NORTHWESTERN UNIVERSITY

Spatial Statistics Analysis with Artificial Neural Network

A DISSERTATION

SUBMITTED TO THE GRADUATE SCHOOL

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS

for the degree

DOCTOR OF PHILOSOPHY

Field of Statistics

By

Wenqian Wang

EVANSTON, ILLINOIS

June 2019

Abstract

The spatial autoregressive model has been widely applied in science, in areas such as economics, public finance, political science, agricultural economics, environmental studies and transportation analyses. The classical spatial autoregressive model is a linear model for describing spatial correlation. In this work, we expand the classical model to include time lagged observations, related exogenous variables, possibly non-Gaussian, high volatility errors, and a nonlinear neural network component. The nonlinear neural network component allows for more model flexibility — the ability to learn and model nonlinear and complex relationships. We use a maximum likelihood approach for model parameter estimation. We establish consistency and asymptotic normality for these estimators under some standard conditions on the spatial/space-time model and neural network component. We investigate the quality of the asymptotic approximations for finite samples by means of numerical simulation studies. Next, we discuss the model selection in the proposed space-time autoregressive model. We employ the Shakeout noise injection method to conduct feature selection and use the likelihood ratio test for the time lag order selection. We evaluate the performance of Shakeout noise injection technique in a simulated dataset and also investigate the asymptotic approximation of the likelihood

ratio test statistics by simulations. Finally, we apply our proposed spatial and space-time autoregressive models to a real world application.

Acknowledgements

First, I would like to express my sincere gratitude to my advisor Prof. Beth Andrews who guided me well throughout the research work from topic selection to developing new results. Even though there were lots of obstacles during the research, it has been a pleasant journey to work with her and to explore something new in spatial analysis and machine learning. She gave me a lot of insightful suggestions and taught me to be more independent in the research. On a personal level, Prof. Andrews inspired me by her hardworking and passionate attitude. I really appreciate her patience in supporting me and helping me be a better person.

Apart from my advisor, I would like to thank the rest of my thesis committee: Prof. Wenxin Jiang and Prof. Thomas A Severini, for giving the encouragement and sharing insightful suggestions. They have tremendous experience in related fields and their suggestions helped me a lot in defining a comprehension research project. They have played a major role in guiding me to widen my research from various perspectives.

I would always remember my fellow in the department of statistics which makes me feel that I am in a big family. Everyone in the department is so friendly that I really enjoy working and living at Northwestern. I would like to thank all faculties in the department who supported me during the five-year study. They gave me lots of advice in how to live a more fulfilling life during the graduate study. Also I should express my gratitude to my colleagues. They accompanied me throughout the graduate program and encouraged me

to go through the tough times. In the department, we shared lots of precious memory together.

In the end, I am grateful to my parents, friends and acquaintances who gave me encouragement and motivation to accomplish my personal goals. During nine years of study since my undergraduate, even though I am studying far away from my hometown, my parents always encourage me throughout the entire study programs. They are my backbone. With their great support, I could fully concentrate on studies and achieving my objective.

Table of Contents

List of Tables

List of Figures

CHAPTER 1

Partially Specified Spatial Autoregressive Model with Artificial Neural Network

One commonly used assumption in regression analysis is that observations are uncorrelated, but this assumption is sometimes impossible to be defended in the analysis of spatial data when one observation may be related to neighboring entities. The nature of the covariance among observations may not be known precisely and researchers have been dedicated for years to building appropriate models to describe such correlation. The collection of techniques to investigate properties in the spatial models is considered to have begun in the domain of spatial econometrics first proposed by Paelinck in the early 1970s [35]. Later, the books by Cliff and Ord [18], Anselin [3] and Cressie [14] detailed research results related to spatial autocorrelation, purely spatial dependence as well as cross-sectional and/or panel data.

So why has estimating the spatial correlation drawn so much attention? In some applications estimating the spatial structure of the dependence may be a subject of interest or provide a key insight; in other contexts, it may be regarded as serial correlations. However, in either case, inappropriate treatment of data with spatial dependence can lead to inefficient or biased and inconsistent estimates. These consequences may result in misleading conclusions in the analysis of real world problems. Therefore, it is important to describe spatial dependence; some standard parametric models are spatial autoregressive

models (SAR), spatial error models (SEM) and spatial Durbin models (LeSage, R. Pace, [26]). According to the spatial autoregressive model, values of the dependent variable are linearly related to observations in neighboring regions. The SAR model has been widely discussed in the literature, and researchers have proposed various parameter estimation techniques such as the method of maximum likelihood by Ord [34] and Smirnov and Anselin [40], the method of moments by Kelejian and Prucha [22, 24, 23] and the method of quasi-maximum likelihood estimation by Lee [25].

In a SAR model with covariates, the observations are modeled as a weighted average of neighboring observations with weights determined by the distance between them plus a function of the covariates:

$$y_s = \rho \sum_{i=1}^{n} w_{si} y_i + x_s' \beta + \varepsilon_s \quad s = 1, 2, \ldots, n$$

where y_s denotes the observation of interest and x_s denotes the value of a p dimensional independent variable at location $s \in \{1, 2, \ldots, n\}$. w_{ij} is the (i, j) entry of a $n \times n$ weight matrix W_n; it is a nonnegative weight which measures the degree of interaction between units i and j. By convention, we always set $w_{ii} = 0$. The random disturbances $\{\varepsilon_s\}_{s=1}^{n}$ are uncorrelated with zero means and equal variances; often in the literature these are taken to be normally distributed. The model has parameter vector $\theta = (\rho, \beta')$. However, parametric models are vulnerable to the preciseness of model specification: a misspecified model can draw misleading inferences. Whereas a nonparametric model is more robust even though it sacrifices the precision. In this sense, to combine the advantages of these two models, we consider a semi-parametric model in the spatial context. The suggested model, a partially specified spatial autoregressive model (PSAR) [44], is defined

as follows:

$$y_s = \rho \sum_{i=1}^{n} w_{si} y_i + x_s' \beta + g(z_s) + \varepsilon_s \quad s = 1, 2, \ldots, n \tag{1.1}$$

where $g(\cdot)$ is an unknown function and z_s denotes a q dimensional vector of explanatory variables at location s. This PSAR model has a more flexible functional form than the ordinary spatial autoregressive model. Methods of parameter estimation for the PSAR model include profile quasi-maximum likelihood estimation by Su and Jin [44] and a sieve method by Zhang [51]. In Su and Jin [44], they used profile quasi-maximum likelihood estimators for independent and identically distributed errors and gave an asymptotic analysis using local polynomials to describe g. This method showed its advantage in dimension reduction when maximizing concentrated likelihood function with respect to one parameter ρ but involved in two-stage maximization if we wanted to obtain other parameter estimators such as β's. However, in Zhang [51], they were using a sieve method (Ai, Chen [1]) to approximate the nonparametric function. They applied a sequence of known basis functions to approximate $g(\cdot)$ in equation (1.1), and used the two-stage least squares estimation with some instrumental variables to obtain consistent estimators for the PSAR model.

Both methods use Gaussian likelihood techniques. But normality is unreasonable in many cases when we observe errors with heavy tails or abnormal patterns. If this is the case, maximum likelihood estimation can be more efficient than Gaussian-based quasi-maximum likelihood estimation. Another difference is that we are using neural network models to estimate the nonlinear function $g(\cdot)$ whereas Su and Jin [44] applied a finite

order of local polynomials about some explanatory variables and Zhang [51] used a linear combination of a sequence of known functions to estimate $g(\cdot)$.

The purpose of this paper is to extend an autoregressive artificial neural network model (Medeiros, Teräsvirta, Rech [32]) developed in the context of time series data to a partially specified spatial autoregressive model and we regard the artificial neural network part as a nonlinear statistical component to approximate the nonparametric function $g(\cdot)$ in the PSAR model (1.1). The use of an ANN (Artificial Neural Network) model is motivated by mathematical results showing that under mild conditions, a relatively simple ANN model is competent in approximating any Borel-measurable function to any given degree of accuracy (see for example Hornik *et al.* [19], Gallant and White [16]). Under this theoretical foundation, we would expect our model to perform well when modeling nonparamatric components in spatial contexts. Another improvement is that, in our model, the random error is independent and identically distributed but does not necessarily follow a normal distribution. We derive parameter estimates by maximizing the corresponding likelihood function and discuss the asymptotic properties of our estimators under conditions that the spatial weight matrix is nonsingular and the log likelihood function has some dominated function with a finite mean.

In Sections 2 and 3, our model PSAR-ANN is given and a likelihood function for corresponding parameters is derived. In Sections 4 and 5, we discuss model identification and establish consistency and asymptotic normality for MLEs of model parameters. In section 6, we describe numerical simulation studies to investigate how well the behavior of estimators for finite samples matches the limiting theory, i.e., the quality of the normal approximation. In the real data example, we would like to explore spatial dynamics in

U.S. presidential elections and a PSAR-ANN model is fit to the proportion of votes cast for 2004 U.S. presidential candidates at the county level.

1.1. Model Specification

The main focus of this paper is to approximate the nonparametric function $g(\cdot)$ in the partially specified spatial autoregressive model (1.1) by an artificial neural network model. The model in matrix form is defined as

$$Y_n = X_n\beta + \rho W_n Y_n + F(X_n\boldsymbol{\gamma}')\lambda + \boldsymbol{\varepsilon}_n \tag{1.2}$$

where $Y_n = \{y_s\}_{s=1}^n$ contains observations of the dependent variable at n locations. The independent variable matrix $X_n = (x_1, x_2, \ldots, x_n)' \in \mathbb{R}^{n\times q}$ contains values of exogenous regressors for the n regions, where for each region, $x_s = (x_{s1}, \ldots, x_{sq})'$, $s = 1, 2, \ldots, n$, is a q dimensional vector. $\boldsymbol{\varepsilon}_n = \{\varepsilon_s\}_{s=1}^n$ denotes a vector of n independent identically distributed random noises with density function $f(\cdot)$, mean 0 and variance $\sigma^2 = 1$.

Exogenous parameters $\beta = (\beta_1, \ldots, \beta_q)' \in \mathbb{R}^q$ and scalar ρ, the spatial autoregressive parameter, are assumed to be the same over all regions. $W_n = \{w_{ij}\} \in \mathbb{R}^{n\times n}$ denotes a spatial weight matrix which characterizes the connections between neighboring regions. For the ease of illustration, we define some additional notations. Given a function $f \in C^1(R^1)$ continuous on $\mathbb{R}$, we define a new matrix mapping $R^n \to R^n$ as $\boldsymbol{f}$ s.t. $\boldsymbol{f}(x_1, \ldots, x_n) = (f(x_1), \ldots, f(x_n))'$. Using this notation, the artificial neural network

component (Medeiros *et al.* [32]) can be written as $\boldsymbol{F}(X_n\boldsymbol{\gamma}')\lambda$ with

$$\boldsymbol{F}(X_n\boldsymbol{\gamma}') = \begin{bmatrix} F(x_1'\boldsymbol{\gamma}_1) & F(x_1'\boldsymbol{\gamma}_2) & \dots & F(x_1'\boldsymbol{\gamma}_h) \\ F(x_2'\boldsymbol{\gamma}_1) & F(x_2'\boldsymbol{\gamma}_2) & \dots & F(x_2'\boldsymbol{\gamma}_h) \\ \vdots & \vdots & \ddots & \vdots \\ F(x_n'\boldsymbol{\gamma}_1) & F(x_n'\boldsymbol{\gamma}_2) & \dots & F(x_n'\boldsymbol{\gamma}_h) \end{bmatrix} \in \mathbb{R}^{n\times h}$$

This matrix represents a single layer neural network with h neurons for every location. The value of h is determined by researchers and can be selected by comparing AIC/BIC. Under this setting, the parameter matrix $\boldsymbol{\gamma} = (\boldsymbol{\gamma}_1, \boldsymbol{\gamma}_2, \dots, \boldsymbol{\gamma}_h)' \in \mathbb{R}^{h\times q}$, $\boldsymbol{\gamma}_i = (\gamma_{i1}, \dots, \gamma_{iq})' \in \mathbb{R}^q$, $i = 1, 2, \dots, h$, contains all the weights in a neural network model. $F(\cdot)$ is called the activation function and we discuss the situation when it is the logistic function with range from 0 to 1 (the logistic activation function is the most common choice in neural network modeling [32]). For given information x_s at region s, the corresponding output of ith neuron in a single layer neural network is

$$F(x_s'\boldsymbol{\gamma}_i) = (1 + e^{-x_s'\boldsymbol{\gamma}_i})^{-1}, \quad s = 1, 2, \dots, n,\ i = 1, 2, \dots, h$$

Parameter vector $\lambda = (\lambda_1, \lambda_2, \dots, \lambda_h)'$ denotes weights for h neurons. So $\boldsymbol{F}(X_n\boldsymbol{\gamma}')\lambda =$

$$\begin{bmatrix} F(x_1'\boldsymbol{\gamma}_1) & F(x_1'\boldsymbol{\gamma}_2) & \dots & F(x_1'\boldsymbol{\gamma}_h) \\ F(x_2'\boldsymbol{\gamma}_1) & F(x_2'\boldsymbol{\gamma}_2) & \dots & F(x_2'\boldsymbol{\gamma}_h) \\ \vdots & \vdots & \ddots & \vdots \\ F(x_n'\boldsymbol{\gamma}_1) & F(x_n'\boldsymbol{\gamma}_2) & \dots & F(x_n'\boldsymbol{\gamma}_h) \end{bmatrix} \begin{bmatrix} \lambda_1 \\ \lambda_2 \\ \vdots \\ \lambda_h \end{bmatrix} = \begin{bmatrix} \sum_{i=1}^h \lambda_i F(x_1'\boldsymbol{\gamma}_i) \\ \sum_{i=1}^h \lambda_i F(x_2'\boldsymbol{\gamma}_i) \\ \vdots \\ \sum_{i=1}^h \lambda_i F(x_n'\boldsymbol{\gamma}_i) \end{bmatrix} \in \mathbb{R}^n$$

One important element in the model (1.2) is the spatial weight matrix W_n. The spatial weights depend on the definition of a neighborhood set for each observation. In our applications we begin by using a square symmetric $n \times n$ matrix with (i, j) element equal to 1 if regions i and j are neighbors and $w_{ij} = 0$ otherwise. The diagonal elements of the spatial neighbors matrix are set to zero. Then we row standardize the weight

matrix, so the nonzero weights are scaled so that the weights in each row sum up to 1. In convention, people usually use the row standardized weight matrices because row standardization creates proportional weights in cases where features have an unequal number of neighbors; also this normalized matrix has nice properties in the range of eigenvalues (this will be mentioned later). As LeSage [27] suggests, there is a vast number of ways to define neighbors and to construct a weight matrix. In the following we discuss some commonly used methods in lattice cases and non-lattice cases. In a lattice case shown in the following Figure 1.1, we have 9 locations and we label them as $1, 2, \ldots, 9$ at left bottom corners in each cell. Suppose i is the target location and j identifies a neighbor of i.

- Rook Contiguity (Fig 1 (a)): two regions are neighbors if they share (part of) a common edge (on any side)
- Bishop Continuity (Fig 1 (b)): two regions are spatial neighbors if they share a common vertex (or a point)
- Queen Contiguity (Fig 1 (c)): this is the union of Rook and Bishop contiguity. Two regions are neighbors in this sense if they share any common edge or vertex

In practice, we may not always have a problem in a lattice. So an analog of an edge and a vertex is called "snap distance" [5] such that any border larger than this "snap distance" will be regarded as an edge or otherwise a vertex. So the Queen contiguity may be interpreted as that two regions are neighbors as long as they are connected no matter how short the common border is. Under the Queen criterion, for example, based on the example illustrated in Figure (1.1(c)), a 9×9 weight matrix for nine locations is shown below.

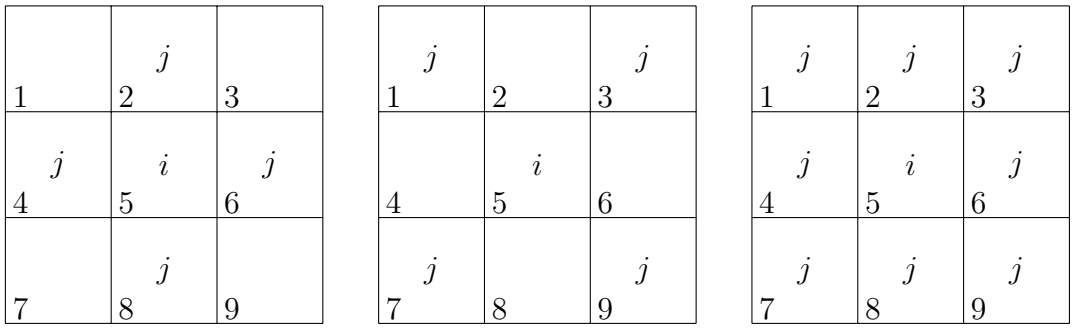


Figure 1.1. Examples of Rook (a), Bishop (b) and Queen Contiguity (c)

$$\begin{pmatrix} 0 & 1 & 0 & 1 & 1 & 0 & 0 & 0 & 0 \\ 1 & 0 & 1 & 1 & 1 & 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 1 & 1 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 & 1 & 0 & 1 & 1 & 0 \\ 1 & 1 & 1 & 1 & 0 & 1 & 1 & 1 & 1 \\ 0 & 1 & 1 & 0 & 1 & 0 & 0 & 1 & 1 \\ 0 & 0 & 0 & 1 & 1 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 & 1 & 1 & 1 & 0 & 1 \\ 0 & 0 & 0 & 0 & 1 & 1 & 0 & 1 & 0 \end{pmatrix} \tag{1.3}$$

However, in a non-lattice case when units, such as cities, are only points, this neighborhood definition does not work because all units/points do not share any common edge or vertex. So a distance based method is utilized to deal with such point case. Denote $d_{ij} \equiv d(i,j)$ as the distance between two units/points i and j, then some commonly used ways to define neighborhoods are

- Minimum Distance Neighbors:

  A neighbor j of unit i satisfies that their distance $d_{ij} \in \left(0, \max\limits_{i=\{1,\ldots,n\}} \min\limits_{j \neq i} d(i,j)\right]$.

This method controls that every unit has at least one neighbor but usually includes a large number of irrelevant connections.

- K-nearest Neighbors:

  Neighbors of i are restricted by the user-defined parameter K. A unit j is a neighbor of i if $j \in N_K(i)$, where $N_K(i)$ defines the K-nearest neighbors of i. This method also guarantees that there is no neighborless unit and has less noise then the Minimum Distance Neighbors. However, the user-choice parameter K may not reflect the true level of connectedness or isolation between points.

- Sphere of Influence Neighbors:

  For each point $i \in S = \{1, \ldots, n\}$, $r_i = \min_{k \neq i} d(i,k)$ and denote C_i as a circle of radius r_i centered at i. Units i and j are neighborhoods whenever C_i and C_j intersect in exactly two points. This graph-based method improves the K-nearest Neighbors in a way that relatively long links are avoided and the number of connections per unit is variable. This method works well even with irregularly located areal entities and precludes the intervention of user-defined parameter K in the previous method (See Figure 1.2).

According to Figure 1.2, the weight matrix for A, B, C and D is:

$$\begin{pmatrix} 0 & 1 & 0 & 0 \\ 1 & 0 & 1 & 1 \\ 0 & 1 & 0 & 1 \\ 0 & 1 & 1 & 0 \end{pmatrix}$$

To write our model (1.2) more explicitly, for each location s, $s = 1, 2, \ldots, n$

$$y_s = x_s'\boldsymbol{\beta} + \rho \sum_{i=1}^{n} w_{si} y_i + \sum_{i=1}^{h} \lambda_i F(x_s'\boldsymbol{\gamma}_i) + \varepsilon_s \tag{1.4}$$

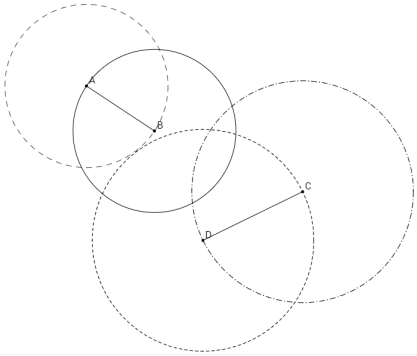


Figure 1.2. Sphere of Influence Graph: A,B,C,D represent four units. Where the circles around each city overlap in at least two points, the cities can be considered neighbors. In the current example, A is a neighbor of only B, B is a neighbor to all, C is a neighbor of B and D, D is a neighbor of B and C but not A.

The term $\sum_{i=1}^{h} \lambda_i F(x_s'\boldsymbol{\gamma}_i)$, a linear combination of logistic functions with weights $\lambda_i, i = 1, 2, \ldots, h$, forms a hidden layer of this neural network with h neurons (Medeiros, Teräsvirta, Rech [32]). This neural network helps discover nonlinear relationship between the response variable and its covariates.

1.2. Likelihood Function

Rewriting the equation in (1.2), we have

$$(I_n - \rho W_n)Y_n - X_n\beta - \boldsymbol{F}(X_n\boldsymbol{\gamma}')\lambda = \boldsymbol{\varepsilon}_n \tag{1.5}$$

where I_n is an $n \times n$ identity matrix. We denote $\boldsymbol{\theta} = (\beta_1, \ldots, \beta_q, \rho, \lambda_1, \ldots, \lambda_h, \boldsymbol{\gamma}_1', \ldots, \boldsymbol{\gamma}_h')' \in \mathbb{R}^{(q+1)(h+1)}$ with true value $\boldsymbol{\theta}_0$.

For the analysis of identification and estimation of this spatial autoregressive model (1.2), we adopt the following assumptions:

Assumption 1. The $(q+1)(h+1)$-dimensional parameter vector

$\boldsymbol{\theta} = (\beta', \rho, \lambda', \boldsymbol{\gamma}_1', \ldots, \boldsymbol{\gamma}_h')' \in \boldsymbol{\Theta}$, where $\boldsymbol{\Theta}$ is a subset of the $(q+1)(h+1)$- dimensional Euclidean space $\mathbb{R}^{(q+1)(h+1)}$. $\boldsymbol{\Theta}$ is a closed and bounded compact set and contains the true parameter value $\boldsymbol{\theta}_0$ as an interior point.

Assumption 2. The spatial correlation coefficient ρ satisfies $\rho \in (-1/\tau, 1/\tau)$, where $\tau = \max\{|\tau_1|, |\tau_2|, \ldots, |\tau_n|\}$, $\tau_1, \ldots, \tau_n$ are eigenvalues of spatial weight matrix W_n. To avoid the non-stationarity issue when ρ approaches to 1, we assume $\sup_{\rho \in \boldsymbol{\Theta}} |\rho| < 1$.

Assumption 3. We assume W_n is defined by queen contiguity and is uniformly bounded in row and column sums in absolute value as $n \to \infty$ so $(I_n - \rho W_n)^{-1}$ is also uniformly bounded in row and column sums as $n \to \infty$.

Assumption 4. X_n is stationary, ergodic satisfying $\mathbb{E}\,|x_s|^2 < \infty, s = 1, \ldots, n$ and X_n is full column rank.

Assumption 5. The error terms ε_s, $s = 1, 2, \ldots, n$ are independent and identically distributed with density function $f(\cdot)$, zero mean and unit variance $\sigma^2 = 1$. The moment $E(|\varepsilon_s|^{2+r})$ exists for some $r > 0$ and $E\,|\ln f(\varepsilon_s)| < \infty$.

Assumption 2 defines the parameter space for ρ as an interval around zero such that $I_n - \rho W_n$ is strictly diagonally dominant. By the Levy-Desplanques theorem [45], it follows that $I_n - \rho W_n$ is nonsingular for any values ρ in that interval.

Note that the diagonal entries in $I_n - \rho W_n$ are all 1 (because $w_{ii} = 0$). Using Gershgorin circle theorem [17, p. 749-754], we can show that the largest eigenvalue of a row-standardized matrix W_n is bounded by 1. Using the 9×9 non-standardized weight matrix (1.3) constructed under Queen's criterion in the section 2, the interval for ρ is $(-0.207, 0.207)$ whereas the row standardized weight matrix corresponds to $(-1, 1)$.

It is natural to consider the neighborhood by connections and in many practical studies, since entries scaled to sum up to 1, each row of W_n sums up to 1, which guarantees that all nonzero weights are in $(0, 1]$. For simplicity, we define the weight matrix W_n using the queen criterion and do row standardization. Assumption 3 is originated by Kelejian and Prucha (1998 [22], 2001 [23]) and is also used in Lee (2004 [25]). Restricting W_n to be uniformly bounded prevents the model prediction from exploding when n goes to infinity. By Lemma A.4 in Lee [25], we can prove that $(I_n - \rho W_n)^{-1}$ is also uniformly bounded in row and column sums for $\rho \in (-1/\tau, 1/\tau)$.

From Assumptions 2 and 3 we can also decompose W_n by its eigenvalue and eigenvector pairs τ_i, v_i: $W_n = P\Lambda P^{-1}$, where Λ is a diagonal matrix with eigenvalues τ_i on its diagonals and $P = [v_1, v_2, \ldots, v_n]$ (we assume v_i's are normalized eigenvectors). So

$$W_n = P \begin{pmatrix} \tau_1 & 0 & \cdots & 0 \\ 0 & \tau_2 & \cdots & 0 \\ 0 & 0 & \ddots & 0 \\ 0 & 0 & \cdots & \tau_n \end{pmatrix} P^{-1}, (I_n - \rho W_n)^{-1} = P \begin{pmatrix} \frac{1}{1-\rho\tau_1} & 0 & \cdots & 0 \\ 0 & \frac{1}{1-\rho\tau_2} & \cdots & 0 \\ 0 & 0 & \ddots & 0 \\ 0 & 0 & \cdots & \frac{1}{1-\rho\tau_n} \end{pmatrix} P^{-1} \tag{1.6}$$

This decomposition will later help us compute the likelihood function.

Assumption 4 guarantees the stationarity of $\{x_s\}$ so we can apply ergodic theorem later in the proofs.

Assumption 5 imposes restrictions for the random error. We assume that errors $\{\varepsilon_s\}_{s=1}^n$ have an identical density function $f(\cdot)$. So to derive the likelihood function of $\boldsymbol{\theta}$, it is necessary to introduce the Jacobian coefficient which allows us to derive the joint distribution of $Y_n = \{y_s\}_{s=1}^n$ from that of $\{\varepsilon_s\}_{s=1}^n$, through equation (1.5):

$$J = det(\partial \boldsymbol{\varepsilon}_n / \partial Y_n) = |I_n - \rho W_n| \tag{1.7}$$

Hence, based on the joint distribution for the vector of independent errors $\{\varepsilon_s\}_{s=1}^n$, and using (1.7) the log-likelihood function for $\boldsymbol{\theta}$ is given by (Anselin [3, p. 63])

$$\begin{aligned} \mathcal{L}_n(\boldsymbol{\theta}) &= \ln|I_n - \rho W_n| + \sum_{s=1}^{n} \ln f(\varepsilon_s(\boldsymbol{\theta})) \\ \varepsilon_s(\boldsymbol{\theta}) &= y_s - x_s'\beta - \rho\sum_{i=1}^{n} w_{si}y_i - \sum_{i=1}^{h}\lambda_i F(x_s'\boldsymbol{\gamma}_i) \end{aligned} \tag{1.8}$$

In practice, the density function f could be chosen by looking at the distribution for observations and model residuals $\varepsilon_s(\boldsymbol{\theta})$. Common choices are normal distribution, t-distribution and Laplace distribution. We examined these three distributions (with unit variances under Assumption 5) and the corresponding log-likelihood functions functions are given below.

When $\varepsilon_s \sim N(0,1)$,

$$f(\varepsilon_s) = \frac{1}{\sqrt{2\pi}}\exp(-\frac{\varepsilon_s^2}{2})$$

$$\mathcal{L}_n(\boldsymbol{\theta}) = \ln|I_n - \rho W_n| - \frac{n}{2}\ln(2\pi) - \frac{1}{2}\sum_{s=1}^{n}\varepsilon_s^2(\boldsymbol{\theta})$$

When ε_s has the rescaled t distribution with degree of freedom ν ($\nu > 2$, known) which is symmetric about zero and has variance 1:

$$f(\varepsilon_s) = \sqrt{\frac{\nu}{\nu-2}}\frac{\Gamma[\frac{1}{2}(\nu+1)]}{\sqrt{\nu\pi}\,\Gamma(\frac{1}{2}\nu)}\cdot\left(1+\frac{\varepsilon_s^2}{\nu-2}\right)^{-\frac{1+\nu}{2}}$$

$$\mathcal{L}_n(\boldsymbol{\theta}) = \ln|I_n - \rho W_n| - \frac{n}{2}\ln(\nu-2)\pi + n\ln\frac{\Gamma[\frac{1}{2}(\nu+1)]}{\Gamma(\frac{1}{2}\nu)} - \frac{1+\nu}{2}\sum_{s=1}^{n}\ln\big(1+\frac{\varepsilon_s^2(\boldsymbol{\theta})}{\nu-2}\big)$$

When $\varepsilon_s \sim$ Laplace distribution with mean $\mu = 0$ and scale parameter $b = \sqrt{2}/2$,

$$f(\varepsilon_s) = \frac{1}{\sqrt{2}}\exp\Big(-\sqrt{2}|\varepsilon_s|\Big)$$

$$\mathcal{L}_n(\boldsymbol{\theta}) = \ln|I_n - \rho W_n| - \frac{n}{2}\ln 2 - \sum_{s=1}^{n}\sqrt{2}|\varepsilon_s(\boldsymbol{\theta})|$$

In the following sections, we will discuss model identifiability and establish asymptotic properties for the maximum likelihood estimator $\hat{\boldsymbol{\theta}} = \arg\max_{\boldsymbol{\theta}\in\boldsymbol{\Theta}} \mathcal{L}_n(\boldsymbol{\theta})$.

1.3. Model Identification

We now investigate the conditions under which our proposed model is identified. By Rothenberg [38], a parameter $\theta_0 \in \boldsymbol{\Theta}$ is *globally identified* if there is no other θ in $\boldsymbol{\Theta}$ that is observationally equivalent to θ_0 such that $f(y;\theta) = f(y;\theta_0)$; or the parameter θ_0 is *locally identified* if there is no such θ in an open neighborhood of θ_0 in $\boldsymbol{\Theta}$. The model (1.4), in principle, is neither globally nor locally identified and the lack of identification of Neural Network models has been discussed in many papers [20], [32]. Here we extend the discussion to our proposed PSAR model. Three characteristics imply non-identification of our model: (a) the interchangeable property: the value in the likelihood function may remain unchanged if we permute the hidden units. For a model with h neurons, this will result in $h!$ different models that are indistinguishable from each other and have equal local maximums of the log-likelihood function; (b) the "symmetry" property: for a logistic function, $F(x) = 1 - F(-x)$ allows two equivalent parametrization for each of the hidden units; (c) the reducible property: the presence of irrelevant neurons in model (1.4) happens when $\lambda_i = 0$ so parameters $\boldsymbol{\gamma}_i$ in this neuron would remain unidentified. Conversely, if $\boldsymbol{\gamma}_i = \mathbf{0}$, the output of that sigmoid function is a constant so λ_i can take any value without affecting the value of likelihood functions.

The problem of interchangeability (as mentioned in (a)) can be solved by imposing the following restriction, as in Medeiros *et al.* [32]:

Restriction 1. *parameters* $\lambda_1, \ldots, \lambda_h$ *are restricted such that:* $\lambda_1 \geq \cdots \geq \lambda_h$.

And to tackle (b) and (c), we can apply another restriction:

Restriction 2. *The parameters λ_i and γ_{i1} should satisfy:*

(1) $\lambda_i \neq 0$, $\forall i \in \{1, 2, \ldots, h\}$; *and*

(2) $\gamma_{i1} > 0$, $\forall i \in \{1, 2, \ldots, h\}$.

To guarantee the non-singularity of model matrices and the uniqueness of parameters, we impose the following basic assumption:

Assumption 6. The true parameter vector $\boldsymbol{\theta}_0$ satisfies Restrictions 1-2.

Referring to the section 4.3 by Medeiros *et al.* [32], we can conclude the identifiability of the PSAR-ANN model

Lemma 1. Under the Assumptions 1-6, this partially specified spatial autoregressive model (1.4) is globally identified.

1.4. Asymptotic Results

1.4.1. Preliminary

Denote the true parameter vector as $\boldsymbol{\theta}_0$ and the solution which maximizes the log-likelihood function (1.8) as $\hat{\boldsymbol{\theta}}_n$. Hence, $\hat{\boldsymbol{\theta}}_n$ should satisfy

$$\hat{\boldsymbol{\theta}}_n \equiv \arg\max_{\boldsymbol{\theta}\in\boldsymbol{\Theta}} \mathcal{L}_n(\boldsymbol{\theta}) \ ,$$

$$\mathcal{L}_n(\boldsymbol{\theta}) = \ln|I_n - \rho W_n| + \sum_{s=1}^{n} \ln f\big(y_s - x_s'\beta - \rho\sum_{i=1}^{n} w_{si}y_i - \sum_{i=1}^{h} \lambda_i F(x_s'\gamma_i)\big)$$

Suppose we have a $n_1 \times n_2$ lattice where we consider asymptotic properties of $\hat{\boldsymbol{\theta}}_n$ when $n = n_1 n_2 \to \infty$. Write the location s as the coordinate (s_x, s_y) in the $[1, n_1] \times [1, n_2]$ lattice space. The distance between two locations s, j is defined as $d(s, j) = \max(|s_x -$

$j_x|, |s_y - j_y|)$. So if observations at s, j locations are neighbors (by queen criterion), their coordinates should satisfy $(s_x - j_x)^2 + (s_y - j_y)^2 \leq 2$ or $d(s, j) = 1$.

In a spatial context, we should notice that the functional form of y_s is not identical for all the locations due to values of the weights w_{si}. For example, in a lattice, units at edges, vertexes or in the interior have different density functions due to different neighborhood structures (Figure 1.3). For an interior point (Figure 1.3(c)), its neighborhood set $\mathcal{N}_s$ contains eight neighbors where $w_{sj} = 1/8$ if $d(s, j) = 1$ otherwise $w_{sj} = 0$, for $j = 1, 2, \ldots, n$. Similarly, an edge point (Figure 1.3(b)) has five neighboring units with $w_{sj} = 1/5$ and the weight of a vertex neighborhood is $1/3$ because a vertex unit has only three neighbors. This is known as an edge effect in spatial problems. To deal with this, referring

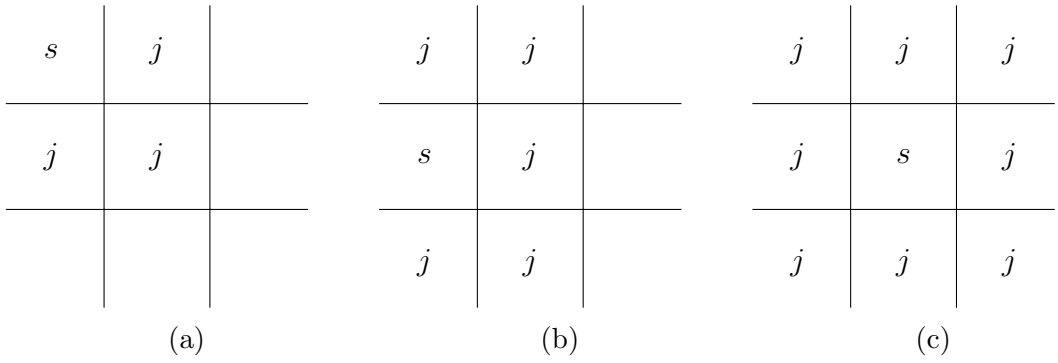


Figure 1.3. Vertex (a), Edge (b) and Interior Points (c) Neighborhood Structures: s is the target location and j represents the neighborhood of s

to Yao and Brockwell [50], we construct an edge effect correction scheme based on the way that the sample size tends to infinity. In a space $[1, n_1] \times [1, n_2]$, we consider its interior area as $\mathcal{S} = \{(s_x, s_y) : b_1 \leq s_x \leq n_1 - b_1, b_2 \leq s_y \leq n_2 - b_2\}$, where $b_1, b_2, n_1, n_2 \to \infty$ satisfying that $b_1/n_1, b_2/n_2 \to 0$ and other locations belong to the boundary areas $\mathcal{M}$. Therefore the set $\mathcal{S}$ contains $n^* = (n_1 - 2b_1)(n_2 - 2b_2)$ interior locations while the set $\mathcal{M}$ contains $n - n^*$ boundary locations. Then $n^*/n \to 1$ and $\mathcal{L}_n(\boldsymbol{\theta})$ can be split into a sum

of two parts (interior $\mathcal{S}$ and boundary $\mathcal{M}$ parts):

$$\mathcal{L}_n(\boldsymbol{\theta}) = \sum_{s\in\mathcal{M}} l(\boldsymbol{\theta}|x_s, y_s) + \sum_{s\in\mathcal{S}} l(\boldsymbol{\theta}|x_s, y_s)$$

$$l(\boldsymbol{\theta}|x_s, y_s) = n^{-1}\ln|I_n - \rho W_n| + \ln f\big(y_s - x_s'\beta - \rho\sum_{i=1}^{n} w_{si}y_i - \sum_{i=1}^{h}\lambda_i F(x_s'\boldsymbol{\gamma}_i)\big)$$

Therefore, given that $\lim_{n_1,n_2\to\infty}\frac{|\mathcal{M}|}{n} = 0$, $n^{-1}\sum_{s\in\mathcal{M}} l(\boldsymbol{\theta}|x_s, y_s)$ vanishes a.s. as n tends to infinity for any $\boldsymbol{\theta}\in\boldsymbol{\Theta}$. Therefore,

$$\begin{aligned}\lim_{n_1,n_2\to\infty} n^{-1}\mathcal{L}_n(\boldsymbol{\theta}) &= \lim_{n_1,n_2\to\infty}(n_1n_2)^{-1}\Big(\sum_{s\in\mathcal{M}} l(\boldsymbol{\theta}|x_s, y_s) + \sum_{s\in\mathcal{S}} l(\boldsymbol{\theta}|x_s, y_s)\Big)\\ &= \lim_{n_1,n_2\to\infty}(n_1n_2)^{-1}\sum_{s\in\mathcal{S}} l(\boldsymbol{\theta}|x_s, y_s) \quad a.s.\end{aligned}$$

In this equation, every location $s\in\mathcal{S}$ has eight neighboring units under the queen criterion with nonzero weights $w_{sj} = 1/8$. Hence for an interior unit $s\in\mathcal{S}$, $\sum_{i=1}^{n} w_{si}y_i = \sum_{j=1}^{n}\frac{1}{8}y_jI_{\{d(s,j)=1\}}$. And the log likelihood function $\mathcal{L}_n(\boldsymbol{\theta})$ is approximately

$$n^{-1}\mathcal{L}_n(\boldsymbol{\theta}) \approx n^{-1}\sum_{s\in\mathcal{S}} l(\boldsymbol{\theta}|x_s, y_s) \quad \text{for large } n \tag{1.9}$$

So the maximum likelihood estimator $\hat{\boldsymbol{\theta}}_n$ approximately maximizes $n^{-1}\sum_{s\in\mathcal{S}} l(\boldsymbol{\theta}|x_s, y_s)$.

$$\hat{\boldsymbol{\theta}}_n \approx \arg\max_{\boldsymbol{\theta}\in\boldsymbol{\Theta}} n^{-1}\sum_{s\in\mathcal{S}} l(\boldsymbol{\theta}|x_s, y_s)$$

1.4.2. Consistency Results

To establish the consistency of $\hat{\boldsymbol{\theta}}_n$, the heuristic insight is that because $\hat{\boldsymbol{\theta}}_n$ maximizes $n^{-1}\mathcal{L}_n(\boldsymbol{\theta})$, it approximately maximizes $n^{-1}\sum_{s\in\mathcal{S}} l(\boldsymbol{\theta}|x_s, y_s)$. By (1.9), $n^{-1}\mathcal{L}_n(\boldsymbol{\theta})$ can generally be shown tending to a real function $\mathcal{L}:\boldsymbol{\Theta}\to\mathbb{R}$ with maximizer $\boldsymbol{\theta}_0$ as $n\to\infty$ under mild conditions on the data generating process, then $\hat{\boldsymbol{\theta}}_n$ should tend to $\boldsymbol{\theta}_0$ almost surely.

Before the formal proof of the consistency, we need the following assumptions on density function $f(\cdot)$ satisfied (similar assumptions are made in White [48], Andrews, Davis and Breidt [2], Lii and Rosenblatt [29]).

Assumption 7. For all $s \in \mathbb{R}$, $f(s) > 0$ and $f(s)$ is twice continuously differentiable with respect to s.

Assumption 8. The density should satisfy the following equations:

- $\int s f'(s)\, ds = s f(s)|_{-\infty}^{\infty} - \int f(s)\, ds = -1$
- $\int f''(s)\, ds = f'(s)|_{-\infty}^{\infty} = 0$
- $\int s^2 f''(s)\, ds = s^2 f'(s)|_{-\infty}^{\infty} - 2\int s f'(s)\, ds = 2$

Assumption 9. The density should follow the following dominance condition: $\left|\frac{f'(s)}{f(s)}\right|$, $\left|\frac{f'(s)}{f(s)}\right|^2$, $\left|\frac{f'(s)}{f(s)}\right|^4$, $\frac{f''(s)}{f(s)}$, and $\frac{f''(s)f'^2(s)}{f^3(s)}$ are dominated by $a_1 + a_2 |s|^{c_1}$, where a_1, a_2, c_1 are non-negative constants and $\int_{-\infty}^{\infty} |s|^{c_1+2} f(s)\, ds < \infty$.

Assumption 10. If $c_1 > 2$ in previous assumption, we further assume $\mathbb{E}\,|x_s|^{c_1} < \infty$.

Discussed in Breidt, Davis, Lii and Rosenblatt [8] and Andrews, Davis and Breidt [2, p. 1642-1645]), these assumptions on the density $f(\cdot)$ are satisfied in the t-distribution case when $\nu > 2$ and the mixed Gaussian distribution. The assumption $\mathbb{E}\,|\ln f(s)| < \infty$ (see Assumption 5) is also checked satisfied in the normal and t-distribution ($\nu > 2$). The Laplace distribution does not strictly satisfy the Assumptions 7-9, since it is not differentiable at 0 but it satisfies these boundedness conditions almost everywhere so we believe the consistency and asymptotic normality results remain valid for parameter estimates. This will be shown in the simulation section. Assumption 10 is a necessary to boundedness conditions in later proof.

Lemma 2. Given Assumptions 1-9,

$$\boldsymbol{\theta}_0 = \max_{\boldsymbol{\theta}\in\boldsymbol{\Theta}} \mathbb{E}\,\mathcal{L}_n(\boldsymbol{\theta}) \equiv \max_{\boldsymbol{\theta}\in\boldsymbol{\Theta}} \mathbb{E}\,\frac{\mathcal{L}_n(\boldsymbol{\theta})}{n} \quad \text{for all } n \tag{1.10}$$

PROOF. $\mathcal{L}_n(\boldsymbol{\theta})$ is the log of the likelihood function $L_n(\boldsymbol{\theta})$,

$$\mathcal{L}_n(\boldsymbol{\theta}) = \ln|I_n - \rho W_n| + \sum_{s=1}^{n} \ln f(\varepsilon_s(\boldsymbol{\theta}))$$

$$\mathbb{E}\,\mathcal{L}_n(\boldsymbol{\theta}) - \mathbb{E}\,\mathcal{L}_n(\boldsymbol{\theta}_0) = \mathbb{E}\ln\frac{L_n(\boldsymbol{\theta})}{L_n(\boldsymbol{\theta}_0)}$$

Denote $Z_n = (Y_n, X_n)$. By Jensen's inequality,

$$\mathbb{E}\ln\frac{L_n(\boldsymbol{\theta})}{L_n(\boldsymbol{\theta}_0)} \le \ln\mathbb{E}\frac{L_n(\boldsymbol{\theta})}{L_n(\boldsymbol{\theta}_0)} = \ln\int_{-\infty}^{\infty}\frac{L_n(\boldsymbol{\theta})}{L_n(\boldsymbol{\theta}_0)}L_n(\boldsymbol{\theta}_0)\,dZ_n = 0$$

So $\mathbb{E}\mathcal{L}_n(\boldsymbol{\theta}) \le \mathbb{E}\mathcal{L}_n(\boldsymbol{\theta}_0)$. By Lemma 1, the PSAR model is globally identified and therefore, $\mathbb{E}\mathcal{L}_n(\boldsymbol{\theta})$ is uniquely maximized at $\boldsymbol{\theta}_0$ for all n. Since the parameter vector $\boldsymbol{\theta}_0$ does not depend on sample size n, it is equivalent to say that, $\boldsymbol{\theta}_0 = \max_{\boldsymbol{\theta}\in\boldsymbol{\Theta}} \frac{1}{n}\mathbb{E}\mathcal{L}_n(\boldsymbol{\theta})$. □

In the following, to simplify the expression, denote $g(x_s, \boldsymbol{\theta}) = x_s'\beta + \boldsymbol{F}(x_s'\boldsymbol{\gamma})\lambda$. Define the Hadamard product $\circ$ as,

$$a \circ B = \begin{bmatrix} a_1b_{11} & a_1b_{21} & \cdots & a_1b_{n1} \\ a_2b_{12} & a_2b_{22} & \cdots & a_2b_{n2} \\ \vdots & \vdots & \ddots & \vdots \\ a_nb_{1n} & a_nb_{2n} & \cdots & a_nb_{nn} \end{bmatrix}, a \circ b_1 = \begin{bmatrix} a_1b_{11} \\ a_2b_{12} \\ \vdots \\ a_nb_{1n} \end{bmatrix}$$

where $a, b_1, \ldots, b_n \in \mathbb{R}^n$, a matrix $B = (b_1, \ldots, b_n) \in \mathbb{R}^{n\times n}$. And let

$$k_0 = \int \left| \frac{f'(s)}{f(s)} \right| f(s)\, ds$$

$$k_1 = \int \left| \frac{f'^2(s)}{f^2(s)} - \frac{f''(s)}{f(s)} \right| f(s)\, ds$$

$$k_2 = \int \left| \frac{sf'^2(s)}{f(s)} - \frac{sf''(s)}{f(s)} \right| f(s)\, ds$$

$$k_3 = \int \left| \frac{s^2f'^2(s)}{f(s)} - \frac{s^2f''(s)}{f(s)} \right| f(s)\, ds$$

Lemma 3. Given Assumptions 1-10

$$\sup_{\boldsymbol{\theta}\in\boldsymbol{\Theta}} \left| \frac{1}{n}\sum_{s=1}^{n} \ln f(\varepsilon_s(\boldsymbol{\theta})) - \mathbb{E}\frac{1}{n}\sum_{s=1}^{n} \ln f(\varepsilon_s(\boldsymbol{\theta})) \right| \xrightarrow{p} 0 \text{ as } n \to \infty \tag{1.11}$$

Proof. As illustrated in equation (1.9), in a lattice with size $n_1 \times n_2$,

$$\sup_{\boldsymbol{\theta}\in\boldsymbol{\Theta}} \left| \frac{1}{n}\sum_{s=1}^{n} \ln f(\varepsilon_s(\boldsymbol{\theta})) - \frac{1}{n}\sum_{s\in\mathcal{S}} \ln f(\varepsilon_s(\boldsymbol{\theta})) \right| \xrightarrow{a.s.} 0 \text{ as } n_1, n_2 \to \infty$$

Therefore, to prove (1.11) is equivalent to show that

$$\sup_{\boldsymbol{\theta}\in\boldsymbol{\Theta}} \left| \frac{1}{n}\sum_{s\in\mathcal{S}} \ln f(\varepsilon_s(\boldsymbol{\theta})) - \mathbb{E}\frac{1}{n}\sum_{s\in\mathcal{S}} \ln f(\varepsilon_s(\boldsymbol{\theta})) \right| \xrightarrow{p} 0 \text{ as } n \to \infty \tag{1.12}$$

where $\mathcal{S}$ denotes the interior units mentioned before. Since the interior units have the same neighboring structure, the space process for them is stationary when n_1, n_2 go to infinity. We first show $\left|\frac{1}{n}\sum_{s\in\mathcal{S}} \ln f(\varepsilon_s(\boldsymbol{\theta})) - \mathbb{E}\frac{1}{n}\sum_{s\in\mathcal{S}} \ln f(\varepsilon_s(\boldsymbol{\theta}))\right| \xrightarrow{p} 0$ for fixed $\boldsymbol{\theta}$ (Similar proof in [25, Theorem 3.1,4.1]).

To prove this, we want to show that $\mathbb{E}|\ln f(\varepsilon_s(\boldsymbol{\theta}))| < \infty, s \in \mathcal{S}$. Expanding $\ln f(\varepsilon_s(\boldsymbol{\theta}))$ around $\boldsymbol{\theta}_0$ with respect to $\boldsymbol{\theta}$,

$$\ln f(\varepsilon_s(\boldsymbol{\theta})) = \ln f(\varepsilon_s(\boldsymbol{\theta}_0)) + \left| \frac{f'(\varepsilon_s(\tilde{\boldsymbol{\theta}}_n))}{f(\varepsilon_s(\tilde{\boldsymbol{\theta}}_n))} \frac{\partial \varepsilon_s(\tilde{\boldsymbol{\theta}}_n)}{\partial \boldsymbol{\theta}'} \right| (\boldsymbol{\theta} - \boldsymbol{\theta}_0)$$

$$\mathbb{E}|\ln f(\varepsilon_s(\boldsymbol{\theta}))| \le \mathbb{E}|\ln f(\varepsilon_s(\boldsymbol{\theta}_0))| + \mathbb{E}\left| \frac{f'(\varepsilon_s(\tilde{\boldsymbol{\theta}}_n))}{f(\varepsilon_s(\tilde{\boldsymbol{\theta}}_n))} \frac{\partial \varepsilon_s(\tilde{\boldsymbol{\theta}}_n)}{\partial \boldsymbol{\theta}'} \right| |\boldsymbol{\theta} - \boldsymbol{\theta}_0|$$

where $\tilde{\boldsymbol{\theta}}_n$ is between $\boldsymbol{\theta}$ and $\boldsymbol{\theta}_0$. Under the true parameter values $\varepsilon_s(\boldsymbol{\theta}_0)$ (denoted as ε_s or $\boldsymbol{\varepsilon}_n$ as its vector form in the following) is independent and identically distributed. From Assumption 5, $\mathbb{E}\left|\ln f(\varepsilon_s)\right| < \infty$. For $\mathbb{E}\left|\frac{f'(\varepsilon_s(\tilde{\boldsymbol{\theta}}))}{f(\varepsilon_s(\tilde{\boldsymbol{\theta}}))}\frac{\partial \varepsilon_s(\tilde{\boldsymbol{\theta}})}{\partial \boldsymbol{\theta}'}\right|$, $\left|\frac{\partial \varepsilon_s(\tilde{\boldsymbol{\theta}})}{\partial \boldsymbol{\theta}}\right|$ can be expressed as

$$
\begin{aligned}
&\left|\frac{\partial \varepsilon_s(\tilde{\boldsymbol{\theta}})}{\partial \beta}\right| = |x_s| \\
&\left|\frac{\partial \varepsilon_s(\tilde{\boldsymbol{\theta}})}{\partial \lambda}\right| = \left|\boldsymbol{F}(x_s'\tilde{\boldsymbol{\gamma}})'\right| \leq \mathbf{1}_h \\
&\left|\frac{\partial \varepsilon_s(\tilde{\boldsymbol{\theta}})}{\partial \boldsymbol{\gamma}_i}\right| = \left|\tilde{\lambda}_i \frac{\partial F(x_s'\tilde{\boldsymbol{\gamma}}_i)}{\partial x_s'\boldsymbol{\gamma}_i} x_s\right| = \left|\tilde{\lambda}_i F(x_s'\tilde{\boldsymbol{\gamma}}_i)(1 - F(x_s'\tilde{\boldsymbol{\gamma}}_i))x_s\right| \\
&\qquad\qquad \leq \max_{\lambda_i \in \boldsymbol{\Theta}} |\lambda_i| \frac{|x_s|}{4} \\
&\left|\frac{\partial \varepsilon_s(\tilde{\boldsymbol{\theta}})}{\partial \rho}\right| = \left|\sum_{i=1}^{n} w_{si} y_i\right| = \left|[M_n(\boldsymbol{g}(X_n, \tilde{\boldsymbol{\theta}}_n) + \boldsymbol{\varepsilon}_n(\tilde{\boldsymbol{\theta}}_n))]_s\right| = \left|\sum_{k=1}^{n} m_{sk}(g(x_k, \tilde{\boldsymbol{\theta}}_n) + \varepsilon_k(\tilde{\boldsymbol{\theta}}_n))\right|
\end{aligned}
\tag{1.13}
$$

where m_{ij} is (i,j) element of $M_n = W_n(I_n - \rho W_n)^{-1}$. M_n is bounded uniformly in column and row sums (see Assumption 3) so $\sum_{j=1}^{n} m_{ij}, \sum_{i=1}^{n} m_{ij}$ are bounded by a constant b for $i, j = 1, \ldots, n$. The logistic function $F(x)$ is bounded by 1 and its derivative $F'(x)$ is also bounded by 1. Consider $\boldsymbol{\varepsilon}_n(\tilde{\boldsymbol{\theta}}_n)$,

$$
\begin{aligned}
|\boldsymbol{\varepsilon}_n(\tilde{\boldsymbol{\theta}}_n)| &= \left|(I_n - \tilde{\rho}W_n)Y_n - \boldsymbol{g}(X_n, \tilde{\boldsymbol{\theta}}_n)\right| \\
&= \left|\boldsymbol{\varepsilon}_n + (\rho_0 - \tilde{\rho})W_n Y_n + (\boldsymbol{g}(X_n, \boldsymbol{\theta}_0) - \boldsymbol{g}(X_n, \tilde{\boldsymbol{\theta}}_n))\right| \\
&= |\boldsymbol{\varepsilon}_n + (\rho_0 - \tilde{\rho})M_n\boldsymbol{\varepsilon}_n + (\rho_0 - \tilde{\rho})M_n X_n\beta_0 \\
&\quad + (\rho_0 - \tilde{\rho})M_n\boldsymbol{F}(X_n\boldsymbol{\gamma}_0')\lambda_0 + X_n(\beta_0 - \tilde{\beta}) + \boldsymbol{F}(X_n\boldsymbol{\gamma}_0')\lambda_0 - \boldsymbol{F}(X_n\tilde{\boldsymbol{\gamma}}')\tilde{\lambda}\Big| \\
&< |\boldsymbol{\varepsilon}_n + (\rho_0 - \tilde{\rho})M_n\boldsymbol{\varepsilon}_n + (\rho_0 - \tilde{\rho})M_n X_n\beta_0 \\
&\quad + (\rho_0 - \tilde{\rho})M_n\boldsymbol{F}(X_n\boldsymbol{\gamma}_0')\lambda_0 + X_n(\beta_0 - \tilde{\beta})\Big| + ||\lambda_0 - \tilde{\lambda}|| \cdot \mathbf{1}_n
\end{aligned}
$$

Denote $P(x^c)$ is a polynomial about x with highest order c. Since we have assumed that M_n is uniformly bounded in column and row sums so M_n is finite. By Assumption 9-10, $\left|\frac{f'(\varepsilon_s(\tilde{\boldsymbol{\theta}}_n))}{f(\varepsilon_s(\tilde{\boldsymbol{\theta}}_n))}\right| < a_1 + a_2|\varepsilon_s(\tilde{\boldsymbol{\theta}}_n)|^{c_1}$ and $\mathbb{E}\left|\frac{f'(\varepsilon_s)}{f(\varepsilon_s)}\right|, \mathbb{E}\left|\frac{f'(\varepsilon_s)}{f(\varepsilon_s)}\right|^2$ are dominated by $a_1 + a_2|\varepsilon_s|^{c_1}$, $\mathbb{E}|\varepsilon_s|^{c_1} < \infty$, $\mathbb{E}|x_s|^{c_1} < \infty$. Let $c^* = \max(1, c_1)$, then,

$$\mathbb{E}\left|\frac{f'(\varepsilon_s(\tilde{\boldsymbol{\theta}}_n))}{f(\varepsilon_s(\tilde{\boldsymbol{\theta}}_n))}\right|^2 < P(\mathbb{E}\ |\varepsilon_s|^{c^*}) + P(\mathbb{E}\ |x_s|^{c^*}) + Constant < \infty$$

So also $\mathbb{E}\left|\frac{f'(\varepsilon_s(\tilde{\boldsymbol{\theta}}_n))}{f(\varepsilon_s(\tilde{\boldsymbol{\theta}}_n))}\right| < \infty$. With Cauchy–Schwarz inequality [43] and the finite second moment of X_n, we can have,

$$\begin{aligned}
&\mathbb{E}\left|\frac{f'(\varepsilon_s(\tilde{\boldsymbol{\theta}}_n))}{f(\varepsilon_s(\tilde{\boldsymbol{\theta}}_n))}\frac{\partial\varepsilon_s(\tilde{\boldsymbol{\theta}}_n)}{\partial\beta}\right| = \mathbb{E}\left|\frac{f'(\varepsilon_s(\tilde{\boldsymbol{\theta}}_n))}{f(\varepsilon_s(\tilde{\boldsymbol{\theta}}_n))}x_s\right| < \left(\mathbb{E}\left|\frac{f'(\varepsilon_s(\tilde{\boldsymbol{\theta}}_n))}{f(\varepsilon_s(\tilde{\boldsymbol{\theta}}_n))}\right|^2 \mathbb{E}\,|x_s|^2\right)^{1/2} < \infty \\
&\mathbb{E}\left|\frac{f'(\varepsilon_s(\tilde{\boldsymbol{\theta}}_n))}{f(\varepsilon_s(\tilde{\boldsymbol{\theta}}_n))}\frac{\partial\varepsilon_s(\tilde{\boldsymbol{\theta}}_n)}{\partial\lambda}\right| = \mathbb{E}\left|\frac{f'(\varepsilon_s(\tilde{\boldsymbol{\theta}}_n))}{f(\varepsilon_s(\tilde{\boldsymbol{\theta}}_n))}\boldsymbol{F}(x_s'\tilde{\boldsymbol{\gamma}})'\right| \le \mathbb{E}\left|\frac{f'(\varepsilon_s(\tilde{\boldsymbol{\theta}}_n))}{f(\varepsilon_s(\tilde{\boldsymbol{\theta}}_n))}\mathbf{1}_h\right| < \infty \qquad (1.14) \\
&\mathbb{E}\left|\frac{f'(\varepsilon_s(\tilde{\boldsymbol{\theta}}_n))}{f(\varepsilon_s(\tilde{\boldsymbol{\theta}}_n))}\frac{\partial\varepsilon_s(\tilde{\boldsymbol{\theta}}_n)}{\partial\boldsymbol{\gamma}_i}\right| \le \mathbb{E}\left|\frac{f'(\varepsilon_s(\tilde{\boldsymbol{\theta}}_n))}{f(\varepsilon_s(\tilde{\boldsymbol{\theta}}_n))}\tilde{\lambda}_i x_s\right| < \infty \\
&\mathbb{E}\left|\frac{f'(\varepsilon_s(\tilde{\boldsymbol{\theta}}_n))}{f(\varepsilon_s(\tilde{\boldsymbol{\theta}}_n))}\frac{\partial\varepsilon_s(\tilde{\boldsymbol{\theta}}_n)}{\partial\rho}\right| = \mathbb{E}\left|\frac{f'(\varepsilon_s(\tilde{\boldsymbol{\theta}}_n))}{f(\varepsilon_s(\tilde{\boldsymbol{\theta}}_n))}\sum_{k=1}^{n} m_{sk}(g(x_k,\tilde{\boldsymbol{\theta}}_n) + \varepsilon_k(\tilde{\boldsymbol{\theta}}_n))\right| \\
&\qquad\qquad < b\cdot\mathbb{E}\left|\frac{\varepsilon_s(\tilde{\boldsymbol{\theta}}_n)f'(\varepsilon_s(\tilde{\boldsymbol{\theta}}_n))}{f(\varepsilon_s(\tilde{\boldsymbol{\theta}}_n))}\right| + k_0\mathbb{E}\left|\sum_{k=1}^{n} m_{sk}g(x_k,\tilde{\boldsymbol{\theta}}_n)\right|
\end{aligned}$$

Since $\mathbb{E}\,|x_s|^2 < \infty$ for all s, $\mathbb{E}|g(x_k,\tilde{\boldsymbol{\theta}}_n)|$ is finite for $\tilde{\boldsymbol{\theta}}_n \in \boldsymbol{\Theta}$. X_n is stationary with finite second moment, so component $\mathbb{E}\left|\sum_{k=1}^{n} m_{sk}g(x_k,\tilde{\boldsymbol{\theta}}_n)\right|$ is finite. $\mathbb{E}\left|\frac{\varepsilon_s(\tilde{\boldsymbol{\theta}}_n)f'(\varepsilon_s(\tilde{\boldsymbol{\theta}}_n))}{f(\varepsilon_s(\tilde{\boldsymbol{\theta}}_n))}\right|$ is dominated by $P(\mathbb{E}\,|\varepsilon_s|^{c^*+1})$ so with the dominance assumption, $\mathbb{E}\left|\frac{\varepsilon_s(\tilde{\boldsymbol{\theta}}_n)f'(\varepsilon_s(\tilde{\boldsymbol{\theta}}_n))}{f(\varepsilon_s(\tilde{\boldsymbol{\theta}}_n))}\right|$ is finite. Hence, with (1.14) finite, $\mathbb{E}|\ln f(\varepsilon_s(\boldsymbol{\theta}_0))| < \infty$, we can conclude that $\mathbb{E}|\ln f(\varepsilon_s(\boldsymbol{\theta}))| < \infty$. Then, by ergodic theorem,

$$\left|\frac{1}{n}\sum_{s\in\mathcal{S}}\ln f(\varepsilon_s(\boldsymbol{\theta})) - \mathbb{E}\frac{1}{n}\sum_{s\in\mathcal{S}}\ln f(\varepsilon_s(\boldsymbol{\theta}))\right| \to_p 0, \quad n\to\infty$$

To complete the proof of uniform convergence, we also need to show the equicontinuity of $\frac{1}{n}\sum_{s\in\mathcal{S}} \ln f(\varepsilon_s(\boldsymbol{\theta}))$, i.e., for all $\boldsymbol{\theta}_1, \boldsymbol{\theta}_2 \in \boldsymbol{\Theta}$,

$$\frac{1}{n}\left|\sum_{s\in\mathcal{S}} \ln f(\varepsilon_s(\boldsymbol{\theta}_1)) - \sum_{s\in\mathcal{S}} \ln f(\varepsilon_s(\boldsymbol{\theta}_2))\right| \leq ||\boldsymbol{\theta}_1 - \boldsymbol{\theta}_2|| O_p(1) \tag{1.15}$$

Applying the mean value theorem to the left side in (1.15):

$$\begin{aligned}\frac{1}{n}\left|\sum_{s\in\mathcal{S}} \ln f(\varepsilon_s(\boldsymbol{\theta}_1)) - \sum_{s\in\mathcal{S}} \ln f(\varepsilon_s(\boldsymbol{\theta}_2))\right| &\leq \frac{1}{n}\left|\sum_{s\in\mathcal{S}} \frac{\partial \ln f(\varepsilon_s(\tilde{\boldsymbol{\theta}}_n))}{\partial \boldsymbol{\theta}'}\right| ||\boldsymbol{\theta}_1 - \boldsymbol{\theta}_2|| \\ &= \frac{1}{n}\left|\sum_{s\in\mathcal{S}} \frac{f'(\varepsilon_s(\tilde{\boldsymbol{\theta}}_n))}{f(\varepsilon_s(\tilde{\boldsymbol{\theta}}_n))} \frac{\partial \varepsilon_s(\tilde{\boldsymbol{\theta}}_n)}{\partial \boldsymbol{\theta}'}\right| ||\boldsymbol{\theta}_1 - \boldsymbol{\theta}_2||\end{aligned}$$

where $\tilde{\boldsymbol{\theta}}_n$ is some value between $\boldsymbol{\theta}_1$ and $\boldsymbol{\theta}_2$.

By the ergodic theorem, $\frac{1}{n}\left|\sum_{s\in\mathcal{S}} \frac{f'(\varepsilon_s(\tilde{\boldsymbol{\theta}}_n))}{f(\varepsilon_s(\tilde{\boldsymbol{\theta}}_n))} \frac{\partial \varepsilon_s(\tilde{\boldsymbol{\theta}}_n)}{\partial \boldsymbol{\theta}}\right| \xrightarrow{a.s.} \mathbb{E}\left|\frac{f'(\varepsilon_s(\tilde{\boldsymbol{\theta}}))}{f(\varepsilon_s(\tilde{\boldsymbol{\theta}}_n))} \frac{\partial \varepsilon_s(\tilde{\boldsymbol{\theta}}_n)}{\partial \boldsymbol{\theta}}\right|$. Since $\boldsymbol{\theta}$ is in a compact set $\boldsymbol{\Theta}$, we show in (1.16) that, for all s, $\varepsilon_s(\boldsymbol{\theta})$ is bounded by some function of Y_n, X_n not depending on $\boldsymbol{\theta}$.

$$\begin{aligned}|\boldsymbol{\varepsilon}_n(\boldsymbol{\theta})| &= \left|Y_n - \rho W_n Y_n - X_n\beta - \boldsymbol{F}(X_n\boldsymbol{\gamma}')\lambda\right| \\ &\leq |(I_n - \rho W_n)Y_n| + |X_n\beta| + \left|\boldsymbol{F}(X_n\boldsymbol{\gamma}')\lambda\right| \\ &\leq (I_n + \max_{\rho\in\boldsymbol{\Theta}} |\rho| W_n)|Y_n| + |X_n| \max_{\beta\in\boldsymbol{\Theta}} |\beta| + \max_{\lambda\in\boldsymbol{\Theta}} ||\lambda|| \mathbf{1}_n\end{aligned} \tag{1.16}$$

Similarly, referring to (1.13), it is easy to show that $\left|\frac{\partial \varepsilon_s(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}}\right|$ is also bounded by some function about Y_n and X_n. Therefore, due to the dominance of $\left|\frac{f'(s)}{f(s)}\right|$ (see Assumption 9) and stationarity of X_n, Y_n, for $\tilde{\boldsymbol{\theta}}_n$ between $\boldsymbol{\theta}_1$ and $\boldsymbol{\theta}_2$, there exists a constant M such that

$$\frac{1}{n}\left|\sum_{s\in\mathcal{S}} \frac{f'(\varepsilon_s(\tilde{\boldsymbol{\theta}}_n))}{f(\varepsilon_s(\tilde{\boldsymbol{\theta}}_n))} \frac{\partial \varepsilon_s(\tilde{\boldsymbol{\theta}}_n)}{\partial \boldsymbol{\theta}'}\right| \leq M \quad \text{for } n \to \infty \tag{1.17}$$

Hence, for $\boldsymbol{\theta}_1, \boldsymbol{\theta}_2 \in \boldsymbol{\Theta}$

$$\frac{1}{n}\left|\sum_{s\in\mathcal{S}} \ln f(\varepsilon_s(\boldsymbol{\theta}_1)) - \sum_{s\in\mathcal{S}} \ln f(\varepsilon_s(\boldsymbol{\theta}_2))\right| = ||\boldsymbol{\theta}_1 - \boldsymbol{\theta}_2|| O_p(1)$$

So $\frac{1}{n}\left|\sum_{s\in\mathcal{S}} \ln f(\varepsilon_s(\boldsymbol{\theta}))\right|$ is equicontinuous for $\boldsymbol{\theta}\in\boldsymbol{\Theta}$. With the pointwise convergence and equicontinuity, we can conclude the uniform convergence in (1.12) and furthermore (1.11) follows. □

We now give a formal statement of consistency of the maximum likelihood estimator $\hat{\boldsymbol{\theta}}_n$.

Theorem 1. Given Assumptions 1-10, $\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0 \xrightarrow{p} 0$ as $n\to\infty$.

PROOF. Similar to the proof in Lung-fei Lee [25], we need to show the stochastic equicontinuity of $\frac{1}{n}\ln|I_n-\rho W_n|$ to have the uniform convergence of the log likelihood function $\mathcal{L}_n(\boldsymbol{\theta})$. Applying the mean value theorem,

$$\left|\frac{1}{n}(\ln|I_n-\rho_1 W_n|-\ln|I_n-\rho_2 W_n|)\right| = \left|(\rho_1-\rho_2)\frac{1}{n}tr(W_n(I_n-\tilde{\rho}_n W_n)^{-1})\right|$$

where $\tilde{\rho}_n$ is between ρ_1 and ρ_2. Since W_n is a row standardized matrix, the row sum equals to 1. By Assumption 2 and 3, $\sup_{\rho\in\boldsymbol{\Theta}}|\rho|<1$, W_n is bounded in both row and column sums uniformly and using (1.6),

$$\left|\frac{1}{n}tr(W_n(I_n-\tilde{\rho}_n W_n)^{-1})\right| = \left|\frac{1}{n}\sum_{i=1}^{n}\frac{\tau_i}{1-\tilde{\rho}_n\tau_i}\right| \le C_1$$

where C_1 is a constant not depending on n. So $\left|\frac{1}{n}(\ln|I_n-\rho_1 W_n|-\ln|I_n-\rho_2 W_n|)\right| \le |\rho_1-\rho_2|C_1$ and with Lemma 3 we can conclude the uniform convergence that

$$\sup_{\boldsymbol{\theta}\in\boldsymbol{\Theta}}\left|\frac{1}{n}\mathcal{L}_n(\boldsymbol{\theta}|Y_n,X_n)-\mathbb{E}\frac{1}{n}\mathcal{L}_n(\boldsymbol{\theta}|Y_n,X_n)\right| \xrightarrow{p} 0. \tag{1.18}$$

With Assumptions 1-9, the parameter space $\boldsymbol{\Theta}$ is compact; $\frac{1}{n}\mathcal{L}_n(\boldsymbol{\theta}|Y_n,X_n)$ is continuous in $\boldsymbol{\theta}\in\boldsymbol{\Theta}$ and is a measurable function of Y_n, X_n for all $\boldsymbol{\theta}\in\boldsymbol{\Theta}$. $\mathbb{E}\frac{1}{n}\mathcal{L}_n(\boldsymbol{\theta}|Y_n,X_n)$ is continuous on $\boldsymbol{\Theta}$ and by Lemma 2, $\mathbb{E}\frac{1}{n}\mathcal{L}_n(\boldsymbol{\theta}|Y_n,X_n)$ has a unique maximum at $\boldsymbol{\theta}_0$. Referring to Theorem 3.5 in White[47], with the uniform convergence in (1.18), we can conclude that $\hat{\boldsymbol{\theta}}_n-\boldsymbol{\theta}_0\xrightarrow{p}0$ as $n\to\infty$. □

1.4.3. Asymptotic Distribution

Assumption 11. The limit $A(\boldsymbol{\theta}_0) = -\lim_{n\to\infty} \mathbb{E}\frac{1}{n}\frac{\partial^2 \mathcal{L}_n(\boldsymbol{\theta}_0)}{\partial \boldsymbol{\theta}\partial \boldsymbol{\theta}'}$ is nonsingular.

Assumption 12. The limit $B(\boldsymbol{\theta}_0) = \lim_{n\to\infty} \mathbb{E}\frac{1}{n}\frac{\partial \mathcal{L}_n(\boldsymbol{\theta}_0)}{\partial \boldsymbol{\theta}}\frac{\partial \mathcal{L}_n(\boldsymbol{\theta}_0)}{\partial \boldsymbol{\theta}'}$ is nonsingular.

These assumptions are to guarantee the existence of the covariance matrix of the limiting distribution of parameters in a PSAR-ANN model. We now give the asymptotic distribution of the maximum likelihood estimator $\hat{\boldsymbol{\theta}}_n$.

Theorem 2. Under Assumptions 1-12,

$$\sqrt{n}(\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0) \xrightarrow{d} N(\mathbf{0}, \boldsymbol{\Omega}_0) \tag{1.19}$$

where $\boldsymbol{\Omega}_0 = A(\boldsymbol{\theta}_0)^{-1}B(\boldsymbol{\theta}_0)A(\boldsymbol{\theta}_0)^{-1} = A(\boldsymbol{\theta}_0)^{-1}$.

Proof. Since $\hat{\boldsymbol{\theta}}_n$ maximizes $\mathcal{L}_n(\boldsymbol{\theta})$, $\frac{\partial \mathcal{L}_n(\hat{\boldsymbol{\theta}}_n)}{\partial \boldsymbol{\theta}} = 0$. By the mean value theorem, expand $\frac{\partial \mathcal{L}_n(\hat{\boldsymbol{\theta}}_n)}{\partial \boldsymbol{\theta}}$ around $\boldsymbol{\theta}_0$ with respect to $\boldsymbol{\theta}$,

$$\frac{\partial \mathcal{L}_n(\hat{\boldsymbol{\theta}}_n)}{\partial \boldsymbol{\theta}} = \frac{\partial \mathcal{L}_n(\boldsymbol{\theta}_0)}{\partial \boldsymbol{\theta}} + \frac{\partial^2 \mathcal{L}_n(\tilde{\boldsymbol{\theta}}_n)}{\partial \boldsymbol{\theta}\partial \boldsymbol{\theta}'}(\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0)$$

$$0 = \frac{\partial \mathcal{L}_n(\boldsymbol{\theta}_0)}{\partial \boldsymbol{\theta}} + \frac{\partial^2 \mathcal{L}_n(\tilde{\boldsymbol{\theta}}_n)}{\partial \boldsymbol{\theta}\partial \boldsymbol{\theta}'}(\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0)$$

where $\tilde{\boldsymbol{\theta}}_n$ is between $\hat{\boldsymbol{\theta}}_n$ and $\boldsymbol{\theta}_0$. Therefore, we can have the following equation:

$$\sqrt{n}(\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0) = \left[-\frac{1}{n}\frac{\partial^2 \mathcal{L}_n(\tilde{\boldsymbol{\theta}}_n)}{\partial \boldsymbol{\theta}\partial \boldsymbol{\theta}'}\right]^{-1} \frac{1}{\sqrt{n}}\frac{\partial \mathcal{L}_n(\boldsymbol{\theta}_0)}{\partial \boldsymbol{\theta}} \tag{1.20}$$

We first show the limiting distribution of $\frac{1}{\sqrt{n}}\frac{\partial \mathcal{L}_n(\boldsymbol{\theta}_0)}{\partial \boldsymbol{\theta}}$. Under $\boldsymbol{\theta}_0$, $\boldsymbol{\varepsilon}_n(\boldsymbol{\theta}_0) = \boldsymbol{\varepsilon}_n$,

$$\boldsymbol{\varepsilon}_n(\boldsymbol{\theta}_0) = (I_n - \rho_0 W_n)Y_n - X_n\beta_0 - F(X_n\boldsymbol{\gamma}_0')\lambda_0 = \boldsymbol{\varepsilon}_n \tag{1.21}$$

Denote $\frac{\boldsymbol{f}'(\boldsymbol{\varepsilon}_n(\boldsymbol{\theta}))}{\boldsymbol{f}(\boldsymbol{\varepsilon}_n(\boldsymbol{\theta}))}$ as $V_n(\boldsymbol{\theta}) \in \mathbb{R}^n$ and $\frac{\boldsymbol{f}'(\boldsymbol{\varepsilon}_n)}{\boldsymbol{f}(\boldsymbol{\varepsilon}_n)}$ as $V_n \in \mathbb{R}^n$, then the first order derivatives are

$$\frac{1}{\sqrt{n}}\frac{\partial \mathcal{L}_n(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} = \begin{pmatrix} -\frac{1}{\sqrt{n}}\left((W_nY_n)'V_n(\boldsymbol{\theta}) + tr(W_n(I_n-\rho W_n)^{-1})\right) \\ -\frac{1}{\sqrt{n}}X_n'V_n(\boldsymbol{\theta}) \\ -\frac{1}{\sqrt{n}}(\boldsymbol{F}(X_n\boldsymbol{\gamma}')'V_n(\boldsymbol{\theta})) \\ -\frac{\lambda_1}{\sqrt{n}}X_n'(\boldsymbol{F}(X_n\boldsymbol{\gamma}_1)\circ V_n(\boldsymbol{\theta})) \\ \vdots \\ -\frac{\lambda_h}{\sqrt{n}}X_n'(\boldsymbol{F}(X_n\boldsymbol{\gamma}_h)\circ V_n(\boldsymbol{\theta})) \end{pmatrix} \tag{1.22}$$

By Lemma 2, the true parameter values maximize $\frac{1}{n}\mathbb{E}(\mathcal{L}_n(\boldsymbol{\theta}))$, so $\frac{1}{n}\frac{\partial \mathbb{E}(\mathcal{L}_n(\boldsymbol{\theta}_0))}{\partial \boldsymbol{\theta}} = \mathbf{0}$. In (1.14) and (1.16), we showed that $\mathbb{E}\left|\frac{\partial \ln f(\varepsilon_s(\boldsymbol{\theta}))}{\partial \boldsymbol{\theta}}\right|$ is dominated by some function not related to $\boldsymbol{\theta}$ and (1.17) indicates that $\mathbb{E}\left|\frac{\partial \ln f(\varepsilon_s(\boldsymbol{\theta}))}{\partial \boldsymbol{\theta}}\right|$ is bounded for interior units in $\mathcal{S}$. Hence, $\mathbb{E}\frac{\partial \ln f(\varepsilon_s(\boldsymbol{\theta}))}{\partial \boldsymbol{\theta}} = \frac{\partial}{\partial \boldsymbol{\theta}}\mathbb{E}\ln f(\varepsilon_s(\boldsymbol{\theta}))$, it follows that, with $\frac{1}{n}\mathcal{L}_n(\boldsymbol{\theta}) = \frac{1}{n}\ln|I_n-\rho_0W_n| + \frac{1}{n}\sum_{s=1}^n \ln f(\varepsilon_s(\boldsymbol{\theta}))$, we can have,

$$\frac{1}{n}\frac{\partial \mathbb{E}\mathcal{L}_n(\boldsymbol{\theta}_0)}{\partial \boldsymbol{\theta}} = \frac{1}{n}\mathbb{E}\frac{\partial \mathcal{L}_(\boldsymbol{\theta}_0)}{\partial \boldsymbol{\theta}} = \mathbf{0}$$

Therefore, with Assumption 12

$$Var\left(\frac{1}{\sqrt{n}}\frac{\partial \mathcal{L}_n(\boldsymbol{\theta}_0)}{\partial \boldsymbol{\theta}}\right) = -\mathbb{E}\frac{1}{n}\frac{\partial^2 \mathcal{L}_n(\boldsymbol{\theta}_0)}{\partial \boldsymbol{\theta}\partial \boldsymbol{\theta}'} = \mathbb{E}\frac{1}{n}\frac{\partial \mathcal{L}_n(\boldsymbol{\theta}_0)}{\partial \boldsymbol{\theta}}\frac{\partial \mathcal{L}_n(\boldsymbol{\theta}_0)}{\partial \boldsymbol{\theta}'} \to B(\boldsymbol{\theta}_0)$$

And under this $A(\boldsymbol{\theta}_0) = B(\boldsymbol{\theta}_0)$ when $n \to \infty$. From (1.22), we can see that $\frac{\partial \mathcal{L}_n(\boldsymbol{\theta}_0)}{\partial \boldsymbol{\theta}}$ is a sum of n identical and ergodic random variables. By the central limit theorem for stationary ergodic processes [33], we can conclude the limiting distribution of $\frac{1}{\sqrt{n}}\frac{\partial \mathcal{L}_n(\boldsymbol{\theta}_0)}{\partial \boldsymbol{\theta}}$ is $N(\mathbf{0}, B(\boldsymbol{\theta}_0))$.

Next, we want to show that $\frac{1}{n}\frac{\partial^2 \mathcal{L}_n(\tilde{\boldsymbol{\theta}}_n)}{\partial \boldsymbol{\theta}\partial \boldsymbol{\theta}'} - \frac{1}{n}\frac{\partial^2 \mathcal{L}_n(\boldsymbol{\theta}_0)}{\partial \boldsymbol{\theta}\partial \boldsymbol{\theta}'} \xrightarrow{p} 0$. Following the results in (1.22), define $U_n(\boldsymbol{\theta}) = \frac{\boldsymbol{f}''(\boldsymbol{\varepsilon}_n(\boldsymbol{\theta}))}{\boldsymbol{f}(\boldsymbol{\varepsilon}_n(\boldsymbol{\theta}))} - \frac{\boldsymbol{f}'^2(\boldsymbol{\varepsilon}_n(\boldsymbol{\theta}))}{\boldsymbol{f}^2(\boldsymbol{\varepsilon}_n(\boldsymbol{\theta}))} \in \mathbb{R}^n$ and $U_n(\boldsymbol{\theta}_0) = U_n$ so the second order derivatives

are given below $-\frac{1}{n}\frac{\partial^2 \mathcal{L}_n(\boldsymbol{\theta})}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}'} =$

(1.23)

$$\frac{1}{n}\begin{pmatrix}
G_0(\boldsymbol{\theta}) & (W_nY_n)'G_1(\boldsymbol{\theta}) & (W_nY_n)'G_2(\boldsymbol{\theta}) & (W_nY_n)'H_1(\boldsymbol{\theta}) & \cdots & (W_nY_n)'H_h(\boldsymbol{\theta}) \\
G_1'(\boldsymbol{\theta})W_nY_n & X_n'G_1(\boldsymbol{\theta}) & X_n'G_2(\boldsymbol{\theta}) & X_n'H_1(\boldsymbol{\theta}) & \cdots & X_n'H_h(\boldsymbol{\theta}) \\
G_2'(\boldsymbol{\theta})W_nY_n & G_2'(\boldsymbol{\theta})X_n & \boldsymbol{F}(X_n\boldsymbol{\gamma}')'G_2(\boldsymbol{\theta}) & \boldsymbol{F}(X_n\boldsymbol{\gamma}')'H_1(\boldsymbol{\theta}) & \cdots & \boldsymbol{F}(X_n\boldsymbol{\gamma}')'H_h(\boldsymbol{\theta}) \\
 & & & +K_1(\boldsymbol{\theta}) & \cdots & +K_h(\boldsymbol{\theta}) \\
H_1'(\boldsymbol{\theta})W_nY_n & H_1'(\boldsymbol{\theta})X_n & H_1'(\boldsymbol{\theta})\boldsymbol{F}(X_n\boldsymbol{\gamma}') & & & \\
 & & +K_1(\boldsymbol{\theta})' & & & \\
\vdots & \vdots & \vdots & & J(\boldsymbol{\theta}) & \\
H_h'(\boldsymbol{\theta})W_nY_n & H_h'(\boldsymbol{\theta})X_n & H_h'(\boldsymbol{\theta})\boldsymbol{F}(X_n\boldsymbol{\gamma}') & & & \\
 & & +K_h(\boldsymbol{\theta})' & & &
\end{pmatrix}$$

$$J_{ij}(\boldsymbol{\theta}) = \begin{cases} \lambda_i X_n'(\boldsymbol{F}''(X_n\boldsymbol{\gamma}_i) \circ V_n(\boldsymbol{\theta}) \circ X_n) + \lambda_i X_n'(\boldsymbol{F}'(X_n\boldsymbol{\gamma}_i) \circ H_i) & i = j \\ \lambda_i(\boldsymbol{F}'(X_n\boldsymbol{\gamma}_i) \circ H_j)'X_n & i > j \\ \lambda_i X_n'(\boldsymbol{F}'(X_n\boldsymbol{\gamma}_i) \circ H_j) & i < j \end{cases} \quad i,j = 1,2,\ldots,h$$

$$G_0(\boldsymbol{\theta}) = \left(-W_nY_n \circ W_nY_n\right)'U_n(\boldsymbol{\theta}) + tr((W_n(I_n - \rho W_n)^{-1})^2)$$

$$G_1(\boldsymbol{\theta}) = -U_n(\boldsymbol{\theta}) \circ X_n$$

$$G_2(\boldsymbol{\theta}) = -U_n(\boldsymbol{\theta}) \circ \boldsymbol{F}(X_n\boldsymbol{\gamma}')$$

$$H_i(\boldsymbol{\theta}) = -U_n(\boldsymbol{\theta}) \circ (\lambda_i \boldsymbol{F}'(X_n\boldsymbol{\gamma}_i) \circ X_n) \quad i = 1,\ldots,h$$

$$K_i(\boldsymbol{\theta}) = [V_n(\boldsymbol{\theta}) \circ \boldsymbol{F}'(X_n\boldsymbol{\gamma}')]'X_n \circ e_i \quad i = 1,\ldots,h \quad k = 1,\ldots,h$$

$$e_{i,k} = \begin{cases} 1 & k = i \\ 0 & k \neq i \end{cases}$$

Since $\tilde{\boldsymbol{\theta}}_n$ is between $\hat{\boldsymbol{\theta}}_n$ and $\boldsymbol{\theta}_0$, $\hat{\boldsymbol{\theta}}_n \xrightarrow{p} \boldsymbol{\theta}_0$ so $\tilde{\boldsymbol{\theta}}_n$ also converges to $\boldsymbol{\theta}_0$ in probability as $n \to \infty$. By Assumption 9, $\left|\frac{f'(s)}{f(s)}\right|, \left|\frac{f''(s)}{f(s)}\right|$ and $\left|\frac{f'^2(s)}{f^2(s)}\right|$ are continuous and are bounded by $a_1 + a_2 |s|^{c_1}$ so $V_n(\boldsymbol{\theta}), U_n(\boldsymbol{\theta})$ are continuous. With $\rho \in (-\frac{1}{\tau}, \frac{1}{\tau})$, $tr((W_n(I_n - \rho W_n)^{-1})^2) = \sum_{i=1}^n \frac{\tau_i^2}{(1-\rho\tau_i)^2}$ is also a continuous function of ρ.

Therefore elements in $\frac{1}{n}\frac{\partial^2 \mathcal{L}_n(\boldsymbol{\theta})}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}'}$ are continuous functions for $\boldsymbol{\theta}$ in $\boldsymbol{\Theta}$. By the continuity,

$$\frac{1}{n}\frac{\partial^2 \mathcal{L}_n(\tilde{\boldsymbol{\theta}}_n)}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}'} - \frac{1}{n}\frac{\partial^2 \mathcal{L}_n(\boldsymbol{\theta}_0)}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}'} \xrightarrow{p} 0, \quad \text{as } \tilde{\boldsymbol{\theta}}_n \xrightarrow{p} \boldsymbol{\theta}_0 \tag{1.24}$$

Finally, show that $\left|\frac{1}{n}\frac{\partial^2 \mathcal{L}_n(\boldsymbol{\theta}_0)}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}'} - \mathbb{E}\frac{1}{n}\frac{\partial^2 \mathcal{L}_n(\boldsymbol{\theta}_0)}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}'}\right| \xrightarrow{p} 0$.

Since Y_n, X_n are stationary, we can first show that for each s,

$$\mathbb{E}\left|\frac{\partial^2}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}'}\left(\frac{1}{n}\sum_{i=1}^n \ln(1-\rho_0\tau_i) + \ln f(\varepsilon_s(\boldsymbol{\theta}_0))\right)\right| < \infty \tag{1.25}$$

We first discuss the expected value of second derivative with respect to ρ in (1.25). By triangular inequality, $\mathbb{E}\left|\frac{\partial^2}{\partial\rho\partial\rho}\left(\frac{1}{n}\sum_{i=1}^n \ln(1-\rho_0\tau_i) + \ln f(\varepsilon_s(\boldsymbol{\theta}_0))\right)\right| < \mathbb{E}\left|\frac{1}{n}\sum_{i=1}^n \frac{\partial^2 \ln(1-\rho_0\tau_i)}{\partial\rho\partial\rho}\right| + \mathbb{E}\left|\frac{\partial^2 \ln f(\varepsilon_s(\boldsymbol{\theta}_0))}{\partial\rho\partial\rho}\right|$. Because $\sum_{i=1}^n \frac{\partial^2 \ln(1-\rho_0\tau_i)}{\partial\rho\partial\rho} = tr(M_n^2)$ (defined in (1.14)), this can be further simplified to

$$\frac{1}{n}tr(M_n^2) + \mathbb{E}\left|\left(\frac{f'^2(\varepsilon_s)}{f^2(\varepsilon_s)} - \frac{f''(\varepsilon_s)}{f(\varepsilon_s)}\right)\left(\sum_{k=1}^n w_{sk}y_k\right)^2\right| \tag{1.26}$$

Because M_n is uniformly bounded in column and row sums, $\frac{1}{n}tr(M_n^2) < \infty$, $\sum_{k=1}^n m_{sk} < b$ so $\sum_{j=1}^n \sum_{k=1}^n m_{sj}m_{sk} < (\sum_{k=1}^n m_{sk})^2 < b^2$.

We need to show $\mathbb{E}\left|\left(\frac{f'^2(\varepsilon_s)}{f^2(\varepsilon_s)} - \frac{f''(\varepsilon_s)}{f(\varepsilon_s)}\right)(\sum_{k=1}^n w_{sk}y_k)^2\right| < \infty$.

Because $Y_n = (I_n - \rho_0 W_n)^{-1}(\boldsymbol{g}(X_n, \boldsymbol{\theta}_0) + \boldsymbol{\varepsilon}_n)$, $W_n Y_n = M_n(\boldsymbol{g}(X_n, \boldsymbol{\theta}_0) + \boldsymbol{\varepsilon}_n)$, $\sum_{k=1}^n w_{sk} y_k = \sum_{k=1}^n m_{sk}(g(x_k, \boldsymbol{\theta}_0) + \varepsilon_k)$. It follows that $\mathbb{E}\left|\left(\frac{f'^2(\varepsilon_s)}{f^2(\varepsilon_s)} - \frac{f''(\varepsilon_s)}{f(\varepsilon_s)}\right)\left(\sum_{k=1}^n w_{sk} y_k\right)^2\right| =$

$$
\begin{aligned}
&\mathbb{E}\left|\left(\frac{f'^2(\varepsilon_s)}{f^2(\varepsilon_s)} - \frac{f''(\varepsilon_s)}{f(\varepsilon_s)}\right)\left(\sum_{k=1}^n m_{sk}(g(x_k, \boldsymbol{\theta}_0) + \varepsilon_k)\right)^2\right| \\
<&\mathbb{E}\left|\left(\frac{f'^2(\varepsilon_s)}{f^2(\varepsilon_s)} - \frac{f''(\varepsilon_s)}{f(\varepsilon_s)}\right)\sum_{k=1}^n m_{sk}^2[g(x_k, \boldsymbol{\theta}_0) + \varepsilon_k]^2\right| \\
&+ \mathbb{E}\left|\left(\frac{f'^2(\varepsilon_s)}{f^2(\varepsilon_s)} - \frac{f''(\varepsilon_s)}{f(\varepsilon_s)}\right)\sum_{j=1, j\neq k}^n \sum_{k=1}^n m_{sk} m_{sj}[g(x_k, \boldsymbol{\theta}_0) + \varepsilon_k][g(x_j, \boldsymbol{\theta}_0) + \varepsilon_j]\right|
\end{aligned}
$$

By assumption, $\mathbb{E}\,\varepsilon_k \varepsilon_j = 0$ if $k \neq j$, $\mathbb{E}\left|\frac{\varepsilon_s f'^2(\varepsilon_s)}{f^2(\varepsilon_s)} - \frac{\varepsilon_s f''(\varepsilon_s)}{f(\varepsilon_s)}\right| < \infty$, $\mathbb{E}\left|\frac{\varepsilon_s^2 f'^2(\varepsilon_s)}{f^2(\varepsilon_s)} - \frac{\varepsilon_s^2 f''(\varepsilon_s)}{f(\varepsilon_s)}\right| < \infty$. Through mathematical computation, we can prove that $\mathbb{E}\left|\frac{\partial^2 \ln f(\varepsilon_s(\boldsymbol{\theta}_0))}{\partial\rho\partial\rho}\right|$ is finite, i.e.,

$$
\mathbb{E}\left|\frac{\partial^2}{\partial\rho\partial\rho}\left(\frac{1}{n}\sum_{i=1}^n \ln(1 - \rho_0\tau_i) + \ln f(\varepsilon_s(\boldsymbol{\theta}_0))\right)\right| < \infty
$$

Because $\frac{1}{n}\sum_{i=1}^n \ln(1 - \rho_0\tau_i)$ in (1.25) only relates to ρ, this term goes away when taken second derivative with respect to other parameters. Hence, other elements in (1.25) equal

to those in $\mathbb{E}\left|\frac{\partial^2 \ln f(\varepsilon_s(\boldsymbol{\theta}_0))}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}'}\right|$ and we can show that those expectations are also finite.

$$\mathbb{E}\left|\frac{\partial^2 \ln f(\varepsilon_s(\boldsymbol{\theta}_0))}{\partial\rho\partial\beta'}\right| \leq |x_s'| \cdot \left(k_2|m_{ss}| + k_1|b - m_{ss}| \cdot \mathbb{E}|\varepsilon_s| + k_1\left|\sum_{k=1}^{n} m_{sk} g(x_k, \boldsymbol{\theta}_0)\right|\right) \tag{1.27}$$

$$\mathbb{E}\left|\frac{\partial^2 \ln f(\varepsilon_s(\boldsymbol{\theta}_0))}{\partial\rho\partial\lambda'}\right| \leq \mathbf{1}_h' \cdot \left(k_2|m_{ss}| + k_1|b - m_{ss}| \cdot \mathbb{E}|\varepsilon_s| + k_1\left|\sum_{k=1}^{n} m_{sk} g(x_k, \boldsymbol{\theta}_0)\right|\right) \tag{1.28}$$

$$\mathbb{E}\left|\frac{\partial^2 \ln f(\varepsilon_s(\boldsymbol{\theta}_0))}{\partial\rho\partial\gamma_i'}\right| \leq \frac{|\lambda_{i0} x_s'|}{4} \cdot \left(k_2|m_{ss}| + k_1|b - m_{ss}| \cdot \mathbb{E}|\varepsilon_s| + k_1\left|\sum_{k=1}^{n} m_{sk} g(x_k, \boldsymbol{\theta}_0)\right|\right) \tag{1.29}$$

$$\mathbb{E}\left|\frac{\partial^2 \ln f(\varepsilon_s(\boldsymbol{\theta}_0))}{\partial\beta\partial\beta'}\right| = k_1|x_s x_s'| \tag{1.30}$$

$$\mathbb{E}\left|\frac{\partial^2 \ln f(\varepsilon_s(\boldsymbol{\theta}_0))}{\partial\beta\partial\lambda'}\right| = k_1|x_s \boldsymbol{F}(x_s'\boldsymbol{\gamma}_0)| \tag{1.31}$$

$$\mathbb{E}\left|\frac{\partial^2 \ln f(\varepsilon_s(\boldsymbol{\theta}_0))}{\partial\beta\partial\boldsymbol{\gamma}_i'}\right| \leq \frac{k_1}{4}|\lambda_{i0} x_s x_s'| \tag{1.32}$$

$$\mathbb{E}\left|\frac{\partial^2 \ln f(\varepsilon_s(\boldsymbol{\theta}_0))}{\partial\lambda\partial\lambda'}\right| = k_1\left|\boldsymbol{F}(x_s'\boldsymbol{\gamma}_0)'\boldsymbol{F}(x_s'\boldsymbol{\gamma}_0)\right| \leq k_1 \cdot \mathbf{1}_{h\times h} \tag{1.33}$$

$$\mathbb{E}\left|\frac{\partial^2 \ln f(\varepsilon_s(\boldsymbol{\theta}_0))}{\partial\lambda\partial\boldsymbol{\gamma}_i'}\right| = \frac{k_1}{4}|\lambda_{i0} F'(x_s'\boldsymbol{\gamma}_{i0})| \cdot |\boldsymbol{F}(x_s'\boldsymbol{\gamma}_0)' x_s'| \leq \frac{k_1|\lambda_{i0}|}{4} \cdot |\boldsymbol{F}(x_s'\boldsymbol{\gamma}_0)' x_s'| \tag{1.34}$$

$$\mathbb{E}\left|\frac{\partial^2 \ln f(\varepsilon_s(\boldsymbol{\theta}_0))}{\partial\boldsymbol{\gamma}_i\partial\boldsymbol{\gamma}_j'}\right| \leq \frac{k_1|\lambda_{i0}\lambda_{j0}|}{16} \cdot |x_s x_s'|, \quad i \neq j \tag{1.35}$$

$$\mathbb{E}\left|\frac{\partial^2 \ln f(\varepsilon_s(\boldsymbol{\theta}_0))}{\partial\boldsymbol{\gamma}_i\partial\boldsymbol{\gamma}_i'}\right| \leq \frac{k_1\lambda_{i0}^2}{16} \cdot |x_s x_s'| + \frac{\sqrt{3}k_0|\lambda_{i0}|}{18}|x_s x_s'| \tag{1.36}$$

With assumptions 1-10, (1.26)-(1.36) are finite. Then we can apply the ergodic theorem [4] and conclude that

$$\left|\frac{1}{n}\frac{\partial^2 \mathcal{L}_n(\boldsymbol{\theta}_0)}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}'} - \mathbb{E}\frac{1}{n}\frac{\partial^2 \mathcal{L}_n(\boldsymbol{\theta}_0)}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}'}\right| \xrightarrow{p} 0$$

We have proved that $\left|\frac{1}{n}\frac{\partial^2 \mathcal{L}_n(\tilde{\boldsymbol{\theta}}_n)}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}'} - \frac{1}{n}\frac{\partial^2 \mathcal{L}_n(\boldsymbol{\theta}_0)}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}'}\right| \xrightarrow{p} \mathbf{0}$ so it is trivial that

$$\left|\frac{1}{n}\frac{\partial^2 \mathcal{L}_n(\tilde{\boldsymbol{\theta}}_n)}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}'} - \mathbb{E}\frac{1}{n}\frac{\partial^2 \mathcal{L}_n(\boldsymbol{\theta}_0)}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}'}\right| \xrightarrow{p} 0 \tag{1.37}$$

Recall the equation (1.20), we have proved that $\frac{1}{\sqrt{n}}\frac{\partial \mathcal{L}_n(\boldsymbol{\theta}_0)}{\partial \boldsymbol{\theta}}$ has the limiting distribution $N(\mathbf{0}, B(\boldsymbol{\theta}_0))$. With (1.37), for $\tilde{\boldsymbol{\theta}}_n$ between $\hat{\boldsymbol{\theta}}_n$ and $\boldsymbol{\theta}_0$, $-\frac{1}{n}\frac{\partial^2 \mathcal{L}_n(\tilde{\boldsymbol{\theta}}_n)}{\partial \boldsymbol{\theta}\partial \boldsymbol{\theta}'} \xrightarrow{p} A(\boldsymbol{\theta}_0)$ so we can conclude that $\sqrt{n}(\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0) \xrightarrow{d} N(\mathbf{0}, \boldsymbol{\Omega}_0)$, where $\boldsymbol{\Omega}_0 = A^{-1}(\boldsymbol{\theta}_0)B(\boldsymbol{\theta}_0)A^{-1}(\boldsymbol{\theta}_0)$. □

1.5. Numerical Results

1.5.1. Simulation Study

In this section, we conduct simulation experiments to examine the estimators' behavior for finite samples. For estimation purposes it is often useful to reparametrize the logistic function $F(x_s'\boldsymbol{\gamma}_i)$ as

$$F\left(||\boldsymbol{\gamma}_i|| \cdot x_s' \frac{\boldsymbol{\gamma}_i}{||\boldsymbol{\gamma}_i||}\right) = \left(1 + e^{-||\boldsymbol{\gamma}_i|| \cdot x_s' \frac{\boldsymbol{\gamma}_i}{||\boldsymbol{\gamma}_i||}}\right)^{-1}, \ i = 1, \ldots, h \tag{1.38}$$

where $||\boldsymbol{\gamma}_i||, i = 1, \ldots, h$ is the L_2-norm of $\boldsymbol{\gamma}_i$. We use a univariate exogenous variable and let $X_n = (x_1, \ldots, x_n)'$. For illustration, we only include the nonlinear component of X_n. Usually we would like to normalize predictors before fitting a neural network model to avoid the computation overflow [32] so we add a centralizing constant γ_0 in this simulation. The model becomes

$$y_s = \rho \sum_{i=1}^{n} w_{si} y_i + \lambda F(\gamma_1(x_s - \gamma_0)) + \varepsilon_s \tag{1.39}$$

For identification reasons mentioned in Restriction 1 and 2, we impose $\gamma_1 > 0$.

We sample $n = 2500, 4900$ random errors respectively from three distributions (standard normal, rescaled t-distribution and Laplace distribution) with variance 1 and X is a univariate exogenous variable, values of which sampled from a normal distribution $N(0.5, 3^2)$. We set the true parameters to be $\rho_0 = 0.6$, $\lambda_0 = 5$, and weights in the neural net $\gamma_{00} = 0.5$, $\gamma_{10} = 1$. The log-likelihood function $\boldsymbol{\mathcal{L}}_n(\boldsymbol{\theta})$ is given in (1.40) and we use L-BFGS-B method[12, 52] (recommended for bound constrained optimization) to find the

parameter estimates $\hat{\boldsymbol{\theta}}$ which maximize (1.40).

$$\mathcal{L}_n(\boldsymbol{\theta}) = \ln|I_n - \rho W_n| + \sum_{s=1}^{n} \ln f(\varepsilon_s(\boldsymbol{\theta})) \tag{1.40}$$

$$\varepsilon_s(\boldsymbol{\theta}) = y_s - \rho \sum_{s=1}^{n} w_{si} y_i - x_s\beta - \lambda F(\gamma_1(x_s - \gamma_0)) \tag{1.41}$$

For the model under consideration, we estimated the covariance of the asymptotic normal distribution equation (1.19). Since matrices $A(\boldsymbol{\theta}_0)$ and $B(\boldsymbol{\theta}_0)$ involve expected values with respect to the true parameter $\boldsymbol{\theta}_0$, given merely observations, in practice they can be estimated as follows:

$$\hat{A}(\hat{\boldsymbol{\theta}}) = \frac{1}{n}\sum_{s=1}^{n} -\frac{\partial^2 l(\hat{\boldsymbol{\theta}}|x_s, y_s)}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}'}$$

$$\hat{B}(\hat{\boldsymbol{\theta}}) = \frac{1}{n}\sum_{s=1}^{n} \frac{\partial l(\hat{\boldsymbol{\theta}}|x_s, y_s)}{\partial\boldsymbol{\theta}} \frac{\partial l(\hat{\boldsymbol{\theta}}|x_s, y_s)}{\partial\boldsymbol{\theta}'}$$

where

$$l(\boldsymbol{\theta}|x_s, y_s) = \frac{1}{n}\ln|I_n - \rho W_n| + \ln f(\varepsilon_s(\boldsymbol{\theta}))$$

Using (1.22) and (1.23), we can calculate $\hat{A}(\boldsymbol{\theta}_0)$, $\hat{B}(\boldsymbol{\theta}_0)$ to assess the asymptotic properties of parameter estimates. Note that the derivative of the log-likelihood with respect to ρ cannot be calculated directly because it requires taking derivative with respect to a log-determinant of $I_n - \rho W_n$. For small sample sizes, we can compute the determinant directly and get the corresponding derivatives; but for large sample sizes, for example a dataset with 3000 observations, W_n is a 3000×3000 weight matrix which makes it impossible to calculate the derivative directly. Since W_n is a square matrix, we can apply the spectral decomposition such that W_n can be expressed in terms of its n eigenvalue-eigenvector pairs in (1.6). So we can apply the following approach to calculate the derivative of

$\ln|I_n - \rho W_n|$, which greatly reduce the burden of computations (Viton [46]).

$$\ln|I_n - \rho W_n| = \ln\left(\prod_{s=1}^{n}(1-\rho\tau_i)\right)$$

Further the derivatives of the log-likelihood function with respect to ρ is

$$\frac{\partial l_s(\boldsymbol{\theta}|x_s,y_s)}{\partial\rho} = \frac{1}{n}\sum_{i=1}^{n}\frac{-\tau_i}{(1-\rho\tau_i)} + \{y_s - \rho\sum_{i=1}^{n}w_{si}y_i - \lambda F(\gamma_1(x_s-\gamma_0)\}\cdot\left(\sum_{i=1}^{n}w_{si}y_i\right)$$

$$\frac{\partial^2 l_s(\boldsymbol{\theta}|x_s,y_s)}{\partial\rho\partial\rho} = -\frac{1}{n}\sum_{i=1}^{n}\left[\frac{\tau_i^2}{(1-\rho\tau_i)^2} + \left(\sum_{i=1}^{n}w_{si}y_i\right)^2\right]$$

Finally we can estimate the covariance matrix by equation (1.42).

$$\hat{\boldsymbol{\Omega}} = \hat{A}^{-1}(\boldsymbol{\theta}_0)\hat{B}(\boldsymbol{\theta}_0)\hat{A}^{-1}(\boldsymbol{\theta}_0) \tag{1.42}$$

In our simulation study, we computed $\hat{\boldsymbol{\theta}}$ the for 200 replicates for each $n = 2500, 4900$. The estimate $\hat{\boldsymbol{\Omega}}$ of the asymptotic covariance matrix is computed based on a sample with 10000 simulated observations. Table 1.1 compares the empirical mean and standard errors (in parentheses) of parameter estimators with the true value and their asymptotic standard deviations (in squared brackets) respectively. Comparing the simulation results when $\boldsymbol{\varepsilon}$ follows a standard normal distribution with simulation results when $\boldsymbol{\varepsilon}$ follows a $t(4)$ distribution, means of the estimates over 200 replicates are closer to the true values and their empirical standard deviations are smaller when $\boldsymbol{\varepsilon}$ follows the heavy tailed distribution. For all these experiments with different error distributions, the empirical standard deviations of $\hat{\boldsymbol{\theta}}$ are close to the asymptotic standard deviations which implies that the estimators' finite sample behavior roughly matches their asymptotic distributions. Note that when $\boldsymbol{\varepsilon}$ is sampled from a Laplace distribution, this covariance matrix cannot be computed because its second order derivative is not differentiable at 0. But the simulated

$\hat{\boldsymbol{\theta}}$'s still appear consistent properties. Normal plots for parameter estimates are shown in Figure 1.4 and give a strong indication of normality.

ε	$n = 2500$			
	$\hat{\rho}$	$\hat{\lambda}$	$\hat{\gamma}_0$	$\hat{\gamma}_1$
$N(0,1)$	0.6178 (0.0075) [0.0046]	4.8504 (0.0812) [0.0639]	0.5410 (0.0425) [0.0417]	1.0576 (0.0431) [0.0354]
$t(4)$	0.6132 (0.0060) [0.0044]	4.8952 (0.0623) [0.0562]	0.5326 (0.0364) [0.0353]	1.0411 (0.0320) [0.0310]
Laplace $(0, \frac{\sqrt{2}}{2})$	0.6107 (0.0053)	4.9132 (0.0562)	0.5283 (0.0295)	1.0358 (0.0291)
ε	$n = 4900$			
	$\hat{\rho}$	$\hat{\lambda}$	$\hat{\gamma}_0$	$\hat{\gamma}_1$
$N(0,1)$	0.6175 (0.0056) [0.0033]	4.8617 (0.0572) [0.0456]	0.5435 (0.0297) [0.0298]	1.0517 (0.0303) [0.0252]
$t(4)$	0.6130 (0.0051) [0.0031]	4.8957 (0.0526) [0.0426]	0.5312 (0.0274) [0.0260]	1.0380 (0.0246) [0.0235]
Laplace $(0, \frac{\sqrt{2}}{2})$	0.6096 (0.0047)	4.9242 (0.0487)	0.5217 (0.0239)	1.0268 (0.0233)

Table 1.1. Empirical mean and standard errors (in parentheses) of parameter estimates when ε is sampled from a standard normal, standardized student t distribution and a Laplace distribution. The asymptotic standard errors are displayed for reference in square brackets.

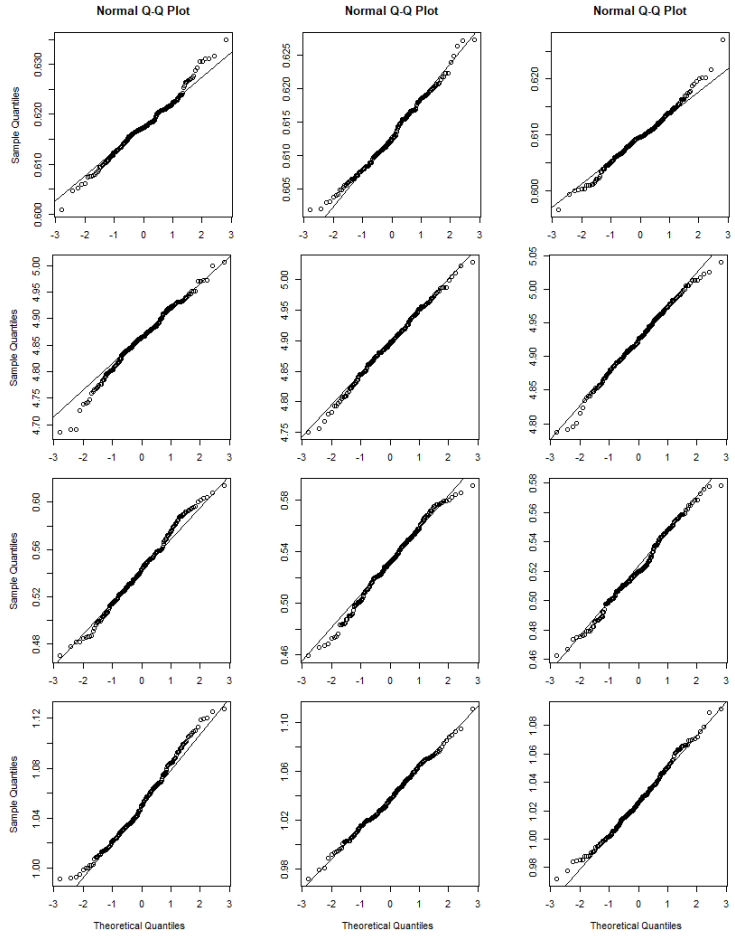


Figure 1.4. Normal plots for parameter estimates ρ (1st row), λ (2nd row), γ_0 (3rd row) and γ_1 (4th row) when ε_s follows a standard normal distribution (first column), standardized t distribution (middle column) and Laplace distribution (last column) $n = 70 \times 70$

1.5.2. Real Data Example

Spatial models have a lot of applications in understanding spatial interactions in cross-sectional data. Among them, the study of electoral behavior has attracted considerable attention by political scientists. Poole and Rosenthal [37] found that the spatial variation plays an important role in presidential electoral dynamics. And mentioned by Braha

and de Aguiar (2017 [7]), most studies in the U.S. consider vote choices as the result of attitudinal factors such as evaluations of the candidates and government performances as well as social factors such as race, social class, and region. Inspired by their research, we would like to understand this electoral dynamics using our proposed partially specified spatial autoregressive model and to help identify how social factors influence people's voting preferences.

Here, we focus on the proportion of votes cast for U.S. presidential candidates at the county level in 2004. Counties are grouped by state, and let Y be the corresponding fraction of votes (vote-share) in a county for the Democratic candidate (John Kerry) in 2004. Predictors X are chosen from economic and social factors covering the living standard, economy development and racial distribution. Figure 1.5 shows the observed

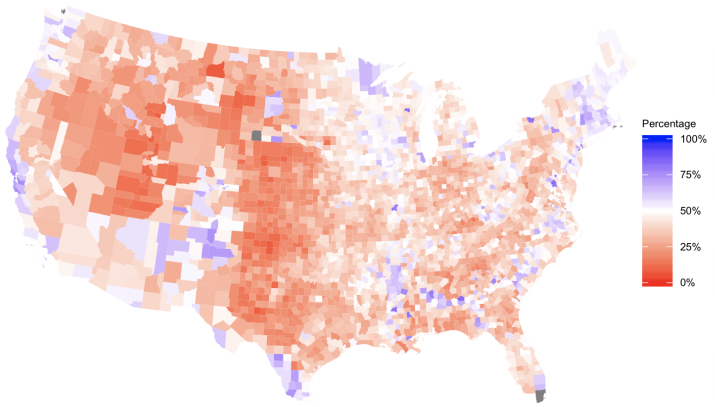


Figure 1.5. Fractions of Vote-shares per County for Democratic presidential candidate in 2004

values of Y_n for 2004. This heat map exhibits strong correlation between observations in neighboring counties which is supported by Moran's Test on Y (test statistic = 52.4,

P-value $< 2.2 \times 10^{-16}$). This indicates that Y, the fraction of vote-share for Democratic candidate, is not independently distributed across the space. So we consider fitting a spatial model to the data.

In our analysis, we exclude the four U.S. counties with no neighbors (San Juan, Dukes, Nantucket, Richmond) to avoid the non-singularity of our spatial weight matrix W_n in the modeling, so the total number of observations is $n = 3107$.

First we fit a linear regression model to see if it is sufficient to explain the voting dynamic using explanatory variables $X = (X_1, \ldots, X_5)$. From the preliminary analysis fitting Y on all the available variables, we chose the five most significant ones for modeling out of more than 20 different variables. The chosen predictors are percent residents under 18 years X_1 (`UNDER18`), percent white residents X_2 (`WHITE`), percent residents below poverty line X_3 (`pctpoor`), per capita income X_4 (`pcincome`) and USDA urban/rural code X_5 (`urbrural`, `0 = most rural`, `9 = most urban`). The corresponding least-squares line is as follows:

$$\hat{Y} = 80.4 - 0.932X_1 - 0.250X_2 + 0.324X_3 + 2.76 \times 10^{-5}X_4 - 1.24X_5 \tag{1.43}$$

These six parameter estimates are all significant at $\alpha = 0.05$ and by looking at signs, it is easy to tell how these covariates relate to the voting behaviors. However, one major drawback of this linear model is that the fitted residuals are still correlated across the space (null hypothesis of independence rejected in Moran's Test, test statistic = 54.1, P-value $< 2.2 \times 10^{-16}$; see Figure 1.6) so a multiple linear regression fails to adequately describe the spatial dependence in Y. Another concern is that a Gaussian estimation procedure was used; it is not most efficient when there appears to be heavy tailed errors. Figure 1.7 shows the histogram of $Y/8$ (men 4.87, standard deviation 1.57) which looks closer to a t-distribution than a normal distribution (scaled to have the same mean and

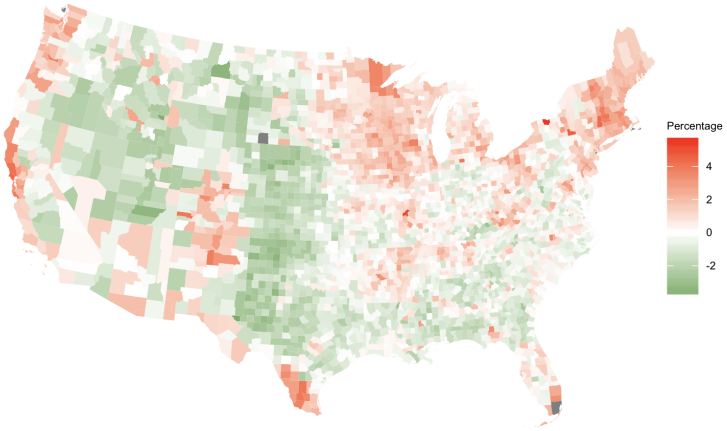


Figure 1.6. Residuals after fitting a linear regression model

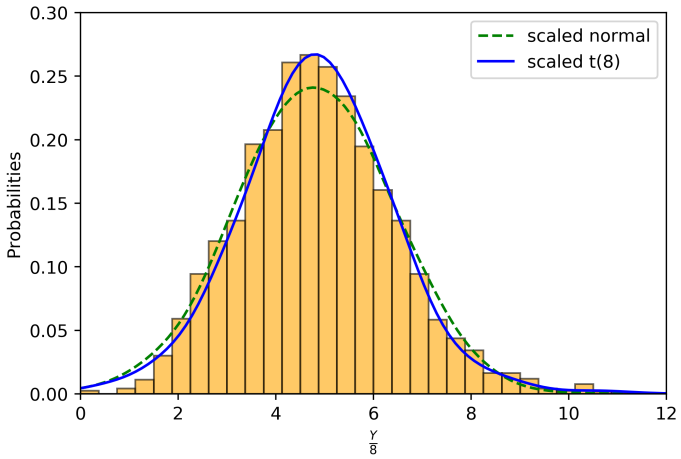


Figure 1.7. Histogram of (scaled by $\frac{1}{8}$) fraction of vote-shares per county for Democratic presidential candidate in 2004 overlaid with a scaled t and a normal density curves. The mean is 4.87 and the standard deviation is 1.57.

standard deviation as those of $Y/8$). Figure 1.8 also demonstrates the tail distribution of $Y/8$, where the vertical axis is the sample quantiles of $Y/8$ and horizontal axis is the theoretical quantiles of scaled $t(8)$ and normal distribution. Clearly the observation $Y/8$ is heavy tailed. To address these, we would like to fit a spatial autoregressive model to those

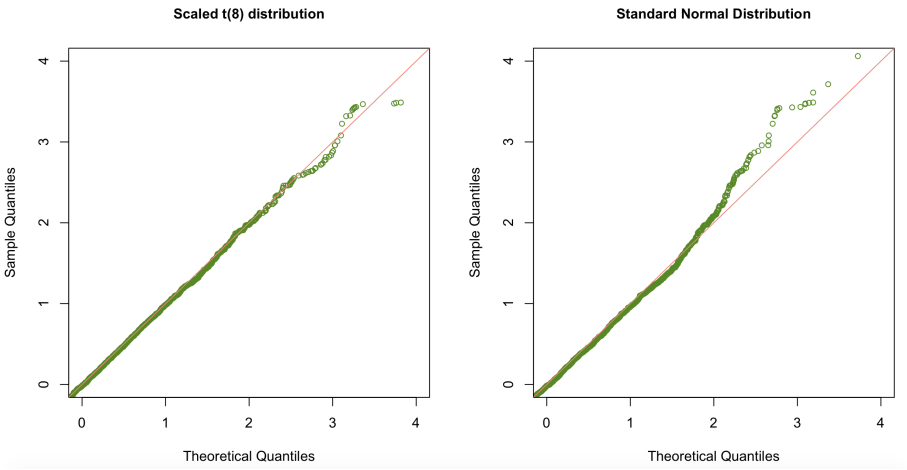


Figure 1.8. Q-Q plots of Y versus scaled $t(8)$ and standard normal distributions: Y-axis is the sample quantiles of $Y/8$ and X-axis is the theoretical quantiles of a t-distribution (left) and a normal distribution (right).

data and assume that the random error follows a scaled $t(8)$ distribution (scaled $t(8)$ has a closer density curve to $Y/8$ shown in Figure 1.7 and 1.8). We maximized the corresponding log-likelihood function to obtain parameter estimates. For simplicity, we then only selected the three most significant variables as predictors X based on the linear regression results; they are `UNDER18`, `WHITE` and `pctpoor`. The weight matrix W_n is generated through a shapefile [42] (a geospatial vector storage format for storing geometric location and associated attribute information) using the queen criterion. Scatter plots of X_1, X_2 and X_3 versus Y are shown in Figure 1.9. We can clearly observe the nonlinear trend between X_2, X_3 and Y. In the linear model X_1 is the most significant variable but despite the linear trend, the scatter plot of `UNDER18` versus Y has lots of noises around the center range from 20 to 30 percent. This may be caused by some spatial correlation in X_1 itself so we try despatializing X_1 by fitting an ordinary spatial autoregressive model $X_1 = \rho_x W_{3107} X_1 + \varepsilon$. The spatial correlation of X_1 is estimated as 0.6 so we define the despatialized variable $\tilde{X_1} = (I_n - 0.6W_{3107})X_1$ (the scatter plot of X_1 in Figure 1.9 does not show specific pattern

even though this variable is significant from our preliminary analysis. So we consider de-spatializing X_1 and $\sum_{s=1}^{n} \hat{\varepsilon}_s^2$ of the model fitted with despatializated X_1 is smaller than that of the model fitted with original X_1). In addition, to avoid the computation overflow when maximizing the corresponding log-likelihood function, we normalized these predictors to have zero means and unit variances and also rescaled Y by $\frac{1}{8}$. We conduct the following analysis using these transformed variables Y^*, $X^* = (X_1^*, X_2^*, X_3^*)$.

$$Y^* = Y/8$$

$$X_1^* = \frac{\tilde{X_1} - Average(\tilde{X_1})}{Std(\tilde{X_1})}$$

$$X_2^* = \frac{X_2 - Average(X_2)}{Std(X_2)}$$

$$X_3^* = \frac{X_3 - Average(X_3)}{Std(X_3)}$$

The first spatial model we tried is the ordinary spatial autoregressive model with X_1^*, X_2^*,

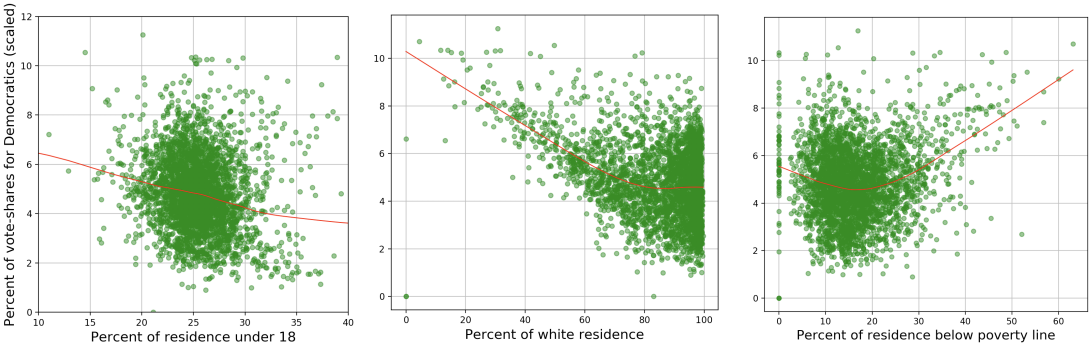


Figure 1.9. Scatter plots: percentage residence under 18 (left), percent white residents (middle) and percent below poverty line (right) in U.S. counties. The red lines are the Lowess smoothing curve between Y^* and predictors.

X_3^* and we assume the error follows a scaled t distribution with $df = 8$ ($Y^* = \rho W_{3107} Y^* + X^*\beta + \varepsilon$; the fitted residual variance is about 1 when we assume the t distribution with $df = 8$ referring to Figure 1.7). The model fit via maximizing log-likelihood function is

shown below:

$$Y^* = 0.744W_{3107}Y^* + 1.222 - 0.284X_1^* - 0.451X_2^* + 0.03X_3^* + \varepsilon \tag{1.44}$$

The spatial correlation parameter $\hat{\rho} = 0.744$ indicates pretty high spatial dependence in Y^* and the spatial dependence in the residuals is insignificant (Moran's test statistics = 1.38, P-value = 0.167). However, in Figure 1.9, there appears to be a nonlinear relationship between Y and X_2, X_3. To address this, we would like to fit our proposed PSAR-ANN model to the same dataset and still we assume a $t(8)$ distributed error. The log-likelihood function in this case should be

$$\mathcal{L}_{3107}(\boldsymbol{\theta}) = \ln|I_{3017} - \rho W_{3017}| - 4.5\sum_{s=1}^{3107}\ln(1 + \frac{\varepsilon_s(\boldsymbol{\theta})^2}{6}) \tag{1.45}$$

$$\boldsymbol{\varepsilon}(\boldsymbol{\theta}) = (I_{3017} - \rho W_{3017})Y^* - X^*\beta - F(X^*\boldsymbol{\gamma}')\lambda \tag{1.46}$$

Since the PSAR-ANN model has both linear and nonlinear components, we optimize these two parts iteratively to find the maximum likelihood estimators. Many optimization algorithm are sensitive to the choice of starting-values and people usually train neural network models starting at very small initial values. So especially each time instead of using the previous parameter estimates for neural network component, we always reinitialize the starting values for the neural network component and use L-BFGS-B algorithm [12] to search the optimum. The optimization steps are outlined below:

- Step 0: Based on some pre-knowledge about the parameters, set starting values $(\rho^0, \beta^0, \lambda^0, \boldsymbol{\gamma}^0)$ and predetermine bounds for parameters in the optimization.
- Step 1: In the ith iteration for the linear component optimization, fixing λ^{i-1}, $\boldsymbol{\gamma}^{i-1}$, use $(\rho^{i-1}, \beta^{i-1}, \lambda^{i-1}, \boldsymbol{\gamma}^{i-1})$ as starting values and apply L-BFGS-B algorithm [12, 52] to find $\rho^{(i)}$ and $\beta^{(i)}$ which maximize $\mathcal{L}_{3107}(\boldsymbol{\theta})$ in (1.45) given λ^{i-1}, $\boldsymbol{\gamma}^{i-1}$.

- Step 2: In the ith iteration for the nonlinear component optimization, fixing $\rho^{(i)}, \beta^{(i)}$ from Step 1, randomly initialize $\lambda, \boldsymbol{\gamma}$ starting values from a small interval $(0, 0.05)$ (to avoid the computation overflow when calculating exponentials) and again use L-BFGS-B algorithm [12] to find $\lambda^{(i)}, \boldsymbol{\gamma}^{(i)}$ which maximize $\mathcal{L}_{3107}(\boldsymbol{\theta})$ in (1.45) given $\rho^{(i)}, \beta^{(i)}$.
- Step 3: Repeat Step 1, 2 until the difference of the corresponding log-likelihood function values in Step 1 and 2 is smaller than some threshold value (for example 10^{-2}).

The following is the estimated PSAR-ANN model

$$\begin{aligned} Y^* =& 0.721 W_{3107} Y^* + 1.693 - 0.185 X_1^* - 0.658 X_2^* + 0.181 X_3^* \\ & - 0.937 F(1.509 X_1^* - 2.544 X_2^* + 2.268 X_3^*) + \hat{\varepsilon} \end{aligned} \tag{1.47}$$

In model (1.47), the correlation estimate is roughly the same as the model (1.44) indicating that people in neighboring counties tend to have similar voting preferences. The Moran's test statistic for the residuals is 1.78 with P-value = 0.0745. We compare the SAR model with our proposed model and find that even though the new model has more parameters, it has lower AIC (AIC = $2(\#\text{parameters}) - 2 \ln L_n(\hat{\boldsymbol{\theta}})$) compared to the original spatial autoregressive model (See table 1.2). Through likelihood ratio test ($\mathcal{H}_0$: SAR model is adequate, $\mathcal{H}_1$: PSAR-ANN model is adequate), the test statistic $-2 \ln L_{SAR} + 2 \ln L_{PSAR-ANN} = 157.45$ with $df = 4$, P-value < 0.05, so we rejected $\mathcal{H}_0$ and conclude that the PSAR-ANN model is a better fit than SAR model. Figure 1.10 shows the residuals (of PSAR-ANN model) heat map and its histogram. Through the residual histogram, assuming the error density as a standardized $t(8)$ (df is chosen by the

	SAR	PSAR-ANN
# Parameters	5	9
Moran's Test	0.167 (1.3808)	0.0745 (1.7836)
$-\ln L$	1958.08	1879.35
AIC	3926.16	3776.17

Table 1.2. Comparison of SAR and PSAR-ANN model by # parameters, Moran's test P-value (test statistics), $-\ln L$ and AIC

shape of the residual histogram) seems to be more appropriate than a standard normal distribution.

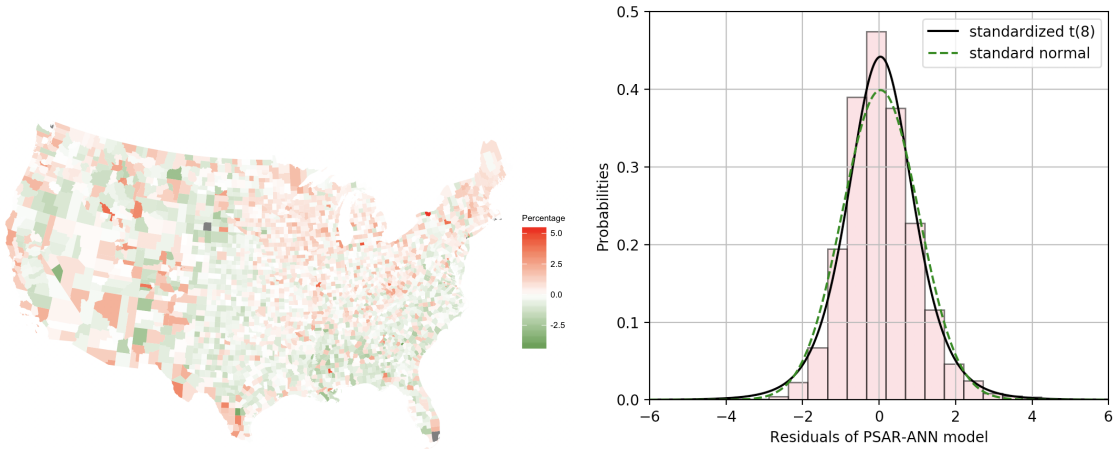


Figure 1.10. Heat map (left) and histogram (right) of residuals of the PSAR-ANN model

The covariance matrix for parameter estimates in model (1.47) is calculated and the 95% confidence intervals for the model parameters are shown in table 1.3. From the table, all the parameters are significant at 95% significance level. Looking at the signs of parameter estimates, we can learn that a county with more young residence under 18 and white residence is more likely to support Republicans while people struggling to

make ends meet are prone to support the Democrat. This opposite effect can also be observed in Figure 1.9. The neural network component in our model helps to capture the nonlinear relationship between X and Y. Parameter λ is significant so the nonlinear component is appreciable in modeling and $\boldsymbol{\gamma}$'s are all significant at 95% confidence level. Figure 1.11 shows scatter plots of X_1^*, X_2^*, X_3^*, where points are colored by the value of the fitted neural network component $-0.937F(1.509X_1^* - 2.544X_2^* + 2.268X_3^*)$. Observations with green color are counties tending to have more voters for the Democratic candidate while the red points represent counties tending to have more voters for the Republican candidate. From the distribution of these colored points, it appears that counties with more people below poverty line and less white residence tend to have more Democratic voters. On the other hand, voters in counties with more children and higher percent white residence tend to be less likely to vote Democratic. These findings also correspond to the trend we can find in the linear component but they are presented in a non-linear way.

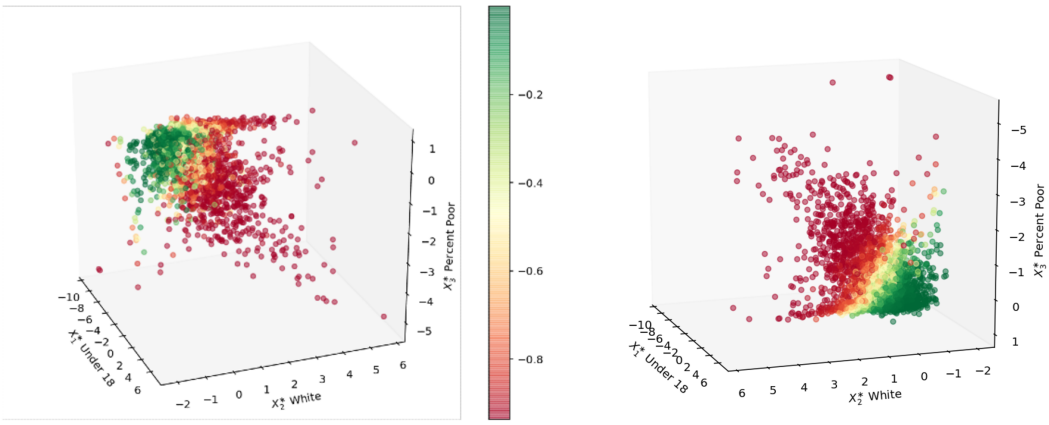


Figure 1.11. Scatter plot of X_1^*, X_2^*, X_3^* colored by the output of fitted neural network component $-0.937F(1.509X_1^* - 2.544X_2^* + 2.268X_3^*)$.

To conclude, our proposed model PSAR-ANN appears to successfully capture some spatial election dynamics. It allows for non-Gaussian random errors and is flexible in learning nonlinear relationships between the response and exogenous variables.

Parameter	Estimate	Std.	95% C.I.
ρ	0.721	0.0102	$(0.7010, 0.7410)$
β_0	1.693	0.0573	$(1.5807, 1.8053)$
β_1	-0.185	0.0219	$(-0.2279, -0.1421)$
β_2	-0.658	0.0288	$(-0.7144, -0.6016)$
β_3	0.181	0.0243	$(0.1334, 0.2286)$
λ	-0.937	0.0581	$(-1.0464, -0.8276)$
γ_1	1.509	0.0239	$(1.4622, 1.5558)$
γ_2	-2.544	0.0137	$(-2.5709, -2.5171)$
γ_3	2.268	0.0157	$(2.2372, 2.2988)$

Table 1.3. Parameter estimates of PSAR-ANN model with 95% confidence intervals

CHAPTER 2

Partially Specified Space Time Autoregressive Model with Artificial Neural Network

In Chapter 1, we proposed a PSAR model enhanced by a neural network component which aims at explaining the spatial dependence through a nonlinear approach. However, sometimes we may collect data across time as well as space. For this type of data, we want to construct a model with dependence over time taken into consideration which has a broad application especially in environmental sciences. One interesting application is forecasting the weather. For example, in a fixed location, the everyday temperature will change from time to time but in the meanwhile, it would also be affected by temperatures in the neighboring locations.

A class of such linear models known as space-time autoregressive (STAR) and space-time autoregressive moving average (STARMA) models was introduced by Cliff and Ord (1973) and Martin and Oeppen (1975) in 1970s. In general, STAR models contain a hierarchical ordering of "neighbors" of each site. For instance, on a regular grid, one can categorize neighbors of a site as first-order and second-order neighborhoods and so on. An observation at each site is then modeled as a linear function of the previous time observations at the same site and of the weighted previous observations at the neighboring sites of each order. Let $\{Y_t : t = 0, \pm 1, \pm 2, \ldots\}$ be a multivariate time series of n location components. Weights are incorporated in weight matrices $W^{(k)}$ for order k. An STAR

model with autoregressive order p and spatial order $(\lambda_1, \ldots, \lambda_p)$ considerded in Borovkova *et.al* (2008) is defined as

$$Y_t = \sum_{i=1}^{p} \sum_{k=0}^{\lambda_i} \phi_{ik} W^{(k)} Y_{t-i} + \varepsilon_t$$

where λ_i is the spatial order of the ith autoregressive term, ϕ_{ik} is the autoregressive parameter at time lag i and spatial lag k. Similarly an STAR model with n space locations and q exogenous variables is given by Stoffer (1985) as, for $Y_t \in \mathbb{R}^n$,

$$Y_t = \sum_{i=1}^{p} \sum_{k=0}^{\lambda_i} \phi_{ik} W^{(k)} Y_{t-i} + \sum_{i=0}^{p'} X_{t-i} \beta_i + \varepsilon_t$$

where values of the exogenous variables $\{X_t : t = 0, \pm 1, \pm 2, \ldots\}$ are $n \times q$ covariate matrices containing q values of exogenous variables for all n locations at time t. $X_t = (x_{1,t}, \ldots, x_{n,t})'$ and $x_{s,t} \in \mathbb{R}^q$. p' is the autoregressive order for $\{X_t\}$ and β_i is a $q \times 1$ model parameter.

STAR models have been widely applied in many areas of science. In genomics, Epperson (1993) analyzed population gene frequencies using STAR models where he assumed genes may vary over space and time. This model is also well known in economics (Giacomini and Granger, 2004) and has been applied to forecasting regional employment (Hernandez and Owyang, 2004) as well as traffic flow (Garrido 2000; Kamarianakis and Prastacos, 2004). For instance, the traffic flow of a road network observed at different fixed locations can be simultaneously modelled as a linear combination of past observations and current observations at neighboring sites. Through weight matrices, an STAR model assumes that near sites exert more influence on each other than distant ones.

In this chapter, we want to extend an STAR model to a semi-parametric model such that this new model can capture nonlinear dependence between covariates and the spatial observations of interest.

2.1. PSTAR-ANN(p) model

We define a Partially Specified Space-Time Autoregressive model with Artificial Neural Network (PSTAR-ANN(p)) as follows.

$$Y_t = \sum_{i=0}^{p} \phi_i W_n Y_{t-i} + X_t\beta + \boldsymbol{F}(X_t\boldsymbol{\gamma}')\lambda + \boldsymbol{\varepsilon}_t, \quad T = 1, \ldots, T \tag{2.1}$$

where $Y_t = \{y_{s,t}\}_{s=1}^{n}$ contains observations of dependent variables at n locations and at time t. The independent variable matrix $X_t = (x_{1,t}, \ldots, x_{n,t})'$ is the covariate matrix at time t, where $x_{s,t} \in \mathbb{R}^{q\times 1}$ is a vector containing exogenous regressors at location s and time t, $s = 1, \ldots, n$. $\boldsymbol{\varepsilon}_t = \{\varepsilon_{s,t}\}_{s=1}^{n}$ denote a vector of n noise terms which are independent identically distributed across s and t with density function f, mean 0 and variance $\sigma^2 = 1$.

Exogenous parameters $\beta = (\beta_1, \ldots, \beta_q)' \in \mathbb{R}^q$ and scalars $\phi_i, i = 0, 1, \ldots, p$, the spatial/space-time autoregressive parameters, are assumed to be the same over all regions. $W_n = \{w_{ij}\} \in \mathbb{R}^{n\times n}$ is a known spatial weight matrix which characterizes the connection between neighboring regions. For the ease of illustration, we define some notations. Given a function $f \in C^1(R^1)$ continuous in $\mathbb{R}$, we define a new matrix map $R^n \to R^n$ as $\boldsymbol{f}$ s.t. $\boldsymbol{f}(x_1, \ldots, x_n) = (f(x_1), \ldots, f(x_n))'$.

Using the notation defined above, the artificial neural network component (Medeiros *et al.* [32]) can be written as

$$\boldsymbol{F}(X_t\boldsymbol{\gamma}')\lambda = \begin{bmatrix} F(x'_{1,t}\boldsymbol{\gamma}_1) & F(x'_{1,t}\boldsymbol{\gamma}_2) & \dots & F(x'_{1,t}\boldsymbol{\gamma}_h) \\ F(x'_{2,t}\boldsymbol{\gamma}_1) & F(x'_{2,t}\boldsymbol{\gamma}_2) & \dots & F(x'_{2,t}\boldsymbol{\gamma}_h) \\ \vdots & \vdots & \cdots & \vdots \\ F(x'_{n,t}\boldsymbol{\gamma}_1) & F(x'_{n,t}\boldsymbol{\gamma}_2) & \dots & F(x'_{n,t}\boldsymbol{\gamma}_h) \end{bmatrix} \begin{bmatrix} \lambda_1 \\ \lambda_2 \\ \vdots \\ \lambda_h \end{bmatrix} \in \mathbb{R}^n$$

$\boldsymbol{F}(X_t\boldsymbol{\gamma}')\lambda$ represents two layer NN component where the first layer has h-neurons with the sigmoid activation function and the second layer has only one neuron with an identity activation function. In the first layer, the input is X_t and weights are $\boldsymbol{\gamma} = (\boldsymbol{\gamma}'_1, \dots, \boldsymbol{\gamma}'_h) \in \mathbb{R}^{h\times q}$ where $\boldsymbol{\gamma}_i = (\gamma_{i1}, \dots, \gamma_{iq})'$ is the weights in the ith neuron. $F(\cdot)$ is the sigmoid activation function in this layer.

$$F(x'_{s,t}\boldsymbol{\gamma}_i) = (1 + e^{-x'_{s,t}\boldsymbol{\gamma}_i})^{-1}, \quad s = 1, 2, \dots, n,\ i = 1, 2, \dots, h$$

In the second layer, the inputs are $F(x'_{s,t}\boldsymbol{\gamma}_i), i = 1, \dots, h$ and the weights are $\lambda_1, \dots, \lambda_h$. So final output is $\sum_{i=1}^{h} \lambda_i F(x'_{s,t}\boldsymbol{\gamma}_i)$ for each $x_{s,t}$.

The weight matrix W_n is a measure of distance between the spatial units, and in our application, we begin by using a square symmetric matrix with (i, j) element equals to 1 if regions i and j are neighbors and 0 otherwise. The diagonal elements of the matrix are set to zero. Then we row standardize this matrix denoted by W_n. For more details on construction of the weight matrix, you can refer to the previous chapter or LeSage [27]. The following plot provides a preview of the data we are working with. This data is generated from a PSTAR-ANN(2) model in a 10 by 10 lattice. The model equation is shown below:

(2.2)

$$Y_t = 0.6W_nY_t - 0.274W_nY_{t-1} + X_t \begin{pmatrix} 0.24 \\ \\ -0.7 \end{pmatrix} + 1.5\boldsymbol{F}(X_t\boldsymbol{\gamma}') + \boldsymbol{\varepsilon}_t, \quad X_t = \begin{pmatrix} x_{11,t} & \dots & x_{1n,t} \\ \\ x_{21,t} & \dots & x_{2n,t} \end{pmatrix}'$$

$\boldsymbol{\gamma} = (0.75, -0.35)$, with $\{x_{1i,t}\}_{i=1}^n, \{x_{2i,t}\}_{i=1}^n$ are generated i.i.d from $N(0, 1.5^2)$, $N(0, 3^2)$ and the error $\boldsymbol{\varepsilon}_t$ is from $N(0, 1)$. Figure 2.1 shows the heatmaps of Y_t simulated at $t = 30, 29, 28$ using (2.2).

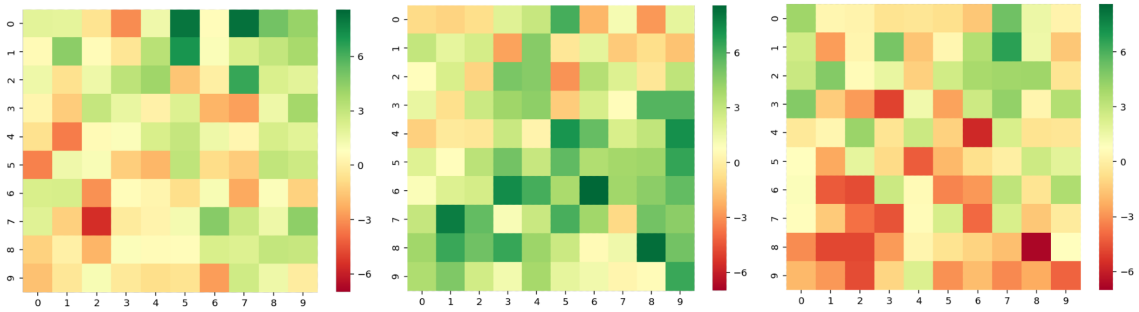


Figure 2.1. Heat map of Y_{30}, Y_{29} and Y_{28} simulated from a PSTAR-ANN(2) model

The color scale represents the value in each cell. We can observe colors in cells changing gradually with the spatial and time dependence ($\phi_1 = -0.274$, there is a little flip in cell color comparing the left figure with the middle one).

2.2. The Model and the Likelihood Function

2.2.1. The Model

Let

$$A_0 = I_n - \phi_0 W_n, \quad A_i = \phi_i W_n \quad i = 1, \dots, p$$

Suppose A_0 is invertible, then model (2.1) can be rewritten as:

$$A_0 Y_t = \sum_{i=1}^{p} A_i Y_{t-i} + X_t\beta + \boldsymbol{F}(X_t\boldsymbol{\gamma}')\lambda + \boldsymbol{\varepsilon}_t$$

$$Y_t = A_0^{-1}\sum_{i=1}^{p} A_i Y_{t-i} + A_0^{-1}X_t\beta + A_0^{-1}\boldsymbol{F}(X_t\boldsymbol{\gamma}')\lambda + A_0^{-1}\boldsymbol{\varepsilon}_t$$

Let L be the usual backshift operator such that $L^i Y_t = Y_{t-i}$, $A(L) = A_0 - \sum_{i=1}^{p} A_i L^i$. Assuming that $A^{-1}(L)$ exists, we can rewrite Y_t as

$$Y_t = A(L)^{-1}(X_t\beta + \boldsymbol{F}(X_t\boldsymbol{\gamma}')\lambda + \boldsymbol{\varepsilon}_t) \tag{2.3}$$

In order to derive asymptotic properties, we also need Y_t to be a causal spatial temporal process. Referring to the definition in Brockwell and Davis [9], the process Y_t is causal if there exists matrices $\{\Psi_j\}$ with absolutely summable components such that $A^{-1}(L) = \sum_{j=0}^{\infty}\Psi_j A_0^{-1}L^j$. Let $A(z) = A_0 - A_1 z - A_2 z^2 - \cdots - A_p z^p = A_0(I_n - A_0^{-1}A_1 z - A_0^{-1}A_2 z^2 - \cdots - A_0^{-1}A_p z^p)$ be a matrix-valued polynomial. Causality is equivalent to the condition $det(A(z)) \neq 0$ for all $z \in \mathbb{C}$ such that $|z| \leq 1$.

The matrices Ψ_j can be found recursively from the equations

$$\Psi_j = \Theta_j + \sum_{k=1}^{\infty} A_0^{-1}A_k\Psi_{j-k} \tag{2.4}$$

where we define $\Theta_0 = I_n$, $\Theta_j = 0_n$ for $j > 0$, $A_j = 0_n$ for $j > p$ and $\Psi_j = 0_n$ for $j < 0$. Therefore, this gives us

$$\Psi_0 = I_n$$

$$\Psi_1 = A_0^{-1}A_1$$

$$\Psi_2 = (A_0^{-1}A_1)^2 + A_0^{-1}A_2$$

$$\cdots$$

Then

$$Y_t = \sum_{j=0}^{\infty} \Psi_j A_0^{-1}(X_{t-j}\beta + \boldsymbol{F}(X_{t-j}\boldsymbol{\gamma}') + \boldsymbol{\varepsilon}_{t-j}) \tag{2.5}$$

With this expansion, we need few assumptions on $\sum_{j=0}^{\infty} \Psi_j A_0^{-1}$ and will be discussed later.

2.2.2. Likelihood Function

Denote $\boldsymbol{\theta} = (\phi_0, \phi_1, \ldots, \phi_p, \beta_1, \ldots, \beta_q, \lambda, \boldsymbol{\gamma}_1', \ldots, \boldsymbol{\gamma}_h')' \in \boldsymbol{\Theta}$. Since $\varepsilon_{s,t}$ has an identical density function f, the conditional joint density of $Y_T, Y_{T-1}, \ldots, Y_1$ conditioned on a finite number of past values $\{Y_0, \ldots, Y_{1-p}\}$ and $\{X_t\}_{t=1}^T$ is

$$f_{Y_T, Y_{T-1}, \ldots, Y_1}(\boldsymbol{\theta}|Y_0, \ldots, Y_{1-p}, \{X_t\}) = \prod_{t=1}^{T} f_{Y_t}(\boldsymbol{\theta}|Y_{t-1}, \ldots, Y_{1-p}, \{X_t\})$$

Since

$$f_{Y_t}(\boldsymbol{\theta}|Y_{t-1}, \ldots, Y_{1-p}, \{X_t\}) = |A_0| \prod_{s=1}^{n} f(\varepsilon_{s,t}(\boldsymbol{\theta}))$$

we have

$$f_{Y_T, Y_{T-1}, \ldots, Y_1}(\boldsymbol{\theta}|Y_0, \ldots, Y_{1-p}, \{X_t\}) = |A_0|^T \prod_{t=1}^{T} \prod_{s=1}^{n} f(\varepsilon_{s,t}(\boldsymbol{\theta}))$$

Hence, the log-likelihood function of $\boldsymbol{\theta}$ is given by [3, p. 63],

$$\mathcal{L}_{n,T}(\boldsymbol{\theta}) = T \ln |A_0| + \sum_{t=1}^{T} \sum_{s=1}^{n} \ln f(\varepsilon_{s,t}(\boldsymbol{\theta})) \tag{2.6}$$

where $\boldsymbol{\varepsilon}_t(\boldsymbol{\theta}) = \{\varepsilon_{s,t}(\boldsymbol{\theta})\}_{s=1}^n = A(L)Y_t - X_t\beta - \boldsymbol{F}(X_t\boldsymbol{\gamma})\lambda$ for $t = 1, \ldots, T$.

For the analysis of identification and estimation of the PSTAR-ANN(p) model, we adopt the following assumptions:

Assumption 13. The $p + (q+1)(h+1)$ parameter vector

$\boldsymbol{\theta} = (\phi_0, \phi_1, \ldots, \phi_p, \beta', \lambda', \boldsymbol{\gamma}_1', \ldots, \boldsymbol{\gamma}_h')' \in \boldsymbol{\Theta}$, where $\boldsymbol{\Theta}$ is a subset of the $p+(q+1)(h+1)$ dimensional Euclidean space, $\mathbb{R}^{p+(q+1)(h+1)}$. $\boldsymbol{\Theta}$ is a closed and bounded compact set and contains the true parameter value $\boldsymbol{\theta}_0$ as an interior point.

Assumption 14. The spatial correlation coefficient ϕ_0 satisfies $|\phi_0| < 1$ and $\phi_0 \in (-1/\tau, 1/\tau)$, where $\tau = max\{|\tau_1|, |\tau_2|, \ldots, |\tau_n|\}$, $\tau_1, \ldots, \tau_n$ are eigenvalues of spatial weight matrix W_n. To avoid the non-stationarity issue when ϕ_0 approaches to 1, we assume $\sup_{\phi_0 \in \boldsymbol{\Theta}} |\phi_0| < 1$.

Assumption 15. We assume W_n is defined by queen contiguity and is uniformly bounded in row and column sums in absolute value as $n \to \infty$ so A_0^{-1} is also uniformly bounded in both column and row sums as $n \to \infty$.

Assumption 16. We assume a causal spatial process Y_t which means that every z which solves

$$\det\left[z^p A_0 - \sum_{i=1}^{p} \phi_i W_n z^{p-i}\right] = 0$$

lie inside a unit circle. So the operator $A(L)$ is causal [36].

Assumption 17. X_t is stationary, ergodic satisfying $\mathbb{E}\,|x_{s,t}|^2 < \infty$ and X_t is full column rank for $t = 1, 2 \ldots, T$.

Assumption 18. The error terms $\varepsilon_{s,t}$, $s = 1, 2, \ldots, n$, $t = 1, 2 \ldots, T$ are independent and identically distributed with density function $f(\cdot)$, zero mean and unit variance $\sigma^2 = 1$. The moment $\mathbb{E}(|\varepsilon_{s,t}|^{2+r})$ exists for some $r > 0$ and $\mathbb{E}|\ln f(\varepsilon_{s,t})| < \infty$.

Assumption 14 defines the parameter space for ϕ_0 such that A_0 is strictly diagonally dominant. By the Levy-Desplanques theorem [45], it follows that A_0^{-1} exists for any values

ϕ_0 in $(-1/\tau, 1/\tau)$. In real applications, since W_n is row standardized, one just searches $\hat{\phi}_0$ over a parameter space on $(-1, 1)$ to find the optimizer [17, p. 749-754].

It is natural to consider the neighborhood by connections and in many practical studies, since entries scaled to sum up to 1, each row of W_n sums up to 1, which guarantees that all nonzero weights are in $(0, 1]$. For simplicity, we define the weight matrix W_n using the queen criterion and do row standardization. Assumption 15 is originated by Kelejian and Prucha [22, 23] and is also used in Lee [25]. With W_n to be uniformly bounded, we can prove that $(I_n - \phi_0 W_n)^{-1}$ is also uniformly bounded in row and column sums for $\phi_0 \in (-1/\tau, 1/\tau)$ and $\sup_{\phi_0 \in \mathbf{\Theta}} |\phi_0| < 1$, by Lemma A.4 in Lee[25]. This result is a necessary condition for Assumption 16.

From Assumption 14 and 15, we can decompose W_n by its eigenvalue and eigenvector pairs τ_i, v_i: $W_n = P\Lambda P^{-1}$, where Λ is a diagonal matrix with eigenvalues τ_i on its diagonals and $P = [v_1, v_2, \ldots, v_n]$ (we assume v_i's are normalized eigenvectors). So

$$W = P \begin{pmatrix} \tau_1 & 0 & \cdots & 0 \\ 0 & \tau_2 & \cdots & 0 \\ 0 & 0 & \ddots & 0 \\ 0 & 0 & \cdots & \tau_n \end{pmatrix} P^{-1}, A_0^{-1} = P \begin{pmatrix} \frac{1}{1-\phi_0\tau_1} & 0 & \cdots & 0 \\ 0 & \frac{1}{1-\phi_0\tau_2} & \cdots & 0 \\ 0 & 0 & \ddots & 0 \\ 0 & 0 & \cdots & \frac{1}{1-\phi_0\tau_n} \end{pmatrix} P^{-1} \tag{2.7}$$

It is trivial that $A_0^{-1} W_n = W_n A_0^{-1}$.

Assumption 16 guarantees that $A(L)$ is a causal operator and there exists a casual solution $\{Y_t\}$ to the system of the model equation (2.1). Then $\sum_{j=0}^{\infty} \Psi_j A_0^{-1}$ is absolutely summable. This requirement serves to determine a region of possible ϕ_i values that will result in a stationary process $\{Y_t\}$.

Assumption 17 is a trivial one when exogenous variables are included in a space time model. Similar to previous chapter, the stationarity of $\{x_{s,t}\}$ is necessary in the ergodic theorem in later proofs.

Assumption 18 imposes restrictions for the random error. In this paper we mainly consider the heavy tailed density functions such scaled t distributions and Laplace distributions. When the degrees of freedom goes to infinity, the scaled t distribution would approximate a standard normal distribution. So we would like to concentrate more on the scaled t distribution with lower degrees of freedom.

2.3. Model Identification

In the previous section, we have some restrictions on the weight matrices W_n and A_i's to guarantee the identification of a classical spatial time autoregressive model. We now investigate the conditions under which PSTAR(p)-ANN model is identified. By Rothenberg [38], a parameter $\theta_0 \in \mathbf{\Theta}$ is *globally identified* if there is no other θ in $\mathbf{\Theta}$ that observationally equivalent to θ_0 such that $f(y, \theta) = f(y, \theta_0)$; or the parameter θ_0 is *locally identified* if there is no such θ in an open neighborhood of θ_0 in $\mathbf{\Theta}$. The model (2.1), in principle, is neither globally nor locally identified due to the neural network component. The lack of identification of neural network models has been discussed in many papers (Hwang and Ding [20]; Medeiros *et al.* [32]). Here we extend the discussion to our proposed PSTAR(p)-ANN model. Three characteristics imply non-identification of our model: (a) the interchangeable property: the value of the likelihood function may remain unchanged if we permute the hidden units. For a model with h neurons, this will result in $h!$ different models that are indistinguishable from each other and have equal

local maximums of the log-likelihood function; (b) the "symmetry" property: for a logistic function, $F(x) = 1 - F(-x)$ allows two equivalent parametrization for each hidden unit; (c) the reducible property: the presence of irrelevant neurons in model (2.1) happens when $\lambda_i = 0$ for at least one i and parameters $\boldsymbol{\gamma}_i$ remain unidentified. Conversely, if $\boldsymbol{\gamma}_i = \mathbf{0}$, $F(X_t\boldsymbol{\gamma}_i)$ is a constant and λ_i can take any value without affecting the value of likelihood functions.

The problem of interchangeability (as mentioned in (a)) can be solved by imposing the following restriction, as in Medeiros *et al.* [32]:

Restriction 1. *parameters $\lambda_1, \ldots, \lambda_h$ are restricted such that: $\lambda_1 \geq \cdots \geq \lambda_h$.*

And to tackle (b) and (c), we can apply another restriction:

Restriction 2. *The parameters λ_i and γ_{i1} should satisfy:*

(1) $\lambda_i \neq 0$, $\forall i \in \{1, 2, \ldots, h\}$; and

(2) $\gamma_{i1} > 0$, $\forall i \in \{1, 2, \ldots, h\}$.

To guarantee the non-singularity of model matrices and the uniqueness of parameters, we impose the following basic assumption:

Assumption 19. The true parameter vector $\boldsymbol{\theta}_0$ satisfies Restrictions 1-2.

Referring to the section 4.3 by Medeiros *et al.* [32], we can conclude the identifiability of the PSAR-ANN model.

Lemma 4. Under the Assumptions 13-19, this PSTAR-ANN(p) model (2.1) is globally identified.

2.4. Asymptotic Results

Let the true parameter vector as $\boldsymbol{\theta}_0$ and the solution which maximizes the log-likelihood function (2.6) as $\hat{\boldsymbol{\theta}}_{n,T}$. Hence, $\hat{\boldsymbol{\theta}}_{n,T}$ should satisfy

$$\hat{\boldsymbol{\theta}}_{n,T} = \arg\max_{\boldsymbol{\theta}\in\boldsymbol{\Theta}} \mathcal{L}_{n,T}(\boldsymbol{\theta})$$

Suppose as n is large enough, T goes to infinity, $\hat{\boldsymbol{\theta}}_{n,T}$ is equivalent to maximizing the average of the likelihood function $\mathcal{L}_{n,T}(\boldsymbol{\theta})$ shown as follows:

$$\frac{1}{nT}\mathcal{L}_{n,T}(\boldsymbol{\theta}) = \frac{1}{n}\ln|A_0| + \frac{1}{nT}\sum_{s=1}^{n}\sum_{t=1}^{T}\ln f(\varepsilon_{s,t}(\boldsymbol{\theta}))$$

$$\hat{\boldsymbol{\theta}}_{n,T} = \arg\max_{\boldsymbol{\theta}\in\boldsymbol{\Theta}} \left(\frac{1}{n}\ln|A_0| + \frac{1}{nT}\sum_{s=1}^{n}\sum_{t=1}^{T}\ln f(\varepsilon_{s,t}(\boldsymbol{\theta}))\right)$$

$$\varepsilon_{s,t}(\boldsymbol{\theta}) = y_{s,t} - \sum_{i=0}^{p}\sum_{k=1}^{n}\phi_i w_{sk} y_{k,t-i} - x'_{s,t}\beta - \sum_{i=1}^{h}\lambda_i F(x'_{s,t}\boldsymbol{\gamma}_i)$$

At specific time t, suppose we have a $n_1 \times n_2$ lattice where we consider asymptotic properties of $\hat{\boldsymbol{\theta}}_{n,T}$ when $n = n_1 n_2 \to \infty$. Write the location s as the coordinate (s_x, s_y) in the $[1, n_1] \times [1, n_2]$ lattice space. The distance between two locations s, j is defined as $d(s, j) = \max(|s_x - j_x|, |s_y - j_y|)$. So if observations at s, j locations are neighbors (by queen criterion), their coordinates should satisfy $(s_x - j_x)^2 + (s_y - j_y)^2 \leq 2$ or $d(s, j) = 1$.

In a spatial context, we should notice that the functional form of $y_{s,t}$ is not identical for all the locations due to values of the weights $\{w_{si}\}_{i=1}^{n}$. For example, in a lattice, units at edges, vertexes or in the interior have different density functions due to different neighborhood structures (Figure 2.2). Denote $\mathcal{N}_s$ as a neighborhood set for location s. For an interior point (Figure 2.2(c)), its neighborhood set $\mathcal{N}_s$ contains eight neighbors where $w_{sj} = 1/8$ if $d(s, j) = 1$ otherwise $w_{sj} = 0$, for $j = 1, 2, \ldots, n$. Similarly, an edge

point (Figure 2.2(b)) has five neighboring units with $w_{sj} = 1/5$ for $j \in \mathcal{N}_s$ and the weight of a vertex neighborhood is 1/3 because a vertex unit has only three neighbors. This is known as an edge effect in spatial problems.

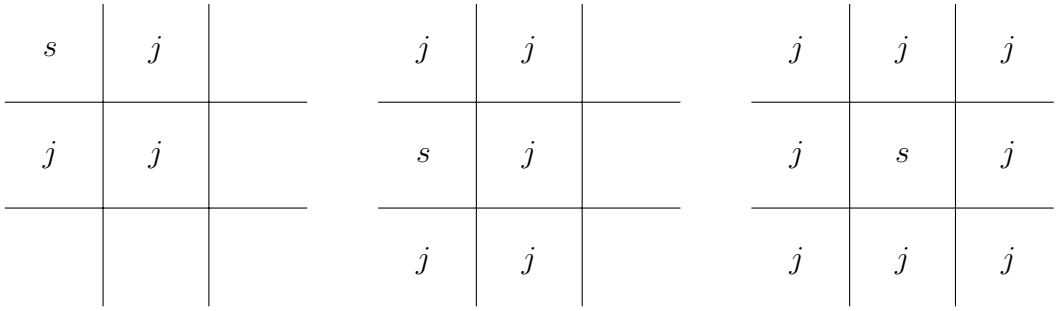


Figure 2.2. Vertex (a), Edge (b) and Interior Points (c) Neighborhood Structures: s is the target location and j represents the neighborhood of s

To deal with this, referring to Yao and Brockwell [50], we construct an edge effect correction scheme based on the way that the sample size tends to infinity. In a space $[1, n_1] \times [1, n_2]$, we consider its interior area as $\mathcal{S} = \{(s_x, s_y) : b_1 \leq s_x \leq n_1 - b_1, b_2 \leq s_y \leq n_2 - b_2\}$, where $b_1, b_2, n_1, n_2 \to \infty$ satisfying that $b_1/n_1, b_2/n_2 \to 0$ and other locations belong to the boundary areas $\mathcal{M}$. Therefore the set $\mathcal{S}$ contains $n^* = (n_1 - 2b_1)(n_2 - 2b_2)$ interior locations while the set $\mathcal{M}$ contains $n - n^*$ boundary locations. Then $n^*/n \to 1$ and $\mathcal{L}_{n,T}(\boldsymbol{\theta})$ can be split into a sum of two parts (interior $\mathcal{S}$ and boundary $\mathcal{M}$ parts):

$$\mathcal{L}_{n,T}(\boldsymbol{\theta}) = \sum_{t=1}^{T} \left(\sum_{s \in \mathcal{M}} l(\boldsymbol{\theta}|z_{s,t}) + \sum_{s \in \mathcal{S}} l(\boldsymbol{\theta}|z_{s,t}) \right)$$

$$l(\boldsymbol{\theta}|z_{s,t}) = \frac{1}{n} \ln |A_0| + \ln f(y_{s,t} - \sum_{i=0}^{p} \sum_{k=1}^{n} \phi_i w_{sk} y_{k,t-i} - x'_{s,t}\beta - \sum_{i=1}^{h} \lambda_i F(x'_{s,t}\boldsymbol{\gamma}_i))$$

where $Z_t = (W_n Y_t, W_n Y_{t-1}, \ldots, W_n Y_{t-p}, X_t)$ and $z_{s,t}$ is the s row of Z_t.

Therefore, given that $\lim_{n_1,n_2\to\infty} \frac{|\mathcal{M}|}{n} = 0$, $n^{-1}\sum_{s\in\mathcal{M}} l(\boldsymbol{\theta}|z_{s,t})$ vanishes a.s. as n tends to infinity for any $\boldsymbol{\theta} \in \boldsymbol{\Theta}$. Therefore,

$$\begin{aligned}\lim_{n,T\to\infty}(nT)^{-1}\mathcal{L}_{n,T}(\boldsymbol{\theta}) &= \lim_{T\to\infty}\frac{1}{T}\sum_{t=1}^{T}\lim_{n_1,n_2\to\infty}\frac{1}{n_1 n_2}\left(\sum_{s\in\mathcal{M}} l(\boldsymbol{\theta}|z_{s,t}) + \sum_{s\in\mathcal{S}} l(\boldsymbol{\theta}|z_{s,t})\right)\\ &= \lim_{T\to\infty}\frac{1}{T}\sum_{i=1}^{T}\lim_{n_1,n_2\to\infty}\frac{1}{n_1 n_2}\sum_{s\in\mathcal{S}} l(\boldsymbol{\theta}|z_{s,t}) \quad a.s.\end{aligned}$$

In this equation, every location $s \in \mathcal{S}$ has eight neighboring units under the queen criterion with nonzero weights $w_{sj} = 1/8$. Hence for an interior unit $s \in \mathcal{S}$, $\sum_{i=1}^{n} w_{si}y_i = \sum_{j=1}^{n} \frac{1}{8} y_j I_{\{d(s,j)=1\}}$. And the log likelihood function $\mathcal{L}_{n,T}(\boldsymbol{\theta})$ is approximately

$$(nT)^{-1}\mathcal{L}_{n,T}(\boldsymbol{\theta}) \approx \frac{1}{nT}\sum_{i=1}^{T}\sum_{s\in\mathcal{S}} l(\boldsymbol{\theta}|z_{s,t}) \quad \text{for } n_1, n_2, T \to \infty \tag{2.8}$$

So the maximum likelihood estimator $\hat{\boldsymbol{\theta}}_{n,T}$ approximately maximizes

$$\hat{\boldsymbol{\theta}}_{n,T} \approx \arg\max_{\boldsymbol{\theta}\in\boldsymbol{\Theta}} \lim_{\substack{T\to\infty \\ n_1,n_2\to\infty}} \frac{1}{nT}\sum_{i=1}^{T}\sum_{s\in\mathcal{S}} l(\boldsymbol{\theta}|z_{s,t})$$

2.4.1. Consistency Results

To establish the consistency of $\hat{\boldsymbol{\theta}}_{n,T}$, the heuristic insight is that because $\hat{\boldsymbol{\theta}}_{n,T}$ maximizes $\frac{1}{nT}\mathcal{L}_{n,T}(\boldsymbol{\theta})$, it approximately maximizes $\frac{1}{nT}\sum_{i=1}^{T}\sum_{s\in\mathcal{S}} l(\boldsymbol{\theta}|z_{s,t})$. By equation (2.8), $\frac{1}{nT}\mathcal{L}_{n,T}(\boldsymbol{\theta})$ can generally be shown tending to a real function $\mathcal{L} : \boldsymbol{\Theta} \to \mathbb{R}$ with maximizer $\boldsymbol{\theta}_0$ as $n, T \to \infty$ under mild conditions on the data generating process, then $\hat{\boldsymbol{\theta}}_{n,T}$ should tend to $\boldsymbol{\theta}_0$ almost surely. Before the formal proof of the consistency, we need the following assumptions on the density function $f(\cdot)$ satisfied (similar assumptions are made in White [48], Andrews, Davis and Breidt [2], Lii and Rosenblatt [29]).

Assumption 20. For all $s \in \mathbb{R}$, $f(s) > 0$ and $f(s)$ is twice continuously differentiable with respect to s.

Assumption 21. The density should satisfy the following equations:

- $\int s f'(s)\, ds = s f(s)|_{-\infty}^{\infty} - \int f(s)\, ds = -1$
- $\int f''(s)\, ds = f'(s)|_{-\infty}^{\infty} = 0$
- $\int s^2 f''(s)\, ds = s^2 f'(s)|_{-\infty}^{\infty} - 2\int s f'(s)\, ds = 2$

Assumption 22. The density should follow the following dominance conditions: $\left|\frac{f'(s)}{f(s)}\right|$, $\left|\frac{f'(s)}{f(s)}\right|^2$, $\left|\frac{f'(s)}{f(s)}\right|^4$, $\frac{f''(s)}{f(s)}$, and $\frac{f''(s)f'^2(s)}{f^3(s)}$ are dominated by $a_1 + a_2 |s|^{c_1}$, where a_1, a_2, c_1 are non-negative constants and $\int |s|^{c_1+2} f(s)\, ds < \infty$.

Assumption 23. If $c_1 > 2$ in previous assumption, we further assume $\mathbb{E}\,|x_{s,t}|^{c_1} < \infty$.

Discussed in Breidt, Davis, Lii and Rosenblatt [8] and Andrews, Davis and Breidt [2, p. 1642-1645], these assumptions on the density $f(\cdot)$ are satisfied by the t-distribution case when $\nu > 2$ and by a mixture of Gaussian distributions. The assumption $\mathbb{E}|\ln f(s)| < \infty$ (see Assumption 18) is also checked satisfied by the normal and t distributions ($\nu > 2$). The Laplace distribution does not strictly satisfy the Assumptions 20-22, since it is not differentiable at 0 but it satisfies these boundedness conditions almost everywhere so we believe the consistency and asymptotic normality results remain valid for parameter estimates. This will be shown in the simulation section. Assumption 23 is a necessary to boundedness conditions in later proof.

Lemma 5. Given Assumptions 13-22,

$$\boldsymbol{\theta}_0 = \max_{\boldsymbol{\theta}\in\boldsymbol{\Theta}} \mathbb{E}\,\mathcal{L}_{n,T}(\boldsymbol{\theta}) \equiv \max_{\boldsymbol{\theta}\in\boldsymbol{\Theta}} \mathbb{E}\,\frac{1}{nT}\mathcal{L}_{n,T}(\boldsymbol{\theta})$$

PROOF. $L_{n,T}$ is the joint density function of Y_t, X_t for $t = 1, \ldots, T$.

$$\mathbb{E}\,\mathcal{L}_{n,T}(\boldsymbol{\theta}) - \mathbb{E}\,\mathcal{L}_{n,T}(\boldsymbol{\theta}_0) = \mathbb{E}\,\ln\frac{L_{n,T}(\boldsymbol{\theta})}{L_{n,T}(\boldsymbol{\theta}_0)}$$

Denote $Z = (Y_T, X_T, \ldots, Y_1, X_1)$. By Jensen's inequality,

$$\mathbb{E}\,\ln\frac{L_{n,T}(\boldsymbol{\theta})}{L_{n,T}(\boldsymbol{\theta}_0)} \leq \ln\mathbb{E}\,\frac{L_{n,T}(\boldsymbol{\theta})}{L_{n,T}(\boldsymbol{\theta}_0)} = \ln\int_{-\infty}^{\infty}\frac{L_{n,T}(\boldsymbol{\theta})}{L_{n,T}(\boldsymbol{\theta}_0)}L_{n,T}(\boldsymbol{\theta}_0)\,dZ = 0$$

So $\mathbb{E}\,\mathcal{L}_{n,T}(\boldsymbol{\theta}) < \mathbb{E}\,\mathcal{L}_{n,T}(\boldsymbol{\theta}_0)$. By Lemma 4, the PSTAR(p)-ANN model is globally identified and therefore $\mathbb{E}\,\mathcal{L}_{n,T}(\boldsymbol{\theta})$ is uniquely maximized at $\boldsymbol{\theta}_0$ for all n, T. Since the parameter vector $\boldsymbol{\theta}$ does not depend on n and T, it is equivalent to say that $\boldsymbol{\theta}_0 = \max_{\boldsymbol{\theta}\in\boldsymbol{\Theta}}\mathbb{E}\,\frac{1}{nT}\mathcal{L}_{n,T}(\boldsymbol{\theta})$.

□

We define a Hadamard product denoted by $\circ$, s.t. for vectors $a, b_1, \ldots, b_n \in \mathbb{R}^n$, a matrix $B = (b_1, \ldots, b_n) \in \mathbb{R}^{n\times n}$,

$$a \circ B = \begin{bmatrix} a_1b_{11} & a_1b_{21} & \cdots & a_1b_{n1} \\ a_2b_{12} & a_2b_{22} & \cdots & a_2b_{n2} \\ \vdots & \vdots & \ddots & \vdots \\ a_nb_{1n} & a_nb_{2n} & \cdots & a_nb_{nn} \end{bmatrix}, a \circ b_1 = \begin{bmatrix} a_1b_{11} \\ a_2b_{12} \\ \vdots \\ a_nb_{1n} \end{bmatrix}$$

And let

$$k_0 = \int\left|\frac{f'(s)}{f(s)}\right| f(s)\,ds$$

$$k_1 = \int\left|\frac{f'^2(s)}{f^2(s)} - \frac{f''(s)}{f(s)}\right| f(s)\,ds$$

$$k_2 = \int\left|\frac{sf'^2(s)}{f(s)} - \frac{sf''(s)}{f(s)}\right| f(s)\,ds$$

$$k_3 = \int\left|\frac{s^2f'^2(s)}{f(s)} - \frac{s^2f''(s)}{f(s)}\right| f(s)\,ds$$

To facilitate the proof later on, we provide a lemma as follows.

Lemma 6. Given Assumptions 13-23,

$$\sup_{\boldsymbol{\theta}\in\boldsymbol{\Theta}}\left|\frac{1}{nT}\sum_{s=1}^{n}\sum_{t=1}^{T}\ln f(\varepsilon_{s,t}(\boldsymbol{\theta}))-\mathbb{E}\,\frac{1}{nT}\sum_{s=1}^{n}\sum_{t=1}^{T}\ln f(\varepsilon_{s,t}(\boldsymbol{\theta}))\right|\xrightarrow{p}0 \text{ as } n,T\to\infty \tag{2.9}$$

Proof. As illustrated in equation (2.8), in a lattice with size $n_1 \times n_2$,

$$\sup_{\boldsymbol{\theta}\in\boldsymbol{\Theta}}\left|\frac{1}{nT}\sum_{t=1}^{T}\sum_{s=1}^{n}\ln f(\varepsilon_{s,t}(\boldsymbol{\theta}))-\frac{1}{nT}\sum_{t=1}^{T}\sum_{s\in\mathcal{S}}\ln f(\varepsilon_{s,t}(\boldsymbol{\theta}))\right|\xrightarrow{a.s.}0 \text{ as } n_1,n_2,T\to\infty$$

Therefore, to prove (2.9) is equivalent to show that

$$\sup_{\boldsymbol{\theta}\in\boldsymbol{\Theta}}\left|\frac{1}{nT}\sum_{t=1}^{T}\left(\sum_{s\in\mathcal{S}}\ln f(\varepsilon_{s,t}(\boldsymbol{\theta}))-\mathbb{E}\,\frac{1}{n}\sum_{s\in\mathcal{S}}\ln f(\varepsilon_{s,t}(\boldsymbol{\theta}))\right)\right|\xrightarrow{p}0 \text{ as } n_1,n_2,T\to\infty \tag{2.10}$$

where $\mathcal{S}$ denotes the interior units mentioned before. Since the interior units have the same neighboring structure, the space process for them is stationary when n_1, n_2 go to infinity. We first show $\left|\frac{1}{nT}\sum_{t=1}^{T}\left(\sum_{s\in\mathcal{S}}\ln f(\varepsilon_s(\boldsymbol{\theta}))-\mathbb{E}\,\frac{1}{n}\sum_{s\in\mathcal{S}}\ln f(\varepsilon_s(\boldsymbol{\theta}))\right)\right|\xrightarrow{p}0$ for fixed $\boldsymbol{\theta}$.

To prove this, we want to show that $\mathbb{E}\,|\ln f(\varepsilon_{s,t}(\boldsymbol{\theta}))|<\infty$. Expanding $\ln f(\varepsilon_{s,t}(\boldsymbol{\theta}))$ around $\boldsymbol{\theta}_0$ with respect to $\boldsymbol{\theta}$,

$$\ln f(\varepsilon_{s,t}(\boldsymbol{\theta}))=\ln f(\varepsilon_{s,t}(\boldsymbol{\theta}_0))+\left|\frac{f'(\varepsilon_{s,t}(\tilde{\boldsymbol{\theta}}_{n,T}))}{f(\varepsilon_{s,t}(\tilde{\boldsymbol{\theta}}_n))}\frac{\partial\varepsilon_{s,t}(\tilde{\boldsymbol{\theta}}_{n,T})}{\partial\boldsymbol{\theta}'}\right|(\boldsymbol{\theta}-\boldsymbol{\theta}_0)$$

$$\mathbb{E}\,|\ln f(\varepsilon_{s,t}(\boldsymbol{\theta}))|\le\mathbb{E}\,|\ln f(\varepsilon_{s,t}(\boldsymbol{\theta}_0))|+\mathbb{E}\left|\frac{f'(\varepsilon_{s,t}(\tilde{\boldsymbol{\theta}}_{n,T}))}{f(\varepsilon_{s,t}(\tilde{\boldsymbol{\theta}}_{n,T}))}\frac{\partial\varepsilon_{s,t}(\tilde{\boldsymbol{\theta}}_{n,T})}{\partial\boldsymbol{\theta}'}\right||\boldsymbol{\theta}-\boldsymbol{\theta}_0|$$

where $\tilde{\boldsymbol{\theta}}_{n,T}$ is between $\boldsymbol{\theta}$ and $\boldsymbol{\theta}_0$. Under the true parameter values, $\varepsilon_{s,t}(\boldsymbol{\theta}_0)$ (denoted as $\varepsilon_{s,t}$ or $\boldsymbol{\varepsilon}_t$ as its vector form in the following) is independent and identically distributed. From Assumption 18, $\mathbb{E}\,|\ln f(\varepsilon_{s,t})|<\infty$. For $\mathbb{E}\left|\frac{f'(\varepsilon_{s,t}(\tilde{\boldsymbol{\theta}}_{n,T}))}{f(\varepsilon_{s,t}(\tilde{\boldsymbol{\theta}}_{n,T}))}\frac{\partial\varepsilon_{s,t}(\tilde{\boldsymbol{\theta}}_{n,T})}{\partial\boldsymbol{\theta}'}\right|$, $\left|\frac{\partial\varepsilon_{s,t}(\tilde{\boldsymbol{\theta}}_{n,T})}{\partial\boldsymbol{\theta}}\right|$ can be

expressed as

$$
\begin{aligned}
&\left|\frac{\partial \varepsilon_{s,t}(\tilde{\boldsymbol{\theta}}_{n,T})}{\partial \beta}\right| = |x_{s,t}| \\
&\left|\frac{\partial \varepsilon_{s,t}(\tilde{\boldsymbol{\theta}}_{n,T})}{\partial \lambda}\right| = \left|\boldsymbol{F}(x'_{s,t}\tilde{\boldsymbol{\gamma}}_{n,T})'\right| \leq \mathbf{1}_h \qquad (2.11)\\
&\left|\frac{\partial \varepsilon_{s,t}(\tilde{\boldsymbol{\theta}}_{n,T})}{\partial \boldsymbol{\gamma}_i}\right| = \left|\tilde{\lambda}_i \frac{\partial F(x'_{s,t}\tilde{\boldsymbol{\gamma}}_i)}{\partial x'_{s,t}\boldsymbol{\gamma}_i} x_{s,t}\right| = \left|\tilde{\lambda}_i F(x'_{s,t}\tilde{\boldsymbol{\gamma}}_i)(1 - F(x'_{s,t}\tilde{\boldsymbol{\gamma}}_i))x_{s,t}\right| \\
&\qquad\qquad \leq \max_{\lambda_i \in \boldsymbol{\Theta}} \frac{|\lambda_i x_{s,t}|}{4} \\
&\left|\frac{\partial \varepsilon_{s,t}(\tilde{\boldsymbol{\theta}}_{n,T})}{\partial \phi_i}\right| = \left|\sum_{k=1}^{n} w_{sk} y_{k,t-i}\right| = \left|\left[W_n A^{-1}(L)(\boldsymbol{g}(X_{t-i}, \boldsymbol{\theta}_0) + \boldsymbol{\varepsilon}_{t-i}(\boldsymbol{\theta}_0)\right]_s\right|
\end{aligned}
$$

where $A^{-1}(L)(\boldsymbol{g}(X_{t-i}, \boldsymbol{\theta}_0) + \boldsymbol{\varepsilon}_{t-i}(\boldsymbol{\theta}_0) = \sum_{j=0}^{\infty} \Psi_j A_0^{-1}(\boldsymbol{g}(X_{t-i-j}, \boldsymbol{\theta}_0) + \boldsymbol{\varepsilon}_{t-i-j}(\boldsymbol{\theta}_0))$. Function $g(x_{s,t}, \boldsymbol{\theta}) = x'_{s,t}\beta + \boldsymbol{F}(x'_{s,t}\gamma)\lambda$. Consider $\boldsymbol{\varepsilon}_t(\tilde{\boldsymbol{\theta}}_{n,T})$,

$$
\begin{aligned}
|\boldsymbol{\varepsilon}_t(\tilde{\boldsymbol{\theta}}_{n,T})| &= \left|(I_n - \tilde{\phi}_0 W_n)Y_t - \sum_{i=1}^{p} \tilde{\phi}_i W_n Y_{t-i} - \boldsymbol{g}(X_t, \tilde{\boldsymbol{\theta}}_{n,T})\right| \\
&= \left|\boldsymbol{\varepsilon}_t(\boldsymbol{\theta}_0) + \sum_{i=0}^{p}(\phi_{i0} - \tilde{\phi}_i)W_n Y_{t-i} + (\boldsymbol{g}(X_t, \boldsymbol{\theta}_0) - \boldsymbol{g}(X_t, \tilde{\boldsymbol{\theta}}_{n,T}))\right| \\
&= \Bigg|\boldsymbol{\varepsilon}_t + \sum_{i=0}^{p}(\phi_{i0} - \tilde{\phi}_i)W_n \sum_{j=0}^{\infty} \Psi_j A_0^{-1} \boldsymbol{\varepsilon}_{t-i-j} + \sum_{i=0}^{p}(\phi_{i0} - \tilde{\phi}_i)W_n \sum_{j=0}^{\infty} \Psi_j A_0^{-1} X_{t-i-j}\beta_0 \\
&\quad + \sum_{i=0}^{p}(\phi_{i0} - \tilde{\phi}_i)W_n \sum_{j=0}^{\infty} \Psi_j A_0^{-1} \boldsymbol{F}(X_{t-i-j}\boldsymbol{\gamma}'_0)\lambda_0 + X_t(\beta_0 - \tilde{\beta}) \\
&\quad + \boldsymbol{F}(X_t\boldsymbol{\gamma}'_0)\lambda_0 - \boldsymbol{F}(X_t\tilde{\boldsymbol{\gamma}}')\tilde{\lambda}\Bigg| \\
&< \Bigg|\boldsymbol{\varepsilon}_t + \sum_{i=0}^{p}(\phi_{i0} - \tilde{\phi}_i)W_n \sum_{j=0}^{\infty} \Psi_j A_0^{-1} \boldsymbol{\varepsilon}_{t-i-j} + \sum_{i=0}^{p}(\phi_{i0} - \tilde{\phi}_i)W_n \sum_{j=0}^{\infty} \Psi_j A_0^{-1} X_{t-i-j}\beta_0 \\
&\quad + \sum_{i=0}^{p}(\phi_{i0} - \tilde{\phi}_i)W_n \sum_{j=0}^{\infty} \Psi_j A_0^{-1} \boldsymbol{F}(X_{t-i-j}\boldsymbol{\gamma}'_0)\lambda_0 + X_t(\beta_0 - \tilde{\beta})\Bigg| + ||\lambda_0 - \tilde{\lambda}|| \cdot \mathbf{1}_n
\end{aligned}
$$

Denote $P(x^c)$ is a polynomial about x with highest order c. Since we have assumed that $A^{-1}(L)$ existed and the expansion $\sum_{j=0}^{\infty}\Psi_j A_0^{-1}$ is absolutely summable so $W_n\sum_{j=0}^{\infty}\Psi_j A_0^{-1}$ is finite. By Assumption 22-23, $\left|\frac{f'(\varepsilon_{s,t}(\tilde{\boldsymbol{\theta}}_{n,T}))}{f(\varepsilon_{s,t}(\tilde{\boldsymbol{\theta}}_{n,T}))}\right| < a_1+a_2|\varepsilon_{s,t}(\tilde{\boldsymbol{\theta}}_{n,T})|^{c_1}$ and $\mathbb{E}\left|\frac{f'(\varepsilon_{s,t})}{f(\varepsilon_{s,t})}\right|, \mathbb{E}\left|\frac{f'(\varepsilon_{s,t})}{f(\varepsilon_{s,t})}\right|^2$ are dominated by $a_1+a_2|\varepsilon_{s,t}|^{c_1}$, $\mathbb{E}|\varepsilon_{s,t}|^{c_1}<\infty$, $\mathbb{E}|x_{s,t}|^{c_1}<\infty$. Let $c^*=\max(1,c_1)$, then,

$$\mathbb{E}\left|\frac{f'(\varepsilon_{s,t}(\tilde{\boldsymbol{\theta}}_{n,T}))}{f(\varepsilon_{s,t}(\tilde{\boldsymbol{\theta}}_{n,T}))}\right|^2 < P(\mathbb{E}\ |\varepsilon_{s,t}|^{c^*}) + P(\mathbb{E}\ |x_{s,t}|^{c^*}) + Constant < \infty$$

So also $\mathbb{E}\left|\frac{f'(\varepsilon_{s,t}(\tilde{\boldsymbol{\theta}}_{n,T}))}{f(\varepsilon_{s,t}(\tilde{\boldsymbol{\theta}}_{n,T}))}\right| < \infty$. With Cauchy–Schwarz inequality [43] and the finite second moment of $x_{s,t}$, we can have,

(2.12)

$$\mathbb{E}\left|\frac{f'(\varepsilon_{s,t}(\tilde{\boldsymbol{\theta}}_{n,T}))}{f(\varepsilon_{s,t}(\tilde{\boldsymbol{\theta}}_{n,T}))}\frac{\partial\varepsilon_{s,t}(\tilde{\boldsymbol{\theta}}_{n,T})}{\partial\beta}\right| = \mathbb{E}\left|\frac{f'(\varepsilon_{s,t}(\tilde{\boldsymbol{\theta}}_{n,T}))}{f(\varepsilon_{s,t}(\tilde{\boldsymbol{\theta}}_{n,T}))}x_{s,t}\right| < \left(\mathbb{E}\left|\frac{f'(\varepsilon_s(\tilde{\boldsymbol{\theta}}_{n,T}))}{f(\varepsilon_s(\tilde{\boldsymbol{\theta}}_{n,T}))}\right|^2\mathbb{E}\,|x_{s,t}|^2\right)^{1/2} < \infty$$

(2.13)

$$\mathbb{E}\left|\frac{f'(\varepsilon_{s,t}(\tilde{\boldsymbol{\theta}}_{n,T}))}{f(\varepsilon_{s,t}(\tilde{\boldsymbol{\theta}}_{n,T}))}\frac{\partial\varepsilon_{s,t}(\tilde{\boldsymbol{\theta}}_{n,T})}{\partial\lambda}\right| = \mathbb{E}\left|\frac{f'(\varepsilon_{s,t}(\tilde{\boldsymbol{\theta}}_{n,T}))}{f(\varepsilon_{s,t}(\tilde{\boldsymbol{\theta}}_{n,T}))}\boldsymbol{F}(x'_{s,t}\tilde{\boldsymbol{\gamma}})'\right| \le \mathbb{E}\left|\frac{f'(\varepsilon_{s,t}(\tilde{\boldsymbol{\theta}}_{n,T}))}{f(\varepsilon_{s,t}(\tilde{\boldsymbol{\theta}}_{n,T}))}\mathbf{1}_h\right| < \infty$$

(2.14)

$$\mathbb{E}\left|\frac{f'(\varepsilon_{s,t}(\tilde{\boldsymbol{\theta}}_{n,T}))}{f(\varepsilon_{s,t}(\tilde{\boldsymbol{\theta}}_{n,T}))}\frac{\partial\varepsilon_{s,t}(\tilde{\boldsymbol{\theta}}_{n,T})}{\partial\boldsymbol{\gamma}_i}\right| \le \mathbb{E}\left|\frac{f'(\varepsilon_{s,t}(\tilde{\boldsymbol{\theta}}_{n,T}))}{f(\varepsilon_{s,t}(\tilde{\boldsymbol{\theta}}_{n,T}))}\tilde{\lambda}_i x_{s,t}\right| < \infty$$

(2.15)

$$\mathbb{E}\left|\frac{f'(\varepsilon_{s,t}(\tilde{\boldsymbol{\theta}}_{n,T}))}{f(\varepsilon_{s,t}(\tilde{\boldsymbol{\theta}}_{n,T}))}\frac{\partial\varepsilon_{s,t}(\tilde{\boldsymbol{\theta}}_{n,T})}{\partial\phi_i}\right| = \mathbb{E}\left|\frac{f'(\varepsilon_{s,t}(\tilde{\boldsymbol{\theta}}_{n,T}))}{f(\varepsilon_{s,t}(\tilde{\boldsymbol{\theta}}_{n,T}))}\Big[W_nA^{-1}(L)(\boldsymbol{g}(X_{t-i},\boldsymbol{\theta}_0)+\boldsymbol{\varepsilon}_{t-i}(\boldsymbol{\theta}_0)\Big]_s\right|$$

(2.16)
$$< \mathbb{E}\left|\frac{f'(\varepsilon_{s,t}(\tilde{\boldsymbol{\theta}}_{n,T}))}{f(\varepsilon_{s,t}(\tilde{\boldsymbol{\theta}}_{n,T}))}\Big[W_nA^{-1}(L)\boldsymbol{\varepsilon}_{t-i}(\boldsymbol{\theta}_0)\Big]_s\right|$$

(2.17)
$$+ k_0\mathbb{E}\left|\Big[W_nA^{-1}(L)\boldsymbol{g}(X_{t-i},\boldsymbol{\theta}_0)\Big]_s\right|\ i=0,\ldots,p$$

Because $W_nA^{-1}(L)$ is well defined and X_t is stationary with finite second moment, so component (2.17) is finite. (2.16) is dominated by $P(\mathbb{E}\,|\varepsilon_{s,t}|^{c^*+1})$ so with the dominance assumption, (2.16) is finite. Hence, with (2.12)-(2.17) finite, $\mathbb{E}\,|\ln f(\varepsilon_{s,t}(\boldsymbol{\theta}_0))|<\infty$, we

can conclude that $\mathbb{E}\,|\ln f(\varepsilon_{s,t}(\boldsymbol{\theta}))| < \infty$. Then by ergodic theorem [4],

$$\left|\frac{1}{nT}\sum_{t=1}^{T}\left(\sum_{s\in\mathcal{S}}\ln f(\varepsilon_{s,t}(\boldsymbol{\theta})) - \mathbb{E}\,\frac{1}{n}\sum_{s\in\mathcal{S}}\ln f(\varepsilon_{s,t}(\boldsymbol{\theta}))\right)\right| \xrightarrow{p} 0, \quad n_1, n_2, T \to \infty$$

To complete the proof of uniform convergence, we also need $\dfrac{1}{nT}\displaystyle\sum_{t=1}^{T}\sum_{s\in\mathcal{S}}\ln f(\varepsilon_{s,t}(\boldsymbol{\theta}))$ is equicontinuous for $\boldsymbol{\theta}\in\boldsymbol{\Theta}$, i.e., for all $\boldsymbol{\theta}_1, \boldsymbol{\theta}_2 \in \boldsymbol{\Theta}$,

$$\frac{1}{nT}\left|\sum_{t=1}^{T}\sum_{s\in\mathcal{S}}\Big(\ln f(\varepsilon_{s,t}(\boldsymbol{\theta}_1)) - \ln f(\varepsilon_{s,t}(\boldsymbol{\theta}_2))\Big)\right| \le ||\boldsymbol{\theta}_1 - \boldsymbol{\theta}_2|| O_p(1) \tag{2.18}$$

Applying the mean value theorem to the left side in (2.18):

$$\begin{aligned}\frac{1}{nT}\left|\sum_{t=1}^{T}\sum_{s\in\mathcal{S}}\Big(\ln f(\varepsilon_{s,t}(\boldsymbol{\theta}_1)) - \ln f(\varepsilon_{s,t}(\boldsymbol{\theta}_2))\Big)\right| &\le \frac{1}{nT}\left|\sum_{t=1}^{T}\sum_{s\in\mathcal{S}}\frac{\partial \ln f(\varepsilon_{s,t}(\tilde{\boldsymbol{\theta}}_{n,T}))}{\partial\boldsymbol{\theta}'}\right| ||\boldsymbol{\theta}_1 - \boldsymbol{\theta}_2|| \\ &= \frac{1}{nT}\left|\sum_{t=1}^{T}\sum_{s\in\mathcal{S}}\frac{f'(\varepsilon_{s,t}(\tilde{\boldsymbol{\theta}}_{n,T}))}{f(\varepsilon_{s,t}(\tilde{\boldsymbol{\theta}}_{n,T}))}\frac{\partial\varepsilon_{s,t}(\tilde{\boldsymbol{\theta}}_{n,T})}{\partial\boldsymbol{\theta}'}\right| ||\boldsymbol{\theta}_1 - \boldsymbol{\theta}_2||\end{aligned}$$

where $\tilde{\boldsymbol{\theta}}_{n,T}$ is some value between $\boldsymbol{\theta}_1$ and $\boldsymbol{\theta}_2$. Since $\boldsymbol{\theta}$ is in a compact set $\boldsymbol{\Theta}$, we show in (2.19) that, for all s, t, $\varepsilon_{s,t}(\boldsymbol{\theta})$ is bounded by some function of Z_t not depending on $\boldsymbol{\theta}$.

$$\begin{aligned}|\boldsymbol{\varepsilon}_t(\boldsymbol{\theta})| &= \left|Y_t - \phi_0 W_n Y_t - \sum_{k=1}^{p}\phi_k W_n Y_{t-k} - X_t\beta - \boldsymbol{F}(X_t\boldsymbol{\gamma}')\lambda\right| \\ &\le |(I_n - \phi_0 W_n)Y_t| + |\sum_{k=1}^{p}\phi_k W_n Y_{t-k}| + |X_n\beta| + \left|\boldsymbol{F}(X_n\boldsymbol{\gamma}')\lambda\right| \qquad (2.19) \\ &\le (I_n + \max_{\phi_0\in\boldsymbol{\Theta}}|\phi_0|W_n)|Y_t| + \sum_{k=1}^{p}\max_{\phi_i\in\boldsymbol{\Theta}} W_n|\phi_i Y_{t-k}| + |X_n|\max_{\beta\in\boldsymbol{\Theta}}|\beta| + \max_{\lambda\in\boldsymbol{\Theta}}||\lambda||\mathbf{1}_n\end{aligned}$$

Similarly, referring to (2.11), it is easy to show that $\left|\frac{\partial\varepsilon_{s,t}(\boldsymbol{\theta})}{\partial\boldsymbol{\theta}}\right|$ is bounded by some function about Y_t and X_t. Therefore, due to the dominance of $\left|\frac{f'(s)}{f(s)}\right|$ (see Assumption 22) and stationarity of X_t, Y_t, for $\tilde{\boldsymbol{\theta}}_{n,T}$ between $\boldsymbol{\theta}_1$ and $\boldsymbol{\theta}_2$, there exists a constant M such that

$$\frac{1}{nT}\left|\sum_{t=1}^{T}\sum_{s\in\mathcal{S}}\frac{f'(\varepsilon_{s,t}(\tilde{\boldsymbol{\theta}}_{n,T}))}{f(\varepsilon_{s,t}(\tilde{\boldsymbol{\theta}}_{n,T}))}\frac{\partial\varepsilon_{s,t}(\tilde{\boldsymbol{\theta}}_{n,T})}{\partial\boldsymbol{\theta}'}\right| \le M \quad \text{for } n_1, n_2, T \to \infty \tag{2.20}$$

Hence, for $\boldsymbol{\theta}_1, \boldsymbol{\theta}_2 \in \boldsymbol{\Theta}$

$$\frac{1}{nT}\left|\sum_{t=1}^{T}\sum_{s\in\mathcal{S}}\Big(\ln f(\varepsilon_{s,t}(\boldsymbol{\theta}_1)) - \ln f(\varepsilon_{s,t}(\boldsymbol{\theta}_2))\Big)\right| = ||\boldsymbol{\theta}_1 - \boldsymbol{\theta}_2|| O_p(1)$$

So $\frac{1}{nT}\sum_{t=1}^{T}\sum_{s\in\mathcal{S}} \ln f(\varepsilon_{s,t}(\boldsymbol{\theta}))$ is equicontinuous for $\boldsymbol{\theta} \in \boldsymbol{\Theta}$. With the pointwise convergence and equicontinuity, we can conclude the uniform convergence in (2.10) and furthermore (2.9) follows. □

Similar to Chapter 1, we now give a formal statement of the consistency results.

Theorem 3. Given Assumptions 13-23, $\hat{\boldsymbol{\theta}}_{n,T} \xrightarrow{p} \boldsymbol{\theta}_0$ as $n, T \to \infty$.

Proof. Similar to the proof by Lung-fei Lee [25], we need to show the stochastic equicontinuity of $\frac{1}{n}\ln|A_0|$ to have the uniform convergence of the log likelihood function $\mathcal{L}_{n,T}(\boldsymbol{\theta})$. Applying the mean value theorem,

$$\left|\frac{1}{n}(\ln|I_n - \phi_0^{\dagger}W_n| - \ln|I_n - \phi_0^{\ddagger}W_n|)\right| = \left|(\phi_0^{\dagger} - \phi_0^{\ddagger})\frac{1}{n}tr(W_n(I_n - \phi^*_{0_{n,T}}W_n)^{-1})\right|$$

where $\phi^*_{0_{n,T}}$ is between $\phi_0^{\dagger}$ and $\phi_0^{\ddagger}$. By Assumption 14 and 15, $\sup_{\phi_0\in\boldsymbol{\Theta}}|\phi_0| < 1$, W_n is bounded in both rows and column sums uniformly and using (2.7),

$$\left|\frac{1}{n}tr(W_n(I_n - \phi^*_{0_{n,T}}W_n)^{-1})\right| = \left|\frac{1}{n}\sum_{i=1}^{n}\frac{\tau_i}{1-\phi^*_{0_{n,T}}\tau_i}\right| \le C_1$$

where C_1 is a constant not depending on n. So $\left|\frac{1}{n}(\ln|I_n - \phi_0^{\dagger}W_n| - \ln|I_n - \phi_0^{\ddagger}W_n|)\right| \le C_1|\phi_0^{\dagger} - \phi_0^{\ddagger}|$ and with Lemma 6 we can conclude the uniform convergence that

$$\sup_{\boldsymbol{\theta}\in\boldsymbol{\Theta}}\left|\frac{1}{nT}\mathcal{L}_{n,T}(\boldsymbol{\theta}) - \mathbb{E}\,\frac{1}{nT}\mathcal{L}_{n,T}(\boldsymbol{\theta})\right| \xrightarrow{p} 0.$$

With the assumptions 13-22, the parameter space $\boldsymbol{\Theta}$ is compact; $\frac{1}{nT}\mathcal{L}_{n,T}(\boldsymbol{\theta})$ is continuous in $\boldsymbol{\theta} \in \boldsymbol{\Theta}$ and is a measurable of $Y_t, X_t, t = 1, \ldots, T$ for all $\boldsymbol{\theta} \in \boldsymbol{\Theta}$. $\mathbb{E}\,\frac{1}{nT}\mathcal{L}_{n,T}(\boldsymbol{\theta})$ is continuous on $\boldsymbol{\Theta}$ and by Lemma 5, $\mathbb{E}\,\frac{1}{nT}\mathcal{L}_{n,T}(\boldsymbol{\theta})$ has a unique maximum at $\boldsymbol{\theta}_0$. Referring

to Theorem 3.5 in White [47] with the uniform convergence in (2.9), we can conclude that $\hat{\boldsymbol{\theta}}_{n,T} \xrightarrow{p} \boldsymbol{\theta}_0$ as $n, T \to \infty$. □

2.4.2. Asymptotic Distribution

Assumption 24. The limit $A(\boldsymbol{\theta}_0) = -\lim_{n,T\to\infty} \mathbb{E}\, \frac{1}{nT} \frac{\partial^2 \mathcal{L}_{n,T}(\boldsymbol{\theta}_0)}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}'}$ is nonsingular.

Assumption 25. The limit $B(\boldsymbol{\theta}_0) = \lim_{n,T\to\infty} \mathbb{E}\, \frac{1}{nT} \frac{\partial \mathcal{L}_{n,T}(\boldsymbol{\theta}_0)}{\partial \boldsymbol{\theta}} \frac{\partial \mathcal{L}_{n,T}(\boldsymbol{\theta}_0)}{\partial \boldsymbol{\theta}'}$ is nonsingular.

These assumptions are to guarantee the existence of the covariance matrix of the limiting distribution of parameters in a PSTAR(p)-ANN model. We now give the asymptotic distribution of the maximum likelihood estimator $\hat{\boldsymbol{\theta}}_{n,T}$.

Theorem 4. Under Assumptions 13-25,

$$\sqrt{nT}(\hat{\boldsymbol{\theta}}_{n,T} - \boldsymbol{\theta}_0) \xrightarrow{d} N(\mathbf{0}, \boldsymbol{\Omega}_0) \tag{2.21}$$

where $\boldsymbol{\Omega}_0 = A(\boldsymbol{\theta}_0)^{-1} B(\boldsymbol{\theta}_0) A(\boldsymbol{\theta}_0)^{-1} = A(\boldsymbol{\theta}_0)^{-1}$

Proof. Since $\hat{\boldsymbol{\theta}}_{n,T}$ maximizes $\mathcal{L}_{n,T}(\boldsymbol{\theta})$, $\frac{\partial \mathcal{L}_{n,T}(\hat{\boldsymbol{\theta}}_{n,T})}{\partial \boldsymbol{\theta}} = 0$. By the mean value theorem, expand $\frac{\partial \mathcal{L}_{n,T}(\hat{\boldsymbol{\theta}}_{n,T})}{\partial \boldsymbol{\theta}}$ around $\boldsymbol{\theta}_0$ with respect to $\boldsymbol{\theta}$,

$$\frac{\partial \mathcal{L}_{n,T}(\hat{\boldsymbol{\theta}}_{n,T})}{\partial \boldsymbol{\theta}} = \frac{\partial \mathcal{L}_{n,T}(\boldsymbol{\theta}_0)}{\partial \boldsymbol{\theta}} + \frac{\partial^2 \mathcal{L}_{n,T}(\tilde{\boldsymbol{\theta}}_{n,T})}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}'} (\hat{\boldsymbol{\theta}}_{n,T} - \boldsymbol{\theta}_0)$$

$$0 = \frac{\partial \mathcal{L}_{n,T}(\boldsymbol{\theta}_0)}{\partial \boldsymbol{\theta}} + \frac{\partial^2 \mathcal{L}_{n,T}(\tilde{\boldsymbol{\theta}}_{n,T})}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}'} (\hat{\boldsymbol{\theta}}_{n,T} - \boldsymbol{\theta}_0)$$

where $\tilde{\boldsymbol{\theta}}_{n,T}$ is between $\hat{\boldsymbol{\theta}}_{n,T}$ and $\boldsymbol{\theta}_0$. Therefore, we can have the following equation:

$$\sqrt{nT}(\hat{\boldsymbol{\theta}}_{n,T} - \boldsymbol{\theta}_0) = \left[-\frac{1}{nT} \frac{\partial^2 \mathcal{L}_{n,T}(\tilde{\boldsymbol{\theta}}_{n,T})}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}'} \right]^{-1} \frac{1}{\sqrt{nT}} \frac{\partial \mathcal{L}_{n,T}(\boldsymbol{\theta}_0)}{\partial \boldsymbol{\theta}} \tag{2.22}$$

From (2.11), denote $\frac{\boldsymbol{f}'(\varepsilon_t \boldsymbol{\theta})}{\boldsymbol{f}(\varepsilon_t(\boldsymbol{\theta}))}$ as $V_t(\boldsymbol{\theta}) \in \mathbb{R}^n$ and $\frac{\boldsymbol{f}'(\varepsilon_t \boldsymbol{\theta}_0)}{\boldsymbol{f}(\varepsilon_t(\boldsymbol{\theta}_0))} = V_t$.

Recall that $Z_t = (W_nY_t, W_nY_{t-1}, \ldots, W_nY_{t-p}, X_t)$ so the first order derivatives can be expressed as

$$\frac{1}{\sqrt{nT}}\frac{\partial \mathcal{L}_{n,T}(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} = \begin{pmatrix} -\frac{1}{\sqrt{nT}}\sum_{t=1}^{T}\left((W_nY_t)'V_t(\boldsymbol{\theta}) + tr(W_nA_0^{-1})\right) \\ -\frac{1}{\sqrt{nT}}\sum_{t=1}^{T} Z_t'V_t(\boldsymbol{\theta}) \\ -\frac{1}{\sqrt{nT}}\sum_{t=1}^{T}\left(\boldsymbol{F}(X_t\boldsymbol{\gamma}')\right)'V_t(\boldsymbol{\theta}) \\ -\frac{\lambda_1}{\sqrt{nT}}\sum_{t=1}^{T} X_t'\left(\boldsymbol{F}'(X_t\boldsymbol{\gamma}_1)\circ V_t(\boldsymbol{\theta})\right) \\ \vdots \\ -\frac{\lambda_h}{\sqrt{nT}}\sum_{t=1}^{T} X_t'\left(\boldsymbol{F}'(X_t\boldsymbol{\gamma}_h)\circ V_t(\boldsymbol{\theta})\right) \end{pmatrix} \tag{2.23}$$

By Lemma 5, the true parameter values maximize $\frac{1}{nT}\mathbb{E}\,\mathcal{L}_{n,T}(\boldsymbol{\theta})$, so $\frac{1}{nT}\frac{\partial \mathbb{E}\,\mathcal{L}_{n,T}(\boldsymbol{\theta}_0)}{\partial \boldsymbol{\theta}} = \mathbf{0}$. In (2.12)-(2.17) and (2.19), we showed that $\mathbb{E}\left|\frac{\partial \ln f(\varepsilon_{s,t}(\boldsymbol{\theta}))}{\partial \boldsymbol{\theta}}\right|$ is dominated by some function not related to $\boldsymbol{\theta}$ and (2.20) indicates that $\mathbb{E}\left|\frac{\partial \ln f(\varepsilon_{s,t}(\boldsymbol{\theta}))}{\partial \boldsymbol{\theta}}\right|$ is bounded for interior units in $\mathcal{S}$. Hence,$\mathbb{E}\frac{\partial \ln f(\varepsilon_{s,t}(\boldsymbol{\theta}))}{\partial \boldsymbol{\theta}} = \frac{\partial}{\partial \boldsymbol{\theta}}\mathbb{E}\ln f(\varepsilon_{s,t}(\boldsymbol{\theta}))$, it follows that, with $\frac{1}{nT}\mathcal{L}_{n,T}(\boldsymbol{\theta}) = \frac{1}{n}\ln|A_0| + \frac{1}{nT}\sum_{s=1}^{n}\sum_{t=1}^{T}\ln f(\varepsilon_{s,t}(\boldsymbol{\theta}))$, we can have,

$$\frac{1}{nT}\frac{\partial \mathbb{E}\,\mathcal{L}_{n,T}(\boldsymbol{\theta}_0)}{\partial \boldsymbol{\theta}} = \frac{1}{nT}\mathbb{E}\,\frac{\partial \mathcal{L}_{n,T}(\boldsymbol{\theta}_0)}{\partial \boldsymbol{\theta}} = \mathbf{0}$$

Therefore, with Assumption 25,

$$\mathrm{Var}\left(\frac{1}{\sqrt{nT}}\frac{\partial \mathcal{L}_{n,T}(\boldsymbol{\theta}_0)}{\partial \boldsymbol{\theta}}\right) = -\mathbb{E}\frac{1}{nT}\frac{\partial^2 \mathcal{L}_{n,T}(\boldsymbol{\theta}_0)}{\partial \boldsymbol{\theta}\partial \boldsymbol{\theta}'} = \mathbb{E}\left(\frac{1}{nT}\frac{\partial \mathcal{L}_{n,T}(\boldsymbol{\theta}_0)}{\partial \boldsymbol{\theta}}\frac{\partial \mathcal{L}_{n,T}(\boldsymbol{\theta}_0)}{\partial \boldsymbol{\theta}'}\right) \to B(\boldsymbol{\theta}_0)$$

And under this $A(\boldsymbol{\theta}_0) = B(\boldsymbol{\theta}_0)$. Since $\frac{\partial \mathcal{L}_{n,T}(\boldsymbol{\theta}_0)}{\partial \boldsymbol{\theta}}$ is the sum of T identical and ergodic random variables, by the central limit theorem for stationary ergodic processes [33], the limiting distribution of $\frac{1}{\sqrt{nT}}\frac{\partial \mathcal{L}_{n,T}(\boldsymbol{\theta}_0)}{\partial \boldsymbol{\theta}}$ is $N(\mathbf{0}, B(\boldsymbol{\theta}_0))$.

Next we would like to show that $\frac{1}{nT}\frac{\partial^2 \mathcal{L}_{n,T}(\tilde{\boldsymbol{\theta}}_{n,T})}{\partial \boldsymbol{\theta}\partial \boldsymbol{\theta}'} - \frac{1}{nT}\frac{\partial^2 \mathcal{L}_{n,T}(\boldsymbol{\theta}_0)}{\partial \boldsymbol{\theta}\partial \boldsymbol{\theta}'} \xrightarrow{p} 0$. Following the results in (2.23), define $U_t(\boldsymbol{\theta}) = \frac{\boldsymbol{f}''(\varepsilon_t(\boldsymbol{\theta}))}{\boldsymbol{f}(\varepsilon_t(\boldsymbol{\theta}))} - \frac{\boldsymbol{f}'^2(\varepsilon_t(\boldsymbol{\theta}))}{\boldsymbol{f}^2(\varepsilon_t(\boldsymbol{\theta}))} \in \mathbb{R}^n$, and write $U_t = U_t(\boldsymbol{\theta}_0)$ so the

second order derivatives are given below $-\frac{1}{nT}\frac{\partial^2 \mathcal{L}_{n,T}(\boldsymbol{\theta})}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}'} =$

$$
\begin{array}{ll}
(2.24) & \frac{1}{nT}\sum_{t=1}^{T}\begin{pmatrix}
G_{0,t}(\boldsymbol{\theta}) & (W_nY_t)'G_{1,t}(\boldsymbol{\theta}) & (W_nY_t)'G_{2,t}(\boldsymbol{\theta}) & (W_nY_t)'H_{1,t}(\boldsymbol{\theta}) & \cdots & (W_nY_t)'H_{h,t}(\boldsymbol{\theta}) \\
G'_{1,t}(\boldsymbol{\theta})W_nY_t & Z_t'G_{1,t}(\boldsymbol{\theta}) & Z_t'G_{2,t}(\boldsymbol{\theta}) & Z_t'H_{1,t}(\boldsymbol{\theta}) & \cdots & Z_t'H_{h,t}(\boldsymbol{\theta}) \\
G'_{2,t}(\boldsymbol{\theta})W_nY_t & G'_{2,t}(\boldsymbol{\theta})Z_t & \boldsymbol{F}(X_t\boldsymbol{\gamma}')'G_{2,t}(\boldsymbol{\theta}) & \boldsymbol{F}(X_t\boldsymbol{\gamma}')'H_{1,t}(\boldsymbol{\theta}) & \cdots & \boldsymbol{F}(X_t\boldsymbol{\gamma}')'H_{h,t}(\boldsymbol{\theta}) \\
 & & & +K_{1,t}(\boldsymbol{\theta}) & \cdots & +K_{h,t}(\boldsymbol{\theta}) \\
H'_{1,t}(\boldsymbol{\theta})W_nY_t & H'_{1,t}(\boldsymbol{\theta})Z_t & H'_{1,t}(\boldsymbol{\theta})\boldsymbol{F}(X_t\boldsymbol{\gamma}') & & & \\
 & & +K_{1,t}(\boldsymbol{\theta})' & & & \\
\vdots & \vdots & \vdots & & J(\boldsymbol{\theta}) & \\
H'_{h,t}(\boldsymbol{\theta})W_nY_t & H'_{h,t}(\boldsymbol{\theta})Z_t & H'_{h,t}(\boldsymbol{\theta})\boldsymbol{F}(X_t\boldsymbol{\gamma}') & & & \\
 & & +K_{h,t}(\boldsymbol{\theta})' & & &
\end{pmatrix}
\end{array}
$$

$$
J_{ij,t}(\boldsymbol{\theta}) = \begin{cases}
\lambda_i X_t'(\boldsymbol{F}''(X_t\boldsymbol{\gamma}_i)\circ V_t(\boldsymbol{\theta})\circ X_t) + \lambda_i X_t'(\boldsymbol{F}'(X_t\boldsymbol{\gamma}_i)\circ H_{i,t}) & i = j \\
\lambda_i(\boldsymbol{F}'(X_t\boldsymbol{\gamma}_i)\circ H_{j,t})'X_t & i > j \quad i,j = 1,2,\ldots,h \\
\lambda_i X_t'(\boldsymbol{F}'(X_t\boldsymbol{\gamma}_i)\circ H_{j,t}) & i < j
\end{cases}
$$

$$G_{0,t}(\boldsymbol{\theta}) = \left(-W_nY_t\circ W_nY_t\right)'U_t(\boldsymbol{\theta}) + tr((W_nA_0^{-1})^2)$$

$$G_{1,t}(\boldsymbol{\theta}) = -U_t(\boldsymbol{\theta})\circ Z_t$$

$$G_{2,t}(\boldsymbol{\theta}) = -U_t(\boldsymbol{\theta})\circ \boldsymbol{F}(X_t\boldsymbol{\gamma}')$$

$$H_{i,t}(\boldsymbol{\theta}) = -U_t(\boldsymbol{\theta})\circ(\lambda_i\boldsymbol{F}'(X_t\boldsymbol{\gamma}_i)\circ X_t) \quad i = 1,\ldots,h$$

$$K_{i,t}(\boldsymbol{\theta}) = [V_t(\boldsymbol{\theta})\circ\boldsymbol{F}'(X_t\boldsymbol{\gamma}')]'X_t\circ e_i \quad i = 1,\ldots,h \quad k = 1,\ldots,h$$

$$e_{i,k} = \begin{cases} 1 & k = i \\ 0 & k \neq i \end{cases}$$

Since $\tilde{\boldsymbol{\theta}}_{n,T}$ is between $\hat{\boldsymbol{\theta}}_{n,T}$ and $\boldsymbol{\theta}_0$, $\hat{\boldsymbol{\theta}}_{n,T} \xrightarrow{p} \boldsymbol{\theta}_0$ so $\tilde{\boldsymbol{\theta}}_{n,T}$ also converges to $\boldsymbol{\theta}_0$ in probability as $n \to \infty$. By Assumption 22, $\left|\frac{f'(s)}{f(s)}\right|$, $\left|\frac{f''(s)}{f(s)}\right|$ and $\left|\frac{f'^2(s)}{f^2(s)}\right|$ are continuous and are bounded by $a_1 + a_2|s|^{c_1}$ so $U_t(\boldsymbol{\theta}), V_t(\boldsymbol{\theta})$ are continuous. With $\phi_0 \in (-\frac{1}{\tau}, \frac{1}{\tau})$, $tr((W_nA_0^{-1})^2) = \sum_{i=1}^{n}\frac{\tau_i^2}{(1-\phi_0\tau_i)^2}$ is also a continuous function of ϕ_0.

Therefore elements in $\frac{1}{nT}\frac{\partial^2 \mathcal{L}_{n,T}(\boldsymbol{\theta})}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}'}$ are continuous functions for $\boldsymbol{\theta}$ in $\boldsymbol{\Theta}$. Then by the continuity,

$$\frac{1}{nT}\frac{\partial^2 \mathcal{L}_{n,T}(\tilde{\boldsymbol{\theta}}_{n,T})}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}'} - \frac{1}{nT}\frac{\partial^2 \mathcal{L}_{n,T}(\boldsymbol{\theta}_0)}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}'} \xrightarrow{p} 0, \quad \text{as } \tilde{\boldsymbol{\theta}}_{n,T} \xrightarrow{p} \boldsymbol{\theta}_0 \tag{2.25}$$

Finally we will prove that $\left|\frac{1}{nT}\frac{\partial^2 \mathcal{L}_{n,T}(\boldsymbol{\theta}_0)}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}'} - \mathbb{E}\frac{1}{nT}\frac{\partial^2 \mathcal{L}_{n,T}(\boldsymbol{\theta}_0)}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}'}\right| \xrightarrow{p} 0$. Since $\ln|A_0|$ can be decomposed as $\sum_{i=1}^{n}\ln(1-\phi_0\tau_i)$, to show $\mathbb{E}\frac{1}{nT}\frac{\partial^2 \mathcal{L}_{n,T}(\boldsymbol{\theta}_0)}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}'} < \infty$ is equivalent to show

$$\mathbb{E}\left|\frac{\partial^2}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}'}\left(\frac{1}{nT}\sum_{t=1}^{T}\sum_{s=1}^{n}\ln(1-\phi_{00}\tau_s) + \ln f(\varepsilon_{s,t}(\boldsymbol{\theta}_0))\right)\right| < \infty \tag{2.26}$$

We first discuss the second derivative with respect to ϕ_0 component in (2.26). By triangular inequality,

$$\mathbb{E}\left|\frac{\partial^2}{\partial\phi_0\partial\phi_0}\frac{1}{nT}\sum_{t=1}^{T}\sum_{s=1}^{n}\Big(\ln(1-\phi_{00}\tau_s) + \ln f(\varepsilon_{s,t}(\boldsymbol{\theta}_0))\Big)\right| < \mathbb{E}\left|\frac{1}{n}\sum_{s=1}^{n}\frac{\partial^2\ln(1-\phi_{00}\tau_s)}{\partial\phi_0\partial\phi_0}\right| + \mathbb{E}\left|\frac{\partial^2 \ln f(\varepsilon_{s,t}(\boldsymbol{\theta}_0))}{\partial\phi_0\partial\phi_0}\right|$$

where ϕ_{00} is the true value of ϕ_0. Consider $\mathbb{E}\left|\frac{1}{n}\sum_{s=1}^{n}\frac{\partial^2\ln(1-\phi_{00}\tau_i)}{\partial\phi_0\partial\phi_0}\right| + \mathbb{E}\left|\frac{\partial^2 \ln f(\varepsilon_{s,t}(\boldsymbol{\theta}_0))}{\partial\phi_0\partial\phi_0}\right|$, under stationarity, it can be simplified as

$$\frac{1}{n}tr(W_nA_0^{-1})^2 + \mathbb{E}\left|\left(\frac{f'^2(\varepsilon_{s,t})}{f^2(\varepsilon_{s,t})} - \frac{f''(\varepsilon_{s,t})}{f(\varepsilon_{s,t})}\right)\left(\sum_{k=1}^{n}w_{sk}y_{k,t}\right)^2\right| \tag{2.27}$$

Define $M_n = \{m_{i,j}\} = W_nA_0^{-1}$ and by assumptions, M_n is uniformly bounded in row and column. Suppose the row sum or column sum of M_n is bounded by a constant b. We know $\frac{1}{n}tr(W_nA_0^{-1})^2 < \infty$. So we only need to show $\mathbb{E}\left|\left(\frac{f'^2(\varepsilon_{s,t})}{f^2(\varepsilon_{s,t})} - \frac{f''(\varepsilon_{s,t})}{f(\varepsilon_{s,t})}\right)\left(\sum_{k=1}^{n}w_{sk}y_{k,t}\right)^2\right| < \infty$.

By simple linear algebra,

$$\begin{aligned} Y_t &= \sum_{j=0}^{\infty}\Psi_jA_0^{-1}(X_{t-j}\beta + \boldsymbol{F}(X_{t-j}\boldsymbol{\gamma}') + \boldsymbol{\varepsilon}_{t-j}) \\ &= \sum_{j=0}^{\infty}\Psi_jA_0^{-1}(\boldsymbol{g}(X_{t-j},\boldsymbol{\theta}_0) + \boldsymbol{\varepsilon}_{t-j}) \end{aligned}$$

So $W_nY_t = W_n\sum_{j=0}^{\infty}\Psi_j A_0^{-1}(\boldsymbol{g}(X_{t-j},\boldsymbol{\theta}_0)+\boldsymbol{\varepsilon}_{t-j})$. Therefore $(\sum_{k=1}^{n} w_{sk}y_{k,t})^2$ is the s^{th} component of $(W_nY_t \circ W_nY_t)$ and we expand $(W_nY_t \circ W_nY_t)_s =$

$$\left[W_n\sum_{j=0}^{\infty}\Psi_j A_0^{-1}\boldsymbol{g}(X_{t-j},\boldsymbol{\theta}_0)\circ W_n\sum_{j=0}^{\infty}\Psi_j A_0^{-1}\boldsymbol{g}(X_{t-j},\boldsymbol{\theta}_0)\right]_s \tag{2.28}$$

$$+\left[2W_n\sum_{j=0}^{\infty}\Psi_j A_0^{-1}\boldsymbol{g}(X_{t-j},\boldsymbol{\theta}_0)\circ W_n\sum_{j=0}^{\infty}\Psi_j A_0^{-1}\boldsymbol{\varepsilon}_{t-j}\right]_s \tag{2.29}$$

$$+\left[W_n\sum_{j=0}^{\infty}\Psi_j A_0^{-1}\boldsymbol{\varepsilon}_{t-j}\circ W_n\sum_{j=0}^{\infty}\Psi_j A_0^{-1}\boldsymbol{\varepsilon}_{t-j}\right]_s \tag{2.30}$$

From assumptions 15 and 16, we know that W_n is uniformly bounded and $\sum_{j=0}^{\infty}\Psi_j A_0^{-1}$ is absolute summable so (2.28) $< \infty$ under the stationary condition of X_t. Hence, $\mathbb{E}\left|\left(\frac{f'^2(\varepsilon_{s,t})}{f^2(\varepsilon_{s,t})}-\frac{f''(\varepsilon_{s,t})}{f(\varepsilon_{s,t})}\right)\cdot(2.28)\right| < \infty$.

For (2.29), when $j>0$, $\boldsymbol{\varepsilon}_{t-j}$ is independent from $\boldsymbol{\varepsilon}_t$. So for all k when $j>0$,

$$\mathbb{E}\left|\left(\frac{f'^2(\varepsilon_{s,t})}{f^2(\varepsilon_{s,t})}-\frac{f''(\varepsilon_{s,t})}{f(\varepsilon_{s,t})}\right)\cdot\left[W_n\sum_{j=1}^{\infty}\Psi_j A_0^{-1}\boldsymbol{\varepsilon}_{t-j}\right]_s\right| = k_1\cdot\left[W_n\sum_{j=1}^{\infty}\Psi_j A_0^{-1}\mathbb{E}\left|\boldsymbol{\varepsilon}_{t-j}\right|\right]_s < \infty$$

when $j=0$, this reduces to $\mathbb{E}\left|\left(\frac{f'^2(\varepsilon_{s,t})}{f^2(\varepsilon_{s,t})}-\frac{f''(\varepsilon_{s,t})}{f(\varepsilon_{s,t})}\right)\cdot\left[W_nA_0^{-1}\boldsymbol{\varepsilon}_t\right]_s\right| < k_1|b-m_{ss}|+k_2|m_{ss}|$. So $\mathbb{E}\left|\left(\frac{f'^2(\varepsilon_{s,t})}{f^2(\varepsilon_{s,t})}-\frac{f''(\varepsilon_{s,t})}{f(\varepsilon_{s,t})}\right)\cdot(2.29)\right| < \infty$.

For (2.30), similar to (2.29), we can have $\mathbb{E}\left|\left(\frac{f'^2(\varepsilon_{s,t})}{f^2(\varepsilon_{s,t})}-\frac{f''(\varepsilon_{s,t})}{f(\varepsilon_{s,t})}\right)\cdot(2.30)\right| < Constant\cdot(k_2+k_3+\mathbb{E}|\varepsilon_{s,t}|) < \infty$.

Therefore combining all these components together,

$$\mathbb{E}\left|\left(\frac{f'^2(\varepsilon_{s,t})}{f^2(\varepsilon_{s,t})}-\frac{f''(\varepsilon_{s,t})}{f(\varepsilon_{s,t})}\right)\left(\sum_{k=1}^{n}w_{sk}y_{k,t}\right)^2\right|$$

$$=\mathbb{E}\left|\left(\frac{f'^2(\varepsilon_{s,t})}{f^2(\varepsilon_{s,t})}-\frac{f''(\varepsilon_{s,t})}{f(\varepsilon_{s,t})}\right)\Big((2.28)+(2.29)+(2.30)\Big)\right| < \infty$$

So equation (2.27) is finite.

Because $\sum_{t=1}^{T}\sum_{s=1}^{n}\ln(1-\phi_0\tau_s)$ in (2.26) only relates to ϕ_0, this term goes away when taken second derivative with respect to other parameters. Similar to the proof of $\mathbb{E}\left|\frac{\partial^2 \ln f(\varepsilon_{s,t}(\boldsymbol{\theta}_0))}{\partial\phi_0\partial\phi_0}\right| < \infty$, we can show that $\mathbb{E}\left|\frac{\partial^2 \ln f(\varepsilon_{s,t}(\boldsymbol{\theta}_0))}{\partial\phi_i\partial\phi_j}\right| < \infty$ for $i = 0, 1, \ldots, p$ and $j = 1, \ldots, p$, i.e.,

$$\mathbb{E}\left|\frac{\partial^2}{\partial\phi_i\partial\phi_j}\frac{1}{nT}\sum_{t=1}^{T}\sum_{s=1}^{n}\Big(\ln(1-\phi_{00}\tau_s)+\ln f(\varepsilon_{s,t}(\boldsymbol{\theta}_0))\Big)\right| < \infty \quad \text{for } i,j = 0,1,\ldots,p$$

Other elements in the matrix (2.26) equal to those in $\mathbb{E}\left|\frac{\partial^2 \ln f(\varepsilon_{s,t}(\boldsymbol{\theta}_0))}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}'}\right|$ and they are also finite.

$$\mathbb{E}\left|\frac{\partial^2 \ln f(\varepsilon_{s,t}(\boldsymbol{\theta}_0))}{\partial\phi_i\partial\beta'}\right| \le Constant\cdot|x'_{s,t}|\,(k_2+k_1\mathbb{E}|\varepsilon_{s,t}|) \tag{2.31}$$

$$\mathbb{E}\left|\frac{\partial^2 \ln f(\varepsilon_{s,t}(\boldsymbol{\theta}_0))}{\partial\phi_i\partial\lambda'}\right| \le Constant\cdot\mathbf{1}'_h\,(k_2+k_1\mathbb{E}|\varepsilon_{s,t}|) \tag{2.32}$$

$$\mathbb{E}\left|\frac{\partial^2 \ln f(\varepsilon_{s,t}(\boldsymbol{\theta}_0))}{\partial\phi_i\partial\gamma'_j}\right| \le Constant\cdot\frac{|\lambda_{j0}x'_{s,t}|}{4}\,(k_2+k_1\mathbb{E}|\varepsilon_{s,t}|) \tag{2.33}$$

$$\mathbb{E}\left|\frac{\partial^2 \ln f(\varepsilon_{s,t}(\boldsymbol{\theta}_0))}{\partial\beta\partial\beta'}\right| = k_1|x_{s,t}x'_{s,t}| \tag{2.34}$$

$$\mathbb{E}\left|\frac{\partial^2 \ln f(\varepsilon_{s,t}(\boldsymbol{\theta}_0))}{\partial\beta\partial\lambda'}\right| = k_1|x_{s,t}\boldsymbol{F}(x'_{s,t}\boldsymbol{\gamma}_0)| \tag{2.35}$$

$$\mathbb{E}\left|\frac{\partial^2 \ln f(\varepsilon_{s,t}(\boldsymbol{\theta}_0))}{\partial\beta\partial\boldsymbol{\gamma}'_j}\right| \le \frac{k_1}{4}|\lambda_{j0}x_{s,t}x'_{s,t}| \tag{2.36}$$

$$\mathbb{E}\left|\frac{\partial^2 \ln f(\varepsilon_{s,t}(\boldsymbol{\theta}_0))}{\partial\lambda\partial\lambda'}\right| = k_1\left|\boldsymbol{F}(x'_{s,t}\boldsymbol{\gamma}_0)'\boldsymbol{F}(x'_{s,t}\boldsymbol{\gamma}_0)\right| \le k_1\cdot\mathbf{1}_{h\times h} \tag{2.37}$$

$$\mathbb{E}\left|\frac{\partial^2 \ln f(\varepsilon_{s,t}(\boldsymbol{\theta}_0))}{\partial\lambda\partial\boldsymbol{\gamma}'_j}\right| = \frac{k_1}{4}|\lambda_{j0}F'(x'_{s,t}\boldsymbol{\gamma}_{j0})|\cdot|\boldsymbol{F}(x'_{s,t}\boldsymbol{\gamma}_0)'x'_{s,t}| \le \frac{k_1|\lambda_{j0}|}{4}\cdot|\boldsymbol{F}(x'_{s,t}\boldsymbol{\gamma}_0)'x'_{s,t}| \tag{2.38}$$

$$\mathbb{E}\left|\frac{\partial^2 \ln f(\varepsilon_{s,t}(\boldsymbol{\theta}_0))}{\partial\boldsymbol{\gamma}_k\partial\boldsymbol{\gamma}'_j}\right| \le \frac{k_1|\lambda_{k0}\lambda_{j0}|}{16}\cdot|x_{s,t}x'_{s,t}|, \quad k\neq j \tag{2.39}$$

$$\mathbb{E}\left|\frac{\partial^2 \ln f(\varepsilon_{s,t}(\boldsymbol{\theta}_0))}{\partial\boldsymbol{\gamma}_j\partial\boldsymbol{\gamma}'_j}\right| \le \frac{k_1\lambda_{j0}^2}{16}\cdot|x_{s,t}x'_{s,t}| + \frac{\sqrt{3}k_0|\lambda_{j0}|}{18}|x_{s,t}x'_{s,t}| \tag{2.40}$$

Then we can apply the ergodic theorem [4] and conclude that

$$\left|\frac{1}{nT}\frac{\partial^2\mathcal{L}_{n,T}(\boldsymbol{\theta}_0)}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}'} - \mathbb{E}\frac{1}{nT}\frac{\partial^2\mathcal{L}_{n,T}(\boldsymbol{\theta}_0)}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}'}\right| \xrightarrow{p} 0 \tag{2.41}$$

Recall the equation (2.22), we have proved that $\frac{1}{\sqrt{nT}}\frac{\partial\mathcal{L}_{n,T}(\boldsymbol{\theta}_0)}{\partial\boldsymbol{\theta}}$ has the limiting distribution $N(\mathbf{0}, B(\boldsymbol{\theta}_0))$. With (2.41), for $\tilde{\boldsymbol{\theta}}_{n,T}$ between $\hat{\boldsymbol{\theta}}_{n,T}$ and $\boldsymbol{\theta}_0$, $-\frac{1}{nT}\frac{\partial^2\mathcal{L}_{n,T}(\tilde{\boldsymbol{\theta}}_{n,T})}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}'} \xrightarrow{p} A(\boldsymbol{\theta}_0)$ so we can conclude that $\sqrt{nT}(\hat{\boldsymbol{\theta}}_{n,T} - \boldsymbol{\theta}_0) \xrightarrow{d} N(\mathbf{0}, \boldsymbol{\Omega}_0)$, where $\boldsymbol{\Omega}_0 = A^{-1}(\boldsymbol{\theta}_0)B(\boldsymbol{\theta}_0)A^{-1}(\boldsymbol{\theta}_0)$. □

2.5. Numerical Results

2.5.1. Simulation Study

In this section, we conduct simulation experiments to examine the estimators' behavior for finite samples. We look at two PSTAR-ANN(1) models with one and two neurons with model parameters specified below:

$$Y_t = \phi_0 W_n Y_t + \phi_1 W_n Y_{t-1} + \boldsymbol{F}(X_t\boldsymbol{\gamma}')\lambda + \boldsymbol{\varepsilon}_t \tag{2.42}$$

$$\phi_0 = 0.6, \quad \phi_1 = -0.274, \quad \lambda = 1.5$$

$$\boldsymbol{\gamma} = (\gamma_1, \gamma_2)' = (0.75, -0.35)'$$

$$Y_t = \phi_0 W_n Y_t + \phi_1 W_n Y_{t-1} + X_t\beta + \boldsymbol{F}(X_t\boldsymbol{\gamma}_1')\lambda_1 + \boldsymbol{F}(X_t\boldsymbol{\gamma}_2')\lambda_2 + \boldsymbol{\varepsilon}_t \tag{2.43}$$

$$\phi_0 = 0.6, \quad \phi_1 = -0.274, \quad \beta = (0.24, -0.7)'$$

$$\lambda_1 = 2, \quad \boldsymbol{\gamma}_1 = (\gamma_{11}, \gamma_{12})' = (0.75, -0.35)'$$

$$\lambda_2 = 0.8, \quad \boldsymbol{\gamma}_2 = (\gamma_{21}, \gamma_{22})' = (0.35, -0.5)'$$

Simulations are conducted in a 30 by 30 lattice grid, so $n = 900$ and $p = 1$, $T = 30$. Random errors are sampled respectively from three distributions (standard normal, rescaled t-distribution and Laplace distribution) with variance 1. We generated data for two exogenous variables, observed at different time points t and location s.

Let

$$X_t = \begin{pmatrix} x_{11,t} & \dots & x_{1n,t} \\ \\ x_{21,t} & \dots & x_{2n,t} \end{pmatrix}'$$

Usually we would like to normalize predictors before fitting a neural network model to avoid the computation overflow [32] so values of $x_{s,t}[i], i = 1, 2$, were generated independently from normal distributions $N(0, 1.5^2)$ and $N(0, 3^2)$ respectively. The log-likelihood function $\mathcal{L}_{n,T}(\boldsymbol{\theta})$ is given in (2.44) and we use L-BFGS-B method [12, 52] (recommended for bound constrained optimization) to find the parameter estimates $\hat{\boldsymbol{\theta}}$ which maximize (2.44).

$$\mathcal{L}_{n,T}(\boldsymbol{\theta}) = T \ln |I_n - \phi_0 W_n| + \sum_{t=1}^{T} \sum_{s=1}^{n} \ln f(\varepsilon_{s,t}(\boldsymbol{\theta})) \tag{2.44}$$

for model (2.42): $\varepsilon_{s,t}(\boldsymbol{\theta}) = y_{s,t} - \sum_{i=0}^{p} \sum_{k=1}^{n} \phi_i w_{sk} y_{k,t-i} - \boldsymbol{F}(x'_{s,t}\boldsymbol{\gamma})\lambda$

for model (2.43): $\varepsilon_{s,t}(\boldsymbol{\theta}) = y_{s,t} - \sum_{i=0}^{p} \sum_{k=1}^{n} \phi_i w_{sk} y_{k,t-i} - \boldsymbol{F}(x'_{s,t}\boldsymbol{\gamma}_1)\lambda_1 - \boldsymbol{F}(x'_{s,t}\boldsymbol{\gamma}_2)\lambda_2$

For the models under consideration, we estimated the covariance of the asymptotic normal distribution equation (2.21). Since matrices $A(\boldsymbol{\theta}_0)$ and $B(\boldsymbol{\theta}_0)$ involve expected values with respect to the true parameter $\boldsymbol{\theta}_0$, given merely observations, in practice they can be estimated as follows:

$$\hat{A}(\boldsymbol{\theta}_0) = \frac{1}{nT} \sum_{t=1}^{T} \sum_{s=1}^{n} -\frac{\partial^2 l_{s,t}(\boldsymbol{\theta}_0)}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}'}$$

$$\hat{B}(\boldsymbol{\theta}_0) = \frac{1}{nT} \sum_{t=1}^{T} \sum_{s=1}^{n} \frac{\partial l_{s,t}(\boldsymbol{\theta}_0)}{\partial \boldsymbol{\theta}} \frac{\partial l_{s,t}(\boldsymbol{\theta}_0)}{\partial \boldsymbol{\theta}'}$$

where

$$l_{s,t}(\boldsymbol{\theta}) = \frac{1}{n} \ln |I_n - \phi_0 W_n| + \ln f(\varepsilon_{s,t}(\boldsymbol{\theta}))$$

Using (2.23) and (2.24), we can calculate $\hat{A}(\boldsymbol{\theta}_0), \hat{B}(\boldsymbol{\theta}_0)$ to assess the asymptotic properties of parameter estimates. Note that the derivative of the log-likelihood with respect to ϕ_0 cannot be calculated directly because it requires taking derivative with respect to a log-determinant of $I_n - \phi_0 W_n$. For small sample sizes, we can compute the determinant directly and get the corresponding derivatives; but for large sample sizes, for example a dataset with $n = 900$ observations, W_n is a 900×900 weight matrix which makes it impossible to calculate the derivative directly. Since W_n is a square matrix, we can apply the spectral decomposition such that W_n can be expressed in terms of its n eigenvalue-eigenvector pairs in (2.7). So we can apply the following approach to calculate the derivative of $\ln|I_n - \phi_0 W_n|$, which greatly reduces the burden of computations (Viton [46]).

$$\ln|I_n - \phi_0 W_n| = \ln\left(\prod_{s=1}^{n}(1 - \phi_0\tau_i)\right)$$

Further the derivatives of the log-likelihood function with respect to ϕ_0 is

$$\frac{\partial l_{s,t}(\boldsymbol{\theta})}{\partial \phi_0} = \frac{1}{n}\sum_{i=1}^{n}\frac{-\tau_i}{(1-\rho\tau_i)} + \{y_{s,t} - \sum_{i=0}^{p}\phi_i\sum_{j=1}^{n}w_{sj}y_{j,t-i} - \lambda F(x'_{s,t}\boldsymbol{\gamma})\}\cdot\left(\sum_{j=1}^{n}w_{sj}y_{j,t}\right)$$

$$\frac{\partial^2 l_{s,t}(\boldsymbol{\theta})}{\partial \phi_0 \partial \phi_0} = -\frac{1}{n}\sum_{i=1}^{n}\left[\frac{\tau_i^2}{(1-\phi_0\tau_i)^2} + \left(\sum_{j=1}^{n}w_{sj}y_{j,t}\right)^2\right]$$

Finally we can estimate the covariance matrix by equation (2.45).

$$\hat{\boldsymbol{\Omega}} = \hat{A}^{-1}(\boldsymbol{\theta}_0)\hat{B}(\boldsymbol{\theta}_0)\hat{A}^{-1}(\boldsymbol{\theta}_0) \tag{2.45}$$

In each simulation study, we compute $\hat{\boldsymbol{\theta}}$ for each of 200 replicates. The estimated $\hat{\boldsymbol{\Omega}}$ of the asymptotic covariance matrix $\hat{\boldsymbol{\Omega}}$ is computed based on a sample with $n = 10000, T = 100$ simulated observations. Table 2.1 and 2.2 compare the empirical mean and standard errors (in parentheses) of $\hat{\boldsymbol{\theta}}$ with the true value and their estimated asymptotic standard deviations. From simulation results of the two models, the empirical standard deviations

of $\hat{\boldsymbol{\theta}}$ are close to the asymptotic standard deviations, which implies that the estimators' large finite sample behavior roughly matches their asymptotic distributions. Note that when ε_t is sampled from a Laplace distribution, this covariance matrix cannot be computed because its second order derivative is not differentiable at 0. But the simulated $\hat{\boldsymbol{\theta}}$'s still exhibit normal properties. Normal plots for parameter estimates are shown in Figure 2.3 and give a strong indication of normality.

Model 1: $Y_t = \phi_0 W_n Y_t + \phi_1 W_n Y_{t-1} + \boldsymbol{F}(X_t\boldsymbol{\gamma}')\lambda + \varepsilon_t$					
ε_t	$\hat{\phi}_0$	$\hat{\phi}_1$	$\hat{\lambda}$	$\hat{\gamma}_1$	$\hat{\gamma}_2$
true value	0.6	-0.274	1.50	0.75	-0.35
$N(0,1)$	0.5997 (0.0065) [0.0079]	-0.2743 (0.0079) [0.0085]	1.5025 (0.0274) [0.0308]	0.7485 (0.0269) [0.0310]	-0.3476 (0.0134) [0.0147]
$t(4)$	0.5994 (0.0059) [0.0068]	-0.2737 (0.0069) [0.0071]	1.5000 (0.0236) [0.0259]	0.7531 (0.0249) [0.0258]	-0.3507 (0.0112) [0.0122]
Laplace $(0, \frac{\sqrt{2}}{2})$	0.5999 (0.0048)	-0.2736 (0.0058)	1.4992 (0.0199)	0.7501 (0.0196)	-0.3504 (0.0097)

Table 2.1. Empirical means and standard errors (in parentheses) of parameter estimates when ε is sampled from a standard normal, standardized student t distribution and a Laplace distribution. The asymptotic standard errors are displayed for reference in square brackets.

2.5.2. Real Data Example

Spatial models have a lot of applications in understanding spatial interactions in cross-sectional data. In our first chapter we applied a partially specified spatial autoregressive model to understand the relationships between vote choices and social factors. In this

Model 2: $Y_t = \phi_0 W_n Y_t + \phi_1 W_n Y_{t-1} + X_t\beta + \boldsymbol{F}(X_t\boldsymbol{\gamma}_1')\lambda_1 + \boldsymbol{F}(X_t\boldsymbol{\gamma}_2')\lambda_2 + \varepsilon_t$					
$\boldsymbol{\varepsilon}_t$	$\hat{\phi}_0$	$\hat{\phi}_1$	$\hat{\beta}$		$\hat{\lambda}_1$
	0.6	-0.274	0.24	-0.70	2
$N(0,1)$	0.6000 (0.0039) [0.0040]	-0.2748 (0.0046) [0.0044]	0.2402 (0.0137) [0.0135]	-0.6985 (0.0140) [0.0141]	1.9927 (0.0928) [0.0921]
$t(4)$	0.5999 (0.0036) [0.0035]	-0.2740 (0.0034) [0.0036]	0.2402 (0.0130) [0.0116]	-0.7005 (0.0106) [0.0113]	2.0008 (0.0727) [0.0759]
Laplace $(0, \frac{\sqrt{2}}{2})$	0.6006 (0.0030)	-0.2743 (0.0030)	0.2408 (0.0100)	-0.6997 (0.00995)	1.9983 (0.0638)
$\boldsymbol{\varepsilon}_t$	$\hat{\gamma}_1$		$\hat{\lambda}_2$	$\hat{\gamma}_2$	
	0.75	0.7	0.8	0.35	-1
$N(0,1)$	0.7503 (0.0962) [0.0920]	0.7030 (0.0369) [0.0390]	0.8076 (0.0450) [0.0449]	0.3577 (0.0899) [0.0835]	-1.0159 (0.1209) [0.1243]
$t(4)$	0.7496 (0.0714) [0.0324]	0.7016 (0.0332) [0.0371]	0.7989 (0.0392) [0.0758]	0.3521 (0.0749) [0.0697]	-1.0078 (0.0972) [0.1026]
Laplace $(0, \frac{\sqrt{2}}{2})$	0.7509 (0.0624)	0.7026 (0.0256)	0.8034 (0.0293)	0.3477 (0.0621)	-1.0089 (0.0873)

Table 2.2. Empirical means and standard errors (in parentheses) of parameter estimates when ε is sampled from a standard normal, standardized student t distribution and a Laplace distribution. The asymptotic standard errors are displayed for reference in square brackets.

chapter, we want to use a partially specified space time autoregressive model to further analyze the time influence in the electoral dynamics.

We focus on the proportion of votes cast for U.S. presidential candidates at the county level in 2004. Counties are grouped by state, and let Y_t, Y_{t-1} (so $t = 1, 2$, i.e., observe Y_1 and Y_2) be the corresponding fraction of votes (vote-share) in a county for the Democratic

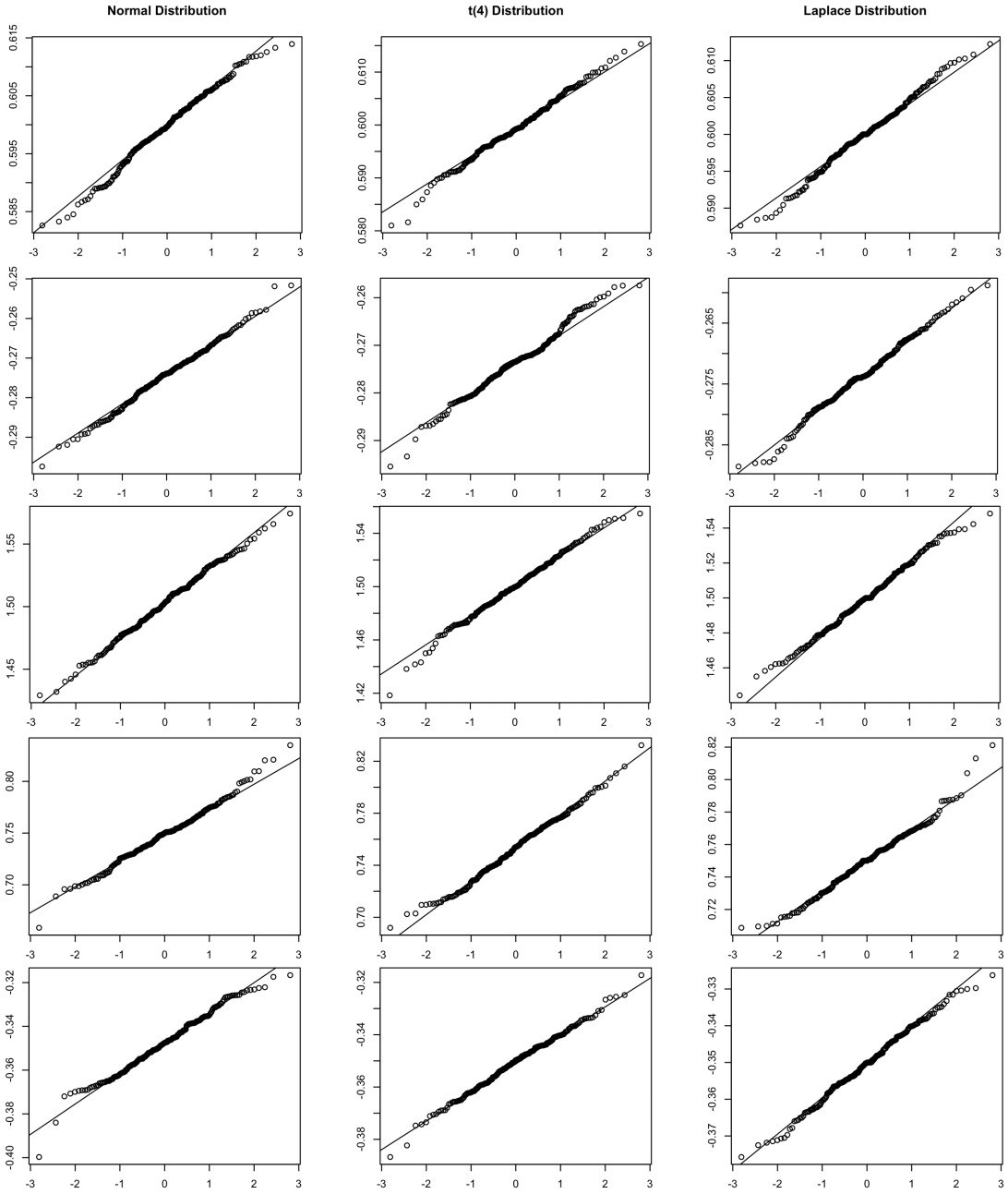


Figure 2.3. Normal plots for parameter estimates ϕ_0 (1st row), ϕ_1 (2nd row), λ (3rd row) and γ_1 (4th row), γ_2 (5th row) of model (2.42) when ε_s follows a standard normal distribution (first column), standardized t distribution (middle column) and Laplace distribution (last column) $n = 30\times30, T = 30$

candidate in 2004 and 2000. Predictors X_t are chosen from economic and social factors covering the living standard, economy development and racial distribution. Figure 2.4

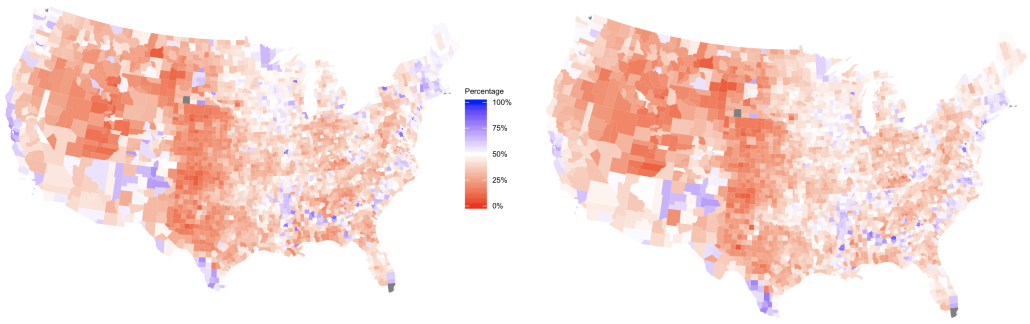


Figure 2.4. Fractions of vote-shares per county for Democratic presidential candidate in 2004 (left) and 2000 (right)

shows the observed values of Y_2 for 2004 and Y_1 for 2000 in a US map. Despite the strong spatial correlation (by Moran's Test on Y_t test statistic $= 52.4$, P-value $< 2.2 \times 10^{-16}$), these heat maps also exhibit the correlation across time since the two heat maps look rather similar. This indicates that Y_t, the fraction of vote-share for Democratic candidate, is not independently distributed across the space or time. Therefore we consider fitting a space time model to the data.

In our analysis, we exclude the four U.S. counties with no neighbors (San Juan, Dukes, Nantucket, Richmond) to avoid the non-singularity of our spatial weight matrix W_n in the modeling, so the total number of observations is $n = 3107$. Continuing our analysis in the first chapter, the selected explanatory variables are percent residents under 18 years in 2004 $X_{1,t}$ (`UNDER18`), percent white residents in 2004 $X_{2,t}$ (`WHITE`), percent residents below poverty line in 2004 $X_{3,t}$ (`pctpoor`).

We also assume the random error follows a scaled $t(8)$ distribution and, similar to previous chapter, perform variable transformations as follows:

$$Y_t^* = Y_t/8$$

$$Y_{t-1}^* = Y_{t-1}/8$$

$$\tilde{X}_{1,t} = (I_{3107} - 0.6W_{3107})X_{1,t}$$

$$X_{1,t}^* = \frac{\tilde{X}_{1,t} - Average(\tilde{X}_{1,t})}{Std(\tilde{X}_{1,t})}$$

$$X_{2,t}^* = \frac{X_{2,t} - Average(X_{2,t})}{Std(X_{2,t})}$$

$$X_{3,t}^* = \frac{X_{3,t} - Average(X_{3,t})}{Std(X_{3,t})}$$

Figure 2.5 illustrates histograms of Y_t^* (first row) and histograms of exogenous variables $X_{1,t}^*, X_{2,t}^*, X_{3,t}^*$ when $t = 1, 2, 3$ ($t = 3$ represents the year 2008) respectively. Comparing their histograms at different years, we can observe that the distributions look similar so we may consider X_t and Y_t as stationary processes across time.

The estimated PSAR-ANN model in chapter 1 is:

$$\begin{aligned} Y_t^* =& 0.721W_{3107}Y_t^* + 1.693 - 0.185X_{1,t}^* - 0.658X_{2,t}^* + 0.181X_{3,t}^* \\ & - 0.937F(1.509X_{1,t}^* - 2.544X_{2,t}^* + 2.268X_{3,t}^*) + \hat{\varepsilon}_t \end{aligned} \tag{2.46}$$

In this chapter we would like to add time into the model and we fit two PSTAR-ANN(1) models with one and two neurons respectively. Similarly we find the parameter estimates by maximizing the corresponding log-likelihood functions and use the L-BFGS-B algorithm to search for the optimum. Detailed optimization steps are similar to those in chapter 1. The model fits are shown below. One is the PSTAR-ANN(1) with one

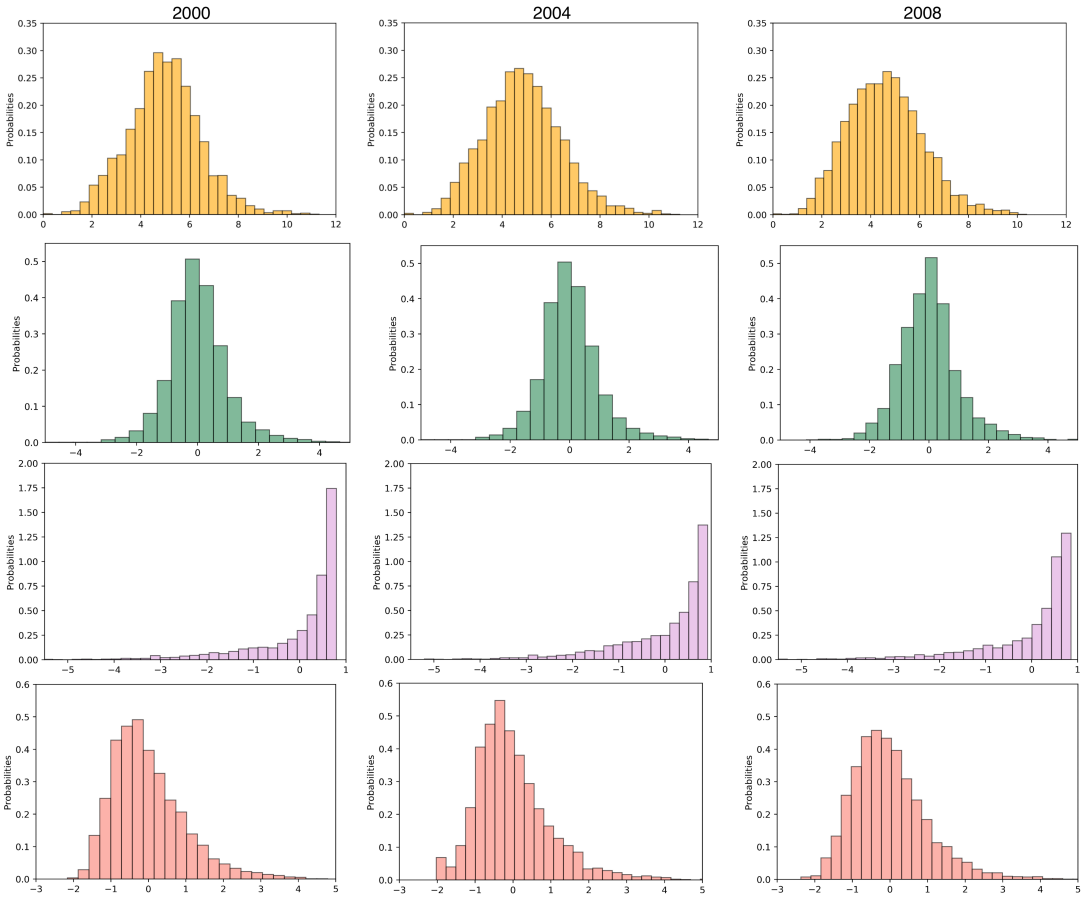


Figure 2.5. Histograms of Y_t^* (1st row), $X_{1,t}^*$ (2nd row), $X_{3,t}^*$ (3rd row) and $X_{3,t}^*$ (4th row) for $t = 1, 2, 3$ corresponding to the year 2000 (left), 2004 (middle) and 2008 (right)

neuron:

$$Y_t^* = 0.425W_{3107}Y_t^* + 0.464W_{3107}Y_{t-1}^* - 1.173 + 0.148X_{1,t}^* - 1.177X_{2,t}^* - 0.153X_{3,t}^*$$

$$+ 3.056F(-0.722X_{1,t}^* + 1.689X_{2,t}^* + 0.248X_{3,t}^*) + \hat{\varepsilon}_t \tag{2.47}$$

Another is the PSTAR-ANN(1) with two neurons:

$$
\begin{aligned}
Y_t^* &= 0.417W_{3107}Y_t^* + 0.467W_{3107}Y_{t-1}^* - 1.576 + 0.203X_{1,t}^* - 1.222X_{2,t}^* - 0.057X_{3,t}^* \\
&+ 0.699F(-1.249X_{1,t}^* + 0.084X_{2,t}^* - 3.247X_{3,t}^*) \\
&+ 3.180F(-0.621X_{1,t}^* + 1.621X_{2,t}^* + 0.495X_{3,t}^*) + \hat{\varepsilon}_t
\end{aligned} \tag{2.48}
$$

Comparing the three models (2.46), (2.47) and (2.48), the coefficients estimates are all positive so it is apparent that there exist a positive space correlation, between $y_{s,t}$ and its neighbors, and also a positive time correlation between Y_t and Y_{t-1}. The P-values of Moran's test statistic of PSTAR-ANN(1) model residuals (residuals of model (2.47) and (2.48)) are higher than that of model (2.46), which indicates that PSTAR-ANN(1) models are able to describe more spatial correlations than the PSAR-ANN model. For

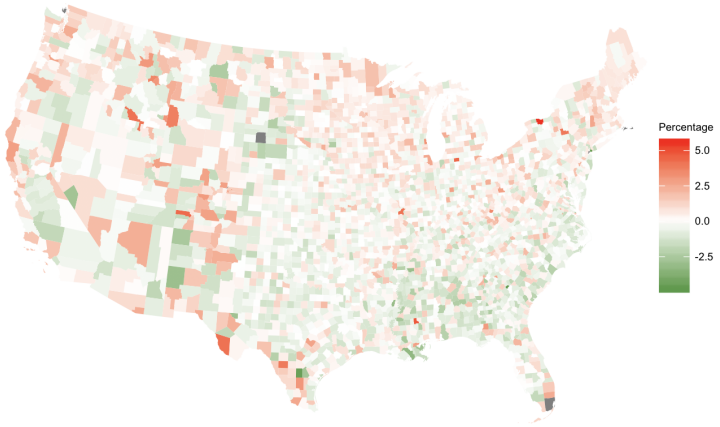


Figure 2.6. Residuals heat map (calculated from the PSTAR-ANN model with one neuron)

the preliminary comparison purpose, we compare the AICs (AIC = 2#parameters $-$ $2\ln L_{n,T}(\hat{\boldsymbol{\theta}})$) of the three models (See table 2.3). For likelihood ratio test ($\mathcal{H}_0$: Model (2.46) is adequate, $\mathcal{H}_1$: model (2.47) is adequate), the test statistic $-2\ln L_{\text{Model (2.46)}} +$

$2 \ln L_{\text{Model (2.47)}} = 287.17$ with $df = 6$, P-value < 0.05, so we rejected $\mathcal{H}_0$ and conclude that the PSTAR-ANN(1) model with one neuron is a better fit. Similarly we apply the same method to compare the two PSTAR-ANN(1) models and conclude that the model with two neurons is better (the test statistic $-2 \ln L_{\text{Model (2.47)}} + 2 \ln L_{\text{Model (2.48)}} = 40.33$ with $df = 4$, P-value < 0.05). The covariance matrices for the parameter estimates of

Models	PSAR-ANN (one neuron)	PSTAR-ANN (one neuron)	PSTAR-ANN (two neurons)
# Parameters	9	10	14
Moran's Test	0.0745 (1.7836)	0.2336 (−1.1910)	0.3368 (−0.9604)
$-\ln L$	1879.35	1734.5	1710.33
AIC	3776.17	3489	3448.669

Table 2.3. Model Comparisons: PSAR-ANN model with one neuron (2.46), PSTAR-ANN models with one (2.47) and two neurons (2.48)

model (2.47) and (2.48) are calculated and the 95% confidence intervals for the model parameters are shown in Tables 2.4 and 2.5. From Table 2.4, all the parameters, except $X_{3,t}$ (`pctpoor`), are significant at 0.05 significance level. Table 2.5 shows the 95% level of parameter estimates in model (2.48).

From Table 2.4 and 2.5, we can see that values of $y_{s,t}$ are positively spatially correlated in both space and time. Looking at the signs of parameter estimates of coefficients, we can see that the sign of variable `UNDER18` in model (2.46) is negative while positive in model (2.47) and (2.48). Considering its parameter estimate significant in all models, this indicates that age and vote-shares for Democratic candidates can be dependent but the percent residents under 18 may not be a good measurement for this social factor.

Parameter	Estimate	Std.	95% C.I.
ϕ_0	0.425	0.0086	$(0.4081, 0.4419)$
ϕ_1	0.464	0.0182	$(0.4283, 0.4997)$
β_0	-1.173	0.3283	$(-1.8165, -0.5295)$
β_1	0.148	0.0697	$(0.0114, 0.2846)$
β_2	-1.177	0.1638	$(-1.4980, -0.8560)$
β_3	-0.153	0.1079	$(-0.3645, 0.0585)^*$
λ	3.056	0.6397	$(1.8022, 4.3098)$
γ_1	-0.722	0.1278	$(-0.9725, -0.4715)$
γ_2	1.689	0.1762	$(1.3436, 2.0344)$
γ_3	0.248	0.1890	$(-0.1224, 0.6184)^*$

Table 2.4. Parameter estimates of PSTAR-ANN model (2.47) parameters with 95% confidence intervals ($*$ indicates the insignificance)

Parameter	Estimate	Std.	95% C.I.
ϕ_0	0.417	0.0086	$(0.4000, 0.4339)$
ϕ_1	0.467	0.0178	$(0.4321, 0.5019)$
β_0	-1.576	0.3063	$(-2.1764, -0.9756)$
β_1	0.203	0.0731	$(0.0598, 0.3462)$
β_2	-1.222	0.1507	$(-1.5174, -0.9266)$
β_3	-0.057	0.0926	$(-0.2385, 0.1245)^*$
λ_1	3.180	1.2624	$(0.7057, 5.6543)$
γ_{11}	-0.621	0.1193	$(-0.8548, -0.3872)$
γ_{12}	1.621	0.1521	$(1.3230, 1.9190)$
γ_{13}	0.495	0.1063	$(0.2866, 0.7034)$
λ_2	0.699	0.2397	$(0.2291, 1.1689)$
γ_{21}	-1.294	0.6060	$(-2.4368, -0.0612)$
γ_{22}	0.084	0.4859	$(-0.8683, 1.0363)^*$
γ_{23}	-3.247	0.3570	$(-3.9469, -2.5471)$

Table 2.5. Parameter estimates of PSTAR-ANN model (2.48) parameters with 95% confidence intervals ($*$ indicates the insignificance)

We should consider using other age related variables to predict Y_t such as the percent young voters between 18 and 30 years old. Variable `WHITE` is negatively correlated with

Y_t in all three fitted models and this negative correlation accords with our common sense that white voters tend to support the Republican candidate. The last variable `pctpoor` is bit tricky because it is not significant in model (2.47) but is significant in the neural network component in model (2.48). Regarding to this, it needs further assessment to decide if `pctpoor` should be included in the model. In chapter 3, we will further discuss the model selection in detail. To conclude, our proposed model PSTAR-ANN appears to successfully capture some presidential election dynamics over both space and time. It allows for non-Gaussian random errors and is flexible in learning nonlinear relationships between the response and exogenous variables.

CHAPTER 3

Model Selection in Partially Specified Space Time Autoregressive Model with Artificial Neural Network

In previous chapters, we discussed the asymptotic properties of maximum likelihood estimators for the parameters of a partially specified spatial and/or temporal autoregressive model with artificial neural network (equation (3.1)) under the assumption of stationarity. When $p = 0$ in equation (3.1), the PSTAR-ANN(p) model will reduce to a PSAR-ANN model. A PSTAR-ANN(p) model is given by

$$Y_t = \sum_{i=0}^{p} \phi_i W_n Y_{t-i} + X_t\beta + \boldsymbol{F}(X_t\boldsymbol{\gamma}')\lambda + \boldsymbol{\varepsilon}_t \tag{3.1}$$

A first problem is to select useful predictors X_t to describe Y_t. One aspect is to select exogenous variables in the linear component and some standard methods include step-wise selection, likelihood ratio test and AIC/BIC criteria. Another aspect is to select exogenous variables in the neural network component and to determine a proper size of the network so as to provide desired outcomes in test data. The complexity of artificial neural network regression makes it difficult to apply many existing analytical variable selection methods. Recent papers [31] [28] [21] [41] showed experimentally that the robustness of neural networks can be remarkably enhanced by adding random noise to the inputs (predictors) during the training. This idea is based on the principle [31] that a robust network should be as insensitive as possible to a reasonable variation of inputs. By injecting noise into inputs, we could add randomness into a training set when we only have limited number

of samples. Mathematical proofs [31] have shown that minimizing a loss function with noise injected in raw data is equivalent to adding a penalty in the loss function with the raw data. In this way noise injection realizes penalizing the number and magnitudes of parameters in a network, which mitigate the over-fitting in neural networks. In this paper, we discuss one of the noise injection methods, Shakeout (Kang and Li [21]), and elaborate its penalization effect in a PSTAR-ANN(p) model: selecting useful exogenous variables, controling the size of the neural network component and stabilizing the learned network.

Secondly we will talk about AR order selection in this spatial context. One way to estimate the time lag p is the likelihood ratio test [49] by which we can test the significance of parameter ϕ_p in model (3.1). Because the goal is to test if a higher time lag is significant, we apply the likelihood ratio test to the nested model selection and prove the asymptotic distribution of the test statistic. In practice, analogous to order selection in the time series, we can also look at the sample autocorrelations (ACF) and partial autocorrelations (PACF) to approximate p. The detailed calculation method will be discussed later.

Section 2 presents the methodology of Shakeout technique in selecting exogenous variables and we demonstrate its penalization effects in the linear and nonlinear component of a PSTAR-ANN(p) model. Section 3 presents autoregressive order selection using techinque sample ACF, sample PACF and the likelihood ratio test for nested model selection. In section 4, we show via simulation that the asymptotic distribution of the likelihood ratio test statistic in previous section is approximately valid for finite large samples and we illustrate the penalization effects of Shakeout by experiments. We also apply the likelihood ratio test to the real data example to find an appropriate autoregressive order in the

US presidential election results. Due to limited access to historical election data and other records of social variables, we do not perform the Shakeout method in the real example. Finally, section 6 discuss potential future development in PSTAR-ANN(p) models.

3.1. Feature Selection

As mentioned in the introduction, due to the neural network component in our model, one important aspect in feature selection is to select a proper size of the network and useful exogenous variables in a net because the redundant neurons in a network can cause non-identifiability and over-fitting problems. When we only have limited training data, the estimated network can be pretty unstable due to random sampling errors. To avoid this, many papers have proposed a noise injection method – injecting noises in the inputs of a network [31], [41], [21], [28]. It has been mathematically shown that the noise injection is essentially a regularization in neural networks such that it can control both the size of a network and magnitudes of weights [31], [21], [28]. "Regularization" is a technique of adding constraints to the model parameters, which attempts to learn a simpler model from the training data. A regularization is equivalent to adding a penalization or a regularizer in the loss function and the resulting estimators are called regularized estimators. For example, in ridge regression, we impose a restriction in the sum of squares of parameters (known as L_2 regularization) while in the lasso, we impose the restriction in the sum of absolute value of parameters (known as L_1 regularization). In this section, we are going to apply the Shakeout, first proposed by Kang et al. [21], to select inputs/predictors and to prune the network in a PSTAR-ANN(p) model. In the meanwhile, we will mathematically show the regularization effect of this technique.

3.1.1. Shakeout Regularization

We now illustrate how to inject "Shakeout" noise in the linear component and it is also the same in the neural network component. Suppose we would like to minimize some loss function, the intermediate output of one neuron is a weighted sum of inputs $x_1, \ldots, x_q$ with a set of starting values for weights denoted by $\beta_j, j = 1, \ldots, q$.

$$\sum_{j=1}^{q} \beta_j x_j$$

When optimizing the loss function, Shakeout [21] adjusts the weights β_j to $\tilde{\beta}_j$ in some way that we can control the variable x_j in or out of this neuron:

$$\text{Step 1: Draw } r_j \text{ from the distribution } \begin{cases} \Pr(r_j = 0) & = \tau \\ \Pr(r_j = \frac{1}{1-\tau}) & = 1 - \tau \end{cases}$$

Step 2: Adjust the weight according to r_j,

$$\begin{cases} \tilde{\beta}_j = -cs_j, & \text{if } r_j = 0 \quad (A) \\ \tilde{\beta}_j = (\beta_j + c\tau s_j)/(1-\tau), & \text{otherwise} \quad (B) \end{cases}$$

where $s_j = sgn(\beta_j)$ takes ± 1 depending on the sign of β_j or take 0 if $\beta_j = 0$, c is a positive constant, $\tau \in (0, 1)$. τ and c are hyper-parameters in Shakeout and r_j is randomly generated from a scaled Bernoulli with the parameter τ. As shown above, Shakeout proposes two different modifications on the weights β_j. (A) sets the weights to constant values with the opposite sign of the original weights. (B) updates the weights by a factor $\frac{1}{(1-\tau)}$ and a bias $c\tau s_j$. The Dropout method [41] is a special case of Shakeout when $c = 0$. One crucial improvement of Shakeout over Dropout is that Shakeout avoids setting the weights directly to zero and preserves the zero values of original weights. Hence, the expected value of updated weighted sum with respect to the noise r is unbiased

$\mathbb{E}_r(\sum_{j=1}^{q} \tilde{\beta}_j x_j) = \sum_{j=1}^{q} \beta_j x_j$. The hyper-parameters $\tau \in (0,1)$ and $c \in (0, +\infty)$ determine the structure of the penalization term (detailed expression will be shown later in the prove) and therefore decide the regularization effect invoked by Shakeout.

In our model, we can utilize this technique to select features X_t in the linear part and prune the neural network part separately. Recalling the model (3.1), parameters we can update using the Shakeout are β and $\lambda, \boldsymbol{\gamma}$. Intuitively we could also update the space time autoregressive parameters $\phi_i, i = 0, 1, \ldots, p$ but additional restrictions are required to guarantee that the model remains stationary after updating parameters using Shakeout. For now, we only consider selecting exogenous variables X_t as well as the size of the neural network $\sum_{k=1}^{h} \lambda_k \boldsymbol{F}(X_t \boldsymbol{\gamma}_k)$ and will discuss alternative methods for selecting the autocorrelation parameters, ϕ_i's, in the later section.

Figure 3.1 shows the weights update in the linear component in our model. The

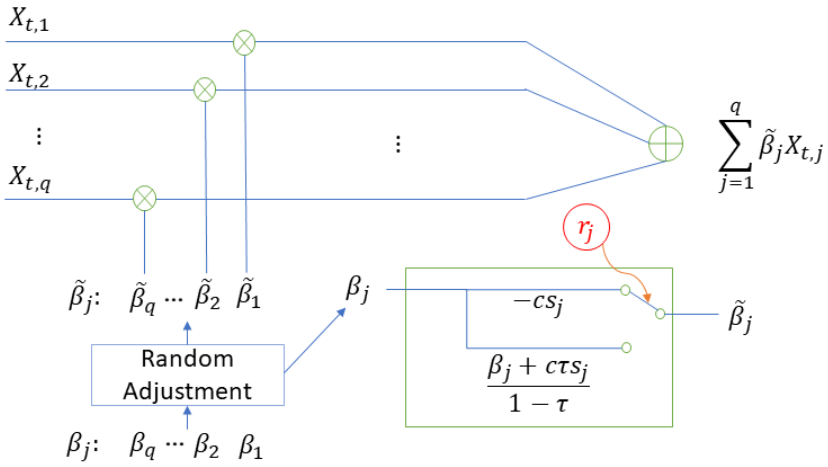


Figure 3.1. Shakeout regularization of the linear component $X_t\beta$ in a PSTAR-ANN model. Shakeout modifies the parameter β through random variable r_j. Define $P(r_j = 0) = \tau, P(r_j = \frac{1}{1-\tau}) = 1 - \tau$. When $r_j = 0$, β_j will be replaced by a constant $\tilde{\beta}_j = cs_j$, where $s_j = sgn(\beta_j)$; otherwise, the weight will be updated to $\tilde{\beta}_j = \beta_j + c\tau s_j$.

nonlinear part $\boldsymbol{F}(X_t\boldsymbol{\gamma}')\lambda$ can be regarded as a two layer neural network: $\boldsymbol{F}(X_t\boldsymbol{\gamma}')$ can be treated as the first layer where the inputs are X_t and weights $\boldsymbol{\gamma}_1, \ldots, \boldsymbol{\gamma}_h$. The activation function is the sigmoid function $\boldsymbol{F}$ so the outputs of this layer are $\boldsymbol{F}(X_t\boldsymbol{\gamma}_1), \ldots, \boldsymbol{F}(X_t\boldsymbol{\gamma}_h)$; the second layer receives the $\boldsymbol{F}(X_t\boldsymbol{\gamma}_1), \ldots, \boldsymbol{F}(X_t\boldsymbol{\gamma}_h)$ as its inputs and the weights are $\lambda_1, \ldots, \lambda_h$ respectively. The activation function of this layer is the identity function so the final output is $\sum_{i=1}^{h} \boldsymbol{F}(X_t\boldsymbol{\gamma}_i)\lambda_i$. Figure 3.2 shows the parameter updates of this two layer neural network.

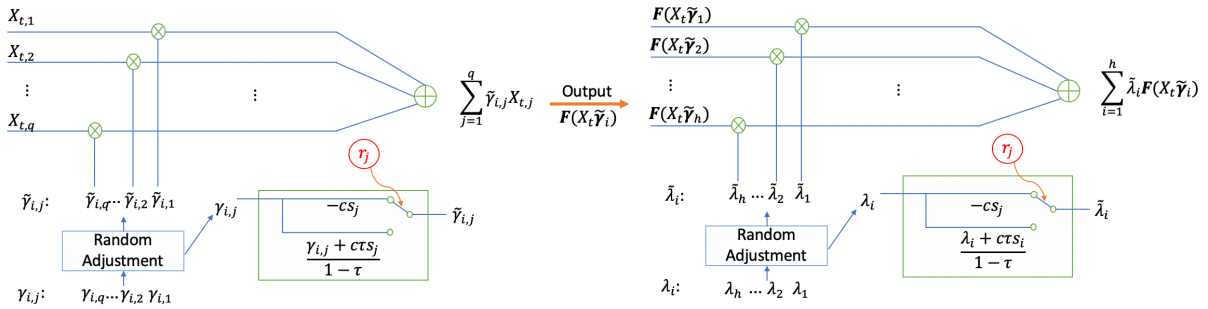


Figure 3.2. Shakeout regularization of nonlinear component in PSTAR-ANN model. (Left) the first layer i^{th} neuron: the input is X_t and the associated parameters in the kth neuron are $\boldsymbol{\gamma}_k = (\gamma_{k,1}, \ldots, \gamma_{k,p})'$. The activation function is the sigmoid function $\boldsymbol{F}$. (Right) the second layer with one neuron: the input is $\boldsymbol{F}(X_t\boldsymbol{\gamma}_1), \ldots, \boldsymbol{F}(X_t\boldsymbol{\gamma}_h)$ and weights are $\lambda_1, \ldots, \lambda_h$ respectively. The activation function is the identity function.

Shakeout, inherently, injects multiplicative noises to the observations or inputs in neurons. For example, for a input $x \in \mathbb{R}^q$ in the kth neuron with associated weight $\gamma_k \in \mathbb{R}^q$, Shakeout randomly updates γ_{kj}, the jth component in γ_k, $j = 1, \ldots, q$. This is equivalent to scale the jth component of x as $\tilde{x}_j$ by $r_j + \frac{c(r_j - 1)}{|\gamma_{kj}|}$ where r_j is randomly draw from $\{0, \frac{1}{1-\tau}\}$ with probability $\tau, 1-\tau$ respectively. One thing to notice is that r is repeatedly and randomly chosen for every component in $x_{s,t}$ for all s, t. In this way, we can guarantee the randomness in all the injected noise. By calculation, it is easy to show

$\mathbb{E}_{r_j}(\tilde{x}_j) = x_j$. We can rewrite this multiplicative noise as an additive noise:

$$\text{In the } k^{th} \text{ neuron, } j^{th} \text{ component of input variable } x,$$

$$\tilde{x}_j = x_j e_j, \text{where } e_j = \begin{cases} \frac{-c}{|\gamma_{kj}|} & \Pr = \tau \\ \frac{1}{1-\tau}\left(1 + \frac{c\tau}{|\gamma_{kj}|}\right) & \Pr = 1-\tau \end{cases}$$

$$= x_j + e_j, \text{where } e_j = \begin{cases} -x_j\left(1 + \frac{c}{|\gamma_{kj}|}\right) & \Pr = \tau \\ x_j\left(\frac{\tau}{1-\tau} + \frac{c\tau}{(1-\tau)|\gamma_{kj}|}\right) & \Pr = 1-\tau \end{cases}$$

This reformulated additive noise depends on the associated variable x_j, which implies that extreme noise values may be generated if $|x_j|$ is large [28]. In return, this extreme value in noise can lead to harsher penalty on parameters connected to large variables and therefore improve the robustness in model fitting (details will be discussed later).

3.1.2. Regularization Effects

Using the Shakeout noise injection, we define the loss function of a PSTAR-ANN(p) model as the negative of its log-likelihood function given $X = \{X_t\}_{t=1}^T$, $Y = \{Y_t\}_{t=1}^T$:

$$loss(\boldsymbol{\theta}) = -T \ln |I_n - \phi_0 W_n| - \sum_{t=1}^{T}\sum_{s=1}^{n} \ln f(\varepsilon_{s,t}(\boldsymbol{\theta})) \tag{3.2}$$

where $\varepsilon_{s,t}(\boldsymbol{\theta}) = y_{s,t} - \sum_{i=1}^{n} \phi_0 w_{si} y_{i,t} - \sum_{i=1}^{p} \phi_i \sum_{j=1}^{n} w_{sj} y_{j,t-i} - x'_{s,t}\beta - \sum_{k=1}^{h} \lambda_k F(x'_{s,t}\boldsymbol{\gamma}_k)$. Considering our model formulation, we justify Shakeout as a noise injection regularization technique from two aspects. On the one hand, in the linear component, τ and c controls the structure of penalty term about model parameters; for example, different choices of τ and c can lead to different combination of L_0, L_1, L_2 penalizations (this will be shown later). On the other hand, in the network component, adding random noise to the inputs can enhance robustness of the learned neural network [21]. In the following, we are going

to demonstrate this mathematically. We define a Hadamard product denoted by $\circ$, s.t. for vectors $a, b_1, \ldots, b_n \in \mathbb{R}^n$, a matrix $B = (b_1, \ldots, b_n) \in \mathbb{R}^{n \times n}$,

$$a \circ B = \begin{bmatrix} a_1 b_{11} & a_1 b_{21} & \cdots & a_1 b_{n1} \\ a_2 b_{12} & a_2 b_{22} & \cdots & a_2 b_{n2} \\ \vdots & \vdots & \ddots & \vdots \\ a_n b_{1n} & a_n b_{2n} & \cdots & a_n b_{nn} \end{bmatrix}, a \circ b_1 = \begin{bmatrix} a_1 b_{11} \\ a_2 b_{12} \\ \vdots \\ a_n b_{1n} \end{bmatrix}$$

Using the expression of multiplicative noises, we denote $(\boldsymbol{u}, \boldsymbol{v})$ as multiplicative noise in linear and NN part respectively. Specifically, in linear part, $\boldsymbol{u} = \{u_{s,t}\}$ where $s = 1, \ldots, n, t = 1, \ldots, T$. For j^{th} component of an observation at location s and time t, denoted as $x_{s,t}[j]$, the multiplicative noise $u_{s,t}[j] = \{\frac{-c}{|\beta_j|}, \frac{1}{1-\tau}(1 + \frac{c\tau}{|\beta_j|})\}$ with probability τ and $1 - \tau$ respectively. $\tilde{x}_{s,t}$ denotes the observation after injecting noises in the linear part.

(3.3)
$$\begin{array}{ccccccc} & \text{Input} & & \text{Shakeout Noise} & & \text{Updated Input} & \text{Output} \\ x_{s,t} = \begin{bmatrix} x_{s,t}[1] \\ x_{s,t}[2] \\ \vdots \\ x_{s,t}[q] \end{bmatrix}_q & u_{s,t} = \begin{bmatrix} u_{s,t}[1] \\ u_{s,t}[2] \\ \vdots \\ u_{s,t}[q] \end{bmatrix}_q & \xrightarrow{\text{Shakeout}} & \tilde{x}_{s,t} = \begin{bmatrix} x_{s,t}[1]u_{s,t}[1] \\ x_{s,t}[2]u_{s,t}[2] \\ \vdots \\ x_{s,t}[q]u_{s,t}[q] \end{bmatrix}_q & \xrightarrow{\text{linear}} & \tilde{x}_{s,t}\beta \end{array}$$

While in the NN part, we have two layers so $\boldsymbol{v} = \{v_{s,t}\} = \{v_{s,t}^{(1)}, v_{s,t}^{(2)}\}$ denotes the noise in the first and second layer in the network. In the first layer, $v_{s,t}^{(1)} = \{v_{s,t}^{(1,1)}, v_{s,t}^{(1,2)}, \ldots, v_{s,t}^{(1,h)}\}$ represents a collection of noises in all h neurons for every input $x_{s,t}$. So in the k^{th} neuron $v_{s,t}^{(1,k)}[j] = \{\frac{-c}{|\gamma_{kj}|}, \frac{1}{1-\tau} + \frac{c\tau}{|\gamma_{kj}|}\}$ for $j = 1, \ldots, q$. $\hat{x}_{s,t}^{(k)}$ denotes the observation after injecting

noises in the k^{th} neuron and the noise is independent among all neurons in this layer.

$$\begin{matrix} & \text{Input} & & \text{Shakeout Noise} & & & \text{Updated Input} & & \text{Output} \\ x_{s,t} = \begin{bmatrix} x_{s,t}[1] \\ x_{s,t}[2] \\ \vdots \\ x_{s,t}[q] \end{bmatrix}_q & v_{s,t}^{(1,k)} = \begin{bmatrix} v_{s,t}^{(1,k)}[1] \\ v_{s,t}^{(1,k)}[2] \\ \vdots \\ v_{s,t}^{(1,k)}[q] \end{bmatrix}_q & \xrightarrow{\text{Shakeout}} & \hat{x}_{s,t}^{(k)} = \begin{bmatrix} x_{s,t}[1]v_{s,t}^{(1,k)}[1] \\ x_{s,t}[2]v_{s,t}^{(1,k)}[2] \\ \vdots \\ x_{s,t}[q]v_{s,t}^{(1,k)}[q] \end{bmatrix}_q & \xrightarrow[\text{function}]{\text{Sigmoid}} & F(\hat{x}_{s,t}^{(k)'}\boldsymbol{\gamma}_k) \end{matrix} \tag{3.4}$$

In the second layer, the inputs are the outputs from the first layer $\boldsymbol{F}(x'_{s,t}\boldsymbol{\gamma})$ so $v_{s,t}^{(2)}[k] = \{\frac{-c}{|\lambda_k|}, \frac{1}{1-\tau} + \frac{c\tau}{|\lambda_k|}\}$, with probability τ and $1-\tau$. $\tilde{\boldsymbol{F}}(\hat{x}'_{s,t}\boldsymbol{\gamma})$ denotes the observation after injecting noises in the second layer.

$$\begin{matrix} & \text{Input} & & \text{Shakeout Noise} & & & \text{Updated Input} & & \text{Output} \\ \boldsymbol{F}(\hat{x}'_{s,t}\boldsymbol{\gamma}) = \begin{bmatrix} F(\hat{x}_{s,t}^{(1)'}\boldsymbol{\gamma}_1) \\ F(\hat{x}_{s,t}^{(2)'}\boldsymbol{\gamma}_2) \\ \vdots \\ F(\hat{x}_{s,t}^{(h)'}\boldsymbol{\gamma}_h) \end{bmatrix}_h & v_{s,t}^{(2)} = \begin{bmatrix} v_{s,t}^{(2)}[1] \\ v_{s,t}^{(2)}[2] \\ \vdots \\ v_{s,t}^{(2)}[h] \end{bmatrix}_h & \xrightarrow{\text{Shakeout}} & \tilde{\boldsymbol{F}}(\hat{x}'_{s,t}\boldsymbol{\gamma}) = \begin{bmatrix} F(\hat{x}_{s,t}^{(1)'}\boldsymbol{\gamma}_1)v_{s,t}^{(2)}[1] \\ F(\hat{x}_{s,t}^{(2)'}\boldsymbol{\gamma}_2)v_{s,t}^{(2)}[2] \\ \vdots \\ F(\hat{x}_{s,t}^{(h)'}\boldsymbol{\gamma}_h)v_{s,t}^{(2)}[h] \end{bmatrix}_h & \xrightarrow[\text{function}]{\text{linear}} & \tilde{\boldsymbol{F}}(\hat{x}'_{s,t}\boldsymbol{\gamma})\lambda \end{matrix} \tag{3.5}$$

Now we can rewrite the loss function with Shakeout noise injected as

$$loss(\boldsymbol{\theta}|\boldsymbol{u},\boldsymbol{v}) = -T\ln|I_n - \phi_0 W_n| - \sum_{t=1}^{T}\sum_{s=1}^{n}\ln f(\varepsilon_{s,t}(\boldsymbol{\theta}|\boldsymbol{u},\boldsymbol{v})) \tag{3.6}$$

$$\varepsilon_{s,t}(\boldsymbol{\theta}|\boldsymbol{u},\boldsymbol{v}) = y_{s,t} - \sum_{j=1}^{n}\phi_0 w_{sj}y_{j,t} - \sum_{i=1}^{p}\phi_i\sum_{j=1}^{n}w_{sj}y_{j,t-i} - \tilde{x}'_{s,t}\beta - \sum_{k=1}^{h}\lambda_k\tilde{F}(\hat{x}_{s,t}^{(k)'}\boldsymbol{\gamma}_k) \tag{3.7}$$

where $\tilde{x}_{s,t}$, $\hat{x}_{s,t}$, $\tilde{F}(\cdot)$ are defined in equations (3.3), (3.4) and (3.5). In the following, the expressions are all conditioned on X, Y so we omit X, Y in the conditional notation for simplicity.

In the following, we are going to demonstrate the Shakeout regularization effects in linear and nonlinear components in a PSTAR-ANN(p) model. These theorems give approximate results. Similar work can be seen in Matsuoka [31], Li and Liu [28], Kang et al. [21].

Define $loss(\boldsymbol{\theta}|\boldsymbol{u})$ as the loss function with Shakeout noise injected in the linear part. Theorem 1 below establishes that the expected value of $loss(\boldsymbol{\theta}|\boldsymbol{u})$ over the distribution of injected noise $\boldsymbol{u}$ is a roughly penalized loss function (3.6) with the original data, more specifically, the loss function with the original data plus a combination of L_0, L_1 and L_2 regularization terms, $\pi(\beta)$.

Theorem 5 (Regularization on β in a PSTAR-ANN model with Shakeout). The expectation of Equation (3.6) over the distribution of noise $\boldsymbol{u}$ is

$$\mathbb{E}_{\boldsymbol{u}}[loss(\boldsymbol{\theta}|\boldsymbol{u})] = loss(\boldsymbol{\theta}) + \pi(\boldsymbol{\theta}) \tag{3.8}$$

where

$$\begin{aligned}\pi(\boldsymbol{\theta}) &= -\sum_{t=1}^{T}\sum_{s=1}^{n}\mathbb{E}_{\boldsymbol{u}}[\ln f(\varepsilon_{s,t}(\boldsymbol{\theta}|\boldsymbol{u})) - \ln f(\varepsilon_{s,t}(\boldsymbol{\theta}))] \\ &\approx \frac{\tau}{2(1-\tau)}\sum_{t=1}^{T}\sum_{s=1}^{n}\left(\frac{f'^2(\varepsilon_{s,t}(\boldsymbol{\theta})}{f^2(\varepsilon_{s,t}(\boldsymbol{\theta}))} - \frac{f''(\varepsilon_{s,t}(\boldsymbol{\theta})}{f(\varepsilon_{s,t}(\boldsymbol{\theta}))}\right)||x_{s,t}\circ(\beta + \frac{c\beta}{|\beta|})||_2^2 \end{aligned} \tag{3.9}$$

where $||x_{s,t}\circ(\beta+\frac{c\beta}{|\beta|})||_2^2 = \sum_{j=1}^{q}(x_{s,t}[j]\beta_j)^2 + 2c\sum_{j=1}^{q}x_{s,t}^2[j]|\beta_j| + c^2\sum_{j=1}^{q}x_{s,t}^2[j]\cdot 1_{\beta_j\neq 0}$ so $\pi(\boldsymbol{\theta})$ is a combination of L_0, L_1 and L_2 regularization [21].

Proof.

$$\begin{aligned}\pi(\boldsymbol{\theta}) &= \mathbb{E}_{\boldsymbol{u}}\, loss(\boldsymbol{\theta}|\boldsymbol{u}) - loss(\boldsymbol{\theta}) \\ &= -\sum_{t=1}^{T}\sum_{s=1}^{n}\mathbb{E}_{\boldsymbol{u}}\,\ln f(\varepsilon_{s,t}(\boldsymbol{\theta}|\boldsymbol{u})) - \ln f(\varepsilon_{s,t}(\boldsymbol{\theta}))\end{aligned}$$

And the residuals after injecting noise $\boldsymbol{u}$ are

$$\varepsilon_{s,t}(\boldsymbol{\theta}|\boldsymbol{u}) = y_{s,t} - \sum_{k=1}^{n} \phi_0 w_{sk} y_{k,t} - \sum_{i=1}^{p} \phi_i \sum_{k=1}^{n} w_{sk} y_{k,t-i} - \tilde{x}'_{s,t}\beta - \sum_{k=1}^{h} \lambda_k F(x'_{s,t}\boldsymbol{\gamma}_k)$$

By the property of Shakeout, $\mathbb{E}_{\boldsymbol{u}}(\tilde{x}'_{s,t}\beta) = x'_{s,t}\beta$. So expanding $\mathbb{E}_{\boldsymbol{u}} \ln f(\varepsilon_{s,t}(\boldsymbol{\theta}|\boldsymbol{u}))$ around $x'_{s,t}\beta$ with regards to the noise distribution [28, Lemma 1],

$$\begin{aligned}
\mathbb{E}_{\boldsymbol{u}} \ln f(\varepsilon_{s,t}(\boldsymbol{\theta}|\boldsymbol{u})) \approx & \ln f(\varepsilon_{s,t}(\boldsymbol{\theta})) + \frac{\partial \ln f(\varepsilon_{s,t}(\boldsymbol{\theta}))}{\partial (x'_{s,t}\beta)} \mathbb{E}_{\boldsymbol{u}}(\tilde{x}'_{s,t}\beta - x'_{s,t}\beta) \\
& + \frac{1}{2} \frac{\partial^2 \ln f(\varepsilon_{s,t}(\boldsymbol{\theta}))}{\partial (x'_{s,t}\beta) \partial (x'_{s,t}\beta)'} \mathbb{E}_{\boldsymbol{u}}(\tilde{x}'_{s,t}\beta - x'_{s,t}\beta)^2 \\
\approx & \ln f(\varepsilon_{s,t}(\boldsymbol{\theta})) + \frac{1}{2} \frac{\partial^2 \ln f(\varepsilon_{s,t}(\boldsymbol{\theta}))}{\partial (x'_{s,t}\beta) \partial (x'_{s,t}\beta)'} \mathrm{Var}_{\boldsymbol{u}}(\tilde{x}'_{s,t}\beta) \qquad (3.10)
\end{aligned}$$

And, we can compute $\mathrm{Var}_{\boldsymbol{u}}(\tilde{x}'_{s,t}\beta)$ by

$$\begin{aligned}
\mathrm{Var}_{\boldsymbol{u}}(\tilde{x}'_{s,t}\beta) &= \mathbb{E}_{\boldsymbol{u}}(\tilde{x}'_{s,t}\beta - x'_{s,t}\beta)^2 \\
&= \sum_{j=1}^{q} (x_{s,t}[j]\beta_j)^2 (1 + \frac{c}{|\beta_j|})^2 \cdot \tau + (x_{s,t}[j]\beta_j)^2 (\frac{\tau}{1-\tau} + \frac{1+c}{|\beta_j|})^2 \cdot (1-\tau) \\
&= \sum_{j=1}^{q} (x_{s,t}[j]\beta_j)^2 (1 + \frac{c}{|\beta_j|})^2 \frac{\tau}{1-\tau} \\
&= \frac{\tau}{1-\tau} ||x_{s,t} \circ (\beta + \frac{c\beta}{|\beta|})||_2^2 \qquad (3.11)
\end{aligned}$$

Substituting (3.10), (3.11) into $\pi(\beta)$,

$$\begin{aligned}
\pi(\boldsymbol{\theta}) &\approx -\frac{1}{2} \sum_{t=1}^{T} \sum_{s=1}^{n} \frac{\partial^2 \ln f(\varepsilon_{s,t}(\boldsymbol{\theta}))}{\partial (x'_{s,t}\beta) \partial (x'_{s,t}\beta)'} \mathrm{Var}_{\boldsymbol{u}}(\tilde{x}'_{s,t}\beta) \\
&\approx \frac{1}{2} \sum_{t=1}^{T} \sum_{s=1}^{n} \left(\frac{f'^2(\varepsilon_{s,t}(\boldsymbol{\theta})}{f^2(\varepsilon_{s,t}(\boldsymbol{\theta}))} - \frac{f''(\varepsilon_{s,t}(\boldsymbol{\theta})}{f(\varepsilon_{s,t}(\boldsymbol{\theta}))} \right) \mathrm{Var}_{\boldsymbol{u}}(\tilde{x}'_{s,t}\beta) \\
&\approx \frac{\tau}{2(1-\tau)} \sum_{t=1}^{T} \sum_{s=1}^{n} \left(\frac{f'^2(\varepsilon_{s,t}(\boldsymbol{\theta})}{f^2(\varepsilon_{s,t}(\boldsymbol{\theta}))} - \frac{f''(\varepsilon_{s,t}(\boldsymbol{\theta})}{f(\varepsilon_{s,t}(\boldsymbol{\theta}))} \right) ||x_{s,t} \circ (\beta + \frac{c\beta}{|\beta|})||_2^2
\end{aligned}$$

where $\beta + \frac{c\beta}{|\beta|} = \left(\beta_1 + \frac{c\beta_1}{|\beta_1|}, \ldots, \beta_q + \frac{c\beta_q}{|\beta_q|}\right)'$. $||x_{s,t} \circ (\beta + \frac{c\beta}{|\beta|})||_2^2$ can be decomposed into [see 21, pg. 1248]:

$$\sum_{j=1}^{q}(x_{s,t}[j]\beta_j)^2 + 2c\sum_{j=1}^{q}x_{s,t}^2[j]|\beta_j| + c^2\sum_{j=1}^{q}x_{s,t}^2[j]\cdot 1_{\beta_j\neq 0}$$

□

From the term $||x_{s,t}\circ(\beta+\frac{c\beta}{|\beta|})||_2^2$, we could also see that, in a linear model, the Shakeout tends to penalize those parameters whose corresponding features' magnitudes are large. On the other hand, parameters whose features have more zeros are less penalized. With $c > 0$, the penalization terms $\sum_{j=1}^{q} x_{s,t}^2[j]|\beta_j|$ and $\sum_{j=1}^{q} x_{s,t}^2[j]\cdot 1_{\beta_j\neq 0}$ can help with feature selection. This penalization can train the parameters to be more confident about its prediction [21].

Next we discuss injecting Shakeout noise in the neural network component. Matsuoka [31] brought up an assumption on the property of neural networks: "the output of the network after learning should be as insensitive as possible to input variation, as long as the error is within a reasonable bound". The mentioned "error" in his assumption refers to the loss function in the OLS estimation, the sum of squared residuals. This statement means that similar inputs tend to produce similar outputs even if the input is not trained before. Suppose the random disturbance for input $x_{s,t}$ is $d_{s,t}$, $\boldsymbol{d} = \{d_{s,t}\}$. Matsuoka [31] and Li and Liu [28] assumed $d_{s,t}$ has identical distributions for all s, t. But in some occasions, larger $|x_{s,t}|$ may have larger value of perturbations and variances so the identical distributed perturbation is not always true. To accommodate this, we define a Shakeout perturbation $d_{s,t}$ as follows:

$$d_{s,t}[j] = \begin{cases} -h_j & \Pr = \tau_d \\ \frac{h_j\tau_d+1}{1-\tau_d} & \Pr = 1-\tau_d \end{cases} \qquad j = 1, 2, \ldots, q \tag{3.12}$$

where h_j, τ_d are hyper-parameters and h_j is positive. We allow h_j varying across different predictors so that the injected noise can vary with predictors. Assume $\mathbb{E}\, d_{s,t} = \mathbf{1}_q$, $\mathrm{Var}(d_{s,t}) = \frac{\tau_d}{1-\tau_d}\mathrm{diag}((h_1+1)^2, \ldots, (h_q+1)^2)$ for $s = 1, \ldots, n, t = 1, \ldots, T$, then the perturbed observation $x_{s,t}$ are

(3.13)
$$\underset{\text{Input}}{x_{s,t} = \begin{bmatrix} x_{s,t}[1] \\ x_{s,t}[2] \\ \vdots \\ x_{s,t}[q] \end{bmatrix}_q} \quad \underset{\text{Perturbance}}{d_{s,t} = \begin{bmatrix} d_{s,t}[1] \\ d_{s,t}[2] \\ \vdots \\ d_{s,t}[q] \end{bmatrix}_q} \xrightarrow[\text{perturbation}]{\text{External}} \underset{\text{Perturbed Input}}{x_{s,t} \circ d_{s,t} = \begin{bmatrix} x_{s,t}[1]d_{s,t}[1] \\ x_{s,t}[2]d_{s,t}[2] \\ \vdots \\ x_{s,t}[q]d_{s,t}[q] \end{bmatrix}_q}$$

We can consider $x_{s,t} \circ d_{s,t}$ as an input pattern similar to the training input $x_{s,t}$. This simulates the situation when the training data is disturbed by external disturbances even if they are sample from a same distribution [31]. Combined with Shakeout noise $v_{s,t}^{(1)}, v_{s,t}^{(2)}$, the output of the ANN component is given by

(3.14)
$$\underset{\text{Input}}{x_{s,t} = \begin{bmatrix} x_{s,t}[1] \\ x_{s,t}[2] \\ \vdots \\ x_{s,t}[q] \end{bmatrix}_q} \quad \underset{\text{Perturbance}}{d_{s,t} = \begin{bmatrix} d_{s,t}[1] \\ d_{s,t}[2] \\ \vdots \\ d_{s,t}[q] \end{bmatrix}_q} \quad \underset{\text{Shakeout Noise}}{v_{s,t}^{(1)} = \begin{bmatrix} v_{s,t}^{(1,1)\prime} \\ v_{s,t}^{(1,2)\prime} \\ \vdots \\ v_{s,t}^{(1,h)\prime} \end{bmatrix}'_{q\times h}} \quad \underset{\text{Shakeout Noise}}{v_{s,t}^{(2)} = \begin{bmatrix} v_{s,t}^{(2)}[1] \\ v_{s,t}^{(2)}[2] \\ \vdots \\ v_{s,t}^{(2)}[h] \end{bmatrix}_h} \xrightarrow{\text{Shakeout}} \underset{\text{Final Output}}{\begin{bmatrix} F\Big((\hat{x}_{s,t}^{(1)} \circ d_{s,t})'\boldsymbol{\gamma}_1\Big)v_{s,t}^{(2)}[1] \\ F\Big((\hat{x}_{s,t}^{(2)} \circ d_{s,t})'\boldsymbol{\gamma}_2\Big)v_{s,t}^{(2)}[2] \\ \vdots \\ F\Big((\hat{x}_{s,t}^{(h)} \circ d_{s,t})'\boldsymbol{\gamma}_h\Big)v_{s,t}^{(2)}[h] \end{bmatrix} \begin{bmatrix} \lambda_1 \\ \lambda_2 \\ \vdots \\ \lambda_h \end{bmatrix}}$$

Hence the variation of the ANN output after Shakeout noises and random perturbation should be

(3.15)
$$\begin{aligned} \delta_{s,t} &= \sum_{k=1}^{h} \lambda_k \tilde{F}\Big((\hat{x}_{s,t}^{(k)} \circ d_{s,t})'\boldsymbol{\gamma}_k\Big) - \lambda F(x_{s,t}'\boldsymbol{\gamma}_k) \\ &= \sum_{k=1}^{h} \lambda_k v_{s,t}^{(2)}[k] F\Big((\hat{x}_{s,t}^{(k)} \circ d_{s,t})'\boldsymbol{\gamma}_k\Big) - \lambda F(x_{s,t}'\boldsymbol{\gamma}_k) \end{aligned}$$

Based on the definition by Matsuoka [31], in a PSTAR-ANN(p) model, we can define the sensitivity of a learned ANN component as the mean ratio of the variance of $\delta_{s,t}$ and $|d_{s,t}|$, where f is the density of random error and $\varepsilon_{s,t}(\boldsymbol{\theta})$ is the residual at location s, time t evaluated at $\boldsymbol{\theta}$.

$$R(\boldsymbol{\theta}) = \sum_{t=1}^{T}\sum_{s=1}^{n} \frac{\partial^2 \ln f(\varepsilon_{s,t}(\boldsymbol{\theta}))}{\partial^2 \varepsilon_{s,t}(\boldsymbol{\theta})} \frac{\mathrm{Var}_{v_{s,t},d_{s,t}}(\delta_{s,t})}{\sum_{j=1}^{q} \mathrm{Var}_{d_{s,t}}(d_{s,t}[j])}$$

Following Li and Liu's work, we want to justify that the Shakeout noise procedure can stabilize a learned network in the sense that external perturbations in the training data can be taken into account. More specifically, because the training data can vary from sample to sample, under proper settings of hyper-parameters, Shakeout noises can be considered as external perturbations in the data generating process. In this sense, training a network with Shakeout can be robust with respect to the outcome parameters. Theorem 2 supports the idea that minimizing the original loss function (with original inputs) plus a penalty for the sensitivity of a network is approximately equivalent to minimizing the original loss function with original observations and a set of Shakeout noise $\boldsymbol{v}^*$ (different parameterization from $\boldsymbol{v}$). And it illustrates that injecting Shakeout noise in the neural network component can simulate the situation when we are training the network with perturbed inputs. The procedure of the proof is shown in Figure 3.3.

Theorem 6 (**Low sensitivity of a learned ANN component in a PSTAR-ANN** (p) **model with Shakeout).** Given X and Y, the expected value of the noise perturbed loss function over the distribution of the injected Shakeout noise $\boldsymbol{v}^*$ is approximately

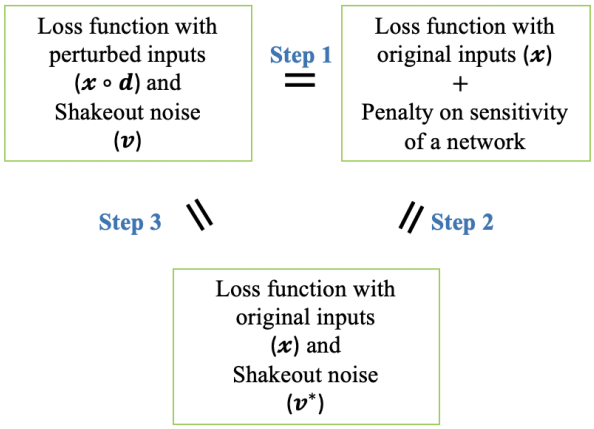


Figure 3.3. Steps of the proof in Theorem 2

equivalent to the original loss function plus the sensitivity of the neural network.

$$\mathbb{E}_{\boldsymbol{v}^*} loss(\boldsymbol{\theta}|\boldsymbol{v}^*) \approx loss(\boldsymbol{\theta}) + aR(\boldsymbol{\theta}) \tag{3.16}$$

where $a = \frac{\tau_d \sum_{i=1}^{q}(h_i+1)^2}{2(1-\tau_d)}$, h_i, τ_d are hyper-parameters of Shakeout noise $d_{s,t}$ ($d_{s,t}$ is the perturbation in the input variable $x_{s,t}$, see definition in (3.12)). The sensitivity is defined as

$$\begin{aligned} R(\boldsymbol{\theta}) &= \sum_{t=1}^{T}\sum_{s=1}^{n} \frac{\partial^2 \ln f(\varepsilon_{s,t}(\boldsymbol{\theta}))}{\partial^2 \varepsilon_{s,t}(\boldsymbol{\theta})} \frac{\mathrm{Var}_{v_{s,t},d_{s,t}}(\delta_{s,t})}{\sum_{j=1}^{q}\mathrm{Var}_{d_{s,t}}(d_{s,t}[j])} \\ &= \sum_{t=1}^{T}\sum_{s=1}^{n} \left(\frac{f'^2(\varepsilon_{s,t}(\boldsymbol{\theta}))}{f^2(\varepsilon_{s,t}(\boldsymbol{\theta}))} - \frac{f''(\varepsilon_{s,t}(\boldsymbol{\theta}))}{f(\varepsilon_{s,t}(\boldsymbol{\theta}))} \right) \frac{\mathbb{E}_{v_{s,t},d_{s,t}}(\delta_{s,t}^2)}{\sum_{j=1}^{q}\mathrm{Var}_{d_{s,t}}(d_{s,t}[j])} \\ &= \frac{1-\tau_d}{\tau_d \sum_{i=1}^{q}(h_i+1)^2} \sum_{t=1}^{T}\sum_{s=1}^{n} \left(\frac{f'^2(\varepsilon_{s,t}(\boldsymbol{\theta}))}{f^2(\varepsilon_{s,t}(\boldsymbol{\theta}))} - \frac{f''(\varepsilon_{s,t}(\boldsymbol{\theta}))}{f(\varepsilon_{s,t}(\boldsymbol{\theta}))} \right) \Psi'_{s,t}\Lambda\Psi_{s,t} \end{aligned}$$

where $\delta_{s,t} = \sum_{k=1}^{h} \lambda_k \tilde{F}\left((\hat{x}_{s,t}^{(k)} \circ d_{s,t})'\boldsymbol{\gamma}_k\right) - \lambda F(x'_{s,t}\boldsymbol{\gamma}_k)$. $v_{s,t}$ is the Shakeout noise with hyper-parameter τ, c defined previously in (3.4) and (3.5). $\Psi_{s,t}$ is the gradient of the ANN structure with regard to $v_{s,t}, d_{s,t}$.

$$\Lambda = \begin{pmatrix} \Big((I_h + \Lambda_2) \otimes I_q\Big)\Lambda_1 & 0 \\ 0 & \Lambda_2 \end{pmatrix}_{h(q+1)\times h(q+1)}$$

The details of this matrix will be discussed in the following proof. The injected Shakeout noise $\boldsymbol{v}^*$ which leads to equation (3.16) is: for the first layer, $v_{s,t}^{*(1)} = \{v_{s,t}^{*(1,1)}, \ldots, v_{s,t}^{*(1,h)}\}$, where $v_{s,t}^{*(1,k)} =$

$$\begin{cases} \frac{h_j c}{|\gamma_{kj}|} & \Pr = \tau_d \tau \\ \frac{-c(h_j \tau_d + 1)}{|\gamma_{kj}|(1-\tau_d)} & \Pr = (1-\tau_d)\tau \\ \frac{-h_j}{1-\tau}(1 + \frac{c\tau}{|\gamma_{kj}|}) & \Pr = \tau_d(1-\tau) \\ \frac{h_j \tau_d + 1}{(1-\tau)(1-\tau_d)}(1 + \frac{c\tau}{|\gamma_{kj}|}) & \Pr = (1-\tau_d)(1-\tau) \end{cases}$$

for $k = 1, \ldots, h$, $j = 1, \ldots, q$. For the kth input nodes in the second layer, $v_{s,t}^{*(2)} =$

$$\begin{cases} \frac{-c}{|\lambda_k|} & \Pr = \tau \\ \frac{1}{1-\tau}(1 + \frac{c\tau}{|\lambda_k|}) & \Pr = 1-\tau \end{cases}$$

for $k = 1, \ldots, h$.

PROOF. First we want to prove that $\mathbb{E}_{\boldsymbol{v},\boldsymbol{d}} loss(\boldsymbol{\theta}|\boldsymbol{v}, \boldsymbol{d}) \approx loss(\boldsymbol{\theta}) + aR(\boldsymbol{\theta})$.

Consider the Shakeout noise $\boldsymbol{v}$, and external disturbed input $x_{s,t} \circ d_{s,t}$,

$$\mathbb{E}_{\boldsymbol{v},\boldsymbol{d}} loss(\boldsymbol{\theta}|\boldsymbol{v}, \boldsymbol{d}) - loss(\boldsymbol{\theta}) = \sum_{t=1}^{T} \sum_{s}^{n} \mathbb{E}_{\boldsymbol{v},\boldsymbol{d}} \ln f(\varepsilon_{s,t}(\boldsymbol{\theta}|\boldsymbol{v}, \boldsymbol{d})) - \ln f(\varepsilon_{s,t}(\boldsymbol{\theta}))$$

Expanding $\mathbb{E}_{\boldsymbol{v},\boldsymbol{d}} \ln f(\varepsilon_{s,t}(\boldsymbol{\theta}|\boldsymbol{v}, \boldsymbol{d}))$ around $\varepsilon_{s,t}(\boldsymbol{\theta})$, we have,

$$\begin{aligned}\mathbb{E}_{\boldsymbol{v},\boldsymbol{d}} \ln f(\varepsilon_{s,t}(\boldsymbol{\theta}|\boldsymbol{v},\boldsymbol{d})) - \ln f(\varepsilon_{s,t}(\boldsymbol{\theta})) \approx & \frac{\partial \ln f(\varepsilon_{s,t}(\boldsymbol{\theta}))}{\partial \varepsilon_{s,t}(\boldsymbol{\theta})} \mathbb{E}_{\boldsymbol{v},\boldsymbol{d}}(\varepsilon_{s,t}(\boldsymbol{\theta}|\boldsymbol{v},\boldsymbol{d}) - \varepsilon_{s,t}(\boldsymbol{\theta})) \\ & + \frac{1}{2} \frac{\partial^2 \ln f(\varepsilon_{s,t}(\boldsymbol{\theta}))}{\partial^2 \varepsilon_{s,t}(\boldsymbol{\theta})} \mathbb{E}_{\boldsymbol{v},\boldsymbol{d}}(\varepsilon_{s,t}(\boldsymbol{\theta}|\boldsymbol{v},\boldsymbol{d}) - \varepsilon_{s,t}(\boldsymbol{\theta}))^2 \\ \approx & \frac{1}{2} \frac{\partial^2 \ln f(\varepsilon_{s,t}(\boldsymbol{\theta}))}{\partial^2 \varepsilon_{s,t}(\boldsymbol{\theta})} \mathbb{E}_{\boldsymbol{v},\boldsymbol{d}}(\delta_{s,t}^2) \end{aligned} \tag{3.17}$$

where $\delta_{s,t}$ is given by

$$\begin{aligned}\delta_{s,t} =& \varepsilon_{s,t}(\boldsymbol{\theta}|\boldsymbol{v},\boldsymbol{d}) - \varepsilon_{s,t}(\boldsymbol{\theta}) \\ =& \sum_{k=1}^{h} \lambda_k \tilde{F}((\hat{x}_{s,t}^{(k)} \circ d_{s,t})' \boldsymbol{\gamma}_k) - \lambda_k F(x_{s,t}' \boldsymbol{\gamma}_k) \\ =& \sum_{k=1}^{h} \lambda_k \tilde{F}((\hat{x}_{s,t}^{(k)} \circ d_{s,t})' \boldsymbol{\gamma}_k) - \lambda_k \tilde{F}(x_{s,t}' \boldsymbol{\gamma}_k) + \lambda_k \tilde{F}(x_{s,t}' \boldsymbol{\gamma}_k) - \lambda_k F(x_{s,t}' \boldsymbol{\gamma}_k) \\ \approx & \sum_{k=1}^{h} \lambda_k \frac{\partial \tilde{F}(x_{s,t}' \boldsymbol{\gamma}_k)}{\partial x_{s,t}'} (\hat{x}_{s,t}^{(k)} \circ d_{s,t} - x_{s,t}) + \lambda_k \Big(\tilde{F}(x_{s,t}' \boldsymbol{\gamma}_k) - F(x_{s,t}' \boldsymbol{\gamma}_k) \Big) \end{aligned} \tag{3.18}$$

where

$$\hat{x}_{s,t}^{(k)} = v_{s,t}^{(1,k)} \circ x_{s,t},\ \tilde{F}(x_{s,t}' \boldsymbol{\gamma}_k) = F(x_{s,t}' \boldsymbol{\gamma}_k) v_{s,t}^{(2)}[k]$$

Hence,

$$\mathbb{E}_{\boldsymbol{v},\boldsymbol{d}} loss(\boldsymbol{\theta}|\boldsymbol{v},\boldsymbol{d}) \approx loss(\boldsymbol{\theta}) + \frac{1}{2} \sum_{t=1}^{T} \sum_{s=1}^{n} \frac{\partial^2 \ln f(\varepsilon_{s,t}(\boldsymbol{\theta}))}{\partial^2 \varepsilon_{s,t}(\boldsymbol{\theta})} \mathbb{E}_{\boldsymbol{v},\boldsymbol{d}}(\delta_{s,t}^2)$$

Then we can expand $\delta_{s,t}$ as follows. Equation (3.18) shows an approximation of $\delta_{s,t}$ through the first order Taylor expansion and we can rewrite this in matrix form $\delta_{s,t} \approx$

$\Psi_{s,t}(\boldsymbol{\theta})'\boldsymbol{n}_{s,t}$, where $\Psi_{s,t}(\boldsymbol{\theta})$, $\boldsymbol{n}_{s,t}$ are:

$$\Psi_{s,t}(\boldsymbol{\theta}_{nn}) = \begin{bmatrix} \lambda_1 F'(x'_{s,t}\boldsymbol{\gamma}_1)(x_{s,t} \circ \boldsymbol{\gamma}_1) \\ \lambda_2 F'(x'_{s,t}\boldsymbol{\gamma}_2)(x_{s,t} \circ \boldsymbol{\gamma}_2) \\ \vdots \\ \lambda_h F'(x'_{s,t}\boldsymbol{\gamma}_h)(x_{s,t} \circ \boldsymbol{\gamma}_h) \\ \lambda_1 F(x'_{s,t}\boldsymbol{\gamma}_1) \\ \lambda_2 F(x'_{s,t}\boldsymbol{\gamma}_2) \\ \vdots \\ \lambda_h F(x'_{s,t}\boldsymbol{\gamma}_h) \end{bmatrix}, \boldsymbol{n}_{s,t} = \begin{bmatrix} v_{s,t}^{(2)}[1]\Big(v_{s,t}^{(1,1)} \circ d_{s,t} - \mathbf{1}_q\Big) \\ v_{s,t}^{(2)}[2]\Big(v_{s,t}^{(1,2)} \circ d_{s,t} - \mathbf{1}_q\Big) \\ \vdots \\ v_{s,t}^{(2)}[h]\Big(v_{s,t}^{(1,h)} \circ d_{s,t} - \mathbf{1}_q\Big) \\ v_{s,t}^{(2)}[1] - 1 \\ v_{s,t}^{(2)}[2] - 1 \\ \vdots \\ v_{s,t}^{(2)}[h] - 1 \end{bmatrix}$$

In the following, we use $\Psi_{s,t}$ in place of $\Psi_{s,t}(\boldsymbol{\theta})$. It is easy to show that $\mathbb{E}_{v_{s,t},d_{s,t}}(\delta_{s,t}) = 0$. Therefore,

$$\frac{1}{2}\sum_{t=1}^{T}\sum_{s=1}^{n}\frac{\partial^2 \ln f(\varepsilon_{s,t}(\boldsymbol{\theta}))}{\partial^2 \varepsilon_{s,t}(\boldsymbol{\theta})}\mathbb{E}_{\boldsymbol{v},\boldsymbol{d}}(\delta_{s,t}^2) = \frac{1}{2}\sum_{t=1}^{T}\sum_{s=1}^{n}\left(\frac{f'^2(\varepsilon_{s,t}(\boldsymbol{\theta}))}{f^2(\varepsilon_{s,t}(\boldsymbol{\theta}))} - \frac{f''(\varepsilon_{s,t}(\boldsymbol{\theta}))}{f(\varepsilon_{s,t}(\boldsymbol{\theta}))}\right)\Psi'_{s,t}\Lambda\Psi_{s,t}$$

where $\Lambda = \begin{pmatrix} \Big((I_h + \Lambda_2) \otimes I_q\Big)\Lambda_1 & 0 \\ 0 & \Lambda_2 \end{pmatrix}$,

$$\Lambda_1 = \begin{pmatrix} \Lambda_{1,1} & & & & \\ & \Lambda_{1,2} & & 0 & \\ & & \Lambda_{1,3} & & \\ & & & \ddots & \\ & & & & \Lambda_{1,h} \end{pmatrix} \Lambda_2 = \frac{\tau}{1-\tau}\begin{pmatrix} (1+\frac{c}{|\lambda_1|})^2 & 0 & \cdots & 0 \\ 0 & (1+\frac{c}{|\lambda_2|})^2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & (1+\frac{c}{|\lambda_h|})^2 \end{pmatrix}$$

where $\Lambda_{1,k}$ is a $q \times q$ diagonal matrix with j^{th} diagonal entry:

$$\Big(\frac{\tau(1+\frac{c}{|\gamma_{kj}|})^2}{1-\tau}+1\Big)\Big(\frac{\tau_d(1+h_j)^2}{1-\tau_d}+1\Big)-1, \quad j=1,\ldots,q$$

Because $\mathrm{Var}(d_{s,t}) = \frac{\tau_d}{1-\tau_d}\mathrm{diag}((h_1+1)^2,\ldots,(h_q+1)^2)$, $\sum_{j=1}^{q}\mathrm{Var}_{d_{s,t}}(d_{s,t}[j]) = \frac{\tau_d\sum_{i=1}^{q}(h_i+1)^2}{1-\tau_d}$. So

$$\begin{aligned}
R(\boldsymbol{\theta}) &= \sum_{t=1}^{T}\sum_{s=1}^{n}\frac{\partial^2 \ln f(\varepsilon_{s,t}(\boldsymbol{\theta})}{\partial^2\varepsilon_{s,t}(\boldsymbol{\theta})}\frac{\mathrm{Var}_{v_{s,t},d_{s,t}}(\delta_{s,t})}{\mathrm{Var}_{d_{s,t}}(|d_{s,t}|)} \\
&= \sum_{t=1}^{T}\sum_{s=1}^{n}\left(\frac{f'^2(\varepsilon_{s,t}(\boldsymbol{\theta}))}{f^2(\varepsilon_{s,t}(\boldsymbol{\theta}))}-\frac{f''(\varepsilon_{s,t}(\boldsymbol{\theta}))}{f(\varepsilon_{s,t}(\boldsymbol{\theta}))}\right)\frac{\mathbb{E}_{v_{s,t},d_{s,t}}(\delta_{s,t}^2)}{\mathrm{Var}_{d_{s,t}}(|d_{s,t}|)} \\
&= \frac{1-\tau_d}{\tau_d\sum_{i=1}^{q}(h_i+1)^2}\sum_{t=1}^{T}\sum_{s=1}^{n}\left(\frac{f'^2(\varepsilon_{s,t}(\boldsymbol{\theta}))}{f^2(\varepsilon_{s,t}(\boldsymbol{\theta}))}-\frac{f''(\varepsilon_{s,t}(\boldsymbol{\theta}))}{f(\varepsilon_{s,t}(\boldsymbol{\theta}))}\right)\Psi'_{s,t}\Lambda\Psi_{s,t} \\
aR(\boldsymbol{\theta}) &= \frac{1}{2}\sum_{t=1}^{T}\sum_{s=1}^{n}\left(\frac{f'^2(\varepsilon_{s,t}(\boldsymbol{\theta}))}{f^2(\varepsilon_{s,t}(\boldsymbol{\theta}))}-\frac{f''(\varepsilon_{s,t}(\boldsymbol{\theta}))}{f(\varepsilon_{s,t}(\boldsymbol{\theta}))}\right)\Psi'_{s,t}\Lambda\Psi_{s,t}
\end{aligned}$$

Therefore, we can show that

$$\mathbb{E}_{\boldsymbol{v},\boldsymbol{d}}loss(\boldsymbol{\theta}|\boldsymbol{v},\boldsymbol{d}) \approx loss(\boldsymbol{\theta}) + aR(\boldsymbol{\theta})$$

This equation indicates that minimizing the loss function with the original data plus a penalty for the sensitivity of a network is approximately equivalent to minimizing the loss function with perturbed inputs (perturbation $\boldsymbol{d}$) and Shakeout noise $\boldsymbol{v}$.

The next step is to prove that there exists a set of Shakeout noise $\boldsymbol{v}^*$ whose regularization effect is equivalent to injecting Shakeout noise $\boldsymbol{v}$ into perturbed inputs. We show this by first prove that

$$\mathbb{E}_{\boldsymbol{v}^*}loss(\boldsymbol{\theta}|\boldsymbol{v}^*) \approx loss(\boldsymbol{\theta}) + aR(\boldsymbol{\theta})$$

And then we can show that $\mathbb{E}_{\boldsymbol{v}^*}loss(\boldsymbol{\theta}|\boldsymbol{v}^*)$ is approximately equivalent to $\mathbb{E}_{\boldsymbol{v},\boldsymbol{d}}loss(\boldsymbol{\theta}|\boldsymbol{v},\boldsymbol{d})$.

To achieve this, we define $\boldsymbol{v}^*$ as follows: in the first layer, $v_{s,t}^{*(1)} = \{v_{s,t}^{*(1,1)}, \ldots, v_{s,t}^{*(1,h)}\}$, where $v_{s,t}^{*(1,k)}[j] =$

$$\begin{cases} \frac{h_j c}{|\gamma_{kj}|} & \Pr = \tau_d \tau \\ \frac{-c(h_j \tau_d + 1)}{|\gamma_{kj}|(1-\tau_d)} & \Pr = (1-\tau_d)\tau \\ \frac{-h_j}{1-\tau}(1 + \frac{c\tau}{|\gamma_{kj}|}) & \Pr = \tau_d(1-\tau) \\ \frac{h_j \tau_d + 1}{(1-\tau)(1-\tau_d)}(1 + \frac{c\tau}{|\gamma_{kj}|}) & \Pr = (1-\tau_d)(1-\tau) \end{cases}$$

for $k = 1, \ldots, h$, $j = 1, \ldots, q$. For the kth input nodes in the second layer, $v_{s,t}^{*(2)}[k] =$

$$\begin{cases} \frac{-c}{|\lambda_k|} & \Pr = \tau \\ \frac{1}{1-\tau}(1 + \frac{c\tau}{|\lambda_k|}) & \Pr = 1-\tau \end{cases}$$

for $k = 1, \ldots, h$. Under this setting, it is easy to see that $v_{s,t}^{*(1,k)} = v_{s,t}^{(1,k)} \circ d_{s,t}$ and $v_{s,t}^{*(2)} = v_{s,t}^{(2)}$. So $\varepsilon_{s,t}(\boldsymbol{\theta}|\boldsymbol{v}^*) = \varepsilon_{s,t}(\boldsymbol{\theta}|\boldsymbol{v}, \boldsymbol{d})$. Let $e_{s,t} = \varepsilon_{s,t}(\boldsymbol{\theta}|\boldsymbol{v}^*) - \varepsilon_{s,t}(\boldsymbol{\theta})$. Then,

$$\begin{aligned} \delta_{s,t} =& \varepsilon_{s,t}(\boldsymbol{\theta}|\boldsymbol{v}, \boldsymbol{d}) - \varepsilon_{s,t}(\boldsymbol{\theta}) \\ =& \varepsilon_{s,t}(\boldsymbol{\theta}|\boldsymbol{v}^*) - \varepsilon_{s,t}(\boldsymbol{\theta}) \\ =& e_{s,t} \end{aligned}$$

And we can validate that $\mathbb{E}_{v_{s,t}^*} e_{s,t} = 0$, $\mathbb{E}_{v_{s,t}^*}(e_{s,t}^2) = \Psi'_{s,t} \Lambda \Psi_{s,t}$. Previously we have shown that

$$aR(\boldsymbol{\theta}) = \frac{1}{2} \sum_{t=1}^{T} \sum_{s=1}^{n} \left(\frac{f'^2(\varepsilon_{s,t}(\boldsymbol{\theta}))}{f^2(\varepsilon_{s,t}(\boldsymbol{\theta}))} - \frac{f''(\varepsilon_{s,t}(\boldsymbol{\theta}))}{f(\varepsilon_{s,t}(\boldsymbol{\theta}))} \right) \Psi'_{s,t} \Lambda \Psi_{s,t}$$

Similar to (3.17), we can show that

$$\mathbb{E}_{\boldsymbol{v}^*} loss(\boldsymbol{\theta}|\boldsymbol{v}^*) \approx loss(\boldsymbol{\theta}) + \frac{1}{2} \sum_{t=1}^{T} \sum_{s=1}^{n} \left(\frac{f'^2(\varepsilon_{s,t}(\boldsymbol{\theta}))}{f^2(\varepsilon_{s,t}(\boldsymbol{\theta}))} - \frac{f''(\varepsilon_{s,t}(\boldsymbol{\theta}))}{f(\varepsilon_{s,t}(\boldsymbol{\theta}))} \right) \mathbb{E}_{v_{s,t}^*}(e_{s,t}^2) \tag{3.19}$$

$$\approx loss(\boldsymbol{\theta}) + aR(\boldsymbol{\theta}) \tag{3.20}$$

Expanding $loss(\boldsymbol{\theta})$ in (3.19), with the second order Taylor expansion,

$$\begin{aligned}
\mathbb{E}_{\boldsymbol{v},\boldsymbol{d}} loss(\boldsymbol{\theta}|\boldsymbol{v},\boldsymbol{d}) \approx & \left(-T \ln |I_n - \phi_0 W_n| + \sum_{t=1}^{T} \sum_{s=1}^{n} \ln f(\varepsilon_{s,t}(\boldsymbol{\theta})) \right) \\
& + \frac{1}{2} \sum_{t=1}^{T} \sum_{s=1}^{n} \frac{\partial^2 \ln f(\varepsilon_{s,t}(\boldsymbol{\theta}))}{\partial^2 \varepsilon_{s,t}(\boldsymbol{\theta})} \mathbb{E}_{v^*_{s,t}} (e^2_{s,t}) \\
\approx & - T \ln |I_n - \phi_0 W_n| + \sum_{t=1}^{T} \sum_{s=1}^{n} \left(\ln f(\varepsilon_{s,t}(\boldsymbol{\theta})) + \frac{1}{2} \frac{\partial^2 \ln f(\varepsilon_{s,t}(\boldsymbol{\theta})}{\partial^2 \varepsilon_{s,t}(\boldsymbol{\theta}))} \mathbb{E}_{v^*_{s,t}} (e^2_{s,t}) \right) \\
\approx & - T \ln |I_n - \phi_0 W_n| + \sum_{t=1}^{T} \sum_{s=1}^{n} \left(\ln f(\varepsilon_{s,t}(\boldsymbol{\theta})) + \frac{\partial \ln \ln f(\varepsilon_{s,t}(\boldsymbol{\theta}))}{\partial \varepsilon_{s,t}(\boldsymbol{\theta})} \mathbb{E}_{v^*_{s,t}} (e^2_{s,t}) \right. \\
& \left. + \frac{1}{2} \frac{\partial^2 \ln f(\varepsilon_{s,t}(\boldsymbol{\theta}))}{\partial^2 \varepsilon_{s,t}(\boldsymbol{\theta})} \mathbb{E}_{v^*_{s,t}} (e^2_{s,t}) \right) \\
\approx & - T \ln |I_n - \phi_0 W_n| + \sum_{t=1}^{T} \sum_{s=1}^{n} \ln f(\varepsilon_{s,t}(\boldsymbol{\theta}) + e_{s,t}) \\
\approx & \mathbb{E}_{\boldsymbol{v}^*} loss(\boldsymbol{\theta}|\boldsymbol{v}^*)
\end{aligned}$$

Hence, $\mathbb{E}_{\boldsymbol{v}^*} loss(\boldsymbol{\theta}|\boldsymbol{v}^*)$ is approximately equivalent to $\mathbb{E}_{\boldsymbol{v},\boldsymbol{d}} loss(\boldsymbol{\theta}|\boldsymbol{v},\boldsymbol{d})$. □

The right side in equation (3.16) combines the loss function with the original data plus a penalty which is the sensitivity of a neural network component. This penalization measures the deviation of a network output when the input variables are perturbed by a multiplicative noise. So this new objective loss function mitigates the overfitting or instability problem in the original loss function (3.2). Note that a is a function of τ_d and h_j and it can be regarded as a tuning parameter by adjusting the values of hyper-parameters τ_d and h_j. When τ_d is larger, a will increase so the learned network will be more stable. When the magnitude of input variable $x_{s,t}[j]$ is large, h_j can be set as some larger constant so a will be larger. In return, the term $aR(\boldsymbol{\theta})$ will be larger so we impose a harsher penalization in the sensitivity of a network. On the other hand, choosing a

smaller τ_d or h_j results in a smaller a and the learned network will be rather vulnerable if inputs are disturbed by external random errors. So by adjusting the hyper-parameters of $d_{s,t}$, we can tune a and are able to find a network minimizing the sum of the original loss function and its sensitivity. Further discussion about hyper-parameter selection can be found in Kang et al. [21].

3.2. Space-time Autoregressive Order

In this section, we are going to discuss some techniques in selecting space-time autoregressive order p. First we will construct a likelihood ratio test for nested model selection and prove the asymptotic distribution of the LR test statistic. In practice, for preliminary analysis, we also suggest plotting sample ACF and PACF, which is an intuitive way to get a rough idea about the choice of p.

3.2.1. Likelihood Ratio Test

We know that the likelihood ratio test [13, pp. 488–492] can be used to select models which are nested so we can compare two PSTAR-ANN(p) models with different time lag p. For example, to test a model with order p versus $p+1$, the null and alternative hypothesis can be:

$$H_0 : Y_t = \phi_0 W_n Y_t + \sum_{i=1}^{p} \phi_i W_n Y_{t-i} + X_t\beta + \boldsymbol{F}(X_t\boldsymbol{\gamma}')\lambda + \boldsymbol{\varepsilon}_t$$

$$H_1 : Y_t = \phi_0 W_n Y_t + \sum_{i=1}^{p+1} \phi_i W_n Y_{t-i} + X_t\beta + \boldsymbol{F}(X_t\boldsymbol{\gamma}')\lambda + \boldsymbol{\varepsilon}_t$$

Or

$$H_0 : \phi_{p+1} = 0$$

$$H_1 : \phi_{p+1} \neq 0$$

Let the parameter estimates under the null and alternative be $\hat{\boldsymbol{\theta}}^{(0)}, \hat{\boldsymbol{\theta}}^{(1)}$ respectively. Then the Likelihood ratio test statistic is

$$LR = -2\Big(\mathcal{L}_{n,T}(\hat{\boldsymbol{\theta}}^{(0)}) - \mathcal{L}_{n,T}(\hat{\boldsymbol{\theta}}^{(1)})\Big)$$

and, under the null, is distributed asymptotically as $\chi^2(1)$. If the null is rejected, we may consider the model with $(p+1)$ time lags; otherwise, we can continue test if $\phi_p = 0$ until rejected. In the following, we are going to derive the asymptotic distribution of this test statistics [49, pg. 76-78].

Assumption 26. The limit $A(\boldsymbol{\theta}_0) = -\lim_{n,T\to\infty} \mathbb{E}\, \frac{1}{nT} \frac{\partial^2 \mathcal{L}_{n,T}(\boldsymbol{\theta}_0)}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}'}$ is nonsingular.

This assumption is to guarantee the existence of the covariance matrix of the limiting distribution of parameters in a PSTAR(p)-ANN model. And we need all other assumptions in chapter 2.

Theorem 7. Consider $\boldsymbol{\theta} \in \boldsymbol{\Theta}_{p+1}$ and rewrite $\boldsymbol{\theta} = (\phi_{p+1}, \boldsymbol{\psi})$. Let $\hat{\boldsymbol{\theta}}_{p+1}$, $\hat{\boldsymbol{\theta}}_p$ be the MLEs of a PSTAR-ANN($p+1$) model and its sub-model PSTAR-ANN(p) (with $\phi_{p+1} = 0$) respectively, $\mathcal{L}_{n,T}(\boldsymbol{\theta})$ is the corresponding log-likelihood function with respect to $\boldsymbol{\theta}$.

$$\hat{\boldsymbol{\theta}}_{p+1} = \arg\max_{\boldsymbol{\theta} \in \boldsymbol{\Theta}_{p+1}} \mathcal{L}_{n,T}(\boldsymbol{\theta}),\ \hat{\boldsymbol{\theta}}_p = \arg\max_{\boldsymbol{\theta} \in \boldsymbol{\Theta}_p} \mathcal{L}_{n,T}(\boldsymbol{\theta})$$

where $\boldsymbol{\Theta}_p = \{\boldsymbol{\theta} \in \boldsymbol{\Theta}_{p+1} : \phi_{p+1} = 0\}$. For the hypothesis test:

$$H_0 : \phi_{p+1} = 0$$

$$H_1 : \phi_{p+1} \neq 0$$

the likelihood ratio test statistic is

$$LR = -2\Big(\mathcal{L}_{n,T}(\hat{\boldsymbol{\theta}}_p) - \mathcal{L}_{n,T}(\hat{\boldsymbol{\theta}}_{p+1})\Big) \tag{3.21}$$

and, under H_0, $LR \xrightarrow{d} \chi^2(1)$ as $n, T \to \infty$.

PROOF. We prove this by following Severini's work [39, p. 113-115].

Expanding equation (3.21) by second order Taylor series around $\hat{\boldsymbol{\theta}}_{p+1}$,

$$LR = -2(\hat{\boldsymbol{\theta}}_p - \hat{\boldsymbol{\theta}}_{p+1})' \frac{\partial \mathcal{L}_{n,T}(\hat{\boldsymbol{\theta}}_{p+1})}{\partial \boldsymbol{\theta}} - (\hat{\boldsymbol{\theta}}_p - \hat{\boldsymbol{\theta}}_{p+1})' \frac{\partial^2 \mathcal{L}_{n,T}(\tilde{\boldsymbol{\theta}})}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}'} (\hat{\boldsymbol{\theta}}_p - \hat{\boldsymbol{\theta}}_{p+1})$$

where $\tilde{\boldsymbol{\theta}}$ is between $\hat{\boldsymbol{\theta}}_p$ and $\hat{\boldsymbol{\theta}}_{p+1}$. Suppose the true unknown parameter under the null is $\boldsymbol{\theta}_0 = (0, \boldsymbol{\psi}_0)'$ in the interior of $\boldsymbol{\Theta}_{p+1}$. From the property of MLE proved in chapter 2, under the null, $\hat{\boldsymbol{\theta}}_{p+1} \xrightarrow{p} \boldsymbol{\theta}_0$ so $\tilde{\boldsymbol{\theta}} \xrightarrow{p} \boldsymbol{\theta}_0$ and we have

$$\frac{\partial \mathcal{L}_{n,T}(\hat{\boldsymbol{\theta}}_{p+1})}{\partial \boldsymbol{\theta}} = 0, \quad -\frac{1}{nT} \frac{\partial^2 \mathcal{L}_{n,T}(\tilde{\boldsymbol{\theta}})}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}'} \xrightarrow{p} A(\boldsymbol{\theta}_0) \text{ as } \tilde{\boldsymbol{\theta}} \xrightarrow{p} \boldsymbol{\theta}_0$$

Hence,

$$\begin{aligned} LR =& -(\hat{\boldsymbol{\theta}}_p - \hat{\boldsymbol{\theta}}_{p+1})' \frac{\partial^2 \mathcal{L}_{n,T}(\tilde{\boldsymbol{\theta}})}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}'} (\hat{\boldsymbol{\theta}}_p - \hat{\boldsymbol{\theta}}_{p+1}) \\ \approx& \sqrt{nT}(\hat{\boldsymbol{\theta}}_p - \hat{\boldsymbol{\theta}}_{p+1})' A(\boldsymbol{\theta}_0) \sqrt{nT}(\hat{\boldsymbol{\theta}}_p - \hat{\boldsymbol{\theta}}_{p+1}) \end{aligned} \tag{3.22}$$

Then term $\hat{\boldsymbol{\theta}}_{p+1} - \hat{\boldsymbol{\theta}}_p$ is of the form $(\hat{\phi}_{p+1} - 0, \hat{\boldsymbol{\psi}}_{p+1} - \hat{\boldsymbol{\psi}}_p)'$ where $\hat{\boldsymbol{\psi}}_p$ denotes the maximum likelihood estimator of $\boldsymbol{\psi}$ for fixed $\phi_{p+1} = 0$.

If we expand the score function $\frac{\partial \mathcal{L}_{n,T}(\hat{\boldsymbol{\theta}}_{p+1})}{\partial \boldsymbol{\theta}}$ around $\boldsymbol{\theta}_0$,

$$\frac{\partial \mathcal{L}_{n,T}(\hat{\boldsymbol{\theta}}_{p+1})}{\partial \boldsymbol{\theta}} - \frac{\partial \mathcal{L}_{n,T}(\boldsymbol{\theta}_0)}{\partial \boldsymbol{\theta}} = \frac{\partial^2 \mathcal{L}_{n,T}(\boldsymbol{\theta}^*)}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}'} (\hat{\boldsymbol{\theta}}_{p+1} - \boldsymbol{\theta}_0), \quad \boldsymbol{\theta}^* \text{ is between } \hat{\boldsymbol{\theta}}_{p+1}, \boldsymbol{\theta}_0$$

Hence,

$$\begin{aligned} \sqrt{nT}(\hat{\boldsymbol{\theta}}_{p+1} - \boldsymbol{\theta}_0) =& \left[-\frac{1}{nT} \frac{\partial^2 \mathcal{L}_{n,T}(\boldsymbol{\theta}^*)}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}'} \right]^{-1} \frac{1}{\sqrt{nT}} \frac{\partial \mathcal{L}_{n,T}(\boldsymbol{\theta}_0)}{\partial \boldsymbol{\theta}} \\ \approx& A^{-1}(\boldsymbol{\theta}_0) \frac{1}{\sqrt{nT}} \frac{\partial \mathcal{L}_{n,T}(\boldsymbol{\theta}_0)}{\partial \boldsymbol{\theta}} \end{aligned}$$

Expand this equation, we have

$$\begin{pmatrix} \hat{\phi}_{p+1} - 0 \\ \hat{\boldsymbol{\psi}}_{p+1} - \boldsymbol{\psi}_0 \end{pmatrix} \approx A^{-1}(\boldsymbol{\theta}_0)\frac{1}{nT}\begin{pmatrix} \frac{\partial \mathcal{L}_{n,T}(\boldsymbol{\theta}_0)}{\partial \phi_{p+1}} \\ \frac{\partial \mathcal{L}_{n,T}(\boldsymbol{\theta}_0)}{\partial \boldsymbol{\psi}} \end{pmatrix}$$

For the model with $\phi_{p+1} = 0$ fixed, similarly,

$$(\hat{\boldsymbol{\psi}}_p - \boldsymbol{\psi}_0) \approx A^{-1}_{\boldsymbol{\psi}\boldsymbol{\psi}}(\boldsymbol{\theta}_0)\frac{1}{nT}\frac{\partial \mathcal{L}_{n,T}(\boldsymbol{\theta}_0)}{\partial \boldsymbol{\psi}}$$

where $A_{\boldsymbol{\psi}\boldsymbol{\psi}}(\boldsymbol{\theta}_0)$ is the bottom right hand corner of the $(\phi_{p+1}, \boldsymbol{\psi})$ partition of $A(\boldsymbol{\theta}_0)$. For simplicity, we define the following notation:

$$A^{-1}(\boldsymbol{\theta}_0) = \begin{pmatrix} A_{\phi_{p+1}\phi_{p+1}}(\boldsymbol{\theta}_0) & A_{\phi_{p+1}\boldsymbol{\psi}}(\boldsymbol{\theta}_0) \\ A_{\boldsymbol{\psi}\phi_{p+1}}(\boldsymbol{\theta}_0) & A_{\boldsymbol{\psi}\boldsymbol{\psi}}(\boldsymbol{\theta}_0) \end{pmatrix}^{-1} = \begin{pmatrix} A^{11}(\boldsymbol{\theta}_0) & A^{12}(\boldsymbol{\theta}_0) \\ A^{21}(\boldsymbol{\theta}_0) & A^{22}(\boldsymbol{\theta}_0) \end{pmatrix}$$

From the results above, it follows that

$$\sqrt{nT}(\hat{\boldsymbol{\psi}}_{p+1} - \hat{\boldsymbol{\psi}}_p) \approx A^{21}(\boldsymbol{\theta}_0)\frac{1}{\sqrt{nT}}\frac{\partial \mathcal{L}_{n,T}(\boldsymbol{\theta}_0)}{\partial \phi_{p+1}} + \left(A^{22}(\boldsymbol{\theta}_0) - A^{-1}_{\boldsymbol{\psi}\boldsymbol{\psi}}(\boldsymbol{\theta}_0)\right)\frac{1}{\sqrt{nT}}\frac{\partial \mathcal{L}_{n,T}(\boldsymbol{\theta}_0)}{\partial \boldsymbol{\psi}}.$$

Therefore,

$$\sqrt{nT}\begin{pmatrix} \hat{\phi}_{p+1} - 0 \\ \hat{\boldsymbol{\psi}}_{p+1} - \hat{\boldsymbol{\psi}}_p \end{pmatrix} \approx (A^{-1}(\boldsymbol{\theta}_0) - H)\frac{1}{\sqrt{nT}}\frac{\partial \mathcal{L}_{n,T}(\boldsymbol{\theta}_0)}{\partial \boldsymbol{\theta}}$$

where $H = \begin{pmatrix} 0 & 0 \\ 0 & A^{-1}_{\boldsymbol{\psi}\boldsymbol{\psi}}(\boldsymbol{\theta}_0) \end{pmatrix}$. Substituting this expression into equation (3.22) yields,

$$LR \approx \frac{1}{\sqrt{nT}}\left(\frac{\partial \mathcal{L}_{n,T}(\boldsymbol{\theta}_0)}{\partial \boldsymbol{\theta}}\right)'(A^{-1}(\boldsymbol{\theta}_0) - H)A(\boldsymbol{\theta}_0)(A^{-1}(\boldsymbol{\theta}_0) - H)\frac{1}{\sqrt{nT}}\left(\frac{\partial \mathcal{L}_{n,T}(\boldsymbol{\theta}_0)}{\partial \boldsymbol{\theta}}\right)$$

Since under the null $\frac{1}{\sqrt{nT}}\frac{\partial \mathcal{L}_{n,T}(\boldsymbol{\theta}_0)}{\partial \boldsymbol{\theta}}$ asymptotically follows $N(\mathbf{0}, A(\boldsymbol{\theta}_0))$ (refer to the proof in chapter 2), it follows that the asymptotic distribution of LR is the distribution of $Z'Z$

where Z has a multivariate normal distribution with mean 0 and covariance matrix

$$A^{1/2}(\boldsymbol{\theta}_0)(A^{-1}(\boldsymbol{\theta}_0) - H)A(\boldsymbol{\theta}_0)(A^{-1}(\boldsymbol{\theta}_0) - H)A^{1/2}(\boldsymbol{\theta}_0) \tag{3.23}$$

This covariance (3.23) can be reduced to

$$\boldsymbol{\Sigma} = A^{1/2}(\boldsymbol{\theta}_0)(A^{-1}(\boldsymbol{\theta}_0) - H)A^{1/2}(\boldsymbol{\theta}_0)$$

It is easy to show that $\boldsymbol{\Sigma} = \boldsymbol{\Sigma}\boldsymbol{\Sigma}$ since $HA(\boldsymbol{\theta}_0)H = H$, and $\text{tr}(\boldsymbol{\Sigma}) = \text{tr}(I_{dim(\boldsymbol{\theta})} - A(\boldsymbol{\theta}_0)H) = 1$.

Therefore under the null, $LR \xrightarrow{d} \chi^2(1)$ as $n, T \to \infty$. □

3.2.2. Sample ACF and PACF

Now we discuss some practical suggestion in estimating p.

Pfeifer and Deutrch [36] suggested a space time autocovariance function which describes the covariance between units lagged both in space and time. Based on their work, we can also calculate the autocorrelations of a PSTAR-ANN(p) model. Recall the PSTAR-ANN(p) model,

$$Y_t = \phi_0 W_n Y_t + \sum_{i=1}^{p} \phi_i W_n Y_{t-i} + X_t\beta + \boldsymbol{F}(X_t\boldsymbol{\gamma}')\lambda + \boldsymbol{\varepsilon}_t$$

where p is the autoregressive order and $\phi_i, i = 1, \ldots, p$ are the associated parameters. Here we only consider first order spatial weights, $W_n \in \mathbb{R}^{n\times n}$. Sample autocorrelations and partial autocorrelation functions are useful for model selection in linear time series. Theoretically, since a PSTAR-ANN(p) model is a nonlinear space-time model, ACF and PACF are not sufficient for lag identification. However, in practice, they can still give us some preliminary knowledge of the lag order p.

Assuming Y_t is stationary over both space and time, we can compute the correlation coefficient at time difference h at a single location by

$$\rho(h) = \frac{\mathbb{E}[(y_{s,t} - \mu)(y_{s,t+h} - \mu)]}{\sigma^2}$$

where μ is the mean of $\{Y_t\}$ and σ^2 is the variance of $\{Y_t\}$. This can be estimated by

$$\hat{\rho}(h) = \frac{1}{n(T-h)} \sum_{t=1}^{T-h} \frac{\sum_{s=1}^{n}(y_{s,t} - \bar{y}_t)(y_{s,t+h} - \bar{y}_{t+h})}{\hat{\sigma}_t \hat{\sigma}_{t+h}}$$

where $\bar{y}_t = \sum_{s=1}^{n} y_{s,t}$ and $\hat{\sigma}_t^2 = \frac{1}{(n-1)} \sum_{s=1}^{n}(y_{s,t} - \bar{y}_t)^2$. For the partial correlation, the 1st order partial autocorrelation equals to the 1st order autocorrelation. For h^{th} order partial autocorrelation, $h > 1$, denote the regression of Y_{t+h} on $\{Y_{t+h-1}, \ldots, Y_{t+1}\}$ as

$$\hat{Y}_{t+h} = a_0 + a_1 Y_{t+h-1} + \cdots + a_{h-1} Y_{t+1}$$

and let $\hat{Y}_t$ denote the regression of Y_t on $\{Y_{t+1}, \ldots, Y_{t+h-1}\}$

$$\hat{Y}_t = b_0 + b_1 Y_{t+1} + \cdots + b_{h-1} Y_{t+h-1}$$

Denote $E_t = Y_t - \hat{Y}_t$, $E_t = \{e_{s,t}\}_{s=1}^{n}$. Then the h^{th} order sample partial autocorrelation of Y_t denoted $\hat{\phi}_{hh}$ is

$$\hat{\phi}_{hh} = \frac{1}{(T-h)} \sum_{t=1}^{T-h} \frac{\sum_{s=1}^{n} e_{s,t+h} e_{s,t}}{\sqrt{\sum_{s=1}^{n} e_{s,t+h}^2 \sum_{s=1}^{n} e_{s,t}^2}}$$

For a linear AR(p) process, the partial autocorrelations beyond lag p equal to 0 [9]. Hence sample PACF can be useful for model selection. Figures 3.4 and 3.5 show the sample ACF and PACF of Y_t generated from model (3.24) and (3.25). $X_{1,t}, X_{2,t}$ are generated independently from $N(0, 1.5^2)$ and $N(0, 3^2)$ respectively. The error $\boldsymbol{\varepsilon}_t$ is generated independently from $N(0, 1)$. Both of these two series are stationary because the autoregressive

polynomial is stationary ($|\phi_0| + |\phi_1| + |\phi_2| < 1$).

$$Y_t = 0.6W_nY_t - 0.274W_nY_{t-1} + 0.24X_{1,t} - 0.7X_{2,t} + 2\boldsymbol{F}(0.75X_{1,t} + 0.7X_{2,t}) + \boldsymbol{\varepsilon}_t \tag{3.24}$$

$$Y_t = 0.6W_nY_t + 0.1W_nY_{t-1} + 0.25W_nY_{t-2} + 0.24X_{1,t} - 0.7X_{2,t} + 2\boldsymbol{F}(0.75X_{1,t} + 0.7X_{2,t}) + \boldsymbol{\varepsilon}_t \tag{3.25}$$

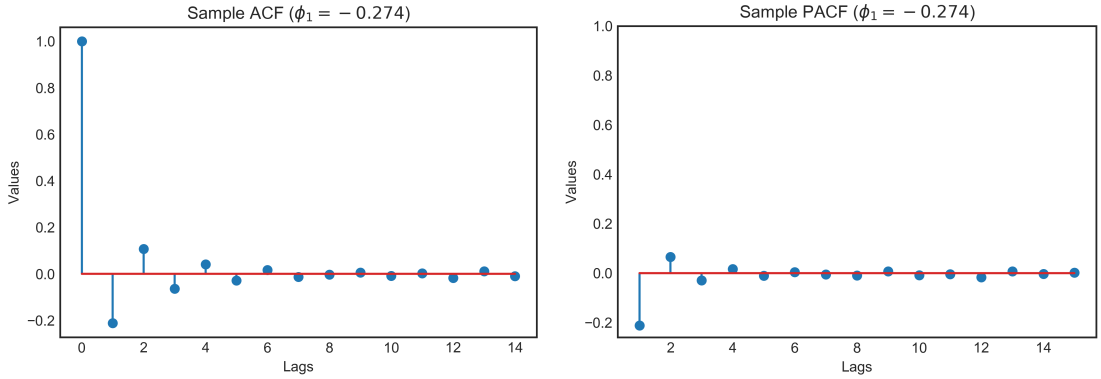


Figure 3.4. Sample ACF and PACF of model (3.24), $p = 1$

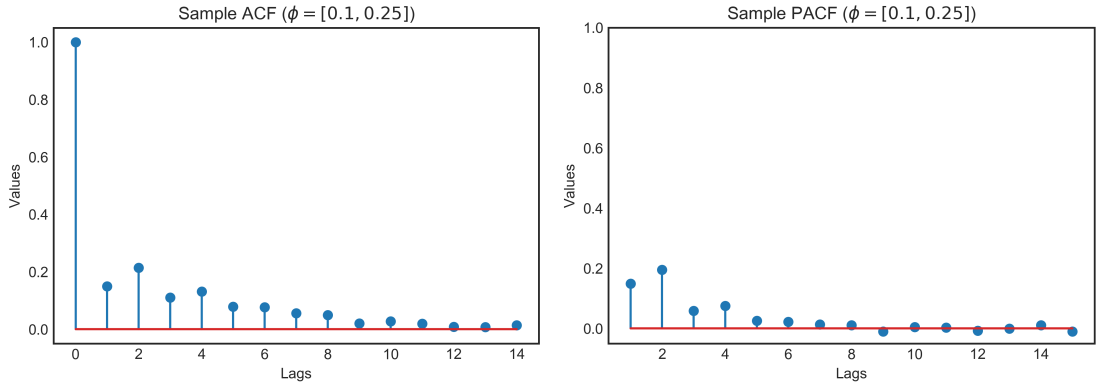


Figure 3.5. Sample ACF and PACF of model (3.25), $p = 2$

In Figures 3.4 and 3.5, the sample autocorrelations die down more gradually than the sample partial autocorrelations, which roughly cut off after time lag $p = 1$ and $p = 2$

respectively. Therefore a sample PACF can be a useful tool to obtain a rough range about the autoregressive order in a PSTAR-ANN(p) model. However, to effectively estimate the order p, we recommend using the likelihood ratio test.

3.3. Simulations and Real Example

3.3.1. LR test for lag p

In section 3, we introduce likelihood ratio test for the time lag order selection in a PSTAR-ANN(p) model and prove its asymptotic distribution. In this subsection, we are going to evaluate the asymptotic distribution of this LR test statistic derived in section 3. We generate observations Y_t from a PSTAR-ANN(1) model in a 30×30 grid, $T = 30$:

(3.26)

$$Y_t = 0.6W_nY_t - 0.274W_nY_{t-1} + 0.24X_{t,1} - 0.7X_{t,2} + 2\boldsymbol{F}(0.75X_{t,1} + 0.7X_{t,2}) + \boldsymbol{\varepsilon}_t,\ t = 1, \ldots, 30$$

where $X_{t,1}, X_{t,2}$ are generated independently from $N(0, 1.5^2)$, $N(0, 3^2)$. $\boldsymbol{\varepsilon}_t$ are generated independently from a heavy tailed distribution, where we use $t(4)$ and Laplace$(0, \frac{\sqrt{2}}{2})$ for illustration. The purpose is to test if a PSTAR-ANN(2) model is a good fit for Y_t:

$$H_0 : \phi_2 = 0$$

$$H_1 : \phi_2 \neq 0$$

We fit the simulated data on models under the null and alternative hypothesis respectively. $\hat{\boldsymbol{\theta}}_1, \hat{\boldsymbol{\theta}}_2$ are parameter estimates of a PSTAR-ANN(1) and PSTAR-ANN(2) model. The LR test statistic is given by $-2\Big(\mathcal{L}_{n,T}(\hat{\boldsymbol{\theta}}_1) - \mathcal{L}_{n,T}(\hat{\boldsymbol{\theta}}_2)\Big)$. We conduct the likelihood ratio tests for 1000 replicates and compute the test statistic in each replicate. Table 3.1 shows the percentage of rejected tests at different significance level α (critical values are based on $\chi^2(1)$) out of 1000 simulated tests with standard normal, $t(4)$ and Laplace $(0, \frac{\sqrt{2}}{2})$

error distributions respectively. We can see that the percentages approximately match the Chi-squared significance levels. Therefore we can believe that under the null, the LR test statistic roughly follows a $\chi^2(1)$ distribution when $n, T \to \infty$.

	$\chi^2(1)$	Rejected Test %		
α	Critical Value	$N(0,1)$	$t(4)$	Laplace$(0, \frac{\sqrt{2}}{2})$
0.01	6.635	0.01	0.01	0.01
0.02	5.411	0.02	0.03	0.02
0.05	3.841	0.05	0.05	0.05
0.10	2.706	0.09	0.10	0.08
0.90	0.016	0.91	0.91	0.89

Table 3.1. Percentage of rejected tests at nominal significance level α out of 1000 simulated tests where data are generated from equation (3.26) with different error densities

3.3.2. Shakeout in Linear Regression

In this subsection, we will evaluate the feature selection effect of the Shakeout in model fitting. Plenty of simulation studies of implementing Shakeout in neural networks were discussed and demonstrated its flexible effects in pruning neural networks in Kang et al. [21]. So in the following we will illustrate the Shakeout regularization in linear models on simulated datasets. Let sample size $n = 100$ and the number of predictors $q = 20$. To better simulate real world problems where we do not always observe independent predictors, we generate 20 correlated predictors $X = (X_1, \ldots, X_{20}) \in \mathbb{R}^{n \times q}$ with $X_i \sim N(0, \Sigma)$, where $\Sigma_{jk} = 0.5^{|j-k|} + 0.2I_{j \neq k}$ (this is a popular setting to generate correlated predictors [30]). We select $X_1, X_2, \ldots, X_7$ as true predictors and the other thirteen variables are not predictors. The coefficients $\beta_j, j = 1, \ldots, 7$, are shown below. Random error ε are

generated independently from $N(0, I_n)$. The dependent variable Y is given by

$$Y = 1.5X_1 + 2X_2 - 2X_3 + 2.5X_4 - 1.5X_5 + 3X_6 + 2.1X_7 + \varepsilon$$

Then we fit a linear regression model on Y with all these 20 predictors and obtain the parameter estimates $\hat{\beta}_{mle}$ by maximizing the log-likelihood function with the raw data . To compare, we also compute the parameter estimates $\hat{\beta}_{sko}$ by maximizing the log-likelihood function with the Shakeout noise injected data under different hyper-parameters τ and c. We calculate the asymptotic covariance matrix of coefficients by $(X'X)^{-1}$ and construct 95% confidence intervals for $\hat{\beta}_{mle}, \hat{\beta}_{sko}$. Predictors whose associated parameters are significant will be selected. Even though $\hat{\beta}_{sko}$ is biased, we can still construct confidence intervals with the asymptotic covariance matrix for $\hat{\beta}_{mle}$ for a reference. For evaluation metrics, we consider the precision, recall and the number of selected predictors, where precision and recall are given by,

$$\text{Precision} = \frac{\#\text{correctly selected predictors}}{\#\text{selected predictors}}$$

$$\text{Recall} = \frac{\#\text{correctly selected predictors}}{\#\text{true predictors}}$$

Table 3.2 compares L_1, L_2 norms ($||\hat{\beta}||_1, ||\hat{\beta}||_2^2$) of $\hat{\beta}_{mle}$ and $\hat{\beta}_{sko}$ and the number of significant parameters using a 95% confidence interval.

From the comparison, we can see that using Shakeout in estimating parameters impose penalty on L_1 and L_2 norms of parameter estimates, which controls the magnitude of the estimators and further assists with feature selection. Comparing τ and c in Table 3.2, fixing $c = 0$, larger values of τ, or fixing $\tau = 0.5$, larger values of c result in smaller values in $||\hat{\beta}||_1$, $||\hat{\beta}||_2^2$ with less variables selected. Hence, when τ or c is large, this penalization is more severe resulting in smaller $||\hat{\beta}||_1$, $||\hat{\beta}||_2^2$ and less predictor selected. Figure 3.6 show distribution curves for values of parameter estimates $\hat{\beta}_{mle}$ and $\hat{\beta}_{sko}$. Looking at these

Method	Hyper-parameters (τ, c)	$\|\|\hat{\beta}\|\|_1$	$\|\|\hat{\beta}\|\|_2^2$	# predictors selected	Recall	Precision
MLE	–	13.92	21.03	15	1.00	0.47
Shakeout	$(0.3, 0)$	9.18	9.75	6	0.71	0.83
	$(0.5, 0)$	8.20	6.72	7	0.86	0.86
	$(0.5, 0.5)$	3.68	2.31	4	0.57	1.00
	$(0.5, 1)$	1.70	0.51	3	0.43	1.00

Table 3.2. Compare $\hat{\beta}_{mle}$ and $\hat{\beta}_{sko}$ by $||\hat{\beta}||_1$, $||\hat{\beta}||_2^2$ and the number of predictors selected

curves, we can see that Shakeout method tends to shrink the magnitudes of parameter estimates, which agrees with our conclusion from Table 3.2.

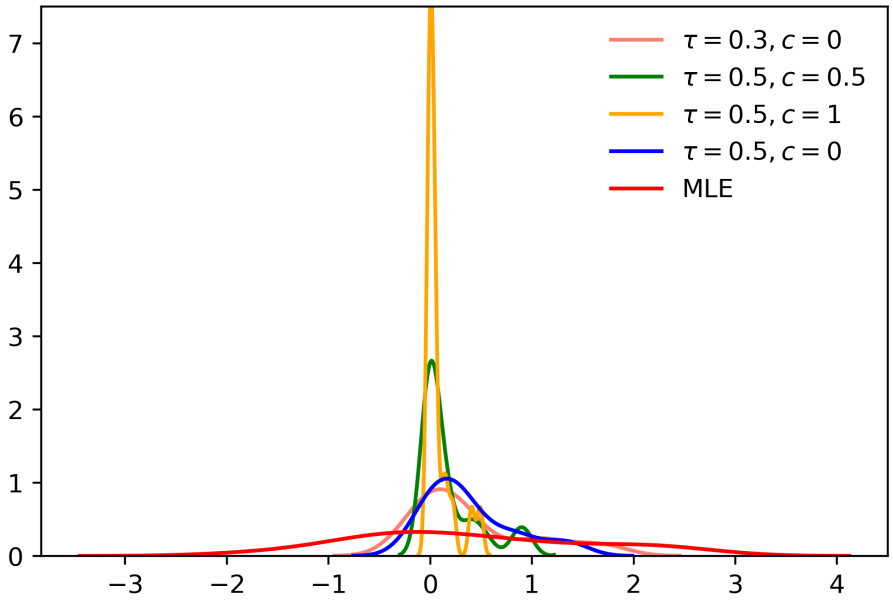


Figure 3.6. Distribution curves for values of parameter estimates $\hat{\beta}_{mle}$ and $\hat{\beta}_{sko}$ with (τ, c) settings in Table 3.2

Regarding the evaluation metrics precision and recall, we can see that when c or τ are getting larger, stronger penalization will lead to less predictors selected and under proper settings, using Shakeout can achieve higher precision and recall for example comparing $\tau = 0.5, c = 0$ and $\tau = 0.3, c = 0$. However, sometimes, too strong penalization can lower the recall ($\tau = 0.5, c = 1$). In practice, if we want to implement the Shakeout for feature selection, we need to carefully select hyper-parameters for a trade-off between different evaluation metrics for example the precision and recall. One of the practical and popular ways is to partition the training data into a training set and a validation set and to evaluate the prediction performance of Shakeout with different τ and c (refer to [21] for further reading).

3.3.3. Likelihood Ratio Test in the Election Example

In previous chapters we applied PSAR-ANN and PSTAR-ANN(1) models to the US presidential election data and investigated the spatial dynamics in the proportion of vote cast for the presidential candidates at county level in 2004. In this subsection, we will employ the likelihood ratio test to compare the two models.

Recall the election problem: the dependent variable Y_t is the fraction of votes in a county for the Democratic candidate at time point t. The number of spatial units is $n = 3107$, excluding those U.S. counties with no neighbors and the weight matrix W_n is the first order spatial matrix by queen contiguity (refer to Chapter 1 for the weight matrix definition). In previous analysis, we choose three exogenous variables to learn their relationship with Y_t: $X_{1,t}$ (`under18`) is the percent residents under 18 years old at time t; $X_{2,t}$ (`white`) is the percent white residents at time t; $X_{3,t}$ (`pctpoor`) is the percent

residents below poverty line at time t. Refer to Chapter 2, we transform $Y_t, X_{1,t}, X_{2,t}, X_{3,t}$ for analysis purposes and denote the transformed variables as $Y_t^*, X_{1,t}^*, X_{2,t}^*, X_{3,t}^*$. Let $t = 1, 2$ represent the years 2000 and 2004 respectively.

The fitted PSAR-ANN model is based on observations in 2004:

$$\begin{aligned} Y_2^* =& 0.721 W_{3107} Y_2^* + 1.693 - 0.185 X_{1,2}^* - 0.658 X_{2,2}^* + 0.181 X_{3,1}^* \\ & - 0.937 \boldsymbol{F}(1.509_{1,2}^* - 2.544 X_{2,2}^* - 2.268 X_{3,2}^*) + \varepsilon_2 \end{aligned} \tag{3.27}$$

The fitted PSTAR-ANN(1) model is based on observations in 2004 and 2000.

$$\begin{aligned} Y_2^* =& 0.425 W_{3107} Y_2^* + 0.464 W_{3107} Y_1^* - 1.173 + 0.148 X_{1,2}^* - 1.177 X_{2,2}^* - 0.153 X_{3,2}^* \\ & + 3.056 \boldsymbol{F}(-0.722 X_{1,2}^* + 1.689 X_{2,2}^* + 0.248 X_{3,2}^*) + \varepsilon_2 \end{aligned} \tag{3.28}$$

Comparing model (3.27) and (3.28), the hypothesis is:

$$H_0 : \phi_1 = 0$$

$$H_1 : \phi_1 \neq 0$$

The LR test statistics is $-2(1734.5 - 1879.35) = 289.7 \gg \chi^2_{0.95}(1)$, which indicates that including the first space time lag is significant and the model (3.28) is preferable to the model (3.27).

3.4. Future Work

In this section, we discuss some additional problems of interest and future directions of the current work. They include detecting non-stationarity in space-time processes and making improved forecasts.

3.4.1. Spatial Nonstationarity

In our work, we derive asymptotic properties of parameter estimates of a PSTAR-ANN(p) model under stationarity. If this assumption is violated, the asymptotic distributions of the parameter estimates are no longer reliable. For stationary processes, corresponding statistical properties are constant over both space and time. Non-stationarity can sometimes be detected by looking the data. Corresponding tests are also useful in practice. In the context of the PSTAR-ANN(p) model, non-stationarity exists, for instance, when $\phi_0 = 1$ or the autoregressive polynomial is non-stationary. Future extensions of this work includes tests for those cases.

To demonstrate stationarity versus non-stationarity, we simulate observations from PSTAR-ANN(1) (time lag parameter ϕ_1) and PSTAR-ANN(2) (time lag parameters ϕ_1, ϕ_2) respectively. Let $\phi_0 = 0.6$, $X_{1,t}, X_{2,t}, \boldsymbol{\varepsilon}_t$ are generated independently from $N(0, 1.5^2)$, $N(0, 3^2)$ and $N(0, 1)$ respectively. Y_t are simulated from equation (3.29) with different p and ϕ_i for $i = 1, 2$ (see Table 3.3).

(3.29)

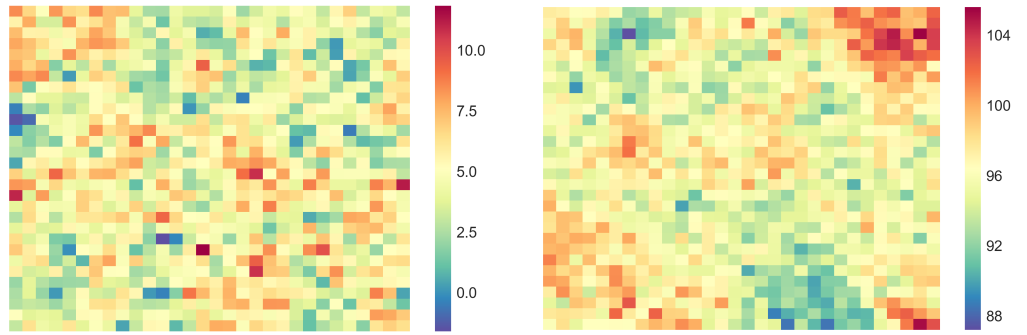
$$Y_t = 0.6W_nY_t + \sum_{i=1}^{p} \phi_i W_n Y_{t-i} + 0.24X_{1,t} - 0.7X_{2,t} + 2\boldsymbol{F}(0.75X_{1,t} + 0.7X_{2,t}) + \boldsymbol{\varepsilon}_t,\ t = 1, \ldots, 30$$

Table 3.3 exemplifies several stationary and non-stationary space-time processes. To evaluate the stationarity of a space-time process, instead of checking $|\phi_0| < 1$, we should consider the autoregressive polynomial stationarity mentioned in chapter 2 assumption 4. So we want to have that every z which solves $\det[z^p(I_n - \phi_0 W_n) - \sum_{i=1}^{p} \phi_i W_n z^{p-i}] = 0$ should lie inside a unit circle. When $\phi_0 = 0.6$ and $p = 1$, series with $\phi_1 = 0.45$ and $\phi_1 = -0.45$ are non-stationary processes; series with $\phi_1 = 0.2$ and $\phi_1 = 0.39$ are stationary.

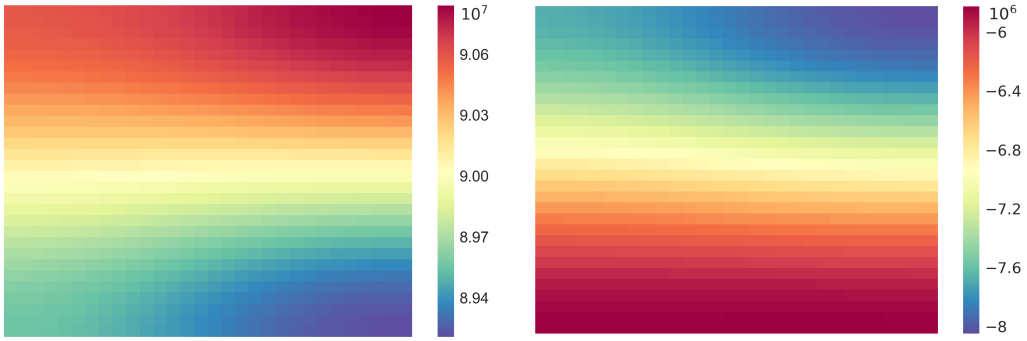
Model	Parameter	Stationary Process	Non-stationary Process
$p = 1$	ϕ_1	0.2	-0.45
		0.39	0.45
$p = 2$	ϕ_1, ϕ_2	$0.1, 0.25$	$0.1, 0.32$
		$-0.1, 0.25$	$-0.1, 0.32$

Table 3.3. Stationary versus non-stationary space time process in PSTAR-ANN(p)

Figure 3.7 display heatmaps of observations in a 30×30 grid at a single time, generated from PSTAR-ANN(1) models in Table 3.3. The color at each unit represents the value



(a) $\phi_0 = 0.6, \phi_1 = 0.2$ (b) $\phi_0 = 0.6, \phi_1 = 0.39$

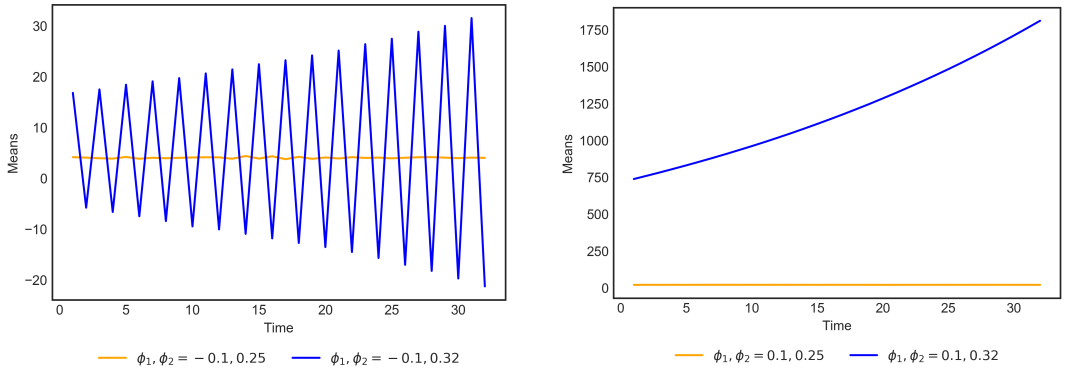


(c) $\phi_0 = 0.6,\ \phi_1 = -0.45$ (d) $\phi_0 = 0.6,\ \phi_1 = 0.45$

Figure 3.7. Simulated observations PSTAR-ANN(1) models: $(a), (b)$ shows the observations from stationary space-time process; $(c), (d)$ show those from non-stationary processes.

of observation at that location. We can clearly see that observations generated from non-stationary space-time processes exhibit clear trends.

When $\phi_0 = 0.6$ and $p = 2$, by checking the stationarity of the autoregressive polynomial (if the roots solving $(1-\phi_0\tau_i)z^2-\phi_1\tau_i z-\phi_2\tau_i = 0$ are in the unit circle for all τ_i, where τ_i are eigenvalues of W_n), series with $(\phi_1, \phi_2) = (-0.1, 0.32)$ and $(\phi_1, \phi_2) = (0.1, 0.32)$ are non-stationary processes while series with $(\phi_1, \phi_2) = (0.1, 0.25)$ and $(\phi_1, \phi_2) = (-0.1, 0.25)$ are both stationary. Figure 3.8 shows the trend of $\bar{Y}_t = \frac{1}{n}\sum_{s=1}^{n} y_{s,t}$ over the time t, where Y_t are generated from PSTAR-ANN(2) models in Table 3.3. It is clear to see that $\bar{Y}_t$ increases exponentially with time t when $(\phi_1, \phi_2) = (0.1, 0.32)$ or increases with high fluctuation when $(\phi_1, \phi_2) = (-0.1, 0.32)$. Compared to the stationary cases, the high fluctuation or large variance of a space time series over the time can be an alert for the non-stationarity.



(a) $\phi_1, \phi_2 = -0.1, 0.25$ is stationary; $\phi_1, \phi_2 = -0.1, 0.32$ is non-stationary

(b) $\phi_1, \phi_2 = 0.1, 0.25$ is stationary; $\phi_1, \phi_2 = 0.1, 0.32$ is non-stationary

Figure 3.8. Means of Y_t versus time t, where Y_t are generated from PSTAR(2)-ANN models with different parameter ϕ_1, ϕ_2

From the comparison shown above, often nonstationary spatial processes, with moments that are not constant, can be detected in practice. Non-stationarity can make it

harder to make predictions. In this sense, it is important to check the space and space-time stationarity of the variable of interest. In the future, we can research more in identifying non-stationary space-time processes.

3.4.2. Spatial Correlation ϕ_0 Localization

In our proposed model, we assume the uniform space correlation ϕ_0 and time lag autoregressive parameters $\phi_i, i = 1, \ldots, p$ for all space units. In practice, this assumption may be violated. For instance in the election example, if we take a closer look into counties in a single state, we can find that the dependence structure can be different. Figure 3.9 shows the fractions of vote-shares per county for 2004 Democratic presidential candidate in Texas and Illinois. Most counties in Texas are positively correlated with their neighbors while few counties near the south corner have negative correlations with some neighbors. On right side, spatial correlations in counties in Illinois looks more consistent than those in Texas.

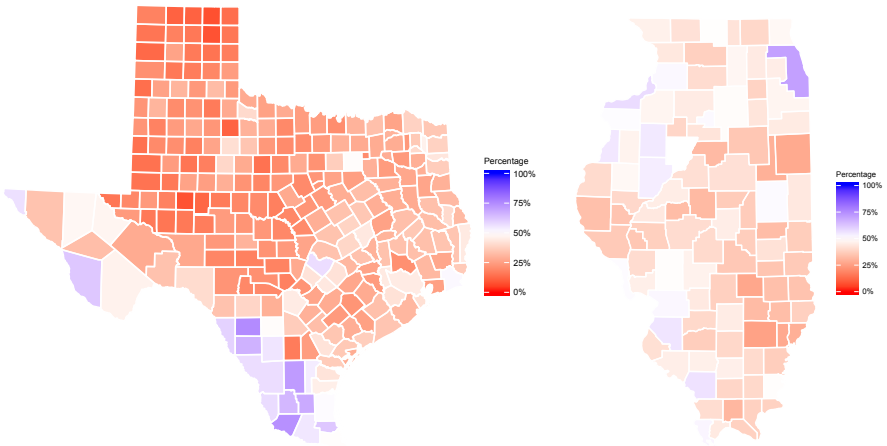


Figure 3.9. Fractions of vote-shares per county for 2004 Democratic presidential candidate in Texas (left) and Illinois (right)

To accommodate this localized spatial dependence, one approach is to use geographically weighted regression (GWR) to estimate $\phi_{i,k}$ for a location k [10], [15]. Brunsdon *et al* [10] and Fotheringham *et al* [15] discussed this calibration in a simple spatial autoregressive model and we can extend their work to a PSTAR-ANN(p) model. Suppose we want to estimate the spatial autoregressive parameter $\phi_{i,k}$ at a sample location k, the technique is to weight the data according to the geographical location with respect to k. We define a diagonal matrix D_k where the diagonals are given by a monotone increasing function of distance between the sample location k and other locations. So different matrix D_k will be initialized based on the geographical properties of the location k. Referring to Brunsdon [11], a PSTAR-ANN(p) model is now modified as,

$$Y_t = \sum_{i=0}^{p} \phi_{i,k} W_n Y_{t-i} + X_t \beta + \boldsymbol{F}(X_n \boldsymbol{\gamma}')\lambda + D_k \boldsymbol{\varepsilon}_t \tag{3.30}$$

Looking at the term $D_k\boldsymbol{\varepsilon}_t$, this modification is equivalent to scaling the identically distributed errors $\boldsymbol{\varepsilon}_t$ by a distance-based function. Also, in the maximum likelihood estimation, observations which have larger variances will be downweighted. So in order to put more weight on locations closer to k, the entries in D_k should be proportional to the distance from k which means that observations in closer neighborhoods around the sample point k have smaller variances [10] and have more weight in the likelihood function. Then the error term in (3.30) can be rewritten as

$$\boldsymbol{\varepsilon}_t = U_k(I_n - \phi_{0,k} W_n) Y_t - \sum_{i=1}^{p} \phi_{i,k} U_k W_n Y_{t-p} - U_k X_t \beta - U_k \boldsymbol{F}(X_t \boldsymbol{\gamma}')\lambda \tag{3.31}$$

where $U_k D_k = I_n$. Therefore the conditional log-likelihood function of the model (3.30) becomes

$$\mathcal{L}_{GW-PSTAR}(\boldsymbol{\theta}) = T \ln |U_k(I_n - \phi_{0,k} W_n)| - \sum_{t=1}^{T} \sum_{s=1}^{n} \ln f(\varepsilon_{s,t}(\boldsymbol{\theta}))$$

Brunsdon [11] provided a convenient approach to re-estimate the spatial autoregressive parameters at any sample point k and his method allowed flexibility in defining the distance-based matrix D_k. To apply this in practice, because it can be expensive to estimate spatial autoregressive parameters for every location, we need to have a good understanding of the geographic property in the space and carefully choose the representative sample locations to estimate their local spatial correlations.

Another method is to use a generalized PSTAR-ANN(p) model referred in Borovkova's work [6]. A generalized PSTAR-ANN(p) is defined as

$$Y_t = \sum_{i=0}^{p} \Phi_i W_n Y_{t-i} + X_t\beta + \boldsymbol{F}(X_t\boldsymbol{\gamma}')\lambda + \boldsymbol{\varepsilon}_t, \quad t = 1, \ldots, T \tag{3.32}$$

where $\Phi_i = \text{diag}(\phi_{i1}, \ldots, \phi_{in})$ for $i = 0, 1, \ldots, p$, containing local autoregressive parameters and W_n is the first order spatial weight matrix.

Borovkova [6] proposed least square estimators in his generalized space-time autoregressive model where he let $\Phi_0 = 0$ in equation (3.32) and derived the asymptotic property of the least square estimates when $T \to \infty$. However, In our case, $\Phi_0 \neq 0$ and we can use maximum likelihood estimation to estimate parameters. Note that we allow $n, T \to \infty$ in a PSTAR-ANN(p) model so due to model identification, we should restrict the rank of Φ_i bounded by some constant $d_i < n$ for $i = 0, \ldots, p$. Then the log-likelihood function of model (3.32) should be:

$$\mathcal{L}_{GPSTAR}(\boldsymbol{\theta}) = T \ln |I_n - \Phi_0 W_n)| - \sum_{t=1}^{T}\sum_{s=1}^{n} \ln f(\varepsilon_{s,t}(\boldsymbol{\theta})) \tag{3.33}$$

$$\varepsilon_{s,t}(\boldsymbol{\theta}) = (I_n - \Phi_0 W_n)Y_t - \sum_{i=1}^{p} \Phi_i W_n Y_{t-i} - X_t\beta - \boldsymbol{F}(X_t\boldsymbol{\gamma}')\lambda$$

Based on the discussion above, there are two approaches that can be applied to consider local spatial correlations in a PSTAR-ANN(p). Further work is needed to find and prove

the limiting distributions of maximum likelihood estimators of parameters in model (3.30) and (3.32). And the problem of having only local stationarity will be considered in the future.

3.4.3. Prediction for the Election Problem

In the election problem, due to the limited access to exogenous variables associated with social factors, we fitted PSTAR-ANN(1) and PSTAR-ANN(2) models using the data observed in year 2000 and 2004. Theoretically the limiting distribution of parameter estimates in a PSTAR-ANN(p) model are derived assuming $n, T \to \infty$. So fitting the model using the dataset with two time points does not exactly satisfy the condition that $T \to \infty$. In chapter 2, even though histograms of Y_t and X_t indicate stationarity over the time, we still suggest including more time points in the analysis to obtain reliable results. In chapter 2, analysis on the parameter estimates suggested that some exogenous variables such as `UNDER18` and `pctpoor` are not great predictors. To improve this, we should also consider finding other measures for age and social economic level. In addition, due to the limited time points in the data, we are not able to perform the variable selection using the Shakeout method because the dataset can not be split into training and test sets.

On the other hand, regarding to the data collection, we can hardly guarantee that exogenous variables included in the analysis are collected in historical records. Some variables recorded at time t may be only estimators based on previous years or they are replaced by other variables. This uncertainty makes it difficult in data collection so we should pay more attention in the future when selecting exogenous variables for the analysis.

To conclude, in the future, we would like to collect more data observed at different time points. With more observations and exogenous variables over the time , we can be more confident in the fitted parameter estimates. In the meanwhile we are able to employ the Shakeout method to discover better exogenous variables to explain the spatial dynamic in the election problem.

Bibliography

[1] Chunrong Ai and Xiaohong Chen. Efficient estimation of models with conditional moment restrictions containing unknown functions. *Econometrica*, 71(6):1795–1843, 2003.

[2] Beth Andrews, Richard A Davis, and F Jay Breidt. Maximum likelihood estimation for all-pass time series models. *Journal of Multivariate Analysis*, 97(7):1638–1659, 2006.

[3] Luc Anselin. *Spatial econometrics: methods and models*, volume 4. Springer Science & Business Media, 2013.

[4] George D Birkhoff. Proof of the ergodic theorem. *Proceedings of the National Academy of Sciences*, 17(12):656–660, 1931.

[5] Roger S Bivand, Edzer J Pebesma, Virgilio Gomez-Rubio, and Edzer Jan Pebesma. *Applied spatial data analysis with R*, volume 747248717. Springer, 2008.

[6] Svetlana Borovkova, Hendrik P Lopuhaä, and Budi Nurani Ruchjana. Consistency and asymptotic normality of least squares estimators in generalized star models. *Statistica Neerlandica*, 62(4):482–508, 2008.

[7] Dan Braha and Marcus A. M. de Aguiar. Voting contagion: Modeling and analysis of a century of u.s. presidential elections. *PLOS ONE*, 12(5):1–30, 05 2017. doi: 10.1371/journal.pone.0177970. URL `https://doi.org/10.1371/journal.pone.0177970`.

[8] F Jay Breid, Richard A Davis, Keh-Shin Lh, and Murray Rosenblatt. Maximum likelihood estimation for noncausal autoregressive processes. *Journal of Multivariate Analysis*, 36(2):175–198, 1991.

[9] Peter J Brockwell, Richard A Davis, and Matthew V Calder. *Introduction to time series and forecasting*, volume 2. Springer, 2002.

[10] Chris Brunsdon, A Stewart Fotheringham, and Martin E Charlton. Geographically weighted regression: a method for exploring spatial nonstationarity. *Geographical*

analysis, 28(4):281–298, 1996.

[11] Chris Brunsdon, A Stewart Fotheringham, and Martin Charlton. Spatial nonstationarity and autoregressive models. *Environment and Planning A*, 30(6):957–973, 1998.

[12] Richard H Byrd, Peihuang Lu, Jorge Nocedal, and Ciyou Zhu. A limited memory algorithm for bound constrained optimization. *SIAM Journal on Scientific Computing*, 16(5):1190–1208, 1995.

[13] George Casella and Roger L Berger. *Statistical inference*, volume 2. Duxbury Pacific Grove, CA, 2002.

[14] Noel Cressie. *Statistics for spatial data*. John Wiley & Sons, 2015.

[15] A Stewart Fotheringham, Martin Charlton, and Christopher Brunsdon. Two techniques for exploring non-stationarity in geographical data. *Geographical Systems*, 4 (1):59–82, 1997.

[16] A Ronald Gallant and Halbert White. On learning the derivatives of an unknown mapping with multilayer feedforward networks. *Neural Networks*, 5(1):129–138, 1992.

[17] Semyon Aranovich Gershgorin. Uber die abgrenzung der eigenwerte einer matrix. *Bulletin de l'Académie des Sciences de l'URSS. Classe des sciences mathématiques et na*, (6):749–754, 1931.

[18] Arthur Getis. Cliff, ad and ord, jk 1973: Spatial autocorrelation. london: Pion. *Progress in Human Geography*, 19(2):245–249, 1995.

[19] Kurt Hornik, Maxwell Stinchcombe, and Halbert White. Universal approximation of an unknown mapping and its derivatives using multilayer feedforward networks. *Neural networks*, 3(5):551–560, 1990.

[20] JT Gene Hwang and A Adam Ding. Prediction intervals for artificial neural networks. *Journal of the American Statistical Association*, 92(438):748–757, 1997.

[21] Guoliang Kang, Jun Li, and Dacheng Tao. Shakeout: A new approach to regularized deep neural network training. *IEEE transactions on pattern analysis and machine intelligence*, 40(5):1245–1258, 2018.

[22] Harry H Kelejian and Ingmar R Prucha. A generalized spatial two-stage least squares procedure for estimating a spatial autoregressive model with autoregressive disturbances. *The Journal of Real Estate Finance and Economics*, 17(1):99–121, 1998.

[23] Harry H Kelejian and Ingmar R Prucha. A generalized moments estimator for the autoregressive parameter in a spatial model. *International economic review*, 40(2): 509–533, 1999.

[24] Harry H Kelejian and Ingmar R Prucha. Specification and estimation of spatial autoregressive models with autoregressive and heteroskedastic disturbances. *Journal of Econometrics*, 157(1):53–67, 2010.

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

[26] J Pace LeSage et al. Introduction to spatial econometrics. Technical report, CRC Press, 2009.

[27] James P LeSage et al. Spatial econometrics, 1999.

[28] Yinan Li and Fang Liu. Whiteout: Gaussian adaptive noise regularization in deep neural networks. *arXiv preprint arXiv:1612.01490*, 2016.

[29] Keh-Shin Lii and Murray Rosenblatt. An approximate maximum likelihood estimation for non-gaussian non-minimum phase moving average processes. *Journal of Multivariate Analysis*, 43(2):272–299, 1992.

[30] Antonio R Linero. Bayesian regression trees for high-dimensional prediction and variable selection. *Journal of the American Statistical Association*, 113(522):626–636, 2018.

[31] Kiyotoshi Matsuoka. Noise injection into inputs in back-propagation learning. *IEEE Transactions on Systems, Man, and Cybernetics*, 22(3):436–440, 1992.

[32] Marcelo C Medeiros, Timo Teräsvirta, and Gianluigi Rech. Building neural network models for time series: a statistical approach. *Journal of Forecasting*, 25(1):49–75, 2006.

[33] M.I.Gordin. The central limit theorem for stationary processes. *Soviet. Math. Dokl.*, 10:1174–1176, 1969.

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

[35] J Paelinck. Spatial econometrics. *Economics Letters*, 1(1):59–63, 1978.

[36] Phillip E Pfeifer and Stuart Jay Deutrch. A three-stage iterative procedure for space-time modeling phillip. *Technometrics*, 22(1):35–47, 1980.

[37] Keith T Poole and Howard Rosenthal. U.s. presidential elections 1968-80: A spatial analysis. *American Journal of Political Science*, pages 282–312, 1984.

[38] Thomas J Rothenberg. Identification in parametric models. *Econometrica: Journal of the Econometric Society*, pages 577–591, 1971.

[39] Thomas A Severini. *Likelihood methods in statistics*. Oxford University Press, 2000.

[40] Oleg Smirnov and Luc Anselin. Fast maximum likelihood estimation of very large spatial autoregressive models: a characteristic polynomial approach. *Computational Statistics & Data Analysis*, 35(3):301–319, 2001.

[41] Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. Dropout: a simple way to prevent neural networks from overfitting. *The Journal of Machine Learning Research*, 15(1):1929–1958, 2014.

[42] Ben Stabler. *shapefiles: Read and Write ESRI Shapefiles*, 2013. URL `https://CRAN.R-project.org/package=shapefiles`. R package version 0.7.

[43] J Michael Steele. *The Cauchy-Schwarz master class: an introduction to the art of mathematical inequalities*. Cambridge University Press, 2004.

[44] Liangjun Su and Sainan Jin. Profile quasi-maximum likelihood estimation of partially linear spatial autoregressive models. *Journal of Econometrics*, 157(1):18–33, 2010.

[45] Olga Taussky. A recurring theorem on determinants. *The American Mathematical Monthly*, 56(10P1):672–676, 1949.

[46] Philip A Viton. Notes on spatial econometric models. *City and regional planning*, 870(03):9–10, 2010.

[47] Halbert White. Parametric statistical estimation with artificial neural networks: A condensed discussion. *From statistics to neural networks: theory and pattern recognition applications*, 136:127, 1994.

[48] Halbert White. *Estimation, inference and specification analysis*. Number 22. Cambridge university press, 1996.

[49] Halbert White. *Asymptotic theory for econometricians*. Academic press, 2014.

[50] Qiwei Yao, Peter J Brockwell, et al. Gaussian maximum likelihood estimation for arma models ii: spatial processes. *Bernoulli*, 12(3):403–429, 2006.

[51] Yuan-qing Zhang and Guang-ren Yang. Statistical inference of partially specified spatial autoregressive model. *Acta Mathematicae Applicatae Sinica, English Series*, 31(1):1–16, 2015.

[52] Ciyou Zhu, Richard H Byrd, Peihuang Lu, and Jorge Nocedal. Algorithm 778: L-bfgs-b: Fortran subroutines for large-scale bound-constrained optimization. *ACM Transactions on Mathematical Software (TOMS)*, 23(4):550–560, 1997.