3.4).

Hense, we apply SGD for parameter estimation and online bootstrapping for CIs construction for SAR quantile regression, focusing on 1SQR and 2SQR. This chapter is organized as follows. In Section 3.2, we introduce the SGD estimator based on 1SQR and then propose the SGD-based estimator based on 2SQR. Section 3.3 studies the finite sample properties of these SGD estimates with a simulation study. Section 3.4 summarizes this chapter and provides directions for future study.

3.2 SGD on SAR Quantile Regression

This section develops the procedure of parameter estimation with SGD and CIs construction with online bootstrapping for the SAR quantile regression model.

3.2.1 One-stage quantile regression

We first work out the SGD-based parameter estimation and confidence interval construction procedure according to 1SQR. For a quantile τ of interest, let $\ell_i = \lambda_\tau(y_i - \rho\bar{y}_i - \boldsymbol{\beta}^T\boldsymbol{x}_i)$ and Equation (3.4) can be written as:

$$(\boldsymbol{\beta}_\tau, \rho_\tau) = \underset{(\boldsymbol{\beta},\rho)}{\operatorname{argmin}} \sum_i \ell_i. \tag{3.9}$$

Following Example 4 in [13], we first get the derivative of ℓ_i:

$$\begin{aligned} \nabla\ell_{\boldsymbol{\beta},i} &= -\boldsymbol{x}_i(\tau - I(y_i - \rho\bar{y}_i - \boldsymbol{\beta}^T\boldsymbol{x}_i)) \\ \nabla\ell_{\rho,i} &= -\bar{y}_i(\tau - I(y_i - \rho\bar{y}_i - \boldsymbol{\beta}^T\boldsymbol{x}_i)). \end{aligned} \tag{3.10}$$

Like with the SAR mean regression, we set $\rho = \sin\eta$ to accommodate the restriction that $-1 < \rho < 1$ and accordingly $\nabla\ell_{\eta,i} = \nabla\ell_{\rho,i}\cos\eta$. Given the starting values $\hat{\boldsymbol{\beta}}_0$ and $\hat{\rho}_0$, the starting value for η is calculated as $\hat{\eta}_0 = \arcsin\hat{\rho}_0$. The SGD estimates $\hat{\boldsymbol{\beta}}_k$ and $\hat{\rho}_k$ are updated as:

$$\begin{aligned} \hat{\boldsymbol{\beta}}_k = \hat{\boldsymbol{\beta}}_{k-1} - \gamma_k\nabla\ell_{\boldsymbol{\beta},k}, \;\; \hat{\eta}_k = \hat{\eta}_{k-1} - \gamma_k\nabla\ell_{\eta,k} \\ \hat{\rho}_k = \sin\hat{\eta}_k, \quad \bar{\boldsymbol{\beta}}_k = \frac{1}{k}\sum_{i=1}^k \hat{\boldsymbol{\beta}}_i, \;\; \bar{\rho}_k = \frac{1}{k}\sum_{i=1}^k \hat{\rho}_i. \end{aligned} \tag{3.11}$$

For confidence interval construction, the perturbed estimates are updated by:

$$\begin{aligned} \hat{\boldsymbol{\beta}^*}_k &= \hat{\boldsymbol{\beta}^*}_{k-1} - \gamma_k W_k \nabla \ell_{\boldsymbol{\beta},k} \quad \hat{\eta^*}_k = \hat{\eta^*}_{k-1} - \gamma_k W_k \nabla \ell_{\eta^*,k} \\ \hat{\rho^*}_k &= \sin \hat{\eta^*}_k, \quad \bar{\boldsymbol{\beta}^*}_k = \frac{1}{k}\sum_{i=1}^{k} \hat{\boldsymbol{\beta}^*}_i, \quad \bar{\rho^*}_k = \frac{1}{k}\sum_{i=1}^{k} \hat{\rho^*}_i, \end{aligned} \tag{3.12}$$

with γ_k and W_k in Equations (3.11) and 3.12 are defined the same as described in Section 1.1.2.

3.2.2 Two-stage quantile regression

For two-stage quantile regression (2SQR), we apply SGD on both stages of the quantile regression to get the estimates. To construct confidence intervals, we generate perturbed estimates during the second stage. The detailed procedure is described as below.

(1) First use SGD to get the estimate for regression in (3.5). The updating procedure is given by:

$$\begin{aligned} \hat{\bar{\boldsymbol{\beta}}}_k &= \hat{\bar{\boldsymbol{\beta}}}_{k-1} + \gamma_k \boldsymbol{x}_k [\tau - I(\bar{y}_i - \hat{\bar{\boldsymbol{\beta}}}_{k-1}^T \boldsymbol{x}_k - \hat{\bar{\boldsymbol{\zeta}}}_{k-1}^T \boldsymbol{z}_k)] \\ \hat{\bar{\boldsymbol{\zeta}}}_k &= \hat{\bar{\boldsymbol{\zeta}}}_{k-1} + \gamma_k \boldsymbol{z}_k [\tau - I(\bar{y}_i - \hat{\bar{\boldsymbol{\beta}}}_{k-1}^T \boldsymbol{x}_k - \hat{\bar{\boldsymbol{\zeta}}}_{k-1}^T \boldsymbol{z}_k)]. \end{aligned} \tag{3.13}$$

Here, $\bar{y}_k$ is the k-th element of $\boldsymbol{WY}$ and $\boldsymbol{x}_k, \boldsymbol{z}_k$ are the k-th row of $\boldsymbol{X}, \boldsymbol{Z}$ respectively.

(2) Calculate the predicted value of $\boldsymbol{WY}$ as follows:

$$\begin{aligned} \bar{\hat{\boldsymbol{\beta}}}_N = \frac{1}{N}\sum_{i=1}^{N} \hat{\bar{\boldsymbol{\beta}}}_i, \quad \bar{\hat{\boldsymbol{\zeta}}}_N = \frac{1}{N}\sum_{i=1}^{N} \hat{\bar{\boldsymbol{\zeta}}}_i \\ \widehat{\boldsymbol{WY}} = \boldsymbol{X}\bar{\hat{\boldsymbol{\beta}}}_N + \boldsymbol{Z}\bar{\hat{\boldsymbol{\zeta}}}_N. \end{aligned} \tag{3.14}$$

(3) Then apply SGD to generate the estimates and perturbed estimates based on Equation (3.6). We use $\eta = \arcsin \rho$ and the SGD estimates can be updated as:

$$\begin{aligned} \hat{\boldsymbol{\beta}}_k &= \hat{\boldsymbol{\beta}}_{k-1} + \gamma_k \boldsymbol{x}_i (\tau - I(y_k - \rho \hat{\bar{y}}_k - \hat{\boldsymbol{\beta}}_{k-1}^T \boldsymbol{x}_k)) \\ \hat{\eta}_k &= \hat{\eta}_{k-1} + \gamma_k \hat{\bar{y}}_k (\tau - I(y_k - \rho \hat{\bar{y}}_k - \hat{\boldsymbol{\beta}}_{k-1}^T \boldsymbol{x}_k)) \cos(\hat{\eta}_{k-1}) \\ \hat{\rho}_k &= \sin(\hat{\eta}_k). \end{aligned} \tag{3.15}$$

The perturbed estimates are updated as:

$$\begin{aligned}
\hat{\boldsymbol{\beta}}_k^* &= \hat{\boldsymbol{\beta}}_{k-1}^* + \gamma_k W_k \boldsymbol{x}_k(\tau - I(y_k - \rho\hat{\bar{y}}_k - \hat{\boldsymbol{\beta}}_{k-1}^{*T}\boldsymbol{x}_k)) \\
\hat{\eta}_k^* &= \hat{\eta}_{k-1}^* + \gamma_k W_k \hat{\bar{y}}_k(\tau - I(y_i - \rho\hat{\bar{y}}_k - \hat{\boldsymbol{\beta}}_{k-1}^{*T}\boldsymbol{x}_k))\cos(\hat{\eta}_{k-1}^*) \\
\hat{\rho}_k^* &= \sin(\hat{\eta}_k^*).
\end{aligned} \tag{3.16}$$

Here, $\boldsymbol{x}_k$ and y_k are from the k-th data point and $\hat{\bar{y}}_k$ is the k-th element of $\widehat{WY}$ calculated in Equation (3.14).

3.3 Simulation Studies

We use simulations to study the finite sample properties of these two SGD-based estimators and the empirical coverage of the confidence intervals. We use '4-neighbors' neighborhood matrix discussed in Section 2.5.1 and generate a total of $N = 90,000$ data points for each independent replicate. We follow [50] for the data generation process:

$$y_i = \rho(v_i)\bar{y}_i + \boldsymbol{\beta}(v_i)^T\boldsymbol{x}_i, \tag{3.17}$$

for $i = 1, 2, \cdots, N$, where v_i are i.i.d $U(0,1)$, $\bar{y}_i = \boldsymbol{w}_i\boldsymbol{Y}$, $\boldsymbol{w}_i$ the i-th row of neighborhood matrix $\boldsymbol{W}$, $\boldsymbol{\beta} = [\beta_0, \beta_1, \beta_2]^T$, $\boldsymbol{x}_i = [1, x_{1i}, x_{2i}]^T$ with x_{1i}, x_{2i} are i.i.d. $U(-1,1)$. We use five settings which are different in terms of how $\boldsymbol{\beta}$ and ρ changes with the quantile τ and whether the error error distribution is symmetric:

- **Setting 1:** $\boldsymbol{\beta}(v_i) = [0.5, 0.5, -0.5]^T + [1,0,0]^T F_1^{-1}(v_i)$, $\rho = 0.3$
- **Setting 2:** $\boldsymbol{\beta}(v_i) = [0.5, 0.5, -0.5]^T + [1,0.1,0.1]^T F_1^{-1}(v_i)$, $\rho = 0.3$
- **Setting 3:** $\boldsymbol{\beta}(v_i) = [0.5, 0.5, -0.5]^T + [1,0.1,0.1]^T F_1^{-1}(v_i)$, $\rho(v_i) = 0.3 + 0.1F_1^{-1}(v_i)$
- **Setting 4:** $\boldsymbol{\beta}(v_i) = [0.5, 0.5, -0.5]^T + [1,0.1,0.1]^T F_2^{-1}(v_i)$, $\rho = 0.3$
- **Setting 5:** $\boldsymbol{\beta}(v_i) = [0.5, 0.5, -0.5]^T + [1,0.1,0.1]^T F_2^{-1}(v_i)$, $\rho(v_i) = 0.3 + 0.1F_2^{-1}(v_i)$,

where, F_1, F_2 are probability distribution functions. We use standard normal distribution for F_1 and standardized χ^2_3 (mean 0 and variance 1) for F_2. We use quantile $\tau = 0.1, 0.25, 0.5, 0.75$, and 0.9. The true parameter values for these quantiles are listed in Table 3.1. Thus, in setting 1, only β_0 changes with τ; in settings 2 and 4, all the β_i change with τ, but ρ is constant. In settings 3 and 5, ρ changes with τ as well. For settings 1,2,3, the error distribution is symmetric, while for settings 4,5, it is skewed. One advantage of quantile regression over mean regression is that it can model skewed data.

Table 3.1 Summary of True Quantile Parameters Used in Simulations

Setting	τ	0.1	0.25	0.5	0.75	0.9
	$\beta_1(\tau)$			0.5		
1	$\beta_2(\tau)$			-0.5		
	$\rho(\tau)$			0.3		
	$\beta_1(\tau)$	0.372	0.433	0.500	0.567	0.628
2	$\beta_2(\tau)$	-0.628	-0.567	-0.500	-0.433	-0.372
	$\rho(\tau)$			0.3		
	$\beta_1(\tau)$	0.372	0.433	0.500	0.567	0.628
3	$\beta_2(\tau)$	-0.628	-0.567	-0.500	-0.433	-0.372
	$\rho(\tau)$	0.172	0.233	0.300	0.367	0.428
	$\beta_1(\tau)$	0.401	0.427	0.474	0.545	0.633
4	$\beta_2(\tau)$	-0.599	-0.537	-0.526	-0.455	-0.367
	$\rho(\tau)$			0.3		
	$\beta_1(\tau)$	0.401	0.427	0.474	0.545	0.633
5	$\beta_2(\tau)$	-0.599	-0.537	-0.526	-0.455	-0.367
	$\rho(\tau)$	0.201	0.207	0.274	0.345	0.433

We first use 1SQR and 2SQR-based SGD and CIs construction for the first setting and for all five quantiles listed above. For each combination we generate 200 independent replications. The mean and standard deviation of estimates from these 200 independent replications and the empirical coverage of confidence intervals are shown in Table 3.2. (Only the results for β_1 and ρ are shown, since results for β_2 is very similar to those for β_1.) For 1SQR based on SGD, we find that the mean of the estimates are close to the true values (i.e., with minimal bias) and the empirical coverages are at the nominal level for regressors coefficient. For the spatial parameter ρ, the estimates are biased and the empirical coverage of CIs are all below the nominal level. This result is consistent with the understanding that 1SQR is biased. For 2SQR-based SGD, simulation results show that the mean of the estimates are close to the true values. This is consistent with the understanding that 2SLS is generally unbiased. The empirical coverage of confidence intervals are all close to the nominal level.

We then use 2SQR-based SGD for Setting 2-5 for all five quantiles. Results are shown in Tables 3.3 - 3.6. For all these settings, the mean of the estimates are close to the true values and empirical coverage of the CIs are at the nominal level for all parameters. These simulation results suggest that our 2SQR-based SGD and CI construction procedure work well for various scenarios regarding how the coefficients depend on quantile τ and whether the error distribution is symmetric or not.

3.4 Summary and Discussion

This chapter develops the SGD-based parameter estimation and confidence interval construction algorithm for the SAR quantile regression model. The algorithms are developed based on one-stage quantile regression (1SQR) and two-stage quantile regression (2SQR). The 1SQR method ignores the endogeneity of the spatial lag term and treats it as an independent covariate. The 2SQR method, on the other

Table 3.2 Simulation Result for SAR Quantile Regression–Setting 1

τ		1SQR			2SQR		
		bias	sd	CI	bias	sd	CI
0.1	$\beta_1(\tau)$	0.001	0.011	0.930	-0.001	0.011	0.945
	$\rho(\tau)$	0.217	0.015	0.485	0	0.032	0.945
0.25	$\beta_1(\tau)$	0	0.008	0.960	-0.001	0.009	0.965
	$\rho(\tau)$	0.215	0.013	0.12	0.004	0.026	0.990
0.5	$\beta_1(\tau)$	0	0.008	0.955	0	0.009	0.960
	$\rho(\tau)$	0.214	0.012	0.005	0.009	0.033	0.965
0.75	$\beta_1(\tau)$	0.001	0.009	0.95	-0.001	0.011	0.935
	$\rho(\tau)$	0.215	0.013	0	0.008	0.025	0.970
0.9	$\beta_1(\tau)$	0.001	0.012	0.940	-0.003	0.018	0.985
	$\rho(\tau)$	0.214	0.044	0.335	0.004	0.025	0.945

hand, first models the spatial lag term with instrumental variables by quantile regression. Then the fitted spatial lag term is used as an independent covariates for modeling the response variable in the second stage. Parameters are estimated based on minimizing the check function for 1SQR and for both stages of 2SQR. Simulations results show that the 2SQR-based SGD parameter estimation method is unbiased and the empirical coverage of constructed confidence interval are at the desired level.

Unlike SAR mean regression model, where the time needed for MLE is much longer than that for the SGD procedure, for SAR quantile regression model, the time for SGD procedure is not necessarily less than directly applying quantile regression function provided in standard packages (e.g., 'statsmodels' package in Python [49], 'quantreg' package in R [26]). However, the SGD-based procedure can still provide benefits in two aspects. First, in theory there is no upper limit for the sample size

Table 3.3 Simulation Result for SAR Quantile Regression–Setting 2

τ		bias	SD	CI
0.1	β_1	-0.001	0.012	0.965
	ρ	-0.011	0.070	0.920
0.25	β_1	-0.003	0.010	0.980
	ρ	-0.005	0.037	0.960
0.5	β_1	-0.002	0.008	0.955
	ρ	0.004	0.032	0.970
0.75	β_1	-0.002	0.010	0.970
	ρ	0.004	0.023	0.975
0.9	β_1	-0.003	0.016	0.935
	ρ	0.006	0.030	0.940

that can be applied, since for each iteration, only one data point and its neighbors are required for updating the estimates. Second, it is possible to extend the algorithm discussed here to an online version. The 1SQR-based method is clearly an online method. For the 2SQR-based SGD method, we can use the current available data to get the parameter estimates and construct the confidence intervals. Denoting these estimates from the first and second stages as $\hat{\theta}_{s1}$ and $\hat{\theta}_{s2}$, when new data are available, we can use $\hat{\theta}_{s1}$ and $\hat{\theta}_{s2}$ as the starting points to continue updating the parameters in each of the two stages. Thus, in this way, we can make the 2SQR-based SGD method an online method as well. This idea can be extended further by considering one data point at a time: i.e., take one data point and update the parameter for first stage, then obtain the predicted spatial lag term, and use it together with the current data point for updating parameters in the second stage. This process is then applied for

Table 3.4 Simulation Result for SAR Quantile Regression–Setting 3

τ		bias	SD	CI
0.1	β_1	0.002	0.022	0.955
	ρ	0.046	0.135	0.93
0.25	β_1	0.001	0.018	0.955
	ρ	0.025	0.084	0.955
0.5	β_1	-0.011	0.028	0.945
	ρ	0.003	0.026	0.970
0.75	β_1	0.003	0.020	0.955
	ρ	-0.002	0.071	0.95
0.9	β_1	0	0.029	0.955
	ρ	-0.037	0.184	0.94

all data points. This way once the data point and all its neighbors are used, it can be discarded. This method is worth further investigation.

As discussed in Section 3.1, besides the 2SLS method, another instrumental variable based on method is the **CH** method [6]. We can derive SGD updating equations based on this method. First, we select one value of ρ from a pre-defined grid of values, say ρ_i. Then we can use SGD to fit the model in Equation (3.7) . This process can be done for each ρ in the given grid and the final estimate of ρ is selected based on Equation (3.8). Perturbed estimates can only be generated for $\boldsymbol{\beta}$ but not for ρ when fitting the model for Equation (3.7). Thus, we can only use the online bootstrapping to construct confidence interval for $\boldsymbol{\beta}$ but not for ρ. Hence, we do not consider the **CH** method here. However, obtaining CIs for ρ using online bootstrapping can be a focus for future work.

Table 3.5 Simulation Result for SAR Quantile Regression–Setting 4

τ		bias	SD	CI
0.1	β_1	0.002	0.004	0.96
	ρ	-0.003	0.015	0.975
0.25	β_1	-0.002	0.005	0.95
	ρ	-0.001	0.005	0.95
0.5	β_1	-0.003	0.009	0.95
	ρ	0.004	0.025	0.945
0.75	β_1	-0.009	0.020	0.945
	ρ	-0.001	0.019	0.95
0.9	β_1	-0.013	0.038	0.945
	ρ	0.002	0.021	0.955

Table 3.6 Simulation Result for SAR Quantile Regression–Setting 5

τ		bias	SD	CI
0.1	β_1	0.002	0.008	0.955
	ρ	-0.006	0.037	0.935
0.25	β_1	0.002	0.011	0.930
	ρ	0.006	0.031	0.955
0.5	β_1	0	0.015	0.955
	ρ	0.015	0.063	0.970
0.75	β_1	-0.004	0.031	0.975
	ρ	0.002	0.047	0.925
0.9	β_1	-0.001	0.052	0.970
	ρ	0.022	0.131	0.935

CHAPTER 4

PUF DATA ANALYSIS

4.1 Introduction

According to US census data, in 2018 alone, 8.5 percent of people, or 27.5 million, did not have health insurance at any point during the year. These numbers were increased from 7.9 percent (25.6 million) in 2017 [2]. Uninsured Americans are especially vulnerable to the high cost of health care. The Physician and Other Supplier Public Use File (PUF) provides information on medical service and procedures provided to Medicare beneficiaries by physicians and other medical professionals. It contains information about submitted charges, Medicare allowed amount, Medicare payment amount and Medicare standardized payment amount [3]. In this chapter, we analyze the PUF data to study the effect of location and other characteristics of medical facilities on medical prices. Results from this analysis can help to improve the transparency in healthcare pricing and thus, benefit both insured and uninsured patients.

4.2 Data Description

Currently, PUF data from year 2013 to 2019 are available. For year 2017, the dataset is of size 3GB and contains more than 9 million records, 2018 dataset 3GB and 10 million records, 2019 dataset 3GB more than 10 million records. To analyze dataset of this size, scaleable statistical methods studied in this dissertation are necessary. Also, spatial information are included in the PUF dataset, being the locations of Physicians and Medical providers. For many other products or services, prices depend on locations. For example, services provided in nearby locations tend to have similar prices. We use the spatial autoregressive model to take spatial location into consideration and investigated its effect on charges of medical services. Both

mean regression and quantile regression are fitted and compared. Quantile regression can help to identify the relationship between charges and other factors in different parts of the distribution.

The data studied are the PUF data from year 2017. Each entry of the PUF data contains variables in the following categories: 1) identification characteristics of the provider 2) location that the service is provided 3) identification characteristics of the medical service 4) average charges of the medical service. The charges depend on the medical service types and for illustration purpose, this dissertation only analyzes one type of medical service: CT scan on face. Using this dataset, we investigate relationship of average charge submitted by the provider with the following variables:

- **Gender**: Gender of the provider. This analysis only considered the providers registered in NEPPS (National Plan and Provider Enumeration System) as individuals. Female is coded as 1 while male is coded as 2.
- **RUCA**: Rural-Urban Commuting Area Codes (RUCAs), are a Census tract-based classification scheme that utilizes the standard Bureau of Census Urbanized Area and Urban Cluster definitions in combination with work commuting information to characterize all of the nation's Census tracts regarding their rural and urban status and relationships. Majority of them are belong to 'Metropolitan area core'. For this analysis we coded 'Metropolitan area core' as 1 and all others as 2.
- **place of service**: The place of service: facility (coded as 1) or non-facility (coded as 2).
- **total number of service**: Total number of services provided by the provider for a certain year.

After removing outliers and missing data, this dataset contains 14464 data points.

We first study the weight average charges in each state. Total number of services is used as the weight. The result is shown in Figure 4.1. California has the highest weighted average charges and South Dakota has the lowest. Clustering exists in this map and this suggests that data are spatially correlated. Figure 4.2 is the scatter plot of average submitted charges and total number of services provided by a certain

provider. This plot shows the negative association between the average submitted charges and number of services provided, since smaller average submitted charges corresponds to a relative higher proportion of large number of services. Figure 4.3 shows the boxplot of average submitted charges grouped by each of these category variables: Gender, RUCA, place of service. The plot suggests that overall gender of male, RUCA of not in metropolitan area, and place of non-facility has a lower average submitted charges compared to the other level for the same covariate.

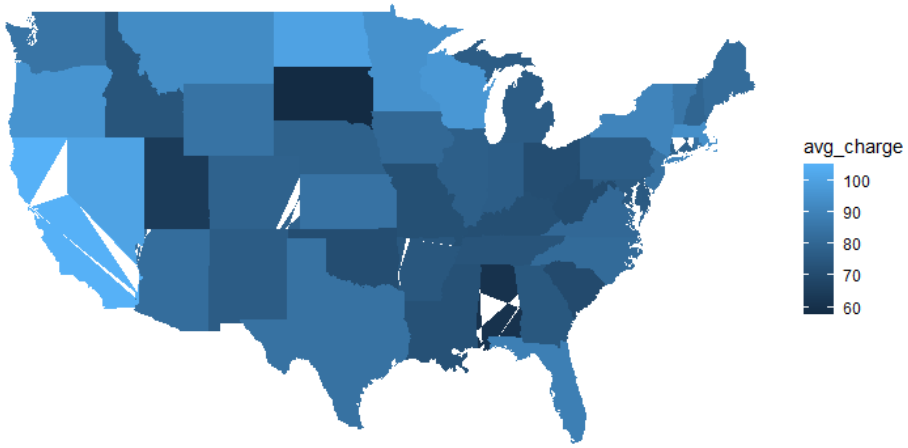


Figure 4.1 Weighted average charge in each state of the Contiguous US.

The neighborhood matrix for this dataset is determined in the following way. We first replace the locations of providers as the centroids of their zipcode and then calculate the distances of all provider pairs. Then we calculate the 1% quantile of all these distances, which is about 20.3 miles. Data points with distance less than this are treated as neighbors. (On average, each data point is neighbors to 1% of all data points, about 144 data points. See Section 4.4 for other ways to generate neighborhood matrix.) We give equal weights for all neighbors and the neighborhood matrix is then row normalized with row sum equal to 1.

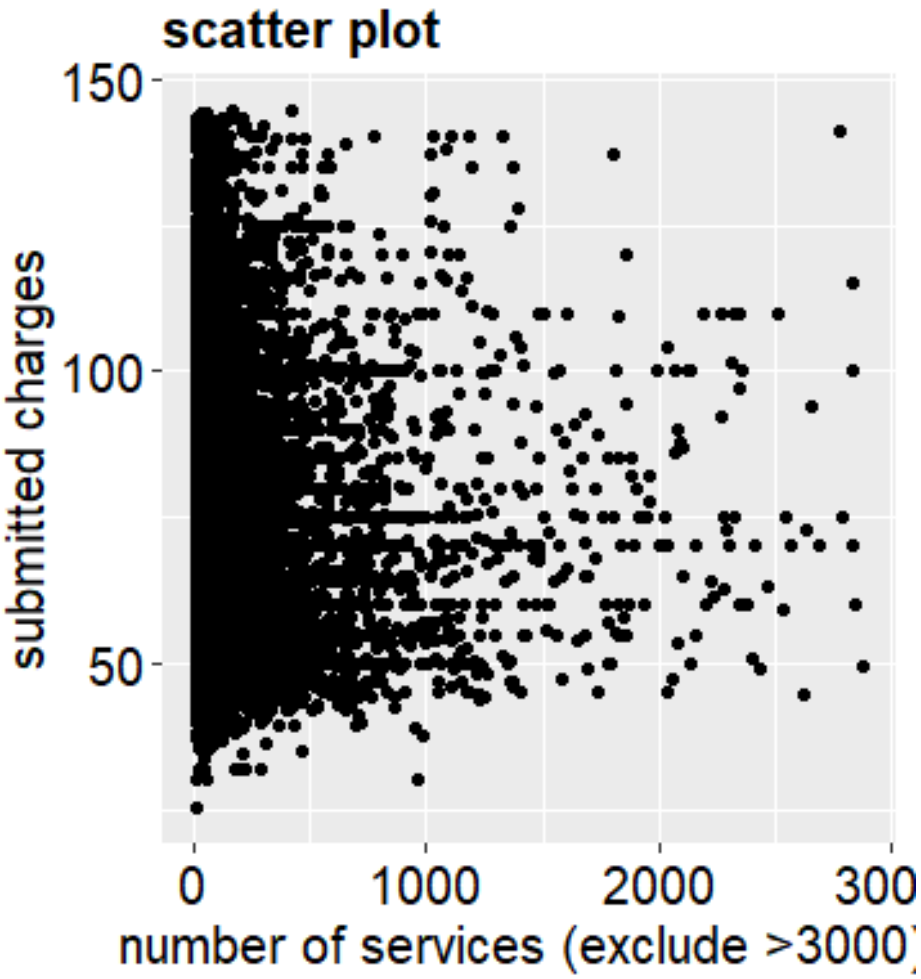


Figure 4.2 Scatter plot of average submitted charges and number of services.

We fit both the SAR mean regression and quantile regressions for the dataset. SGD based on two stage quantile regression procedures are used for parameter estimation and CIs construction for quantile regressions.

4.3 Models

4.3.1 SAR mean regression

We first estimate parameters and construct CIs for the SAR mean regression model with SGD algorithm developed in Section 2.3. The result is shown in Table 4.1. The confidence interval of ρ not covering 0 suggests that data are spatial correlated. The

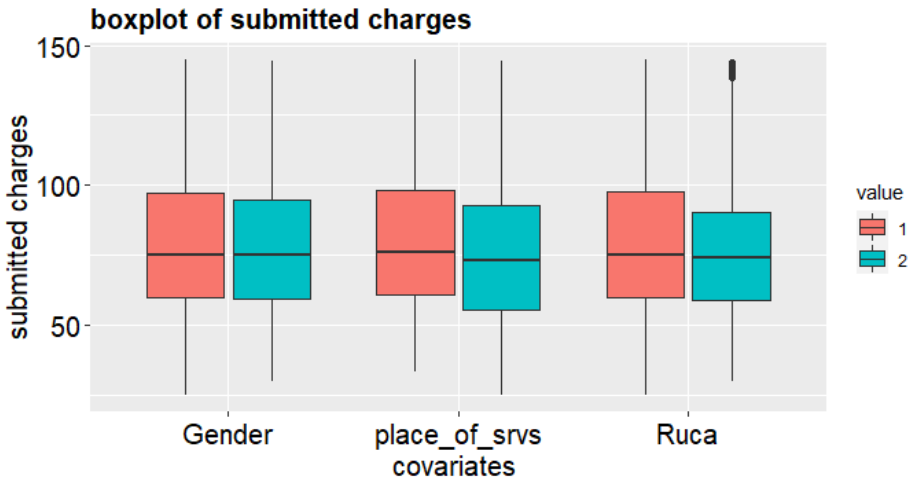


Figure 4.3 Boxplot of average submitted plot grouped by Gender, RUCA, and place of service

effect of all these covariates on submitted charges are significant since 0 is not inside in any of these confidence intervals. The effect of these covariates suggested by this model is consistent with the result of descriptive statistical analysis. For examples, it suggests that providers providing more services a year charges less given other factors are the same. Also, provides not in the metropolitan area charges less given other factors are the same.

4.3.2 SAR quantile regression

For SAR quantile regression, SGD based on two-stage quantile regression procedures are used. We fitted for the following quantiles: 0.1, 0.25, 0.5, 0.75, 0.9 (Table 4.2. The contribution of values of neighbors and other variables are consistent across different quantiles in terms of the sign of coefficients. And the sign is the same as the result of mean regression. As mentioned before, quantile regression provides a broader information about the effect of these factors on submitted charges.

Table 4.1 Point Estimates and 95% CIs for Mean Regression

Variable	Estimate	95% CI
Gender	-1.84	(-2.09, -1.59)
RUCA	-3.23	(-3.47, -2.99)
place	-4.8	(-5.05, -4.56)
total number	-43.2	(-43.21, -43.20)
ρ	-0.25	(-0.27, -0.23)

4.4 Summary and Discussion

This chapter investigates the relationship between submitted charges and some factors in PUF data for the service of CT scan on face. We fit spatial autoregressive models to incorporate spatial correlations. Both the mean regression and quantile regression models are built and all models suggest the existence of spatial correlation between data points. Also, the effect of these factors are consistent between mean regression and quantile regressions. Besides the analysis done in this chapter, this dataset can be further analyzed by the following ways:

- Considering other ways to generate neighborhood matrix. For this chapter, neighborhood matrix is generated by replacing the location of providers with the centroids of their zipcode and treating two data points as neighbors if their distance is less than a threshold. Also equal weights are given for all neighbors. Alternatives include directly calculating the distance between providers without replacing them with centroids of zipcode or give different weights for neighbors based on the distances. Also, different thresholds can be chosen to determine whether two locations are neighbors.
- Considering more covariates. Examples are charges of other services provided by the same provider, some social and economic variables associated with the district that the provider is located, etc.
- We can analyze data for the same medical services from multiple years and see if there are any changes between different years. What's more, the method

Table 4.2 Point Estimates and 95% CIs for Quantile Regression

Estimate					
τ	Gender	RUCA	place	tot num	ρ
0.1	-9.64	-9.89	-12.15	-59.54	-0.73
0.25	-6.17	-7.09	-9.00	-59.50	-0.75
0.5	-1.55	-2.69	-4.93	-59.44	-0.78
0.75	-2.07	-2.79	-4.92	-59.46	-0.45
0.9	-1.80	-2.23	-4.27	-59.43	-0.24
95% CI					
τ	Gender	RUCA	place	tot num	ρ
0.1	(-12.55, -6.73)	(-12.37, -7.40)	(-14.98, -9.31)	(-59.62, -59.46)	(-0.91, -0.54)
0.25	(-9.47, -2.86)	(-9.93, -4.26)	(-12.23, -5.77)	(-59.58, -59.41)	(-0.97, -0.52)
0.5	(-2.86, -0.23)	(-3.84, -1.53)	(-6.27, -3.59)	(-59.47, -59.41)	(-0.87, -0.70)
0.75	(-3.35, -0.79)	(-3.94, -1.64)	(-6.23, -3.61)	(-59.48, -59.44)	(-0.53, -0.37)
0.9	(-3.00, -0.61)	(-3.35, -1.12)	(-5.51, -3.03)	(-59.45, -59.41)	(-0.31, -0.17)

developed here can be used for online learning. We can update the parameter estimates when new data for a certain year is available.

- The analysis can be easily extended to other medical services provided. Also, we can analyze several related services together, For example, we can analyze all types of CT scans, including CT scan on face, with a dummy variable for the type of CT scans.

CHAPTER 5

CONCLUSION

Spatial correlation exists in many types of data and ignoring it can affect the estimation of parameters (e.g., see Section 2.5.3). Spatial autoregressive (SAR) models are usually used to study spatially correlated data. For the SAR model, the response variable depends not only on the covariates but also the values of its neighbors. Stochastic gradient descent (SGD) is an iterative parameter estimation method that minimizes a target function by processing each data point in turn. Thus, SGD can scale up easily for large datasets and is suitable for online learning. Fang et al. developed an online bootstrapping method for statistical inference of estimates obtained by SGD in the case of independent data[13]. In this research, we consider SGD and online bootstrapping algorithms for parameter estimation and inference in the presence of spatially correlation, specifically for the SAR model. In particular, we study: (1) parameter estimation and inference for the SAR mean regression model; (2) parameter estimation and inference for the SAR quantile regression model; (3) analysis of the PUF data using the SAR model.

For the SAR mean regression model, the MLE is unbiased and most efficient. However, getting MLE for large dataset is computationally heavy. We propose a modified SGD algorithm based on MLE for parameter estimation and statistical inference to accommodate the spatial data correlation. Results show that SGD based estimation is at least 40 times faster than MLE, with the SGD estimators for all parameters close to the true values. The empirical coverages of CIs are at the nominal level for the coefficients of the covariates but not for the spatial parameter. Two methods are proposed to improve the empirical coverages of ρ CI. Also, we develop the 2SLS-based SGD algorithm for parameter estimation and CIs

construction. Simulation results show that estimates of this method are unbiased and empirical coverages of CIs constructed are at the nominal levels. Lastly, we also develop the SGD algorithm for the spatial autoregressive model with random disturbance.

The second part of this dissertation investigates SAR quantile regression model. Compared with SAR mean regression model, this quantile regression model can provide a more detailed information on the distribution of the response variable. Also, it provides more flexibility on model specification, as it allows the coefficients and spatial parameter to be dependent on quantiles. We focus on two parameter estimation methods: one-stage quantile regression (1SQR) and two-stage quantile regression (2SQR). 1SQR ignores the endogeneity of spatial lag term and treats it the same as the exogenous regressors. The parameters are estimated in the same way as linear quantile regression. In the first stage of 2SQR, a quantile regression is fit for the spatial lag term using the exogenous regressors and instrumental variables. Then in the second stage, the fitted spatial lag term is used to fit the quantile regression for the response variable. We develop parameter estimation and inference algorithms with SGD based on these two algorithms. Simulation results show that SGD estimator based on 2SQR is unbiased while that based on 1SQR is biased. Also, the empirical coverages of CIs constructed based on 2SQR are all at the nominal levels.

Finally, we analyze the Physician and Other Supplier Public Use File (PUF) data using the methods described in Chapters 2 and 3. This dataset contains information about charges submitted for medical services provided to Medicare beneficiaries by physicians and healthcare professionals at medical facilities. The results suggest that the locations of facilities have significant effect in modelling the medical charges. Also, the models find that charges depend on the total number of services provided yearly, gender of the provider, facility type, and whether the provider is in a metropolitan area.

As discussed in the end of previous chapters, several work are the focus of the future work:

- Derive the asymptotic properties for MLE-based and 2SLS-based SGD estimates and perturbed estimates for the SAR mean regression model
- Develop methods to improve the empirical coverage for ρ CI constructed for the SAR mean regression model
- Derive the asymptotic properties for 2SQR-based SGD estimates and perturbed estimates for the SAR quantile regression model
- Derive the 2SQR-based SGD algorithm using one data point at a time for both stages for the SAR quantile regression model

Besides the future work mentioned above, this work can also be extended to other models. One example is the spatial autoregressive with autoregressive disturbance model (SARAR model):

$$\boldsymbol{y} = \rho_1 \boldsymbol{W}_1 \boldsymbol{y} + \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{u}, \quad \boldsymbol{u} = \rho_2 \boldsymbol{W}_2 \boldsymbol{u} + \boldsymbol{\epsilon} \tag{5.1}$$

Here, $\boldsymbol{y}, \boldsymbol{X}, \boldsymbol{\beta}$ and $\boldsymbol{\epsilon}$ are defined the same as in (2.1). $\boldsymbol{W}_1, \boldsymbol{W}_2$, $n \times n$, are neighborhood matrices; ρ_1, ρ_2, are autoregressive parameters. This model provides more flexibility in modeling spatial correlated data [31, 46] and is worth further investigation. Similar extensions can be made to the conditional autoregressive (CAR) model.

APPENDIX A

PROOFS OF THEOREMS

Detailed proof of theorems in this dissertation are shown here.

A.1 Proof of Theorem 1

We prove this theorem by verifying the assumption A1 to A5 in [13] (label them as FA1 to FA5). Note our Assumption A4 is FA5, thus, we only need to verify FA1 to FA4. First list the log-likelihood and its derivative:

$$\ell(\boldsymbol{\theta}) = -\frac{\ln(\sigma^2)}{2}n - \frac{(\boldsymbol{Ay} - \boldsymbol{X\beta})^T(\boldsymbol{Ay} - \boldsymbol{X\beta})}{2\sigma^2} + \ln|\boldsymbol{A}|$$

$$\nabla\ell_{\boldsymbol{\beta}} = \frac{1}{\sigma^2}(\boldsymbol{X}^T\boldsymbol{Ay} - \boldsymbol{X}^T\boldsymbol{X\beta})$$

$$\nabla\ell_{\sigma^2} = -\frac{n}{2\sigma^2} + \frac{1}{2(\sigma^2)^2}(\boldsymbol{Ay} - \boldsymbol{X\beta})^T(\boldsymbol{Ay} - \boldsymbol{X\beta})$$

$$\nabla\ell_{\rho} = -tr(\boldsymbol{A}^{-1}\boldsymbol{W}) + \frac{(\boldsymbol{Ay} - \boldsymbol{X\beta})^T\boldsymbol{Wy}}{\sigma^2}$$

A.1.1 Verification of assumption FA1

Assumption FA1. The objective function $L(\boldsymbol{\theta})$ is convex, continuously differentiable over $\boldsymbol{\theta} \in \Theta$, and twice continuously differentiable at $\boldsymbol{\theta} = \boldsymbol{\theta}_0$, where, $\boldsymbol{\theta}_0$ is the unique minimizer of $L(\boldsymbol{\theta})$.

As defined in Section 2.4, $S(\boldsymbol{\theta}) = \nabla^2 L(\boldsymbol{\theta})$ and clearly $S(\theta)$ is positive semi-definite, thus, $L(\theta)$ is a convex function. Also, according to [33], under some Assumption A3, θ_0 is the unique minimizer of $L(\theta)$ (Other assumptions listed in [33]

can be easily verified). Clearly $L(\theta)$ is continuous differentiable over θ and twice continuous differentiable at $\theta = \theta_0$. Thus, we finish the verification of A1.

A.1.2 Verification of assumption FA2

Assumption FA2 The gradient of $L(\theta), R(\theta) = \nabla L(\theta)$, is Lipschitz continuous with constant $L_1 > 0$; that is, for any θ_1 and θ_2, $\|R(\theta_1) - R(\theta_2)\| \leq L_1 \|\theta_1 - \theta_2\|$.

We can write $R(\boldsymbol{\theta})$ as $[R_1^T, R_2, R_3]^T$, and $R_1 = -\mathbb{E}[\nabla \ell_{\boldsymbol{\beta}}]$, $R_2 = -\mathbb{E}[\nabla \ell_{\sigma^2}]$, and $R_3 = -\mathbb{E}[\nabla \ell_\rho]$. To prove $R(\boldsymbol{\theta})$ is Lipschitz continuous, we only need to show R_1, R_2 and R_3 are Lipschitz continuous.

For $\boldsymbol{R_1}$:

$$R_1 = -\mathbb{E}[\nabla \ell_{\boldsymbol{\beta}}] = -\mathbb{E}[\frac{1}{\sigma^2}(\boldsymbol{X}^T\boldsymbol{A}\boldsymbol{y} - \boldsymbol{X}^T\boldsymbol{X}\boldsymbol{\beta})]. \tag{A.1}$$

Clearly R_1 is a linear function of ρ and $\boldsymbol{\beta}$, thus, it is Lipschitz continuous w.r.t. ρ and $\boldsymbol{\beta}$. Also, easy to show R_1 is Lipschitz continuous w.r.t. σ^2 as long as $\sigma^2 \in [a, \infty)$ for some $a > 0$. And this is guaranteed by Assumption A2.

For $\boldsymbol{R_2}$

$$R_2 = -\mathbb{E}[\nabla \ell_{\sigma^2}] = -\mathbb{E}[-\frac{n}{2\sigma^2} + \frac{1}{2(\sigma^2)^2}(\boldsymbol{A}\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})^T(\boldsymbol{A}\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})]. \tag{A.2}$$

Clearly, R_2 is a quadratic function for $\boldsymbol{\beta}$ and ρ. Easy to show R_2 is Lipschitz continuous w.r.t. them as long as $\|\boldsymbol{\beta}\| \in [\ 0, b]\ , b \geq 0$ (Guaranteed by Assumption A2. Also, note we already have the restriction that $\rho \in (-1, 1)$). To show R_2 is Lipschitz continuous w.r.t. σ^2, is the same as to show a function $y = \frac{C_1}{x} + \frac{C_2}{x^2}$ is Lipschitz continuous w.r.t. x. Easy to show as long as $\sigma^2 \in [\ a, \infty), a > 0$, R_2 is Lipschitz continuous w.r.t. σ^2.

For $\boldsymbol{R_3}$

$$R_3 = -\mathbb{E}[\nabla \ell_\rho] = \mathbb{E}[tr(\boldsymbol{A}^{-1}\boldsymbol{W})] - \mathbb{E}[\frac{(\boldsymbol{A}\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})^T\boldsymbol{W}\boldsymbol{y}}{\sigma^2]} \tag{A.3}$$

Easy to see, the second part of R_3 (the part does not include $tr(\boldsymbol{A}^{-1}\boldsymbol{W})$ is Lipschitz continuous w.r.t. $\boldsymbol{\beta}, \sigma^2$, and ρ given $\sigma^2 \in [a, \infty), a > 0$. For the first part, note,

$$\begin{aligned}
\boldsymbol{A}^{-1} &= (\boldsymbol{I} - \rho\boldsymbol{W})^{-1} = \sum_{k=0}^{\infty}(\rho\boldsymbol{W})^k = \sum_{k=0}^{\infty}\rho^k\boldsymbol{W}^k \\
\text{let } f(\rho) &= (\boldsymbol{A}^{-1}\boldsymbol{W}) = (\sum_{k=0}^{\infty}\rho^k\boldsymbol{W}^{k+1}) = \sum_{k=0}^{\infty}\rho^k tr(\boldsymbol{W}^{k+1}) \\
|f(\rho_1) - f(\rho_2)| &= |\sum_{k=0}^{\infty}(\rho_1^k - \rho_2^k)tr(\boldsymbol{W}^{k+1})| \leq \sum_{k=0}^{\infty}|\rho_1^k - \rho_2^k|tr(\boldsymbol{W}^{k+1}) \\
&\leq N\sum_{k=0}^{\infty}|\rho_1^k - \rho_2^k|, \text{ since } tr(\boldsymbol{W}^{k+1}) \leq N \\
&= N\sum_{k=1}^{\infty}|\rho_1^k - \rho_2^k| = N\sum_{k=1}^{\infty}|(\rho_1 - \rho_2)\sum_{m=0}^{k-1}(\rho_1^m\rho_2^{k-1-m})| \\
&\leq |\rho_1 - \rho_2|\sum_{k=1}^{\infty}k\rho_m^{k-1}, \text{ given } |\rho| \leq \rho_m \\
&= |\rho_1 - \rho_2|C, \; C = \sum_{k=1}^{\infty}k\rho_m^{k-1} \text{which converges given } 0 < \rho_m < 1
\end{aligned}$$

Easy to see $f(\rho)$ is Lipschitz continous w.r.t. ρ given $|\rho| \in [\rho_{\min}, \rho_{\max}]$, $0 < \rho_{\min} \leq \rho_{\max} < 1$. This is guaranteed by Assumption A2. This R_3 is Lipschitz continuous w,r.t. ρ. (Note: $tr(\boldsymbol{W}^k) \leq N$ can be argued as following. All the elements of $\boldsymbol{W}$ between 0 and 1 and the sum of each row of $\boldsymbol{W}$ is 1. Thus, each element of $\boldsymbol{W}^2$ is just a weighted average of elements in a certain column of $\boldsymbol{W}$. Thus, each element of $\boldsymbol{W}^2$ is between 0 and 1. Follow the same procedure, we can show elements of $\boldsymbol{W}^k$ is between 0 and 1.)

A.1.3 Verification of assumption FA3

Assumption FA3 The Hessian matrix of $L(\theta), S(\theta) = \nabla^2 L(\theta)$, exists and is positive definite at θ_0 with $S_0 = S(\theta_0) > 0$ and is Lipschitz continuous at θ_0 with constant $L_2 > 0$.

$\boldsymbol{S(\theta)} = \nabla \boldsymbol{R}(\boldsymbol{\theta})$ and since $\boldsymbol{R}(\theta)$ is a smooth function, $\boldsymbol{S(\theta)}$ exists. First to show $\boldsymbol{S}_0$ is Lipschitz continuous, we only need to show all its elements (elements in this matrix) are Lipschitz continuous w.r.t. $\boldsymbol{\beta}, \sigma^2$ and ρ respectively. We show this by showing the derivative of R_1, R_2 and R_3 w.r.t. $\boldsymbol{\beta}, \sigma^2$ and ρ are Lipschitz continuous.

For derivatives of $\boldsymbol{R_1}$

R_1 is a linear function for ρ and $\boldsymbol{\beta}$, thus, $\frac{\partial R_1}{\partial \rho}$ and $\frac{\partial \boldsymbol{R_1}}{\partial \boldsymbol{\beta}}$ are Lipschitz continuous. $\frac{\partial R_1}{\partial \sigma^2}$ is in the format of $\frac{C}{(\sigma^2)^2}$. It is Lipschitz continuous as long as $\sigma^2 \in [\ a, \infty), a > 0$. And this is guaranteed by Assumption A2.

For derivatives of $\boldsymbol{R_2}$

R_2 is a quadratic function for $\boldsymbol{\beta}$ and ρ, the derivative is a linear function, which is Lipschitz continuous. $\frac{\partial R_2}{\partial \sigma^2}$ is in the format of $\frac{C_1}{(\sigma^2)^2} + \frac{C_2}{(\sigma^2)^3}$. Easy to show it is Lipschitz continuous w.r.t. x as long as $\sigma^2 \in [\ a, \infty), a > 0$.

For derivatives of $\boldsymbol{R_3}$

Following the similar argument for showing R_3 is Lipschitz continuous, easy to see derivative of R_3 is Lipschitz continuous w.r.t. $\boldsymbol{\beta}$ and σ^2. Derivative second half of R_3 is also Lipschitz continuous w.r.t. ρ. If we differentiate the first half of R_3, we will get the term of $k\rho_{k-1}$. We can show the derivative is also Lipschitz continuous by following the same argument as showing the second half of R_3 is Lipschitz continuous.

Next we show $\boldsymbol{S}_0$ is positive definite. We first we write down expression of S_0 explicitly, $\boldsymbol{S} = \nabla \boldsymbol{R} = \nabla^2(-L)$. Take derivative of $\boldsymbol{R}$ with respect to $\boldsymbol{\beta}, \sigma^2$ and ρ,

$$-\frac{\partial L}{\partial \boldsymbol{\beta}} = -\frac{1}{\sigma^2} E\{\boldsymbol{X}^T(\boldsymbol{A}\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})\} \tag{A.4}$$

$$-\frac{\partial L}{\partial \sigma^2} = \frac{n}{2\sigma^2} - \frac{1}{2(\sigma^2)^2} E\{(\boldsymbol{A}\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})^T(\boldsymbol{A}\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})\} \tag{A.5}$$

$$-\frac{\partial L}{\partial \rho} = \mathrm{tr}(\boldsymbol{A}^{-1}\boldsymbol{W}) - \frac{1}{\sigma^2} E\{(\boldsymbol{A}\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})^T \boldsymbol{W}\boldsymbol{y}\} \tag{A.6}$$

Take second derivative and evaluate at $\boldsymbol{\theta}_0$,

$$-\left.\frac{\partial^2 L}{\partial \boldsymbol{\beta}^2}\right|_{\boldsymbol{\theta}_0} = \frac{\mathbb{E}\{\boldsymbol{X}^T\boldsymbol{X}\}}{\sigma_0^2} \tag{A.7}$$

$$\begin{aligned} -\left.\frac{\partial^2 L}{\partial \boldsymbol{\beta}\partial\sigma^2}\right|_{\boldsymbol{\theta}_0} &= \left.(\frac{\partial^2 L}{\partial\sigma^2\partial\boldsymbol{\beta}})^T\right|_{\boldsymbol{\theta}_0} \\ &= \left.\frac{1}{(\sigma^2)^2}\mathbb{E}\{\boldsymbol{X}^T(\boldsymbol{A}\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\beta})\}\right|_{\boldsymbol{\theta}_0} \\ &= \frac{1}{(\sigma_0^2)^2}\mathbb{E}(\boldsymbol{X}^T\boldsymbol{\epsilon}) = \mathbf{0} \end{aligned} \tag{A.8}$$

$$-\left.\frac{\partial^2 L}{\partial \boldsymbol{\beta}\partial\rho}\right|_{\boldsymbol{\theta}} = \left.(\frac{\partial^2 L}{\partial\rho\partial\boldsymbol{\beta}})^T\right|_{\boldsymbol{\theta}} = \frac{1}{\sigma_0^2}\mathbb{E}\{\boldsymbol{X}^T\boldsymbol{W}\boldsymbol{y}\} \tag{A.9}$$

$$\begin{aligned} -\left.\frac{\partial^2 L}{\partial(\sigma^2)^2}\right|_{\boldsymbol{\theta}_0} &= -\frac{n}{2(\sigma^2)^2} + \left.\frac{1}{(\sigma^2)^3}\mathbb{E}\{(\boldsymbol{A}\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\beta})^T(\boldsymbol{A}\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\beta})\}\right|_{\boldsymbol{\theta}_0} \\ &= -\frac{n}{2(\sigma_0^2)^2} + \frac{1}{(\sigma_0^2)^3}\mathbb{E}(\boldsymbol{\epsilon}^T\boldsymbol{\epsilon}) = \frac{n}{2(\sigma_0^2)^2} \end{aligned} \tag{A.10}$$

$$\begin{aligned} -\left.\frac{\partial^2 L}{\partial\sigma^2\partial\rho}\right|_{\boldsymbol{\theta_0}} &= \left.\frac{\partial^2 L}{\partial\rho\partial\sigma^2}\right|_{\boldsymbol{\theta_0}} = \frac{\mathbb{E}\{(\boldsymbol{A}_0\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\beta}_0)^T\boldsymbol{W}\boldsymbol{y}\}}{(\sigma_0^2)^2} \\ &= \frac{\mathbb{E}[\boldsymbol{\epsilon}^T\boldsymbol{W}\boldsymbol{y}]}{(\sigma_0^2)^2} = \frac{tr(\boldsymbol{W}\boldsymbol{A}_0^{-1})}{\sigma_0^2} \end{aligned} \tag{A.11}$$

$$\begin{aligned} -\left.\frac{\partial^2 L}{\partial\rho^2}\right|_{\boldsymbol{\theta_0}} &= tr(\boldsymbol{A}_0^{-1}\boldsymbol{W}\boldsymbol{A}_0^{-1}\boldsymbol{W}) + tr((\boldsymbol{A}_0^{-1})^T\boldsymbol{W}^T\boldsymbol{W}\boldsymbol{A}_0^{-1}) \\ &\quad + \frac{1}{\sigma_0^2}\boldsymbol{\beta}^T\boldsymbol{X}^T(\boldsymbol{A}_0^{-1})^T\boldsymbol{W}^T\boldsymbol{W}\boldsymbol{A}_0^{-1}\boldsymbol{X}\boldsymbol{\beta} \end{aligned} \tag{A.12}$$

Let $\boldsymbol{S}_0 = \nabla^2 L|_{\boldsymbol{\theta}_0}$, thus,

$$\begin{aligned}
\boldsymbol{S}_0 &= -\left[\begin{array}{ccc}
\frac{\partial^2 L}{\partial \boldsymbol{\beta}^2}_{p\times p} & \frac{\partial^2 L}{\partial \boldsymbol{\beta}\partial\sigma^2}_{p\times 1} & \frac{\partial^2 L}{\partial \boldsymbol{\beta}\partial\rho}_{p\times 1} \\
\frac{\partial^2 L}{\partial\sigma^2\partial\boldsymbol{\beta}}_{1\times p} & \frac{\partial^2 L}{\partial(\sigma^2)^2}_{1\times 1} & \frac{\partial^2 L}{\partial\sigma^2\partial\rho}_{1\times 1} \\
\frac{\partial^2 L}{\partial\rho\partial\boldsymbol{\beta}}_{1\times p} & \frac{\partial^2 L}{\partial\rho\partial\sigma^2}_{1\times 1} & \frac{\partial^2 L}{\partial\rho^2}_{1\times 1}
\end{array}\right]\Bigg|_{\boldsymbol{\theta}_0} \\
&= \left[\begin{array}{ccc}
\frac{\mathbb{E}\{\boldsymbol{X}^T\boldsymbol{X}\}}{\sigma_0^2}_{p\times p} & \boldsymbol{0}_{p\times 1} & \frac{1}{\sigma_0^2}\mathbb{E}\{\boldsymbol{X}^T\boldsymbol{W}\boldsymbol{y}\}_{p\times 1} \\
\boldsymbol{0}_{1\times p} & \frac{n}{2(\sigma_0^2)^2}_{1\times 1} & \frac{tr(\boldsymbol{W}\boldsymbol{A}_0^{-1})}{\sigma_0^2}_{1\times 1} \\
\frac{1}{\sigma_0^2}\mathbb{E}\{\boldsymbol{y}^T\boldsymbol{W}^T\boldsymbol{X}\}_{1\times p} & \frac{tr(\boldsymbol{W}\boldsymbol{A}_0^{-1})}{\sigma_0^2}_{1\times 1} & T_{1\times 1}
\end{array}\right]
\end{aligned} \tag{A.13}$$

where, $T = tr(\boldsymbol{A}_0^{-1}\boldsymbol{W}\boldsymbol{A}_0^{-1}\boldsymbol{W}) + tr((\boldsymbol{A}_0^{-1})^T\boldsymbol{W}^T\boldsymbol{W}\boldsymbol{A}_0^{-1})$

$$+ \frac{1}{\sigma_0^2}\boldsymbol{\beta}^T\boldsymbol{X}^T(\boldsymbol{A}_0^{-1})^T\boldsymbol{W}^T\boldsymbol{W}\boldsymbol{A}_0^{-1}\boldsymbol{X}\boldsymbol{\beta}$$

$$= 2tr(\boldsymbol{A}_0^{-1}\boldsymbol{W}\boldsymbol{A}_0^{-1}\boldsymbol{W}) + \frac{1}{\sigma_0^2}\boldsymbol{\beta}^T\boldsymbol{X}^T(\boldsymbol{A}_0^{-1})^T\boldsymbol{W}^T\boldsymbol{W}\boldsymbol{A}_0^{-1}\boldsymbol{X}\boldsymbol{\beta}$$

(given $\boldsymbol{W}$ is symmetric by Assumption A1)

To show $\boldsymbol{S}_0$ is positive definite, we only need to show all its leading principal minors are positive. For the first p leading principle minors, consider a nonzero vector $\boldsymbol{z}_{[p]}$,

$$\boldsymbol{z}^T E(\boldsymbol{X}\boldsymbol{X}^T)\boldsymbol{z} = E((\boldsymbol{X}^T\boldsymbol{z})^T\boldsymbol{X}^T\boldsymbol{z})$$

$$> 0, \text{ (given columns of } \boldsymbol{X} \text{ are linearly independent)}$$

This means the first p order leading principal minor of $\boldsymbol{S}$ are positive. And this also means that $|\frac{\mathbb{E}\{\boldsymbol{X}^T\boldsymbol{X}\}}{\sigma_0^2}| > 0$. Denote $\boldsymbol{M}$ as the upper left $(p+1)\times(p+1)$ part of $\boldsymbol{S}_0$. Then $|\boldsymbol{M}| = |\frac{\mathbb{E}\{\boldsymbol{X}^T\boldsymbol{X}\}}{\sigma_0^2}|\frac{n}{2(\sigma_0^2)^2}_{1\times 1} > 0$. Thus, the $(p+1)$-th order leading principal minor of $\boldsymbol{S}_0$ is also positive.

Next we show the last leading principle minor of $\boldsymbol{S}_0$, i.e., the determine of $\boldsymbol{S}_0$ is also positive. We calculate the determinant along the $(p+1)$-th row.

$$
\begin{aligned}
Det = & \frac{n}{2(\sigma_0)^2}\begin{vmatrix} \frac{\mathbb{E}\{\boldsymbol{X}^T\boldsymbol{X}\}}{\sigma_0^2}_{3\times 3} & \frac{1}{\sigma_0^2}\mathbb{E}\{\boldsymbol{X}^T\boldsymbol{W}\boldsymbol{y}\}_{3\times 1} \\ \frac{1}{\sigma_0^2}\mathbb{E}\{\boldsymbol{y}^T\boldsymbol{W}^T\boldsymbol{X}\}_{1\times 3} & T_{1\times 1} \end{vmatrix} \\
& - \frac{tr(\boldsymbol{W}\boldsymbol{A}_0^{-1})}{\sigma_0^2}\begin{vmatrix} \frac{\mathbb{E}\{\boldsymbol{X}^T\boldsymbol{X}\}}{\sigma_0^2}_{3\times 3} & \boldsymbol{0}_{3\times 1} \\ \frac{1}{\sigma_0^2}\mathbb{E}\{\boldsymbol{y}^T\boldsymbol{W}^T\boldsymbol{X}\}_{1\times 3} & \frac{tr(\boldsymbol{W}\boldsymbol{A}_0^{-1})}{\sigma_0^2}_{1\times 1} \end{vmatrix} \\
= & \frac{n}{2(\sigma_0^2)^2}\frac{|\mathbb{E}(\boldsymbol{X}^T\boldsymbol{X})|}{\sigma_0^2}T - \frac{tr(\boldsymbol{W}\boldsymbol{A}_0^{-1})}{\sigma_0^2}\frac{tr(\boldsymbol{W}\boldsymbol{A}_0^{-1})}{\sigma_0^2}\frac{|\mathbb{E}(\boldsymbol{X}^T\boldsymbol{X})|}{\sigma_0^2} \\
= & [ntr(\boldsymbol{A}_0^{-1}\boldsymbol{W}\boldsymbol{A}_0^{-1}\boldsymbol{W}) - tr(\boldsymbol{W}\boldsymbol{A}_0^{-1})tr(\boldsymbol{W}\boldsymbol{A}_0^{-1})]\frac{\mathbb{E}(\boldsymbol{X}^T\boldsymbol{X})}{(\sigma_0^2)^3} \\
& + \frac{1}{\sigma_0^2}\boldsymbol{\beta}^T\boldsymbol{X}^T(\boldsymbol{A}_0^{-1})^T\boldsymbol{W}^T\boldsymbol{W}\boldsymbol{A}_0^{-1}\boldsymbol{X}\boldsymbol{\beta}\frac{n\mathbb{E}(\boldsymbol{X}^T\boldsymbol{X})}{2(\sigma_0^2)^3}
\end{aligned}
\tag{A.14}
$$

Easy to see that the second part of of the last expression in the above equation is non-negative. And we only need to show $ntr(\boldsymbol{A}_0^{-1}\boldsymbol{W}\boldsymbol{A}_0^{-1}\boldsymbol{W}) - tr(\boldsymbol{W}\boldsymbol{A}_0^{-1})tr(\boldsymbol{W}\boldsymbol{A}_0^{-1}) > 0$. To prove this, first note that all eigenvalues of $\boldsymbol{W}$ are real (given $\boldsymbol{W}$ is symmetric) and if λ is an eigenvalue of $\boldsymbol{W}$, then λ^k is an eigenvalue of $\boldsymbol{W}^k$ for $k = 1, 2, \cdots$. Also, $\boldsymbol{A}^{-1}\boldsymbol{W} = (\sum_{k=0}^{\infty}\rho^k\boldsymbol{W}^k)\boldsymbol{W} = \sum_{k=0}^{\infty}\rho^k\boldsymbol{W}^{k+1}$. Thus, all the eigenvalues of $\boldsymbol{A}^{-1}\boldsymbol{W}$ are real. Let $\lambda_i, i = 1, \cdots, n$ be eigenvalues of $\boldsymbol{A}_0^{-1}\boldsymbol{W}$, According to Lemma 2.1 in [20], λ_i cannot be all identical. By Cauchy-Schwarz inequality, we have

$$
\begin{aligned}
& (1\cdot\lambda_1 + 1\cdot\lambda_2 + \cdots + 1\cdot\lambda_n)^2 < (1+1+\cdots+1)(\lambda_1^2+\lambda_2^2+\cdots+\lambda_n^2) \\
& \text{i.e., } (\sum\lambda_i)^2 = [tr(\boldsymbol{A}_0^{-1}\boldsymbol{W})]^2 < n\cdot\sum\lambda_i^2 = n\cdot tr[(\boldsymbol{A}_0^{-1}\boldsymbol{W})^2]
\end{aligned}
\tag{A.15}
$$

Thus, we prove the determinant of $\boldsymbol{S}_0$ is positive and finish the proof of $\boldsymbol{S}_0$ is positive definite.

A.1.4 Verification of assumption FA4

Assumption FA4 Assume $\mathbb{E}\|\nabla\ell(\theta;Z)\|^2 \le C\,(1+\|\theta\|^2)$ for some C and $\mathbb{E}\,\|\nabla\ell(\theta;Z) - \nabla\ell\,(\theta_0;Z)\|^2 \le \delta\,(\|\theta-\theta_0\|)$ for some $\delta(\cdot)$ with $\delta(x)\to 0$ as $x\to 0$.

First, we prove $\mathbb{E}\|\nabla\ell(\boldsymbol{\theta})\|^2 \leq C(1+\|\boldsymbol{\theta}\|^2)$ for some constant C. We prove this by showing that $\mathbb{E}\|\nabla\ell_{\boldsymbol{\beta}}\|^2 \leq C_1(1+\|\boldsymbol{\theta}\|^2)$, $\mathbb{E}\|\nabla\ell_{\sigma^2}\|^2 \leq C_2(1+\|\boldsymbol{\theta}\|^2)$, and $\mathbb{E}\|\nabla\ell_{\rho}\|^2 \leq C_3(1+\|\boldsymbol{\theta}\|^2)$, here, C_1, C_2, and C_3 are constants.

Prove $\mathbb{E}\|\nabla\ell_{\boldsymbol{\beta}}\|^2 \leq C_1(1+\|\boldsymbol{\theta}\|^2)$

Assume σ^2 is bounded below by $\tilde{\sigma}^2$ and $\|\boldsymbol{\beta}\|$ is bounded above.

$$\begin{aligned}
\mathbb{E}\|\nabla\ell_{\boldsymbol{\beta}}\|^2 &\leq \frac{1}{\tilde{\sigma}^2}\mathbb{E}[(\boldsymbol{X}^T\boldsymbol{A}\boldsymbol{y}-\boldsymbol{X}^T\boldsymbol{X}\boldsymbol{\beta})^T(\boldsymbol{X}^T\boldsymbol{A}\boldsymbol{y}-\boldsymbol{X}^T\boldsymbol{X}\boldsymbol{\beta})] \\
&\leq \frac{2}{\tilde{\sigma}^2}\mathbb{E}[(\boldsymbol{X}^T\boldsymbol{A}\boldsymbol{y})^T(\boldsymbol{X}^T\boldsymbol{A}\boldsymbol{y})+(\boldsymbol{X}^T\boldsymbol{X}\boldsymbol{\beta})^T(\boldsymbol{X}^T\boldsymbol{X}\boldsymbol{\beta})], \text{ by triangle inequality} \\
&= K_1\mathbb{E}[\boldsymbol{y}^T\boldsymbol{A}\boldsymbol{X}\boldsymbol{X}^T\boldsymbol{A}\boldsymbol{y}+\boldsymbol{\beta}^T\boldsymbol{X}^T\boldsymbol{X}\boldsymbol{X}^T\boldsymbol{X}\boldsymbol{\beta}], \text{ let } K_1=\frac{2}{\tilde{\sigma}^2} \\
&= K_1[Tr(\boldsymbol{A}\boldsymbol{X}\boldsymbol{X}^T\boldsymbol{A}\sigma_0^2(\boldsymbol{A}_0^{-1})^2)+(\boldsymbol{A}_0^{-1}\boldsymbol{X}\boldsymbol{\beta}_0)^T\boldsymbol{A}\boldsymbol{X}\boldsymbol{X}^T\boldsymbol{A}(\boldsymbol{A}_0^{-1}\boldsymbol{X}\boldsymbol{\beta}_0) \\
&\qquad +\boldsymbol{\beta}^T\boldsymbol{X}^T\boldsymbol{X}\boldsymbol{X}^T\boldsymbol{X}\boldsymbol{\beta}]
\end{aligned} \tag{A.16}$$

Here, $\sigma_0^2(\boldsymbol{A}_0^{-1})^2, \boldsymbol{A}_0^{-1}\boldsymbol{X}\boldsymbol{\beta}_0, \boldsymbol{X}^T\boldsymbol{X}\boldsymbol{X}^T\boldsymbol{X}$ are constants, and $\boldsymbol{A}\boldsymbol{X}\boldsymbol{X}^T\boldsymbol{A}=(\boldsymbol{I}-\rho\boldsymbol{W})\boldsymbol{X}\boldsymbol{X}^T(\boldsymbol{I}-\rho\boldsymbol{W})$ is a second order function of ρ, so we can bound $Tr(\boldsymbol{A}\boldsymbol{X}\boldsymbol{X}^T\boldsymbol{A}\sigma_0^2(\boldsymbol{A}_0^{-1})^2)+(\boldsymbol{A}_0^{-1}\boldsymbol{X}\boldsymbol{\beta}_0)^T\boldsymbol{A}\boldsymbol{X}\boldsymbol{X}^T\boldsymbol{A}(\boldsymbol{A}_0^{-1}\boldsymbol{X}\boldsymbol{\beta}_0)$ with $K_2(1+\|\boldsymbol{\theta}\|^2)$. Also, easy to see we can bound $\boldsymbol{\beta}^T\boldsymbol{X}^T\boldsymbol{X}\boldsymbol{X}^T\boldsymbol{X}\boldsymbol{\beta}$ with $K_3(1+\|\boldsymbol{\theta}\|^2)$. Thus, $\mathbb{E}\|\nabla\ell_{\boldsymbol{\beta}}\|^2 \leq C_1(1+\|\boldsymbol{\theta}\|^2)$.

Prove $\mathbb{E}(\nabla\ell_{\sigma^2})^2 \leq C_2(1+\|\boldsymbol{\theta}\|^2)$

$$\begin{aligned}
\mathbb{E}(\nabla\ell_{\sigma^2})^2 &= \mathbb{E}[-\frac{n}{2\sigma^2}+\frac{1}{2(\sigma^2)^2}(\boldsymbol{A}\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\beta})^T(\boldsymbol{A}\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\beta})]^2 \\
&\leq 2\frac{n^2}{(2\sigma^2)^2}+2(\frac{1}{2(\sigma^2)^2})^2\mathbb{E}[(\boldsymbol{A}\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\beta})^T(\boldsymbol{A}\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\beta})]^2 \\
&\leq K_4+K_5\mathbb{E}[(\boldsymbol{A}\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\beta})^T(\boldsymbol{A}\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\beta})]^2, \text{ since } \sigma^2 \geq \sigma_0^2 \\
&\leq K_4+K_5\mathbb{E}[2\|\boldsymbol{A}\boldsymbol{y}\|^2+2\|\boldsymbol{X}\boldsymbol{\beta}\|^2]^2 \\
&\leq K_4+K_5\mathbb{E}[8\|\boldsymbol{A}\boldsymbol{y}\|^4+8\|\boldsymbol{X}\boldsymbol{\beta}\|^4]
\end{aligned} \tag{A.17}$$

Clearly, $\mathbb{E}[8\|\boldsymbol{Ay}\|^4 + 8\|\boldsymbol{X\beta}\|^4]$ is a 4th order function of ρ and $\boldsymbol{\beta}$, ρ is between -1 and 1 and $\|\boldsymbol{\beta}\|$ is bounded above, so that $\mathbb{E}(\nabla\ell_{\sigma^2})^2 \leq C_2(1 + \|\boldsymbol{\theta}\|^2)$.

Prove $\mathbb{E}(\boldsymbol{\nabla\ell_\rho}) \leq \boldsymbol{C_3(1 + \|\theta\|^2)}$

$$\begin{aligned}\mathbb{E}(\nabla\ell_\rho)^2 &= \mathbb{E}[-tr(\boldsymbol{A}^{-1}\boldsymbol{W}) + \frac{(\boldsymbol{Ay} - \boldsymbol{X\beta})^T\boldsymbol{Wy}}{\sigma^2}]^2 \\ &\leq 2[tr(\boldsymbol{A}^{-1}\boldsymbol{W}]^2 + K_6\mathbb{E}(\boldsymbol{y}^T\boldsymbol{AWy} - \boldsymbol{\beta}^T\boldsymbol{X}^T\boldsymbol{Wy})^2, \text{ since } \sigma^2 \geq \tilde{\sigma}^2 \\ &\leq 2[tr(\boldsymbol{A}^{-1}\boldsymbol{W}]^2 + 2K_6\mathbb{E}[(\boldsymbol{y}^T\boldsymbol{AWy})^2 + (\boldsymbol{\beta}^T\boldsymbol{X}^T\boldsymbol{Wy})^2]\end{aligned} \tag{A.18}$$

Note that,

$$\begin{aligned}|tr(\boldsymbol{A}^{-1}\boldsymbol{W})| &= |tr(\sum_{k=0}^{\infty}\rho^k\boldsymbol{W}^{k+1})| \leq n\sum_{k=0}^{\infty}|\rho|^k, \\ &\qquad\qquad \text{since elements of } \boldsymbol{W}^k \text{ are positive and less than 1} \\ &= \frac{n}{1-|\rho|} \leq \frac{n}{1-\tilde{\rho}} = K_7, \tilde{\rho} = min(|\rho|).\end{aligned} \tag{A.19}$$

Also, $(\boldsymbol{y}^T\boldsymbol{AWy})^2$ and $(\boldsymbol{\beta}^T\boldsymbol{X}^T\boldsymbol{Wy})^2$ are 2nd order polynomial function of ρ and $\boldsymbol{\beta}$, thus, $\mathbb{E}(\nabla\ell_\rho) \leq C_3(1 + \|\boldsymbol{\theta}\|^2)$.

Then we prove $\mathbb{E}\|\nabla\ell(\boldsymbol{\theta}, Z) - \nabla\ell(\boldsymbol{\theta_0}, Z)\|^2 \leq \delta(\|\boldsymbol{\theta} - \boldsymbol{\theta_0}\|)$ for some $\delta(\cdot)$ with $\delta(x) \to 0$ as $x \to 0$. Let $\delta_1(x), \delta_2(x)$ and $\delta_3(x)$ are functions go to 0 as $x \to 0$. We prove this by show that (1) $\mathbb{E}\|\nabla\ell_{\boldsymbol{\beta}}(\boldsymbol{\theta}) - \nabla\ell_{\boldsymbol{\beta}}(\boldsymbol{\theta_0})\|^2 \leq \delta_1(\|\boldsymbol{\theta} - \boldsymbol{\theta_0}\|)$, (2) $\mathbb{E}\|\nabla\ell_{\sigma^2}(\boldsymbol{\theta}) - \nabla\ell_{\sigma^2}(\boldsymbol{\theta_0})\|^2 \leq \delta_2(\|\boldsymbol{\theta} - \boldsymbol{\theta_0}\|)$ and (3) $\mathbb{E}\|\nabla\ell_\rho(\boldsymbol{\theta}) - \nabla\ell_\rho(\boldsymbol{\theta_0})\|^2 \leq \delta_3(\|\boldsymbol{\theta} - \boldsymbol{\theta_0}\|)$.

Prove $\mathbb{E}\|\nabla\ell_{\boldsymbol{\beta}}(\boldsymbol{\theta}) - \nabla\ell_{\boldsymbol{\beta}}(\boldsymbol{\theta_0})\|^2 \leq \delta_1(\|\boldsymbol{\theta} - \boldsymbol{\theta_0}\|)$

$$\begin{aligned}
\nabla\ell_{\boldsymbol{\beta}}(\boldsymbol{\theta}) - \nabla\ell_{\boldsymbol{\beta}}(\boldsymbol{\theta_0}) &= \frac{1}{\sigma^2}(\boldsymbol{X}^T\boldsymbol{A}\boldsymbol{y} - \boldsymbol{X}^T\boldsymbol{X}\boldsymbol{\beta}) - \frac{1}{\sigma_0^2}(\boldsymbol{X}^T\boldsymbol{A}_0\boldsymbol{y} - \boldsymbol{X}^T\boldsymbol{X}\boldsymbol{\beta}_0) \\
&= (\frac{1}{\sigma^2}\boldsymbol{X}^T\boldsymbol{A}\boldsymbol{y} - \frac{1}{\sigma_0^2}\boldsymbol{X}^T\boldsymbol{A}_0\boldsymbol{y}) - (\frac{1}{\sigma^2}\boldsymbol{X}^T\boldsymbol{X}\boldsymbol{\beta} - \frac{1}{\sigma_0^2}\boldsymbol{X}^T\boldsymbol{X}\boldsymbol{\beta}_0) \\
&= [(\frac{1}{\sigma^2} - \frac{1}{\sigma_0^2})\boldsymbol{X}^T\boldsymbol{A}\boldsymbol{y} + \frac{1}{\sigma_0^2}\boldsymbol{X}^T(\boldsymbol{A} - \boldsymbol{A}_0)\boldsymbol{y}] \\
&\qquad - [(\frac{1}{\sigma^2} - \frac{1}{\sigma_0^2})\boldsymbol{X}^T\boldsymbol{X}\boldsymbol{\beta} + \frac{1}{\sigma_0^2}\boldsymbol{X}^T\boldsymbol{X}(\boldsymbol{\beta} - \boldsymbol{\beta}_0)]
\end{aligned} \tag{A.20}$$

$$\begin{aligned}
&\mathbb{E}\|\nabla\ell_{\boldsymbol{\beta}}(\boldsymbol{\theta}) - \nabla\ell_{\boldsymbol{\beta}}(\boldsymbol{\theta_0})\|^2 \\
&\leq 2\mathbb{E}\|(\frac{1}{\sigma^2} - \frac{1}{\sigma_0^2})\boldsymbol{X}^T\boldsymbol{A}\boldsymbol{y} + \frac{1}{\sigma_0^2}\boldsymbol{X}^T(\boldsymbol{A} - \boldsymbol{A}_0)\boldsymbol{y}\|^2 \\
&\qquad + 2\|(\frac{1}{\sigma^2} - \frac{1}{\sigma_0^2})\boldsymbol{X}^T\boldsymbol{X}\boldsymbol{\beta} + \frac{1}{\sigma_0^2}\boldsymbol{X}^T\boldsymbol{X}(\boldsymbol{\beta} - \boldsymbol{\beta}_0)\|^2 \\
&\leq 4(\frac{1}{\sigma^2} - \frac{1}{\sigma_0^2})^2\mathbb{E}\|\boldsymbol{X}^T\boldsymbol{A}\boldsymbol{y}\|^2 + \frac{4}{(\sigma_0^2)^2}\mathbb{E}\|\frac{1}{\sigma_0^2}\boldsymbol{X}^T(\rho - \rho_0)\boldsymbol{y}\|^2 \\
&\qquad + 4(\frac{1}{\sigma^2} - \frac{1}{\sigma_0^2})^2\mathbb{E}\|\boldsymbol{X}^T\boldsymbol{X}\boldsymbol{\beta}\|^2 + \frac{4}{(\sigma_0^2)^2}\mathbb{E}\|\boldsymbol{X}^T\boldsymbol{X}(\boldsymbol{\beta} - \boldsymbol{\beta}_0)\|^2 \\
&\leq 4K_8\frac{(\sigma_0^2 - \sigma^2)^2}{\sigma^2\sigma_0^2} + \frac{4K_9}{(\sigma_0^2)^2}(\rho - \rho_0)^2 + \leq 4K_{10}\frac{(\sigma_0^2 - \sigma^2)^2}{\sigma^2\sigma_0^2} + \frac{4K_{11}}{(\sigma_0^2)^2}\|\boldsymbol{\beta} - \boldsymbol{\beta}_0\|^2
\end{aligned} \tag{A.21}$$

K_8 is the maximum of $E\|\boldsymbol{X}^T\boldsymbol{A}\boldsymbol{y}\|^2$, $E\|\boldsymbol{X}^T\boldsymbol{A}\boldsymbol{y}\|^2$ is a smooth function of ρ, and ρ is in a closed set, thus, maximum exist and is finite. $K_9 = \mathbb{E}\|\frac{1}{\sigma_0^2}\boldsymbol{X}^T\boldsymbol{y}\|^2$. K_10 is the maximum of $\mathbb{E}\|\boldsymbol{X}^T\boldsymbol{X}\boldsymbol{\beta}\|^2$ and this maximum exist and is finite since $\|\boldsymbol{\beta}\|$ is bounded above.

Note,

$$\|\boldsymbol{X}^T\boldsymbol{X}(\boldsymbol{\beta} - \boldsymbol{\beta}_0)\|^2 = (\boldsymbol{\beta} - \boldsymbol{\beta}_0)\boldsymbol{X}^T\boldsymbol{X}\boldsymbol{X}^T\boldsymbol{X}(\boldsymbol{\beta} - \boldsymbol{\beta}_0) \leq \lambda_{max}\|\boldsymbol{\beta} - \boldsymbol{\beta}_0\| \tag{A.22}$$

Here, λ_{max} is the largest eigenvalue of $\boldsymbol{X}^T\boldsymbol{X}\boldsymbol{X}^T\boldsymbol{X}$ and $K_11 = \lambda_{max}$. Thus, easy to show $\mathbb{E}\|\nabla\ell_{\boldsymbol{\beta}}(\boldsymbol{\theta}) - \nabla\ell_{\boldsymbol{\beta}}(\boldsymbol{\theta_0})\|^2 \leq \delta_1(\|\boldsymbol{\theta} - \boldsymbol{\theta_0}\|) = C_4\|(\|\boldsymbol{\theta} - \boldsymbol{\theta_0}\|)\|^2$ for some constant C_4.

Prove $\mathbb{E}\|\nabla\ell_{\sigma^2}(\theta) - \nabla\ell_{\sigma^2}(\theta_0)\|^2 \leq \delta_2(\|\theta - \theta_0\|)$

$$
\begin{aligned}
&\nabla\ell_{\sigma^2}(\boldsymbol{\theta}) - \nabla\ell_{\sigma^2}(\boldsymbol{\theta_0}) \\
&= -\frac{n}{2\sigma^2} + \frac{1}{2(\sigma^2)^2}(\boldsymbol{Ay} - \boldsymbol{X\beta})^T(\boldsymbol{Ay} - \boldsymbol{X\beta}) + \frac{n}{2\sigma_0^2} \\
&\qquad - \frac{1}{2(\sigma^2)^2}(\boldsymbol{A}_0\boldsymbol{y} - \boldsymbol{X\beta}_0)^T(\boldsymbol{A}_0\boldsymbol{y} - \boldsymbol{X\beta}_0),\ \text{let}\sigma^2 = \gamma, \sigma_0^2 = \gamma_0 \\
&= \frac{n}{2}(\gamma_0 - \gamma) + \frac{1}{2}\gamma^2(\boldsymbol{Ay} - \boldsymbol{X\beta})^T(\boldsymbol{Ay} - \boldsymbol{X\beta}) - \frac{1}{2}\gamma_0^2(\boldsymbol{Ay} - \boldsymbol{X\beta})^T(\boldsymbol{Ay} - \boldsymbol{X\beta}) \\
&\qquad + \frac{1}{2}\gamma_0^2(\boldsymbol{Ay} - \boldsymbol{X\beta})^T(\boldsymbol{Ay} - \boldsymbol{X\beta}) - \frac{1}{2}\gamma_0^2(\boldsymbol{A}_0\boldsymbol{y} - \boldsymbol{X\beta}_0)^T(\boldsymbol{A}_0\boldsymbol{y} - \boldsymbol{X\beta}_0) \\
&= \frac{n}{2}(\gamma_0 - \gamma) + \frac{1}{2}(\gamma^2 - \gamma_0^2)(\boldsymbol{Ay} - \boldsymbol{X\beta})^T(\boldsymbol{Ay} - \boldsymbol{X\beta}) \\
&\qquad + \frac{1}{2}\gamma_0^2[(\boldsymbol{Ay} - \boldsymbol{X\beta})^T(\boldsymbol{Ay} - \boldsymbol{X\beta}) - (\boldsymbol{A}_0\boldsymbol{y} - \boldsymbol{X\beta}_0)^T(\boldsymbol{A}_0\boldsymbol{y} - \boldsymbol{X\beta}_0)].
\end{aligned}
\tag{A.23}
$$

Thus,

$$
\begin{aligned}
&\mathbb{E}[\|\nabla\ell_{\sigma^2}(\boldsymbol{\theta}) - \nabla\ell_{\sigma^2}(\boldsymbol{\theta_0})\|^2] \\
&\leq 3[\frac{n}{2}(\gamma_0 - \gamma)]^2 + 3\mathbb{E}[\frac{1}{2}(\gamma^2 - \gamma_0^2)(\boldsymbol{Ay} - \boldsymbol{X\beta})^T(\boldsymbol{Ay} - \boldsymbol{X\beta})]^2 \\
&\quad + 3\mathbb{E}\{\frac{1}{2}\gamma_0^2[(\boldsymbol{Ay} - \boldsymbol{X\beta})^T(\boldsymbol{Ay} - \boldsymbol{X\beta}) - (\boldsymbol{A}_0\boldsymbol{y} - \boldsymbol{X\beta}_0)^T(\boldsymbol{A}_0\boldsymbol{y} - \boldsymbol{X\beta}_0)]\}^2
\end{aligned}
\tag{A.24}
$$

And,

$$
(\gamma_0 - \gamma)^2 = (\frac{(\sigma_0^2)^2 - (\sigma^2)^2}{(\sigma_0^2)^2(\sigma^2)^2})^2 \leq K_{13}\|\boldsymbol{\theta} - \boldsymbol{\theta_0}\| \tag{A.25}
$$

$$
(\gamma^2 - \gamma_0^2)^2 = \frac{(\sigma_0^2 + \sigma^2)^2(\sigma_0^2 - \sigma^2)^2}{(\sigma^2)^4(\sigma_0^2)^4} \leq K_{14}(\sigma^2 - \sigma_0^2)^2 \tag{A.26}
$$

$\mathbb{E}[(\boldsymbol{Ay}-\boldsymbol{X\beta})^T(\boldsymbol{Ay}-\boldsymbol{X\beta})]^2$ can be bounded above, since it is a continuous function of ρ and $\boldsymbol{\beta}$ and both ρ and $\boldsymbol{\beta}$ are in a closed set.

$$\begin{aligned}
&[(\boldsymbol{Ay}-\boldsymbol{X\beta})^T(\boldsymbol{Ay}-\boldsymbol{X\beta})-(\boldsymbol{A}_0\boldsymbol{y}-\boldsymbol{X\beta}_0)^T(\boldsymbol{A}_0\boldsymbol{y}-\boldsymbol{X\beta}_0)]]^2\\
&=[\boldsymbol{y}^T\boldsymbol{AAy}-\boldsymbol{y}^T\boldsymbol{A}_0\boldsymbol{A}_0\boldsymbol{y}-(2\boldsymbol{y}^T\boldsymbol{AX\beta}-2\boldsymbol{y}^T\boldsymbol{A}_0\boldsymbol{X\beta}_0)\\
&\qquad+\boldsymbol{\beta}^T\boldsymbol{X}^T\boldsymbol{X\beta}-\boldsymbol{\beta}_0\boldsymbol{X}^T\boldsymbol{X\beta}_0]^2\\
&\le 3[\boldsymbol{y}^T(\boldsymbol{AA}-\boldsymbol{A}_0\boldsymbol{A}_0)\boldsymbol{y}]^2+12(\boldsymbol{y}^T\boldsymbol{AX\beta}-\boldsymbol{y}^T\boldsymbol{A}_0\boldsymbol{X\beta}_0)^2\\
&\qquad+3(\boldsymbol{\beta}^T\boldsymbol{X}^T\boldsymbol{X\beta}-\boldsymbol{\beta}_0\boldsymbol{X}^T\boldsymbol{X\beta}_0)^2
\end{aligned}\tag{A.27}$$

Note,

$$\begin{aligned}
&\mathbb{E}[\boldsymbol{y}^T(\boldsymbol{AA}-\boldsymbol{A}_0\boldsymbol{A}_0)\boldsymbol{y}]^2\\
&=Tr[(\boldsymbol{AA}-\boldsymbol{A}_0\boldsymbol{A}_0)\sigma_0^2\boldsymbol{A}_0\boldsymbol{A}_0]+(\boldsymbol{A}_0^{-1}\boldsymbol{X\beta}_0)^T(\boldsymbol{AA}-\boldsymbol{A}_0\boldsymbol{A}_0)(\boldsymbol{A}_0^{-1}\boldsymbol{X\beta}_0)
\end{aligned}\tag{A.28}$$

and

$$\begin{aligned}
\|\boldsymbol{AA}-\boldsymbol{A}_0\boldsymbol{A}_0\|&=\|(\boldsymbol{I}-\rho\boldsymbol{W})^2-(\boldsymbol{I}-\rho_0\boldsymbol{W})^2\|\\
&=\|-2(\rho-\rho_0)\boldsymbol{W}+(\rho-\rho_0)(\rho+\rho_0)\boldsymbol{W}^2\|^2\\
&\le 4(\rho-\rho_0)^2\|\boldsymbol{W}\|^2+4(\rho-\rho_0)^2\|\boldsymbol{W}^2\|^2\le K_{15}(\rho-\rho_0)^2
\end{aligned}\tag{A.29}$$

Thus, $E[\boldsymbol{y}^T(\boldsymbol{AA}-\boldsymbol{A}_0\boldsymbol{A}_0)\boldsymbol{y}]^2\le K_{16}\|\boldsymbol{\theta}-\boldsymbol{\theta_0}\|^2$. Note,

$$\begin{aligned}
&(\boldsymbol{y}^T\boldsymbol{AX\beta}-\boldsymbol{y}^T\boldsymbol{A}_0\boldsymbol{X\beta}_0)^2\\
&=(\boldsymbol{y}^T\boldsymbol{AX\beta}-\boldsymbol{y}^T\boldsymbol{A}_0\boldsymbol{X\beta}+\boldsymbol{y}^T\boldsymbol{A}_0\boldsymbol{X\beta}-\boldsymbol{y}^T\boldsymbol{A}_0\boldsymbol{X\beta}_0)^2\\
&\le 2(\boldsymbol{y}^T\boldsymbol{AX\beta}-\boldsymbol{y}^T\boldsymbol{A}_0\boldsymbol{X\beta})^2+2(\boldsymbol{y}^T\boldsymbol{A}_0\boldsymbol{X\beta}-\boldsymbol{y}^T\boldsymbol{A}_0\boldsymbol{X\beta}_0)^2\\
&=2[\boldsymbol{y}^T(\rho_0-\rho)\boldsymbol{WX\beta}]^2+2[\boldsymbol{y}^T\boldsymbol{A}_0\boldsymbol{X}(\boldsymbol{\beta}-\boldsymbol{\beta}_0)]^2\\
&\le K_{17}\|\boldsymbol{\theta}-\boldsymbol{\theta_0}\|^2,\text{ since }\|\beta\|\text{ is bounded above.}
\end{aligned}\tag{A.30}$$

Note,

$$
\begin{aligned}
&(\boldsymbol{\beta}^T\boldsymbol{X}^T\boldsymbol{X}\boldsymbol{\beta} - \boldsymbol{\beta}_0\boldsymbol{X}^T\boldsymbol{X}\boldsymbol{\beta}_0)^2 \\
&= [(\boldsymbol{\beta} - \boldsymbol{\beta}_0 + \boldsymbol{\beta}_0)^T\boldsymbol{X}^T\boldsymbol{X}(\boldsymbol{\beta} - \boldsymbol{\beta}_0 + \boldsymbol{\beta}_0) - \boldsymbol{\beta}_0^T\boldsymbol{X}^T\boldsymbol{X}\boldsymbol{\beta}_0]^2 \\
&= [(\boldsymbol{\beta} - \boldsymbol{\beta}_0)^T\boldsymbol{X}^T\boldsymbol{X}(\boldsymbol{\beta} - \boldsymbol{\beta}_0) + 2(\boldsymbol{\beta} - \boldsymbol{\beta}_0)^T\boldsymbol{X}^T\boldsymbol{X}\boldsymbol{\beta}_0]^2 \\
&\leq K_{18}\|\boldsymbol{\theta} - \boldsymbol{\theta_0}\|^4 + K_{19}\|\boldsymbol{\theta} - \boldsymbol{\theta_0}\|^3 + K_{20}\|\boldsymbol{\theta} - \boldsymbol{\theta_0}\|^2
\end{aligned}
\tag{A.31}
$$

Combined all the argument above, we proved that $\mathbb{E}\|\nabla\ell_{\sigma^2}(\boldsymbol{\theta}) - \nabla\ell_{\sigma^2}(\boldsymbol{\theta_0})\|^2 \leq \delta_2(\|\boldsymbol{\theta} - \boldsymbol{\theta_0}\|) = K_{21}\|\boldsymbol{\theta} - \boldsymbol{\theta_0}\|^4 + K_{22}\|\boldsymbol{\theta} - \boldsymbol{\theta_0}\|^3 + K_{23}\|\boldsymbol{\theta} - \boldsymbol{\theta_0}\|^2$.

Prove $\mathbb{E}\|\boldsymbol{\nabla\ell_\rho(\theta) - \nabla\ell_\rho(\theta_0)}\|^2 \leq \boldsymbol{\delta_3}(\|\boldsymbol{\theta} - \boldsymbol{\theta_0}\|)$

$$
\begin{aligned}
&\nabla\ell_\rho(\boldsymbol{\theta}) - \nabla\ell_\rho(\boldsymbol{\theta_0}) \\
&= -tr(\boldsymbol{A}^{-1}\boldsymbol{W}) + tr(\boldsymbol{A}_0^{-1}\boldsymbol{W}) + \frac{1}{\sigma^2}(\boldsymbol{A}\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})^T\boldsymbol{W}\boldsymbol{y} \\
&\qquad - \frac{1}{\sigma_0^2}(\boldsymbol{A}_0\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}_0)^T\boldsymbol{W}\boldsymbol{y}) \\
&= -tr[(\boldsymbol{A}^{-1} - \boldsymbol{A}_0^{-1})\boldsymbol{W}] + \{[(\frac{1}{\sigma^2} - \frac{1}{\sigma_0^2})\boldsymbol{A}\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})^T\boldsymbol{W}\boldsymbol{y} \\
&\qquad + \frac{1}{\sigma_0^2}[(\boldsymbol{A}\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})^T - (\boldsymbol{A}_0\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}_0)^T]\boldsymbol{W}\boldsymbol{y}\}
\end{aligned}
\tag{A.32}
$$

We have shown above $(\frac{1}{\sigma^2} - \frac{1}{\sigma_0^2})^2 \leq K_{13}\|\boldsymbol{\theta} - \boldsymbol{\theta_0}\|^2$. Also, $\|\boldsymbol{A}\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})^T\boldsymbol{W}\boldsymbol{y}\|^2$ is bounded above. Note,

$$
\begin{aligned}
&\mathbb{E}\{[(\boldsymbol{A}\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})^T - (\boldsymbol{A}_0\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}_0)^T]\boldsymbol{W}\boldsymbol{y}\}^2 \\
&= \mathbb{E}\{\boldsymbol{y}^T(\boldsymbol{A} - \boldsymbol{A}_0)\boldsymbol{W}\boldsymbol{y} - (\boldsymbol{\beta} - \boldsymbol{\beta}_0)^T\boldsymbol{X}^T\boldsymbol{W}\boldsymbol{y}\}^2 \\
&= \mathbb{E}\{\boldsymbol{y}^T(\rho_0 - \rho)\boldsymbol{W}^2\boldsymbol{y} - (\boldsymbol{\beta} - \boldsymbol{\beta}_0)^T\boldsymbol{X}^T\boldsymbol{W}\boldsymbol{y}\}^2 \\
&\leq 2(\rho - \rho_0)^2\mathbb{E}\|\boldsymbol{y}^T\boldsymbol{W}\boldsymbol{W}\boldsymbol{y}\|^2 + 2\mathbb{E}\|(\boldsymbol{\beta} - \boldsymbol{\beta}_0)^T\boldsymbol{X}^T\boldsymbol{W}\boldsymbol{y}\|^2 \\
&\leq 2K_{24}(\rho - \rho_0)^2 + 2\mathbb{E}[(\boldsymbol{\beta} - \boldsymbol{\beta}_0)^T\boldsymbol{X}^T\boldsymbol{W}\boldsymbol{y}\boldsymbol{y}^T\boldsymbol{W}^T\boldsymbol{X}(\boldsymbol{\beta} - \boldsymbol{\beta}_0)] \\
&\leq 2K_{24}(\rho - \rho_0)^2 + 2\tilde{\lambda_{max}}\|\boldsymbol{\beta} - \boldsymbol{\beta}_0\|^2,
\end{aligned}
\tag{A.33}
$$

$\tilde{\lambda}_{max}$ is the max eigenvalue of $\boldsymbol{X}^T\boldsymbol{W}\boldsymbol{y}\boldsymbol{y}^T\boldsymbol{W}^T\boldsymbol{X}$

$$
\begin{aligned}
&\{tr[(\boldsymbol{A}^{-1} - \boldsymbol{A}_0^{-1})\boldsymbol{W}]\}^2 \\
&= \{tr[\sum_{k=0}^{\infty}(\rho_0^k - \rho^k)\boldsymbol{W}^{k+1}]\}^2 \\
&\leq \{tr[\sum_{k=0}^{\infty}|\rho_0^k - \rho^k|\boldsymbol{W}^{k+1}]\}^2 \\
&= \{tr[\sum_{k=0}^{\infty}|\rho_0 - \rho||\rho_0^{k-1} + \rho_0^{k-2}\rho + \cdots + \rho^{k-1}|\boldsymbol{W}^{k+1}]\}^2 \\
&\leq \{tr[\sum_{k=0}^{\infty}|\rho_0 - \rho|k\tilde{\rho}\boldsymbol{W}^{k+1}]\}^2, \tilde{\rho} = min(|\rho|) \\
&\leq K_{25}(\rho - \rho_0)^2, \sum_{k=0}^{\infty} k\tilde{\rho}^{k-1} \text{ converges since } \tilde{\rho} < 1
\end{aligned}
\tag{A.34}
$$

Thus, $\mathbb{E}\|\nabla\ell_\rho(\boldsymbol{\theta}) - \nabla\ell_\rho(\boldsymbol{\theta_0})\|^2 \leq \delta_3(\|\boldsymbol{\theta} - \boldsymbol{\theta_0}\|) = C_6\|\boldsymbol{\theta} - \boldsymbol{\theta_0}\|^2$.

APPENDIX B

PROOFS OF PROPOSITIONS

Detailed proof of propositions in this dissertation are shown here.

B.1 Proof of Proposition 1

Proposition. *A one-to-one correspondence exists between each diagonal element of* $\boldsymbol{A}^{-1}$ *and each data point of the SAR model.*

Proof. To show this we only need to show that if we swap the location of two data in the neighborhood matrix $\boldsymbol{W}$, the location of these two corresponding diagonal elements of $\boldsymbol{A}^{-1}$ switches accordingly and other diagonal elements do not change. Suppose we want to swap the location of ith and jth data point to generate a new neighborhood matrix $\boldsymbol{W}'$ and $\boldsymbol{A}' = \boldsymbol{I} - \rho\boldsymbol{W}'$. We want to see how $\boldsymbol{A}'^{-1}$ changes compared to $\boldsymbol{A}^{-1}$.

Indeed, to generate $\boldsymbol{W'}$, we only need to swap the ith and jth row of $\boldsymbol{W}$, and then swap the ith and jth column of $\boldsymbol{W}$. Let $\boldsymbol{P}$ be a permutation matrix, which is generated by swapping the ith and jth row of the identity matrix $\boldsymbol{I}$, note that $\boldsymbol{PP} = \boldsymbol{I}$. Thus, we have $\boldsymbol{W'} = \boldsymbol{PWP}$, and accordingly we have

$$\begin{aligned} \boldsymbol{A}' &= \boldsymbol{I} - \rho\boldsymbol{PWP} = \boldsymbol{PIP} - \rho\boldsymbol{PWP} \\ &= \boldsymbol{P}(\boldsymbol{I} - \rho\boldsymbol{W})\boldsymbol{P} = \boldsymbol{PAP}. \end{aligned} \tag{B.1}$$

Then,

$$\boldsymbol{A'}^{-1} = (\boldsymbol{PAP})^{-1} = \boldsymbol{P}^{-1}\boldsymbol{A}^{-1}\boldsymbol{P}^{-1} = \boldsymbol{PA}^{-1}\boldsymbol{P}. \tag{B.2}$$

Thus, compared to $\boldsymbol{A}^{-1}$, $\boldsymbol{A}'^{-1}$ just swaps the ith and jth rows, and also, the ith and jth columns. These two swaps will lead to the swap of the ith and jth diagonal element of $\boldsymbol{A}^{-1}$, and keep other diagonal elements unchanged. □

B.2 Proof of Proposition 3

Proposition 3. *Let the SAR model defined as Equation* (2.1)*, let* $\boldsymbol{\theta} = [\boldsymbol{\beta}^T, \sigma^2, \rho]^T$*,* $\boldsymbol{\theta}_0 = [\boldsymbol{\beta}_0^T, \sigma_0^2, \rho_0]^T$ *be the true parameter, and* $\nabla\ell_{\boldsymbol{\beta},i}, \nabla\ell_{\sigma^2,i}, \nabla\ell_{\rho,i}$ *are the contribution of i-th data unit to the derivative of log-likelihood w.r.t.* $\boldsymbol{\beta}, \sigma^2$ *and* ρ*, respectively. Then,* $\mathbb{E}[\nabla\ell_{\boldsymbol{\beta},i}(\boldsymbol{\theta}_0)] = \mathbf{0}, \mathbb{E}[\nabla\ell_{\sigma^2,i}(\boldsymbol{\theta}_0)] = \mathbb{E}[\nabla\ell_{\rho,i}(\boldsymbol{\theta}_0)] = 0$*. Here, the expectation is with respect to* $\boldsymbol{Y}$*, the conclusion is still true if the expectation is with respect to* $\boldsymbol{Y}$ *and* $\boldsymbol{X}$*.*

Proof. As discussed in Section 2.3, the expression for $\nabla\ell_{\boldsymbol{\beta},i}, \nabla\ell_{\rho,i}$ and $\nabla\ell_{\sigma^2,i}$ are:

$$\nabla\ell_{\boldsymbol{\beta},i} = \frac{1}{\sigma^2}\boldsymbol{x}_i(y_i - \rho\bar{y}_i - \boldsymbol{x}_i^T\boldsymbol{\beta}) \tag{B.3a}$$

$$\nabla\ell_{\rho,i} = -\frac{1}{\rho}((\boldsymbol{A}^{-1})_{ii} - 1) + \frac{1}{\sigma^2}(y_i - \rho\bar{y}_i - \boldsymbol{x}_i^T\boldsymbol{\beta})\bar{y}_i \tag{B.3b}$$

$$\nabla\ell_{\sigma^2,i} = -\frac{1}{2\sigma^2} + \frac{1}{2(\sigma^2)^2}(y_i - \rho\bar{y}_i - \boldsymbol{x}_i^T\boldsymbol{\beta})^2 \tag{B.3c}$$

It is easy to see that $\mathbb{E}[\nabla\ell_{\boldsymbol{\beta},i}(\boldsymbol{\theta}_0)] = \mathbf{0}$ and $\mathbb{E}[\nabla\ell_{\sigma^2,i}(\boldsymbol{\theta}_0)] = 0$, since $y_i - \rho_0\bar{y}_i - \boldsymbol{\beta}_0^T\boldsymbol{x}_i = \epsilon_i$ and $\mathbb{E}[\epsilon_i] = 0, Var(\epsilon_i) = \sigma^2$. For $\nabla\ell_{\rho,i}(\boldsymbol{\theta}_0)$,

$$\mathbb{E}[\nabla\ell_{\rho,i}(\boldsymbol{\theta}_0)] = -\frac{1}{\rho_0}((\boldsymbol{A_0}^{-1})_{ii} - 1) + \frac{1}{\sigma_0^2}\mathbb{E}[(y_i - \rho_0\bar{y}_i - \boldsymbol{\beta}_0^T\boldsymbol{x}_i)\bar{y}_i] \tag{B.4}$$

and,

$$\begin{aligned}
&\mathbb{E}[(y_i - \rho_0\bar{y}_i - \boldsymbol{\beta}_0^T\boldsymbol{x}_i)\bar{y}_i] = \mathbb{E}[\epsilon_i\boldsymbol{w}_i\boldsymbol{Y}] = \mathbb{E}[\epsilon_i\boldsymbol{w}_i(\boldsymbol{A}_0^{-1}\boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{A}_0^{-1}\boldsymbol{\epsilon})] \\
=&\mathbb{E}[\epsilon_i\boldsymbol{w}_i\boldsymbol{A}_0^{-1}\boldsymbol{X}\boldsymbol{\beta}] + \mathbb{E}[\epsilon_i\boldsymbol{w}_i\boldsymbol{A}_0^{-1}\boldsymbol{\epsilon}] = \mathbb{E}[(\boldsymbol{e}_i\boldsymbol{\epsilon})\boldsymbol{w}_i\boldsymbol{A}_0^{-1}\boldsymbol{\epsilon}] = \mathbb{E}[\boldsymbol{w}_i\boldsymbol{A_0}^{-1}\boldsymbol{\epsilon}(\boldsymbol{\epsilon}^T\boldsymbol{e}_i^T)] \\
=&\boldsymbol{w}_i\boldsymbol{A}_0^{-1}\mathbb{E}[\boldsymbol{\epsilon}\boldsymbol{\epsilon}^T]\boldsymbol{e}_i^T = \boldsymbol{w}_i\boldsymbol{A}_0^{-1}\sigma_0^2\boldsymbol{I}\boldsymbol{e}_i^T = \sigma_0^2\boldsymbol{w}_i\boldsymbol{A}_0^{-1}\boldsymbol{e}_i^T = \sigma_0^2\boldsymbol{e}_i\boldsymbol{W}\boldsymbol{A}_0^{-1}\boldsymbol{e}_i^T,
\end{aligned} \tag{B.5}$$

where, $\boldsymbol{e}_i$ is a $1 \times n$ vector and $\boldsymbol{e}_i = [0, \cdots, 0, 1, 0, \cdots, 0]$, i.e., the ith element is 1 and the rest are 0.

Thus,

$$\mathbb{E}[\nabla\ell_{\rho,i}(\boldsymbol{\theta}_0)] = -\frac{1}{\rho_0}((\boldsymbol{A_0}^{-1})_{ii} - 1) + \boldsymbol{e}_i\boldsymbol{W}\boldsymbol{A}_0^{-1}\boldsymbol{e}_i^T. \tag{B.6}$$

To show $\mathbb{E}[\nabla \ell_{\rho,i}(\boldsymbol{\theta}_0)] = 0$, we only need to show

$$(\boldsymbol{A_0}^{-1})_{ii} - 1 = \rho_0 \boldsymbol{e}_i \boldsymbol{W} \boldsymbol{A}_0^{-1} \boldsymbol{e}_i^T. \tag{B.7}$$

Note $\boldsymbol{A}_0^{-1} = (\boldsymbol{I} - \rho_0 \boldsymbol{W})^{-1} = \boldsymbol{I} + \sum_{k=1}^{\infty} (\rho_0 \boldsymbol{W})^k$, thus, the LHS of (B.7) equals to $\sum_{k=1}^{\infty} \rho_0^k (\boldsymbol{W}^k)_{ii}$.

Now the RHS of (B.7):

$$\begin{aligned} \text{RHS} &= \rho_0 \boldsymbol{e}_i \boldsymbol{W} (\boldsymbol{I} + \sum_{k=1}^{\infty} \rho_0^k \boldsymbol{W}^k) \boldsymbol{e}_i^T = \boldsymbol{e}_i (\rho_0 \boldsymbol{W} + \sum_{k=1}^{\infty} \rho_0^{k+1} \boldsymbol{W}^{k+1}) \boldsymbol{e}_i^T \\ &= \boldsymbol{e}_i (\sum_{k=1}^{\infty} \rho_0^k \boldsymbol{W}^k) \boldsymbol{e}_i^T = \sum_{k=1}^{\infty} \rho_0^k (\boldsymbol{e}_i \boldsymbol{W}^k \boldsymbol{e}_i^T) \\ &= \sum_{k=1}^{\infty} \rho_0^k (\boldsymbol{W}^k)_{ii} = \text{LHS of (B.7)}. \end{aligned} \tag{B.8}$$

□

APPENDIX C

S_0 AND V_0 FOR A SINGLE DATA POINT

Let $\nabla \ell_i = [\nabla \ell_{\boldsymbol{\beta},i}^T, \nabla \ell_{\sigma^2,i}, \nabla \ell_{\rho,i}]^T$ be the contribution from the i-th data unit to the derivative of the likelihood of the SAR model. Let $\boldsymbol{\theta}_0$ be the true parameter value, $\boldsymbol{S}_{0,i} = -\mathbb{E}[\nabla^2 \ell_i(\boldsymbol{\theta}_0)]$ and $\boldsymbol{V}_{0,i} = \mathbb{E}[\nabla \ell_i(\boldsymbol{\theta}_0)^T \nabla \ell_i(\boldsymbol{\theta}_0)]$. In this chapter we derive the explicit expression of $\boldsymbol{S}_{0,i}$ and $\boldsymbol{V}_{0,i}$ for the SAR model Equation (2.1).

Note that both $\boldsymbol{S}_0$ and $\boldsymbol{V}_0$ are symmetric and we can write explicitly as:

$$\boldsymbol{S}_0 = -\mathbb{E}\begin{bmatrix} \frac{\partial^2 \ell_i}{\partial \boldsymbol{\beta} \partial \boldsymbol{\beta}^T}|_{\boldsymbol{\theta}_0} & * & * \\ \frac{\partial^2 \ell_i}{\partial \boldsymbol{\beta}^T \partial \sigma^2}|_{\boldsymbol{\theta}_0} & \frac{\partial^2 \ell_i}{\partial (\sigma^2)^2}|_{\boldsymbol{\theta}_0} & * \\ \frac{\partial^2 \ell_i}{\partial \boldsymbol{\beta}^T \partial \rho}|_{\boldsymbol{\theta}_0} & \frac{\partial^2 \ell_i}{\partial \sigma^2 \partial \rho}|_{\boldsymbol{\theta}_0} & \frac{\partial^2 \ell_i}{\partial \rho^2}|_{\boldsymbol{\theta}_0} \end{bmatrix}$$

and

$$\boldsymbol{V}_0 = \mathbb{E}\begin{bmatrix} \frac{\partial \ell_i}{\partial \boldsymbol{\beta}} (\frac{\partial \ell_i}{\partial \boldsymbol{\beta}})^T|_{\boldsymbol{\theta}_0} & * & * \\ (\frac{\partial \ell_i}{\partial \boldsymbol{\beta}})^T \frac{\partial \ell_i}{\partial \sigma^2}|_{\boldsymbol{\theta}_0} & (\frac{\partial \ell_i}{\partial \sigma^2})^2|_{\boldsymbol{\theta}_0} & * \\ (\frac{\partial \ell_i}{\partial \boldsymbol{\beta}})^T \frac{\partial \ell_i}{\partial \rho}|_{\boldsymbol{\theta}_0} & \frac{\partial \ell_i}{\partial \sigma^2} \frac{\partial \ell_i}{\partial \rho}|_{\boldsymbol{\theta}_0} & (\frac{\partial \ell_i}{\partial \rho})^2|_{\boldsymbol{\theta}_0} \end{bmatrix}$$

Below we workout the expression of each element in these two matrices.

$$\begin{aligned}
&\mathbb{E}[\frac{\partial^2 \ell_i}{\partial \boldsymbol{\beta} \partial \boldsymbol{\beta}^T}|_{\boldsymbol{\theta}_0}] \\
&= \mathbb{E}[\frac{\partial (\frac{1}{\sigma^2} \boldsymbol{x}_i (y_i - \rho \bar{y}_i - \boldsymbol{\beta}^T \boldsymbol{x}_i))}{\partial \boldsymbol{\beta}}|_{\boldsymbol{\theta}_0}] = \mathbb{E}[-\frac{1}{\sigma_0^2} \boldsymbol{x}_i \boldsymbol{x}_i^T] = -\frac{1}{\sigma_0^2} E(\boldsymbol{x}_i \boldsymbol{x}_i^T) \\
&\mathbb{E}[\frac{\partial \ell_i}{\partial \boldsymbol{\beta}} (\frac{\partial \ell_i}{\partial \boldsymbol{\beta}})^T | \boldsymbol{\theta}_0] \\
&= E[(\frac{1}{\sigma^2} \boldsymbol{x}_i (y_i - \rho \bar{y}_i - \boldsymbol{\beta}^T \boldsymbol{x}_i))(\frac{1}{\sigma^2} \boldsymbol{x}_i (y_i - \rho \bar{y}_i - \boldsymbol{\beta}^T \boldsymbol{x}_i))^T|_{\boldsymbol{\theta}_0}] \\
&= \mathbb{E}[\frac{1}{\sigma_0^2} \boldsymbol{x}_i \epsilon_i (\frac{1}{\sigma_0^2} \boldsymbol{x}_i \epsilon_i)^T] = \frac{1}{(\sigma_0^2)^2} \mathbb{E}[\boldsymbol{x}_i \boldsymbol{x}_i^T] \mathbb{E}[\epsilon_i^2] = \frac{1}{\sigma_0^2} \mathbb{E}[\boldsymbol{x}_i \boldsymbol{x}_i^T]
\end{aligned}$$

$$
\begin{aligned}
&\mathbb{E}[\frac{\partial^2 \ell_i}{\partial \boldsymbol{\beta}^T \partial \sigma^2}|_{\boldsymbol{\theta}_0}] \\
&= \mathbb{E}[\frac{\partial(\frac{1}{\sigma^2}\boldsymbol{x}_i(y_i - \rho\bar{y}_i - \boldsymbol{\beta}^T\boldsymbol{x}_i)^T)}{\partial\sigma^2}|_{\boldsymbol{\theta}_0}] = -\frac{1}{(\sigma_0^2)^2}\mathbb{E}[\boldsymbol{x}_i^T\epsilon_i] = \mathbf{0}^T \\
&\mathbb{E}[(\frac{\partial \ell_i}{\partial \boldsymbol{\beta}})^T\frac{\partial \ell_i}{\partial \sigma^2}|_{\boldsymbol{\theta}_0}] \\
&= \mathbb{E}[(\frac{1}{\sigma^2}\boldsymbol{x}_i(y_i - \rho\bar{y}_i - \boldsymbol{\beta}^T\boldsymbol{x}_i)^T(-\frac{1}{2\sigma^2} + \frac{1}{2(\sigma^2)^2}(y_i - \rho\bar{y}_i - \boldsymbol{\beta}^T\boldsymbol{x}_i)^2)|_{\boldsymbol{\theta}_0}] \\
&= \mathbb{E}[\frac{1}{\sigma_0^2}\boldsymbol{x}_i^T\epsilon_i(-\frac{1}{2\sigma_0^2} + \frac{1}{2(\sigma_0^2)^2})\epsilon_i^2] = \mathbf{0}^T
\end{aligned}
$$

$$
\begin{aligned}
&\mathbb{E}[\frac{\partial^2 \ell_i}{\partial(\sigma^2)^2}|_{\boldsymbol{\theta}_0}] \\
&= \mathbb{E}[\frac{\partial(-\frac{1}{2\sigma^2} + \frac{1}{2(\sigma^2)^2}(y_i - \rho\bar{y}_i - \boldsymbol{\beta}^T\boldsymbol{x}_i)^2)}{\partial\sigma^2}|_{\boldsymbol{\theta}_0}] \\
&= \frac{1}{2(\sigma_0^2)^2} - \frac{1}{(\sigma_0^2)^3}E(\epsilon_i^2) = -\frac{1}{2(\sigma_0^2)^2} \\
&E(\frac{\partial \ell_i}{\partial \sigma^2})^2|_{\boldsymbol{\theta}_0}] \\
&= \mathbb{E}[(-\frac{1}{2\sigma^2} + \frac{1}{2(\sigma^2)^2}(y_i - \rho\bar{y}_i - \boldsymbol{\beta}^T\boldsymbol{x}_i)^2)^2|_{\boldsymbol{\theta}_0}] \\
&= \mathbb{E}[(-\frac{1}{2\sigma_0^2} + \frac{1}{2(\sigma_0^2)^2}\epsilon_i^2)^2] \\
&= \frac{1}{4(\sigma_0^2)^2} - \frac{1}{2(\sigma_0^2)^3}\mathbb{E}[\epsilon_i^2] + \frac{1}{4(\sigma_0^2)^4}\mathbb{E}[\epsilon_i^4] \\
&= \frac{1}{4(\sigma_0^2)^2} - \frac{1}{2(\sigma_0^2)^3}\sigma_0^2 + \frac{1}{4(\sigma_0^2)^4}3(\sigma_0^2)^2 = \frac{1}{2(\sigma_0^2)^2}
\end{aligned}
$$

Let $\boldsymbol{e}_i = [0, \cdots, 1, \cdots, 0]^T$, i.e., $\boldsymbol{e}_i$ is a n by 1 column vector whose ith element is 1 and the rest are 0. Also, note

$$
\bar{y}_i = w_i\boldsymbol{y} = \boldsymbol{e}_i^T\boldsymbol{W}\boldsymbol{y} = \boldsymbol{e}_i^T\boldsymbol{W}(\boldsymbol{A}_0^{-1}\boldsymbol{X}\boldsymbol{\beta}_\mathbf{0} + \boldsymbol{A}_0^{-1}\boldsymbol{\epsilon})
$$

And easy to verify that, $-\frac{1}{\rho}((\boldsymbol{A})_{ii}^{-1} - 1) = -\boldsymbol{e}_i^T\boldsymbol{W}\boldsymbol{A}^{-1}\boldsymbol{e}_i$

$$
\begin{aligned}
&\mathbb{E}[\frac{\partial^2 \ell_i}{\partial \boldsymbol{\beta}^T \partial \rho}|_{\boldsymbol{\theta}_0}] \\
&= \mathbb{E}[\frac{\partial(\frac{1}{\sigma^2}\boldsymbol{x}_i(y_i - \rho\bar{y}_i - \boldsymbol{\beta}^T\boldsymbol{x}_i)^T}{\partial \rho}|_{\boldsymbol{\theta}_0}] \\
&= -\frac{1}{\sigma_0^2}\mathbb{E}[\bar{y}_i\boldsymbol{x}_i^T] = -\frac{1}{\sigma_0^2}\mathbb{E}[\boldsymbol{e}_i^T\boldsymbol{W}(\boldsymbol{A}_0^{-1}\boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{A}_0^{-1}\boldsymbol{\epsilon})\boldsymbol{e}_i^T\boldsymbol{X}] \\
&= -\frac{1}{\sigma_0^2}\mathbb{E}[(\boldsymbol{e}_i^T\boldsymbol{W}\boldsymbol{A}_0^{-1}\boldsymbol{X}\boldsymbol{\beta_0})\boldsymbol{e}_i^T\boldsymbol{X}] = -\frac{1}{\sigma_0^2}\boldsymbol{e}_i^T\boldsymbol{W}\boldsymbol{A}_0^{-1}\mathbb{E}[\boldsymbol{X}\boldsymbol{\beta_0}\boldsymbol{e}_i^T\boldsymbol{X}] \\
&\mathbb{E}[(\frac{\partial \ell_i}{\partial \boldsymbol{\beta}})^T\frac{\partial \ell_i}{\partial \rho}|_{\boldsymbol{\theta}_0}] \\
&= \mathbb{E}[(\frac{1}{\sigma^2}\boldsymbol{x}_i(y_i - \rho\bar{y}_i - \boldsymbol{\beta}^T\boldsymbol{x}_i)^T(-\boldsymbol{e}_i^T\boldsymbol{W}\boldsymbol{A}^{-1}\boldsymbol{e}_i + \frac{1}{\sigma^2}(y_i - \rho\bar{y}_i - \boldsymbol{\beta}^T\boldsymbol{x}_i)\bar{y}_i)|_{\boldsymbol{\theta}_0}] \\
&= \frac{1}{\sigma_0^2}\mathbb{E}[\epsilon_i\boldsymbol{x}_i^T(-\boldsymbol{e}_i^T\boldsymbol{W}\boldsymbol{A}^{-1}\boldsymbol{e}_i + \frac{1}{\sigma_0^2}\epsilon_i\bar{y}_i)] = \frac{1}{(\sigma_0^2)^2}\mathbb{E}[\boldsymbol{x}_i^T\epsilon_i^2\bar{y}_i] \\
&= \frac{1}{(\sigma_0^2)^2}\mathbb{E}[\boldsymbol{x}_i^T\epsilon_i^2\boldsymbol{e}_i^T\boldsymbol{W}(\boldsymbol{A}_0^{-1}\boldsymbol{X}\boldsymbol{\beta_0} + \boldsymbol{A}_0^{-1}\boldsymbol{\epsilon})] = \frac{1}{(\sigma_0^2)^2}\mathbb{E}[\epsilon_i^2\boldsymbol{e}_i^T\boldsymbol{W}\boldsymbol{A}_0^{-1}\boldsymbol{X}\boldsymbol{\beta}_0\boldsymbol{x}_i^T] \\
&= \frac{1}{\sigma_0^2}\boldsymbol{e}_i^T\boldsymbol{W}\boldsymbol{A}_0^{-1}\mathbb{E}[\boldsymbol{X}\boldsymbol{\beta}_0\boldsymbol{e}_i^T\boldsymbol{X}]
\end{aligned}
$$

$$
\begin{aligned}
&\mathbb{E}[\frac{\partial^2 \ell_i}{\partial \sigma^2 \partial \rho}|_{\boldsymbol{\theta}_0}] \\
&= \mathbb{E}[\frac{\partial(-\boldsymbol{e}_i^T\boldsymbol{W}\boldsymbol{A}^{-1}\boldsymbol{e}_i + \frac{1}{\sigma^2}(y_i - \rho\bar{y}_i - \boldsymbol{\beta}^T\boldsymbol{x}_i)\bar{y}_i)}{\partial \sigma^2}|_{\boldsymbol{\theta}_0}] \\
&= \mathbb{E}[-\frac{1}{(\sigma_0)^2}\epsilon_i\bar{y}_i] = -\frac{1}{(\sigma_0^2)^2}\mathbb{E}[\epsilon_i\boldsymbol{e}_i^T\boldsymbol{W}(\boldsymbol{A}_0^{-1}\boldsymbol{X}\boldsymbol{\beta_0} + \boldsymbol{A}_0^{-1}\boldsymbol{\epsilon})] \\
&= -\frac{1}{(\sigma_0^2)^2}\mathbb{E}[\boldsymbol{e}_i^T\boldsymbol{W}\boldsymbol{A}_0^{-1}\boldsymbol{\epsilon}\epsilon_i] = -\frac{1}{(\sigma_0^2)^2}\mathbb{E}[\boldsymbol{e}_i^T\boldsymbol{W}\boldsymbol{A}_0^{-1}\boldsymbol{\epsilon}\boldsymbol{\epsilon}^T\boldsymbol{e}_i] \\
&= -\frac{1}{\sigma_0^2}\boldsymbol{e}_i^T\boldsymbol{W}\boldsymbol{A}_0^{-1}\boldsymbol{e}_i
\end{aligned}
$$

$$
\begin{aligned}
&\mathbb{E}[\frac{\partial \ell_i}{\partial \sigma^2}\frac{\partial \ell_i}{\partial \rho}|_{\boldsymbol{\theta}_0}] \\
&= \mathbb{E}[(-\frac{1}{2\sigma^2}+\frac{1}{2(\sigma^2)^2}(y_i-\rho\bar{y}_i-\boldsymbol{\beta}^T\boldsymbol{x}_i)^2)(-\boldsymbol{e}_i^T\boldsymbol{W}\boldsymbol{A}^{-1}\boldsymbol{e}_i+\frac{1}{\sigma^2}(y_i-\rho\bar{y}_i-\boldsymbol{\beta}^T\boldsymbol{x}_i)\bar{y}_i)|_{\boldsymbol{\theta}_0}] \\
&= \mathbb{E}[(-\frac{1}{2\sigma_0^2}+\frac{1}{2(\sigma_0^2)^2}\epsilon_i^2)(-\boldsymbol{e}_i^T\boldsymbol{W}\boldsymbol{A}^{-1}\boldsymbol{e}_i+\frac{1}{\sigma_0^2}\epsilon_i\bar{y}_i)] \\
&= \mathbb{E}[(-\frac{1}{2\sigma_0^2}+\frac{1}{2(\sigma_0^2)^2}\epsilon_i^2)(-\boldsymbol{e}_i^T\boldsymbol{W}\boldsymbol{A}^{-1}\boldsymbol{e}_i)]+\mathbb{E}[(-\frac{1}{2\sigma_0^2}+\frac{1}{2(\sigma_0^2)^2}\epsilon_i^2)\frac{1}{\sigma_0^2}\epsilon_i\bar{y}_i] \\
&= \mathbb{E}[(-\frac{1}{2\sigma_0^2}+\frac{1}{2(\sigma_0^2)^2}\epsilon_i^2)\frac{1}{\sigma_0^2}\epsilon_i\bar{y}_i] = -\frac{1}{2(\sigma_0^2)^2}\mathbb{E}[\epsilon_i\bar{y}_i]+\frac{1}{2(\sigma_0^2)^3}\mathbb{E}[\epsilon_i^3\bar{y}_i] \\
&= -\frac{1}{2\sigma_0^2}\boldsymbol{e}_i^T\boldsymbol{W}\boldsymbol{A}_0^{-1}\boldsymbol{e}_i+\frac{1}{(2\sigma_0^2)^3}\mathbb{E}[\epsilon_i^3\boldsymbol{e}_i^T\boldsymbol{W}(\boldsymbol{A}_0^{-1}\boldsymbol{X}\boldsymbol{\beta_0}+\boldsymbol{A}_0^{-1}\boldsymbol{\epsilon})] \\
&= -\frac{1}{2\sigma_0^2}\boldsymbol{e}_i^T\boldsymbol{W}\boldsymbol{A}_0^{-1}\boldsymbol{e}_i+\frac{1}{(2\sigma_0^2)^3}\boldsymbol{e}_i^T\boldsymbol{W}\boldsymbol{A}_0^{-1}\mathbb{E}[\epsilon_i^3\boldsymbol{\epsilon}] \\
&= -\frac{1}{2\sigma_0^2}\boldsymbol{e}_i^T\boldsymbol{W}\boldsymbol{A}_0^{-1}\boldsymbol{e}_i+\frac{1}{(2\sigma_0^2)^3}\boldsymbol{e}_i^T\boldsymbol{W}\boldsymbol{A}_0^{-1}E(\epsilon_i^4)\boldsymbol{e}_i \\
&= -\frac{1}{2\sigma_0^2}\boldsymbol{e}_i^T\boldsymbol{W}\boldsymbol{A}_0^{-1}\boldsymbol{e}_i+\frac{1}{(2\sigma_0^2)^3}\boldsymbol{e}_i^T\boldsymbol{W}\boldsymbol{A}_0^{-1}(3(\sigma_0^2)^2)\boldsymbol{e}_i \\
&= \frac{1}{\sigma_0^2}\boldsymbol{e}_i^T\boldsymbol{W}\boldsymbol{A}_0^{-1}\boldsymbol{e}_i
\end{aligned}
$$

Note that $\frac{\partial \boldsymbol{A}^{-1}}{\partial \rho} = -\boldsymbol{A}^{-1}\frac{\partial \boldsymbol{A}}{\partial \rho}\boldsymbol{A}^{-1} = \boldsymbol{A}^{-1}\boldsymbol{W}\boldsymbol{A}^{-1}$. To simplify notation, let

$$
P = \boldsymbol{e}_i^T\boldsymbol{W}\boldsymbol{A}_0^{-1}\boldsymbol{X}\boldsymbol{\beta}\boldsymbol{\beta}^T\boldsymbol{X}^T(\boldsymbol{A}_0^{-1})^T\boldsymbol{W}^T\boldsymbol{e}_i
$$

$$
\begin{aligned}
&\mathbb{E}[\frac{\partial^2 \ell_i}{\partial \rho^2}|_{\boldsymbol{\theta}_0}] \\
&= -\boldsymbol{e}_i^T\boldsymbol{W}\boldsymbol{A}_0^{-1}\boldsymbol{W}\boldsymbol{A}_0^{-1}\boldsymbol{e}_i-\frac{1}{\sigma_0^2}P-\boldsymbol{e}_i^T\boldsymbol{W}\boldsymbol{A}_0^{-1}(\boldsymbol{A}_0^{-1})^T\boldsymbol{W}^T\boldsymbol{e}_i \\
&\mathbb{E}[(\frac{\partial \ell_i}{\partial \rho})^2|_{\boldsymbol{\theta}_0}] \\
&= (\boldsymbol{e}_i^T\boldsymbol{W}\boldsymbol{A}_0^{-1}\boldsymbol{e}_i)^2+\frac{1}{\sigma_0^2}P+\boldsymbol{e}_i^T\boldsymbol{W}\boldsymbol{A}_0^{-1}(\boldsymbol{A}_0^{-1})^T\boldsymbol{W}^T\boldsymbol{e}_i
\end{aligned}
$$

APPENDIX D

R PACKAGE

This appendix briefly describes a R package, 'SGDCI', developed for applying SGD algorithm for parameter estimation and perturbed estimates for CIs construction. The most updated version of the package is available at `http://github.com/ganluannj/Spatial_SGD_Inference`. It contains the parameter estimation function and CIs construction function for the following models:

- Linear mean regression
- Logistic regression
- linear quantile regression
- SAR mean regression

REFERENCES

[1] Mordecai Avriel and Douglass J Wilde. Optimally proof for the symmetric fibonacci search technique. *Fibonacci Quarterly Journal*, 1966.

[2] Edward R Berchick, Emily Hood, and Jessica C Barnett. *Health insurance coverage in the United States: 2018*. Washington, DC: US Department of Commerce, 2019.

[3] Centers for Medicare and Medicaid Services (Baltimore, MD). Medicare provider utilization and payment data: Physician and other supplier. `https://www.cms.gov/Research-Statistics-Data-and-Systems/Statistics-Trends-and-Reports/Medicare-Provider-Charge-Data/Physician-and-Other-Supplier`, retrieved on 9/20/2021.

[4] Colin Chen and Ying Wei. Computational issues for quantile regression. *Sankhyā: The Indian Journal of Statistics*, pages 399–417, 2005.

[5] Xi Chen, Jason D Lee, Xin T Tong, and Yichen Zhang. Statistical inference for model parameters in stochastic gradient descent. *arXiv preprint arXiv:1610.08637*, 2016.

[6] Victor Chernozhukov and Christian Hansen. Instrumental quantile regression inference for structural and treatment effect models. *Journal of Econometrics*, 132(2):491–525, 2006.

[7] Andrew D. Clif and Keith Ord. *Spatial autocorrection*. London, UK: Pion, 1973.

[8] Andrew D. Clif and Keith Ord. *Spatial processes: models and applications*. London, UK: Pion, 1981.

[9] Noel Cressie. *Statistics for spatial data, revised edition*. Hoboken, NJ: Wiley, 2015.

[10] Cristina Davino, Marilena Furno, and Domenico Vistocco. *Quantile regression: theory and applications*, volume 988. Hoboken, NJ: Wiley, 2013.

[11] Michael John de Smith. *Statistical analysis handbook*. London, UK: The Winchelsea Press, 2021.

[12] Yixin Fang. Scalable statistical inference for averaged implicit stochastic gradient descent. *Scandinavian Journal of Statistics*, 46(4):987–1002, 2019.

[13] Yixin Fang, Jinfeng Xu, and Lei Yang. Online bootstrap confidence intervals for the stochastic gradient descent estimator. *The Journal of Machine Learning Research*, 19(1):3053–3073, 2018.

[14] Jerome Friedman, Trevor Hastie, and Robert Tibshirani. *The elements of statistical learning*, volume 1. New York, NY: Springer, 2001.

[15] Gene H. Golub and Charles F. Van Loan. *Matrix Computations, 4th Edition*. Baltimore, MD: Johns Hopkins University Press, 1996.

[16] Ian Goodfellow, Yoshua Bengio, and Aaron Courville. *Deep learning*. Cambridge, MA: MIT press, 2016.

[17] Murtaza Haider and Eric J Miller. Effects of transportation infrastructure and location on residential real estate values: application of spatial autoregressive techniques. *Transportation Research Record*, 1722(1):1–8, 2000.

[18] Grant Hillier and Federico Martellosio. Properties of the maximum likelihood estimator in spatial autoregressive models. Technical report, Eemmap Working Paper, 2013.

[19] Pavlos S Kanaroglou, Matthew D Adams, Patrick F De Luca, Denis Corr, and Nazmul Sohel. Estimation of sulfur dioxide air pollution concentrations with a spatial autoregressive model. *Atmospheric Environment*, 79:421–427, 2013.

[20] Baris M Kazar and Mete Celik. *Spatial autoregression (SAR) model: Parameter estimation techniques*. Springer Science and Business Media, 2012.

[21] Harry H. Kelejian and Ngmar R. Prucha. A generalized moments estimator for the autoregressive parameter in a spatial model. *International Economic Review*, 40(2):509–33, 1999.

[22] Harry H Kelejian and Dennis P Robinson. A suggested method of estimation for spatial interdependent models with autocorrelated errors, and an application to a county expenditure model. *Papers in Regional Science*, 72(3):297–312, 1993.

[23] Jack Kiefer. Sequential minimax search for a maximum. *Proceedings of the American Mathematical Society*, 4(3):502–506, 1953.

[24] Tae-Hwan Kim and Christophe Muller. Two-stage quantile regression when the first stage is based on quantile regression. *The Econometrics Journal*, 7(1):218–231, 2004.

[25] Roger Koenker and Gilbert Bassett Jr. Regression quantiles. *Econometrica: Journal of the Econometric Society*, pages 33–50, 1978.

[26] Roger Koenker, Stephen Portnoy, Pin Tian Ng, Achim Zeileis, Philip Grosjean, and Brian D Ripley. Package 'quantreg'. *Cran R-project. org*, 2018.

[27] Philip Kostov. A spatial quantile regression hedonic model of agricultural land prices. *Spatial Economic Analysis*, 4(1):53–72, 2009.

[28] Philip Kostov. Empirical likelihood estimation of the spatial quantile regression. *Journal of Geographical Systems*, 15(1):51–69, 2013.

[29] Maria Kyriacou, Peter CB Phillips, and Francesca Rossi. Indirect inference in spatial autoregression. *The Econometrics Journal*, 20(2):168–189, 2017.

[30] Kenneth C Land and Glenn Deane. On the large-sample estimation of regression models with spatial-or network-effects terms: A two-stage least squares approach. *Sociological Methodology*, pages 221–248, 1992.

[31] Lung-Fei Lee. Consistency and efficiency of least squares estimation for mixed regressive, spatial autoregressive models. *Econometric Theory*, 18(2):252–277, 2002.

[32] Lung-fei Lee. Best spatial two-stage least squares estimators for a spatial autoregressive model with autoregressive disturbances. *Econometric Reviews*, 22(4):307–335, 2003.

[33] Lung-Fei Lee. Asymptotic distributions of quasi-maximum likelihood estimators for spatial autoregressive models. *Econometrica*, 72(6):1899–1925, 2004.

[34] Lung-fei Lee. GMM and 2SLS estimation of mixed regressive, spatial autoregressive models. *Journal of Econometrics*, 137(2):489–514, 2007.

[35] Hongfei Li, Catherine A Calder, and Noel Cressie. Beyond Moran's I: testing for spatial dependence based on the spatial autoregressive model. *Geographical Analysis*, 39(4):357–375, 2007.

[36] Songnian Li, Suzana Dragicevic, Francesc Antón Castro, Monika Sester, Stephan Winter, Arzu Coltekin, Christopher Pettit, Bin Jiang, James Haworth, Alfred Stein, et al. Geospatial big data handling theory and methods: A review and research challenges. *ISPRS Journal of Photogrammetry and Remote Sensing*, 115:119–133, 2016.

[37] Tianyang Li, Liu Liu, Anastasios Kyrillidis, and Constantine Caramanis. Statistical inference using SGD. *arXiv preprint arXiv:1705.07477*, 2017.

[38] Federico Martellosio and Grant Hillier. Adjusted qmle for the spatial autoregressive parameter. *Journal of Econometrics*, 2020.

[39] José-María Montero, Gema Fernández-Avilés, and Jorge Mateu. *Spatial and spatio-temporal geostatistical modeling and kriging*, volume 998. Hoboken, NJ: Wiley, 2015.

[40] Patrick AP Moran. Notes on continuous stochastic phenomena. *Biometrika*, 37(1/2):17–23, 1950.

[41] Mor Shpigel Nacson, Nathan Srebro, and Daniel Soudry. Stochastic gradient descent on separable data: Exact convergence with a fixed learning rate. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pages 3051–3059, 2019.

[42] Arkadi Nemirovski, Anatoli Juditsky, Guanghui Lan, and Alexander Shapiro. Robust stochastic approximation approach to stochastic programming. *SIAM Journal on Optimization*, 19(4):1574–1609, 2009.

[43] Keith Ord. Estimation methods for models of spatial interaction. *Journal of the American Statistical Association*, 70(349):120–126, 1975.

[44] Syarifah Diana Permai, Ronald Jauri, and Andry Chowanda. Spatial autoregressive (SAR) model for average expenditure of Papua province. *Procedia Computer Science*, 157:537–542, 2019.

[45] Boris T Polyak and Anatoli B Juditsky. Acceleration of stochastic approximation by averaging. *SIAM Journal on Control and Optimization*, 30(4):838–855, 1992.

[46] Yongsong Qin. Empirical likelihood for spatial autoregressive models with spatial autoregressive disturbances. *Sankhya A*, pages 1–25, 2019.

[47] Herbert Robbins and Sutton Monro. A stochastic approximation method. *The Annals of Mathematical Statistics*, pages 400–407, 1951.

[48] David Ruppert. Efficient estimations from a slowly convergent robbins-monro process. Technical report, Cornell University Operations Research and Industrial Engineering, 1988.

[49] Skipper Seabold and Josef Perktold. Statsmodels: econometric and statistical modeling with python. In *9th Python in Science Conference*, 2010.

[50] Liangjun Su and Zhenlin Yang. Instrumental variable quantile estimation of spatial autoregressive models. *Working paper, School of Economics, Singapore Management University*, 2007.

[51] Weijie J Su and Yuancheng Zhu. Uncertainty quantification for online learning and stochastic approximation via hierarchical incremental gradient descent. *arXiv preprint arXiv:1802.04876*, 2018.

[52] Jonathan Taylor and Robert J Tibshirani. Statistical learning and selective inference. *Proceedings of the National Academy of Sciences*, 112(25):7629–7634, 2015.

[53] Grażyna Trzpiot and Agnieszka Orwat-Acedańska. The classification of spatial quantile regression models for healthy life years in european countries. *Argumenta Oeconomica*, (43):115–136, 2019.

[54] Melanie M Wall. A close look at the spatial structure implied by the car and sar models. *Journal of Statistical Planning and Inference*, 121(2):311–324, 2004.

[55] Chun Wang, Ming-Hui Chen, Elizabeth Schifano, Jing Wu, and Jun Yan. Statistical methods and computing for big data. *Statistics and Its Interface*, 9(4):399, 2016.

[56] Jacob Wolfowitz et al. On the stochastic approximation method of robbins and monro. *The Annals of Mathematical Statistics*, 23(3):457–461, 1952.

[57] Chenghu Zhou, Fenzhen Su, Francis Harvey, and Jun Xu. *Spatial Data Handling in Big Data Era*. New York, NY: Springer, 2016.