Modeling and Inference for Spatial Processes with Ordinal Data

Thesis Proposal
by
Vadim Kutsyy

Advisor
Professor Vijayan N. Nair

Department of Statistics
The University of Michigan
Ann Arbor, MI 48105

March 13, 2001

Contents

List of Figures

Chapter 1

Introduction

1.1 Motivation

Ordinal data appear in many areas of applications including manufacturing, social science, communications and medical research. Often the only way to measure some variables is in the form of ratings (e.g. good, fair and bad). Such data have been studied extensively in the independent case or regression situation and methods of inference have been developed (e.g. Johnson and Albert. [1999]). In this research we consider ordinal data that are spatially dependent . For example, suppose we have recorded some geographical data by counties. We would expect data from nearby counties to be dependent. Similarly in manufacturing settings we may measure a variable in different locations and be concerned about dependence between these locations. For example, in integrated circuit (IC) manufacturing hundreds of integrated circuits are fabricated simultaneously on a disk of silicone called a *wafer* (Figure 1.1). At the end of the process all chips are tested for various failure modes. Some of the tests correspond to tighter specification limits on the same characteristics, so the responses are ordinal in nature.

The simplest type of ordinal data occurs when we have only two categories. Autologistic models for binary data have been known for some time, and inference methods have been studied. Ising [1925] discussed a simple homogeneous first-order auto-logistic model on a countable regular lattice. These models have been of interest to physicists for some time (e.g., Ruelle [1969]). We will describe these models later.

We can think of binary data as a black and white picture. One of the

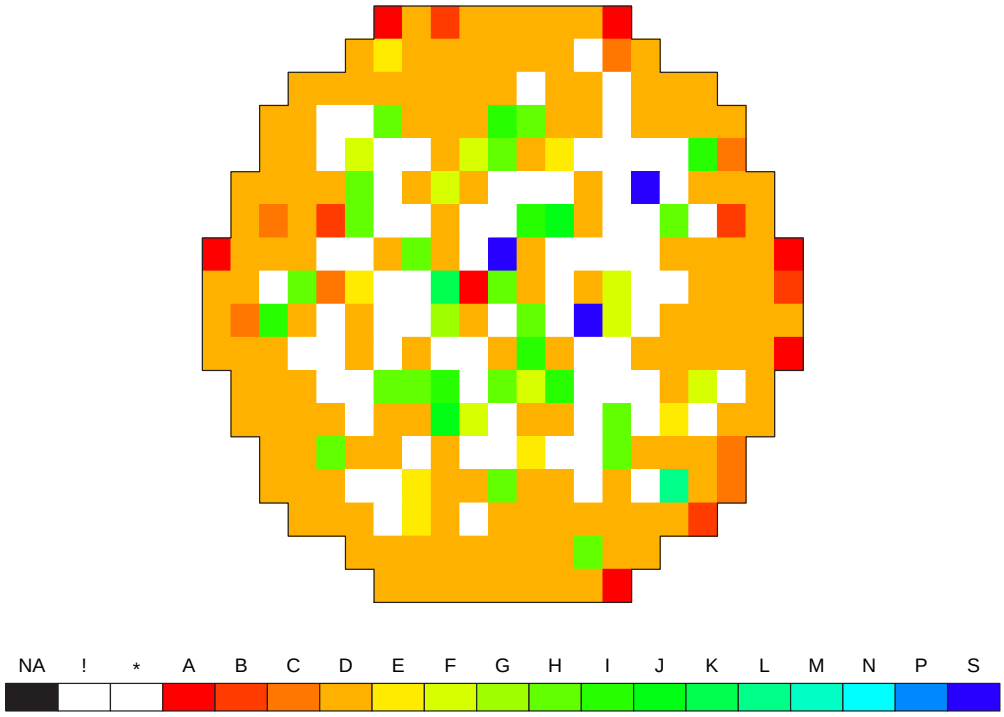

Figure 1.1: Example of a wafer. Each square represents one chip. Colors represent different failure modes. White color represents good chips. Tests are done in alphabetical order.

extensions to such data model is a colored or multinomial model. Such a model also could be seen as a simple way to work with ordinal data, ignoring the ordinal nature and assuming that instead we have categorical data with c categories or colors. Strauss [1977] introduced a model with dependence coefficients β_{kl} between observed variables $Y_i = k$ and $Y_j = l$, were i and j are neighborhood sites. Strauss was able to find approximate maximum likelihood estimators. We will discuss this model in detail later.

Ordinal structure provides a lot of information and ignoring such structure would be inefficient. Strauss' model can be modified to handle ordinal data by putting additional conditions on the β_{ij}'s. Unfortunately, such modification adds additional assumptions. Instead consider the following latent variable model. Let $\{X_i\}$ be a latent continuous variable and let

$$Y_i = k \quad \text{if and only if} \quad \theta_{k-1} \le X_i < \theta_k, \qquad k \in \{1, \dots, c\},$$

for some set of $\theta_0, \dots, \theta_c$. Then instead of the spatial dependence structure of $\{Y_i\}$ we may assume spatial dependence structure on $\{X_i\}$. Often such a variable makes sense. For example, if we measure noise level as loud, normal and quiet, then the underlying level could be a real continuous noise level. In other cases that variable could be thought as combination of a large number of factors. Spatial dependence is usually thought of as local dependence for some neighborhood $\mathcal{N}_i$ around i. In this case $\{X_i, i \in S\}$ is Markov random fields (MRF). MRF and Gaussian MRF are well studied (e.g. Ord [1975]) and we will review them later.

In addition to dependence structure, there are often an explanatory variables that could be used in the model. For example, if additional information has been collected about counties, a prediction could be based on such information. In this case we can think about conditional regression:

$$(X_i | X_{\mathcal{N}_i} = x_{\mathcal{N}_i}) \sim N\left(\phi \sum_{j \in \mathcal{N}_i} x_j + \sum_{k=0}^{p} \beta_k z_{ik}, \sigma^2\right),$$

where Z_i is a vector of covariates for site s_i, and β is a vector of regression coefficients. Such model was also studied by Ord [1975].

However, even if $\{X_i, i \in S\}$ is an MRF, $\{Y_i, i \in S\}$, which is a hidden Markov random field (HMRF), does not have a nice structure. There are some Bayesian studies of HMRF (e.g. Künsch et al. [1995]), but their concern is reconstruction of $\{X_i, i \in S\}$ instead of inference of underlying parameters.

1.2 Scope of this Study

In this study, we view the ordinal spatial process $\{Y_i : i \in S\}$ as an indicator variable obtained by clipping the latent continuous spatial process $\{X_i : i \in S\}$, where S is an enumerable set of sites (generally $S \in \mathbb{Z}^2$ for spatial models). We assume that all sites are numbered $1, \ldots, |S|$, where $|S|$ is total number of sites in S. For notational simplicity we will write $i \in S$, meaning that i is a site of the set S. Then $\{Y_i : i \in S\}$ and $\{X_i; i \in S\}$ are spatial processes observed over the set S defined above. Let $\Theta = (\theta_0, \ldots, \theta_c)$ be set of unknown cutoff points of the continuous latent variable $\mathbf{X}$, with $\theta_0 = -\infty$ and $\theta_c = \infty$. Then we can write the model for obtaining $\{Y_i : i \in S\}$ as follows:

$$Y_i = k \qquad \text{if and only if} \qquad \theta_{k-1} \leq X_i < \theta_k, \quad i \in S; \quad k \in \{1, \ldots, c\}. \tag{1.1}$$

We will assume that $\{X_i : i \in S\}$ is a Gaussian spatial process

$$(X_i | X_{-i} = x_{-i}) \sim N\left(\phi \sum_{j \in \mathcal{N}_i} w_{ij} x_j + \mu_\varepsilon, \sigma^2_{\varepsilon_i}\right), \qquad i \in S, \tag{1.2}$$

where w_{ij} are weights, usually taken to be $w_{ij} = \frac{1}{n_i}$ if $j \in \mathcal{N}_i$ and 0 otherwise for equally spaced lattice. Notation $-i$ means all sites in S except i. Examples of graphical representation of such latent dependence are shown in Figure 1.2. Our goal is to make inference about $\psi = (\phi, \theta_1, \ldots, \theta_{c-1})$, the parameters associated with Gaussian spatial process model with only the indicator process $\{Y_i : i \in S\}$ observed.

Note that we did not say anything about the mean μ_ε and variance $\sigma^2_{\varepsilon_i}$. Without loss of generality one may assume that $\mu_\varepsilon = 0$ and $\sigma^2_{\varepsilon_i} = 1 - \frac{\phi^2}{n_i}$, where n_i is the size of the Markovian neighborhood for site i, which will be discussed later. Otherwise, consider the new spatial process $\{X_i^* : i \in S\}$, where

$$X_i^* = \frac{X_i - \frac{\mu_\varepsilon}{1 - \frac{\phi}{\sqrt{n_i}}}}{\frac{\sigma_{\varepsilon_i}}{\sqrt{1 - \frac{\phi^2}{n_i}}}}$$

with cut points

$$\Theta^* = \left\{ \theta_k^* = \frac{\theta_k - \frac{\mu_\varepsilon}{1 - \frac{\phi}{\sqrt{n_i}}}}{\frac{\sigma_{\varepsilon_i}}{\sqrt{1 - \frac{\phi^2}{n_i}}}} : k = 1, \ldots, c-1 \right\}.$$

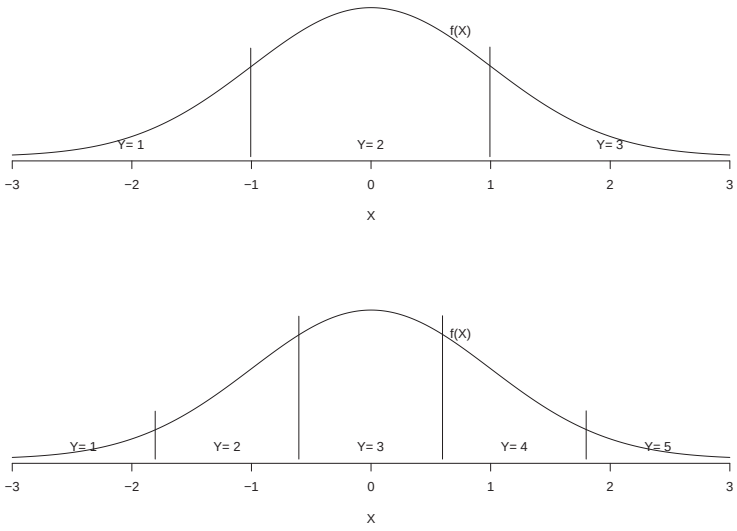

Figure 1.2: Example of a visual representation of a latent variable model for spatial process with ordinal data. In this two examples we can see the relation between latent Gaussian variable X and observed ordinal variable Y. In the first case we have three categories with cut points $\Theta = \left\{-\infty, \Phi^{-1}\left(\frac{1}{3}\right), \Phi^{-}\left(\frac{2}{3}\right), \infty\right\}$, where $\Phi^{-1}(\cdot)$ is inverse of CDF of standard normal variable. In the second example, we have five categories and set of cut points $\Theta = \left\{-\infty, \Phi^{-1}\left(\frac{1}{5}\right), \Phi^{-1}\left(\frac{2}{5}\right), \Phi^{-1}\left(\frac{3}{5}\right), \Phi^{-}\left(\frac{4}{5}\right), \infty\right\}$.

This process will generate exactly the same $\{Y_i : i \in S\}$ spatial process. This phenomena suggest that cut points could be estimated up to a scale only, and the variance σ_{ε_i} and location μ_{ε_i} are not estimable.

With above assumptions we can rewrite $\{X_i : i \in S\}$ spatial process (1.2) as:

$$(X_i | X_{-i} = x_{-i}) \sim N\left(\phi \sum_{j \in \mathcal{N}_i} w_{ij} x_j, 1 - \frac{\phi^2}{n_i}\right), \qquad i \in S.$$

This model can be rewritten in a matrix form:

$$\mathbf{X} \sim N_n(\mathbf{0}, \Lambda^{-1}), \tag{1.3}$$

where $\Lambda = (\mathbf{I}_n - \phi\mathbf{W})'(\mathbf{I}_n - \phi\mathbf{W})$ and $w_{ij} = \begin{cases} \frac{1}{n_i} & j \in \mathcal{N}_i \\ 0 & \text{otherwise} \end{cases}$ for regular equally spaced lattice.

We can extend this model to spatial auto-regression process. For example suppose that Z is a covariate matrix which affects the latent process according to the following model:

$$\mathbf{X} \sim N_n(\mathbf{Z}\beta, \Lambda^{-1}), \tag{1.4}$$

and the response $\{Y_i : i \in S\}$ is generated by the same clipping process (1.1). Of course, we still do not observe $\{X_i : i \in S\}$ and therefore we need to make an inference about (ϕ, Θ, β).

This study is an extension of the research by Wang [1999] on ordinal model for time series data.

The rest of this paper is organized as follows. In Chapter 2 we will review known random fields, models and inference methods. In Chapter 3 we will propose three methods for estimating parameters ψ, investigate their properties and study extension to the auto-regression case. We also will investigate the asymptotic efficiency of the estimators using simulation study. We will conclude our paper by exploring potential future research topics.

Chapter 2

Literature Review

In this section we will review models and inference for random fields. We will start with introduction of Gibbs random fields (GRF), then we will discuss Markov random fields (MRF), and we will introduce hidden Markov random fields (HMRF). We will also discuss important special cases and some known inference methods. Finally, we will look at ordinal data model without spatial correlation.

In this section we will let $\mathbf{X} = \{X_i; i \in S\}$ be a continuous-valued spatial process observed over the set S defined above, and $\mathbf{Y} = \{Y_i; i \in S\}$ be a discrete spatial process observed over the set S. We will also use the lower case $\mathbf{x} = \{x_i; i \in S\}$ and $\mathbf{y} = \{y_i; i \in S\}$ notation for observed data.

2.1 Random Fields

2.1.1 Gibbs Random Fields (GRF)

The most general mathematical model of random field that we will study here is Gibbs random fields (GRF) introduced by Dobrushin and Folguera [1968]. It was proposed as a natural mathematical description of an equilibrium state of a physical system consisting of a very large number of interacting components. A Gibbs measure on the field is a distribution of a countably infinite family of random variables which admit some prescribed conditional probabilities. GRF is defined by conditional probabilities which yield a unique joint distribution. Technical details of GRF are given in Chapter 4.1.

2.1.2 Markov Random Fields (MRF)

Often spatial dependence is thought of as local dependence for some neighborhood $\mathcal{N}_i$ around i.

Definition (Neighborhood). A site j is defined to be a *neighbor* of site i if the conditional distribution of X_i, given all other values, depends functionally on x_j, for $j \neq i$. Also let us define

$$\mathcal{N}_i \equiv \{j : j \quad \text{is a neighbor of} \quad i\} \tag{2.1}$$

which we will call *neighborhood* set of site i. Examples of such neighborhoods will be given later.

For example, in the time series $AR(1)$ case, a neighborhood set will consist only of a previous observation $\mathcal{N}_i \equiv \{i-1\}, i > 1$ and $\mathcal{N}_1 \equiv \emptyset$. There are a lot of different ways to define the neighborhood structure for spatial models. We will discuss a few examples later. Now we can define a process for a given neighborhood structure.

Definition (Markov Random Fields). Any probability measure whose conditional distribution defines a neighborhood structure $\mathcal{N}_i$ through (2.1) is defined to be a *Markov random field*

The Hammersley-Clifford theorem (Besag [1974]) shows that MRF is a special case of GRF. Properties of MRF have been studied extensively (e.g. Guyon [1995]). MRF are such important spatial processes that special cases have been studied extensively as well. We will review some special cases later in this section.

Negpotential Function. In order to study properties of MRF, the joint probability function has to be computed. Often the joint probability of the MRF can be only computed up to a constant. As a result, it makes sense to define a function equal to joint probability only up to a constant as well.

Definition (Negpotential Function). Without loss of generality, assume that reference value 0 can be observed as each site. Then

$$Q(\mathbf{x}) = \ln\left\{\frac{P(\mathbf{x})}{P(\mathbf{0})}\right\} \tag{2.2}$$

is called the *Negpotential Function*.

Here without loss of generality we assume that $P(\mathbf{0}) > 0$ or any other reference level could be used. It is easy to see that knowledge of $Q(\cdot)$ is equivalent to knowledge of $P(\cdot)$. For example, in the discrete case

$$P(\mathbf{x}) = \frac{\exp\{Q(\mathbf{x})\}}{\sum_{\mathbf{z}} \exp\{Q(\mathbf{z})\}}.$$

Following property of Q will be useful for defining MRF models later.

Proposition (Properties of Q).

$$\begin{aligned} Q(\mathbf{x}) = &\sum_{i=1}^{n} x_i G_i(x_i) + \sum_{i=1}^{n-1} \sum_{j=i+1}^{n} x_i x_j G_{ij}(x_i, x_j) + \\ &\sum_{i=1}^{n-2} \sum_{j=i+1}^{n-1} \sum_{k=j+1}^{n} x_i x_j x_k G_{ijk}(x_i, x_j, x_k) + \cdots + \\ &x_1 \ldots x_n G_{1\ldots n}(x_1 \ldots x_n), \quad n = |S|. \end{aligned} \tag{2.3}$$

A proof of this Proposition could be found in Cressie [1993]. Note that $G_{ij\ldots}(x_i, x_j, \ldots)$ in (2.3) are not uniquely defined. By defining $G_{ij\ldots}(x_i, x_j, \ldots) \equiv 0$ whenever $x_i = 0$, or $x_j = 0$, or $\ldots$, uniqueness is obtained.

Pairwise-Only Dependence. Prior to working with MRF we have to choose a neighborhood structure. It makes sense to choose a structure that would make (2.3) simpler.

Definition. *Pairwise-only* dependence between sites is a dependence when $G_A(\cdot) \equiv 0$ in (2.3) for any A whose number of distinct elements is 3 or more.

An example of pairwise-only dependence would be "nearest-neighbor" neighborhood structure of order 1 (Figure 2.1). In this case given site's neighborhood contains one site above, one below, one to the left, and one to the right.

Figure 2.3 gives an example of sites numbering for 10×10 field. Neigh-

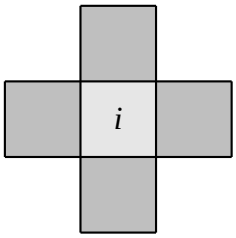

Figure 2.1: "Nearest-neighbor" neighborhood structure.

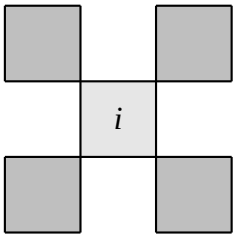

Figure 2.2: "Checker board" neighborhood structure.

borhood sets for "nearest-neighbor" structure are

$$\begin{aligned}
\mathcal{N}_1 &= \{2, 11\} \\
&\dots \\
\mathcal{N}_8 &= \{7, 9, 18\} \\
&\dots \\
\mathcal{N}_{47} &= \{37, 46, 48, 57\} \\
&\dots \\
\mathcal{N}_{73} &= \{63, 72, 74, 83\} \\
&\dots \\
\mathcal{N}_{100} &= \{90, 99\}.
\end{aligned}$$

These sets are also highlighted in the Figure 2.3 for visual demonstration. Neighborhood sets for "checker board" (Figure 2.2) structure for the same

1	2	3	4	5	6	7	8	9	10
11	12	13	14	15	16	17	18	19	20
21	22	23	24	25	26	27	28	29	30
31	32	33	34	35	36	37	38	39	40
41	42	43	44	45	46	47	48	49	50
51	52	53	54	55	56	57	58	59	60
61	62	63	64	65	66	67	68	69	70
71	72	73	74	75	76	77	78	79	80
81	82	83	84	85	86	87	88	89	90
91	92	93	94	95	96	97	98	99	100

Figure 2.3: Example of 10×10 field with total of 100 sites. Examples of "nearest-neighbor" neighborhood structure (Figure 2.1) are shown.

field are

$$\begin{aligned}
\mathcal{N}_1 &= \{12\} \\
&\dots \\
\mathcal{N}_8 &= \{17, 19\} \\
&\dots \\
\mathcal{N}_{47} &= \{36, 38, 56, 58\} \\
&\dots \\
\mathcal{N}_{73} &= \{62, 64, 82, 84\} \\
&\dots \\
\mathcal{N}_{100} &= \{89\}.
\end{aligned}$$

These sets are highlighted in the Figure 2.4.

2.1.3 Hidden Markov Random Fields (HMRF)

Often instead of observing MRF process $\{X_i : i \in S\}$ we observe a process $\{Y_i : i \in S\}$ which is some stochastic or deterministic function of $\{X_i : i \in S\}$. The ordinal latent variable model defines Y_i as a deterministic function of X_i (1.1) which is a Markov random field. As a result $\{Y_i : i \in S\}$ is a hidden Markov random field.

2.2 Gaussian Markov Random Fields

An important special case of MRF is the Gaussian Markov random field. In the latent variable model, we will assume that latent variable is a Gaussian Markov random field. Gaussian fields were originally defined as 2-dimensional generalization of Gaussian time series. It can be defined simultaneously or conditionally. We will describe both, and show equivalence between them. We will also describe inference for GRFs.

2.2.1 Simultaneously Specified Gaussian Markov Random Fields

Whittle [1954] introduced the following class of stationary processes on plane as a generalization of time series. Suppose $\{\varepsilon(u, v) : u = \dots, -1, 0, 1, \dots; v =$

1	2	3	4	5	6	7	8	9	10
11	12	13	14	15	16	17	18	19	20
21	22	23	24	25	26	27	28	29	30
31	32	33	34	35	36	37	38	39	40
41	42	43	44	45	46	47	48	49	50
51	52	53	54	55	56	57	58	59	60
61	62	63	64	65	66	67	68	69	70
71	72	73	74	75	76	77	78	79	80
81	82	83	84	85	86	87	88	89	90
91	92	93	94	95	96	97	98	99	100

Figure 2.4: Example of 10×10 field with total of 100 sites. Examples of "checker board" neighborhood structure (Figure 2.2) are shown.

$\ldots,-1,0,1,\ldots\}$ is a process of independent and identically distributed random variables. Then we can define a process $\{X(u,v)\}$ by

$$\phi(T_1,T_2)X(u,v)=\varepsilon(u,v), \tag{2.4}$$

where T_1 and T_2 are translation operator defined by

$$\begin{aligned} T_1^i X(u,v) &= X(u+i,v), \\ T_2^j X(u,v) &= X(u,v+j), \end{aligned}$$

and ϕ is given by

$$\phi(T_1,T_2)=\sum_i\sum_j a_{ij}T_1^iT_2^j, \tag{2.5}$$

where summation is done over all integers. This simultaneous specification of the variables $X(u,v)$ in (2.4) is analogous to the autoregressive model in time series. The range of summation in (2.5) defines a neighborhood structure. As an example, nearest-neighbor dependence would be specified by

$$\phi(T_1,T_2)=1-\xi(T_1+T_1^{-1}+T_2+T_2^{-1}).$$

Ord [1975] summarized properties and studied estimation methods of Whittle's model on finite lattices for the normal (Gaussian) case. We can write a Gaussian model on finite lattice S as:

$$(X_i|X_{-i}=x_{-i})\sim N\left(\phi\sum_{j\in\mathcal{N}_i}w_{ij}x_j,\sigma\right), \qquad i\in S, \tag{2.6}$$

where w_{ij} are weights, usually taken to be $w_{ij}=\frac{1}{n_i}$ if $j\in\mathcal{N}_i$ and 0 otherwise for equally spaced lattice and S is an enumerable set of sites (generally $S\in\mathbb{Z}^2$). Model (2.6) can be rewritten in matrix form as:

$$\mathbf{X}=\phi\mathbf{W}X+\varepsilon. \qquad \varepsilon\sim N_n(\mathbf{0},\sigma^2I_n). \tag{2.7}$$

Clearly, $E(\mathbf{X})=\mathbf{0}$ and $Var(\mathbf{X})=\sigma^2\left((\mathbf{I}_n-\phi\mathbf{W})'(\mathbf{I}_n-\phi\mathbf{W})\right)^{-1}=\sigma^2\Lambda^{-1}$, where $\Lambda=(\mathbf{I}_n-\phi\mathbf{W})'(\mathbf{I}_n-\phi\mathbf{W})$. Since $\mathbf{X}$ is just a linear combination of ε that is Gaussian,

$$\mathbf{X}\sim N(\mathbf{0},\sigma^2\Lambda^{-1}), \qquad \Lambda=(\mathbf{I}_n-\phi\mathbf{W})'(\mathbf{I}_n-\phi\mathbf{W}).$$

The likelihood of (ϕ,σ) is given by

$$L(\phi,\sigma^2)\propto\frac{|\Lambda|^{\frac{1}{2}}}{\sigma}e^{-\frac{1}{2\sigma^2}\mathbf{x}\Lambda\mathbf{x}}. \tag{2.8}$$

From (2.8) we get ML estimators as:

$$\hat{\sigma}^2 = n^{-1}\mathbf{x}\Lambda\mathbf{x}$$

and $\hat{\phi}$ as the value that maximizes (Mead [1967])

$$l(\phi, \hat{\sigma}^2) \propto -\frac{n}{2}\ln\hat{\sigma}^2|\Lambda|^{-\frac{1}{n}}. \tag{2.9}$$

Unfortunately, there is no closed form solution for (2.9) and numerical maximization requires evaluation of $|\Lambda|$ at each step. Ord [1975] noticed that if $\mathbf{W}$ has eigenvalues $\lambda_1, \ldots, \lambda_n$, then

$$|\Lambda| = |\mathbf{I}_n - \phi\mathbf{W}|^2 = \left\{\prod_{i=1}^{n}(1-\phi\lambda_i)\right\}^2.$$

Using the fact that $\{\lambda_i\}$'s need be determined just once, prior to maximization, ML estimator of ϕ is the value $\hat{\phi}$ that minimizes

$$\left\{\prod_{i=1}^{n}(1-\phi\lambda_i)\right\}^{-\frac{2}{n}}\left(\mathbf{x}'\mathbf{x} - 2\phi\mathbf{x}'\mathbf{x}'_L + \phi^2\mathbf{x}'_L\mathbf{x}_L\right), \qquad \mathbf{x}_L = \mathbf{W}x. \tag{2.10}$$

This can be found fairly quickly.

2.2.2 Conditionally Specified Gaussian Markov Random Fields

Instead of defining a model simultaneously for all sites, it makes sense to define a model for a given site based on all other sites. We can write the conditional distribution of the variable $\mathbf{X}$ at some site i given all other sites. For notational purposes we will use $-i$ to be the set of all sites except site i. Under the condition of "pairwise-only dependence"

$$f(x_i|x_{-i}) = N\left(\phi\sum_{j=1}^{n}c_{ij}x_j, \sigma^2\right), \qquad i = 1, \ldots, n. \tag{2.11}$$

For regular equally spaced lattices (e.g. Figure 2.3) conditional weights are usually defined as

$$c_{ij} = \frac{1}{n_i}, \quad i, j \in S. \tag{2.12}$$

For non-regular lattices (e.g. geographical data) the weights are defined to be inversely proportional to the distances between sites. Cressie [1993] showed that (2.11) is identical to

$$\mathbf{X} \sim N_n(\mathbf{0}, \sigma^2(\mathbf{I}_n - \phi\mathbf{C})^{-1}). \tag{2.13}$$

2.2.3 Comparison of Simultaneously and Conditionally Specified Gaussian Markov Random Fields

Brook [1964] was the first one to make a distinction between the simultaneous specification and conditional specification of the spatial model. Cressie [1993] showed that both models (2.7 and 2.13) are identical given that their variance matrices are equal:

$$\sigma^2(\mathbf{I}_n - \phi\mathbf{C}) = \sigma^2(\mathbf{I}_n - \phi\mathbf{W})'(\mathbf{I}_n - \phi\mathbf{W})$$

or

$$\mathbf{C} = \phi\mathbf{W}'\mathbf{W} - \mathbf{W} - \mathbf{W}' \tag{2.14}$$

Equation (2.14) gives us relation between the two models. Note that any simultaneously specified model can be written as a conditionally specified model, but a conditionally specified model can correspond to more than one simultaneously specified model. Another interesting fact is that if $\sum_{j\in N_i} w_{ij}x_j$ and ε_i are uncorrelated, i.e. $E\{\varepsilon_i | X_j = x_j, j \in \mathcal{N}_i\} = 0$, both models are identical (Ord [1975]).

2.2.4 Spatial Auto-regression

The previous models could be extended by introducing variation in the mean level. Ord [1975] considered the mixed regressive-autoregressive model for the simultaneously defined model

$$\mathbf{X} = \mathbf{A}\beta + \phi\mathbf{W}X + \varepsilon, \qquad \varepsilon \sim N_n(\mathbf{0}, \sigma^2\mathbf{I}_n), \tag{2.15}$$

where $\mathbf{A}$ is an $(n \times p)$ matrix of covariates, and β is a $(p \times 1)$ vector of regression parameters. Ord showed that ML estimators for β and σ^2 are

$$\begin{aligned} \hat{\beta} &= (\mathbf{A}'\mathbf{A})^{-1}\mathbf{A}'\hat{\mathbf{z}}, \\ \hat{\sigma}^2 &= \frac{1}{n}\hat{\mathbf{z}}'\mathbf{M}\hat{\mathbf{z}}, \end{aligned}$$

where $\hat{\mathbf{z}} = (\mathbf{I}_n - \hat{\phi}\mathbf{W})\mathbf{x}$ and $\mathbf{M} = \mathbf{I}_n - \mathbf{A}(\mathbf{A}'\mathbf{A})^{-1}\mathbf{A}'$. $\hat{\phi}$ is the value which maximizes, as before,

$$-\frac{n}{2}\ln\hat{\sigma}^2|\Lambda|^{-\frac{1}{n}}, \qquad \Lambda = (\mathbf{I}_n - \phi\mathbf{W})'(\mathbf{I}_n - \phi\mathbf{W}).$$

For computational purposes, an equation similar to (2.10) is used:

$$\left\{\prod_{i=1}^{n}(1-\phi\lambda_i)\right\}^{-\frac{2}{n}} \left(\mathbf{x}'\mathbf{M}x - 2\phi\mathbf{x}'\mathbf{M}x_L' + \phi^2\mathbf{x}_L'\mathbf{M}x_L\right), \qquad \mathbf{x}_L = \mathbf{W}x,$$

It is important to note that instead of (2.15) auto-regressive models are often defined as

$$(\mathbf{X} - \mathbf{A}\beta) = \phi\mathbf{W}(\mathbf{X} - \mathbf{A}\beta) + \varepsilon, \qquad \varepsilon \sim N_n(\mathbf{0}, \sigma^2\mathbf{I}_n),$$

which is just a one-to-one reparametrization of (2.15).

2.3 Conditionally Specified Spatial Models for Binary Data

Often instead of continuous data, discrete data is observed. In this section we will review a few special conditionally specified spatial models for discrete data.

2.3.1 Auto-Logistic Model for Binary Data

In order to model independed binary data an auto-logistic model is used. A similar auto-logistic model is also defined for spatially dependent data with a pairwise-dependent neighborhood structure. Binary spatial data can be thought as a black and white picture. Assume that the observed data is either 0 (white) or 1 (black). This data often arises from the absence or presence of some characteristic. Because of the nature of the data, the only important values of the G function in (2.3) are, assuming pairwise only dependence between sites, $G_i(1) \equiv \alpha_i$ and $G_{ij}(1,1) \equiv \theta_{ij}$. Because of the pair-wise only dependence all higher order terms are equal to 0. Thus,

$$Q(\mathbf{x}) = \sum_{i=1}^{n}\alpha_i x_i + \sum_{i=1}^{n-1}\sum_{j=i+1}^{n}\theta_{ij}x_i x_j, \tag{2.16}$$

where $\theta_{ik} \equiv 0$ if $k \notin \mathcal{N}_i$. It is easy to see (Cressie [1993]) that:

$$P(x_i|x_{-i}) = P(y_i|y_{\mathcal{N}_i}) = \frac{\exp\left\{\alpha_i x_i + \sum_{j\in\mathcal{N}_i} \theta_{ij} x_i x_j\right\}}{1+\exp\left\{\alpha_i + \sum_{j\in\mathcal{N}_i} \theta_{ij} y_j\right\}}, \quad x_i \in \{0,1\}, \quad i \in S, \tag{2.17}$$

This is called an *auto-logistic model for spatial data*.

2.3.2 Ising Model

MRF has been of interest to physicists for long time. Ising [1925] discussed a simple homogeneous first-order auto-logistic model on a countable regular lattice $D = \{(u,v) : u = \ldots,-1,0,1,\ldots; v = \ldots,-1,0,1,\ldots\}$. This model has been extensively studied in statistical physics and is referred as the classical *Ising model* (e.g., Ruelle [1969]). The model reduces to:

$$P\left(x(u,v)|\{x(k,l) : (k,l) \neq (u,v)\}\right) = \frac{\exp\{g\}}{1+\exp\{g\}}, \quad u,v = \ldots,-1,0,1,\ldots, \tag{2.18}$$

where

$$g \equiv x(u,v)\left\{\alpha + \gamma(x(u-1,v) + x(u+1,v) + x(u,v-1) + x(u,v+1))\right\}.$$

2.4 Conditionally Specified Spatial Models for Multicolored Data

Models for binary data can be extended to categorical data. Categorical spatial data can be seen as a multicolored picture, where each color represent a category.

2.4.1 Strauss [1977] Model

Strauss [1977] generalized Ising model (2.18) to multicolored case. Suppose instead of black and white image each site can have more than 2 colors. i.e.

$x_i = 1, \ldots, c$, where each number represents a color. Strauss [1977] showed that for the pair-wise dependence neighborhood structure

$$Q(\mathbf{x}) = \sum_{i=1}^{n} x_i G_1(x_i) + \sum_{i=1}^{n-1} \sum_{j=i+1}^{n} x_i x_j G_{1,2}(x_i, x_j) = \sum_{r=1}^{l} m_r u_r + \sum_{r=1}^{l} \sum_{s=1}^{l} n_{rs} v_{rs}, \tag{2.19}$$

where $u_r \equiv rG(r)$, $v_{rs} = rsG(r,s)$, $r,s = 1, \ldots, l$ and $m_r \equiv$ number of sites that have color r, $n_{rs} \equiv$ number of pairs of neighboring sites where one has color r and the other has color s. Under two further assumptions (color indifference and equal strength of attraction for all colors), Strauss was able to find approximate maximum likelihood estimates; the approximation is necessary due to the normalizing constant $\sum_{\mathbf{x}} \exp(Q(\mathbf{x}))$

2.5 Inference for Markov Random Fields

Inference for MRF has been studied extensively, and a number of different techniques have been developed. Below we will review a few common methods.

2.5.1 Likelihood Estimation

Let us call the collection of all unknown parameters ψ. The log-likelihood $l(\psi)$ is defined as

$$l(\psi) \equiv \ln\{L(\psi)\} = \ln\{P(\mathbf{X}; \psi)\}. \tag{2.20}$$

By maximizing $l(\psi)$ over ψ we can find the maximum likelihood estimator (MLE) $\hat{\psi}$. Unfortunately, for spatial data the MLE is difficult to compute. Even for the Gaussian case discussed above, the MLE computation requires calculating the determinant of an $|S|$-dimensional matrix. The MLE could be found for some other special cases, but in general, the MLE cannot be computed easily.

2.5.2 Pseudo-Likelihood Estimation

One possibility is to use an approximation to the likelihood, which is easy to compute. Besag [1975] considered the log *pseudo-likelihood*

$$\begin{aligned} l_p(\psi) &\equiv \ln\{\prod_{i=1}^{n} P(\mathbf{x}_i|\mathbf{x}_{-i};\psi)\} \\ &= \sum_{i=1}^{n} \ln\{P(\mathbf{x}_i|\mathbf{x}_{-i};\psi)\} = \sum_{i=1}^{n} \ln\{P(\mathbf{x}_i|\mathbf{x}_j, j \in \mathcal{N}_i;\psi)\}. \end{aligned} \tag{2.21}$$

The log pseudo-likelihood (2.21) is to be maximized with respect to ψ, to yield the maximum pseudo-likelihood estimator $\hat{\psi}_P$. For conditionally specified Gaussian model (2.11), maximum pseudo-likelihood estimators of these parameters are simply the ordinary least squares estimators. For auto-logistic model (2.16) assuming a so-called isotropic Ising model on $\mathbb{Z}^2$ (i.e. assume (2.18) with $\gamma_1 = \gamma_2$) the maximum pseudo likelihood estimator is equivalent to formal maximum likelihood estimator for the logistic regression model (Strauss and Ikeda [1990]). This observation allows one to use the logistic-regression option in computer packages.

2.5.3 Estimation Based on Coding

For conditionally specified models with pair-wise dependence, Besag [1974] proposed a method of estimation called *coding*. The idea is to divide up the lattice D into two disjoint sub-lattices D_0 and D_1, where the neighborhood structure of D_0 is the trivial one of no two sites being neighbors of each other. Coding estimates are obtained by minimizing

$$C(\psi) \equiv -\sum_{i \in D_0} \ln\{P(\mathbf{x}_i|\mathbf{x}_{-i};\psi)\}, \tag{2.22}$$

which is the conditional likelihood of $\{\mathbf{X}_i : i \in D_0\}$ given $\{\mathbf{X}_j : j \in D_1\}$. For example, for the first order dependence model on $\mathbb{Z}^2$, the neighborhood structures are given by Figure 2.1. Then D_0 can be constructed by deleting every other point in D. This is clearly an inefficient way to use the data (half of the data is deleted), but note that in this example the other half of the data could be used in the same way. Thus, there are two possible coding estimators that could be averaged. Usually the coding estimator is easy to compute, so in some situations where other estimators are unavailable, this estimator is an alternative.

Bayesian Estimation Methods The Bayesian approach was first considered in the Geman and Geman [1984] article. They considered maximum a posteriori (MAP) restoration of latent variable. Suppose we have observed a spatial noisy process $\mathbf{Z}$ that is related to some underlying unobserved variable $\mathbf{X}$ through probability function $P(\mathbf{Z}|\mathbf{X};\psi)$. Let $\pi(\cdot)$ denote the prior distribution of $\mathbf{X} \in \mathbb{X}$. Then posterior distribution for $\mathbf{X}$ given $\mathbf{Z}$ is

$$P(\mathbf{X}|\mathbf{Z}) = \frac{f(\mathbf{Z}|\mathbf{X})\pi(\mathbf{X})}{\int_{\mathbb{X}} f(\mathbf{Z}|\tau)\pi(\tau)d\tau} \tag{2.23}$$

Assuming a 0-1 loss function, the joint prediction of $\mathbf{X}$ given $\hat{\mathbf{X}} \in \mathbb{X}$ that maximized (2.23) is a Bayes rule and is called the maximum a posteriori (MAP) estimator. More detailed description with examples can be found in Geman [1988]

Chapter 3

A Latent Variable Model for Spatial Process with Ordinal Data

Here we will study inference for the model introduced in Section 1.2.

For the case when $\{X_i : i \in S\}$ is a spatial Gaussian process, inference has been studied by Ord [1975]. In our case $\{X_i : i \in S\}$ is not observed, and inference has to be based on the $\{Y_i : i \in S\}$ process. The $\{Y_i : i \in S\}$ process is a HMRF.

3.0.4 Maximum Likelihood Estimator (MLE)

The maximum likelihood estimator is commonly used, so we will consider this estimator first. For the model described above we can write the likelihood function:

$$L(\Theta, \phi | \mathbf{Y} = y) \propto \int_{\theta_{y_1-1}}^{\theta_{y_1}} \int_{\theta_{y_2-1}}^{\theta_{y_2}} \dots \int_{\theta_{yn-1}}^{\theta_{yn}} |\Lambda|^{1/2} \exp\left\{-\frac{1}{2}\mathbf{X}'\Lambda X\right\} dx_1 dx_2 \dots dx_n \quad (3.1)$$

where $\Lambda = (I - \phi W)'(I - \phi W) = \Sigma^{-1}$. The likelihood function is to be be maximized with respect to $\psi' = (phi, \theta')$, to yield the maximum likelihood estimator $\hat{\psi}$. Evaluation of (3.1) involves an n-dimensional integral of a multivariate normal. For a simple lattice of size 20×20, $n = 400$ which could be numerically evaluated using numerical methods described above, but calculation

time even for a single integration takes very long. Numerical maximization is not feasible at the present time.

3.0.5 Numerical Computations of Multidementional Normal Integral

During computation of the MLE and other estimators, numerical computation of the multivariate normal integral is required. A fast and highly accurate algorithm for univariate normal could be found in Johnson et al. [1994]. There is also reliable and efficient software available for the bivariate case (e.g. Donnelly [1973]). However, higher dimensional cases are not that easy to compute.

Numerical Integration Method Schervish [1984] developed an algorithm for evaluation of multivariate normal integral based on numerical integral evaluation methods. This particular algorithm is based on Simpson's rule integration. Efficient error control permits computation of integrals up to a small error. The algorithm is fairly efficient for the two and three dimensional case, but the amount of the computation increases exponentially as the number of variables increases. As a result, even 5-dimensional integral is not practical to compute today. The nearest neighborhood structure described above consists of 4 sites per neighborhood which results in 5-dimensional integration for quasi-likelihood estimation.

Gaussian Quadratures Method Drezner [1990] developed a more efficient algorithm for evaluation of multivariate normal integrals based on Gaussian quadratures. The computational speed of this method is linearly proportional to the number of variables. However, the error rate is not as easily controlled because of the irregular set of weights that have to be computed initially. As a result, this algorithm performs fairly well if the value of the integral is not "very" close to 0 or 1.

MCMC Based Method Genz [1992] developed an algorithm of evaluation for multivariate normal integral based on MCMC. He also developed a method for easy control of the proportional error rate. The algorithm is very fast for probabilities away from 0 and slower as the integral value decreases to 0. As a result, computation time of high dimensional integrals ($n > 10$) that usually have very small values is very long, but it still possible to find the values.

3.0.6 The EM Algorithm

In order to evaluate MLE, the EM algorithm is often used. The EM algorithm is an iterative algorithm for finding MLE numerically in situations with missing data. It consists of two steps, the **E** step and the **M** step. In our model we may consider $\{X_i : i \in S\}$ to be missing data. The **E** step consists of computing

$$\begin{aligned} Q(\psi,\psi^i) &= \int_X \log[L(\psi|X,Y)]\,P(X|\psi^i,Y)\mathrm{d}X = \\ &\int_{\theta_{y_1-1}}^{\theta_{y_1}} \int_{\theta_{y_2-1}}^{\theta_{y_2}} \dots \int_{\theta_{y_n-1}}^{\theta_{y_n}} |\Lambda|^{1/2} \exp\left\{-\frac{1}{2}\mathbf{X}'\Lambda X\right\} \\ &\left[|\Lambda^i|^{1/2} \exp\left\{-\frac{1}{2}\mathbf{X}'\Lambda^i X\right\} \prod_{j\in S} I_{\theta^i_{y_j-1}\le x_j<\theta^i_{y_j}}\right] \mathrm{d}x_1\mathrm{d}x_2\dots\mathrm{d}x_n, \end{aligned}$$

where ψ^i is a current guess. In the **M** step the Q function is maximized with respect to ψ to obtain $\psi+i$. This algorithm also requires n-dimensional integration and as a result is not feasible either due to computational problems discussed above.

3.0.7 Quasi-Likelihood Estimator (QLE)

Similar to Besag's [1975] approach, we will consider approximations that make a compromise between efficiency and computational tractability. Because of Markovian property of $\{X_i; i \in S\}$, i.e. $P(X_i|X_{-i}=x_{-i}) = P(X_i|X_{\mathcal{N}_i} = x_{\mathcal{N}_i})$, one may expect that similar approximation will work for $\{Y_i; i \in S\}$ as well. Following (2.21) we can define the individual log quasi-likelihood for site i as :

$$\begin{aligned} l^i_{QL}(\Theta,\phi|y_i,y_{\mathcal{M}}) &= \ln P(y_i|y_{\mathcal{M}};\Theta,\phi) = \ln\frac{P(y_i,y_{\mathcal{M}};\Theta,\phi)}{P(y_{\mathcal{M}};\Theta,\phi)} = \\ &= \ln\frac{\int_{\theta_{y_i-1}}^{\theta_{y_i}} \int_{\theta_{y_{\mathcal{M}_1}-1}}^{\theta_{y_{\mathcal{M}_1}}} \dots \int_{\theta_{y_{\mathcal{M}_m}-1}}^{\theta_{y_{\mathcal{M}_m}}} |\Lambda_{i,\mathcal{M}}|^{1/2} \exp\left\{-\frac{1}{2}\mathbf{X}'_{i,\mathcal{M}}\Lambda_{i,\mathcal{M}}\mathbf{X}_{i,\mathcal{M}}\right\}\mathrm{d}x_i\mathrm{d}x_{\mathcal{M}_1}\dots\mathrm{d}x_{\mathcal{M}_m}}{\int_{\theta_{y_{\mathcal{M}_1}-1}}^{\theta_{y_{\mathcal{M}_1}}} \dots \int_{\theta_{y_{\mathcal{M}_m}-1}}^{\theta_{y_{\mathcal{M}_m}}} |\Lambda_{\mathcal{M}}|^{1/2} \exp\left\{-\frac{1}{2}\mathbf{X}'_{\mathcal{M}}\Lambda_{\mathcal{M}}\mathbf{X}_{\mathcal{M}}\right\}\mathrm{d}x_{\mathcal{M}_1}\dots\mathrm{d}x_{\mathcal{M}_m}}, \end{aligned}$$

where $\mathcal{M} \equiv \mathcal{N}_i$, $m = n_i$ and $\Lambda_A[i,j] = \Lambda[A_i, A_j]$, $X_A[i] = X[A_i]$ for any set A. Then the log quasi-likelihood could be computed as:

$$l_{QL}(\Theta, \phi | \mathbf{Y}) = \sum_{i \in S} l^i_{QL}(\Theta, \phi | y_i, y_{\partial i}) \tag{3.2}$$

As a result, we have reduced calculation of the n-dimensional integral to calculation of $2n$ integrals whose dimensions are at most one more than size of the neighborhood structure. For example, if we use "nearest-neighbor" neighborhood (Figure 2.1), then the integral dimensions are at most 5 which is feasible to do on today's computers. It is still time consuming.

"Nearest-neighbor" neighborhood structure (Figure 2.1) assume that vertical and horizontal dependence are identical, and any estimation methods for this neighborhood estimate vertical and horizontal dependence at once. Instead we may try to estimate vertical dependence first, then estimate hor-

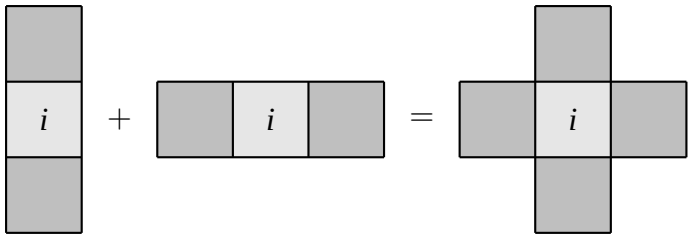

Figure 3.1: "Nearest-neighbor" neighborhood structure decomposition into vertical and horizontal neighborhoods.

izontal and take an average of two. Such decomposition is illustrated in Figure 3.1. Each of the resulting neighborhoods is of size 3, so numerical calculations simplify. Additional research is needed to assess properties of such decomposition.

3.0.8 Mean Based Approximation (*MnA*)

Another way to approximate the likelihood is to impute values of X. The EM algorithm is one such approach. However, the EM algorithm is not computationally feasible as described above. First notice that if value of X in the neighborhood of i is known, then Y_i depends only on these values:

$$P(Y_i | Y_{-i} = y_{-i}, X_{-i} = x_{-i}) = P(Y_i | X_{\mathcal{N}_i} = x_{\mathcal{N}_i}).$$

The values of X can be imputed using the marginal expected values, i.e. $\hat{\mu}_j^X = E(X_j|Y_j;\Theta,\phi)$:

$$\hat{\mu}_j^X = \frac{\phi(\theta_{y_j-1}) - \phi(\theta_{y_j})}{\Phi(\theta_{y_j-1}) - \Phi(\theta_{y_j})} \tag{3.3}$$

where $\phi(\cdot)$ is the standard normal PDF, and $\Phi(\cdot)$ is the standard normal CDF (Johnson et al. [1994]). So we can use the likelihood function

$$\begin{aligned} l_{MnA}(\mathbf{Y}) &= \sum_{i\in S} \ln\left\{P(Y_i|Y_{-i}=y_{-i}, X_{-i}=\hat{\mu}^X_{\mathcal{N}_i})\right\} \\ &= \sum_{i\in S} \ln\left\{P(Y_i|X_{\mathcal{N}_i}=\hat{\mu}^X_{\mathcal{N}_i})\right\} \qquad (3.4) \\ &= \sum_{i\in S} \ln\left\{\Phi\left(\frac{\theta_{y_i} - \phi\times\sum\limits_{j\in\mathcal{N}_i} w_{ij}\hat{\mu}_i^X}{1-\frac{\phi^2}{n_i}}\right) - \Phi\left(\frac{\theta_{y_i-1} - \phi\times\sum\limits_{j\in\mathcal{N}_i} w_{ij}\hat{\mu}_i^X}{1-\frac{\phi^2}{n_i}}\right)\right\}. \end{aligned}$$

This approximate likelihood function can be maximized with respect to ψ, to yield the mean based estimator $\hat{\psi}_{MnA}$.

3.0.9 Median Based Approximation (MdA)

Instead of imputing x_i as a marginal mean of x_i given y_i we can impute it as a conditional median for a given y_i. Then approximate likelihood function will be as follows:

$$\begin{aligned} l_{MdA}(\mathbf{Y}) = \sum_{i=1}^{n} \ln\Bigg\{ &\Phi\left(\frac{\theta_{y_i} - \phi\times\sum\limits_{j\in\mathcal{N}_i} w_{ij}\hat{M}_X^i}{1-\frac{\phi^2}{n_i}}\right) - \\ &\Phi\left(\frac{\theta_{y_i-1} - \phi\times\sum\limits_{j\in\mathcal{N}_i} w_{ij}\hat{M}_X^i}{1-\frac{\phi^2}{n_i}}\right)\Bigg\}, \end{aligned} \tag{3.5}$$

where

$$\hat{M}(x_j \mid y_j) = \Phi^{-1}\left(\frac{\Phi(\theta_{y_j-1}) + \Phi(\theta_{y_j})}{2}\right),$$

is a median of truncated normal distribution (Johnson et al. [1994]). We will compare the performances of MnA and MdA later.

3.0.10 Bayesian Estimation

Bayesian methods for image restoration have been studied (e.g. Geman [1988]). These methods were used for predicting unobserved latent variable X. Instead we will discuss the data augmentation technique (Tanner [1996]) which allow us to study posterior distribution of unknown parameters. Data augmentation is a computational device for obtaining posterior distributions of the parameters of interest. It exploits the feature that the likelihood function or posterior distribution of the parameters is very simple when the "missing" data are augmented. The joint distribution of X and Y given ψ is:

$$f(y,x|\psi) = \prod_{i\in S} I_{\{\theta_{y_i-1}\le x_i<\theta_{y_i}\}} \prod_{i\in S} g(x_i|x_{\mathcal{N}_i}:\phi).$$

Now suppose that Θ and ϕ have prior densities $h_\Theta(\Theta)$ and $\pi_\phi(\phi)$ respectively. It is also reasonable to assume that prior distributions of Θ and ϕ are independent and hence the prior distribution of ψ is $\pi(\psi) = \pi(\Theta)\pi(\phi)$. This gives the posterior distribution of ψ given x and y as

$$\pi(\psi|x,y) \sim \pi(\Theta) \prod_{i\in S} I_{\{\theta_{y_i-1}\le x_i<\theta_{y_i}\}} \times \pi(\phi) \prod_{i\in S} g(x_i|x_{\mathcal{N}_i}:\phi).$$

It is worth noticing that Θ and ϕ have independent posterior distributions. This enables us to sample Θ and ϕ separately. The choice of the priors and numerical implementation will be considered for future research.

3.1 Extension to Spatial Regression Model

An interesting question arises when we have covariates. In that case, our spatial model becomes:

$$(\mathbf{X} - \beta A) = \phi \mathbf{W} (\mathbf{X} - \beta A) + \varepsilon,$$

where $\mathbf{A}$ is the covariate matrix, and β is the matrix of covariate coefficients. However, let $\Theta_{l(x)} = (\theta_{y_1-1}, \ldots, \theta_{y_n-1})$, $\Theta_{u(x)} = (\theta_{y_1}, \ldots, \theta_{y_n})$ and $\Theta_{u(x)_i}\} = \theta_{y_i-1}$, $\Theta_{l(x)_i}\} = \theta_{y_i}$. Then our model could be written as:

$$Y_i = k \qquad \text{if and only if} \qquad \{\Theta_{l(x)_i} < X_i < \Theta_{u(x)_i}\}.$$

Now, let $\mathbf{Z} = X - \beta A$, $\Theta_{u(z)} = \Theta_{u(x)} - \beta A$ and $\Theta_{l(z)} = \Theta_{l(x)} - \beta A$. Then we can re-write our model as:

$$Y_i = k \qquad \text{if and only if} \qquad \{\Theta_{l(z)_i} < Z_i < \Theta_{u(z)_i}\},$$

$$\mathbf{Z} = \phi\mathbf{Z} + \varepsilon.$$

This is different from the previous models only in the limits of integration, so numerical calculation of the quasi-likelihoods is unchanged.

3.2 Hypothesis Testing

One may be interested in testing whether given coefficient (ϕ or β) is different from 0. In order to test this null hypotheses we may use the likelihood ratio test. As discussed above, MLEs cannot be obtained numerically. So we may use one of the approximation methods described.

Use of QLE, MnA and MdA makes hypothesis testing possible; however, performance of these estimators is unknown and will be studied in future research.

3.3 Asymptotic Properties of QLE

In the next two sections we will establish some theoretical results. Most of the basic results are taken from Georgii [1988] and Guyon [1995].

Notation Let $S = \mathbb{Z}^2$, Let D_n be a sequence of toruses, $|D_n| \to \infty$. Let a *random field* over S be a probability measure μ over $D_n \in \mathbb{Z}^2$.

Suppose the field $\{X_i; i \in S\}$ and $\{Y_i; i \in S\}$ are defined by 1.2 and 1.1 respectively. Then it is easy to see that X is a MRF and Y is a HMRF.

3.3.1 Ergodic Theorem

If $\{Y_i : i \in s\}$ is ergodic then asymptotic properties are easier to study. It is easy to see that $\{Y_i : i \in s\}$ as defined in (1.1) and (1.2) is a GRF and satisfies weak dependence Dobrushin's condition. The proof of that result is given in Section 4.2. However, ergodicity is a stronger result.

Theorem 1 (Ergodic Theorem for $\{Y_i : i \in S\}$). *If $\{X_i : i \in S\}$ is defined by (1.2), then $\{Y_i : i \in S\}$ defined by (1.1) is stationary and ergodic.*

Proof. $\{X_i : i \in S\}$ is Gaussian MRF, which is stationary and ergodic, then the result follows from **Example 24.6** in Billingsley [1995] similar to Theorem 36.4 in Billingsley [1995]. □

3.3.2 Consistency of QLE

To prove consistency we will use results in Guyon [1995]. It worth noticing that results in Guyon [1995] are more general and do not require ergodicity.

Theorem 2 (Strong Consistency of QLE). *If $\hat{\psi}_n$ is QLE over D_n, $|D_n| \to \infty$, then*

$$\lim_{n\to\infty} \hat{\psi}_n = \psi_0 \qquad a.e.$$

Proof. Let $K(\psi_0, \psi) = E_{\psi_0}\left(-\ln \frac{P(Y_i=y_i|Y_{\mathcal{N}_i}=y_{\mathcal{N}_i};\psi_0)}{P(Y_i=y_i|Y_{\mathcal{N}_i}=y_{\mathcal{N}_i};\psi)}\right)$, where ψ_0 is the true value of the parameters. We notice that

$$\begin{aligned}\lim_n\left(\frac{1}{|D_n|}l_{QL}(\psi|Y)\right) &= \lim_n\left(\frac{1}{|D_n|}\sum_{i\in D_n}\ln P(Y_i=y_i|Y_{\mathcal{N}_i}=y_{\mathcal{N}_i};\psi)\right)\\ &= \mathrm{E}_{\psi_0}\left(\ln P(Y_1=y_1|Y_{\mathcal{N}_1}=y_{\mathcal{N}_1};\psi)\right),\end{aligned}$$

where the last equality follows from the fact that $\{Y_i, i\in D_n\}$ is ergodic and $l_{QL}(\psi|Y) = l_{QL}(\Theta,\phi|Y)$ is defined in (3.2). Then

$$\begin{aligned}&\lim_n\left(\frac{1}{|D_n|}l_{QL}(\psi|Y) - \frac{1}{|D_n|}l_{QL}(\psi_0|Y)\right)\\ &= \mathrm{E}_{\psi_0}\left(\ln P(Y_1=y_1|Y_{\mathcal{N}_1}=y_{\mathcal{N}_1};\psi)\right)\\ &\quad - \mathrm{E}_{\psi_0}\left(\ln P(Y_1=y_1|Y_{\mathcal{N}_1}=y_{\mathcal{N}_1};\psi_0)\right)\\ &= E_{\psi_0}\left(-\ln\frac{P(Y_i=y_i|Y_{\mathcal{N}_i}=y_{\mathcal{N}_i};\psi_0)}{P(Y_i=y_i|Y_{\mathcal{N}_i}=y_{\mathcal{N}_i};\psi)}\right)\\ &= K(\psi_0,\psi)\geq 0, \quad =0 \quad \text{only if} \quad \psi=\psi_0,\end{aligned}$$

where the inequality follows from the fact that $L_{QL}(\psi|Y) > L_{QL}(\psi_0|Y)$ by definition of QLE ψ_n. We also may notice that for latent variable model for spatial process with ordinal data

$$\begin{aligned}&P(y_i|y_{\mathcal{M}};\psi'=(\phi,\theta')) = \ln\frac{\pi(y_i,y_{\mathcal{M}};\psi'=(\phi,\theta'))}{\pi(y_{\mathcal{M}};\psi'=(\phi,\theta'))} =\\ &= \ln\frac{\int_{\theta_{y_{i-1}}}^{\theta_{y_i}}\int_{\theta_{y_{\mathcal{M}_1}-1}}^{\theta_{y_{\mathcal{M}_1}}}\dots\int_{\theta_{y_{\mathcal{M}_m}-1}}^{\theta_{y_{\mathcal{M}_m}}}|\Lambda_{i,\mathcal{M}}|^{1/2}\exp\left\{-\frac{1}{2}\mathbf{X}'_{i,\mathcal{M}}\Lambda_{i,\mathcal{M}}\mathbf{X}_{i,\mathcal{M}}\right\}\mathrm{d}x_i\mathrm{d}x_{\mathcal{M}_1}\dots\mathrm{d}x_{\mathcal{M}_m}}{\int_{\theta_{y_{\mathcal{M}_1}-1}}^{\theta_{y_{\mathcal{M}_1}}}\dots\int_{\theta_{y_{\mathcal{M}_m}-1}}^{\theta_{y_{\mathcal{M}_m}}}|\Lambda_{\mathcal{M}}|^{1/2}\exp\left\{-\frac{1}{2}\mathbf{X}'_{\mathcal{M}}\Lambda_{\mathcal{M}}\mathbf{X}_{\mathcal{M}}\right\}\mathrm{d}x_{\mathcal{M}_1}\dots\mathrm{d}x_{\mathcal{M}_m}},\end{aligned}$$

which is uniformly continuous in ψ.

The result follows from Theorem 3.4.3 in Guyon [1995]. □

3.3.3 Asymptotic Distributions of QLE

For notation purpose let us deinfe

$$\begin{aligned} g_i(\hat{\psi}_n) &= \left(\ln P_{\hat{\psi}_n}(y_i|y_{\mathcal{N}_i})\right)^{(1)}, \\ H_i(\hat{\psi}_n) &= \left(\ln P_{\hat{\psi}_n}(y_i|y_{\mathcal{N}_i})\right)^{(2)}, \end{aligned}$$

where (1) and (2) represent the vector of first derivatives, and the matrix of the second derivatives over $\hat{\psi}_n$ respectively. We can define *quasi-information* matrices:

$$\begin{aligned} J_{\hat{\psi}_n} &= \frac{1}{|D_n|} \sum_{i \in D_n} \mathrm{Var}_{\psi_0}\left[g_i(\hat{\psi}_n)\right] \\ I_{\hat{\psi}_n} &= -\frac{1}{|D_n|} \sum_{i \in D_n} \mathrm{E}_{\psi_0}\left[H_i(\hat{\psi}_n)\right], \end{aligned}$$

Following the fact that $\{Y_i : i \in S\}$ is ergodic,

$$\begin{aligned} \lim_n J_{\psi_n} &= J_{\psi_0}, \\ \lim_n I_{\psi_n} &= I_{\psi_0}, \end{aligned}$$

were J_{ψ_0} and I_{ψ_0} are positive definite matrixes.

Theorem 3. *If $D_n \nearrow S$ and $\hat{\psi}_n$ is a QLE estimator of a latent variable model for spatial process with ordinal data as described above, then*

$$|D_n|^{1/2}(\hat{\psi}_n - \psi_0) \xrightarrow{\mathcal{D}} N_p(0, I_{\psi_0} J_{\psi_0}^{-1} I_{\psi_0}) \tag{3.6}$$

Proof. We may note that

$$|D_n|^{-1/2} J_{\hat{\psi}_n}^{\frac{1}{2}} g_i(\hat{\psi}_n) \xrightarrow{\mathcal{D}} N_p(0, I),$$

which can be proved analogous to Theorem 5.3.2 in Guyon [1995] without the additional difficulty of having to examine the complimentary term R_n. Direct application of Theorem 3.4.5 in Guyon [1995] yields result. □

3.4 Simulation Results

We have discussed three estimators for a latent variable model for spatial process with ordinal data. As a result, we have conducted a simulation study for the comparison. The speed of QLE estimator depends on the neighborhood size. It is fast for time series data (1 point neighborhood).As a result, we will first compare *QLE* with *MdA* for time series data for data consisted of 100 points. Results can be found in Figures 3.2-3.4. These figures compare MLE_X based on unobserved, but simulated X, *QLE* and *MdA*. Figure 3.2 is based on 1000 simulation of time series of length 100 with cut off points $\theta = \left(\Phi^{-1}\left(\frac{1}{3}\right), \Phi^{-1}\left(\frac{2}{3}\right)\right)$ and correlation coefficient $\phi = \frac{1}{3}$. Comparing means of the estimates, there is a small bias is introduced by *QLE* and *MdA* comparing with MLE_X. Variance doubles when we compare *QLE* and *MdA* with MLE_X. However, there is no increase in variance between *QLE* and *MdA*, the reason for this is probably are due to numerical approximations. Calculation of *QLE* involves bivariate normal CDF, which have larger computational errors than univariate CDF required by *MdA* calculations. Figure 3.3 shows result of similar simulations, but for correlation coefficient $\phi = -\frac{1}{3}$. We can observe symmetry with case $\phi = \frac{1}{3}$. Figure 3.4 shows result of similar simulations, but for correlation coefficient $\phi = 0$. Here *QLE* performs worse. The reason will be studied. In summary we can see that there are no significant losses observed between QLE and *MdA*.

Next, we will compare *MdA* with *MnA*. We will use simulated data similar to one discussed above, but on squares of size 20×20 and 40×40. Results can be found in Figures 3.5-3.9. We can see that there is no significant difference observed between *MnA* and *MdA*. In these examples $MLE(X)$ represents the MLE from the unobserved (but simulated) X variable. *MnA* and *MdA* perform reasonably well comparing to $MLE(X)$.

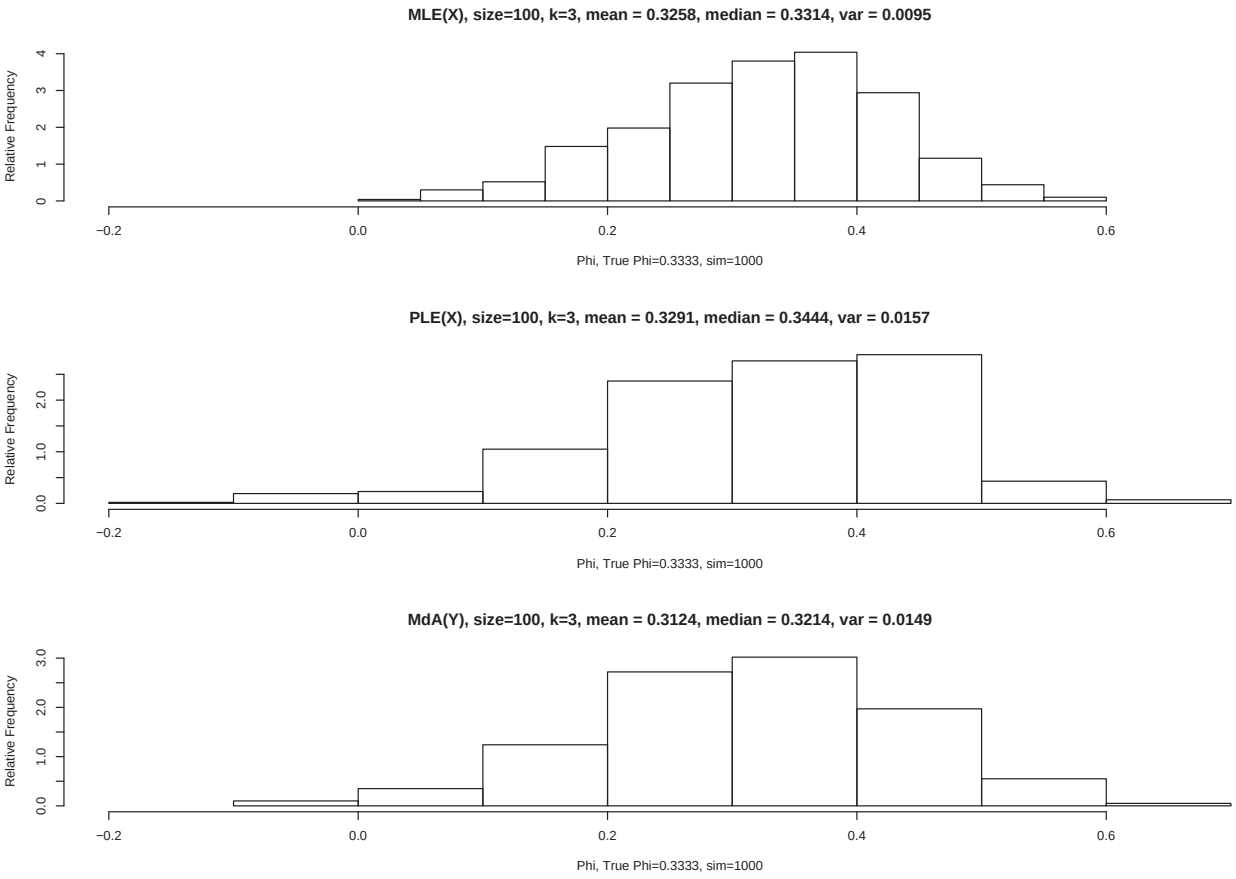

Figure 3.2: Simulation result for QLE and *MdA* comparison for time series data of 100 observation and three category with cut off points $\theta = \left(\Phi^{-1}\left(\frac{1}{3}\right), \Phi^{-1}\left(\frac{2}{3}\right)\right)$ and correlation coefficient $\phi = \frac{1}{3}$ based on 1000 simulations.

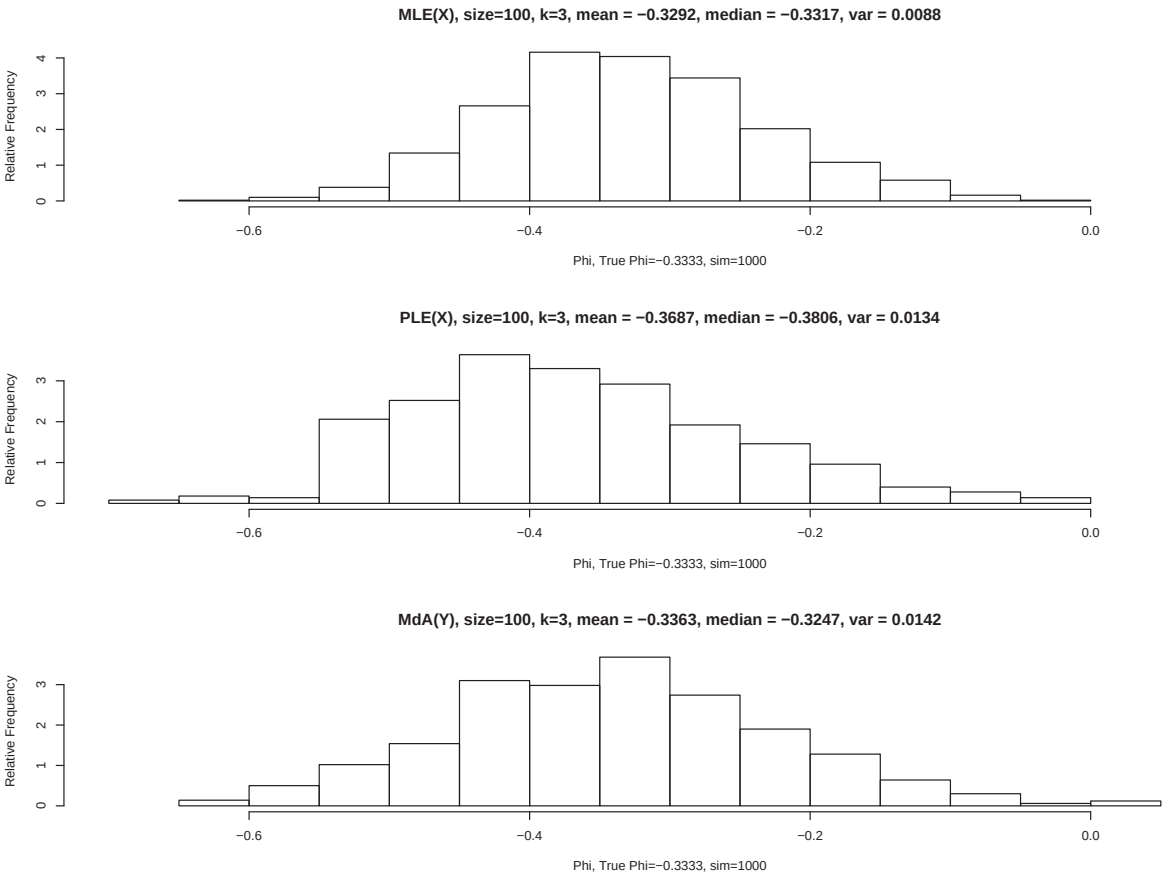

Figure 3.3: Simulation result for QLE and *MdA* comparison for time series data of 100 observation and three category, with cut off points $\theta = \left(\Phi^{-1}\left(\frac{1}{3}\right), \Phi^{-1}\left(\frac{2}{3}\right)\right)$ and correlation coefficient $\phi = -\frac{1}{3}$ based on 1000 simulations.

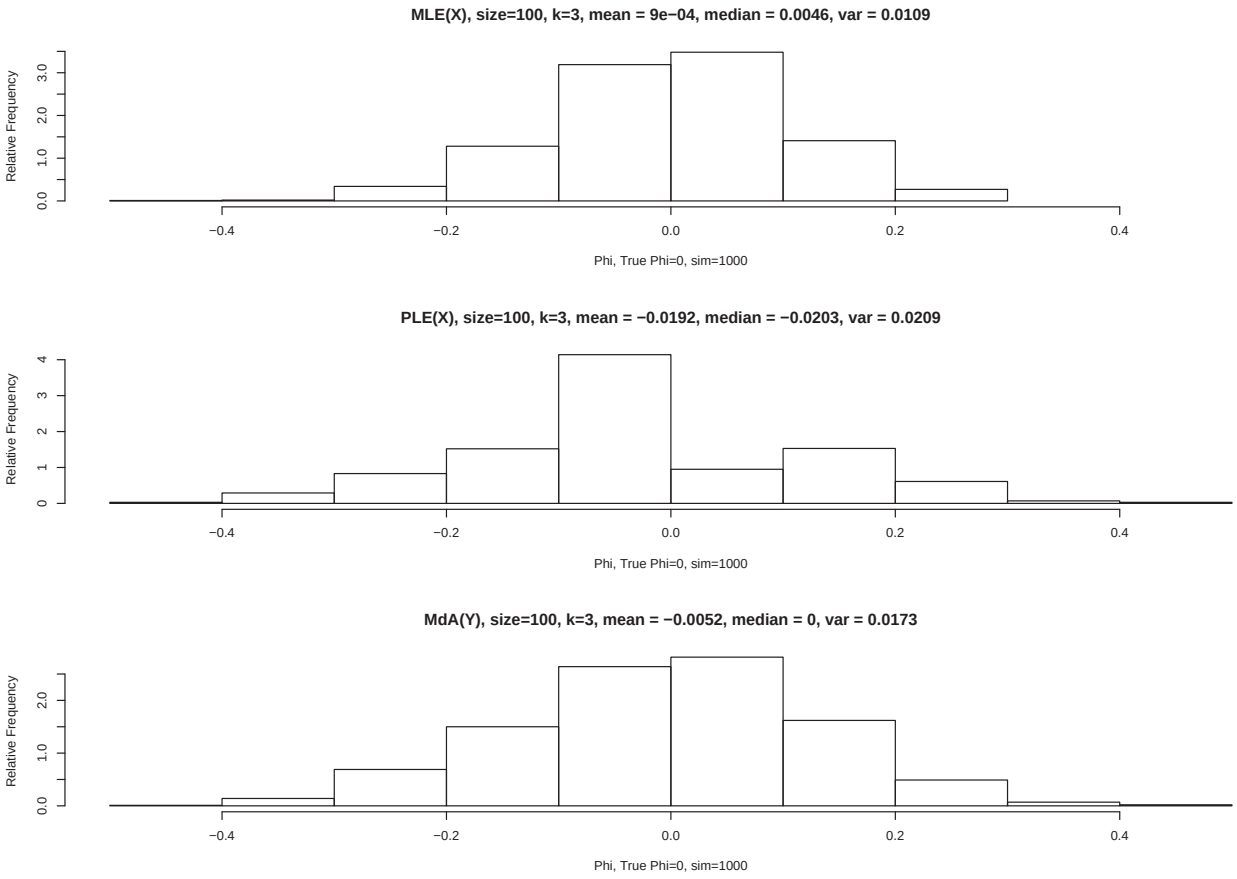

Figure 3.4: Simulation result for QLE and *MdA* comparison for time series data of 100 observation and three category, with cut off points $\theta = \left(\Phi^{-1}\left(\frac{1}{3}\right), \Phi^{-1}\left(\frac{2}{3}\right)\right)$ and correlation coefficient $\phi = 0$ (i.e. no spatial dependence), based on 1000 simulations.

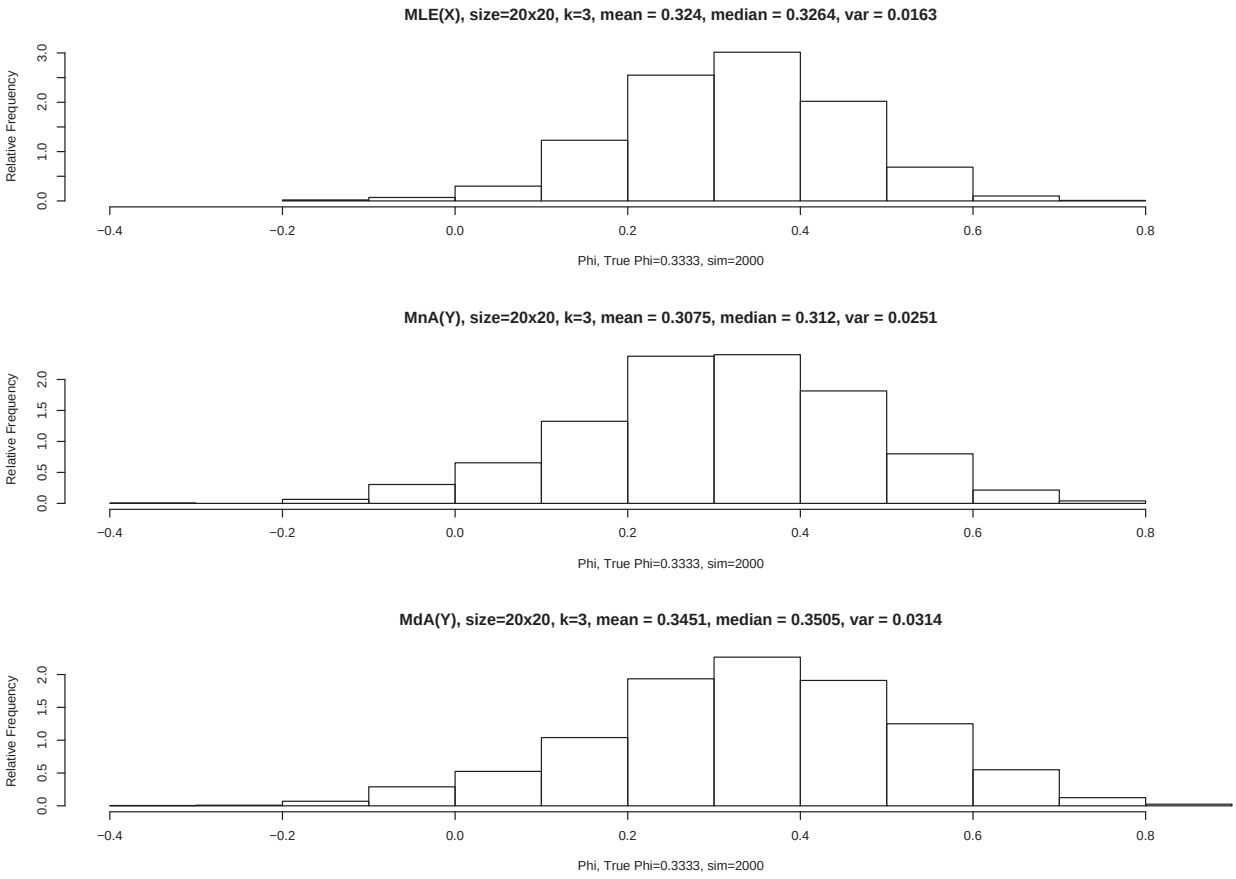

Figure 3.5: Simulation result for *MnA* and *MdA* comparison for time series data of 20×20 observation and three category, with cut off points $\theta = \left(\Phi^{-1}\left(\frac{1}{3}\right), \Phi^{-1}\left(\frac{2}{3}\right)\right)$ and correlation coefficient $\phi = \frac{1}{3}$, based on 2000 simulations.

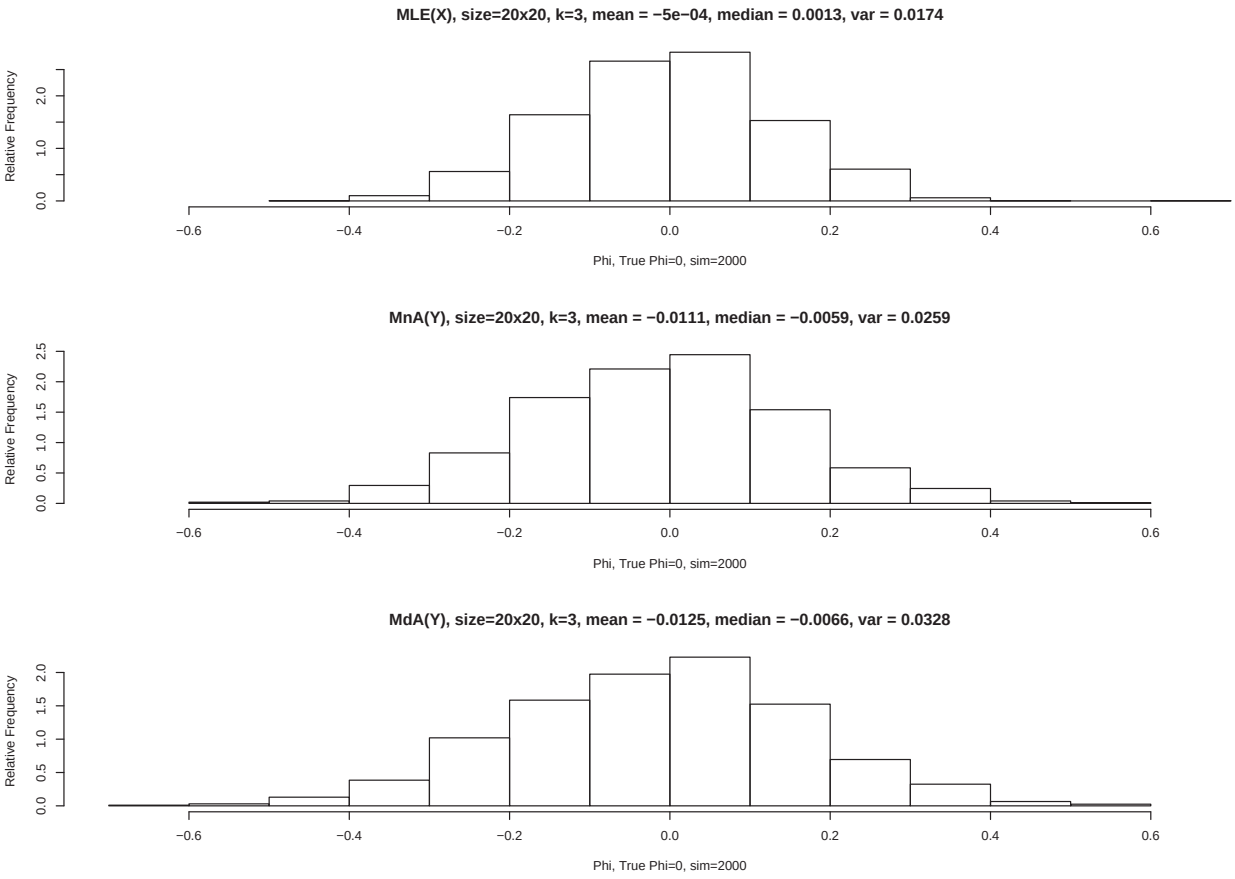

Figure 3.6: Simulation result for *MnA* and *MdA* comparison for time series data of 20×20 observation and three category, with cut off points $\theta = \left(\Phi^{-1}\left(\frac{1}{3}\right), \Phi^{-1}\left(\frac{2}{3}\right)\right)$ and correlation coefficient $\phi = 0$, based on 2000 simulations.

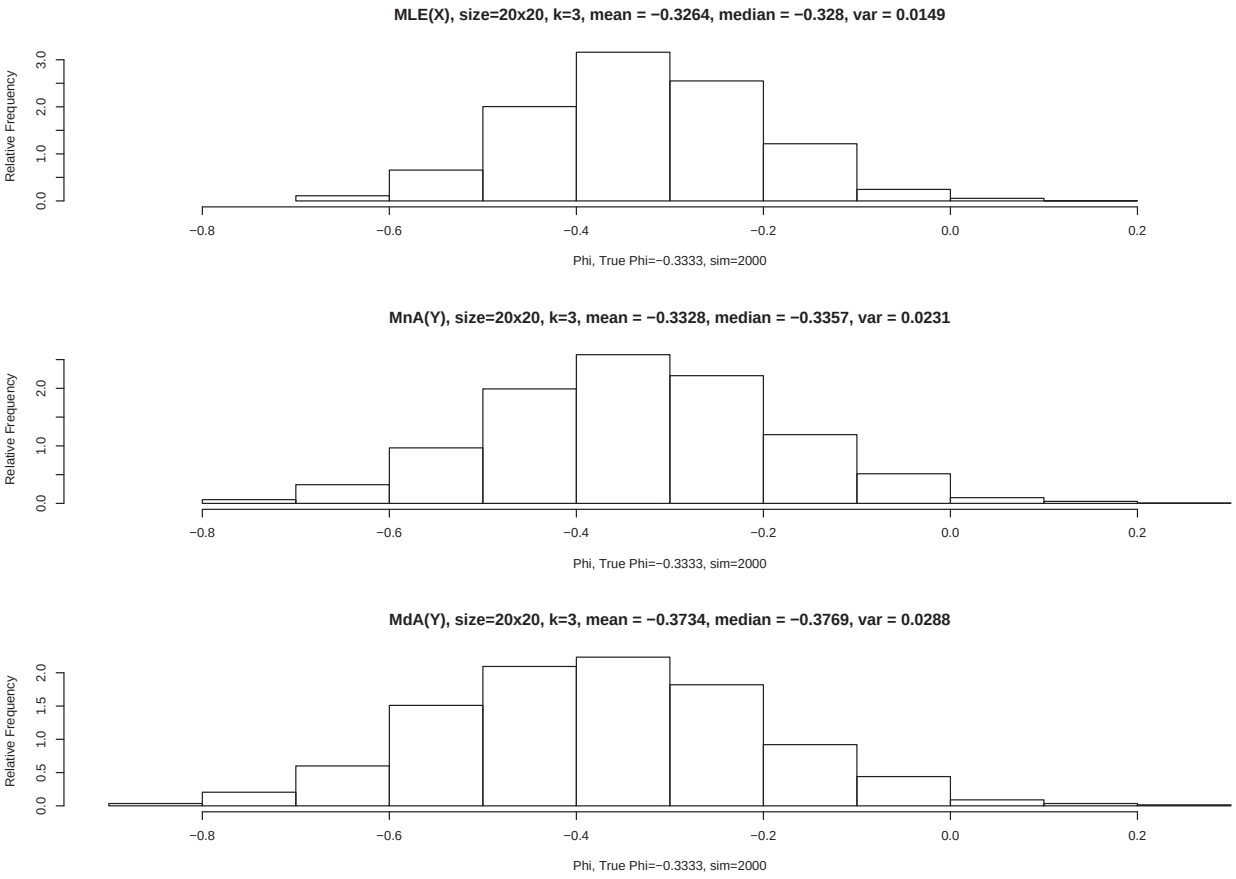

Figure 3.7: Simulation result for *MnA* and *MdA* comparison for time series data of 20×20 observation and three category, with cut off points $\theta = \left(\Phi^{-1}\left(\frac{1}{3}\right), \Phi^{-1}\left(\frac{2}{3}\right)\right)$ and correlation coefficient $\phi = -\frac{1}{3}$, based on 2000 simulations.

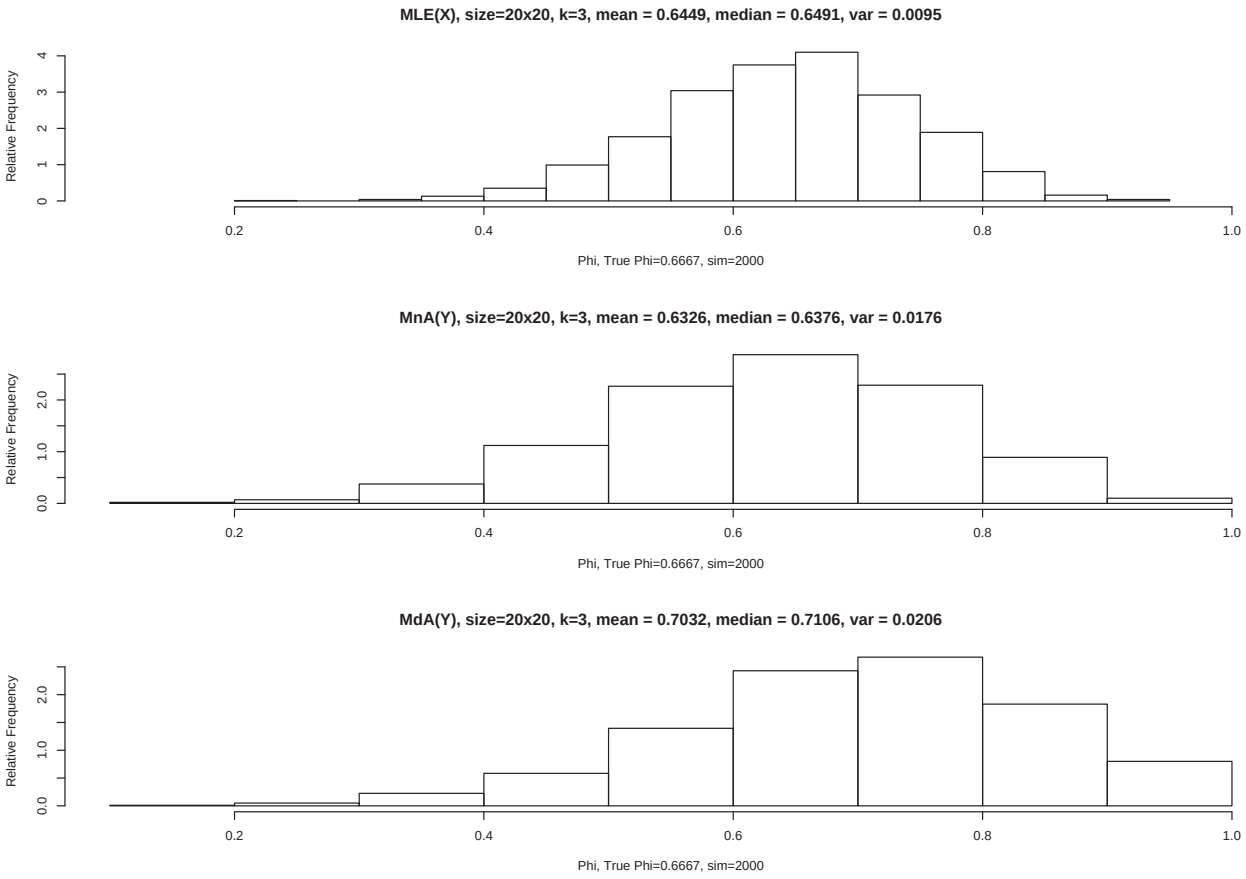

Figure 3.8: Simulation result for *MnA* and *MdA* comparison for time series data of 20×20 observation and three category, with cut off points $\theta = \left(\Phi^{-1}\left(\frac{1}{3}\right), \Phi^{-1}\left(\frac{2}{3}\right)\right)$ and correlation coefficient $\phi = \frac{1}{6}$, based on 2000 simulations.

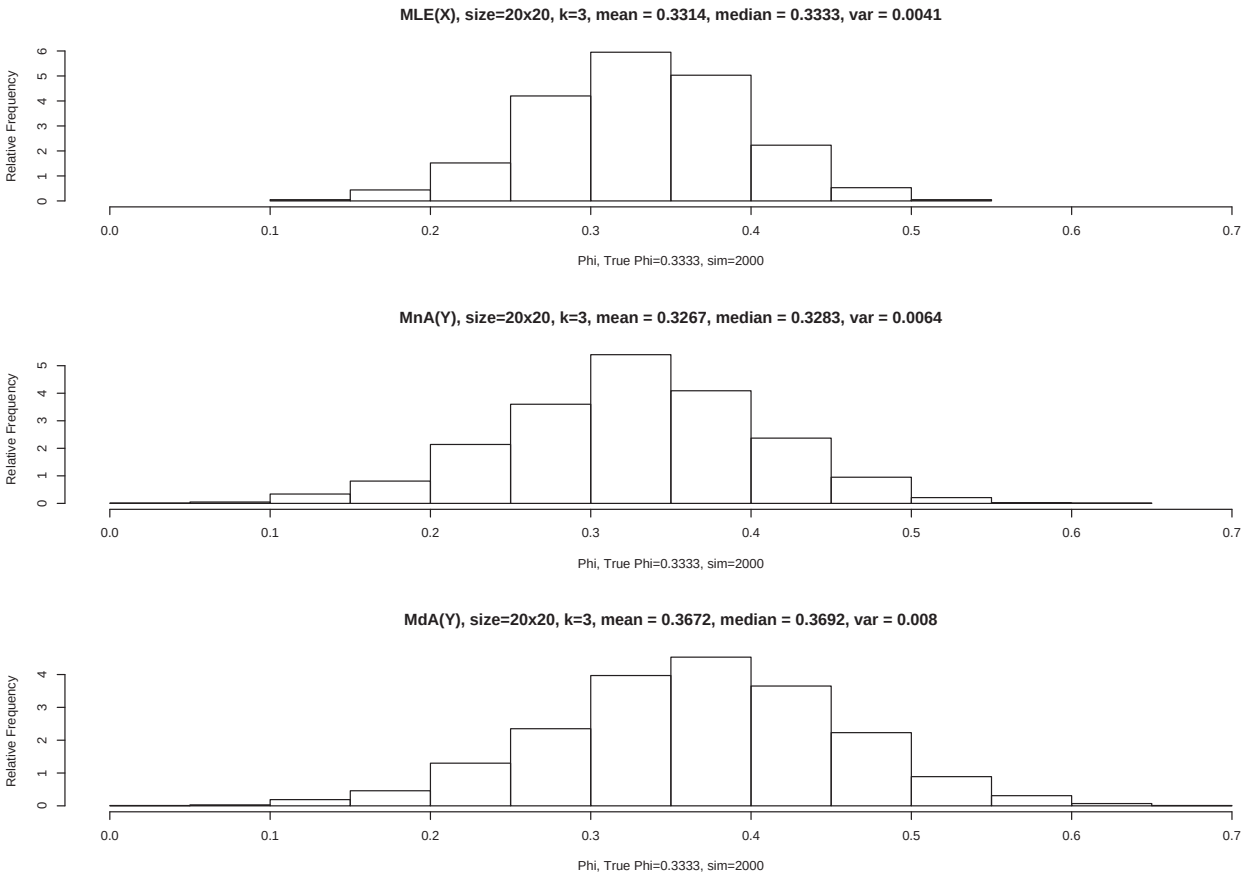

Figure 3.9: Simulation result for *MnA* and *MdA* comparison for time series data of 40×40 observation and three category, with cut off points $\theta = \left(\Phi^{-1}\left(\frac{1}{3}\right), \Phi^{-1}\left(\frac{2}{3}\right)\right)$ and correlation coefficient $\phi = \frac{1}{3}$, based on 2000 simulations.

Chapter 4

Technical Details

4.1 Technical Details for Gibbs Random Fields (GRF)

The Gibbs field (or measure) dates back to Dobrushin and Folguera [1968] and it is highly used in Statistical Physics. In probabilistic term, a Gibbs measure is nothing other than the distribution of conditional probabilities. We will look at two specifications of Gibbs Fields.

4.1.1 Conditional Specification

Definition (Conditional Kernel). If μ is a random field, for each $V \in \mathcal{S}$, we can define the *conditional kernel*

$$\mu_V(\cdot|\cdot) : \mathcal{F}(V) \times \Omega(S\backslash V) \to [0,1]. \tag{4.1}$$

Here $\mathcal{F}(V)$ stands for σ algebra of configuration over V, and $\Omega(S\backslash V)$ is the space of V external configuration. In general this kernel is not well defined.

Definition (Coherence of Kernels). Let $V \subseteq U$ be two elements of $\mathcal{S}$, and let π_V and π_U be two conditional kernels relative to V and U, $A \in \mathcal{F}(V)$, $B \in \mathcal{F}(U\backslash V)$ and $z \in \Omega(S\backslash U)$. Composition of both kernels is another kernel over U defined by

$$(\pi_U \pi_V)(AB|z) = \int_B \pi_V(A|yz)\pi_U(E^V, dy|z). \tag{4.2}$$

For the family $\{\mu_V\}$ defined by the field μ, one has

$$\mu_U \mu_V = \mu_U \tag{4.3}$$

We say this family is coherent.

Definition (Conditional Specification). Any family of kernels $\pi = \{\pi_V, V \in \mathcal{S}\}$ is called conditional specification if it satisfies condition (4.3).

4.1.2 Gibbs Specification

Definition (Interaction Potential). Interaction potential is defined by a family $\phi = \{\phi_A, A \in \mathcal{S}\}$ of applications

$$\phi_A : \Omega(A) \to \mathbb{R} \qquad \text{s.t.} \tag{4.4}$$

(i) For every A, ϕ_A is $\mathcal{F}(A)$ measurable

(ii) If $\Lambda \in \mathcal{S}$ and $w \in \Omega$ then the sum

$$U_\Lambda^\phi(w) = \sum_{A \in \mathcal{S}: A \cap \Lambda \neq 0} \phi_A(w) \qquad \text{exists.} \tag{4.5}$$

$-U_\Lambda^\phi(w)$ is called the energy of w in Λ. ϕ is $\lambda-$admissible if for all $\Lambda \in \mathcal{S}$, $w \in \Omega$,

$$Z_\Lambda^\phi(w) = \int_{\Omega(\Lambda)} \exp U_\Lambda^\phi(w_\Lambda, w_{S\setminus\Lambda}) \lambda^\Lambda(dw_\lambda) < \infty. \tag{4.6}$$

Definition (Gibbs Specification Associated to a Potential ϕ). If ϕ is admissible, the family

$$\pi_\Lambda^\phi(w_\Lambda | w_{S\setminus\Lambda}) = Z_\Lambda^\phi(w) \exp U_\Lambda^\phi(w_\Lambda, w_{S\setminus\Lambda}), \Lambda \in \mathcal{S} \tag{4.7}$$

is coherent. $\{\pi_V^\phi, V \in \mathcal{S}\}$ is called the Gibbs specification associated to a potential ϕ.

The proof of coherence can be found in Guyon [1995]

4.1.3 Unicity: Dobrushin's Condition of Weak Dependence

Definition (Total Variation Norm). Let $(E, \mathcal{E})$ be a measurable space and
$h, g \in \mathcal{P}(E, \mathcal{E})$. Then the total variation norm is defined as:

$$\|h - g\| = \sup_{A \in \mathcal{E}} |h(A) - g(A)|$$

Definition (Dobrushin's Influence Measure). Let a and b be two sites of S, $a \neq b$. Define for a specification π

$$\gamma_{a,b}(\pi) = \sup \frac{1}{2} \|\pi_b(\cdot|w) - \pi_b(\cdot|w')\|. \tag{4.8}$$

Here the total variation is taken over all configurations w, w' that are identical except at site a. $\gamma_{a,b}(\pi)$ is a measure of the influence of a site a over the conditional distribution $\pi_b(\cdot|\cdot)$ in b

Condition (Dobrushin's Condition). Gibbs fields satisfy Dobrushin's condition if each potential is quasi-local and if

$$\alpha(\phi) = \sup_{a \in S} \sum_{b \in S} \gamma_{a,b}(\phi) < 1. \tag{4.9}$$

Dobrushin's condition guarantees unicity of the Gibbs state (Georgii [1988]). However, it is helpful to note that Dobrushin's condition is only a sufficient and not a necessary condition. In practice, most of the useful GRF satisfies this condition, which also defines weak dependence necessary for asymptotic convergence of estimators.

4.1.4 Reconstruction of a Specification π Given By Its Specification $\pi_{\{i\}}$ for Each Site $i \in S$

In practice, a random field is often specified by a conditional distribution of a site i, given all other sites. A question arises whether this specification will yield a joint distribution. Guyon [1995] provides a Theorem 2.5.1, stated below:

Theorem (Guyon [1995]). Next two conditions are true:

(a) Assume that for all $x \in \Omega$, $\pi(x) > 0$. Then π is only determined by its conditional distributions π_i, $i \in S$.

(b) A family of conditional distribution $\{\pi_i, i \in S\}$ in general does not induce a joint distribution.

More general result also can be found in Georgii [1988]

4.2 Technical Details for HMRF

4.2.1 HMRF is GRF

Lemma 1. *Let X be a MRF over S. Let Y be HMRF. Then Y is also a Gibbs field.*

Proof. Joint distribution of Y is defined by uniquely $\pi(Y) = \int \prod_{i \in S} \pi(y_i|x_i, \theta)\pi(X, \psi)\, \mathrm{d}X$, which makes Y a GRF by definition. However, potentials are usually numerically untraceable. □

4.2.2 Dobrushin Condition for HMRF

Lemma 2. *Assume X is a MRF then $Y : \pi(y|x) = \prod_{i \in S} \pi(y_i|x_i)$ satisfies the Dobrushin condition (4.9). In addition the joint GRF $\{X, Y\}$ also satisfies Dobrushin condition (4.9).*

Proof. The intuition is that the Dobrushin condition defines weak dependence and dependence among Y is defined through X, so it is weaker than dependence along X. As a result, Y will satisfy any dependence condition for X.

First, we will prove that $\{Y, X\}$ satisfies (4.9). Let us look at Dobrushin

measure:

$$\begin{aligned}
\gamma_{\{i,j\}}^{\{Y,X\}} &= \sup_{x,y} \frac{1}{2} \| \pi(x_i, y_i | x_j, y_j, x_{-\{i,j\}}, y_{-\{i,j\}}) \\
&\qquad -\pi(x_i, y_i | x'_j, y'_j, x_{-\{i,j\}}, y_{-\{i,j\}}) \| \\
&= \sup_{x,y} \frac{1}{2} \| \pi(y_i | x_i, x_j, y_j, x_{-\{i,j\}}, y_{-\{i,j\}}) \pi(x_i | x_j, y_j, x_{-\{i,j\}}, y_{-\{i,j\}}) \\
&\qquad -\pi(y_i | x_i, x'_j, y'_j, x_{-\{i,j\}}, y_{-\{i,j\}}) \pi(x_i | x'_j, y'_j, x_{-\{i,j\}}, y_{-\{i,j\}}) \| \\
&= \sup_{x,y} \frac{1}{2} \| \pi(y_i | x_i) \pi(x_i | x_j \cup x_{\partial i}) - \pi(y_i | x_i) \pi(x_i | x'_j \cup x_{\partial i}) \| \\
&\leq \sup_{x} \frac{1}{2} \| \pi(x_i | x_j \cup x_{\partial i}) - \pi(x_i | x'_j \cup x_{\partial i}) \| = \gamma_{\{i,j\}}^{\{X\}},
\end{aligned}$$

then by definition of the Dobrushin condition (4.9),

$$\alpha^{\{X,Y\}}(\phi) = \sup_{a \in S} \sum_{b \in S} \gamma_{a,b}^{\{X,Y\}}(\phi) \leq \sup_{a \in S} \sum_{b \in S} \gamma_{a,b}^{\{X\}}(\phi) = \alpha^{\{X,Y\}}(\phi)$$

Now we will prove that $\{Y\}$ satisfies (4.9). Again looking at the Dobrushin

measure:

$$\begin{aligned}
\gamma^{\{Y\}}_{\{i,j\}} &= \sup_y \frac{1}{2}\|\pi(y_i|y_j,y_{-\{i,j\}}) - \pi(y_i|y'_j,y_{-\{i,j\}})\| \\
&= \sup_y \frac{1}{2}\Bigg\| \int \pi(y_i|y_j,y_{-\{i,j\}},x)\pi(x|y_j,y_{-\{i,j\}})\,\mathrm{d}x \\
&\qquad - \int \pi(y_i|y'_j,y_{-\{i,j\}},x)\pi(x|y'_j,y_{-\{i,j\}})\,\mathrm{d}x\Bigg\| \\
&= \sup_y \frac{1}{2}\Bigg\| \int \pi(y_i|x_i)\pi(y_j,y_{-\{i,j\}},x)\pi(y_j,y_{-\{i,j\}})\,\mathrm{d}x \\
&\qquad - \int \pi(y_i|x_i)\pi(y'_j,y_{-\{i,j\}},x)\pi(y'_j,y_{-\{i,j\}})\,\mathrm{d}x\Bigg\| \\
&= \sup_y \frac{1}{2}\Bigg\| \int \pi(y_i|x_i)\pi(y_j|x_j)\prod_{k\neq i,j}\pi(y_k|x_k)\pi(y_j,y_{-\{i,j\}})\pi(x)\,\mathrm{d}x \\
&\qquad - \int \pi(y_i|x_i)\pi(y'_j|x_j)\prod_{k\neq i,j}\pi(y_k|x_k)\pi(y'_j,y_{-\{i,j\}})\pi(x)\,\mathrm{d}x\Bigg\| \\
&= \sup_y \frac{1}{2}\Bigg\| \int \left[\pi(y_j|x_j)\pi(y_j,y_{-\{i,j\}}) - \pi(y'_j|x_j)\pi(y'_j,y_{-\{i,j\}})\right]\times \\
&\qquad \times\pi(y_i|x_i)\prod_{k\neq i,j}\pi(y_k|x_k)\pi(x)\,\mathrm{d}x\Bigg\| \\
&\leq \sup_y \frac{1}{2}\Bigg\| \int \pi(y_j|x_j)\pi(y_i|x_i)\prod_{k\neq i,j}\pi(y_k|x_k)\pi(x)\,\mathrm{d}x\Bigg\| \\
&= \sup_y \frac{1}{2}\Bigg\| \int \pi(y|x)\pi(x)\,\mathrm{d}x\Bigg\| \\
&\leq \sup_{x,y} \frac{1}{2}\|\pi(x,y)\| = \gamma^{\{Y,X\}}_{\{i,j\}} \leq \gamma^{\{X\}}_{\{i,j\}}
\end{aligned}$$

And finally,

$$\alpha^{\{Y\}}(\phi) = \sup_{a\in S}\sum_{b\in S}\gamma^{\{Y\}}_{a,b}(\phi) \leq \sup_{a\in S}\sum_{b\in S}\gamma^{\{X\}}_{a,b}(\phi) = \alpha^{\{X\}}(\phi)$$

□

4.2.3 Simon's Condition for HMRF

The Dobrushin condition is not always easy to obtain. A stricter condition, based directly on the potentials was given by Simon [1979]. If λ is a finite measure over $(E, \mathcal{E})$, and ϕ is a continuous λ-admissible potential such that:

$$\sup_{i \in S} \sum A \ni i (|A| - 1) \delta(\phi_A) < 2$$

then the Dobrushin's condition is satisfied.

Lemma 3. *Assume X is MRF then $Y : \pi(y|x) = \prod_{i \in S} \pi(y_i|x_i)$ satisfies Simon's condition. In addition the joint GRF $\{X, Y\}$ also satisfies Simon's condition.*

Proof. Proof is identical to the proof of the **Lemma 2**. □

Chapter 5

Summary and Future Research

This section describes some of the areas I plan to explore as part of future research.

I have introduced several methods of inference for the ordinal data under the latent variable method. The consistency and asymptotic normality of the quasi-likelihood estimator has been established. Asymptotic properties of the other methods have to be studied. Limited simulation results show that the mean and median-based approximations have reasonably good efficiency in addition to being computationally fast. A more extensive study will be conducted to compare the properties of these methods and also of other quasi-likeliehood based methods.

An alternative approach to analyzing ordinal data is to use Strauss' model for categorial (mulitcolor) spatial data with suitable constraints on the dependence coefficients β_{kl}'s between adjacent categories. I will compare the advantages and disadvantages of this approach with that based on the latent variable model.

As noted earlier, in many applications, the mean structure of the latent variable can vary according to some covariates. It is then of interest to model and make inferences for the regression relationship based on the ordinal data. One application is the study of spatial patterns in wafermaps in semiconductor manufacturing. Methods for estimation and tests of hypothesis will be developed. Since likelihood ratio procedures are not computationally tractable, we will study the use of quasi-likelihood and other approximations.

Finally, Bayesian inference using data augmentation is a natural alternative in this problem. One can consider the unboserved latent varaible as missing data. I have already made some progress in this direction.

References

J. Besag. Spatial interaction and the statistical analysis of lattice systems (with discussion). *Journal of the Royal Statistical Society, Series B, Methodological*, 36:192–236, 1974.

J. Besag. Statistical analysis of non-lattice data. *The Statistician*, 24:179–196, 1975.

P. Billingsley. *Probability and Measure (Third Edition)*. Wiley, 1995.

D. Brook. On the distinction between the conditional probability and the joint probability approaches in the specification of nearest-neighbour systems. *Biometrika*, 51:481–483, 1964.

N. A. C. Cressie. *Statistics for spatial data*. New York : J. Wiley, revised edition, 1993.

R. L. A. Dobrushin and H. C. T. Folguera. The description of a random field by means of conditional probabilities and conditions of its regularity. *Theory of Probability and its Applications (Transl of Teorija Verojatnostei i ee Primenenija)*, 13:197–224, 1968.

T. G. Donnelly. Algorithm 462: Bivariate normal distribution. *Communications of the ACM*, 16:638, 1973.

Z. Drezner. Approximations to the multivariate normal integral. *Communications in Statistics, Part B – Simulation and Computation*, 19:527–534, 1990.

D. Geman. *Random Fields and Inverse Problem in Imaging*, volume 1427 of *Lecture Notes in Statistics*. Springer, 1988.

S. Geman and D. Geman. Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 6:721–741, 1984.

A. Genz. Numerical computation of multivariate normal probabilities. *Journal of Computational and Graphical Statistics*, 1:141–149, 1992.

H.-O. Georgii. *Gibbs Measures and Phase Transitions*. Walter de Gruyter, 1988.

X. Guyon. *Random Fields on a Network: Modelling, Statistics, and Applications*. Springer-Verlag, 1995.

E. Ising. Beitrag zur theorie des ferromagnetismus. *Zeitschrift ür Physik*, 31: 253–258, 1925.

N. L. Johnson, S. Kotz, and N. Balakrishnan. *Continuous Univariate Distributions*, volume 1. Wiley-Interscience, second edition, 1994.

V. E. Johnson and J. H. Albert. *Ordinal Data Modeling*. Springer , New York, 1999.

H. Künsch, S. Geman, and A. Kehagias. Hidden Markov random fields. *Annals of Applied Probability*, 5:577–602, 1995.

R. Mead. A mathematical model for the estimation of inter-plant competition. *Biometrics*, 23:189–205, 1967.

K. Ord. Estimation methods for models of spatial interaction. *Journal of the American Statistical Association*, 70:120–126, 1975.

D. Ruelle. *Statistical mechanics: Rigorous results.* W. A. Benjamin, Inc., New York-Amsterdam, 1969.

M. J. Schervish. [algorithm As 195] Multivariate normal probabilities with error bound (corr: 85v34 p103-104). *Applied Statistics*, 33:81–94, 1984.

B. Simon. A remark on dobrushin's uniqueness theorem. *Comm. Math. Phys.*, 68:183–185, 1979.

D. Strauss and M. Ikeda. Pseudolikelihood estimation for social networks. *Journal of the American Statistical Association*, 85:204–212, 1990.

D. J. Strauss. Clustering on coloured lattices. *Journal of Applied Probability*, 14:135–143, 1977.

M. A. Tanner. *Tools for Statistical Inference. Methods for the Exploration of Posterior Distributions and Likelihood Functions (Third Edition).* Springer-Verlag, 1996.

C. Wang. *Modeling Temporally Dependent Ordinal Processes.* PhD thesis, The Univeristy of Michigan, 1999.

P. Whittle. On stationary processes in the plane. *Biometrika*, 41:434–449, 1954.