Analysis of Spatial Data

Xiang Zhang
University of Kentucky, xiang.zhang0206@gmail.com

Recommended Citation
https://uknowledge.uky.edu/statistics_etds/4

This Doctoral Dissertation is brought to you for free and open access by the Statistics at UKnowledge. It has been accepted for inclusion in Theses and Dissertations--Statistics by an authorized administrator of UKnowledge. For more information, please contact UKnowledge@lsv.uky.edu.
STUDENT AGREEMENT:

I represent that my thesis or dissertation and abstract are my original work. Proper attribution has been given to all outside sources. I understand that I am solely responsible for obtaining any needed copyright permissions. I have obtained and attached hereto needed written permission statements(s) from the owner(s) of each third-party copyrighted matter to be included in my work, allowing electronic distribution (if such use is not permitted by the fair use doctrine).

I hereby grant to The University of Kentucky and its agents the non-exclusive license to archive and make accessible my work in whole or in part in all forms of media, now or hereafter known. I agree that the document mentioned above may be made available immediately for worldwide access unless a preapproved embargo applies.

I retain all other ownership rights to the copyright of my work. I also retain the right to use in future works (such as articles or books) all or part of my work. I understand that I am free to register the copyright to my work.

REVIEW, APPROVAL AND ACCEPTANCE

The document mentioned above has been reviewed and accepted by the student’s advisor, on behalf of the advisory committee, and by the Director of Graduate Studies (DGS), on behalf of the program; we verify that this is the final, approved version of the student’s dissertation including all changes required by the advisory committee. The undersigned agree to abide by the statements above.

Xiang Zhang, Student

Dr. Mai Zhou, Major Professor

Dr. Constance Wood, Director of Graduate Studies
ANALYSIS OF SPATIAL DATA

DISSERTATION

A dissertation submitted in partial fulfillment of the requirements for the degree of Doctor of Philosophy in the College of Arts and Sciences at the University of Kentucky

By
Xiang Zhang
Lexington, Kentucky

Director: Dr. Yanbing Zheng and Dr. Mai Zhou, Professor of Statistics
Lexington, Kentucky

2013

Copyright© Xiang Zhang 2013
ABSTRACT OF DISSERTATION

ANALYSIS OF SPATIAL DATA

In many fields of the agriculture, biological, physical and social sciences, spatial lattice data are becoming increasingly common. In addition, a large amount of lattice data shows not only visible spatial pattern but also temporal pattern (see, Zhu et al. 2005). An interesting problem is to develop a model to systematically model the relationship between the response variable and possible explanatory variable, while accounting for space and time effect simultaneously.

Spatial-temporal linear model and the corresponding likelihood-based statistical inference are important tools for the analysis of spatial-temporal lattice data. We propose a general asymptotic framework for spatial-temporal linear models, namely increasing domain asymptotics, infill asymptotics and hybrid asymptotics and investigate the property of maximum likelihood estimates under such framework. Mild regularity conditions on the spatial-temporal weight matrices will be put in order to derive the asymptotic properties (consistency and asymptotic normality) of maximum likelihood estimates. A simulation study is conducted to examine the finite-sample properties of the maximum likelihood estimates.

For spatial data, aside from traditional likelihood-based method, a variety of literature has discussed Bayesian approach to estimate the correlation (autocovariance function) among spatial data, especially Zheng et al. (2010) proposed a nonparametric Bayesian approach to estimate a spectral density of a random field. We will also discuss nonparametric Bayesian approach in analysing spatial data. We will propose a general procedure for constructing a multivariate Feller prior and establish its theoretical property as a nonparametric prior. A blocked Gibbs sampling algorithm is also proposed for computation since the posterior distribution is analytically manageable.

In summary, the main contributions of this dissertation are

- The development of simultaneous spatial-temporal autoregressive model and validate the asymptotic property of maximum likelihood estimates under such model.
• The establishment of a new nonparametric Bayesian prior and corresponding computational algorithm.

KEYWORDS: Autoregressive models; Spatial-temporal process; Multivariate Feller prior; Blocked Gibbs sampling
ANALYSIS OF SPATIAL DATA

By
Xiang Zhang

Director of Dissertation: Dr. Yanbing Zheng and Dr. Mai Zhou
Director of Graduate Studies: Dr. Constance Wood

Date: May 07, 2012
I would like to thank my parents, for their love, patience and support during my whole life. They gave me my life, raised me up so that I can have the opportunity to live, to learn, and to understand my own life. I wish they could be here with me and enjoy the moment that their son finally earn his Ph.D. degree in statistics.

I would like to express my deep appreciations and gratitude to my advisors, Dr. Yanbing Zheng and Dr. Mai Zhou, for their time, support, and mentorship during my research. They helped me to develop my research ability, creative thinking, and analytical skills. Without them, I cannot even approach the myth of statistical science and I would still wander outside the door as I used to be. They opened the door for me so that I can enjoy the beauty of statistics inside now.

I would like to acknowledge my committee member and former Director of Graduate Studies, Dr. Arne Bathke, for his friendly guidance, thoughtful provoking suggestions during my stay in University of Kentucky.

I would also like to thank other committee members, Dr. Simon Bonner and Dr. Qiang Ye for their valuable comments on my dissertation.

Many thanks to Dr. Arnold J. Stromberg, Dr. Constance Wood and Dr. William Griffith, for the excellent statistical background they provided in teaching and research. They are always willing to help student whenever needed.

I am very grateful to my friends, Yifan Yang, Zilong Wang, Weining Shen, Xianyang
Zhang and Gongjun Xu, for the help they provided during my dissertation research.
Sincere thanks to all of my friends who have made my life meaningful. It is impossible
to list all of your names here but you are always in my mind.

My appreciations also go to the Department of Statistics, which provided me with a
good environment to study and research.
TABLE OF CONTENTS

Acknowledgments ............................................................... iii
Table of Contents .............................................................. v
List of Figures ................................................................. vii
List of Tables ................................................................. viii

Chapter 1  Introduction ....................................................... 1
  1.1  From Ideal Data to Realistic Data .................................. 1
  1.2  Spatial Data ................................................................ 2

Chapter 2  Spatial-Temporal Lattice Modelling and Maximum Likelihood Estimation ...................................................... 6
  2.1  Overview .................................................................. 6
  2.2  Spatial-Temporal Linear Model ...................................... 7
  2.3  Maximum Likelihood Estimations .................................... 9
  2.4  Asymptotic Properties ................................................ 10
  2.5  Simulation .................................................................. 17
  2.6  Conclusions and Discussion ......................................... 22

Chapter 3  Multivariate Feller Priors ....................................... 23
  3.1  Motivation .................................................................. 23
  3.2  Spectral Density Estimation .......................................... 24
  3.3  Multivariate Feller Operators ....................................... 26
  3.4  Multivariate Feller Prior .............................................. 33
  3.5  Application of Multivariate Feller Priors ......................... 35
  3.6  Discussion and Future Work ......................................... 50

Appendices ........................................................................ 53
  Appendix A: Proof of Theorem 1 ......................................... 53
  Appendix B: Proof of Theorem 2 ......................................... 58
  Appendix C: Proof of Theorem 3 ......................................... 72
  Appendix D: Proof of Proposition 1 ..................................... 73
  Appendix E: Proof of Proposition 2 ..................................... 74
  Appendix F: Proof of Proposition 3 ..................................... 75
  Appendix G: Proof of Proposition 4 ..................................... 77
  Appendix H: Proof of Theorem 4 ......................................... 79
  Appendix I: Proof of Theorem 5 ......................................... 81

Bibliography ....................................................................... 84
LIST OF FIGURES

<table>
<thead>
<tr>
<th>Figure Number</th>
<th>Description</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>2.1</td>
<td>Increasing Domain</td>
<td>11</td>
</tr>
<tr>
<td>2.2</td>
<td>Infill Domain</td>
<td>12</td>
</tr>
<tr>
<td>2.3</td>
<td>Hybrid Domain</td>
<td>12</td>
</tr>
<tr>
<td>3.1</td>
<td>Trace of k</td>
<td>42</td>
</tr>
<tr>
<td>3.2</td>
<td>Density function estimate using Feller-Dirichlet(2) prior for data from bivariate normal. The red curve on 3 graphs are contour plot of true density, contour plot of estimated density are (i) S=2, black curve (ii) S=5, green curve (iii) S=10, blue curve.</td>
<td>45</td>
</tr>
<tr>
<td>3.3</td>
<td>Density function estimate using Feller-PY(0.25) prior for data from bivariate normal. The red curve on 3 graphs are contour plot of true density, contour plot of estimated density are (i) S=2, black curve (ii) S=5, green curve (iii) S=10, blue curve.</td>
<td>45</td>
</tr>
<tr>
<td>3.4</td>
<td>Density function estimate using Feller-Dirichlet(2) prior for data from bivariate logistic. The red curve on 3 graphs are contour plot of true density, contour plot of estimated density are (i) S=2, black curve (ii) S=5, green curve (iii) S=10, blue curve.</td>
<td>46</td>
</tr>
<tr>
<td>3.5</td>
<td>Density function estimate using Feller-PY(0.25) prior for data from bivariate logistic. The red curve on 3 graphs are contour plot of true density, contour plot of estimated density are (i) S=2, black curve (ii) S=5, green curve (iii) S=10, blue curve.</td>
<td>46</td>
</tr>
<tr>
<td>3.6</td>
<td>Bivariate Kernel Density function estimate for data from bivariate normal. The green curve on is contour plot of true density, contour plot of estimated density are black curve</td>
<td>47</td>
</tr>
<tr>
<td>3.7</td>
<td>Bivariate Kernel Density function estimate for data from bivariate logistic. The green curve on is contour plot of true density, contour plot of estimated density is black curve</td>
<td>48</td>
</tr>
<tr>
<td>3.8</td>
<td>Density estimate using Feller-Dirichlet(2) prior for data iris. Contour plot of estimated density are (i) S=2, black curve (ii) S=5, green curve (iii) S=10, blue curve.</td>
<td>49</td>
</tr>
<tr>
<td>3.9</td>
<td>Density estimate using Feller-PY(0.25) prior for data iris. Contour plot of estimated density are (i) S=2, black curve (ii) S=5, green curve (iii) S=10, blue curve.</td>
<td>49</td>
</tr>
<tr>
<td>Table</td>
<td>Description</td>
<td>Page</td>
</tr>
<tr>
<td>-------</td>
<td>------------------------------------------------------------------------------------------------------</td>
<td>------</td>
</tr>
<tr>
<td>2.1</td>
<td>Means and standard deviations (SD) of maximum likelihood estimates (MLE) of the model parameters based on 100 simulated data. Here the lattice size is $4 \times 4$ and $8 \times 8$ with varying sub-lattice sizes $1 \times 1$, $2 \times 2$ and $4 \times 4$ within each cell of the lattice, and the number of time points is 2.</td>
<td>19</td>
</tr>
<tr>
<td>2.2</td>
<td>Means and standard deviations (SD) of maximum likelihood estimates (MLE) of the model parameters based on 100 simulated data. Here the lattice size is $4 \times 4$ and $8 \times 8$ with varying sub-lattice sizes $1 \times 1$, $2 \times 2$ and $4 \times 4$ within each cell of the lattice, and the number of time points is 5.</td>
<td>20</td>
</tr>
<tr>
<td>2.3</td>
<td>Means and standard deviations (SD) of maximum likelihood estimates (MLE) of the model parameters based on 100 simulated data. Here the lattice size is $4 \times 4$ and $8 \times 8$ with varying sub-lattice sizes $1 \times 1$, $2 \times 2$ and $4 \times 4$ within each cell of the lattice, and the number of time points is 10.</td>
<td>21</td>
</tr>
<tr>
<td>3.1</td>
<td>Hellinger error of estimated density functions using Feller-Dirichlet prior and Feller-PY prior for data from bivariate normal distribution and bivariate logistic distribution</td>
<td>44</td>
</tr>
</tbody>
</table>
Chapter 1 Introduction

1.1 From Ideal Data to Realistic Data

As known to all, start learning statistics always assume that the observations are taken from identical conditions and that each observation is drawn independently of the others. Then the observations form a random sample, i.e., they are independent and identically distributed. Standard statistical methods can be applied to such "ideal" data to build a statistical model and to estimate the model’s parameters, e.g., linear model and maximum likelihood estimates. The question is: can we treat all kinds of data as ”ideal”? I think the answer is NO. Think about the following scenario: during an experiment, we measure each experimental unit repeatedly and record those measurements. The data should be obviously combined in some way since we measure the same physical constant over and over. Thus, it is unfair to always assume the data we meet in the real world are ”ideal”, and miss-specifying the data type may lead to questionable statistical results. So if not all data are ”ideal”, what is ”realistic” data? Well, it contains, but not limited to, dependent data, inhomogeneous data, or even worse, dependent and inhomogeneous data.

Independence is a very common assumption in statistical analysis, especially in those courses I have learned during the first two years of my graduate study. What can I say? Independence makes things easier! Under independent assumption, plenty of mathematical and statistical methodologies can be used, e.g., Law of large numbers, Central Limit Theorem, and etc. Those techniques are of great use in dealing with problems such as establishing asymptotic properties of a certain parameter estimate, conducting appropriate statistical test and constructing confidence intervals. However, as we discussed above, dependent data is very common in the real-world
problem, so models involving statistical dependence are often more reliable and useful.

Lack of homogeneity in data is another important issue. It is usually accounted for a nonconstant-mean assumption in statistical analysis. In most cases, the nonconstant-mean is described as a linear combination of potential explanatory variables. In some other cases, data is assumed to be independent realization from distributions whose mean are constant but whose variance differ markedly. The above two scenarios (nonconstant-mean or non-constant variance) shows the necessity of relaxing the identically-distributed assumption.

Now we can reach the agreement that not all data are ”ideal”, and we need more general (therefore more complicated) models in our analysis. However, someone, statistician or scientist, may have this question: are these general models of any scientific meaning? If not, why we abandon those nice assumptions and bring troubles to ourselves in statistical analysis by inducing more general models? Well, in the pages to follow, I want to briefly introduce spatial data so that you can see the benefits of using general models.

1.2 Spatial Data

What is spatial data? Simply speaking, if the data are close together in time or space, then it is natural to treat these data correlated rather than independent. For data close together in time, we usually consider time series models (purely temporal models). These models are based on identically distributed observations that are dependent and occur at equally spaced time points, e.g., Autoregressive models, Moving Average Models, Autoregressive Moving Average models. For data close together in space, their dependence is present in all directions and becomes weaker as data locations become more dispersed. Statisticians were aware of spatial dependence for
a long time. For example, R. A. Fisher was clearly aware of spatial dependence in agriculture field experiments because he established the principles of randomization, blocking and replication to neutralize the effect of spatial correlation as well as controlling bias.

Two basic concepts in modeling spatial data: spatial locations $s_1, \ldots, s_n$ and data $z(s_1), \ldots, z(s_n)$ observed at those locations. Now we can give a general definition of spatial data: let $s$ be a general data location in $d$-dimensional Euclidean space and suppose the potential datum $Z(s)$ at spatial location $s$ is a random quantity. Let $s$ vary over set $D \in \mathbb{R}^d$ so as to the generate the multivariate random field

$$\{Z(s) : s \in D\} \quad (1.1)$$

A realization of (1.1) is denoted as $\{z(s) : s \in D\}$. Usually, $D$ is assumed to be fixed subset of $\mathbb{R}^d$, but more generally, we can assume $D$ is a random set. And we have the following three types of data (Cressie, 1993).

- **Geostatistical data**: $D$ is a fixed subset of $\mathbb{R}^d$ that contains a $d$-dimensional rectangular of positive volume; $Z(s)$ is a random vector at location $s \in D$.

- **Lattice data**: $D$ is fixed (regular or irregular) collection of countably many points of $\mathbb{R}^d$; $Z(s)$ is a random vector at location $s \in D$.

- **Point Process**: $D$ is a point process in $\mathbb{R}^d$ or a subset of $\mathbb{R}^d$; $Z(s)$ is a random vector at location $s \in D$.

From the descriptions above, there exist clear difference between lattice/geostatistical data and point process according to whether the spatial domain $D$ is fixed or random. In addition, the main difference between lattice data and geostatistical data is that for geostatistical data, the possible locations are spatially continuous in the region.
while for lattice data, the spatial domain consists of finite or countably many possible locations, i.e., *spatially discrete*.

Moreover, if we consider the time effect into the above spatial process, then the above spatial data can be naturally extended to spatial-temporal data. A purely temporal data (or commonly called time series data) is defined as follows.

\[ \{ Z(t) : -\infty < t < \infty \} \]  

(1.2)

And a *Spatial-temporal* data is described as.

\[ \{ Z(s, t) : s \in D(t), t \in T \} \]  

(1.3)

- \( Z(s, t) \): random variable/vector/set located at \( s \) and occurred at \( t \);
- \( D(\cdot) \): temporal process of random/fixed set in \( \mathbb{R}^d \);
- \( T \): random/fixed set in time (in my dissertation, I will consider \( T \) as discrete time points).

In the real world, data cannot be as "ideal" as we introduced at the very beginning of this chapter, instead, data containing time and space effect are much more common in fields like agriculture, biology, physics and social sciences. Though the spatial data structure is not as simple as the "ideal" data, statistical analysis for spatial data has been developed drastically during past years and a variety of literature has discussed spatial analysis from both theoretical and computational aspects (see Cressie, 1993 and Gelfand, 2005). My dissertation will focus on statistical analysis for spatial data and it will consist of two parts. We will introduce a simultaneous spatial-temporal autoregressive model in the first part and discuss the asymptotic property of the maximum likelihood estimates (MLEs) under such model. In the second part, we will propose a new nonparametric approach in estimating the the
spectral density function for spatial random field, more detailed, we will develop a multivariate nonparametric Bayesian prior and establish its theoretical property and generate a efficient computational algorithm in which we can stimulate the MCMC samples from the posterior.
Chapter 2 Spatial-Temporal Lattice Modelling and Maximum Likelihood Estimation

2.1 Overview

Likelihood-based methods have been widely used in analyzing different types of data since R. A. Fisher proposed the concept "likelihood" in 1930's. For the analysis of spatial-temporal data, spatial-temporal linear models are important tools and have been applied in a wide range of disciplines (see, e.g., Anselin, 2001, Baltagi, 2005 and Cressie, 1993). A spatial-temporal linear model relates the response variable of interest to covariates via a linear regression component and models the spatial-temporal dependence in data via a random error component that is assumed to be a zero-mean Gaussian process. In this chapter, we focus on simultaneous-autoregressive (SAR) type of spatial-temporal linear models for random error components and study the asymptotic properties of statistical inference via a maximum likelihood method.

For statistical inference of spatial-temporal linear models, maximum likelihood estimation is often adopted. For the spatial-only case, Mardia and Marshall (1984) established that the maximum likelihood estimates (MLE) of the parameters are consistent and asymptotically normal as the sample size tends to infinity for general spatial linear models, which corresponds to the increasing domain case discussed in the following sections. Lee (2004) studied the asymptotic properties of quasi-maximum likelihood estimators (QMLE) for a lag SAR model. Recently, Zheng and Zhu (2011) considered an error SAR model with general neighborhood structures and explored the asymptotic properties of MLEs under a unified asymptotic framework. Robinson and Thawornkaiwong (2012) developed the asymptotic normality of ordinary least squares (LS) and instrumental variables (IV) estimates of linear and semi-parametric
partly linear regression models for spatial data and discussed the consistency of the estimates of the spatial covariance matrix. However, it is not clear how general the asymptotic framework is. For the spatial-temporal case, Yu et al. (2008) investigated the asymptotic properties of QMLEs for spatial dynamic panel data with fixed effects and proposed a bias-adjusted estimator. Lee and Yu (2010a) and Lee and Yu (2010b) generalized the spatial dynamic panel data model to include both time and individual fixed effects and studied the asymptotic properties of QMLEs. Here, we consider the asymptotic properties of maximum likelihood estimation for spatial-temporal linear models. Across space, we consider the three types of asymptotic frameworks defined in Zheng and Zhu (2012), namely increasing domain, infill and hybrid of increasing domain and infill asymptotics. Over time, we assume that the number of time points tends to infinite as is traditionally done in time series, but will discuss the case with fixed time points.

For studying the asymptotics of parameter estimates of geostatistical data, three asymptotic frameworks were proposed, increasing domain, infill domain and a combination of increasing domain and infill. There has been active research on infill asymptotics in geostatistics in recent years (see Zhang, 2004). In contrast, for lattice data and SAR models under consideration here, little is known about the asymptotic properties of MLEs under infill and hybrid asymptotics. This is a void that we try to fill in this chapter.

### 2.2 Spatial-Temporal Linear Model

Let $Y_{it}$ denote the response variable at site $i$ and time $t$, where $i = 1, \ldots, n$ and $t = 1, \ldots, m$. Let

$$Y_{it} = X_{it}' \beta + \epsilon_{it},$$

(2.1)
where \( X_{it} = (X_{1it}, \ldots, X_{pit})' \) is a vector of the covariates and \( \beta = (\beta_1, \ldots, \beta_p)' \) a vector of the regression coefficients. To formulate a spatial-temporal model for the random errors \( \{\epsilon_{it} : i = 1, \ldots, n, t = 1, \ldots, m\} \), we focus on a SAR-type model and specify the model for \( \epsilon_t = (\epsilon_{1t}, \ldots, \epsilon_{nt})' \) in terms of the errors from time \( \max\{1, t-s\} \) to \( t \),

\[
\epsilon_t = \sum_{l=0}^{\min\{s, t-1\}} C_l \epsilon_{t-l} + \nu_t, \tag{2.2}
\]

where \( C_t = \{c_{ij}^{(l)}\}_{i,j=1}^n \), \( l = 0, \ldots, \min\{s, t-1\}, s \geq 0 \), are spatial-temporal dependence matrices and \( \nu_t = (\nu_{1t}, \ldots, \nu_{nt})' \sim N(0, \sigma^2 I_n) \) is a vector of white noise. Let \( Y_t = (Y_{1t}, \ldots, Y_{nt})' \) denote the \( n \)–dimensional vector of response variables on the entire spatial lattice for a given time point \( t \) and \( Y_{nm} = (Y_1', \ldots, Y_m')' \) the \( nm \)–dimensional vector of response variables at \( m \) time points.

Now we will let \( \epsilon_{nm} = (\epsilon_{11}, \ldots, \epsilon_{n1}, \ldots, \epsilon_{1m}, \ldots, \epsilon_{nm})' \) be a \( nm \)-dimensional vector of random errors and \( \nu_{nm} = (\nu_{11}, \ldots, \nu_{n1}, \ldots, \nu_{1m}, \ldots, \nu_{nm})' \) a \( nm \)-dimensional vector of white noises. We have, under (2.1) and (2.2),

\[
Y_{nm} = X_{nm} \beta + \epsilon_{nm}, \quad \text{and} \quad \epsilon_{nm} = C \epsilon_{nm} + \nu_{nm},
\]

where \( X_{nm} \) is a \( nm \times p \) design matrix, \( C \) is a lower-triangular matrix with \( C_0 \) as the diagonal blocks and \( C_l \) as the \( l \)th sub-diagonal blocks for \( l = 1, \ldots, s \). So the joint distribution of the response variable is

\[
Y_{nm} \sim N \left( X_{nm} \beta, \sigma^2(I_{nm} - C)'^{-1}(I_{nm} - C')^{-1} \right). \tag{2.3}
\]

where \( I_{nm} \) is an \( nm \times nm \) identity matrix.

The model (2.2) is quite general and features a variety of neighborhood structures over space and time. Let \( N_n(i) = \{j : \text{site } j \text{ is a neighbor of site } i\} \) denote the neighborhood of site \( i \). The neighborhood can be further partitioned into \( q \) orders, such that
\[ N_n(i) = \bigcup_{k=1}^{q} N_{n,k}(i) \] where \( N_{n,k}(i) = \{ j : \text{site } j \text{ is a } k\text{th order neighbor of site } i \} \).

Let \( W_{nk} = [w_{nk}^{i,j}]_{i,j=1}^{n} \) be a \( n \times n \) spatial weight matrix with zero diagonal elements for all \( 1 \leq k \leq q \). An example is binary spatial weights such that \( w_{nk}^{i,j} = 1 \) if \( j \in N_{n,k}(i) \) and 0 otherwise. Some special cases of model (2.2) are as follows:

- **Spatial independence**: \( s \geq 1 \), \( C_0 = 0 \) and \( C_l = \alpha_l I_n \) for \( l = 1, \ldots, s \).
- **Temporal independence**: \( s = 0 \) and \( C_0 = \sum_{k=1}^{q} \theta_k W_{nk} \).
- **Spatial-temporal separable neighborhood structure**: \( s \geq 1 \), \( C_0 = \sum_{k=1}^{q} \theta_k W_{nk} \) and \( C_l = \alpha_l I_n \) for \( l = 1, \ldots, s \).
- **Spatial-temporal non-separable neighborhood structure**: \( s \geq 1 \), \( C_0 = \sum_{k=1}^{q} \theta_k W_{nk} \) and \( C_l = \alpha_l I_n + \sum_{k=1}^{q} \theta_k W_{nk}^l \) for \( l = 1, \ldots, s \).

Regularity conditions on the parameter space are needed to ensure that the model specified under (2.1) and (2.2) is valid. Since \( C \) is lower-triangular, a sufficient condition to ensure the non-singularity of \( I_n - C \) is that its diagonal blocks \( I_n - \sum_{k=1}^{q} \theta_k W_{nk} \) are non-singular, where \( I_n \) is an \( n \times n \) identity matrix. For row standardized spatial weight matrices \( W_{nk} \), if \( \sum_{k=1}^{q} |\theta_k| < 1 \), then \( I_n - \sum_{k=1}^{q} \theta_k W_{nk} \) is nonsingular and thus the covariance matrix in (2.3) is positive definite (Corollary 5.6.16, Horn and Johnson, 1985). We let \( \theta = (\theta_1, \ldots, \theta_q)' \in \Theta \), where \( \Theta \) is a compact subset of \( \mathbb{R}^q \). In the following, we will focus on a spatial-temporal linear model with a separable spatial-temporal neighborhood structure and \( s = 1 \). The results can be extended to general \( s \geq 1 \) readily. We let \( \alpha \in A_{\alpha} \), where \( A_{\alpha} \) is a compact set of \((-1, 1)\).

### 2.3 Maximum Likelihood Estimations

Let \( \eta = (\beta', \xi', \sigma^2)' \) denote the \( \{p + (q + 1) + 1\} \)-dimensional vector of unknown parameters under the model specified in (2.1) and (2.2), where \( \xi = (\theta', \alpha)' \). The
log-likelihood function, up to a constant, is

$$
\ell(\eta) = -(nm/2) \log \sigma^2 + \log |S_{nm}(\xi)| - (2\sigma^2)^{-1} \nu_{nm}'\nu_{nm}. \tag{2.4}
$$

where $S_{nm}(\xi) = I_{nm} - C$ and $\nu_{nm} = S_{nm}(\xi)(Y_{nm} - X_{nm}\beta)$. The first-order derivatives of $\ell(\eta)$ with respect to $\beta$ and $\sigma^2$ are, respectively,

$$
\frac{\partial \ell(\eta)}{\partial \beta} = (\sigma^2)^{-1}X_{nm}'S_{nm}'(\xi)\nu_{nm}, \quad \frac{\partial \ell(\eta)}{\partial \sigma^2} = (2\sigma^4)^{-1}(\nu_{nm}'\nu_{nm} - nm\sigma^2).
$$

By setting the score functions equal to zero, we obtain the maximum likelihood estimate (MLE) of $\beta$ and $\sigma^2$,

$$
\hat{\beta}_{nm}(\xi) = \left\{X_{nm}'S_{nm}'(\xi)S_{nm}(\xi)X_{nm}\right\}^{-1}X_{nm}'S_{nm}'(\xi)S_{nm}(\xi)Y_{nm},
\hat{\sigma}_{nm}^2(\xi) = (nm)^{-1}\{Y_{nm} - X_{nm}\hat{\beta}_{nm}(\xi)\}'S_{nm}'(\xi)S_{nm}(\xi)\{Y_{nm} - X_{nm}\hat{\beta}_{nm}(\xi)\}.
$$

We define a profile log-likelihood function of $\xi$ as

$$
\ell(\xi) = \ell\{\hat{\beta}_{nm}(\xi), \xi, \sigma_{nm}^2(\xi)\} = -(nm/2) \log \hat{\sigma}_{nm}^2(\xi) + \log |S_{nm}(\xi)| - nm/2. \tag{2.5}
$$

The above profile log-likelihood function is similar to the log-likelihood function of weighted least squares: regardless of the constant term in (2.5), the first term can be viewed as the logarithm of the weighted neighborhood residuals, and the second term can be treated as the logarithm of variance-covariance structure of the SAR-type spatial-temporal model, while in weighted least squares, the two additives after removing the constant are the residuals and the logarithm of the weighted variance-covariance matrix. Then the MLE of $\xi$ maximizes the profile log-likelihood $\ell(\xi)$ and is denotes as $\hat{\xi}_{nm}$.

2.4 Asymptotic Properties

In this section, we will discuss the asymptotic properties (consistency and asymptotic normality) of maximum likelihood estimates under the SAR models specified above.
First, we will introduce three asymptotic frameworks. Then we will express some mild regularities in order to establish the asymptotic properties. We will present the main theorem along with some comments at the last subsection. Limited to the paragraph, we won’t attach any detailed proof here but will leave the proof of main theorems in the Appendices.

In Zheng and Zhu (2011), three types of asymptotic frameworks are defined for spatial linear models on a lattice in terms of the volume of the spatial domain (i.e., the Lebesgue measure) and that of the individual cells.

- Increasing domain asymptotics: The volume of the spatial lattice tends to infinity while the volume of each cell on the lattice is fixed (Fig. 2.1).
- Infill asymptotics: The volume of the spatial lattice is fixed while the volume of each cell on the lattice tends to zero (Fig. 2.2).
- Hybrid asymptotics (increasing domain combined with infill asymptotics): The volume of the spatial lattice tends to infinity and the volume of each cell on the lattice tends to zero (Fig. 2.3).

![Figure 2.1: Increasing Domain](image-url)
2.4.1 Regularities

In this section, we study the asymptotic properties of the MLEs of the model parameters for the spatial-temporal linear model defined in (2.1) and (2.2). We consider all three types asymptotics in the spatial domain and assume that time tends to infinity.

Let \( \eta_0 = (\beta_0', \xi_0', \sigma_0^2)' \) denote the \( \{p + (q + 1) + 1\} \)-dimensional vector of true parameters and \( S_{0nm} = S_{nm}(\xi_0) \). The model evaluated at the true parameters \( \eta_0 \) is \( Y_{nm} = X_{nm}\beta_0 + S_{0nm}^{-1}v_{onm} \). To establish the asymptotic properties of MLEs of the model parameters, we impose the following regularities.

(A.1) The elements \( w_{nk}^{ij} \) of the spatial weight matrix \( W_{nk} \) are at most of order \( h_n^{-1} \)
uniformly for all \( j \neq i \) and \( w_{nk}^{ij} = 0 \), for all \( i = 1, \ldots, n \) and \( k = 1, \ldots, q \) and \( h_n \) is bounded away from zero uniformly.

(A.2) The sequence of spatial weight matrices \( \{W_{nk} : k = 1, \ldots, q\} \) are uniformly bounded in matrix norms \( ||\cdot||_1 \) and \( ||\cdot||_\infty \).

(A.3) The matrix \( S_n(\theta) \) is nonsingular for \( \theta \in \Theta \) and \( n \in \mathbb{N} \), where \( S_n(\theta) = I_n - \sum_{h=1}^q \theta_k W_{nk} \).

(A.4) The sequence of matrices \( \{S_n^{-1}(\theta)\} \) is uniformly bounded in matrix norms \( ||\cdot||_1 \) and \( ||\cdot||_\infty \) for \( \theta \in \Theta \). The true parameter \( \xi_0 \) is in the interior of \( \Xi \), where \( \Xi = \Theta \times A_\alpha \).

(A.5) The elements of \( X_{nm} \) are uniformly bounded constants. The limit of
\[
(nm)^{-1}X'_{nm}S'_{nm}(\xi)S_{nm}(\xi)X_{nm} \text{ as } nm \to \infty
\]
exists and is nonsingular for \( \xi \in \Xi \).

(A.6) For \( \xi \neq \xi_0 \), \( \lim_{nm \to \infty} h_n(nm)^{-1}\{\log |\sigma_{nm}^2(\xi)S'_{nm}(\xi)S_{nm}(\xi)| - \log |\sigma_0^2S'_{0nm}S_{0nm}|\} \neq 0 \), where \( \sigma_{nm}^2(\xi) = (nm)^{-1}\sigma_0^2tr \left\{ S'_{0nm}S'_{nm}(\xi)S_{nm}(\xi)S_{0nm}^{-1} \right\} \).

Let \( V_n \) denote an \( n \times n \) matrix with elements \( [v_n^{ij}]_{i,j=1}^n \). The sequence of matrices \( V_n \) is uniformly bounded in matrix norm \( ||\cdot||_\infty \), if \( \sup_{1 \leq i \leq n, n \geq 1} \sum_{j=1}^n |v_n^{ij}| < \infty \). The sequence of matrices \( V_n \) is uniformly bounded in matrix norm \( ||\cdot||_1 \), if \( \sup_{1 \leq j \leq n, n \geq 1} \sum_{i=1}^n |v_n^{ij}| < \infty \).

Assumptions (A.1) and (A.2) are regularity conditions on the spatial weight matrices, where assumption (A.2) is generally satisfied under row standardization. Here the order of the elements in the spatial weight matrices \( h_n^{-1} \) is an essential element in the specification of an asymptotic type in the spatial domain. We assume \( w_{nk}^{ij} = \mathcal{O}(h_n^{-1}) \), where \( h_n \) could be bounded or tend to infinity. Consider an example with distance-based neighbors on a regular spatial lattice (Example 2 in Zheng and Zhu, 2012). Let \( d_{ij} \) denote the Euclidean distance between sites \( i \) and
and $a_{nk}^{i,j} = I\{d_{ij} \sim (\delta_{k-1}, \delta_k]\}$ with pre-specified threshold values $\delta_0 = 0 < \delta_1 < \cdots < \delta_q$. Then row standardized weight matrices based on $A_{nk} = [a_{nk}^{i,j}]_{n_i,j=1}^n$ are $W_{nk}$ with elements $w_{nk}^{i,j} = a_{nk}^{i,j}/\sum_{j=1}^n a_{nk}^{i,j}$. For this example, $h_n = \mathcal{O}(\max\{\sum_j^n a_{nk}^{i,j} : k = 1, \ldots, q, i = 1, \ldots, n\})$, which is bounded under increasing domain asymptotics. Under infill asymptotics $h_n \rightarrow \infty$ and $h_n/n$ does not tend to 0 as $n \rightarrow \infty$, whereas under hybrid asymptotics, $h_n \rightarrow \infty$ and $h_n/n \rightarrow 0$ as $n \rightarrow \infty$. Assumptions (A.3) and (A.4) are standard assumptions made about the sequence of matrices $S_n(\theta)$, where assumption (A.4) is needed to ensure the that the variance of $Y_n$ is bounded. For row standardized spatial weight matrices $W_{nk}$, if $\sum_{k=1}^q |\theta_k| < 1$, then assumption (A.3) and (A.4) are satisfied (Corollary 5.6.16, Horn and Johnson, 1985). Assumption (A.5) is a standard assumption of the design matrix and implies that the elements of $nm\{X'_{nm}S'_{nm}(\xi)S_{nm}(\xi)X_{nm}\}^{-1}$ are uniformly bounded.

Assumption (A.6) is needed to establish identifiable uniqueness when establishing consistency of the MLE (see, e.g., Lee, 2004, Zheng and Zhu, 2012). For the case with spatial independence, it can be shown that assumption (A.6) is simplified to

$$
\lim_{nm \rightarrow \infty} h_n \log \left[ (nm)^{-1} tr \left\{ S'_{0nm}^{-1}S'_{nm}(\xi)S_{nm}(\xi) S_{0nm}^{-1} \right\} \right] \neq 0 \quad \text{for } \xi \neq \xi_0,
$$

which is satisfied if $h_n$ converges to a non-zero point.

### 2.4.2 Asymptotic Properties of Maximum Likelihood Estimates

Let $\hat{\beta} = \hat{\beta}_{nm}(\hat{\xi}_{nm})$, $\hat{\sigma}_2^{nm} = \hat{\sigma}_2^{nm}(\hat{\xi}_{nm})$, and $\hat{\eta}_{nm} = (\hat{\beta}'_{nm}, \hat{\xi}'_{nm}, \hat{\sigma}_2^{nm})$ be the MLE of $\eta$. We have the asymptotic properties of the MLE of $\eta$ as follows.

**Theorem 1** Assume that (A.1)-(A.6) hold and $h_n/(nm) \rightarrow 0$ as $nm \rightarrow \infty$. Then the MLE of $\eta$ is consistent such that, as $nm \rightarrow \infty$, $\hat{\eta}_{nm} \xrightarrow{p} \eta_0$.

Theorem 1 gives the consistency of the MLE of model parameters. It shows that, under the regularity conditions, when the size of the spatial domain $n \rightarrow \infty$,

- if the number of time points $m$ is fixed, then the MLE of $\eta$ is consistent when
either \( h_n = O(1) \) or \( h_n \to \infty \) but \( h_n/n \to 0 \) as \( n \to \infty \), which correspond to increasing domain asymptotics and hybrid asymptotics in the spatial domain, respectively. This result is the same as for the spatial-only case in Zheng and Zhu (2012).

- if the number of time points \( m \to \infty \), the MLE of \( \eta \) is consistent in cases (i) \( h_n = O(1) \); (ii) \( h_n \to \infty \) but \( h_n/n \to 0 \); or (iii) \( h_n \to \infty \) but \( h_n/n \to c \in (0, \infty] \), where the case (iii) corresponds to infill asymptotics in the spatial domain.

When the size of spatial domain \( n \) is fixed, the MLE of \( \eta \) is consistent only if \( m \to \infty \).

**Theorem 2** Assume that (A.1)-(A.6) hold.

(i) If \( h_n = O(1) \) and the limit of \( -(nm)^{-1}E\left\{ \frac{\partial^2 \ell(\eta)}{\partial \eta \partial \eta'} \right\} \) as \( nm \to \infty \) exists and is positive definite for \( \eta \in \mathbb{R}^p \times \Xi \times \mathbb{R}^+ \), then the MLE of \( \eta \) is asymptotic normal such that, as \( nm \to \infty \),

\[
(nm)^{1/2}(\hat{\eta}_{nm} - \eta_0) \xrightarrow{D} N(0, \Sigma_{\eta_0})
\]

where \( \Sigma_{\eta_0}^{-1} = -\lim_{nm \to \infty} (nm)^{-1}E\left\{ \frac{\partial^2 \ell(\eta_0)}{\partial \eta \partial \eta'} \right\} \).

(ii) If \( h_n \to \infty \) and \( h_n^{1+\delta}/(nm) \to 0 \) for some \( \delta > 0 \) as \( nm \to \infty \), and if the limit of \( -h_n(nm)^{-1}E\left\{ \frac{\partial^2 \ell(\eta)}{\partial \theta \partial \theta'} \right\} \) as \( nm \to \infty \) exists and is positive definite for \( \theta \in \Theta \), then the MLE of \( \eta \) is asymptotically normal such that, as \( nm \to \infty \),

\[
(nm)^{1/2}(\hat{\eta}_{nm} - \eta_0) \xrightarrow{D} N(0, \Sigma_{\eta_0}), \quad (nm)^{1/2}(\hat{\sigma}_{nm}^2 - \sigma_0^2) \xrightarrow{D} N(0, 2\sigma_0^4)
\]

\[
(nm/h_n)^{1/2}(\hat{\beta}_{nm} - \beta_0) \xrightarrow{D} N(0, \Sigma_{\beta_0}), \quad (nm)^{1/2}(\hat{\alpha}_{nm}^2 - \alpha_0^2) \xrightarrow{D} N(0, \Sigma_{\alpha_0})
\]

where

\[
\Sigma_{\eta_0} = \sigma_0^2 \lim_{nm \to \infty} nm \left( X_{nm}' S_{0nm} S_{0nm} X_{nm} \right)^{-1},
\]

\[
\Sigma_{\theta_0}^{-1} = \lim_{nm \to \infty} E\left\{ -h_n(nm)^{-1} \frac{\partial^2 \ell(\eta_0)}{\partial \theta \partial \theta'} \right\},
\]

\[
\Sigma_{\alpha_0} = \lim_{nm \to \infty} nm \left\{ tr(H_{nm}' H_{nm}) + H_{nm}' H_{nm} \right\}^{-1}.
\]
Here $H_{nm} = F_{nm}S_{nm}^{-1}$, where $F_{nm}$ is a matrix with all elements equal to zero except that the first-order lower sub-diagonal blocks are $I_n$.

Theorem 2 is about the normality of the MLE of model parameters. Theorem 2(i) shows that

- if the sample size $n \to \infty$ and time points $m \to \infty$, then the MLE of $\eta$ is asymptotically normal at a convergence rate of square root of $nm$ when $h_n = O(1)$.

- if the time points $m$ is fixed but $n \to \infty$, the convergence rate is reduced to the square root of $n$ as for the spatial-only case.

- if the size of spatial lattice $n$ is fixed but $m \to \infty$, then the MLE of $\eta$ is asymptotically normal at a convergence rate of square root of $m$, which is consistent with results in Yu et al. (2008).

Theorem 2(ii) requires that the size of spatial lattice $n \to \infty$. It shows that

- if $m \to \infty$, when $h_n \to \infty$ but $h_n/n \to 0$, or $h_n \to \infty$ but $h_n/n \to c \in (0, \infty]$, then the MLE of $\eta$ is asymptotically normal with a convergence rate $\sqrt{nm}$ for the regression coefficients $\beta$, the temporal autoregressive coefficient $\alpha$ and the variance component $\sigma^2$ and a convergence rate $\sqrt{nm/h_n}$ for the spatial autoregressive coefficients $\theta$.

- if $m$ is fixed, then similarly to the spatial-only case in Zheng and Zhu (2012), the MLE of $\eta$ is asymptotically normal with a reduced convergence rate $\sqrt{n}$ for the regression coefficients $\beta$, the temporal autoregressive coefficient $\alpha$ and the variance component $\sigma^2$ and a convergence rate $\sqrt{n/h_n}$ for the spatial autoregressive coefficients $\theta$ when $h_n \to \infty$ and $h_n/n \to 0$. 
For the large-sample case, according to the consistency of MLEs, under the scenario of Theorem 2 (i) we can estimate the covariance structure $\Sigma_{\eta_i}$ by using
\[
\left[ - (nm)^{-1} E \left\{ \frac{\partial^2 \ell(\eta)}{\partial \eta \partial \eta'} \right\} \right]^{-1}_{\eta = \hat{\eta}_{nm}}.
\]
The detailed formats of the elements in $-(nm)^{-1} E \left\{ \frac{\partial^2 \ell(\eta)}{\partial \eta \partial \eta'} \right\}$ are given in Appendices. The covariance structure under the scenario of Theorem 2 (ii) can be estimated similarly.

2.5 Simulation

We now conduct a simulation study to examine the finite-sample properties of the MLEs under the three types of asymptotics across space and with an increasing number of time points. We consider an $r \times r$ square lattice with a unit resolution and $m$ temporal points. We vary the number of time points by letting $m = 2, 5$ or $10$. For each value of $m$, we vary the lattice size by letting $r = 4$ or $8$. For each lattice size, we further divide each cell into an $r^* \times r^*$ sub-lattice and vary the sub-lattice size by letting $r^* = 1, 2, 4$. Thus, for each time point, the sample size $n$ ranges from $16$ ($r = 4; r^* = 1$) to $1024$ ($r = 8; r^* = 4$).

For a given lattice size $r$, sub-lattice size $r^*$ and temporal length $m$, we simulate data from the spatial-temporal model defined in (2.1) and (2.2). For the linear regression, we let $E(Y_{ni}) = \beta_0 + \beta_1 X_i$, where $X_i = \sin(i)$, $\beta_0 = 2$, and $\beta_1 = 2$ for the $i$th cell, $i = 1, \ldots, n$ and $t = 1, \ldots, m$. For the spatial dependence, we consider distance-based neighborhood with order $q = 1$. We let $a_{ni}^{ij} = I\{d_{ij} \in (0,1]\}$, where $d_{ij}$ denotes the Euclidean distance between sites $i$ and $j$, and then define a row standardized weight matrix $w_{ni}^{ij} = a_{ni}^{ij} / \sum_{j=1}^{n} a_{ni}^{ij}$. The parameter values are set at $\theta_1 = 0.8$, $\alpha = 0.2$, and $\sigma^2 = 1$. For each simulated data, we estimate the model parameters by maximum likelihood and obtain $\hat{\beta}_0, \hat{\beta}_1, \hat{\theta}_1, \hat{\alpha}$, and $\hat{\sigma}^2$. We repeat this
procedure 100 times.

Table 2.1, 2.2 and 2.3 gives the means and standard deviations of the MLEs. First, for a given number of time points \( m \), we note that in general, the biases and standard deviations of all five parameter estimates decrease as the lattice size \( r \) increases from 4 to 8 for any given sub-lattice size \( r^* \) or as both \( r \) and \( r^* \) increase, which correspond to the increasing domain asymptotics and hybrid asymptotics across, respectively. Next, for a given number of time points \( m \) and given lattice size \( r \), we consider the results over all sub-lattice size \( r^* \), which corresponds to infill asymptotic, in general the biases and standard deviations of regression coefficient estimates \( \hat{\beta}_0, \hat{\beta}_1 \), variance component \( \hat{\sigma}^2 \) and temporal autoregressive coefficient \( \hat{\alpha} \) decrease as the sub-lattice size \( r^* \) increases from \( 1 \times 1 \) to \( 4 \times 4 \). However, for the spatial autoregressive coefficient estimate \( \hat{\theta} \), its biases and standard deviations remain similar as \( r^* \) increases, which is indicating \( \hat{\theta} \) is inconsistent in infill asymptotic when the number of time points \( m \) is fixed, and this result agree with the asymptotic property of infill asymptotic for spatial-only case in Zheng and Zhu (2011). Last, as the number of time points \( m \) increases from 2 to 10, we note that the overall performance of all five parameter estimates improves with either fixed or increasing lattice size \( r \) and sub-lattice size \( r^* \).
Table 2.1: Means and standard deviations (SD) of maximum likelihood estimates (MLE) of the model parameters based on 100 simulated data. Here the lattice size is $4 \times 4$ and $8 \times 8$ with varying sub-lattice sizes $1 \times 1$, $2 \times 2$ and $4 \times 4$ within each cell of the lattice, and the number of time points is 2.

<table>
<thead>
<tr>
<th>Truth</th>
<th>MLE</th>
<th>4 × 4</th>
<th>8 × 8</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>$n = 16$</td>
<td>$n = 32$</td>
</tr>
<tr>
<td>$\beta_0$</td>
<td>2.0</td>
<td>Mean 2.2553</td>
<td>1.6990</td>
</tr>
<tr>
<td></td>
<td></td>
<td>SD (0.2745)</td>
<td>(0.0240)</td>
</tr>
<tr>
<td>$\beta_1$</td>
<td>2.0</td>
<td>Mean 1.8913</td>
<td>2.2048</td>
</tr>
<tr>
<td></td>
<td></td>
<td>SD (0.3293)</td>
<td>(0.1026)</td>
</tr>
<tr>
<td>$\sigma^2$</td>
<td>1.0</td>
<td>Mean 1.4461</td>
<td>1.3757</td>
</tr>
<tr>
<td></td>
<td></td>
<td>SD (0.3590)</td>
<td>(0.1890)</td>
</tr>
<tr>
<td>$\theta$</td>
<td>0.8</td>
<td>Mean 0.9623</td>
<td>0.7071</td>
</tr>
<tr>
<td></td>
<td></td>
<td>SD (0.1953)</td>
<td>(0.0692)</td>
</tr>
<tr>
<td>$\alpha$</td>
<td>0.2</td>
<td>Mean 0.0225</td>
<td>0.0649</td>
</tr>
<tr>
<td></td>
<td></td>
<td>SD (0.1389)</td>
<td>(0.0201)</td>
</tr>
</tbody>
</table>
Table 2.2: Means and standard deviations (SD) of maximum likelihood estimates (MLE) of the model parameters based on 100 simulated data. Here the lattice size is 4×4 and 8×8 with varying sub-lattice sizes 1×1, 2×2 and 4×4 within each cell of the lattice, and the number of time points is 5.

<table>
<thead>
<tr>
<th></th>
<th>Truth</th>
<th>MLE</th>
<th>4 × 4</th>
<th></th>
<th>8 × 8</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>n = 16</td>
<td>n = 64</td>
<td>n = 256</td>
<td>n = 64</td>
<td>n = 256</td>
</tr>
<tr>
<td>β₀</td>
<td>2.0</td>
<td>Mean</td>
<td>2.2430</td>
<td>1.7536</td>
<td>2.2494</td>
<td>2.1473</td>
</tr>
<tr>
<td></td>
<td></td>
<td>SD</td>
<td>(0.0254)</td>
<td>(0.0058)</td>
<td>(0.0044)</td>
<td>(0.0072)</td>
</tr>
<tr>
<td>β₁</td>
<td>2.0</td>
<td>Mean</td>
<td>1.7381</td>
<td>2.1886</td>
<td>2.1069</td>
<td>2.1472</td>
</tr>
<tr>
<td></td>
<td></td>
<td>SD</td>
<td>(0.0616)</td>
<td>(0.0191)</td>
<td>(0.0188)</td>
<td>(0.0167)</td>
</tr>
<tr>
<td>σ²</td>
<td>1.0</td>
<td>Mean</td>
<td>1.7640</td>
<td>1.2659</td>
<td>1.2123</td>
<td>1.3342</td>
</tr>
<tr>
<td></td>
<td></td>
<td>SD</td>
<td>(0.5033)</td>
<td>(0.0824)</td>
<td>(0.0690)</td>
<td>(0.1344)</td>
</tr>
<tr>
<td>θ</td>
<td>0.8</td>
<td>Mean</td>
<td>0.7175</td>
<td>0.7076</td>
<td>0.7195</td>
<td>0.6907</td>
</tr>
<tr>
<td></td>
<td></td>
<td>SD</td>
<td>(0.0627)</td>
<td>(0.0246)</td>
<td>(0.0258)</td>
<td>(0.0354)</td>
</tr>
<tr>
<td>α</td>
<td>0.2</td>
<td>Mean</td>
<td>0.0536</td>
<td>0.0774</td>
<td>0.0902</td>
<td>0.0804</td>
</tr>
<tr>
<td></td>
<td></td>
<td>SD</td>
<td>(0.1026)</td>
<td>(0.0053)</td>
<td>(0.0141)</td>
<td>(0.0188)</td>
</tr>
</tbody>
</table>
Table 2.3: *Means and standard deviations (SD) of maximum likelihood estimates (MLE) of the model parameters based on 100 simulated data. Here the lattice size is $4 \times 4$ and $8 \times 8$ with varying sub-lattice sizes $1 \times 1$, $2 \times 2$ and $4 \times 4$ within each cell of the lattice, and the number of time points is 10.*

<table>
<thead>
<tr>
<th>Truth</th>
<th>MLE</th>
<th>4 × 4</th>
<th></th>
<th>8 × 8</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>$n = 16$</td>
<td>$n = 64$</td>
<td>$n = 256$</td>
<td>$n = 64$</td>
</tr>
<tr>
<td>$\beta_0$ 2.0</td>
<td>Mean</td>
<td>1.7480</td>
<td>2.1839</td>
<td>1.8039</td>
<td>1.8889</td>
</tr>
<tr>
<td></td>
<td>SD</td>
<td>(0.0146)</td>
<td>(0.0221)</td>
<td>(0.0161)</td>
<td>(0.0111)</td>
</tr>
<tr>
<td>$\beta_1$ 2.0</td>
<td>Mean</td>
<td>1.8046</td>
<td>2.1780</td>
<td>1.9044</td>
<td>2.1789</td>
</tr>
<tr>
<td></td>
<td>SD</td>
<td>(0.0686)</td>
<td>(0.0319)</td>
<td>(0.0092)</td>
<td>(0.0240)</td>
</tr>
<tr>
<td>$\sigma^2$ 1.0</td>
<td>Mean</td>
<td>1.3016</td>
<td>1.2165</td>
<td>1.2162</td>
<td>1.2812</td>
</tr>
<tr>
<td></td>
<td>SD</td>
<td>(0.1227)</td>
<td>(0.1708)</td>
<td>(0.0738)</td>
<td>(0.1390)</td>
</tr>
<tr>
<td>$\theta$ 0.8</td>
<td>Mean</td>
<td>0.7575</td>
<td>0.7159</td>
<td>0.7158</td>
<td>0.7341</td>
</tr>
<tr>
<td></td>
<td>SD</td>
<td>(0.0339)</td>
<td>(0.0254)</td>
<td>(0.0212)</td>
<td>(0.0321)</td>
</tr>
<tr>
<td>$\alpha$ 0.2</td>
<td>Mean</td>
<td>0.0654</td>
<td>0.0781</td>
<td>0.1146</td>
<td>0.0973</td>
</tr>
<tr>
<td></td>
<td>SD</td>
<td>(0.0675)</td>
<td>(0.0346)</td>
<td>(0.0368)</td>
<td>(0.0543)</td>
</tr>
</tbody>
</table>
2.6 Conclusions and Discussion

According to the discussions above, we have studied the asymptotic properties of MLEs under general asymptotic framework for spatial-temporal linear models. We have considered three asymptotics in the spatial domain and let the number of time points tend to infinity. Under mild regularity conditions on the spatial-temporal weight matrices, we have derived the asymptotic properties (consistency and asymptotic normality) of maximum likelihood estimates. The results can be easily extended to models with temporal lags $s > 1$. It is plausible that the asymptotics of MLEs for models with a general non-separable spatial-temporal neighborhood structure can be developed in a similar technique, which is currently under investigation.

In our spatial-temporal autoregressive models, we assume that the errors are zero with $\epsilon_t = 0$ at initial time points $1 - s \leq l \leq 0$. An alternative way to formulate the process is to pre-specify a distribution for the errors at the initial time points. An analogy is an AR(1) model in time series $\epsilon_t = \rho \epsilon_{t-1} + \nu_t$, where $\nu_t \sim iid \ N(0, \sigma^2_{\nu})$. It is conventional to let $\epsilon_1 \sim N(0, \sigma^2_{\nu}/(1 - \rho^2))$ such that $var(\epsilon_t) = \sigma^2_{\nu}/(1 - \rho^2)$. However, this would be challenging for general spatial-temporal process. In addition, a spatial-temporal process can be proposed to condition on the initial $s$ time points with $\epsilon_t = \sum_{l=0}^{s} C_t \epsilon_{t-l} + \nu_t$, where $\nu_t = (\nu_{1t}, \ldots, \nu_{nt})' \sim iid \ N(0, \sigma^2)$ for $t = s+1, \ldots, m$. When $m$ goes to infinity, MLEs will behave similarly to those under our model specification.

Copyright© Xiang Zhang, 2013.
Chapter 3 Multivariate Feller Priors

3.1 Motivation

Parametric models and traditional methods based on such models have a long history in analyzing data. As we discussed the asymptotic properties of maximum likelihood estimates under SAR-type spatial-temporal models in chapter 2, likelihood-based methods are very useful in investigating data mechanism. However, there may not be enough information in many scenarios for specifying a particular parametric model, and concentrating inferences on a specific parametric form may short the research scope and limit possible conclusions that can be drawn from the data. In addition, if the restrictive parameter assumptions for the model are not appropriate, then inferences based on such parametric form may mislead us. For example, if the data have heavy tails, then assuming the density is normally distributed is obviously incorrect. Thus, a nonparametric or semi-parametric approach is more reasonable when the data structure are complicated.

Traditional nonparametric inferences includes histogram smoothers, splines, kernel density estimations, etc (see, Wasserman, 2005 and reference therein if you want to know more about traditional nonparametric methods). Simply speaking, nonparametric methods will let the data determine the possible statistical models instead of specifying a parametric model in advance. Under a Bayesian framework, a nonparametric model is a probability model with infinitely many parameters and requires specifying a prior on a infinite dimensional function space, e.g., the family of all distribution functions on the sample space. During past years, the literature about nonparametric Bayesian methodologies has grown rapidly with applications in density estimation, regression, survival analysis and hierarchical models (See, e.g.,
Walker et al., 1999, Muller and Quintana, 2004). In spatial statistics, Gelfand et al. (2005) proposed a nonparametric Bayesian method using mixed Dirichlet process, Reich and Fuentes (2007) applied a semi-parametric Bayesian model in environmental statistics. In this chapter, I want to develop a nonparametric Bayesian prior for estimating spectral density functions in spatial data.

3.2 Spectral Density Estimation

For analysis of spatially referenced data, random fields provide a flexible modeling framework other than the SAR-type spatial-temporal modeling as we discussed above. An autocovariance function on a spatial domain can reflect the spatial dependence of the data and a lot of research has been focused on properly inferring and modeling the autocovariance function. Another powerful alternative would be using spectral density instead, since the spectral density is a one-to-one transformation of the autocovariance function.

Let \( \{X(s) : s \in D \subset \mathbb{R}^d\} \) denote a random field on the spatial domain \( D \). We assume that the random field is second-order stationary with an autocovariance function \( C(h) = Cov\{X(s), X(s+h)\} \), where \( s \) represent a spatial location on the spatial domain \( D \) and \( h \) denote the distance. The spectral density, as a Fourier transformation of the autocovariance function, is defined as

\[
f(\omega) = (2\pi)^{-d} \int_{\mathbb{R}^d} \exp(-ih'\omega)C(h)dh\tag{3.1}
\]

where \( \omega = (\omega_1, \ldots, \omega_d) \in \mathbb{R}^d \).

Further, if the spatial domain \( D \) is an integer lattice in \( \mathbb{Z}^d \), the frequency \( \omega \) is restricted to a finite-frequency band \( (-\pi, \pi]^d \) and (3.1) can be written as:

\[
f(\omega) = \frac{1}{2\pi} \sum_{h \in \mathbb{Z}^d} \exp\{-ih'\omega\}C(h).\tag{3.2}
\]
Let $D_n$ denote the spatial sampling locations in $D$ and we assume $D_n$ consists of an $n_1 \times \cdots \times n_d$ lattice in $\mathbb{Z}^d$. A periodogram provides a nonparametric estimate of spectral density as:

$$I(\omega) = \frac{1}{2\pi \prod_{i=1}^d n_i} \left| \sum_{s \in D_n} X(s) \exp\{-is'\omega\} \right|^2. \quad (3.3)$$

Some previous literature shows that the periodogram is an asymptotically unbiased estimator of the spectral density estimation under mild regularities on the autocovariance function, i.e., the periodogram, as a process, is randomly fluctuating around the true spectral density estimation. Thus, statistical inference of spectral density estimation, especially through nonparametric approach, is largely based on smoothing of the periodogram given in (3.3).

The main difference between the spectral density and a probability density is that $f(\omega)$ does not integrate to 1. However, we can do a normalization $q(\omega) = f(\omega)/\tau$ first with the normalizing constant $\tau = \int f$. After normalizing, $q(\omega)$ will integrate to 1 like a common probability density function. A prior on $f$ may be induced by first expressing a nonparametric Bayesian prior on $q$ and then independently expressing a prior on $\tau$. Generally speaking, the prior on $f$ can described as follows:

- Assign a nonparametric prior on $q(\omega)$.
- Let normalizing constant $\tau$ have a Lebesgue density on $(0, \infty)$.
- $q$ and $\tau$ are priori independent.
According to the discussion above, a key point in spectral density estimation is to assign an appropriate nonparametric Bayesian prior on the normalized spectral density $q(\omega)$. A variety of nonparametric Bayesian priors for density functions have been used since Ferguson proposed the well-known Dirichlet process prior in 1973. Some continuous work include Dirichlet Process Mixtures (DP mixtures) (see, Lo, 1984, Escobar and West, 1995 and Gasparini, 1996), Polya trees (Lavine, 1992 and Lavine, 1994) and Bernstein Polynomials (Petrone, 1999a). For time series, various nonparametric Bayesian methods were developed to estimate the spectral density (Choudhuri et al., 2004 and the references therein). In particular, Choudhuri et al. (2004) proposed a nonparametric Bayesian method to estimate the spectral density of a stationary time series, where the nonparametric Bayesian prior on the spectral density was based on one-dimensional Bernstein polynomials. Recently, Zheng et al. (2010) proposed multivariate Bernstein Polynomial priors for estimating spectral density of a random field, which leads to an innovative thinking of constructing a multivariate nonparametric prior which include multivariate Bernstein Polynomials as a special case.

### 3.3 Multivariate Feller Operators

#### 3.3.1 Multi-dimensional Feller-Type Approximation

Petrone and Veronese (2002) discussed some theoretical properties of one-dimensional Feller operators, and we will extend to the multi-dimensional case. The idea is originated from Feller (1971) of defining a constructive way to approximate a given bounded and continuous function on a (bounded or unbounded) interval $E \subset \mathbb{R}$. We will simply extend the idea to multivariate case. Two important definitions are given below before discussing any theoretical property. Assume $\chi \subseteq \mathbb{R}^d$ is a convex set.

**Definition 1 (Random Scheme)** A $D$-dimensional Random Scheme is defined as
a sequence of $d$-dimensional random vectors \( \{ Z_{kx} = (Z_{1,k1,x1}, \ldots, Z_{d,k_d,x_d}) : k = (k_1, \ldots, k_d), k_1, \ldots, k_d = 1, 2, \ldots, x = (x_1, \ldots, x_d) \in \chi \} \) where \( Z_{kx} \) has distribution function (d.f.) \( P_{kx} \) such that for \( x = (x_1, \ldots, x_d) \in \chi \),

\[
\lim_{k_i \to \infty} E(Z_{i,k_i,x_i}) = x_i, \quad i = 1, \ldots, d, \tag{3.4}
\]

\[
\lim_{\min\{k\} \to \infty} \sum_{i=1}^{d} V(Z_{i,k_i,x_i}) = 0, \tag{3.5}
\]

where \( \min\{k\} = \min_{1 \leq i \leq d} k_i \).

**Definition 2 (Multivariate Feller operators)** Let \( G : \mathbb{R}^d \to \mathbb{R} \) be a real function. The \( d \)-dimensional Feller operator with random scheme \( Z_{kx} \) is defined as:

\[
B(x; k, G) = E(G(Z_{kx})) = \int G(z)dP_{kx}(z).
\]

The following theorem states the approximation property of multivariate Feller operators.

**Theorem 3** If \( G : \mathbb{R}^d \to \mathbb{R} \) is bounded on \( \chi \) and let \( B(x; k, G) \) be the associated multi-dimensional Feller operator defined above, then as \( \min\{k\} \to \infty \), we have \( B(x; k, G) \to G(x) \) at every continuity point \( x \in \chi \) of \( G \). If \( G \) is continuous, then \( B(x; k, G) \to G(x) \) uniformly on every compact set \( \Delta \) of \( \chi \) in which (3.4) and (3.5) hold uniformly.

Theorem 3 indicates that if a multivariate real function \( G \) is bounded, then \( G \) can be approximated by a sequence of multivariate Feller operators \( B(x; k, G) \). Further, if restricted on a compact set of the function domain, then the approximation is uniform.
3.3.2 Exponential Random Scheme and Multivariate Distribution Function Approximation

In the previous subsection, we discussed the approximation property of Feller operators for a bounded d-dimensional real function. In statistical inference, we are interested in a particular type of bounded functions, the distribution function. Since distribution function itself is bounded in $[0, 1]$, so a natural thinking is that: can we use a multivariate Feller operators to approximate multivariate distribution function? The answer is "Yes", but not all Feller operators can be used for such approximation, only multivariate Feller operators associated with a certain random scheme can. Therefore, I will introduce a specific random scheme family called Exponential Random Scheme family in order to explore the approximation of a multivariate distribution function.

**Definition 3 (Natural Exponential Family)** Given a non-degenerate $\sigma$–finite measure $\nu$ on the Borel sets of $\mathbb{R}$, then the Natural Exponential Family (NEF), with natural parameter $\theta$, is the family of probability measures $\mathcal{F}$ on $\mathbb{R}$ with density

$$p_\theta(y) = \exp\{\theta y - M(\theta)\}, \quad \theta \in \Theta$$

with respect to $\nu$, where $M(\theta) = \ln \int \exp(\theta y) \nu(dy)$ and $\theta \in \Theta = \{\theta : M(\theta) < +\infty\}$.

Denote the mean parameter by $x$. From the properties of exponential family, we have $x = E_\theta(Y) = dM(\theta)/d\theta = x(\theta)$ and the variance is $\sigma^2 = \sigma^2(x)$. Since the transformation from $x$ to $\theta$ is one-to-one, we can also write $\theta = \theta(x)$, and the family can be re-parametrized in the mean parameter $x$.

**Definition 4 (Exponential Random Scheme)** Consider the average

$$Z_{kx} = (1/k) \sum_{i=1}^{k} Y_i,$$

where $Y_1, \ldots, Y_k$ are i.i.d. random variables with distributions in the natural exponential family. Then $Z_{kx}$ has d.f. $P_{k,\theta(x)}$ which is still in the NEF with mean $E(Z_{kx}) = x$ and variance $V(Z_{kx}) = \sigma^2/k$ converges to zero as $k \to \infty$. 28
Therefore, \( \{Z_{kx} \sim P_{k, \theta(x)}, k = 1, 2, \ldots, x \in \chi^0\} \) is a random scheme with \( \chi^0 \) the interior of \( \chi \), called exponential random scheme (ERS).

Consider a d-dimensional distribution function \( G \) with convex support \( \chi \subseteq \mathbb{R}^d \). Let \( x = (x_1, \ldots, x_d) \in \chi^0 \). A random vector \( Z_{kx} = (Z_{1,kx_1}, \ldots, Z_{d,kx_d}) \) whose component are independent, and defined as:

\[
Z_{i,kx_i} = \frac{1}{k_i} \sum_{j=1}^{k_i} Y_{i,j},
\]

where for any \( i = 1, \ldots, d \), \( (Y_{i,1}, Y_{i,2}, \ldots) \) are i.i.d with a distribution in the NEF such that \( E(Y_{i,j}) = x_i \) and \( V(Y_{i,j}) = \sigma^2(x_i) \). Then \( Z_{i,kx_i} \) has a distribution function \( \mathcal{P}_{k_i, \theta(x_i)}(x_i) \) in the NEF with

\[
E(Z_{i,kx_i}) = x_i \quad \text{and} \quad V(Z_{i,kx_i}) = \frac{\sigma^2(x_i)}{k_i} \to 0 \quad \text{as} \quad k_i \to \infty.
\]

Hence \( \{Z_{kx} = (Z_{1,kx_1}), \ldots, (Z_{d,kx_d}) : k = (k_1, \ldots, k_d), k_1, \ldots, k_d = 1, 2, \ldots, x = (x_1, \ldots, x_d) \in \chi^0 \subseteq \mathbb{R}^d\} \) is a d-dimensional exponential random scheme (ERS).

Since \( G \) is a d-dimensional distribution function with convex support \( \chi \), a multivariate Feller operator with a d-dimensional ERS for \( G \) is:

\[
B(x; k, G) = E(G(Z_{kx})) = \int \cdots \int G(z_1, \ldots, z_d) dP_{k_1, \theta(x_1)}(z_1) \cdots dP_{k_d, \theta(x_d)}(z_d),
\]

for \( x \in \chi \) and the following properties hold for multivariate Feller operators.

**Proposition 1**

(i) \( B(\cdot; k, G) \) is a d.f.

(ii) The derivative of \( B(x; k, G) \) for \( x \in \chi^0 \) is given by

\[
b(x; k, G) = \frac{\partial}{\partial x_1} \cdots \frac{\partial}{\partial x_d} B(x_1, \ldots, x_d; k, G)
\]

\[
= \int \cdots \int \prod_{i=1}^{d} f_{k_i}(x_i, z_i) dG(z_1, \ldots, z_d),
\]

where

\[
f_{k_i}(x_i, z_i) = \frac{k_i}{\sigma^2(x_i)} \int_{[z_i, \infty)} (t - x_i) dP_{k_i, \theta(x_i)}(t).
\]
Proposition 1 shows that we use a sequence of multivariate operators associated with exponential random scheme to approximate a multivariate distribution function, and the operator $B(\cdot; k, G)$ itself is also a distribution function. In addition, the Feller operator has derivative and we call such derivative Feller density. Further, if $G$ has no mass concentrated on the frontier of $\chi$, then $B(x; k, G)$ can be extended to $\mathbb{R}^d$ properly as an absolutely continuous distribution function with density given by (3.7). For Feller density, We have the following property.

**Proposition 2** If $G$ is absolutely continuous with density $g(x_1, \ldots, x_d)$ continuous and bounded in each argument, then $b(x; k, G)$ is a $d$-dimensional Feller operator. Furthermore, $b(x; k, G) \rightarrow g(x)$ for every $x \in \chi^o$, and the convergence is uniformly on every compact set $\Delta \subseteq \chi^o \subseteq \mathbb{R}^d$ in which (3.4) and (3.5) hold uniformly for the corresponding random scheme.

Proposition 2 shows that if $G$ is a probability distribution function with a continuous and bound probability density $g$, then similar to the distribution function $G$, the corresponding density $g$ can also be approximated by a sequence of Feller densities.

### 3.3.3 Examples

In this section, we will provide some examples of multivariate Feller operator with discrete or continuous ERS. Let $k = (k_1, \ldots, k_n)$ and $l = (l_1, \ldots, l_n)$ denote two non-negative vectors. Define $l \leq k$ if $l_i \leq k_i$ for any $i$. Define $\sum_{j=l}^k = \sum_{j_1=l_1}^{k_1} \cdots \sum_{j_d=l_d}^{k_d}$ when $j = (j_1, \ldots, j_d)$. Let $j/k$ denote the vector $(j_1/k_1, \ldots, j_d/k_d)$. Let cube$(j, k)$ denote the $d$-dimensional cube of the form $\left(\frac{j_1}{k_1}, \frac{j_1}{k_1}\right] \times \cdots \times \left(\frac{j_d}{k_d}, \frac{j_d}{k_d}\right]$ with the convention that if $j_i = 0$, then the interval $\left(\frac{j_i}{k_i}, \frac{j_i}{k_i}\right]$ is replaced by the point $\{0\}$.

**Example 1 (Binomial Random Scheme)** This one is the well-know Bernstein polynomial operators. Let $G$ be a $d$-dimensional distribution function with convex support $\chi = [0, 1]^d$. For any $x \in \chi^o$ and $i = 1, 2, \ldots$, let $(Y_{i,1}, Y_{i,2}, \cdots)$ be independent
Bernoulli random variables with parameter \( x_i \). Then

\[
k_i Z_{i,k_i} = \sum_{j_i=1}^{k_i} Y_{i,j_i} \sim \text{Bin}(\cdot; k_i, x_i), \quad i = 1, \ldots, d.
\]

The \( d \)-dimensional Feller operator with \( d \)-dimensional binomial random scheme is defined by

\[
B(x; k, G) = \sum_{j=0}^{k} G(j/k) \prod_{i=1}^{d} \left( \frac{k_i}{j_i} \right) x_i^{j_i} (1 - x_i)^{1-j_i},
\]

and the corresponding derivative for \( x \in (0,1)^d \) is

\[
b(x; k, G) = \sum_{j=1}^{k} u_{k,G}(j) \prod_{i=1}^{d} \text{Beta}(x_i; j_i, k_i - j_i + 1),
\]

where \( \text{Beta}(\cdot; j_i, k_i - j_i + 1) \) denotes the probability density function of a Beta distribution with parameter \( j_i \) and \( k_i - j_i + 1 \) and \( u_{k,G}(j) = P_G((\frac{j_1 - 1}{k_1}, \frac{j_1}{k_1}) \times \cdots \times (\frac{j_d - 1}{k_d}, \frac{j_d}{k_d})) \) with \( P_G \) denoting the probability measure of the d.f. \( G \).

**Example 2 (Poisson Random Scheme)** If \( G \) is a \( d \)-dimensional distribution function defined on \([0, \infty)^d\). Let \((Y_{i,1}, Y_{i,2}, \cdots)\) be independent Poisson random variables with parameter \( x_i \). Then

\[
k_i Z_{i,k_i,x_i} = \sum_{j_i=1}^{k_i} Y_{i,j_i} \sim \text{Poi}(k_i x_i), \quad i = 1, \ldots, d.
\]

Therefore, the Feller operator with \( d \)-dimensional Poisson random scheme is defined by

\[
B(x; k, G) = \sum_{j=1}^{\infty} G(j/k) \prod_{i=1}^{d} (k_i x_i)^{j_i} e^{-k_i x_i} \frac{e^{-k_i x_i}}{j_i!},
\]

and the corresponding derivative of \( B(x; k, G) \) is

\[
b(x; k, G) = \sum_{j=1}^{\infty} u_{k,G}(j) \prod_{i=1}^{d} \text{Ga}(x_i; j_i, k_i),
\]

where \( \text{Ga}(\cdot; j_i, k_i) \) denotes the probability density function of a Gamma distribution with parameter \( j_i \) and \( k_i \), and \( u_{k,G}(j) \) is the same as in Example 1.
Example 3 (Gaussian Random Scheme) The previous two examples are based on discrete exponential random schemes, and we will provide an example with a d-dimensional continuous exponential random scheme here. Let $G$ be a bounded function defined on $\mathbb{R}^d$. Then a Feller operator with ERS having mean in $\mathbb{R}^d$ can be used to approximate $G$. A natural choice is a Gaussian family with mean $x = (x_1, \cdots, x_d)$ and fixed variance-covariance matrix. Let $(Y_{i,1}, Y_{i,2}, \cdots)$ be independent Gaussian random variables with mean $x_i$ and fixed variance $\sigma_i^2$. Then

$$k_i Z_{i,k_i x_i} = \sum_{j=1}^{k_i} Y_{i,j} \sim N(k_i x_i, k_i \sigma_i^2).$$

This gives

$$B(x; k, G) = \int \cdots \int G(z_1, \cdots, z_d) \prod_{i=1}^{d} \varphi(z_i; x_i, \sigma_i^2/k_i) dz_1 \cdots dz_d,$$

and the derivative $b(x; k, G)$ is

$$b(x; k, G) = \int \cdots \int \prod_{i=1}^{d} \varphi(z_i; x_i, \sigma_i^2/k_i) dG(z_1, \cdots, z_d),$$

where $\varphi$ is the density function for normal distribution with mean $x_i$ and variance $\sigma_i^2/k_i$.

Furthermore, if $G$ has a continuous and bounded density $g$, then

$$b(x; k, G) = \int \cdots \int g(z_1, \cdots, z_d) \prod_{i=1}^{d} \varphi(z_i; x_i, \sigma_i^2/k_i) dz_1 \cdots dz_d.$$

3.3.4 Shape properties of Feller operator with ERS

For statistical inference, it is important that the model can be restricted to an admissible class of densities with certain geometric/shape properties of the target distribution. It’s well known that the one-dimensional Bernstein polynomial operators retain monotonicity of the function generating the coefficients (Lorentz, 1986). Zheng (2011) extend such property to multi-dimensional Bernstein polynomial operators, called coordinate-wise monotonicity. We can prove that, with certain exponential random
scheme, i.e. Poisson random scheme and Gaussian random scheme, the corresponding multivariate Feller operators also have coordinate-wise monotonicity.

**Definition 5 (Coordinate-wise monotonicity)** A multivariate function $f : \mathbb{R}^d \to \mathbb{R}$ is said to be coordinate-wise monotone if for each $i = 1, \ldots, d$, $f$ is a non-decreasing or nonincreasing function of the $i$th coordinate while the other coordinates are fixed.

**Proposition 3** Let $B(x; k, G)$ be the feller operator with Poisson random scheme and $b(x; k, G)$ be the corresponding derivative. If the density of $G$ is coordinate-wise monotone, the the coordinate-wise monotonicity is retained by $b(x; k, G)$.

**Proposition 4** Let $B(x; k, G)$ be the feller operator with Gaussian random scheme and $b(x; k, G)$ be the corresponding derivative. If the density of $G$ is coordinate-wise monotone, the the coordinate-wise monotonicity is retained by $b(x; k, G)$.

The above two properties ensure that Feller operators with Poisson or Gaussian random scheme are monotone along the $i$th coordinate.

### 3.4 Multivariate Feller Prior

Till now, the function $G$ to be approximated is deterministic. If $G$ is random, the approximation of $G$ by a sequence of multivariate Feller operator $B(x; k, G)$ are also random functions whose probability law can be used as a prior in nonparametric Bayesian inference. We call such priors "multivariate Feller prior", denoted by $\pi_F$. The following theorem establish the existence of such priors.

**Theorem 4** Let $\Pi$ denote the space of probability distribution functions on the convex support $\chi \subseteq \mathbb{R}^d$ equipped with the Borel $\sigma$-field $\Gamma$ generated by the topology of weak convergence. Then multi-dimensional Feller operators induce a probability prior $\pi_F$ on $(\Pi, \Gamma)$. 

33
The probability measure induced in theorem 4 of prior is the multivariate Feller prior \( \pi_F \) with parameter a joint distribution of \( k \) and \( G \).

In order for the proposed multivariate Feller prior \( \Pi_F \) to be a valid nonparametric prior, a well-accepted criteria is that it should have *large* or *full* topological support (Ferguson, 1973). Next, we will establish the validity of the proposed multivariate Feller prior. Let \( G_n = \{g_1, \ldots, g_n\} \) denote a collection of real-valued continuous functions on a compact set \( \Delta \subseteq \chi \subseteq \mathbb{R}^d \), where \( n = 1, 2, \cdots \). Let \( \Pi_1 \) denote the class of all probability distribution functions on \( \Delta \) and \( \Pi_2 \) denote the class of all absolute continuous probability distribution functions (therefore they have densities) on \( \Delta \).

For any \( \epsilon > 0 \), any \( Q \in \Pi_1 \), we define a weak neighborhood of \( Q \) as:

\[
N_{G_n,\epsilon}(Q) = \{ Q^* \in \Pi_1 : \max_{g \in G_n} \left| \int g dQ^* - \int g dQ \right| < \epsilon \}. \tag{3.8}
\]

For any \( \epsilon > 0 \), any \( Q \in \Pi_2 \) with density \( f_Q(\cdot) \), define a strong neighborhood of \( Q \) as:

\[
N_{\epsilon}(Q) = \{ Q^* \in \Pi_2 : \sup_{\omega \in \Delta} \left| Q^*(\omega) - Q(\omega) \right| < \epsilon \}, \tag{3.9}
\]

and a total-variation (TV) neighborhood of \( Q \) as:

\[
N_{f_Q,\epsilon}(Q) = \{ Q^* \in \Pi_2 : \left| \int f_{Q^*}(z) dz - \int f_Q(z) dz \right| < \epsilon \},
\]

where \( f_Q \) is the density function of the distribution function \( Q \).

**Theorem 5** Let the prior of \( k \) satisfy \( p(k) > 0 \) for all \( k \in \mathbb{N}^d \). Given \( k \),

(i) if for any \( Q \in \Pi_1 \), the conditional prior \( p(N_{G_n,\epsilon}(Q)|k) > 0 \) for any weak neighborhood \( N_{G_n,\epsilon}(Q) \) of \( Q \). Then, \( \pi_F \) has full weak topological support on \( (\Pi_1, \Gamma_1) \), i.e. \( \pi_F(N_{G_n,\epsilon}(Q)) > 0 \) for any \( n, G_n, Q \in \Pi_1 \) and \( \epsilon > 0 \), where \( N_{G_n,\epsilon}(Q) \) is defined in (3.8).
(ii) if for any $Q \in \Pi_2$, the conditional prior $p(N_{fQ,\epsilon}(Q)|k) > 0$ for any TV neighborhood of $Q$. Then, $\pi_F$ has full strong topological support on $(\Pi_2, \Gamma_2)$, i.e. $\pi_F(N_{\epsilon}(Q)) > 0$ for any $n, Q \in \Pi_2$ and $\epsilon > 0$, where $N_{\epsilon}(Q)$ is defined in (3.9).

Theorem 5(i) shows that if the probability distribution of $k$ is positive over $\mathbb{N}^d$ and for any $k \in \mathbb{N}^d$, the conditional distribution of $p(N_{Gn,\epsilon}(Q)|k) > 0$ for any weak neighbourhood of $Q$, then the Feller prior $\pi_F$ has full topological support on $\Pi_1$. Theorem 5(ii) states that if the probability distribution of $k$ is positive over $\mathbb{N}^d$ and for any $k \in \mathbb{N}^d$, the conditional distribution of $p(N_{fQ,\epsilon}(Q)|k) > 0$ for any TV neighbourhood of $Q$, then any given neighbourhood of a given absolute continuous probability distribution function has positive $\pi_F$-probability, where the neighbourhood is defined in terms of Kolmogorv-Smirnov distance. The two parts of the theorem ensure that every probability distribution function on compact set $\Delta$ is in the topology of weak convergence of the Feller prior. Moreover, every absolutely continuous distribution is in the topology of uniform convergence of Feller prior, given $p(k) > 0$ and the conditional probability $p(\cdot|k)$ has full support on some given neighbourhood.

Since any multivariate continuous density on $\Delta$ can be uniformly approximated by a sequence of multivariate Feller densities, we can define a measure on the space of multivariate continuous densities on $\Delta$ by defining a measure on the set of multivariate Feller densities. Similar to the proof of Theorem 4, we can define a random multivariate Feller density as a random function of the form $b(x; k, G)$ in (3.7), and use it as a prior on the space of multivariate continuous densities on $\Delta$. The validity of the proposed prior can be proved similarly to the proof in Theorem 5.

3.5 Application of Multivariate Feller Priors

According to previous discussions, multivariate Feller operators with ERS can be used for inference on an unknown multivariate density function. Remember the spectral
density, after normalizing, behaves like a common density function. Therefore, we will first briefly discuss the nonparametric Bayesian approach for multivariate density estimation in the following pages. Then a computational algorithm for stimulating from the posterior of the multivariate density function will be given.

3.5.1 A Brief Introduction of Density Estimation

Density estimation plays an important role on statistical inference problems. A variety of nonparametric Bayesian models have been used to estimate density functions. For example, Lo (1984), Escobar and West (1995) and Gasparini (1996) investigated Dirichlet Process Mixtures (DP mixtures). Lavine (1992) and Lavine (1994) proposed Polya trees. Another nonparametric Bayesian approach for density estimation is proposed by Petrone (1999) using Bernstein polynomial priors. However, these nonparametric Bayesian approaches listed above focus on univariate case. For multivariate case, Scott (1992) summarized plenty of traditional parametric and nonparametric approaches for multivariate density estimation like frequency polygons, averaged shifted histogram and kernel density estimators, but little has been explored for multivariate density estimation using nonparametric Bayesian approach. That is a void we try to fill in this dissertation.

Let \( \{X_1, X_2, \ldots\} \) be a sequence of exchangeable random variables with values in a set \( E \subset \mathbb{R} \). Then, de Finetti’s representation theorem shows that, the joint distribution function can be written as:

\[
P(X_1 \leq x_1, \ldots, X_n \leq x_n) = \int_{\mathcal{P}} \prod_{i=1}^{n} F(x_i) d\pi(F),
\]

where \( \pi \) is the prior on the class \( \mathcal{P} \) of all the distribution functions on \( E \). More generally, Hewitt and Savage (1955) have generalized de Finetti’s result. Let \( \{X_1, X_2, \ldots\} \) be a sequence of exchangeable random vectors with values in a set \( \chi \subset \mathbb{R}^d \). Then similar to de Finetti’s representation, the joint distribution function of \( X_1, \ldots, X_n \)
can be written as:

\[ P(X_1 \in A_1, \ldots, X_n \in A_n) = \int_{\mathcal{P}} \prod_{i=1}^{n} F(A_i) d\pi(F), \tag{3.10} \]

where \( \pi \) is the prior on the class \( \mathcal{P} \) of all the distribution functions on \( \chi \).

As shown in section 4, to define a multivariate Feller prior \( \pi_F \), we consider \((k, G)\) as random quantities and assign them a probability law. In particular, we will focus on the following two cases.

1. Consider \( k \) and \( G \) are independent. \( k \) has a probability function \( p \) and \( G \) is a Dirichlet process with parameters \((\alpha, G_0)\), where \( \alpha \) is the scale parameter and \( G_0 \) is the base function expressing the initial guess on \( G \). Under this scenario, we call such \( \pi_F \) a Feller-Dirichlet prior with parameters \((p, \alpha, G_0)\).

2. Consider \( k \) and \( G \) are independent. \( k \) has a probability function \( p \) and \( G \) is a Pitman-Yor process with parameters \((\alpha, 0, G_0)\), where \( \alpha \) is the discount parameter and \( G_0 \) is the base distribution, we call such \( \pi_F \) a Feller-PY prior with parameters \((p, \alpha, 0, G_0)\).

We assume that the probability law \( \mathcal{P} \) does not concentrate on its boundary, so the random distribution function \( B(\cdot; k, G) \) is almost surely absolute continuous with density \( b(\cdot; k, G) \). In such case, choosing a Feller prior with random scheme \( \{Z_{kx}, k = (k_1, \ldots, k_d), k_1, \ldots, k_d = 1, 2, \ldots, x \in \chi \subseteq \mathbb{R}^d\} \) in (3.10) is equivalent to the following hierarchical structure.

(i) For any \( n, X_1, \ldots, X_n|k; G \) are conditionally i.i.d with common density

\[
\begin{align*}
    b(x; k, G) &= \int \cdots \int f_{k_1}(x_1; z_1) dG(z_1, \ldots, z_d) \\
                 &= \int f(x; z, k) dG(z),
\end{align*}
\]

where \( f_{k_i}(x_i; z_i) \) is the kernel density associated to the given random scheme.
(ii) \((k, G)\) have the joint probability law \(P\).

Computation in the above structure is in general complicated. However, MCMC approximations of posterior or of other quantities of interest are possible. In order to simplify the computation, we will extend the algorithm proposed by Ishwaren and James (2001) for stimulating from the posterior corresponding to a Feller prior.

### 3.5.2 Blocked Gibbs Sampling

The algorithm is called *blocked gibbs sampling* (Ishwaren and James, 2001) since such algorithm can update parameters "blockwise", and it is based on the well-known characterization of the Dirichlet process as infinite sum (Sethuraman, 1994). If \(G \sim \mathcal{D}(\alpha, G_0)\), then \(G\) is almost surely a discrete distribution function

\[
DP(\alpha G_0)(\cdot) = \sum_{j=1}^{\infty} p_j \delta_{(-\infty,z_j]}(x),
\]  

(3.11)

where \(\delta_A(\cdot)\) is the indicator function of the set \(A\), \(Z_1, Z_2, \ldots\) are independent draws from the base function \(G_0\), \(p_1 = V_1\), \(p_j = (1-V_1)(1-V_2)\cdots(1-V_{j-1})V_j\), \(j \geq 2\) and \(V_1, V_2, \ldots\) are i.i.d chosen from a beta density with parameters \((1, \alpha)\), \(\alpha > 0\).

For Pitman-Yor process, in particular, for \(PY(\alpha, 0, G_0)\) process, Pitman and Yor (1997) provided a similar representative form:

\[
PY(\alpha, 0, G_0)(x) = \sum_{j=1}^{\infty} p_j \delta_{(-\infty,z_j]}(x),
\]  

(3.12)

where \(Z_1, Z_2, \ldots\) are independent draws from the base function \(G_0\), \(p_1 = V_1\), \(p_j = (1-V_1)(1-V_2)\cdots(1-V_{j-1})V_j\), \(j \geq 2\) and \(V_1, V_2, \ldots\) are i.i.d chosen from a beta density with parameters \((1-\alpha, j\alpha)\), \(0 < \alpha < 1\).

The computational algorithm is based on the truncation of the above infinite sum.
representation of Dirichlet process and Pitman-Yor process.

\[ DP_N(x)(\alpha, G_0)(x) = \sum_{j=1}^{N} p_j \delta_{(-\infty, z_j]}(x). \]

of series (3.11) and the truncation

\[ PY_N(\alpha, 0, G_0)(x) = \sum_{j=1}^{N} p_j \delta_{(-\infty, z_j]}(x). \]

of series (3.12) to a finite \( N \), such that the residual probability is negligible.

The hierarchical structure proposed above is not computationally manageable since it involves integral in stage (i), which makes fairly difficult in implementing efficient computational method. So we will rewrite the above hierarchical structure as follows by inducing auxiliary random variables \( Y_1, \ldots, Y_n \) in our model:

\[
(X_i|Y_1, \ldots, Y_n, k) \sim f(x; y, k), \quad i = 1, \ldots, n,

(Y_i|k, G) \sim G_i, \quad i = 1, \ldots, n,

k \sim p(k),

G \sim \mathcal{P},
\]

where \( \mathcal{P} \) stands for our stick-breaking prior \( DP_N(\alpha, G_0)(\cdot) \) or \( PY_N(\alpha, 0, G_0)(\cdot) \).

Further, the above model can be rewritten as

\[
(X_i|Z, J, k) \sim f(x; Z_{j_i}, k), \quad i = 1, \ldots, n,

(J_i|p) \sim \sum_{j=1}^{N} p_j \delta_{(\cdot)}(\cdot), \quad i = 1, \ldots, n,

(p, Z) \sim \pi(p) \times G_0^N(Z),

k \sim p(k).
\]

where \( J = (J_1, \ldots, J_n) \), \( Z = (Z_1, \ldots, Z_n) \), \( p = (p_1, \ldots, p_N) \) are the random weights in truncated Dirichlet process or Pitman-Yor process, and \( Z_j \) are i.i.d from \( G_0 \). Notice that the relationship between the above two structure is \( Y_i = Z_{J_i} \). That is,
Let \( J_i, i = 1, \ldots, n \) act as classification variables to identify which \( Z_{J_i} \) is associated with each \( Y_i \).

Let \( \{J_{1}^*, \ldots, J_{m}^*\} \) denote the set of current unique values of \( J = (J_1, \ldots, J_n) \), to run *blocked Gibbs sampling*, we draw values in the following order:

(i) Conditional for \( Z \): simulate \( Z_j \sim G_0 \) for each \( j \in J - \{J_{1}^*, \ldots, J_{m}^*\} \). Also, draw \((Z_{J_l^*} | J, k, X)\) from the density

\[
 f(Z_{J_l^*} | J, k, X) \propto g_0(Z_{J_l^*}) \prod_{\{i:j_i = J_l^*\}} f(X_i | Z_{J_i}, k), \ l = 1, \ldots, m,
\]

where \( g_0 \) is the density of \( G_0 \).

(ii) Conditional for \( p \): \( p_1 = V_1, p_j = (1 - V_1)(1 - V_2) \cdots (1 - V_{j-1})V_j, j = 2, \ldots, N - 1 \)

where \( V_j \sim Beta(1 + m_j, \alpha + m_j + \cdots + m_N) \) and \( m_j \) is the number of \( j_1, \ldots, j_n \) which equal to \( j, j = 1, \ldots, N - 1, \sum_{j=1}^{N} p_j = 1 \).

(iii) Conditional for \( J \): Draw value

\[
 (J_i | Z, p, k, X) \sim \sum_{j=1}^{N} p_{j,i} \delta_j(\cdot), \ i = 1, \ldots, n,
\]

where

\[
 (p_{1,i}, \ldots, p_{N,i}) \propto (p_1 f(X_i | Z_1, k), \ldots, (p_N f(X_i | Z_N, k))
\]

(iv) Conditional for \( k \): Draw \( k \) from the density (notice that \( Y_i = Z_{J_i} \))

\[
 f(k | Z, J, X) \propto p(k) \prod_{i=1}^{n} f(X_i | Y_i, k).
\]
Given a MCMC sample \( \{k^{(s)}, p_1^{(s)}, \ldots, p_N^{(s)}, Z_1^{(s)}, \ldots, Z_N^{(s)}\} \) from the posterior of 
\((Z, j, p, k|X)\), \( s = 1, \ldots, S \), the number of Chains, we can approximate the density estimation
\[
b_n(x) = \frac{1}{S} \sum_{s=1}^{S} \sum_{j=1}^{N} p_j^{(s)} f(x; Z_j^{(s)}, k) \tag{3.13}
\]
Notice in step(i), the draws from the conditional posterior could be done exactly if 
\( G_0 \) is the conjugate prior for \( f \). However, if a non-conjugate prior is given, we use 
Metroplis-Hasting algorithm.

### 3.5.3 Simulation Study

In this section, we present the performance of estimated density described in (3.13) for data simulated from different distributions using the Feller priors proposed. For simplicity and for better visualization, we focus on the case \( d = 2 \) and set \( k = k_1 \) for all simulations. To calibrate the accuracy of \( b_n(x) \) as an estimate of true density \( f^* \), we define a Hellinger error as
\[
||b_n(x) - f^*(x)||_H = \sqrt{\frac{1}{2} \int (\sqrt{b_n(x)} - \sqrt{f^*(x)})^2 dx}
\]
The different cases we considered are as follows.

- **Case1:** In this case, we simulated \( n = 90 \) observations \( X_i \) from a bivariate normal distribution with mean vector \( \mu_0 = (0, 0) \) and variance-covariance matrix
\[
\Sigma_0 = \begin{pmatrix}
1 & 0.5 \\
0.5 & 1
\end{pmatrix}.
\]

- **Case2:** In this case, we simulated \( n = 90 \) observations \( X_i \) from a standard bivariate logistic distribution with pdf
\[
f^*(x, y) = \frac{2e^{-x}e^{-y}}{(1 + e^{-x} + e^{-y})^3}, \quad x, y \in \mathbb{R}.
\]
Since the simulated data are in $(-\infty, \infty)$, we find it reasonable to use a multivariate Feller prior with a Gaussian random scheme, which is discussed in Example 3 of section 2.3, here we set $\sigma_1 = \sigma_2 = 25$, $k = 300$. The *blocked gibbs sampling* are applied to two different priors.

1. **Feller-Dirichlet prior:** We set truncation $N = 100$, the scale parameter $\alpha = 2$ and the base function $G_0$ bivariate normal with mean vector $\mu = (0, 0)$ and variance-covariance matrix $\Sigma = \begin{pmatrix} 10 & 3 \\ 3 & 10 \end{pmatrix}$.

2. **Feller-PY prior:** We set truncation $N = 100$, the discount parameter $\alpha = 0.25$ and the base function $G_0$ the same as used in Feller-Dirichlet prior.

We vary the number of MCMC chains $S$ by letting $S = 2, 5, \text{or } 10$. For each individual chain, we use the *blocked gibbs sampling* described in previous section. After a burn-in of 500, the Markov Chains become stable. Figure 3.1 plots the first 1000 samples of $k$ in a single MCMC chain, indicating the chain quickly converges after burn-in.

![Figure 3.1: Trace of k](image)
Table 3.1 gives the Hellinger distance (Hellinger error) between the estimated density function and the true density function using 2-dimensional Feller-Dirichlet prior and 2-dimensional Feller-PY prior for data simulated from bivariate normal distribution and bivariate logistic distribution. Hellinger distance is used to quantify the similarity between two probability distributions. The smaller the value, the more similar between the two distributions. We note that the overall performance of the estimated density functions improves as the number of MCMC chains increases for both priors (The increases of number of MCMC chains means that we have a larger MCMC samples, and we use several short MCMC chains instead of a long MCMC chain so that we can do parallel computation). Figures 3.2, 3.3, 3.4, 3.5 give us a visual way to evaluate the performance of the estimated density functions. It’s clear that estimated density is closer to the true density as the number of MCMC chains increases from 2 to 10. Notice that when data are simulated from the bivariate logistic distribution, the base distribution (our guess about the true distribution) we use is not even from the family of bivariate logistic distributions but a general bivariate normal distribution. However, as shown in table 3.1 and Figures 3.4 and 3.5, we still get a good estimated density. This result indicates the flexibility and robustness of the algorithm we propose here. For comparison, we also use a kernel density estimator (Duong and Hazelton, 2003) for the two target distributions considered here. The hellinger distance is 0.158 when target distribution is bivariate normal and is 0.162 when target distribution is bivariate logistic. Figures 3.6 and 3.7 show the performance of kernel density estimator. The comparison indicates that the performance of multi-dimensional Feller prior is comparable to the well-established kernel density estimator.
Table 3.1: *Hellinger error of estimated density functions using Feller-Dirichlet prior and Feller-PY prior for data from bivariate normal distribution and bivariate logistic distribution*

<table>
<thead>
<tr>
<th># of chains</th>
<th>Bivariate normal</th>
<th>Bivariate logistic</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Feller-Diri(2)</td>
<td>Feller-PY(0.25)</td>
</tr>
<tr>
<td>S=2</td>
<td>0.157</td>
<td>0.108</td>
</tr>
<tr>
<td>S=5</td>
<td>0.120</td>
<td>0.107</td>
</tr>
<tr>
<td>S=10</td>
<td>0.090</td>
<td>0.111</td>
</tr>
</tbody>
</table>

**Real Data Application**

As a supplement to the simulation study, we also present a real data application here. The data is the well-known IRIS data introduced by R. A. Fisher (Fisher, 1936). The data set consists of 50 samples from each of three species of Iris (Iris setosa, Iris virginica and Iris versicolor). Four features were measured from each sample: the length and the width of the sepals and petals, in centimetres, and we choose the petal length and width as our bivariate observations. We consider both 2-dimensional Feller-Dirichlet prior and 2-dimensional Feller-PY prior with Gaussian random scheme for density estimation. Here we set $\sigma_1 = \sigma_2 = 25$, $k = 300$. For both priors, we set $N = 150$, base function $G_0$ the bivariate normal distribution with mean vector $\mu = \hat{X}$, the sample mean vector, variance-covariance matrix $\Sigma = \hat{\Sigma}$, the sample variance-covariance matrix. For the 2-dimensional Feller-Dirichlet prior, we set $\alpha = 2$, for the 2-dimensional Feller-PY prior, we set $\alpha = 0.25$. Figures 3.8 and 3.9 shows the estimated density functions. The contour plots accurately reflect the two clusters in the data set, as described in Fisher’s paper.
Figure 3.2: Density function estimate using Feller-Dirichlet(2) prior for data from bivariate normal. The red curve on 3 graphs are contour plot of true density, contour plot of estimated density are (i) S=2, black curve (ii) S=5, green curve (iii) S=10, blue curve.

Figure 3.3: Density function estimate using Feller-PY(0.25) prior for data from bivariate normal. The red curve on 3 graphs are contour plot of true density, contour plot of estimated density are (i) S=2, black curve (ii) S=5, green curve (iii) S=10, blue curve.
Figure 3.4: Density function estimate using Feller-Dirichlet(2) prior for data from bivariate logistic. The red curve on 3 graphs are contour plot of true density, contour plot of estimated density are (i) S=2, black curve (ii) S=5, green curve (iii) S=10, blue curve.

Figure 3.5: Density function estimate using Feller-PY(0.25) prior for data from bivariate logistic. The red curve on 3 graphs are contour plot of true density, contour plot of estimated density are (i) S=2, black curve (ii) S=5, green curve (iii) S=10, blue curve.
Figure 3.6: Bivariate Kernel Density function estimate for data from bivariate normal. The green curve on is contour plot of true density, contour plot of estimated density are black curve.
Figure 3.7: Bivariate Kernel Density function estimate for data from bivariate logistic. The green curve on is contour plot of true density, contour plot of estimated density is black curve.
Figure 3.8: Density estimate using Feller-Dirichlet(2) prior for data iris. Contour plot of estimated density are (i) S=2, black curve (ii) S=5, green curve (iii) S=10, blue curve.

Figure 3.9: Density estimate using Feller-PY(0.25) prior for data iris. Contour plot of estimated density are (i) S=2, black curve (ii) S=5, green curve (iii) S=10, blue curve.
3.6 Discussion and Future Work

3.6.1 Conclusion and Discussion

In this chapter, we propose a nonparametric prior, Feller prior, and establish its existence and validity. As Petrone (1999b) pointed out, a "good" class of prior distributions should satisfy the following properties.

(i) it should have full topological support;

(ii) it can easily select absolute continuous distribution function with continuous derivatives;

(iii) The quantities (e.g. density functions) on which one should express a prior on should have a natural interpretation;

(iv) For computational aspects, it should have easily manageable posterior.

For (i), according to Theorem 5, the proposed Feller prior has full topological support under some certain kind of topology. For (ii), Proposition 1 shows that the Feller prior can select absolute continuous distribution function and the selected distribution function has derivative which we called "Feller density". For (iii), the quantity we considered here is multivariate density function, and the Feller prior can be specified in a fairly natural way by first expressing a initial guess on the unknown density function, then by smoothing the estimated density function using Feller-type approximation. For (iv), the proposed blocked gibbs sampling algorithm provides a simple way to stimulate MCMC samples from the posterior distribution even if the posterior distribution doesn’t have a simple analytic expression.

In addition, from the application of Feller prior in multivariate density estimation, we can see that the proposed hierarchical structure is quite general and flexible. Since we have both discrete and continuous exponential random scheme, we can easily handle
different types of data. Bernstein Polynomials can only deal with data restricted on $[0, 1]^d$, otherwise, a transformation is needed before we apply Bernstein Polynomials. However, Feller priors can be applied directly to observations with different range. Furthermore, the simulation study shows that the performance of this hierarchical structure is still good even if the initial guess of the unknown distribution is not correct, which indicates that the proposed method is quite robust.

Further, the application of multivariate density estimation is of great use in real-world problems. For example, observational study is becoming more and more popular in Pharmaceutical industry since it is much more time-saving and cost-saving. However, lack of randomization makes the conclusion unreliable: is the outcome due to treatment effect, or due to the difference of important characteristics between the treatment and comparison group? Statisticians have developed several methods to balance the groups and reduce the possible bias like propensity score method, but these methods are indirect, e.g., propensity score is the conditional probability of receiving a given exposure (treatment) given a vector of measured covariates. Unlike those methods, multivariate density estimation can give a direct comparison between treatment and comparison group given possible confounding covariates.

3.6.2 Future Work

We have briefly introduced the spectral density of a random field at the beginning of this chapter. It is fairly complicated for a nonparametric Bayesian approach on the spectral density. If we assume the random fields are second-order stationary, then we can use Whittle’s Likelihood (Whittle, 1954) to evaluate the smoothing of the
periodogram (3.3), and the Whittle’s approximation is defined as follows.

\[ \frac{1}{2} \sum_{\omega \in \Omega} \{ \log f(\omega) + I(\omega)/f(\omega) \} \]

where the summation is over the set \( \Omega \), \( I \) is the periodogram (3.3) and \( f \) is the spectral density. Whittle’s likelihood is as follows.

\[ L(f|X_1, \ldots, X_n) = \prod_{\omega \in \Omega} \exp\left[ -\frac{1}{2} \{ \log f(\omega) + I(\omega)/f(\omega) \} \right] \]

When applied certain ERS, we can write the posterior distribution and then apply blocked gibbs sampling for stimulating samples from the corresponding posterior.

Whittle’s likelihood is useful when we assume the random fields are second-order stationary. An interesting problem would be what if the random fields are nonstationary. Dahlhaus (1997) proposed a concept ”locally stationary” in fitting time series models to nonstationary process. He suggested that we can treat a nonstationary time series ”locally stationary”, i.e., split the entire time interval into several different sub-intervals, and within each sub-interval, the process can be viewed as ”stationary”. Under such assumption, a ”local periodogram” will replace the usual periodogram over data segments. Though the properties of local periodogram are not fully developed, it still leads us to consider the extension of the local periodogram to spatial case in order for the nonstationary random fields. This part is still under investigation and we will investigate other nonparametric Bayesian approach in our future work as well.
Appendices

Appendix A: Proof of Theorem 1

From (2.4), we have that, under the true parameters $\eta_0$,

$$
E\{\ell(\eta)\} = -(nm/2) \log \sigma^2 + \log |S_{nm}(\xi)| - (2\sigma^2)^{-1} [(\beta_0 - \beta)' X'_{nm} S'_{nm}(\xi) S_{nm}(\xi) X_{nm}(\beta_0 - \beta)] + (2\sigma^2)^{-1} \left[ \sigma_0^2 \text{tr} \left\{ S'_{0nm} S'_{nm}(\xi) S_{nm}(\xi) S_{0nm}^{-1} \right\} \right].
$$

The first derivatives of $E\ell(\eta)$ with respect to $\beta$ and $\sigma^2$ are, respectively,

$$
\frac{\partial E\ell(\eta)}{\partial \beta} = (\sigma^2)^{-1} X'_{nm} S'_{nm}(\xi) X_{nm}(\beta_0 - \beta),
$$

$$
\frac{\partial E\ell(\eta)}{\partial \sigma^2} = (2\sigma^2)^{-1} [E(\nu'_{nm} \nu_{nm}) - nm\sigma^2].
$$

Thus the maximizers of $E\ell(\eta)$ are

$$
\beta^*_{nm}(\xi) = \beta_0,
$$

$$
\sigma^*_{nm}^{2}(\xi) = (nm)^{-1} \sigma_0^2 \text{tr} \{ S'_{0nm} S'_{nm}(\xi) S_{nm}(\xi) S_{0nm}^{-1} \}.\n$$

Let $g_{nm}(\xi) = E[\ell(\beta^*_{nm}(\xi), \xi, \sigma^*_{nm}^{2}(\xi))] = -(nm/2) \log \sigma^*_{nm}^{2}(\xi) + \log |S_{nm}(\xi)| - nm/2$.

We establish the consistency of $\hat{\xi}_{nm}$ by showing that $\sup_{\xi} h_n(nm)^{-1}|\ell(\xi) - g_{nm}(\xi)| = o_p(1)$ and that $h_n(nm)^{-1} g_{nm}(\xi)$ is identifiably unique (White, 1994).
To show sup $\xi \in \Xi \ h_n(nm)^{-1}|\ell(\xi) - g_{nm}(\xi)| = o_p(1)$, we have

$$h_n(nm)^{-1}(\ell(\xi) - g_{nm}(\xi))$$

$$= \frac{-h_n}{2} \{ \log \hat{\sigma}_{nm}^2(\xi) - \log \sigma_{nm}^2(\xi) \}$$

$$= -h_n \{ 2\hat{\sigma}_{nm}^2(\xi) \}^{-1} \{ \sigma_{nm}^2(\xi) - \sigma_{nm}^2(\xi) \},$$

where $\hat{\sigma}_{nm}^2(\xi) = \lambda \sigma_{nm}^2(\xi) + (1 - \lambda) \hat{\sigma}_{nm}^2(\xi)$ for some $\lambda \in (0, 1)$

and $\hat{\sigma}_{nm}^2(\xi) = (nm)^{-1} \nu_{0nm} B_{nm}(\xi) \nu_{0nm}$ with

$$B_{nm}(\xi)$$

$$= S_{0nm}^{-1} S_{nm}(\xi)' [ I_{nm} - S_{nm}(\xi) X_{nm} \{ X_{nm}' S_{nm}(\xi) S_{nm}(\xi) X_{nm} \}^{-1} X_{nm}' S_{0nm}^{-1} S_{nm}(\xi) S_{0nm}^{-1} \nu_{0nm}.$$

First, we have that $h_n \{ \hat{\sigma}_{nm}^2(\xi) - \sigma_{nm}^2(\xi) \} = o_p(1)$. By (A.1)-(A.4), we know

$$h_n(nm)^{-1} \nu_{0nm} S_{0nm}^{-1} S_{nm}(\xi) S_{nm}(\xi) S_{0nm}^{-1} \nu_{0nm}$$

$$- h_n(nm)^{-1} E \left\{ \nu_{0nm} S_{0nm}^{-1} S_{nm}(\xi) S_{nm}(\xi) S_{0nm}^{-1} \nu_{0nm} \right\} = o_p(1),$$

as $nm/h_n \to \infty$, where the convergence is uniform on $\Xi$ because of the linear-quadratic form in $\xi$ and by corollary 2.2 of Newey (1991).

By (A.1)-(A.5),

$$h_n(nm)^{-1} \nu_{0nm} S_{0nm}^{-1} S_{nm}(\xi) S_{nm}(\xi) X_{nm} \{ X_{nm}' S_{nm}(\xi) S_{nm}(\xi) X_{nm} \}^{-1} X_{nm}' S_{nm}(\xi) S_{nm}(\xi) S_{0nm}^{-1} \nu_{0nm}$$

$$= h_n(nm)^{-1} \left\{ (nm)^{-1/2} X_{nm}' S_{nm}(\xi) S_{nm}(\xi) S_{0nm}^{-1} \nu_{0nm} \right\}'$$

$$\times \left\{ (nm)^{-1} X_{nm}' S_{nm}(\xi) S_{nm}(\xi) X_{nm} \right\}^{-1}$$

$$\times \left\{ (nm)^{-1/2} X_{nm}' S_{nm}(\xi) S_{nm}(\xi) S_{0nm}^{-1} \nu_{0nm} \right\}$$

$$= o_p(1)$$

Again, the convergence is uniform on $\Xi$, since $(nm)^{-1} X_{nm}' S_{nm}(\xi) S_{nm}(\xi) S_{0nm}^{-1} \nu_{0nm} = $
\[ o_p(1) \text{ uniformly on } \Xi, \quad \{(nm)^{-1}X'_{nm}S''_{nm}(\xi)S_{nm}(\xi)X_{nm}\}^{-1} \text{ is uniformly bounded in } l_\infty, \text{ and the boundedness is uniform on } \Xi \text{ by } (A.5). \text{ Thus it follows that, uniformly on } \Xi, \]

\[
\begin{align*}
& h_n \left\{ \hat{\sigma}_{nm}^2(\xi) - \sigma_{nm}^2(\xi) \right\} \\
= & \quad h_n(nm)^{-1}\nu'_{nm}S'_{0nm}(\xi)S_{nm}(\xi)S_{0nm}^{-1}\nu_{0nm} \\
& - h_n(nm)^{-1}E\{\nu'_{nm}S'_{0nm}(\xi)S_{nm}(\xi)S_{0nm}^{-1}\nu_{0nm}\} \\
& - h_n(nm)^{-1}\nu'_{nm}S'_{0nm}(\xi)S_{nm}(\xi)X_{nm}\{X'_{nm}S''_{nm}(\xi)S_{nm}(\xi)X_{nm}\}^{-1} \\
& X'_{nm}S''_{nm}(\xi)S_{nm}(\xi)S_{0nm}^{-1}\nu_{0nm} \\
= & \quad o_p(1)
\end{align*}
\]

Next, we show the uniform boundedness of \( \hat{\sigma}_{nm}^2(\xi) \). By Jensen’s inequality,

\[
\begin{align*}
&(nm)^{-1}\{g_{nm}(\xi) - g_{nm}(\xi_0)\} \\
= & \quad (nm)^{-1}(\log |S_{nm}(\xi)| - \log |S_{0nm}|) - 1/2\{\log \sigma_{nm}^2(\xi) - \log \sigma_0^2\} \\
= & \quad \frac{1}{2} \log \frac{|S'_{0nm}S''_{nm}(\xi)S_{nm}(\xi)S_{0nm}^{-1}|^{nm^{-1}}}{(nm)^{-1}tr\{S'_{0nm}S''_{nm}(\xi)S_{nm}(\xi)S_{0nm}^{-1}\}} \leq 0
\end{align*}
\]

for \( \xi \in \Xi \). Under (A.1)-(A.4),

\[
\begin{align*}
&(nm)^{-1}(\log |S_{nm}(\xi)| - \log |S_{0nm}|) \\
= & \quad -(nm)^{-1}\left[ tr\{S_{nm}(\tilde{\xi}) diag W^*_{n1}, \ldots, tr\{S_{nm}(\tilde{\xi}) diag W^*_{nq}, tr\{1/\alpha S_{nm}(\tilde{\xi})A(\alpha)\} \} \right] \\
(\xi - \xi_0) \\
= & \quad - \sum_{k=1}^q O(h_n^{-1})(\theta_k - \theta_{0k}),
\end{align*}
\]

where \( W^*_nk = I_m \otimes W_{nk}, k = 1, \ldots, q, \tilde{\xi} = \lambda \xi + (1 - \lambda)\xi_0 \) for some \( \lambda \in (0, 1) \) and \( A(\alpha) = \alpha F_{nm} \). Thus,

\[
\begin{align*}
\log \sigma_{nm}^2(\xi) & \quad = \quad -2(nm)^{-1}\{g_{nm}(\xi) - g_{nm}(\xi_0)\} + 2(nm)^{-1}\{\log |S_{nm}(\xi)| - \log |S_{0nm}|\} \\
& \quad + \log \sigma_0^2 \\
& \quad \geq \quad 2(nm)^{-1}\{\log |S_{nm}(\xi)| - \log |S_{0nm}|\} + \log \sigma_0^2
\end{align*}
\]
which is bounded from below uniformly on $\Xi$, and $\sigma_{nm}^2(\xi)$ is bounded away from 0 uniformly on $\Xi$. Since $\hat{\sigma}_{nm}(\xi) - \sigma_{nm}(\xi) = o_p(1)$ uniformly on $\Xi$, $\hat{\sigma}_{nm}(\xi)$ is bounded away from 0 in probability uniformly on $\Xi$. Hence, $\sup_{\xi \in \Xi} h_n(nm)^{-1}|\ell(\xi) - g_{nm}(\xi)| = o_p(1)$.

To show the identifiable uniqueness of $h_n(nm)^{-1}g_{nm}(\xi)$, we note that $h_n(nm)^{-1}g_{nm}(\xi)$ is uniformly equicontinuous. In

$$h_n(nm)^{-1}\{g_{nm}(\xi_1) - g_{nm}(\xi_2)\} = h_n(nm)^{-1}\{\log |S_{nm}(\xi_1)| - \log |S_{nm}(\xi_2)|\} - h_n(nm)^{-1}\{2\log \sigma_{nm}(\xi_1) - \log \sigma_{nm}(\xi_2)\} - h_n(nm)^{-1}\{\sigma_{nm}(\xi_1) - \sigma_{nm}(\xi_2)\},$$

where $\sigma_{nm}^2 = \lambda \sigma_{nm}^2(\xi_1) + (1 - \lambda) \sigma_{nm}^2(\xi_2)$ for some $\lambda \in (0, 1)$ and is bounded away from 0, both terms are uniformly equicontinuous. Since by (3.14), $h_n(nm)^{-1}(\log |S_{nm}(\xi_1)| - \log |S_{nm}(\xi_2)|) = -O(1) \sum_{k=1}^q (\theta_k - \theta_{0k})$ and with $\bar{\xi} = \lambda \xi_1 + (1 - \lambda) \xi_2$ for some $\lambda \in (0, 1)$, we have

$$h_n \{\sigma_{nm}(\xi_1) - \sigma_{nm}(\xi_2)\} = -h_n(nm)^{-1}\sigma_0^2 \sum_{k=1}^q tr \left\{ S_{0nm}^{-1} diag(I_m \otimes W_{nk}) S_{nm}(\bar{\xi}) S_{0nm}^{-1} + S_{0nm}^{-1} S_{nm}(\bar{\xi}) diag(I_m \otimes W_{nk}) S_{0nm}^{-1} \right\} (\theta_{1k} - \theta_{2k}) + \sigma_0^2 O(1)(\alpha_1 - \alpha_2)$$

$$= -\sigma_0^2 O(1) \sum_{k=1}^q (\theta_{1k} - \theta_{2k}) + \sigma_0^2 O(1)(\alpha_1 - \alpha_2).$$

Thus together with (A.6) and (3.14), $h_n(nm)^{-1}g_{nm}(\xi)$ is identifiably unique. Thus, the MLE of $\xi$ is a consistent estimator.

The consistency of $\hat{\sigma}_{nm}^2(\bar{\xi}_{nm})$ can be derived directly from the consistency of $\sigma_{nm}^2(\bar{\xi}_{nm})$.  

56
Further,

\[
\hat{\beta}_{nm}(\hat{\xi}_{nm}) = \beta_0 + \{X'_{nm} S'_{nm}(\hat{\xi}_{nm}) S_{nm}(\hat{\xi}_{nm}) X_{nm}\}^{-1} X'_{nm} S'_{nm}(\hat{\xi}_{nm}) S_{nm}(\hat{\xi}_{nm}) S^{-1}_{0nm} \nu_{0nm}
\]

\[
= \beta_0 + \{X'_{nm} S'_{nm}(\hat{\xi}_{nm}) S_{nm}(\hat{\xi}_{nm}) X_{nm}\}^{-1} X'_{nm} S'_{nm}(\hat{\xi}_{nm}) \nu_{0nm}
\]

\[
+ \sum_{k=1}^{q} (\theta_{0k} - \hat{\theta}_{nmk}) \{X'_{nm} S'_{nm}(\hat{\xi}_{nm}) S_{nm}(\hat{\xi}_{nm}) X_{nm}\}^{-1} X'_{nm} S'_{nm}(\hat{\xi}_{nm}) 
\]

\[
+ \{X'_{nm} S'_{nm}(\hat{\xi}_{nm}) S_{nm}(\hat{\xi}_{nm}) X_{nm}\}^{-1} X'_{nm} \text{diag}(\hat{\alpha}^2_{nm} I_n, \ldots, \hat{\alpha}^2_{nm} I_n, 0) S^{-1}_{0nm} \nu_{0nm}
\]

where the last three terms are of order \(o_p(1)\) by (A.1)-(A.5).
Appendix B: Proof of Theorem 2

The case $h_n = O(1)$

By (A.4), $\xi_0$ is in the interior of $\Xi$. Thus, for sufficiently small $\epsilon > 0$, we have $A_\epsilon = \{\eta : ||\eta - \eta_0|| < \epsilon\} \subset \mathbb{R}^p \times \Xi \times \mathbb{R}^+$ and $P(\hat{\eta}_{nm} \in A_\epsilon) \to 1$ as $n \to \infty$, where $|| \cdot ||$ denotes the Euclidean norm. We establish the asymptotic normality of the MLE by showing asymptotic normality of $(nm)^{-1}\frac{\partial \ell(\eta_0)}{\partial \eta}$ and convergence in probability of $(nm)^{-1}\frac{\partial^2 \ell(\hat{\eta}_{nm})}{\partial \eta \partial \eta'}$, where $\hat{\eta}_{nm} = \lambda \eta_0 + (1 - \lambda) \tilde{\eta}_{nm}$ for $\lambda \in (0, 1)$ converges to $\eta_0$ in probability.

For convergence of $(nm)^{-1}\frac{\partial^2 \ell(\hat{\eta}_{nm})}{\partial \eta \partial \eta'}$, we show that, under (A.1)-(A.5), $(nm)^{-1}\{\frac{\partial^2 \ell(\hat{\eta}_{nm})}{\partial \eta \partial \eta'} - \frac{\partial^2 \ell(\eta_0)}{\partial \eta \partial \eta'}\} = o_p(1)$ and $(nm)^{-1}\{\frac{\partial^2 \ell(\eta_0)}{\partial \eta \partial \eta'} - E\frac{\partial^2 \ell(\eta_0)}{\partial \eta \partial \eta'}\} = o_p(1)$. Here a matrix is said to be $O_p(1)$ (or $o_p(1)$) if all of its elements are of order $O_p(1)$ (or $o_p(1)$).

The second-order derivatives of $l(\eta)$ are

$$\frac{\partial^2 \ell(\eta)}{\partial \beta \partial \beta'} = - (\sigma^2)^{-1} X'_n S'_n(\xi) S_n(\xi) X_n,$$

$$\frac{\partial^2 \ell(\eta)}{\partial \beta \partial \sigma^2} = - (\sigma^4)^{-1} X'_n S'_n(\xi) \nu_{nm},$$

$$\frac{\partial^2 \ell(\eta)}{\partial \beta \partial \theta_k} = - (\sigma^2)^{-1} X'_n \text{diag}\{I_m \otimes (W'_{nk} S_n(\theta) + S'_n(\theta) W_{nk}) W_{nk}\}$$

$$S^{-1}_n(\xi) \nu_{nm}, \quad k = 1, \ldots, q,$$

$$\frac{\partial^2 \ell(\eta)}{\partial \beta \partial \alpha} = - (\sigma^2)^{-1} X'_n \{B_{nm} B'_{nm}\} S^{-1}_n(\xi) \nu_{nm}, \quad k = 1, \ldots, q$$

where

$$B_{nm} = \begin{pmatrix}
-\alpha I_n & S_n(\theta) \\
0 & \ddots & \ddots \\
0 & -\alpha I_n & S_n(\theta) \\
0 & \cdots & 0 & 0
\end{pmatrix}$$
We denote $S_n(\theta_0) = S_{0n}$

$$
\frac{\partial^2 \ell(\eta)}{\partial \sigma^2 \partial \sigma^2} = (2\sigma^0)^{-1}(-2\nu_{nm}^\prime \nu_{nm} + n\sigma^0),
$$

$$
\frac{\partial^2 \ell(\eta)}{\partial \theta_k \partial \sigma^2} = -(\sigma^4)^{-1}\nu_{nm}^\prime \text{diag}(W_{nk}S_n^{-1}(\theta), \ldots, W_{nk}S_n^{-1}(\theta))\nu_{nm},
$$

$$
\frac{\partial^2 \ell(\eta)}{\partial \theta_k \partial \theta_l} = -m \times \text{tr}\{W_{nk}S_n^{-1}(\theta)W_{nl}S_n^{-1}(\theta)\}
$$

\begin{align*}
&= -(\sigma^2)^{-1}\nu_{nm}^\prime S_{nm}^{-1}(\xi)\text{diag}(W_{nl}W_{nk}, \ldots, W_{nl}W_{nk})S_{nm}^{-1}(\xi)\nu_{nm}, \\
&\frac{\partial^2 \ell(\eta)}{\partial \alpha \partial \sigma^2} = -(\sigma^4)^{-1}\nu_{nm}^\prime F_{nm}S_{nn}^{-1}(\xi)\nu_{nm}, \\
&\frac{\partial^2 \ell(\eta)}{\partial \alpha \partial \alpha} = -(\sigma^2)^{-1}\nu_{nm}^\prime S_{nm}^{-1}(\xi)\text{diag}(I_n, \ldots, I_n, 0)S_{nm}^{-1}(\xi)\nu_{nm}, \\
&\frac{\partial^2 \ell(\eta)}{\partial \alpha \partial \theta_k} = 0.
\end{align*}

By (A.1) – (A.5), we have

\begin{align*}
(nm)^{-1}\left\{ \frac{\partial^2 \ell(\tilde{\eta}_{nm})}{\partial \beta \partial \beta^\prime} - \frac{\partial^2 \ell(\eta_0)}{\partial \beta \partial \beta^\prime} \right\} \\
= (nm)^{-1}\left\{ -(\sigma^2_{nm})^{-1}X_{nm}^\prime S_{nm}^\prime(\xi)S_{nm}(\xi)X_{nm} + (\sigma^2_{0})^{-1}X_{nm}^\prime S_{0nm}^\prime S_{0nm}X_{nm} \right\} \\
= (nm)^{-1}X_{nt}^\prime S_{0nm}^\prime S_{0nm}X_{nm}(\frac{1}{\sigma_0^2} - \frac{1}{\sigma_{nm}^2}) \\
\quad + (nm\tilde{\sigma}^2_{nm})^{-1}\{X_{nm}^\prime S_{0nm}^\prime S_{0nm}X_{nm} - X_{nm}^\prime S_{nm}^\prime(\tilde{\xi}_{nm})S_{nm}(\tilde{\xi}_{nm})X_{nm}\} \\
= o_p(1)
\end{align*}

\begin{align*}
(nm)^{-1}\left\{ \frac{\partial^2 \ell(\tilde{\eta}_{nm})}{\partial \beta \partial \sigma^2} - \frac{\partial^2 \ell(\eta_0)}{\partial \beta \partial \sigma^2} \right\} \\
= (nm)^{-1}X_{nm}^\prime S_{0nm}^\prime \nu_{0nm}(\frac{1}{\sigma_0^4} - \frac{1}{\sigma_{nm}^4}) + (nm\tilde{\sigma}^4_{nm})^{-1}X_{nm}^\prime E_{nm}S_{nm}^{-1}\nu_{0nm} \\
\quad -(\tilde{\sigma}^4_{nm})^{-1}X_{nm}^\prime S_{nm}^\prime(\tilde{\xi}_{nm})S_{nm}(\tilde{\xi}_{nm})X_{nm}(\beta_0 - \tilde{\beta}_{nm}) \\
= o_p(1)
\end{align*}

where

$$
E_{nm} = \begin{pmatrix} E_{1n} & E_{2n} \\
E_{2n}^\prime & \ddots & \ddots \\
\vdots & \ddots & \ddots & \ddots \\
E_{nm}^\prime & E_{nm} & \ddots & \ddots \\
E_{nm}^1 & E_{nm}^2 & \cdots & \cdots & \cdots & \ddots \\
\end{pmatrix}
$$
\[ E_{1m} = S_{0m}S_{0n} - S_{n}^*(\tilde{\theta}_{nm})S_n(\tilde{\theta}_{nm}) \]
\[ E_{2m} = \alpha \sum_{k=1}^{q} (\theta_{0k} - \tilde{\theta}_{nk})W_{nk} \]

\[
(nm)^{-1}\left\{ \frac{\partial^2 \ell(\tilde{\eta}_{nm})}{\partial \beta \partial \theta_k} - \frac{\partial^2 \ell(\eta_0)}{\partial \beta \partial \theta_k} \right\}
= (nm)^{-1}\left\{ -(\tilde{\sigma}_{nm})^{-1}X_{nm}'\{J_{nm}(\tilde{\xi}_{nm}) + J'_{nm}(\tilde{\xi}_{nm})\}S_{nm}(\tilde{\xi}_{nm})\nu_{nm}(\tilde{\xi}_{nm}) \right.
\]
\[ + (\sigma_0^2)^{-1}X_{nm}'(J_{0nm} + J'_{0nm})S_{0nm}\nu_{0nm} \}
\]
\[
= (nm)^{-1}\left\{ -(\tilde{\sigma}_{nm})^{-1}X_{nm}'\{J_{nm}(\tilde{\xi}_{nm}) + J'_{nm}(\tilde{\xi}_{nm})\}(Y_{nm} - X_{nm}\beta_0) \right.
\]
\[ + (\sigma_0^2)^{-1}X_{nm}'(J_{0nm} + J'_{0nm})(Y_{nm} - X_{nm}\beta_0) \]
\[ -(\tilde{\sigma}_{nm})^{-1}X_{nm}'\{J_{nm}(\tilde{\xi}_{nm}) + J'_{nm}(\tilde{\xi}_{nm})\}X_{nm}(\beta_0 - \tilde{\beta}_{nm}) \}
\]
\[ = o_p(1) \]

where

\[ J_{nm} = \text{diag}(W_{nm}'S_n(\theta), \cdots , W_{nm}'S_n(\theta)) \]

\[
(nm)^{-1}\left\{ \frac{\partial^2 \ell(\tilde{\eta}_{nm})}{\partial \beta \partial \alpha} - \frac{\partial^2 \ell(\eta_0)}{\partial \beta \partial \alpha} \right\}
= (nm)^{-1}\left\{ -(\tilde{\sigma}_{nm})^{-1}X_{nm}'\{B_{nm}(\tilde{\xi}_{nm}) + B'_{nm}(\tilde{\xi}_{nm})\}(Y_{nm} - X_{nm}\beta_{nm}) \right.
\]
\[ + (\sigma_0^2)^{-1}X_{nm}'(B_{0nm} + B'_{0nm})(Y_{nm} - X_{nm}\beta_0) \}
\]
\[
= (nm)^{-1}\left\{ -(\tilde{\sigma}_{nm})^{-1}X_{nm}'\{B_{nm}(\tilde{\xi}_{nm}) + B'_{nm}(\tilde{\xi}_{nm})\}(Y_{nm} - X_{nm}\beta_0) \right.
\]
\[ + (\sigma_0^2)^{-1}X_{nm}'(B_{0nm} + B'_{0nm})(Y_{nm} - X_{nm}\beta_0) \]
\[ -(\tilde{\sigma}_{nm})^{-1}X_{nm}'\{B_{nm}(\tilde{\xi}_{nm}) + B'_{nm}(\tilde{\xi}_{nm})\}X_{nm}(\beta_0 - \tilde{\beta}_{nm}) \}
\]
\[ = o_p(1) \]
\begin{align*}
(nm)^{-1}\{\frac{\partial^2 \ell(\tilde{\eta}_{nm})}{\partial \theta_k \partial \theta_l} - \frac{\partial^2 \ell(\eta_0)}{\partial \theta_k \partial \theta_l}\}
&= (nm)^{-1}\{-m \times \text{tr}\{W_{nk}S_{nm}^{-1}(\theta)W_{nl}S_{ln}^{-1}(\theta)\} + m \times \text{tr}\{W_{nk}S_{on}^{-1}W_{nl}S_{on}^{-1}\}\}
&\quad - (nm)^{-1}\nu'_{0nm}S_{0nm}^{-1} \text{diag}(W'_{nl}W_{nk}, \ldots, W'_{nl}W_{nk})S_{0nm}^{-1} \nu_{0nm}(\frac{1}{\sigma_0^2} - \frac{1}{\sigma_{nm}^2})
&\quad - (nm)^{-1}\{(\tilde{\sigma}_{nm})^2(\beta_0 - \tilde{\beta}_{nm})'X'_{nm} \text{diag}(W'_{nl}W_{nk}, \ldots, W'_{nl}W_{nk})X_{nm}(\beta_0 - \tilde{\beta}_{nm})
&\quad - 2(\tilde{\sigma}_{nm})^2(\beta_0 - \tilde{\beta}_{nm})'X'_{nm} \text{diag}(W'_{nl}W_{nk}, \ldots, W'_{nl}W_{nk})S_{0nm}^{-1} \nu_{0nm}\}\}
&= o_p(1)
\end{align*}
\[(nm)^{-1}\{\frac{\partial^2 \ell(\tilde{\eta}_{nm})}{\partial \alpha \partial \sigma^2} - \frac{\partial^2 \ell(\eta_0)}{\partial \alpha \partial \sigma^2}\} \]

\[= (nm)^{-1}\{-(\tilde{\sigma}_{nm}^4)^{-1}\nu'_{nm}(\xi_{nm})F_{nm}S_{nm}^{-1}(\tilde{\xi}_{nm})\nu_{nm} + (\sigma_0^4)^{-1}\nu'_{0nm}F_{nm}S_{0nm}^{-1}\nu_{0nm}\} \]

\[= (nm)^{-1}\{\{(\tilde{\sigma}_{nm}^4)^{-1} - (\sigma_0^4)^{-1}\}(Y_{nm} - X_{nm}\beta_0)'diag(\alpha_0 I_n, \cdots, \alpha_0 I_n, 0)\} \]

\[(Y_{nm} - X_{nm}\beta_0)\}

\[+ (nm)^{-1}(\tilde{\sigma}_{nm}^4)^{-1}\{(Y_{nm} - X_{nm}\tilde{\beta}_{nm})'diag(\tilde{\alpha}_{nm} I_n, \cdots, \tilde{\alpha}_{nm} I_n, 0)X_{nm}(\beta_0 - \tilde{\beta}_{nm})\}

\[+ (\beta_0 - \tilde{\beta}_{nm})X'_{nm}diag(\tilde{\alpha}_{nm} I_n, \cdots, \tilde{\alpha}_{nm} I_n, 0)(Y_{nm} - X_{nm}\beta_0)\}

\[= o_p(1) \]

\[\frac{\partial^2 \ell(\eta)}{\partial \alpha \partial \theta_k} = 0 \]

\[(nm)^{-1}\{\frac{\partial^2 \ell(\tilde{\eta}_{nm})}{\partial \alpha \partial \theta_k} - \frac{\partial^2 \ell(\eta_0)}{\partial \alpha \partial \theta_k}\} = o_p(1) \]

Therefore

\[(nm)^{-1}\{\frac{\partial^2 \ell(\tilde{\eta}_{nm})}{\partial \eta \partial \eta'} - \frac{\partial^2 \ell(\eta_0)}{\partial \eta \partial \eta'}\} = o_p(1) \]
Further, under (A.1) – (A.5), we have

\[(nm)^{-1}\{ \frac{\partial^2 \ell(x_0)}{\partial \beta \partial \beta'} - E \frac{\partial^2 \ell(x_0)}{\partial \beta \partial \beta'} \} = 0, \]

\[(nm)^{-1}\{ \frac{\partial^2 \ell(x_0)}{\partial \beta \partial \sigma^2} - E \frac{\partial^2 \ell(x_0)}{\partial \beta \partial \sigma^2} \} = -(nm)^{-1}(\sigma_0^4)^{-1}X_{nm}S_{0nm}^{-1}\nu_{0nm} \]

\[= (nm)^{-1/2} \times O_p(1) = o_p(1), \]

\[(nm)^{-1}\{ \frac{\partial^2 \ell(x_0)}{\partial \beta \partial \theta_k} - E \frac{\partial^2 \ell(x_0)}{\partial \beta \partial \theta_k} \} = -(nm)^{-1}(\sigma_0^2)^{-1}X_{nm}^r(J_{0nm} + J_{0nm}^r)S_{0nm}^{-1}\nu_{0nm} \]

\[= (nm)^{-1} \times O_p(1) = o_p(1), \]

\[(nm)^{-1}\{ \frac{\partial^2 \ell(x_0)}{\partial \beta \partial \alpha} - E \frac{\partial^2 \ell(x_0)}{\partial \beta \partial \alpha} \} = -(nm)^{-1}(\sigma_0^2)^{-1}(\nu_{0nm}^r\nu_{0nm} - nm) = o_p(1). \]

\[(nm)^{-1}\{ \frac{\partial^2 \ell(x_0)}{\partial \theta_k \partial \sigma^2} - E \frac{\partial^2 \ell(x_0)}{\partial \theta_k \partial \sigma^2} \}
\]

\[= -(nm)^{-1}(\sigma_0^4)^{-1}\{ \nu_{0nm}^r \text{diag}(W_{nk}S_{0n}^{-1}, \ldots, W_{nk}S_{0n}^{-1})\nu_{0nm} \}
\]

\[= o_p(1) \]

\[(nm)^{-1}\{ \frac{\partial^2 \ell(x_0)}{\partial \theta_k \partial \theta_l} - E \frac{\partial^2 \ell(x_0)}{\partial \theta_k \partial \theta_l} \}
\]

\[= -(nm)^{-1}(\sigma_0^2)^{-1}\{ \nu_{0nm}^r \text{diag}(W_{nl}W_{nk}, \ldots, W_{nl}W_{nk})S_{0nm}^{-1}\nu_{0nm} \}
\]

\[= o_p(1) \]

\[(nm)^{-1}\{ \frac{\partial^2 \ell(x_0)}{\partial \alpha \partial \sigma^2} - E \frac{\partial^2 \ell(x_0)}{\partial \alpha \partial \sigma^2} \}
\]

\[= -(nm)^{-1}(\sigma_0^4)^{-1}\{ \nu_{0nm}^r F_{nm}S_{0nm}^{-1}\nu_{0nm} - \sigma_0^2 \times \text{tr}\{ F_{nm}S_{0nm}^{-1} \} \}
\]

\[= o_p(1) \]

63
\[(nm)^{-1}\{ \frac{\partial^2 \ell(\eta_0)}{\partial \alpha \partial \alpha} - E \frac{\partial^2 \ell(\eta_0)}{\partial \alpha \partial \alpha} \} = -(nm)^{-1}(\sigma_0^2)^{-1}\{ \nu_{0mn} S_{0mn}^{-1} \text{diag}(I_n, \cdots, I_n, 0) S_{0mn}^{-1} \nu_{0mn} - \sigma_0^2 \times \text{tr}\{ S_{0mn}^{-1} \text{diag}(I_n, \cdots, I_n, 0) S_{0mn}^{-1} \} \} = o_p(1) \]

Furthermore, the first-order derivatives of \( \ell(\eta) \) at \( \eta_0 \) are linear or quadratic forms of \( \nu_{0mn} \) since

\[
\frac{\partial \ell(\eta_0)}{\partial \beta} = (\sigma_0^2)^{-1} X_{nm} S_{0mn}^{-1} \nu_{0mn}, \quad \frac{\partial \ell(\eta_0)}{\partial \sigma^2} = (2\sigma_0^4)^{-1}(\nu_{0mn}^t \nu_{0mn} - nm\sigma_0^2)
\]

\[
\frac{\partial \ell(\eta_0)}{\partial \theta_k} = -\text{tr}(G_k) + (\sigma_0^2)^{-1} \nu_{0mn}^t G_k \nu_{0mn}, \quad \frac{\partial \ell(\eta_0)}{\partial \alpha} = (\sigma_0^2)^{-1} \nu_{0mn}^t F_{nm} S_{0mn}^{-1} \nu_{0mn}
\]

where \( G_k = \text{diag}(I_m \otimes W_{nk} S_n^{-1}(\theta_0)) \), for \( k = 1, \ldots, q \).

By (A.5), we have

\[
(nm)^{-1/2} \frac{\partial \ell(\eta_0)}{\partial \beta} \overset{D}{\to} N(0, \lim_{nm \to \infty} (nm)^{-1}(\sigma_0^2)^{-1} X_{nm}^t S_{0mn} \nu_{0mn} X_{nm}).
\]

By a classic central limit theorem,

\[
(nm)^{-1/2} \frac{\partial \ell(\eta_0)}{\partial \sigma^2} \overset{D}{\to} N(0, (2\sigma_0^4)^{-1})
\]

For asymptotic normality of \( (nm)^{-1/2} \frac{\partial \ell(\eta_0)}{\partial \beta} \), (A.2) and (A.4) ensure that \( G_k \) is uniformly bounded in matrix norm \( || \cdot ||_1 \) and \( || \cdot ||_\infty \) and the positive definiteness of
\(\Sigma_{\eta_0}^{-1}\) ensures that \((nm)^{-1} Var(\frac{\partial \ell(\eta_0)}{\partial \theta_k}) = (nm)^{-1} tr(G_k^2 + G'_k G_k)\) is bounded away from 0. By a central limit theorem for linear-quadratic forms (Theorem 1, Kelejian and Prucha, 2001), we have

\[
(nm)^{-1/2} \frac{\partial \ell(\eta_0)}{\partial \theta_k} \Rightarrow N(0, \lim_{nm \to \infty} (nm)^{-1} tr(G_k^2 + G'_k G_k))
\]

for \(k = 1, \ldots, q\). Similarly, we can get the asymptotic normality of \((nm)^{-1/2} \frac{\partial \ell(\eta_0)}{\partial \alpha}\)

\[
(nm)^{-1/2} \frac{\partial \ell(\eta_0)}{\partial \alpha} \Rightarrow N(0, \lim_{nm \to \infty} (nm)^{-1} tr(H_{nm}^2 + H'_{nm} H_{nm}))
\]

By Cramer-Wold Theorem and the fact that \(\frac{\partial \ell(\eta_0)}{\partial \beta}\) is asymptotically independent of \(\frac{\partial \ell(\eta_0)}{\partial \alpha}, \frac{\partial \ell(\eta_0)}{\partial \sigma^2}\) and \(\frac{\partial \ell(\eta_0)}{\partial \theta_k}\), for \(k = 1, \ldots, q\), we have

\[
(nm)^{-1/2} \frac{\partial \ell(\eta_0)}{\partial \eta} \Rightarrow N(0, \Sigma_{\eta_0}^{-1})
\]

where \(\Sigma_{\eta_0}^{-1} = \lim_{nm \to \infty} E \left( - (nm)^{-1} \frac{\partial^2 \ell(\eta_0)}{\partial \eta \partial \eta'} \right) \) and

\[
E \left( - (nm)^{-1} \frac{\partial^2 \ell(\eta_0)}{\partial \beta \partial \beta'} \right) = (nm \sigma_0^2)^{-1} X'_{nm} S'_{nm} X_{nm},
\]

\[
E \left( - (nm)^{-1} \frac{\partial^2 \ell(\eta_0)}{\partial \beta \partial \xi'} \right) = 0,
\]

\[
E \left( - (nm)^{-1} \frac{\partial^2 \ell(\eta_0)}{\partial \beta \partial \sigma^2} \right) = 0,
\]

\[
E \left( - (nm)^{-1} \frac{\partial^2 \ell(\eta_0)}{\partial \theta \partial \alpha} \right) = 0,
\]

\[
E \left( - (nm)^{-1} \frac{\partial^2 \ell(\eta_0)}{\partial \theta_k \partial \theta_l} \right) = (nm)^{-1} tr(G_k G_l + G'_k G_k), \quad k, l = 1, \ldots, q
\]

\[
E \left( - (nm)^{-1} \frac{\partial^2 \ell(\eta_0)}{\partial \theta_k \partial \sigma^2} \right) = (nm \sigma_0^2)^{-1} tr(G_k), \quad k = 1, \ldots, q
\]

\[
E \left( - (nm)^{-1} \frac{\partial^2 \ell(\eta_0)}{\partial \alpha \partial \sigma^2} \right) = (nm \sigma_0^2)^{-1} tr(H_{nm}),
\]

\[
E \left( - (nm)^{-1} \frac{\partial^2 \ell(\eta_0)}{\partial (\sigma^2)^2} \right) = (2 \sigma_0^2)^{-1}.
\]
Assumptions (A.1)-(A.5) ensure that all the elements in $\Sigma^{-1}_{\eta_0}$ exist. Then it follows that

$$(nm)^{1/2} (\hat{\eta}_{nm} - \eta_0) = - \left\{ (nm)^{-1} \frac{\partial^2 \ell(\eta_0)}{\partial \eta \partial \eta'} + o_p(1) \right\}^{-1} (nm)^{-1/2} \frac{\partial \ell(\eta_0)}{\partial \eta} \xrightarrow{p} N(0, \Sigma_{\eta_0}).$$

and thus the result of this theorem holds.

In this theorem, we assume the existence and positive definiteness of the covariance matrix $\Sigma_{\eta_0}$. Here we discuss a simple example about the validation of the existence of $\Sigma_{\eta_0}$. Suppose we only consider the first-order spatial neighborhood, i.e. $q = 1$. Let $\tau = (\theta', \sigma^2)'$. Then

$$\Sigma_{\eta_0}^{-1} = \lim_{nm \to \infty} E \left( -(nm)^{-1} \frac{\partial^2 \ell(\eta_0)}{\partial \eta \partial \eta'} \right),$$

$$= \lim_{nm \to \infty} \begin{pmatrix} L_{nm} & 0 & 0 \\ 0 & (nm)^{-1} \frac{\partial^2 \ell(\eta_0)}{\partial \tau \partial \tau'} & 0 \\ 0 & 0 & (nm)^{-1} \{ \text{tr} (H^2_{nm} + H'_{nm} H_{nm}) \} \end{pmatrix}$$

where $L_{nm} = (nm\sigma^2_0)^{-1} X'_{nm} S'_{0nm} S_{0nm} X_{nm}$ and

$$E \left( -(nm)^{-1} \frac{\partial^2 \ell(\eta_0)}{\partial \tau \partial \tau'} \right) = \begin{pmatrix} \text{tr}(G^2_1 + G'_1 G_1) & (\sigma^2_0)^{-1} \text{tr}(G_1) \\ (\sigma^2_0)^{-1} \text{tr}(G_1) & nm(2\sigma^4_0)^{-1} \end{pmatrix}$$

Assumption (A.5) guarantees the existence and nonsingularity of $\lim_{nm \to \infty} (nm\sigma^2_0)^{-1} X'_{nm} S'_{0nm} S_{0nm} X_{nm}$. Assumptions (A.2)-(A.4) ensure that $\lim_{nm \to \infty} (nm)^{-1} \text{tr}(H^2_{nm} + H'_{nm} H_{nm})$ is bounded and bounded away from zero. Also we have

$$\left| E \left( -(nm)^{-1} \frac{\partial^2 \ell(\eta_0)}{\partial \tau \partial \tau'} \right) \right| = (2nm\sigma^4_0)^{-1} \left[ \text{tr}(G^2_1 + G'_1 G_1) - 2(nm)^{-1} \text{tr}(G_1) \right]$$

$$= (4nm\sigma^4_0)^{-1} \text{tr}(G^b G^b) = (4nm\sigma^4_0)^{-1} ||G^b||^2_F > 0,$$

where $G^b = G_1 + G'_1 - 2(nm)^{-1} \text{tr}(G_1) I_{nm}$ and $|| \cdot ||_F$ denote the Frobenius norm. Hence, the matrix $E \left( -(nm)^{-1} \frac{\partial^2 \ell(\eta_0)}{\partial \eta \partial \eta'} \right)$ is positive definite in the large-sample case.
Furthermore, assumption (A.6) guarantees its positive definiteness in the limit (see, e.g., Lee, 2004).

**The case** $h_n \to \infty$

Now, we establish the asymptotic normality of the MLE $\hat{\theta}_{nm}$ by showing the asymptotic normality of $\{h_n/(nm)\}^{1/2} \partial \ell(\xi_0) / \partial \theta$ and the convergence in probability of $h_n^{-1} \partial^2 \ell(\tilde{\xi}_{nm}) / \partial \theta \partial \theta'$, where $\tilde{\xi}_{nm} = \lambda \xi_0 + (1 - \lambda) \hat{\xi}_{nm}$ for $\lambda \in (0, 1)$ converges to $\xi_0$ in probability. The asymptotic normality of the MLE $\hat{\alpha}_{nm}$ will be established by showing the asymptotic normality of $\{nm\}^{-1/2} \partial^2 \ell(\xi_0) / \partial \alpha^2$ and the convergence in probability of $\{nm\}^{-1} \partial^2 \ell(\tilde{\xi}_{nm}) / \partial \alpha^2$.

For convergence of $h_n^{-1} \partial^2 \ell(\tilde{\xi}_{nm}) / \partial \theta \partial \theta'$, we show that $h_n^{-1} \{\partial^2 \ell(\tilde{\xi}_{nm}) / \partial \theta \partial \theta' - \partial^2 \ell(\xi_0) / \partial \theta \partial \theta'\} = o_p(1)$ and $h_n^{-1} \{\partial^2 \ell(\xi_0) / \partial \alpha^2 - E \partial^2 \ell(\xi_0) / \partial \alpha^2\} = o_p(1)$, under (A.1)–(A.5). On the other hand, we have that $\{nm\}^{-1} \{\partial^2 \ell(\tilde{\xi}_{nm}) / \partial \theta \partial \theta' - \partial^2 \ell(\xi_0) / \partial \alpha^2\} = o_p(1)$ and $\{nm\}^{-1} \{\partial^2 \ell(\xi_0) / \partial \alpha^2 - E \partial^2 \ell(\xi_0) / \partial \alpha^2\} = o_p(1)$, under (A.1)–(A.5). Under (A.1)–(A.5), we note that

\[
\frac{\partial \ell(\xi)}{\partial \theta_1} = -\{2\hat{\sigma}_{nm}^2(\xi)\}^{-1} \nu'_{0nm} \frac{\partial P_{nm}(\xi)}{\partial \theta_1} \nu_{0nm} - tr\{(I_m \otimes W_{n1})S_{nm}^{-1}(\xi)\}
\]

\[
= \{\hat{\sigma}_{nm}^2(\xi)\}^{-1} \nu'_{0nm} T_{nm,1}(\xi) \nu_{0nm} - tr\{(I_m \otimes W_{n1})S_{nm}^{-1}(\xi)\}
\]

\[
\frac{\partial^2 \ell(\xi)}{\partial \theta_1^2} = 2\{nm\hat{\sigma}_{nm}^4(\xi)\}^{-1} \{\nu'_{0nm} T_{nm,1}(\xi) \nu_{0nm}\}^2 + \{\hat{\sigma}_{nm}^2(\xi)\}^{-1} \nu'_{0nm} \frac{\partial T_{nm,1}(\xi)}{\partial \theta_1} \nu_{0nm}
\]

\[
-t_{n_1}(\xi)(I_m \otimes W_{n1})S_{nm}^{-1}(\xi)(I_m \otimes W_{n1})
\]

with

\[
T_{nm,1}(\xi) = S_{nm}^{-1}(\xi) P_{nm} S_{nm}^{-1}(\xi),
\]
\[ P_{nm}(\xi) \]
\[ = (I_m \otimes W_{n1})M_{nm}(\xi)S_{nm}(\xi) - S'_{nm}(\xi)(I_m \otimes W_{n1})X_{nm} \]
\[ \{X'_{nm}S'_{nm}(\xi)S_{nm}(\xi)X_{nm}\}^{-1}X'_{nm}S'_{nm}(\xi)S_{nm}(\xi) \]
\[ + S_{nm}(\xi)X_{nm}\{X'_{nm}S'_{nm}(\xi)S_{nm}(\xi)X_{nm}\}^{-1}X'_{nm}(I_m \otimes W_{n1})S_{nm}(\xi)X_{nm} \]
\[ \{X'_{nm}S'_{nm}(\xi)S_{nm}(\xi)X_{nm}\}^{-1}X'_{nm}S'_{nm}(\xi)S_{nm}(\xi), \]
\[ M_{nm}(\xi) = I_{nm} - S_{nm}(\xi)X_{nm}\{X'_{nm}S'_{nm}(\xi)S_{nm}(\xi)X_{nm}\}^{-1}X'_{nm}S'_{nm}(\xi). \]

By (A.1)-(A.5)
\[ h_n(nm)^{-1}\nu_{0n_{nm}}T_{nm,1}(\xi)\nu_{0nm} = O_p(1) \]
\[ h_n(nm)^{-1}\nu_{0n_{nm}} \frac{\partial T_{nm,1}(\xi)}{\partial \theta_1} \nu_{0nm} = h_n(N_n m)^{-1} o_p^2 tr\{ \frac{\partial T_{nm,1}(\xi)}{\partial \theta_1} \} + o_p(1) \]

Since \( T_{nm,1} \) and \( \frac{\partial T_{nm,1}(\xi)}{\partial \theta_1} \) are uniformly bounded in either matrixnorm \( ||\cdot||_1 \) or \( ||\cdot||_{\text{inf}}, \) thus, under (A.1) - (A.5),
\[ \frac{\partial^2 (\xi)}{\partial \theta_1^2} = 2\{nm \hat{\sigma}_{nm}^4(\xi)\}^{-1}\nu_{0n_{nm}}T_{nm,1}(\xi)\nu_{0nm}^2 + \{\hat{\sigma}_{nm}^2(\xi)\}^{-1}\nu_{0n_{nm}} \frac{\partial T_{nm,1}(\xi)}{\partial \theta_1} \nu_{0nm} \]
\[ -h_n(nm)^{-1} tr\{(I_m \otimes W_{n1})S_{nm}^{-1}(\xi)(I_m \otimes W_{n1})S_{nm}^{-1}(\xi)\} \]
\[ = h_n(nm)^{-1} tr\{ \frac{\partial T_{nm,1}(\xi)}{\partial \theta_1} \} \]
\[ -h_n(nm)^{-1} tr\{(I_m \otimes W_{n1})S_{nm}^{-1}(\xi)(I_m \otimes W_{n1})S_{nm}^{-1}(\xi)\} + o_p(1) \]

and for \( \xi_{nm} = \lambda \xi_0 + (1 - \lambda) \xi_{nm}^\dagger, \)
\[ h_n(nm)^{-1}\frac{\partial^2 (\xi_{nm}^\dagger)}{\partial \theta_1^2} - \frac{\partial^2 (\xi_0)}{\partial \theta_1^2} \]
\[ = h_n(nm)^{-1}\{ tr\{ \frac{\partial T_{nm,1}(\xi_{nm}^\dagger)}{\partial \theta_1} \} - tr\{ \frac{\partial T_{nm,1}(\xi_0)}{\partial \theta_1} \} \} \]
\[ -h_n(nm)^{-1} [tr\{(I_m \otimes W_{n1})S_{nm}^{-1}(\xi_{nm}^\dagger)(I_m \otimes W_{n1})S_{nm}^{-1}(\xi_{nm}^\dagger)\} \]
\[ -tr\{(I_m \otimes W_{n1})S_{0nm}^{-1}(I_m \otimes W_{n1})S_{0nm}^{-1}\}] + o_p(1) \]
\[ = (\xi_{nm}^\dagger - \xi_0)^\dagger O(1) - (\xi_{nm}^\dagger - \xi_0)^\dagger O(1) + o_p(1) = o_p(1) \]
By similar argument for $\theta_k$ and $\theta_{k',k}, k' = 1, \ldots, q$, we have

\[ h_n(nm)^{-1} \{ \frac{\partial^2 \ell(\xi_{nm}^-)}{\partial \theta \partial \theta'} - \frac{\partial^2 \ell(\xi_0)}{\partial \theta \partial \theta'} \} = o_p(1) \]

Furthermore,

\[ -h_n(nm)^{-1} \frac{\partial^2 \ell(\xi_0)}{\partial \theta_1^2} = -h_n(nm)^{-1} tr \{ \frac{\partial T_{nm,1}(\xi_0)}{\partial \theta_1} \} + h_n(nm)^{-1} tr(G_1^2) + o_p(1) \]

\[ = -h_n(nm)^{-1} \{ tr(G_1'G_1) + tr(G_1^2) \} + o_p(1) \]

where $G_1 = (I_m \otimes W_{n1})S_{0nm}^{-1}$.

By similar argument for $\theta_k$ and $\theta_{k',k'}, k', k = 1, \ldots, q$, we have $h_n(nm)^{-1} \{ \frac{\partial^2 \ell(\xi_0)}{\partial \theta \partial \theta'} - E \frac{\partial^2 \ell(\xi_0)}{\partial \theta \partial \theta'} \} = o_p(1)$. Thus, $h_n(nm)^{-1} \frac{\partial^2 \ell(\xi_{nm})}{\partial \theta \partial \theta'} \to \Sigma^{-1}_{\theta_0} = \lim_{n \to \infty} h_n(nm)^{-1} E \frac{\partial^2 \ell(\xi_0)}{\partial \theta \partial \theta'}$.

To establish the asymptotic normality of $\{h_n/(nm)\}^{1/2} \frac{\partial \ell(\xi_0)}{\partial \theta_1}$, we apply the central limit theorem for linear-quadratic forms in Appendix A of Lee (2004). Note that

\[ \{h_n/(nm)\}^{1/2} \frac{\partial \ell(\xi_0)}{\partial \theta_1} = \begin{align*}
\{h_n/(nm)\}^{1/2} & \left[ \sigma^2_{nm}(\theta_0) \right]^{-1} \nu'_{0nm} T_{nm,1}(\xi_0) \nu_{0nm} - tr \left\{ (I_m \otimes W_{n1})S_{0nm}^{-1} \right\} \\
= \{ \sigma^2_{nm}(\theta_0) \}^{-1} \{h_n/(nm)\}^{1/2} & \left\{ \nu'_{0nm} G_1' \nu_{0nm} - \sigma^2_{0} tr(G_1) \right\} + o_p(1),
\end{align*} \]

where

\[ T_{nm,1}(\xi_0) = S_{0nm}^{-1} P_{0nm} S_{0nm}^{-1}, \]

\[ P_{0nm} = (I_m \otimes W_{n1})M_{0nm} S_{0nm} - S_{0nm}'(I_m \otimes W_{n1})X_{nm} \]

\[ \{X_{nm}' S_{0nm}' S_{0nm} X_{nm} \}^{-1} X_{nm}' S_{0nm}' S_{0nm} \]

\[ + S_{0nm}' S_{0nm} X_{nm} \{X_{nm}' S_{0nm}' S_{0nm} X_{nm} \}^{-1} X_{nm}' (I_m \otimes W_{n1}) S_{0nm} X_{nm} \]

\[ \{X_{nm}' S_{0nm}' S_{0nm} X_{nm} \}^{-1} X_{nm}' S_{0nm}' S_{0nm}, \]
\[
M_{0\text{nm}} = I_{\text{nm}} - S_{0\text{nm}} X_{\text{nm}} \{ X'_{\text{nm}} S'_{0\text{nm}} S_{0\text{nm}} X_{\text{nm}} \}^{-1} X'_{\text{nm}} S'_{0\text{nm}}.
\]

(A.2) and (A.4) ensure that \( G_1 \) is uniformly bounded in matrix norms \( || \cdot ||_1 \) and \( || \cdot ||_{\infty} \) and the positive definiteness of \( \Sigma_{\theta_0}^{-1} \) ensures that \( \{ h_n/(nm) \} \operatorname{Var}(\nu'_{0n} G_1 \nu_{0n}) = \{ h_n/(nm) \} \sigma_0^2 \operatorname{tr}(G_1^2 + G_1' G_1) \) is bounded away from zero. By Lee (2004), we have

\[
\{ h_n/(nm) \}^{1/2} \{ \nu'_{0nm} G_1 \nu_{0nm} - \sigma_0^2 \operatorname{tr}(G_1) \} \xrightarrow{D} N \left( 0, \lim_{n \to \infty} h_n(nm)^{-1} \sigma_0^4 \{ \operatorname{tr}(G_1 G_1') + \operatorname{tr}(G_1^2) \} \right).
\]

Hence

\[
\{ h_n/(nm) \}^{1/2} \frac{\partial \ell(\xi_0)}{\partial \theta_1} \xrightarrow{D} N \left( 0, \lim_{n \to \infty} h_n(nm)^{-1} \{ \operatorname{tr}(G_1 G_1') + \operatorname{tr}(G_1^2) \} \right)
\]

Then similarly for \( \theta_k, k = 2, \ldots, q \), we have

\[
\{ h_n/(nm) \}^{1/2} \frac{\partial \ell(\xi_0)}{\partial \theta_k} \xrightarrow{D} N \left( 0, \lim_{n \to \infty} h_n(nm)^{-1} \{ \operatorname{tr}(G_k G_k') + \operatorname{tr}(G_k^2) \} \right).
\]

Thus, by the Cramér-Wold theorem,

\[
\{ h_n/(nm) \}^{1/2} \frac{\partial \ell(\xi_0)}{\partial \theta} \xrightarrow{D} N(0, \Sigma_{\theta_0}^{-1})
\]

where

\[
\Sigma_{\theta_0}^{-1} = \lim_{nm \to \infty} -h_n(nm)^{-1} E \left\{ \frac{\partial^2 \ell(\eta_0)}{\partial \theta \partial \theta'} \right\}
\]

\[
= \lim_{nm \to \infty} \begin{bmatrix}
    h_n(nm)^{-1} \operatorname{tr}(G_1^2 + G_1' G_1) & \ldots & h_n(nm)^{-1} \operatorname{tr}(G_1 G_q + G_q' G_1) \\
    \vdots & \ddots & \vdots \\
    h_n(nm)^{-1} \operatorname{tr}(G_q G_1 + G_1' G_q) & \ldots & h_n(nm)^{-1} \operatorname{tr}(G_q^2 + G_q' G_q)
\end{bmatrix}
\]

Similarly, we can show that

\[
(nm)^{-1/2} \frac{\partial \ell(\xi_0)}{\partial \alpha} \xrightarrow{D} N \left( 0, \lim_{nm \to \infty} (nm)^{-1} \operatorname{tr}\{H_{nm}^2 + H_{nm}' H_{nm}\} \right)
\]

\[
h_n(nm)^{-1} \frac{\partial^2 \ell(\xi_0)}{\partial \alpha \partial \theta} = o_p(1).
\]
Thus it follows that

\[
(nm/h_n)^{1/2}(\hat{\theta}_n - \theta_0) = - \left\{ h_n(nm)^{-1} \frac{\partial^2 \ell(\xi_0)}{\partial \theta \partial \theta'} + o_p(1) \right\}^{-1} \{ h_n/(nm) \}^{1/2} \frac{\partial \ell(\xi_0)}{\partial \theta} \\
- h_n^{-1/2} \left\{ h_n(nm)^{-1} \frac{\partial^2 \ell(\xi_0)}{\partial \theta \partial \theta'} + o_p(1) \right\}^{-1} \left\{ h_n(nm)^{-1} \frac{\partial^2 \ell(\xi_0)}{\partial \theta \partial \alpha} + o_p(1) \right\} \\
\frac{(nm)^{-1} \frac{\partial^2 \ell(\xi_0)}{\partial \alpha^2} + o_p(1)}{\partial \alpha}^{-1} \left( (nm)^{-1/2} \frac{\partial \ell(\xi_0)}{\partial \alpha} \right)
\]

\[ \xrightarrow{D} N(0, \Sigma_{\theta_0}). \]

and similarly,

\[
(nm)^{1/2}(\hat{\alpha}_n^2 - \alpha_0^2) \xrightarrow{D} N(0, \Sigma_{\alpha_0})
\]

with \( \Sigma_{\alpha_0} = \lim_{nm \to \infty} nm \{ \text{tr}(H_{nm}^2 + H_{nm}'H_{nm}) \}^{-1} \).

To establish the asymptotic normality of \( \hat{\beta}_{nm}(\hat{\xi}_{nm}) \) and \( \hat{\sigma}_{nm}^2(\hat{\xi}_{nm}) \), we have, under (A.1)–(A.5),

\[
(nm)^{1/2}(\hat{\beta}_{nm}(\hat{\xi}_{nm}) - \beta_0) = (nm)^{-1/2} \left\{ (nm)^{-1} X_{nm}' S_{nm}'(\hat{\xi}_{nm}) S_{nm}(\hat{\xi}_{nm}) X_{nm} \right\}^{-1} \\
X_{nm}' S_{nm}'(\hat{\xi}_{nm}) S_{nm}(\hat{\xi}_{nm}) S_{0nm}^{-1} \nu_{0nm} \\
= (nm)^{-1/2} \left( (nm)^{-1} X_{nm}' S_{0nm}' S_{0nm} X_{nm} \right)^{-1} X_{nm}' S_{0nm}' \nu_{0nm} + o_p(1) \\
\xrightarrow{D} N(0, \lim_{n \to \infty} ((nm)^{-1} X_{nm}' S_{0nm}' S_{0nm} X_{nm})^{-1}).
\]

and

\[
(nm)^{1/2}(\hat{\sigma}_{nm}^2(\hat{\xi}_{nm}) - \sigma_0^2) = (nm)^{-1/2} \left( \nu_{0nm}' \nu_{0nm} - nm \sigma_0^2 \right) + o_p(1) \xrightarrow{D} N(0, 2 \sigma_0^4).
\]
Appendix C: Proof of Theorem 3

\[ |E(G(Z_{kx})) - G(x)| = |\int (G(z) - G(x))dP_{kx}(z)| \]
\[ \leq \int |G(z) - G(x)|dP_{kx}(z) \]
\[ = \int_{\{z: \max_{i=1,\ldots,d} |z_i - x_i| \leq \delta\}} |G(z) - G(x)|dP_{kx}(z) \]
\[ + \int_{\{z: \max_{i=1,\ldots,d} |z_i - x_i| > \delta\}} |G(z) - G(x)|dP_{kx}(z). \]

Since \( G(\cdot) \) is continuous in \( x \), for any \( \epsilon > 0 \), we can choose \( \delta \) such that \( |G(z) - G(x)| < \epsilon/2 \) for \( z \) with \( \max_{i=1,\ldots,d} |z_i - x_i| \leq \delta \). Then the first addend above is smaller than \( \epsilon/2 \). And since \( G(\cdot) \) is bounded, i.e. smaller than a \( M \) for some \( M > 0 \), we have

\[ |E(G(Z_{kx})) - G(x)| \leq \epsilon/2 + 2MP(\max_{i=1,\ldots,d} |z_i,k,x_i - x_i| > \delta). \]

By multivariate Markov inequality,

\[ |E(G(Z_{kx})) - G(x)| \leq \epsilon/2 + 2M \sum_{i=1}^{d} \left[ V(Z_{i,k,x_i}) + (E(z_i,k,x_i) - x_i)^2 \right]. \]

which can be arbitrarily small when \( \min\{k\} \to \infty \) by the properties of the random scheme.

Moreover, if \( G \) is continuous on the compact set \( \Delta \), it is uniformly continuous on \( \Delta \). Then for any \( \epsilon > 0 \), we can find a \( \delta^* > 0 \) such that for two points \( x \) and \( x^* \), we have \( |G(x^*) - G(x)| < \epsilon \) if \( |x^* - x| < \delta^* \). Thus, Theorem \[3.6.2\] holds for some \( \delta^* \) which is independent of \( x \) and then the convergence is uniformly on \( \Delta \).
Appendix D: Proof of Proposition 1

For property (i), we consider the marginal’s of $B(x; k, G)$. We can get the $i$th marginal is:

$$B(x_i; k, G) = \int G(x_1, \ldots x_{i-1}, z_i, x_{i+1}, \ldots, x_d) dP_{k_i, \theta(x_i)}(z_i),$$

where $G(x_1, \ldots x_{i-1}, z_i, x_{i+1}, \ldots, x_d)$ is the $i$th marginal d.f. of $G(x_1, \ldots, x_d)$. According to the proposition 1 of univariate Feller operators (Petrone, 2002), we have that all marginal $B(x_i; k, G)$ are d.f.’s. Thus $B(x; k, G)$ is a d.f.

For property (ii), Let $G(x_1, \ldots, x_d) = \int_{-\infty}^{x_1} \ldots \int_{-\infty}^{x_d} dG(t_1, \ldots, t_d)$, for $(x_1, \ldots, x_d) \in \chi^o$. Since $G(\cdot)$ is a probability distribution function (therefore monotone, nondecreasing, and not constant), by interchanging the role of $G$ and $P_{k_i, \theta(x_i)}$, we get

$$B(x; k, G) = \int \cdots \int G(z_1, \ldots, z_d) dP_{k_i, \theta(x_i)}(z_1) \cdots dP_{k_d, \theta(x_d)}(z_d)$$

$$= \int \cdots \int \int_{-\infty}^{z_1} \ldots \int_{-\infty}^{z_d} dG(t_1, \ldots, t_d) dP_{k_1, \theta(x_1)}(z_1) \cdots dP_{k_d, \theta(x_d)}(z_d)$$

$$= \int \cdots \int \prod_{i=1}^d P_{k_i, \theta(x_i)}([z_i, \infty)) dG(z_1, \ldots, z_d).$$

Therefore

$$f_{k_i}(x_i, z_i) = \frac{\partial P_{k_i, \theta(x_i)}([z_i, \infty))}{\partial x_i}$$

$$= \frac{d\theta}{dx_i} \int_{[z_i, \infty)} \frac{d}{d\theta} \exp\{k_i(\theta(x_i)t - M(\theta(x_i)))\} d\nu(t)$$

$$= \frac{k_i}{\sigma^2(x_i)} \int_{[z_i, \infty)} (t - x_i) dP_{k_i, \theta(x_i)}(t),$$

where $x_i = dM(\theta)/d\theta$, $\sigma^2(x_i) = d^2 M(\theta)/d\theta^2$. 

73
Appendix E: Proof of Proposition 2

If G is absolutely continuous with density g, then by \[3.7\], we have:
\[
b(x; k, G) = \int \cdots \int g(z_1, \ldots, z_d) \prod_{i=1}^{d} f_{k_i}(x_i, z_i) dz_1 \ldots dz_d.
\]

According to the proposition 1 in Petrone (2002), for any \(x_i \in \chi_i^o\) where \(\chi_i^o\) is the interior of \(i\)th argument of \(\chi\), \(f_{k_i}(x_i; \cdot)\) is a probability density function. In particular, if \(Z_{i,k_i,x_i}^* \sim f_{k_i}(x_i; \cdot)\), then \(E(Z_{i,k_i,x_i}^*) \to x_i\) and \(V(Z_{i,k_i,x_i}^*) \to 0\) as \(k_i \to \infty\). So \(\{Z_{k_x}^* = (Z_{1,k_1,x_1}^*, \ldots, Z_{d,k_d,x_d}^*), k_1, \ldots, k_d = 1, 2, \ldots, x = (x_1, \ldots, x_d) \in \chi^o\}\) is a d-dimensional random scheme. Since \(g(x_1, \ldots, x_d)\) is bounded in \(\chi^o\), then \(b(x; k, G)\) is a d-dimensional Feller operator with d-dimensional random scheme \(Z_{k_x}^*\). According to Theorem 3, we get this property proved.
Appendix F: Proof of Proposition 3

W.L.O.G, we only consider the case that density function is coordinate-wise monotone increasing. For Feller operators with Poisson random scheme,

\[ B(x; k, G) = \sum_{j}^{\infty} G(j/k) \prod_{i=1}^{d} (k_i x_i)^{j_i} e^{-k x_i} 
\]

\[ b(x; k, G) = \sum_{j=1}^{\infty} (G(j/k) - G((j - 1)/k)) \prod_{i=1}^{d} G(x_i; j_i, k_i) 
\]

\[ = \prod_{i=1}^{d} k_i \sum_{j=0}^{\infty} (G(j + 1/k) - G(j/k)) \prod_{i=1}^{d} P(k x_i; j_i) 
\]

We first consider \( d = 1 \) case,

\[ b(x; k, G) = \sum_{j}^{\infty} (G(j/k) - G((j - 1)/k)) G(x; j, k) 
\]

\[ \frac{\partial b(x; k, G)}{\partial x} = \frac{\partial}{\partial x} \sum_{j=1}^{\infty} (G(j/k) - G((j - 1)/k)) G(x; j, k) 
\]

\[ = \frac{\partial}{\partial x} \sum_{j=0}^{\infty} G((j + 1)/k) G(x; j + 1, k) - \frac{\partial}{\partial x} \sum_{j=0}^{\infty} G(j/k) G(x; j + 1, k) 
\]

\[ = k \sum_{j=0}^{\infty} ((G((j + 2)/k) - (G((j + 1)/k)) G(x; j, k) 
\]

\[ - k \sum_{j=0}^{\infty} ((G((j + 1)/k) - G(j/k)) G(x; j, k) 
\]

\[ = k^2 \sum_{j=0}^{\infty} \Delta^2 G(j/k) P(k x; j) 
\]

where \( \Delta \) is the forward difference operator and \( P(k x; j) \) is the probability density function of Poisson distribution. More generally, the derivatives of feller operators with \( d \)-dimensional Poisson random scheme is

\[ \frac{\partial^{\alpha}}{\partial x_1^{\alpha_1} \cdots \partial x_d^{\alpha_d}} B(x; k, G) = \left\{ \prod_{i=1}^{d} k_{i}^{\alpha_i} \right\} \sum_{j_1=0}^{\infty} \cdots \sum_{j_d=0}^{\infty} \Delta^\alpha G(j/k) \prod_{i=1}^{d} P(k_i x_i; j_i) \] (3.14)
where $\alpha = (\alpha_1, \ldots, \alpha_d)$ is a multi-index with $|\alpha| = (\alpha_1, \ldots, \alpha_d)$.

Let $e_i$ denote the vector with $i$th coordinate 2 and 0 elsewhere. The density of feller operator under poisson random scheme is monotonically increasing along the $i$th coordinate if

$$\frac{\partial^{|\alpha|}}{\partial x_1^{\alpha_1} \ldots \partial x_d^{\alpha_d}} \geq 0.$$ 

From (3.14), the density is monotonically increasing along the $i$th coordinate if $\Delta e_i G(j/k) \geq 0$, which can be ensured by the coordinate-wise monotonicity of $g$. 

76
Appendix G: Proof of Proposition 4

\[ B(x; k, G) = \int \cdots \int G(z_1, \ldots, z_d) \prod_{i=1}^{d} \varphi(z_i; x_i, \sigma_i^2/k_i) dz_1 \cdots dz_d \]

\[ b(x; k, G) = \int \cdots \int g(z_1, \ldots, z_d) \prod_{i=1}^{d} \varphi(z_i; x_i, \sigma_i^2/k_i) dz_1 \cdots dz_d \]

W.L.O.G, we only consider the case that density function is coordinate-wise monotone increasing. Similar to the Poisson random case discussed above, first we consider \( d = 1 \) and \( k = 1 \) case,

\[
\frac{\partial}{\partial x} b(x; k, G) = \int_{-\infty}^{\infty} \frac{\partial}{\partial x} g(z) \exp\{- (z - x)^2 / \sigma^2\} dz \tag{3.15}
\]

\[
= C \int_{-\infty}^{\infty} g(z)(z - x) \exp\{- (z - x)^2 / 2\sigma^2\} dz
\]

\[
= C \int_{-\infty}^{\infty} g(u + x)u \exp\{- u^2 / 2\sigma^2\} du
\]

\[
= C \int_{0}^{\infty} (g(u + x) - g(-u + x))u \exp\{- u^2 / 2\sigma^2\} du
\]

It’s clearly that if \( g \) is monotonically increasing, then (3.15) \( \geq 0 \), i.e., \( b(x; k, G) \) is monotonically increasing. Now let’s consider the general case.

\[ b(x; k, G) = \int g(z_1, \cdots, z_d) \prod_{i=1}^{d} \frac{\sqrt{k_i}}{\sqrt{2\pi\sigma_i}} \exp\{-k_i(z_i - x_i)^2 / 2\sigma_i^2\} dz_1 \cdots dz_d \]

\[ \frac{\partial}{\partial x_i} b(x; k, G) = \int_{-\infty}^{\infty} l(z_i; k_i, G) \frac{\sqrt{k_i}}{\sqrt{2\pi\sigma_i}} \exp\{-k_i(z_i - x_i)^2 / 2\sigma_i^2\} dz_i \]

where

\[ l(z_i; k_i, G) \]

\[
= \int g(z_1, \cdots, z_d) \prod_{j \neq i} \frac{\sqrt{k_j}}{\sqrt{2\pi\sigma_j}} \exp\{-k_j(z_j - x_j)^2 / 2\sigma_j^2\} dz_1 \cdots dz_{i-1}dz_{i+1} \cdots dz_d
\]
since \( g \) is monotonically increasing along the \( \text{ith} \) direction, \( l(z_i; k_i, G) \) is a monotone increasing function, apply the conclusion of one-dimensional case and we got the property proved.
Appendix H: Proof of Theorem 4

Equip the space \( \Pi \) of probability distribution functions on the convex support \( \chi \) with a Borel \( \sigma \)-field \( \Gamma \) generated by the topology of weak convergence. Let \( \mathbb{N}^d \) be the set of all positive integer lattice in \( d \)-dimension, equipped with its power set \( \mathcal{P}(\mathbb{N}^d) \). Let \( \Omega = \mathbb{N}^d \times \Pi \), and \( \mathcal{B}(\Omega) \) be the corresponding \( \sigma \)-field \( \mathcal{P}(\mathbb{N}^d) \times \Gamma \), \( P \) be the probability measure on \( (\Omega, \mathcal{B}(\Omega)) \). For each \( k \in \mathbb{N}^d \) and \( G \in \Pi \), the multi-dimensional Feller operator from \( \Omega \) to \( \Pi \) is,

\[
B(x; k, G) = \int \cdots \int G(z_1, \ldots, z_d) dP_{k_1, \theta(x_1)}(z_1) \cdots P_{k_d, \theta(x_d)}(z_d), \quad (z_1, \ldots, z_n) \in \chi.
\]

For fixed \( k \) and \( G \), \( B(x; k, G) \) is a probability distribution function in \( \Pi \).

We can show that the real function \( B(x; k, \cdot) \) is \( \Pi \)-measurable for every fixed \( k \) and \( x \). For any \( G \in \Pi \), if there exist \( \{G_n\} \subseteq \Pi \) such that \( G_n \to G \) almost surely with respect to the measure \( \nu \) (the measure on space \( (\Pi, \Gamma) \)), then

\[
|B(x; k, G_n) - B(x; k, G)| = |\int \cdots \int G_n(z_1, \ldots, z_d) dP_{k_1, \theta(x_1)}(z_1) \cdots P_{k_d, \theta(x_d)}(z_d) - \int \cdots \int G(z_1, \ldots, z_d) dP_{k_1, \theta(x_1)}(z_1) \cdots dP_{k_d, \theta(x_d)}(z_d)| \to 0
\]

since \( \{G_n\} \) are all probability distribution functions, i.e., \( |G_n(z_1, \ldots, z_d)| \leq 1 \) (Dominated Convergence Theorem). So \( B(x; k, G) \) is continuous (therefore \( \Pi \)-measurable) in \( G \) for fixed \( k \) and \( x \). Thus, the mapping \( B(x; k, G) \) from \( (\Omega, \mathcal{B}, P) \) to \( \Pi \) is a random d.f., with probability law induced by the probability law of \( (k, G) \) (Billingsley, 1999).

Furthermore, if the random distribution function \( H \) is also a measurable map from
$(\Omega, \mathcal{B}, P)$ to $\Gamma$, the distribution of $H$ is a prior on $(\Pi, \Gamma)$. In this case, $B(x; k, G)$ induces a prior $\pi_F$ on $(\Pi, \Gamma)$. 
Appendix I: Proof of Theorem 5

(i) In order to show that $\Pi_F$ has a full topological support on $(\Pi_1, \Gamma_1)$, it suffice to show that $\pi_F(N_{G_n,\epsilon}(Q)) > 0$ for any $n, G_n, Q \in \Pi_1$ and $\epsilon > 0$. From Theorem 3, we have

$$\pi_F(\{Q^* \in \Pi_1 : \max_{g \in G_n} |\int gdQ^* - \int gdQ| < \epsilon\})$$

$$\geq \pi_F(\{B(x; k, Q^*) \in \Pi_1 : \max_{g \in G_n} |\int gdB(x; k, Q^*) - \int gdQ| < \epsilon\})$$

Now

$$\max_{g \in G_n} |\int gdB(x; k, Q^*) - \int gdQ|$$

$$\leq \max_{g \in G_n} |\int gdB(x; k, Q^*) - \int gdB(x; k, Q)|$$

$$+ \max_{g \in G_n} |\int gdB(x; k, Q) - \int gdQ|.$$ 

From Theorem 3 for any $\epsilon > 0$, there exists a $k^*$ such that the second additive $< \epsilon/2$ for $k \geq k^*$.

For the first additive, let us consider any fixed $k \geq k^*$. From the proof of Proposition 1, we know that $B(x; k, Q)$ can be written as

$$B(x; k, Q) = \int \cdots \int \Delta \prod_{i=1}^d P_{k_i, \theta(x_i)}([z_i, \infty))dQ(z_1, \ldots, z_d).$$

According to proof of Theorem 2 in Petrone (2002), $P_{k_i, \theta(x_i)}([z_i, \infty))$ is continuous or piecewise continuous at most countably many discontinuity points, for $i = 1, \ldots, d$. So the product of $P_{k_i, \theta(x_i)}([z_i, \infty))$ is also continuous or piecewise continuous with at most countably many discontinuity points.

Now let $\{Q_n\}$ be a sequence of distribution function in $\Pi$ that weakly con-
verge to \( Q \). Then by the generalized Helly-Bray theorem, we have,

\[
\int \cdots \int_{\Delta} \prod_{i=1}^{d} P_{k_i, \theta(x_i)}([z_i, \infty)) dQ_n(z_1, \ldots, z_d)
\]

\[
\rightarrow \int \cdots \int_{\Delta} \prod_{i=1}^{d} P_{k_i, \theta(x_i)}([z_i, \infty)) dQ(z_1, \ldots, z_d), \quad n \to \infty
\]

i.e.

\[
B(x; k, Q_n) \to B(x; k, Q), \quad n \to \infty.
\]

Hence for any \( \epsilon > 0 \) and any \( k \), we can find \( \delta_0 \) such that

\[
\max_{g \in G_n} | \int g dF(x; k, Q^*) - \int g dF(x; k, Q) | < \epsilon / 2
\]

for any \( Q^* \) in the set

\[
N_{\delta_0} = \{ Q^* \in \Pi_1 : \max_{g \in G_n} | \int g dQ^* - \int g dQ | < \delta_0 \}.
\]

Therefore, from the previous result we have

\[
\pi_F(\{ Q^* \in \Pi_1 : \max_{g \in G_n} | \int g dQ^* - \int g dQ | < \epsilon \} > p(k)p(N_{\delta_0}|k) > 0
\]

for \( k \geq k^* \).

(ii) For simplicity, define \( d(Q^*, Q) = \sup_{\omega \in \Delta} |Q^*(\omega) - Q(\omega)| \). Then similar to the proof of part (i), we have,

\[
\pi_F(\{ Q^* \in \Pi_2 : d(Q^*, Q) < \epsilon \}) = \pi_F(\{ B(x; k, Q^*), k \in \mathbb{N}^d, Q^* \in \Pi_2 : d(B(x; k, Q^*), Q) < \epsilon \}).
\]

Now \( d(B(x; k, Q^*), Q) \leq d(B(x; k, Q^*), B(x; k, Q)) + d(B(x; k, Q), Q) \). Since \( Q \) is an absolutely continuous probability distribution function, from Theorem 3, \( B(x; k, Q) \) converges to \( Q \) uniformly on the compact set \( \Delta \). Thus,
\(d(B(x; k, Q), Q)\) can be arbitrarily small, e.g. \(\leq \epsilon/2\) when \(\min\{k\}\) is large enough. For \(d(B(x; k, Q^*), B(x; k, Q))\), we have,

\[
|B(x; k, Q^*) - B(x; k, Q)|
= \bigg| \int \cdots \int_{\Delta} \prod_{i=1}^{d} P_{k_i, \theta(x_i)}([z_i, \infty))dQ^*(z_1, \ldots, z_d) \bigg|
- \bigg| \int \cdots \int_{\Delta} \prod_{i=1}^{d} P_{k_i, \theta(x_i)}([z_i, \infty))dQ(z_1, \ldots, z_d) \bigg|
= \bigg| \int \cdots \int_{\Delta} \prod_{i=1}^{d} P_{k_i, \theta(x_i)}([z_i, \infty)) [f_{Q^*}(z) - f_{Q}(z)]dz_1 \cdots dz_d \bigg|
\leq \bigg| \int \cdots \int_{\Delta} [f_{Q^*}(z) - f_{Q}(z)]dz_1 \cdots dz_d \bigg|.
\]

Using a similar argument shown in proof of part(i), we have, when \(k\) is large enough, for any \(\epsilon > 0\), we can find \(\delta^*\) such that \(d(B(x; k, Q^*), B(x; k, Q)) < \epsilon/2\) for any \(Q^*\) in \(N_{f_Q, \delta^*}(Q)\). Therefore,

\[
\pi_F(\{Q^* \in \Pi_2 : d(Q^*, Q) < \epsilon\}) > p(k)p(N_{f_Q, \delta^*}(Q)|k) > 0
\]
Bibliography


Copyright© Xiang Zhang, 2013.
Vita

Xiang Zhang

University of Kentucky Department of Statistics

Place of Birth: Hefei, Anhui Province, China

Education

Master of Science in Statistics, University of Kentucky, 2010
Bachelor of Science in Statistics, University of Sci & Tech of China, 2008

Employment

Research Assistant
Applied Statistical Laboratory, University of Kentucky  Aug 2011 to May 2013

Teaching Assistant
Department of Statistics, University of Kentucky  Aug 2008 to Jul 2011

Selected Publications


Copyright© Xiang Zhang, 2013.