1997

Specification of dependence structures and simulation-based estimation for conditionally specified statistical models

Jaehyung Lee

Iowa State University

Follow this and additional works at: https://lib.dr.iastate.edu/rtd

Part of the Statistics and Probability Commons

Recommended Citation

https://lib.dr.iastate.edu/rtd/11477

This Dissertation is brought to you for free and open access by the Iowa State University Capstones, Theses and Dissertations at Iowa State University Digital Repository. It has been accepted for inclusion in Retrospective Theses and Dissertations by an authorized administrator of Iowa State University Digital Repository. For more information, please contact digirep@iastate.edu.
This manuscript has been reproduced from the microfilm master. UMI films the text directly from the original or copy submitted. Thus, some thesis and dissertation copies are in typewriter face, while others may be from any type of computer printer.

The quality of this reproduction is dependent upon the quality of the copy submitted. Broken or indistinct print, colored or poor quality illustrations and photographs, print bleedthrough, substandard margins, and improper alignment can adversely affect reproduction.

In the unlikely event that the author did not send UMI a complete manuscript and there are missing pages, these will be noted. Also, if unauthorized copyright material had to be removed, a note will indicate the deletion.

Oversize materials (e.g., maps, drawings, charts) are reproduced by sectioning the original, beginning at the upper left-hand corner and continuing from left to right in equal sections with small overlaps. Each original is also photographed in one exposure and is included in reduced form at the back of the book.

Photographs included in the original manuscript have been reproduced xerographically in this copy. Higher quality 6” x 9” black and white photographic prints are available for any photographs or illustrations appearing in this copy for an additional charge. Contact UMI directly to order.
Specification of dependence structures and simulation-based estimation for conditionally specified statistical models

by

Jaehyung Lee

A dissertation submitted to the graduate faculty
in partial fulfillment of the requirements for the degree of
DOCTOR OF PHILOSOPHY

Major: Statistics
Major Professors: Mark Kaiser and Noel Cressie

Iowa State University
Ames, Iowa
1997
This is to certify that the Doctoral dissertation of
Jaehyung Lee
has met the dissertation requirements of Iowa State University

Signature was redacted for privacy.
Co-major Professor
Signature was redacted for privacy.
Co-major Professor
Signature was redacted for privacy.
For the Major Program
Signature was redacted for privacy.
For the Graduate College
Dedicated to my wife, Mikyung,
my kids, Inyoung and Dongho,
and my parents.
# TABLE OF CONTENTS

## CHAPTER 1 GENERAL INTRODUCTION
1.1 Introduction ........................................................................................................ 1  
1.2 Dissertation Organization .................................................................................. 3

## CHAPTER 2 LITERATURE REVIEW
2.1 General Distribution Theory ............................................................................. 4  
2.2 Conditional Specifications Based on Markov Random Fields ......................... 5  
2.3 Estimation in Conditionally Specified Models .................................................. 6

## CHAPTER 3 MULTIWAY DEPENDENCE IN EXPONENTIAL FAMILY
  CONDITIONAL DISTRIBUTIONS ........................................................................... 8
  Abstract .................................................................................................................. 8  
  3.1 Introduction ....................................................................................................... 8  
  3.2 Model Formulation from MRF Structure .......................................................... 10  
  3.3 Conditional Exponential Family Distributions .................................................. 12  
    3.3.1 Pairwise-Only Dependence ....................................................................... 12  
    3.3.2 Multiway Dependence ............................................................................. 13  
  3.4 Examples ........................................................................................................... 17  
  3.5 Discussion ......................................................................................................... 19  
  Acknowledgment .................................................................................................... 19

## CHAPTER 4 MAXIMUM LIKELIHOOD ESTIMATION IN CONDITIONALLY
  SPECIFIED STATISTICAL MODELS ...................................................................... 20
  Abstract .................................................................................................................. 20  
  4.1 Introduction ....................................................................................................... 20  
  4.2 Conditional Model Specification for Markov Random Fields ......................... 23  
  4.3 The Overall Estimation Strategy .................................................................... 24
4.4 Importance Sampling .................................................. 27
  4.4.1 Selection of a Sampling Distribution ......................... 27
  4.4.2 Precision in Monte Carlo Log Likelihood .................... 28
  4.4.3 Assessing Convergence in Practice .......................... 31
  4.4.4 Determination of Baseline Monte Carlo Error ............... 32
4.5 Starting Value ....................................................... 32
4.6 Example Using a Winsorized Poisson Model ..................... 34
4.7 Concluding Remarks ................................................ 40
Acknowledgment .......................................................... 42

CHAPTER 5 MAXIMUM LIKELIHOOD ESTIMATION IN CONDITIONALLY
SPECIFIED STATISTICAL MIXTURE MODELS ......................... 43
5.1 Introduction ......................................................... 43
5.2 Conditionally Specified Mixture Models ........................ 46
  5.2.1 Definition ..................................................... 46
  5.2.2 Some Useful Classes of Models .............................. 48
  5.2.3 Properties ..................................................... 49
5.3 The Overall Estimation Strategy .................................. 50
5.4 Importance Sampling ................................................. 53
  5.4.1 Selection of Sampling Distributions .......................... 53
  5.4.2 Precision in Monte Carlo Log Likelihood .................... 55
  5.4.3 Assessing Convergence in Practice .......................... 58
  5.4.4 Determination of Baseline Monte Carlo Error ............... 58
5.5 Starting Value ....................................................... 59
5.6 Examples .......................................................... 61
  5.6.1 Data Description ............................................... 61
  5.6.2 Modeling the Birch Tree Data ................................ 62
  5.6.3 Selection of Importance Densities ........................... 65
  5.6.4 Estimation Process ............................................ 66
  5.6.5 Further Analysis .............................................. 67
5.7 Discussion .......................................................... 70
LIST OF TABLES

Table 4.1 Observed counts of sweet birch trees for sampling locations of Figure 4.1 ........................................ 34
Table 4.2 Neighborhoods defined for sampling locations of Figure 4.1 using various values of the threshold distance $\kappa$ ................................................................. 37
Table 4.3 Maximum likelihood estimation summary for the neighborhoods of Table 4.2, defined with various values of $\kappa$ .................................................................................. 38
Table 4.4 Detailed estimation summary for $\kappa = 48$ .................................................................................. 39
Table 5.1 Summary for the Birch Data ............................................................................................................. 63
Table 5.2 Maximum likelihood estimation summary when $\kappa = 48$ .................................................................. 67
Table 5.3 Distances and Correlations for Birch Data ..................................................................................... 69
LIST OF FIGURES

Figure 4.1  Latitude and longitude for sampling locations used in the analysis of sweet birch
tree counts .................................................... 35

Figure 5.1  Observed probabilities and site numbers for Birch trees .................. 62
CHAPTER 1 GENERAL INTRODUCTION

1.1 Introduction

For sets of random variables that exhibit statistical dependence, formulation of a model through direct specification of a joint distribution is difficult, unless a Gaussian distribution is appropriate for the problem. An alternative approach toward the construction of statistical models is conditional model specification. A conditional model specification involves writing forms for full conditional distributions for each of a set of random variables. In order for such a model to be valid, there must exist a joint probability distribution for the random variables that has the set of specified conditional distributions. In order for a valid model to be useful, we must be able to identify the joint distribution and use it for estimation and inference.

This dissertation concerns two topics in the formulation and analysis of conditionally specified statistical models. The first topic deals with the manner in which the dependencies among random variables in a set are identified. In particular, the development of models from one-parameter exponential family conditional distributions is considered under a relaxation of a commonly used restriction that dependencies are modeled only among pairs of variables in the set. The second topic concerns maximum likelihood estimation of parameters in conditionally specified models. Such estimation is complicated by the fact that many conditional model specifications allow identification of an associated joint distribution only up to an unknown normalizing constant. Estimation is considered for both basic models and mixture (or random parameter) models.

Besag (1974) formulated conditionally specified models that allow identification of appropriate joint distributions by defining dependence structures among variables in terms of Markov random fields. Besag considered conditional distributions specified as one-parameter exponential families, and restricted dependencies to be expressed only among pairs of random variables, the so-called 'pairwise-only' dependence assumption. The resulting class of models were called exponential family 'auto-models' by Besag (1974). In Chapter 3 of this dissertation, I generalize the results of Besag to situations in which dependencies may be expressed among groups of more than two random variables, while continuing to
make use of one-parameter exponential family conditional distributions. The potential for developing new classes of models by relaxing the pairwise-only dependence assumption has not previously been addressed.

Typically, the joint distribution corresponding to a conditionally specified model can be identified only up to a constant of proportionality, the 'normalizing constant'. In general, the normalizing constant will be a function of unknown parameters, and this poses a difficult problem for maximum likelihood estimation. Besag (1975) proposed the use of a 'pseudo-likelihood', defined as the product of the conditional distributions specified in the model. The efficiency of estimation based on Besag's pseudo-likelihood has been examined closely only in the Gaussian case, for which it is possible to derive a closed-form expression for the log likelihood. Besag (1977) and Kashyap and Chellappa (1983) examined this problem independently, showing that pseudo-likelihood estimation is consistent, but suffers from poor efficiency in situations that involve strong dependence. But, aside from the Gaussian case, maximum likelihood estimators have not been available. It is desirable to develop general strategies for maximum likelihood estimation for conditionally specified models. The fundamental problem in developing such strategies is the intractable normalizing constant, which typically is in the form of a high dimensional integral equation.

Monte Carlo integration is an attractive technique for evaluation of high dimensional integrals, and recent research into the use of Markov chain samplers has produced new tools to accomplish maximum likelihood estimation in complex statistical models. Geyer and Thompson (1992) introduced the concept of locating maximum likelihood estimates through repeated maximization of Monte Carlo approximations to the log likelihood function. There are a number of unresolved issues that render the approach of Geyer and Thompson difficult to apply to conditionally specified models. Chief among these is selection of an appropriate Monte Carlo sampling distribution for evaluation of the log likelihood at any given parameter value. What is needed is a sampling distribution that is functionally independent of current parameter estimates at any step in the estimation (so that it can be used for repeated evaluations of the likelihood) and yet reflects dependence in the model (so that it gives adequately precise Monte Carlo approximations). The formulation of sampling distributions that have these properties is addressed for conditionally specified models in Chapter 4 of this dissertation, along with the associated questions of assessing Monte Carlo error and convergence of an overall algorithm.

In some problems, the use of models with random parameters provides additional flexibility for modeling dependencies among random variables. For example, count data that exhibit overdispersion are difficult to model using one distribution with fixed parameter values, such as Poisson or binomial
distributions. In situations involving independent random variables, gamma-Poisson and beta-binomial mixture models have been successfully used to deal with overdispersion. To develop versions of these mixture models for use in problems with dependent random variables requires that a choice be made as to which model 'stage', data or random parameters, will be used to incorporate dependence information. One general structure is to use conditionally independent distributions to model observable random variables (conditional here means conditional on data model parameters) and a conditionally specified mixing distribution that incorporates dependence to model the random parameters. Such models might be called 'conditionally independent mixture models' (CIMMs). While the concept of CIMMs is fundamentally concerned with the manner in which dependence structures are modeled, the major question in allowing their use is how estimation may be accomplished. In Chapter 5, the basic methodology developed in Chapter 4 is extended and adapted for application to CIMMs. A number of theoretical results for the estimation strategies proposed are presented in Chapter 6.

1.2 Dissertation Organization

This dissertation consists of a literature review, two research papers, an extension of the second paper, and theoretical backgrounds for the two papers followed by a general conclusion, a bibliography listing references cited throughout the dissertation, and an acknowledgments.
CHAPTER 2 LITERATURE REVIEW

The formulation and estimation of parameters in conditionally specified models are areas of research that have developed rapidly in recent years, in part because of availability of powerful computers, and in part due to modeling considerations. As indicated by Arnold, Castillo, and Sarabia (1992), in complex settings it is often easier to visualize features of conditional distributions than it is to visualize features of marginal or joint distributions. But the ease of 'thinking conditionally' in the development of a statistical model is accompanied by the need to make use of a valid joint distribution for estimation and inference. Ensuring that a valid joint distribution corresponding to a set of specified conditionals does in fact exist, and using that joint distribution for the purposes of statistical analysis are not trivial matters.

Two major streams of research into the formulation and analysis of conditionally specified models may be identified in the literature. The first deals with general distributional theory, addressing the questions of existence and uniqueness of joint distributions that possess a set of specified conditional distributions. A brief overview of the major results from this work is presented in Section 2.1. The other branch of research into conditional modeling concerns the use of Markov random fields (MRFs) as a basis for specification of sets of conditional distributions, and the identification of joint distributions that may be developed from those specified conditionals. This approach to the formulation of conditionally specified models is the central focus of this dissertation, and is discussed in Section 2.2. Results relevant for the estimation of conditionally specified models are discussed in Section 2.3.

2.1 General Distribution Theory

For models formulated through specification of conditional distributions that have discrete support, sufficient conditions for existence and uniqueness of a corresponding joint distribution have been given by Patil (1965), Amemiya (1975), Nerlove and Press (1986), and Gourieroux and Montfort (1979). Kocherlakota and Kocherlakota (1992) consider various discrete bivariate distributions generated from conditional distributions and probability generating functions. Abrahaama and Thomas (1984) provide
necessary as well as sufficient conditions for existence of a joint having specific conditionals in the bivariate case with general support. Arnold and Press (1989) investigated necessary and sufficient conditions for the existence of bivariate distributions having conditionals within certain families, and indicated extensions of these conditions to higher dimensions. These authors called such joint distributions 'compatible' with the specified conditionals. Gelman and Speed (1993) derived conditions under which sets of conditional and marginal distributions uniquely determine a positive joint distribution.

Models for conditional distributions having a number of particular forms have also been developed. James (1975) considered multivariate distributions that have beta conditionals and Dirichlet conditionals. Arnold and Strauss (1988) developed the most general class of bivariate distributions such that both conditional distributions have exponential forms, and showed that the marginals of such bivariate distributions are not exponential.

Much of the work conducted on characterizing joint distributions having conditionals in prescribed families or of particular forms has been summarized and presented in a monograph by Arnold, Castillo, and Sarabia (1992). These authors provide a useful theorem for characterization of joint distributions having conditionals in prescribed families.

2.2 Conditional Specifications Based on Markov Random Fields

The use of Markov random fields as a basis for the development of conditional model formulation was pioneered by Besag (1974). Besag developed models for data observed on spatial lattice structures by using conditional specifications from one-parameter exponential families. Until recently, the basic results of Besag had not been expanded on in a substantial way. Cressie and Lele (1992) establish sufficient conditions for a set of conditionally specified distributions to define a MRF, and make initial attempts to formulate models from conditional distributions having multiparameter exponential family distributions. Kaiser and Cressie (1996) make explicit the constructive nature of the MRF approach to conditional model specification. They also address several technical issues that relax assumptions made by Besag (1974). These results allow, for example, the full development of models from conditional distributions having multiparameter exponential family form, such as that used in the spatial beta-binomial model of Chapter 5.
2.3 Estimation in Conditionally Specified Models

To make the MRF approach to conditional model specification feasible, methods for effective estimation must be developed. The basic difficulty in estimation of conditionally specified models is an unknown normalizing constant in the joint distribution. A great deal of research into approximation of intractable integrals has been done. Naylor and Smith (1982) outlined a numerical integration method using Gaussian quadrature for efficient calculation of posterior densities. They assumed that the posterior can be approximated by the product of Gaussian densities and a polynomial in parameters. Tierney and Kadane (1986) provided a method for the approximation of posterior moments and marginal densities through the use of Laplace expansions. Tierney, Kass, and Kadane (1989) provided fully exponential Laplace approximations to expectations and variances of nonpositive functions. Morris (1988) established a way to approximate moments of a univariate posterior using the Pearson family of distributions. Estimation of the ratio of two intractable integrals was first developed in physics. Valleau and Card (1972); Torrie, Valleau, and Bain (1973); Torrie and Valleau (1974); and Bennet (1976) used Monte Carlo simulation in estimation of free energy differences in which two intractable integrations are involved. The Gibbs sampler, introduced to statistics by Geman and Geman (1984) makes it possible to sample from conditionally specified models without the knowledge of normalizing constants, resulting in the wide application of Markov Chain Monte Carlo (MCMC) methods in statistical fields. Meng and Wong (1996) adopted Bennet's method for estimating the ratio of intractable integrals and enhanced its applicability to statistical data. Newton and Raftery (1994) provided a method to sample from posterior distributions using a weighted likelihood bootstrap, and various ways to estimate the marginal likelihood. Chib (1995) exploited the fact that the marginal density can be expressed as the product of the prior and the likelihood function, divided by the posterior density. Using this identity, which holds for any parameter value, he showed that estimation of the posterior density is available if all complete conditional densities used in the Gibbs sampler have closed form. Ogata (1990,1996) reduced the dimension of integration to one by scaling family, expressed the one-dimensional integration through a trapezoidal formula, and estimated the function value contained in that trapezoidal formula by MCMC methods. Estimation of intractable integrals is important, particularly in the estimation of posterior distributions, but for maximum likelihood estimation optimization of the likelihood is of primary concern. Any of these methods, among others, may be used to estimate integrals involved in optimization problems. However, many numerical integration methods become ineffective as dimension of the integrals to be evaluated increases. Without using Monte Carlo methods, Besag (1975) suggested pseudo maximum likelihood estimates (PMLEs). Besag’s pseudo-likelihood is just the product of the
conditional densities specified in a model. Besag (1975) showed that PMLEs are inefficient for conditionally specified Gaussian models when dependence is strong. Diggle and Gratton (1984) proposed evaluating integrations on a grid, smoothing these evaluations using a kernel smoother, and then maximizing the resulting smooth approximation. Many samples are required to evaluate integrations over a complete grid. Penttinen (1984) proposed Newton-Raphson iteration to maximize actual log likelihood with the gradient and Hessian matrix of the log likelihood estimated at each iteration by Monte Carlo integration, using a separate Monte Carlo sample for each evaluation of these quantities. Neither of these procedures allow proof of convergence of the algorithms or provide estimates of the size of the Monte Carlo error. Younes (1988) and Moyeed and Baddeley (1991) attempt to maximize the log likelihood through the use of stochastic approximation (Robbins-Munro algorithm), taking at each iteration a small step in the direction indicated by the estimated gradient and with step length unrelated to the distance from the MLE. These strategies make poor use of the MCMC samples. Geyer and Thompson (1992) suggested constrained Monte Carlo MLE, approximating the log likelihood function expressed as a ratio with a constant value through one MCMC realization, then maximizing the approximation. Geyer and Thompson were able to prove the convergence of this estimation strategy to the true MLE for exponential family models. Geyer (1994) also proved convergence outside the exponential family and provided the asymptotic behavior of the difference between the MC MLE and the true MLE. Gelfand and Carlin (1993) used Geyer and Thompson's MC MLE method for constrained and missing data problems, including certain two stage mixture models.
CHAPTER 3 MULTIWAY DEPENDENCE IN EXPONENTIAL FAMILY CONDITIONAL DISTRIBUTIONS

A paper submitted to the Journal Bernoulli
Jaehyung Lee, Mark S. Kaiser, and Noel Cressie

Abstract

Conditionally specified statistical models are frequently constructed from conditional one-parameter exponential family distributions. One way to formulate such a model is to specify the dependence structure among random variables through the use of a Markov random field. When this is done, a common assumption is that dependence is expressed only through pairs of random variables, the 'pairwise-only dependence' assumption. Using a Markov random field structure and the pairwise-only dependence assumption, Besag (1974) formulated exponential family 'auto-models', and showed the form that conditional one-parameter exponential family densities must have in such models. We extend those results under relaxation of the pairwise-only dependence assumption, and give a necessary form for conditional one-parameter exponential family densities under more general conditions of multiway dependence.

Keywords: Conditional exponential family distributions, Markov random fields, Multiway dependence, Pairwise dependence

3.1 Introduction

The construction of statistical models through specification of conditional distributions has seen rapid advancement in recent years. It is frequently difficult or impossible in complex situations to specify a model through formulation of a joint distribution for a complete set of response variables. Even in the Gaussian case, where it may be possible to write a such a joint distribution, the relative
merits of conditional specification versus simultaneous specification of a statistical model may lead one to prefer the conditional approach (Brook 1964, Besag 1974).

One approach to the formulation of conditionally specified models is to define a dependence structure among individual random variables such that the joint collection of variables forms a Markov Random Field (MRF). This approach was pioneered by Besag (1974), who inter alia developed results for conditional distributions from one-parameter exponential families. An important assumption in Besag's development was that individual random variables in the MRF exhibit 'pairwise-only' dependence. That is, interactions among random variables affect the joint probability density (or mass) function only through sums in its exponent that involve pairs of values. Thus, pairwise-only dependence yields simple expressions for the unnormalized joint probability density function of all random variables that form a MRF, and Besag (1974) supplies these expressions for one-parameter exponential family conditional distributions.

In this article, we extend Besag's results described above to one-parameter exponential families with dependence structures that are more complex than those allowed under pairwise-only dependence. In particular, our main result requires no assumptions on the way that dependence is expressed among random variables in exponential family conditional distributions, other than those required to ensure a valid joint distribution.

In what is to follow, we restrict attention to models formulated on the basis of an explicit MRF structure; see, in particular, Section 3.2. More generally, a considerable amount of research has been conducted on requirements for valid multivariate models obtained from conditional distributions. For example, see the monograph by Arnold et al. (1992). Further, James (1975) gives a necessary condition for a multivariate distribution to have beta conditionals, and Kocherlakota and Kocherlakota (1992) consider many aspects of models that lead to bivariate discrete distributions, with an emphasis on the use of probability generating functions and compounding.

The remainder of the paper is organized as follows. In Section 3.2, we define the problem and briefly review the fundamental quantities used in the formulation of multivariate models from a MRF structure. Section 3.3 presents the main result of this article, an extension of Besag's (1974) result for one-parameter exponential family conditional distributions, under relaxation of the pairwise-only dependence assumption. We demonstrate the use of this result in construction of models with normal and Poisson conditional distributions in Section 3.4, and Section 3.5 contains a short discussion.
3.2 Model Formulation from MRF Structure

The collection of random variables to be modeled will be denoted as the vector

$$Z = (Z(s_1), \ldots, Z(s_n))^T,$$  \hspace{1cm} (3.1)

where $s_i$ specifies the 'position' of $Z(s_i)$ in a random field $i = 1, \ldots, n$. It is perhaps easiest to visualize the notion of position for a random variable by considering spatial problems, in which a position corresponds to a physical location. We shall use this spatial conceptualization throughout this paper, although there is nothing inherently spatial about MRFs, and the results that follow apply equally to situations that are not spatial in nature. A dependence structure is specified by defining a 'neighborhood' for each component of $Z$. A site $s_j$ is called a neighbor of site $s_i$ if the conditional distribution of $Z(s_i)$, given values for $\{z(s_1), \ldots, z(s_{i-1}), z(s_{i+1}), \ldots, z(s_n)\}$, depends functionally on $z(s_j)$, where $j \neq i$. Define the 'neighborhood index set' for a variable $Z(s_i)$ as

$$N_i = \{j : s_j \text{ is a neighbor of } s_i\}.$$

The quantity of central concern in MRF models is the 'negpotential function' (e.g., Cressie 1993, p. 415), defined here in terms of a joint probability density (or mass) function $p$ as

$$Q(z) \equiv \log\{p(z)/p(0)\}; \quad z \in \zeta,$$  \hspace{1cm} (3.2)

where $\zeta \equiv \{z : p(z) > 0\}$ and it is assumed that $z = 0$ is a possible joint realization. Further, we assume the 'positivity condition' of Hammersley and Clifford (1971), which says that $\zeta = \zeta_1 \times \ldots \times \zeta_n$, where $\zeta_i$ is the support of $Z(s_i); i = 1, \ldots, n$. Under the positivity condition, $0 \in \zeta$ if $0 \in \zeta_i; i = 1, \ldots, n$. The negpotential function (3.2) is the focus of model formulation because it allows recovery of the joint probability density function $p(z)$, through

$$p(z) = \exp\{Q(z)\}/\int_{\zeta} \exp\{Q(t)\} d\mu(t); \quad z \in \zeta,$$  \hspace{1cm} (3.3)

for the appropriate measure $\mu$ (e.g., Lebesque or counting) and provided the denominator in (3.3) is finite.

Let $j_m \equiv (j(1), j(2), \ldots, j(m))$ denote a generic index of $m \leq n$ sites; that is, $j_m$ is any permutation of any $m$ site indices. Also, define $z(s_{j_m}) \equiv (z(s_{j(1)}), z(s_{j(2)}), \ldots, z(s_{j(m)}))^T$. Define the sets

$$T_m(p) \equiv \{\text{all distinct } p\text{-tuples formed from } j_m\}; \quad p \leq m,$$
and, for \( m = 1, \ldots, n \),
\[
z(s_{j(1)}) \cdots z(s_{j(m)}) G_{j(1)} \ldots j(m)(z(s_{j(1)}), \ldots, z(s_{j(m)}))
\]
\[
= \sum_{t=0}^{m-1} \sum_{j_{m-t} \in T_{m-t}} (-1)^t Q(\{z : z(s_h) = 0, \text{ if } h \text{ is not an element of } j_{m-t}\}) . \tag{3.4}
\]

For example,
\[
z(s_i)z(s_j)G_{i,j}(z(s_i), z(s_j)) = Q(0, \ldots, 0, z(s_i), 0, \ldots, 0, z(s_j), 0, \ldots, 0)
\]
\[
- Q(0, \ldots, 0, z(s_i), 0, \ldots, 0)
\]
\[
- Q(0, \ldots, 0, z(s_j), 0, \ldots, 0); \quad i \neq j .
\]

Notice that the \( G \)-functions defined by (3.4) have the property of being invariant to permutation of the associated indices. For example, \( G_{i,j}(z(s_i)z(s_j)) \equiv G_{j,i}(z(s_j)z(s_i)) \).

Besag (1974) showed that, given the positivity condition and that \( 0 \in \zeta ; \quad i = 1, \ldots, n \), the negpotential function may be expanded uniquely on \( \zeta \) as
\[
Q(z) = \sum_{1 \leq i \leq n} z(s_i)G_i(z(s_i)) + \sum_{1 \leq i < j \leq n} z(s_i)z(s_j)G_{i,j}(z(s_i), z(s_j))
\]
\[
+ \cdots + z(s_1) \cdots z(s_n)G_{1,2, \ldots, n}(z(s_1), \ldots, z(s_n)); \quad z \in \zeta . \tag{3.5}
\]

Note here that, although \( G \)-functions such as \( G_{2,1} \) have been defined in (3.4), they are not used in this expansion, all sums in (3.5) being over ordered subscripts. To simplify the expression (3.5) further, an important concept is that of a ‘clique’ (e.g., Besag 1974), defined to be a set of locations that consists either of an individual site or of sites that are all neighbors of each other. A theorem due to Hammersley and Clifford (see Besag 1974) shows that any collection of sites that do not form a clique yield corresponding components in (3.5) that are zero. Cressie and Lele (1992) show that any model yielding permutation invariant \( G \)-functions implies a valid joint probability distribution. A further simplification of (3.5) is obtained by assuming that all \( G \)-functions of order more than two in (3.5) equal zero (pairwise-only dependence). As a result of these simplifications, the expansion of \( Q(z) \) in (3.5) contains only terms involving \( G_i(\cdot); i = 1, \ldots, n \) and \( G_{i,j}(\cdot, \cdot); i = 1, \ldots, n, \ j = i + 1, \ldots, n, \ j \in N_i \).

In Section 3.3, we relax the pairwise-only dependence assumption, so that the clique structure comes into play for terms in (3.5) of higher order than two.

An important quantity in Besag’s development of conditionally specified models is the negpotential difference, \( Q(z) - Q(z_i) \), where \( z_i \equiv (z(s_1), \cdots, z(s_{i-1}), 0, z(s_{i+1}), \cdots, z(s_n))^T \). Under pairwise-only dependence, this difference allows identification of the forms that will be taken by the quantities
\{z(s_1)z(s_j)G_{i,j}(z(s_i), z(s_j)) : j \in N_i\}. To build conditionally specified models with higher-order dependence requires careful identification of the terms in (3.5) making up cliques of sizes 3, 4, ..., n. Fix site i and consider all cliques of size \((g + 1)\) containing i. Let \(j_g \equiv (j(1), j(2), \ldots, j(g))\) denote a generic index of the \(g\) sites other than \(i\) that are in the clique, and let \(z(s_{j_g}) \equiv (z(s_{j(1)}), z(s_{j(2)}), \ldots, z(s_{j(g)})^T\).

Then, differencing the negpotential function and application of (3.5) yields:

\[
Q(z) - Q(z_i) = z(s_i)G_i(z(s_i)) + \sum_{g=1}^{n-1} \sum_{\beta \in S_g(t,i)} \left\{ \prod_{k=1}^{g} z(s_{j(k)}) \right\} G_{i,j_g}(z(s_i), z(s_{j_g})) \right\},
\]

where

\[
S_g(t, i) \equiv \{j_g : 1 \leq j(1) < \ldots < j(t) < i < j(t + 1) < \ldots < j(g) \leq n\} \quad t = 2, \ldots, (g - 1),
\]

and

\[
S_g(0, i) \equiv \{j_g : (i - 1) < j(1) < j(2) < \ldots < j(g) \leq n\},
\]

\[
S_g(g, i) \equiv \{j_g : 1 \leq j(1) < j(2) < \ldots < j(g) \leq (i - 1)\}.
\]

Note that \(t\) indexes the number of sites in \(j_g\) with indices less than \(i\).

### 3.3 Conditional Exponential Family Distributions

To formulate a model from a MRF structure, one must choose a specific form for the conditional distribution at each site. Besag (1974) suggested the use of one-parameter exponential families to model conditional probability density (or mass) functions. That is, for \(i = 1, \ldots, n,
\]

\[
p(z(s_i) | \{z(s_j) : j \in N_i\}) = \exp[\{A_i(z(s_j) : j \in N_i)B_i(z(s_i)) + C_i(z(s_i)) + D_i(z(s_j) : j \in N_i)\}].
\]

where \(B_i(\cdot)\) and \(C_i(\cdot)\) have specified forms and \(A_i(\cdot)\) and \(D_i(\cdot)\) are functions of the conditioning random variables whose sites are in \(N_i\).

#### 3.3.1 Pairwise-Only Dependence

An important result of Besag (1974) established that, under the exponential family specification (3.7) and pairwise-only dependence, the functions \(\{A_i(\cdot)\}\) must be of the form

\[
A_i(z(s_j) : j \in N_i) = \alpha_i + \sum_{j \in N_i} \theta_i,j B_j(z(s_j)),
\]

(3.8)
where \( \theta_{i,j} = \theta_{j,i} \). A consequence of this result is that, up to an additive constant, the negpotential function (3.5) takes the form

\[
Q(z) = \sum_{1 \leq i \leq n} \{ \alpha_i B_i(z(s_i)) + C_i(z(s_i)) \} + \sum_{1 \leq i < j \leq n} \{ \theta_{i,j} B_i(z(s_i)) B_j(z(s_j)) \}, \tag{3.9}
\]

where \( \theta_{i,j} = \theta_{j,i} \), \( \theta_{i,i} = 0 \), and \( \theta_{i,j} = 0 \) unless \( j \in N_i \).

### 3.3.2 Multiway Dependence

We now give parallel results to those in Section 3.3.1, for models formed from the conditional specification (3.7), without imposing the restriction of pairwise-only dependence.

**Theorem** Let \( \{Z(s_i) : i = 1, \ldots, n\} \) have conditional probability density or mass functions of the form (3.7), and assume that a MRF has been specified through the neighborhoods \( \{N_i : i = 1, \ldots, n\} \). Then the functions \( A_i(\cdot) \) must be of the form

\[
A_i(\{z(s_j) : j \in N_i\}) = \alpha_i + \sum_{g=1}^{n-1} \sum_{I(g)} \left[ \theta_{i,j(1), \ldots, j(g)} \prod_{k=1}^g \{ B_j(z(s_j(k))) - B_j(0) \} \right], \tag{3.10}
\]

where \( \theta_{i,j(1), \ldots, j(g)} = \theta_{p(i,j(1), \ldots, j(g))} \) for any permutation \( p(i,j(1), \ldots, j(g)) \), \( \theta_{i,j(1), \ldots, j(g)} = 0 \) if any \( j(k) = i \) or if \( \{i, j(1), \ldots, j(g)\} \) does not form a clique, and

\[
l(g) \equiv \{j(1), \ldots, j(g) : 1 \leq j(1) < j(2) < \ldots < j(g) \leq n\}.
\]

**Proof:** From the definition of the negpotential function (3.2) and the conditional specifications (3.7),

\[
\begin{align*}
\exp(Q(z) - Q(z_{s_i})) &= \frac{p(z(s_i)) \{z(s_j) : j \neq i\}}{p(0(s_i)) \{z(s_j) : j \neq i\}} \\
&= \exp[A_i(\{z(s_j) : j \in N_i\})] B_i(z(s_i)) - B_i(0)] \\
&\quad + C_i(z(s_i)) - C_i(0), \tag{3.11}
\end{align*}
\]

where \( 0(s_i) \) denotes a realization of 0 at site \( i \). Then, evaluating the logarithm of (3.11) at \( z = (0(s_1), \ldots, 0(s_{i-1}), z(s_i), 0(s_{i+1}), \ldots, 0(s_n))^T \), and equating the result with equation (3.6), yields

\[
z(s_i) C_i(z(s_i)) = A_i(0) \{B_i(z(s_i)) - B_i(0)\} + C_i(z(s_i)) - C_i(0). \tag{3.12}
\]

Similarly, for \( j \in N_i \) (and thus \( i \in N_j \)), evaluation of (3.6) and the logarithm of (3.11) at

\[
z = (0(s_1), \ldots, 0(s_{i-1}), z(s_i), 0(s_{i+1}), \ldots, 0(s_j), z(s_j), 0(s_{j+1}), \ldots, 0(s_n))^T
\]

yields
\[ z(s_j)G_i(z(s_i)) + z(s_i)z(s_j)G_{i,j}(z(s_i), z(s_j)) \]
\[ = A_i(0, \cdots, 0, z(s_j), 0, \cdots, 0)\{ B_i(z(s_i)) - B_i(0) \} + C_i(z(s_i)) - C_i(0), \]
from \( \exp(Q(z) - Q(z_i)) \), and

\[ z(s_j)G_j(z(s_j)) + z(s_i)z(s_j)G_{i,j}(z(s_i), z(s_j)) \]
\[ = A_j(0, \cdots, 0, z(s_i), 0, \cdots, 0)\{ B_j(z(s_j)) - B_j(0) \} + C_j(z(s_i)) - C_j(0), \]
from \( \exp(Q(z) - Q(z_j)) \). Thus,

\[ z(s_i)z(s_j)G_{i,j}(z(s_i), z(s_j)) \]
\[ = \{ B_i(z(s_i)) - B_i(0) \}\{ A_i(0, \cdots, 0, z(s_j), 0, \cdots, 0) - A_i(0) \} \]
\[ = \{ B_j(z(s_j)) - B_j(0) \}\{ A_j(0, \cdots, 0, z(s_i), 0, \cdots, 0) - A_j(0) \}, \]
which implies that
\[ \frac{A_i(0, \cdots, 0, z(s_j), 0, \cdots, 0) - A_i(0)}{B_j(z(s_j)) - B_j(0)} = \frac{A_j(0, \cdots, 0, z(s_i), 0, \cdots, 0) - A_j(0)}{B_i(z(s_i)) - B_i(0)}. \]

For the left hand side, which is a function of \( z(s_j) \in \zeta_j \), to be equal to the right hand side, a function of \( z(s_i) \in \zeta_i \), both must be equal to a constant, \( \theta_{i,j} \). Hence,

\[ z(s_i)z(s_j)G_{i,j}(z(s_i), z(s_j)) = \theta_{i,j}\{ B_i(z(s_i)) - B_i(0) \}\{ B_j(z(s_j)) - B_j(0) \}. \quad (3.13) \]

Now, (3.4) gives
\[ G_{ij}(z(s_i), z(s_j)) \equiv G_{j,i}(z(s_j), z(s_i)), \]
and hence \( \theta_{i,j} = \theta_{j,i}, j \neq i \).

The forms of higher-order interactions in (3.5) follow from an inductive argument on the clique size. Without loss of generality, assume that the collection of \( m \) sites \( \{ s_1, \ldots, s_m \} \) forms a clique. The induction assumption is that, for any \( 1 \leq h \leq m \),

\[ z(s_{j(1)}) \cdots z(s_{j(h)})G_{j(1),\ldots,j(h)}(z(s_{j(1)}), \ldots, z(s_{j(h)})) \]
\[ = \theta_{j(1),\ldots,j(h)}\prod_{k=1}^{h} \{ B_{j(k)}(z(s_{j(k)})) - B_{j(k)}(0) \}, \quad (3.14) \]

where \( \theta_{j(1),\ldots,j(h)} = \theta_{p(j(1),\ldots,j(h))} \) for any permutation \( p(j(1), \ldots, j(h)) \).

Now, let \( \{ s_1, \ldots, s_{m+1} \} \) be a clique of size \( m + 1 \) and define
\[ z^* \equiv (z(s_1), \ldots, z(s_{m+1}), 0(s_{m+2}), \ldots, 0(s_n))^T. \]

For \( i \in \{ 1, \ldots, (m + 1) \} \), define
\[ z^*_i \equiv (z(s_1), \ldots, z(s_{i-1}), 0(s_i), z(s_{i+1}), \ldots, z(s_{m+1}), 0(s_{m+2}), \ldots, 0(s_n))^T. \]
Equation (3.6) gives
\[
Q(z^*) - Q(z_i^*) = z(s_i)G_i(z(s_i)) + z(s_i) \sum_{g=1}^{m-1} \sum_{t=0}^{g} \left\{ \prod_{k=1}^{g} z(s_j(k)) \right\} G_{i,j_s}(z(s_i), z(s_j))
+ H(z(s_i), z(s_{j(1)}), \ldots, z(s_{j(m)})),
\]
where
\[
H(z(s_i), z(s_{j(1)}), \ldots, z(s_{j(m)}))
= z(s_i) \sum_{t=0}^{m} \sum_{j_s \in S_m(t,i)} \left\{ \prod_{k=1}^{g} z(s_j(k)) \right\} G_{i,j_s}(z(s_i), z(s_j))
= z(s_i)z(s_{j(1)}) \ldots z(s_{j(m)})G_{i,j(1)}, \ldots, j(m)(z(s_i), z(s_{j(1)}), \ldots, z(s_{j(m)})).
\]
The latter equality holds, since for \( g = m \) there can be only one nonempty set \( S_m(t, i) \), and this set can contain only one element. The definitions immediately following equation (3.6) yield,
\[
S_m(t, 1) = \{(2, \ldots, m + 1)\} \text{ if } t = 0,
S_m(t, 2) = \{(1, 3, \ldots, m + 1)\} \text{ if } t = 1,
\vdots
\vdots
\vdots
S_m(t, m + 1) = \{(1, \ldots, m)\} \text{ if } t = m;
\]
otherwise, \( S_m(t, i) = \emptyset \). Thus, the induction assumption (3.14) yields,
\[
Q(z^*) - Q(z_i^*) = z(s_i)G_i(z(s_i))
+ \left\{ B_i(z(s_i)) - B_i(0) \right\} \sum_{g=1}^{m-1} \sum_{t=0}^{g} \sum_{j_s \in S_m(t, i)} \theta_{i,j_s} \prod_{k=1}^{g} \left\{ B_{j(k)}(z(s_j(k))) - B_{j(k)}(0) \right\}
+ z(s_i)z(s_{j(1)}) \ldots z(s_{j(m)})G_{i,j(1)}, \ldots, j(m)(z(s_i), z(s_{j(1)}), \ldots, z(s_{j(m)})).
\]
Using the conditional specification (3.7), evaluation of the logarithm of (3.11) at the value \( z^* \) also yields,
\[
Q(z^*) - Q(z_i^*) = A_i(z_i^* - z_i(0)) + C_i(z_i) - C_i(0),
\]
where \( z_i^* = (z(s_1), \ldots, z(s_{i-1}), z(s_{i+1}), \ldots, z(s_m), 0, 0, \ldots, 0)^T \). Finally, using (3.12) in equation (3.15) and equating the result with (3.16), we have that
\[
z(s_i)z(s_{j(1)}) \ldots z(s_{j(m)})G_{i,j(1)}, \ldots, j(m)(z(s_i), z(s_{j(1)}), \ldots, z(s_{j(m)}))
= A_i(z_i - z_i(0)) - \sum_{g=1}^{m-1} \sum_{t=0}^{g} \sum_{j_s \in S_m(t, i)} \theta_{i,j_s} \prod_{k=1}^{g} \left\{ B_{j(k)}(z(s_j(k))) - B_{j(k)}(0) \right\} \left\{ B_i(z(s_i)) - B_i(0) \right\}.
\]
The choice of \( i \) in the clique \( \{1, \ldots, (m + 1)\} \) is arbitrary, so that there are \((m + 1)\) expressions of the form (3.17) for the term
\[
z(s_i)z(s_{j(1)}) \ldots z(s_{j(m)})G_{i,j(1)}, \ldots, j(m)(z(s_i), z(s_{j(1)}), \ldots, z(s_{j(m)})).
\]
Dividing each of these by \( \prod_{k=1}^{m+1} \{ B_k(z(s_k)) - B_k(0) \} \) yields \((m + 1)\) functions

\[
\frac{A_i(z^*_{i-1}) - A_i(0) - \sum_{k=1}^{m+1} \sum_{j \in S(k)} \theta_{i,j} \prod_{k=1}^{m+1} \{ B_j(z(s_{j(k)})) - B_j(0) \}}{\prod_{k \neq i} \{ B_k(z(s_k)) - B_k(0) \}} \quad ; \quad i = 1, \ldots, (m + 1),
\]

which are all of which are equal.

For each \( i \), the function given in (3.18) is independent of \( z(s_i) \) and may be written as a function of \( z^*_i \) alone, say \( H_i(z^*_i) \); \( i = 1, \ldots, (m + 1) \). Then we have that

\[
H_1(z^*_1) = H_2(z^*_2) = \ldots = H(z^*_{(m+1)}),
\]

for all \( z^* \in \prod \{ \zeta_1 \times \ldots \times \zeta_{m+1} \times \{ 0 \} \times \ldots \times \{ 0 \} \} \).

Consider the value \( z^T = (z(s_1), z(s_2), 0(s_3), \ldots, 0(s_n))^T \). Then (3.19) gives

\[
H_1(z^T_{-1}) = H_1(z(s_1), 0) = H_2(z(s_1), 0) = H_2(z^T_{-2}).
\]

Since this must hold for all \( z(s_1) \in \zeta_1 \) and \( z(s_2) \in \zeta_2 \), \( H_1(z^*_1) \) and \( H_2(z^*_2) \) must be functions of \( \{ z(s_3), \ldots, z(s_{m+1}) \} \) alone.

Now consider \( z^O = (z(s_1), z(s_2), z(s_3), 0(s_4), \ldots, 0(s_n))^T \). Then

\[
H_3(z^O_3) = H_3(z(s_1), z(s_2), 0) = H_i(z(s_2), z(s_3), 0),
\]

which in fact depends only on \( z(s_3) \), since \( H_1(z^*_1) \) has been shown not to depend on \( z(s_1) \) or \( z(s_2) \).

Thus, \( H_3(z^*_3) \) (and, similarly, \( H_1(z^*_1) \) and \( H_2(z^*_2) \)) must be a function of \( \{ z(s_4), \ldots, z(s_{m+1}) \} \) alone.

Continuing this process shows that \( H_1(z^*_1), \ldots, H_m(z^*_m) \) are functions of only \( z(s_{m+1}) \). By (3.19), these functions must also be equal to \( H_{m+1}(z^*_m) \), which is independent of \( z(s_{m+1}) \). Thus,

\[
H_1(z^*_1) = \ldots = H_{m+1}(z^*_m) = \theta_1, \ldots, (m + 1),
\]

for \( \theta_{1, \ldots, (m+1)} \) constant.

Thus, for any choice of \( i \in \{ 1, \ldots, (m + 1) \} \), we have that

\[
z(s_i)z(s_{j(1)}) \ldots z(s_{j(m)})G_{i,j(1),\ldots,j(m)}(z(s_i), z(s_{j(1)}), \ldots, z(s_{j(m)}))
\]

\[
= \theta_{1, \ldots, (m+1)} \prod_{k=1}^{(m+1)} \{ B_k(z(s_k)) - B_k(0) \}.
\]

Now, \( G_{i,j(1),\ldots,j(m)}(z(s_i), z(s_{j(1)}), \ldots, z(s_{j(m)})) \) is invariant to permutation of its subscripts by (3.4), so that \( \theta_{p(1,\ldots,(m+1))} = \theta_{1,\ldots,(m+1)} \) for any permutation \( p(1,\ldots,(m+1)) \). The result of equation (3.20)
holds for any clique of size \( (m + 1) \) given that the induction assumption (3.14) is true for any clique of size \( m \). Thus, from (3.6), (3.12), (3.13), and (3.20), for any \( z \in \zeta \),

\[
Q(z) - Q(z_i) = C_i(z(s_i)) - C_i(0) + \{B_i(z(s_i)) - B_i(0)\}
\]

\[
+ \left[ A_i(0) + \sum_{g=1}^{n-1} \sum_{s_r, t_r} \sum_{\theta_{i,j,k}} \prod_{k=1}^{g} \{B_{j(k)}(z(s_{j(k)})) - B_{j(k)}(0)\} \right].
\]

(3.21)

Note that \( A_i(\{z(s_j) : j \in N_i\}) \) is the coefficient of \( \{B_i(z(s_i)) - B_i(0)\} \) in the expression for \( Q(z) - Q(z_i) \) given by the logarithm of (3.11). Now, in (3.10), define \( \alpha_i \equiv A_i(0) \). Then, with \( \theta_{i,j,k} \) as in the statement of the Theorem, \( A(\cdot) \) is equal to the term in square brackets in (3.21), and hence the result of the theorem follows.

**Corollary** Assume (3.7) and the conditions of the Theorem. Then, under the same conditions on the constants \( \{\theta_{i,j(1),...j(g)}\} \) as in the Theorem, \( Q(z) \) is, up to an additive constant, given by

\[
Q(z) = \sum_{i=1}^{n} \{\alpha_i(B_i(z(s_i)) - B_i(0)) + C_i(z(s_i))\}
\]

\[
+ \sum_{1 \leq j(1)<j(2) \leq n} \theta_{j(1),j(2)} \prod_{k=1}^{g} \{B_{j(k)}(z(s_{j(k)})) - B_{j(k)}(0)\}
\]

\[
+ \cdots + \sum_{1 \leq j(1)<...<j(g) \leq n} \theta_{j(1),...j(g)} \prod_{k=1}^{g} \{B_{j(k)}(z(s_{j(k)})) - B_{j(k)}(0)\}
\]

(3.22)

\[
+ \cdots + \theta_{1,2,...,n} \prod_{k=1}^{n} \{B_k(z(s_k)) - B_k(0)\}.
\]

**Proof:** The proof of this Corollary follows immediately from substitution of (3.12), (3.13), and (3.20) into (3.5).

Notice that if \( Q(z) \) is integrable, then the joint probability density function of \( Z \) is given by (3.3).

### 3.4 Examples

The results of Section 3.3 allow conditional specification of models, provided the denominator of (3.3) is finite. Here, we show that some new models may be formed through the expression of conditional dependence in groups of random variables greater than size two. For simplicity, we relax the pairwise-only dependence assumption to allow three-way dependence.

Consider first a model formed with Gaussian conditional distributions \( \{N(\mu_i(z(s_j) : j \in N_i), \sigma^2) : i = 1, \ldots, n\} \). The conditional specifications (3.7) have

\[
A_i(\cdot) = \mu_i(\cdot)/\sigma^2, \quad B_i(z(s_i)) = z(s_i),
\]

\[
D_i(\cdot) = -\mu_i(\cdot)^2/(2\sigma^2), \quad C_i(z(s_i)) = (-z^2(s_i)/(2\sigma^2)) - (1/2) \log(2\pi\sigma^2).
\]
Under the assumption that all G-functions of order more than three in (3.5) equal zero, the Theorem of Section 3.3 gives that
\[ A_i(\{z(s_j) : j \in N_i\}) = \alpha_i + \sum_{j=1}^{n} \theta_{i,j} z(s_j) + \sum_{1 \leq j < k \leq n} \theta_{i,j,k} z(s_j)z(s_k), \]
where \( \theta_{i,j} = \theta_{j,i}, \theta_{i,i} = 0, \theta_{i,j} = 0 \) unless \( j \in N_i, \theta_{i,j,k} = \theta_{p(i,j,k)}, \) for any permutation \( p(i,j,k). \)
\( \theta_{i,i,k} = \theta_{i,j,i} = 0, \) and \( \theta_{i,j,k} = 0 \) if \( \{i, j, k\} \) does not form a clique. From the Corollary, the joint density of \( Z = (Z(s_1), \ldots, Z(s_n))^T \) is given by (3.3) with
\[ Q(z) = \sum_{i=1}^{n} (\alpha_i z(s_i) - \frac{z(s_i)^2}{2\sigma^2}) + \sum_{1 \leq i < j \leq n} \theta_{i,j} z(s_i)z(s_j) + \sum_{1 \leq i < j < k \leq n} \theta_{i,j,k} z(s_i)z(s_j)z(s_k). \]

Now, under pairwise-only dependence (i.e., \( \theta_{i,j,k} = 0, \) for all \( i, j, k), \) the joint distribution is multivariate normal. Adding three-way interactions leads to a new multivariate distribution. The properties of this distribution, as well as distributions formed through the inclusion of fourth and higher-order terms, remain to be investigated.

Consider now a model for count data. A model having Poisson conditional distributions,
\[ \{ \text{Po}(\lambda_i(\{z(s_j) : j \in N_i\})) : i = 1, \ldots, n \}, \]
yields
\[ A_i(\cdot) = \log(\lambda_i(\cdot)), \quad B_i(z(s_i)) = z(s_i), \]
\[ D_i(\cdot) = -\lambda_i(\cdot), \quad C_i(z(s_i)) = -\log(z(s_i)!) \]
Assuming that all G-functions of order more than three in (3.5) equal zero, the negpotential function for a model with Poisson conditionals is, up to an additive constant, given as
\[ Q(z) = \sum_{i=1}^{n} (\alpha_i z(s_i) - \log(z(s_i)!) + \sum_{1 \leq i < j \leq n} \theta_{i,j} z(s_i)z(s_j) + \sum_{1 \leq i < j < k \leq n} \theta_{i,j,k} z(s_i)z(s_j)z(s_k). \]
A question of immediate interest with this model is whether multiway dependence allows any of the spatial dependence parameters to be positive. Recall that, for the pairwise-only dependence Poisson auto-model, Besag (1974) observed that the spatial dependence parameters must be negative for the denominator of (3.3) to be finite. Unfortunately, none of the \( \{\theta_{i,j}\} \) or \( \{\theta_{i,j,k}\} \) can be positive, as may be seen for the \( \{\theta_{i,j}\} \) by setting \( \{z(s_k) : k \neq 1, 2\} \) equal to zero in the expression above for \( Q(z) \) and summing over \( \zeta_1 \times \zeta_2; \zeta_i = \{0, 1, \ldots\}, i = 1, 2. \) Also, the \( \{\theta_{i,j,k}\} \) must be negative, as may be seen by setting \( \{z(s_i) : l \neq 1, 2, 3\} \) equal to zero. Thus, a model with Poisson conditional distributions and three-way dependence may only be used to model negative spatial dependence, and this remains true for general multiway dependence.
3.5 Discussion

Exponential family conditional distributions are becoming increasingly important in the statistical analysis of multivariate and spatial problems, and show great promise for the development of spatio-temporal models. The results given in this paper show that the dependence present must be embodied in a finite set of dependence parameters, more general than those contained in the auto-models of Besag (1974). Thus, having all dependence parameters equal to zero implies an independence model that, in principle, could be tested for during model development. We show that pairwise-only dependence is a special case that, although more general than the independence model, is nested within a class of models exhibiting multiway dependence.

Our results are extensions of those presented by Besag (1974), and models that result from applications of our Theorem and its Corollary bear much in common with Besag's auto-models. There is, however, an interesting distinction seen to arise in the form of the functions $A_i(\cdot)$, given in (3.8) for Besag's original auto-models, and those given in (3.10). Our results certainly apply to the pairwise-only dependence case, and correspond exactly to Besag's results in that situation. Following the result (3.10), a pairwise-only dependence model would have

$$A_i(\{z(s_j) : j \in N_i\}) = \alpha_i + \sum_{j \in N_i} \theta_{i,j}(B_j(z(s_j)) - B_j(0)),$$

where $\theta_{i,j} = \theta_{j,i}$. In this case, $\sum_{j \in N_i} \theta_{i,j} B_j(0)$ would simply be absorbed into $\alpha_i$ giving the expression (3.8). But, in models with cliques of size greater than two, there is no similar simplification, and the differences $B_j(z(s_j)) - B_j(0)$ need to remain in the product term of equation (3.10). (In the examples presented in Section 4, this is of no importance since, in both cases, $B_i(0) = 0; i = 1, \ldots, n$).

Inference for the parameters in the multiway dependence models defined by (3.22) is complicated by an intractable normalizing constant. These models are well suited to inference based on Markov Chain Monte Carlo methods (e.g., Besag and Green 1993).

Acknowledgment

This research was supported by the U.S. Environmental Protection Agency under Assistance Agreement No. CR822919-01-0, and the Office of Naval Research (Cressie) under Grant N00014-93-1-0001.
CHAPTER 4  MAXIMUM LIKELIHOOD ESTIMATION IN CONDITIONALLY SPECIFIED STATISTICAL MODELS

A paper to be submitted to the Journal of Computational and Graphical Statistics
Jaehyung Lee and Mark S. Kaiser

Abstract

We propose a strategy for maximum likelihood estimation of parameters appearing in the joint distribution of a set of random variables modeled through the specification of full conditional probability density or mass functions. This strategy relies on maximization of a sequence of Monte Carlo approximations to the log likelihood function. The fundamental issue addressed in our strategy is formulation of an importance sampling distribution as a product of marginal functions, where those marginals are chosen in a way that reflects the influence of dependence on the first two moments of the actual statistical model under consideration. We address a number of practical issues in the use of Monte Carlo methods to locate maximum likelihood estimates, including criteria for when an additional sampling distribution should be selected and the selection of appropriate starting values. The estimation strategy proposed is applied to a Winsorized Poisson auto-model to demonstrate the presence of positive spatial dependence in counts of a particular tree species.

Keywords: Conditional exponential family distributions, Markov random fields, Winsorized Poisson auto-model

4.1 Introduction

In this article we consider the problem of obtaining full maximum likelihood estimates (MLEs) from statistical models formulated through specification of a set of conditional distributions. We consider this problem without restricting attention to joint distributions of low dimension. As is well known, the major difficulty in obtaining parameter estimates from conditionally specified models is the presence of
an intractable normalizing constant in the joint distribution, a feature that this problem shares with Bayesian analyses of many hierarchical models. Thus, it is natural to apply the same methodologies used successfully in Bayesian estimation, namely Markov Chain Monte Carlo (MCMC) methods, to the problem of obtaining MLEs from conditionally specified models. We assume that, for a likelihood resulting from a joint density or mass function with \( p \) parameters, and having an intractable normalizing constant, an iterative numerical solution of the likelihood equations (e.g., direct search or Newton-Raphson) is appropriate for location of the MLEs.

There are several aspects of this problem that render straightforward use of MCMC techniques difficult. In contrast with simulation from a joint posterior, Monte Carlo integrations must be performed at least once for the normalizing constant at each iteration, and possibly also for the \( p \) first and \( p(p+1)/2 \) second derivatives of the log likelihood function. The use of samples obtained from a different Markov chain at each iteration is not efficient, and suffers other deficiencies related to the size of the Monte Carlo error in estimation of integrals (Geyer, 1996). A second aspect of likelihood analysis that complicates the use of MCMC methodology lies in the need to estimate an appropriate asymptotic sampling distribution for the purpose of constructing inferential quantities such as standard errors. If standard maximum likelihood theory applies, this involves estimation of the negative Hessian matrix, the matrix of second derivatives of the log likelihood with respect to the parameter vector. It is quite possible that estimation of these second derivatives involves integrals in which the integrands distribute their mass differently over the sample space than does the normalizing constant. Also, if one wishes to estimate the Monte Carlo error involved in the overall estimation procedure, estimation of an additional covariance matrix is required (Geyer 1992, 1994). All of these potential difficulties boil down to the fact that, for the use of MCMC methods in likelihood analysis, one needs repeated Monte Carlo estimation of a number of integrals that may involve integrands of varying shapes.

A considerable amount of progress has been made in solving these difficulties. A cornerstone for the use of MCMC methods in optimization of likelihoods (i.e., finding MLEs) is an approach based on importance sampling that uses a single MCMC simulation for multiple evaluations of the quantities needed in an iterative estimation algorithm. In this approach, introduced by Geyer and Thompson (1992), a sampling distribution is chosen that does not depend on the values of parameters contained in the normalizing constant. A sample of values simulated from this distribution is then not dependent on the current parameter estimates in an iterative optimization algorithm, and thus may be used for repeated evaluations of the log likelihood (and, possibly, derivatives). But, unless the starting values of the overall algorithm have been chosen exceedingly well, it is unlikely that only one MCMC sample
will be required. Geyer and Thompson (1992) suggested that one MCMC sample be used for repeated iterations of an overall optimization algorithm as long as current parameter estimates lie within a certain arbitrarily specified region of the parameter space. In their paper, this region was specified through the use of a 'maximum step size' allowable in movement of the current parameter estimate. If maximization of the log likelihood moves the parameter to the border of this region, the sampling distribution is changed to give a better approximation to the log likelihood and another MCMC sample is taken.

In order for the constrained maximization approach of Geyer and Thompson (1992) to be effective, one needs to choose an appropriate importance distribution. Many of the suggestions currently available for choosing an importance distribution make use of sampling from finite mixtures of distributions (Torrie and Valleau 1977, Marinari and Parisi 1992, Geyer 1993), the fundamental idea being to construct a sampling distribution that places sufficient mass in all 'interesting' areas of the regions being integrated over. The use of mixtures has also been proposed as a technique to further remove the MCMC sampling process from the overall maximization algorithm for locating MLEs (Geyer 1996). An older idea, apparently proposed by Penttinen (1984), is to use a Monte Carlo sample from an importance sampling distribution that has independent marginals. According to Geyer and Thompson (1992), this restricts the method to cases allowing only weak dependence in the joint distribution. Clearly, however, the idea of sampling from independent distributions reduces the proportion of the total estimation effort (i.e., programming time plus computing time) that must be devoted to the sampling process.

In this article, we revive Penttinen's concept of using importance sampling distributions formed from independent marginal distributions. We do so under the condition that the (estimated) first two moments of marginal distributions for each component from the actual joint distribution are matched by the independent components of the sampling distribution. Combined with a procedure similar to the constrained maximization of Geyer and Thompson (1992), this provides MLEs for joint distributions of high dimension, even those known only in unnormalized form. For the class of models considered, we show that good starting values are available from Laplace approximations, that Monte Carlo precision may be computed, and that the need to select an arbitrary 'maximum step size' can be eliminated.

The models with which we are concerned are conditionally specified models formed from the use of Markov random fields. Such models are appropriate for a large number of problems involving dependent random variables, including spatial lattice models, longitudinal models, and random parameter or 'mixture' models. We restrict attention here to the simplest case, a conditionally specified data model with fixed parameters for which MLEs are desired. An outline of the basic theory needed in the
formulation of such models is contained in Kaiser and Cressie (1996). We give only a brief overview in Section 4.2 for the sake of completeness. Section 4.3 contains a description of the overall estimation process proposed here. Section 4.4 presents the crucial portion of this process, selection of importance sampling distributions for Monte Carlo evaluations. In Section 4.5 we deal with the choice of starting value, and Section 4.6 contains an example in which we apply full maximum likelihood estimation to a spatial lattice model having Winsorized Poisson conditional distributions. This model allows positive spatial dependence to be estimated for random variables with conditional Poisson distributions.

4.2 Conditional Model Specification for Markov Random Fields

Consider a collection of random variables \( Y \equiv (Y(s_1), \ldots, Y(s_n))^T \) where \( s_i \) denotes the location of a random variable in a random field with finite index. With \( f(\cdot | \cdot) \) denoting a generic conditional probability density or mass function, a conditionally specified model consists of the \( n \) functions,

\[
f(y(s_i) | \{y(s_j) : j \neq i\}) ; \quad i = 1, \ldots, n.
\]  

(4.1)

Dependence is modeled by defining, for each \( Y(s_i) \), a dependence set or neighborhood,

\[ N_i \equiv \{s_h : f(y(s_i) | \{y(s_j) : j \neq i\}) \text{ depends on } y(s_h); \ h \neq i\}, \]

and the corresponding dependence index set \( D_i \equiv \{h : s_h \in N_i\} \). Given this information, our goal is to identify a joint distribution corresponding to the set of specified conditionals, assuming that such a joint exists. This approach to formulation of statistical models was pioneered by Besag (1974). Kaiser and Cressie (1996) outline the constructive process involved, and give conditions under which the joint density or mass function may be identified up to a normalizing constant as,

\[
f(y) = \frac{\exp\{Q(y)\}}{\int \exp\{Q(t)\} d\mu(t)} ; \quad t \in \Omega,
\]

(4.2)

where \( \Omega \) is the support of \( f(y) \), \( Q(\cdot) \) is a linear combination of functions constructed from the conditionals in (4.1), and \( \mu(\cdot) \) is an appropriate measure; in this article, \( \mu(\cdot) \) will be either Lebesgue or counting measure. While delineation of the neighborhood structure for \( Y \) defines a Markov random field and is important in identification of the function \( Q(\cdot) \) appearing in (4.2), nothing in our estimation procedure will make use of this fact. Thus, expression (4.2) may be considered to be quite general since any distribution may be expressed in this manner.
An important class of conditionally specified models are the exponential family auto-models of Besag (1974). For these models the conditional specifications (4.1) have the form

\[ f(y(s_i) | \{ y(s_j) : j \in D_i \}) = \exp \left[ A_i(\{ y(s_j) : j \in D_i \}) B_i(y(s_i)) - D_i(\{ y(s_j) : j \in D_i \}) + C_i(y(s_i)) \right], \]

where the \( A_i(\cdot) \) satisfy

\[ A_i(\{ y(s_j) : j \in D_i \}) = a_i + \sum_{j \in D_i} \eta_{i,j} B_j(y(s_j)), \]

with \( \eta_{i,j} = \eta_{j,i} \) for all \( i \) and \( j \). It turns out that \( Q(\cdot) \) in (2) may then be written as

\[ Q(y) = \sum_{1 \leq i \leq n} \left( a_i B_i(y(s_i)) + C_i(y(s_i)) \right) + \sum_{1 \leq i < j \leq n} \eta_{i,j} B_i(y(s_i)) B_j(y(s_j)), \quad (4.3) \]

where \( \eta_{i,j} = \eta_{j,i}, \eta_{i,i} = 0, \) and \( \eta_{j,j} = 0 \) if \( j \notin D_i \). An important feature of these exponential family auto-models is that the dependence among elements of \( Y \) is captured entirely in the dependence parameters \( \{ \eta_{i,j} : 1 \leq i < j \leq n \} \). Typically, one reduces the number of free parameters further by specifying, for example, \( a_i = \alpha \) and \( \eta_{i,j} = \eta \), for all \( i \) and \( j \). This will be the case for the example of Section 4.6, leaving two parameters, \( \alpha \) and \( \eta \), for which MLEs are desired.

4.3 The Overall Estimation Strategy

In this section we describe an overall strategy that may be used to find MLEs of parameters appearing in joint distributions of the form (4.2). To emphasize the role of such parameters, let \( \theta \equiv (\theta_1, \ldots, \theta_p)^T \in \Theta \subset \mathbb{R}^p \), and write (4.2) in the more specific form,

\[ f(y|\theta) = \frac{\exp \left\{ Q(y|\theta) \right\}}{\int \exp \left\{ Q(y|\theta) \right\} d\mu(\theta)} = \frac{1}{k(\theta)} \exp \left\{ Q(y|\theta) \right\}; \quad y \in \Omega, \ \theta \in \Theta. \quad (4.4) \]

From (4.4) it is easily seen that the log likelihood may be written as

\[ L(\theta) = Q(y|\theta) - \log \{ k(\theta) \}. \quad (4.5) \]

Let \( \hat{\theta} \in \Theta \) denote the maximizing value of (4.5), that is, the true MLE of \( \theta \). Our strategy is to construct a sequence of estimators \( \{ L_{\hat{\theta}(q)}(\theta(q)) : q = 1, \ldots \} \), such that \( L_{\hat{\theta}(q)}(\theta(q)) \) becomes a sufficiently precise estimator of \( L(\hat{\theta}) \) as \( q \) increases; in practice we desire \( q \) to be small (i.e., \( q = 2 \) or \( q = 3 \)). This sequence of estimators will be based on Monte Carlo estimation of \( L(\theta) \), following the concept of constrained maximization introduced by Geyer and Thompson (1992). For this purpose, let \( m(y|\lambda) \) be an importance distribution (density) such that \( \lambda \) does not depend on \( \theta \), and \( m(y|\lambda) \) dominates \( f(y|\theta) \).
in that \( m(y|\lambda) = 0 \Rightarrow f(y|\theta) = 0 \) (see Geyer, 1996). For any value of \( \theta \in \Theta \), an Monte Carlo estimator of \( k(\theta) \) is,

\[
k_{M(q)}(\theta) = \frac{1}{M(q)} \sum_{r=1}^{M(q)} \left[ \exp \{ Q(x_r|\theta) \}/m(x_r|\lambda^{(q)}) \right], \tag{4.6}
\]

where \( x_1, x_2, \ldots, x_{M(q)} \) is a sample of \( M(q) \) values from \( m(y|\lambda^{(q)}) \). A Monte Carlo estimator of the log likelihood (4.5) is then

\[
L_{M(q)}(\theta) = Q(y|\theta) - \log k_{M(q)}(\theta), \tag{4.7}
\]

and an estimator of the maximized log likelihood \( L(\hat{\theta}) \) is \( L_{M(q)}(\theta^{(q)}) \), where \( \theta^{(q)} \) is the value of \( \theta \in \Theta \) that maximizes \( L_{M(q)}(\theta) \).

Notice that, while (4.7) is a Monte Carlo estimator of \( L(\theta) \) over the entire parameter space, for a given Monte Carlo sample size \( M(q) \) it will be sufficiently precise for estimation purposes only for some subset of that space, \( \theta \in \Theta_{\lambda^{(q)}} \) say; the index \( \lambda^{(q)} \) indicates dependence of this subset on the parameters of the importance distribution \( m(y|\lambda^{(q)}) \). Thus, the value \( \theta^{(q)} \) that maximizes \( L_{M(q)}(\theta) \) will be a sufficiently precise estimator of \( \hat{\theta} \) only if \( \hat{\theta} \in \Theta_{\lambda^{(q)}} \), and hence the need for a sequence of such estimators, indexed here by \( q \). Because identification of the appropriate subset \( \Theta_{\lambda^{(q)}} \) is a difficult task \textit{a priori}, it is crucial to have available a means by which to assess how well \( L_{M(q)}(\theta) \) approximates \( L(\theta) \) for various values \( \theta \in \Theta \). In addition, construction of the sequence \( \{ L_{M(q)}(\theta^{(q)}) : q = 1, \ldots \} \) requires a way to maximize the Monte Carlo log likelihood (4.7). The remainder of this section deals with this second issue, maximization of a given \( L_{M(q)}(\theta) \). In Section 4.4 we introduce a way to select \( m(y|\lambda) \) such that the precision of (4.7) as an estimator of the true log likelihood (4.5) may be easily accomplished.

Our approach to maximization of (4.7) in \( \theta \) is simply to use a Newton-Raphson algorithm. This is facilitated by the fact that the first and second derivatives of (4.7) are equivalent to Monte Carlo estimators of the first and second derivatives of the true log likelihood (4.5). The first derivatives of \( L(\theta) \) may be written, for \( j = 1, \ldots, p \), as

\[
\frac{\partial L(\theta)}{\partial \theta_j} = \frac{\partial Q(y|\theta)}{\partial \theta_j} - E_\theta \left\{ \frac{\partial Q(Y|\theta)}{\partial \theta_j} \right\}, \tag{4.8}
\]

and the second derivatives may be written, for \( j, k = 1, \ldots, p \), as

\[
\frac{\partial^2 L(\theta)}{\partial \theta_j \partial \theta_k} = \frac{\partial^2 Q(y|\theta)}{\partial \theta_j \partial \theta_k} - E_\theta \left\{ \frac{\partial^2 Q(Y|\theta)}{\partial \theta_j \partial \theta_k} \right\} - E_\theta \left\{ \frac{\partial Q(Y|\theta)}{\partial \theta_j} \right\} \frac{\partial Q(Y|\theta)}{\partial \theta_k} + E_\theta \left\{ \frac{\partial Q(Y|\theta)}{\partial \theta_j} \right\} E_\theta \left\{ \frac{\partial Q(Y|\theta)}{\partial \theta_k} \right\}. \tag{4.9}
\]

Use of these derivatives in an iterative algorithm would require evaluation of integrals of the form

\[
E_\theta \left\{ h_{(j)}(Y, \theta) \right\} = \frac{\int_{\Omega} h_{(j)}(t, \theta) \exp(Q(t|\theta)) \, d\mu(t)}{\int_{\Omega} \exp(Q(t|\theta)) \, d\mu(t)}, \tag{4.10}
\]
where \( h_{(j)}(Y, \theta) \) is one of \( h_{(j)}(Y, \theta) = \partial Q(Y | \theta) / \partial \theta_j \), \( h_{(j,k)} = \{ \partial Q(Y | \theta) / \partial \theta_j \} \{ \partial Q(Y | \theta) / \partial \theta_k \} \). or \( h_{(j,k)}(Y, \theta) = \partial^2 Q(Y | \theta) / \partial \theta_j \partial \theta_k; j, k = 1, \ldots, p \). Estimators of these expected values, based on a Monte Carlo sample of size \( M^{(q)} \) from \( m(y | \lambda^{(q)}) \), are

\[
E_{M^{(q)}} \{ h_{(j)}(Y, \theta) \} = \frac{1}{M^{(q)} \cdot k_{M^{(q)}(\theta)}} \sum_{r=1}^{M^{(q)}} \left[ h_{(j)}(x_r, \theta) \exp \{ Q(x_r | \theta) / m(x_r | \lambda^{(q)}) \} \right].
\] (4.11)

Thus, derivatives of the Monte Carlo log likelihood (4.7) are given by expressions (4.8) and (4.9) with \( L(\theta) \) replaced by \( L_{M^{(q)}}(\theta) \) and expectations of the form (4.10) replaced by the corresponding expressions in (4.11).

Our overall estimation strategy consists of three 'levels' of iterative procedures. At the first level, a Monte Carlo sample is used to estimate the log likelihood and evaluate the derivatives of the Monte Carlo estimator. At the second level, a Newton-Raphson algorithm is used to maximize the Monte Carlo log likelihood constructed from the current Monte Carlo sample. Finally, at the third level, the essential concept underlying the constrained maximization procedure of Geyer and Thompson (1992) is employed, resulting in a sequence of Monte Carlo log likelihoods. This strategy may be stated as follows:

1. Choose a starting value \( \theta^{(0)} \).
2. For any current parameter value \( \theta^{(q-1)}; q \geq 1 \), choose a sampling density \( m(y | \lambda^{(q)}) \).
3. Draw a sample of size \( M^{(q)} \) from \( m(y | \lambda^{(q)}) \), assessing the error in Monte Carlo log likelihood \( L_{M^{(q)}}(\theta^{(q-1)}) \) to ensure sufficient precision. Increase \( M^{(q)} \) as needed.
4. Use the sample from \( m(y | \lambda^{(q)}) \) in (4.6) and (4.11) for evaluation of the Monte Carlo log likelihood (4.7) and its derivatives.
5. Update the parameter estimate through use of a Newton-Raphson algorithm, repeating step 4 as needed until convergence, always with the same sample from \( m(y | \lambda^{(q)}) \). Call the resulting estimate of the parameter from this Newton-Raphson cycle \( \theta^{(q)} \).
6. Assess the error in Monte Carlo estimation of the maximized log likelihood \( L(\theta^{(q)}) \). If this error has increased from that of \( L_{M^{(q)}}(\theta^{(q-1)}) \) in step 3, update \( q \) to \( q+1 \) and repeat, beginning at step 2.

Despite the use made of constrained maximization, our strategy differs from the approach of Geyer and Thompson (1992) and Geyer (1994). We approach maximization through the use of a direct Monte
Carlo estimator of the log likelihood, rather than through the use of a ratio of un-normalized log densities with a fixed parameter value in the denominator (cf., Geyer, 1994). That is, we estimate the normalizing constant of the joint density \( k(\theta) \) directly through (4.6), and make use of this estimate in other Monte Carlo evaluations of the form (4.11). Related to this is the fact that we assume that the importance distribution \( m(y|\lambda) \) is known completely and nothing in our overall strategy depends explicitly on the use of a Markov chain sampler. The importance distribution we will construct in Section 4.4 may be sampled from without the use of Markov chain methods, although the selection of \( m(y|\lambda) \) will depend in a critical way on a preliminary sample selected through use of a Markov chain.

4.4 Importance Sampling

In this section we propose a method for selecting an importance sampling distribution \( m(\cdot|\cdot) \) that is easy to sample from, appropriate for use with a cycle of Newton-Raphson iterations in the overall estimation strategy, and that allows determination of precision in Monte Carlo estimates of the log likelihood.

4.4.1 Selection of a Sampling Distribution

Consider a model specified through a set of conditional probability density functions as in (4.1). Typically, these conditional specifications will be in parameterized form, \( f(y(s_i)|\theta_i(y(s_j): j \in D_i)): i = 1, \ldots, n \). We will assume the positivity condition of Besag (1974) that, for \( y(s_i) \in \Omega_i, y \in \Omega = \Omega_1 \times \ldots \times \Omega_n \). We will not make a corresponding assumption about the \( \theta_i(\cdot) \), since construction of a joint distribution in the form of expression (4.4) may require additional constraints on allowable parameter values to ensure that the normalizing constant \( k(\theta) \) exists. For the purposes of this article, we will simply assume that the \( \theta_i(\cdot) \) appearing in the conditional probability density functions are compatible with \( \theta \in \Theta \) in the joint density. Given such a set of specified conditional density functions, it will be possible to produce a sample of size \( S \) from the joint density (4.4) through the use of a Gibbs Sampling algorithm. For \( i = 1, \ldots, n \), let \( \hat{\mu}_S(s_i) \) denote the sample mean of these values at location \( s_i \), and \( \hat{\sigma}_S^2(s_i) \) the corresponding sample variance. To construct a sampling distribution for use in the overall estimation strategy of Section 4.3, we develop a 'dummy' model for \( Y \) using the same set of locations \( \{s_i: i = 1, \ldots, n\} \) and the same support \( \Omega \). This dummy model takes the random variables \( Y(s_1), \ldots, Y(s_n) \) to be independent with marginal densities of the same form as the conditional functions specified in the actual model. The dummy model is then given by the collection of marginals.
\{f(y(s_i)|\lambda_i) : i = 1, \ldots, n\}. The parameters \lambda_i; i = 1, \ldots, n are chosen such that

\[ E_\lambda \{Y(s_i)\} = \mu_\lambda(s_i) \quad \text{and} \quad \text{var}_\lambda \{Y(s_i)\} = \sigma^2_\lambda(s_i). \]

We then construct a sampling distribution as

\[ m(y|\lambda) = \prod_{i=1}^{n} f(y(s_i)|\lambda_i); \quad y \in \Omega. \quad (4.12) \]

As pointed out by Geyer (1994) it is not necessary that \lambda have the same parameter space as does \theta, as long as \( m(y|\lambda) \) dominates \( f(y|\theta) \), although this will often be the case simply by default. A sample \((x_1(s_1), \ldots, x_1(s_n))^T, \ldots, (x_M(s_1), \ldots, x_M(s_n))^T \) generated from \( m(y|\lambda) \) will not reflect dependence among its values relative to the dummy 'independence' model. However, because the mean and variance of each \( f(y(s_i)|\lambda_i) \); \( i = 1, \ldots, n \) have been matched with a sample from the actual dependence model, these values will reflect dependence among the \{ \( Y(s_i) : i = 1, \ldots, n \} \) under the actual model of interest, up to the effect of that dependence on the first two moments. In this way, we have constructed an importance sampling distribution for use in the overall estimation strategy of Section 4.3 that is easily sampled from, because it consists of independent components, and yet captures the essential impact of dependence on the pattern of realizations from the actual model. The independence of random variables that have a joint distribution given by the dummy model also provides the necessary theory for evaluation of the precision of the Monte Carlo log likelihood (4.7) as an estimator of the true log likelihood (4.5).

### 4.4.2 Precision in Monte Carlo Log Likelihood

To assess precision of the Monte Carlo log likelihood for a given value of the parameter \( \theta \), we make use of the fact that the sampling distribution \( m(\cdot|\cdot) \) in (4.12) was constructed as a product of marginal probability density functions. In the first part of this section, we suppress dependence of the Monte Carlo quantities on the current iteration of the overall estimation strategy and write \( M \equiv M^{(\theta)} \) and \( \lambda \equiv \lambda^{(\theta)} \) throughout. Write the importance density (4.12) as

\[ m(y|\lambda) = \exp \left[ Q^*(y|\lambda) - \log \{k^* (\lambda)\} \right]; \quad y \in \Omega, \]

where

\[ Q^*(y|\lambda) \equiv \sum_{i=1}^{n} \log \{ g(y(s_i)|\lambda_i) \} \quad \text{and} \quad k^* (\lambda) \equiv \int \exp \{ Q^*(t|\lambda) \} \, d\mu(t), \]

with \( g(y(s_i)|\lambda_i) \propto f(y(s_i)|\lambda_i) \). If the conditional density functions of the actual model in (4.4) have been chosen to correspond to a common distributional form (e.g., Poisson or Gaussian), then we would
typically take \( g(y(s_i)|\lambda_i) = f(y(s_i)|\lambda_i) \), and have \( k^*(\lambda) = 1 \). Alternatively, it may be more convenient for some exponential family auto-models to take \( g(y(s_i)|\lambda_i) = \exp\{\lambda_i B_i(y(s_i)) + C_i(y(s_i))\} \) so that \( Q^* \) is equal to the first summation of terms in (4.3). Notice that in this case we have only one free parameter in each component of \( m(y|\lambda) \), stemming from the fact that the auto-models of Besag (1974) are essentially restricted to models for natural exponential families.

For random variables \( \{X_r : r = 1, \ldots, M\} \) with density or mass function \( m(y|\lambda) \), define the random variables

\[
D(X_r, \theta, \lambda) = \exp\left[Q(X_r, \theta) - (Q(X_r|\lambda) - \log\{k^*(\lambda)\})\right].
\]

Recall here that \( Q(\cdot|\cdot) \) is the log of the unnormalized joint density of the actual model under analysis, while \( Q^*(\cdot|\cdot) \) is defined from the dummy model as given above. Monte Carlo estimators (4.6) and (4.7) of the normalizing constant and log likelihood may then be written as,

\[
k_M(\theta) = \frac{1}{M} \sum_{r=1}^{M} D(X_r, \theta, \lambda),
\]

and,

\[
L_M(\theta) = Q(y|\theta) - \log\left[\frac{1}{M} \sum_{r=1}^{M} D(X_r, \theta, \lambda)\right].
\]

The terms \( D(X_r, \theta, \lambda) \) inside the summations of (4.13) and (4.14) are \( iid \) random variables that have expected values, with respect to \( m(y|\lambda) \), of

\[
E_{\lambda}\{D(X_r, \theta, \lambda)\} = \int_{\Omega} \exp\{Q(t|\theta)\} d\mu(t) = k(\theta).
\]

The \( iid \) structure of \( X_1, X_2, \ldots, \) allows application of the central limit theorem in (4.13) to give,

\[
M^{1/2}[k_M(\theta) - k(\theta)] \xrightarrow{d} N(0, V_1(\theta)),
\]

and an additional log transformation yields, from (14),

\[
M^{1/2}[L_M(\theta) - L(\theta)] \xrightarrow{d} N(0, V_2(\theta)),
\]

where \( V_2(\theta) = V_1(\theta)/k^2(\theta) \). In equations (4.15) and (4.16), \( V_1(\theta) \) and \( V_2(\theta) \) may easily be shown to exist, although their explicit form will generally be unavailable, and will require estimation in practice. We would like to make use of these results to assess (1) the precision in Monte Carlo estimates \( L_M(\theta) \) and (2) the possible need to select a new importance sampling distribution at step 6 of the overall estimation strategy.

In principle, one could assess the precision of \( L_M(\theta) \) through the use of (4.16) for samples drawn from one importance distribution and any value of \( \theta \in \Theta \). Thus, returning to the use of the superscript
(q), at the beginning of a Newton-Raphson cycle in step 5 of the overall estimation procedure (with current parameter $\theta^{(q-1)}$) we could select a sample size $M^{(q)}$ based on the resulting level of precision in $L_{M^{(q)}}(\theta^{(q-1)})$, and base decisions of convergence on whether that level of precision had changed by the end of the cycle, at $L_{M^{(q)}}(\theta^{(q)})$; recall that $\theta^{(q)}$ maximizes $L_{M^{(q)}}(\theta)$. Unfortunately, two aspects of the interplay between the Monte Carlo sample size, the importance sampling distribution used, and the asymptotics of (4.16) vitiate the usefulness of this approach. First, in the actual analysis of data, $V_2(\theta)$ must be estimated as a function of $k(\theta)$, which will itself be estimated as $k_{M^{(q)}}(\theta)$. Thus, use of (4.16) to estimate Monte Carlo error for the purpose of determining whether or not $M$ is sufficiently large involves a circular argument. That is, the Monte Carlo error in $L_{M^{(q)}}(\theta)$ can only be estimated on the basis of (4.16) if we are already assured of a sufficiently precise estimate of $k(\theta)$, in which case the issue of sample size determination is moot. The second problematic aspect in practical use of (4.16) is that, for a given importance distribution $m(y|\lambda^{(q)})$, the Monte Carlo sample size $M^{(q)}$ required for this result to provide a reasonable approximation to the error in $L_{M^{(q)}}(\theta^{(q-1)})$ may no longer be adequate for a similar assessment of $L_{M^{(q)}}(\theta^{(q)})$. Nevertheless, we believe that the combined use of (4.15) and (4.16) can form the basis for a practical assessment of Monte Carlo error, and convergence in the overall estimation strategy, under the following heuristic argument.

Under the assumption that a Monte Carlo sample size $M^{(q)}$ is large enough for (4.16) to provide an adequate indication of the error in $L_{M^{(q)}}(\theta^{(q-1)})$, we can estimate $V_1(\theta^{(q-1)})$ as the sample variance of $D(x_r, \theta^{(q-1)}, \lambda^{(q)})$, where $x_1, \ldots, x_{M^{(q)}}$ are sampled values from the importance distribution $m(y|\lambda^{(q)})$. Call this estimate $\hat{V}_1(\theta^{(q-1)})$. Then $V_2(\theta^{(q-1)})$ is estimated as $\hat{V}_2(\theta^{(q-1)}) = \hat{V}_1(\theta^{(q-1)})/k_{M^{(q)}}^2(\theta^{(q-1)})$, and an $(1-\alpha) \times 100\%$ interval estimate of the Monte Carlo error $L_{M^{(q)}}(\theta^{(q-1)}) - L(\theta^{(q-1)})$ formed as $\pm z_{\alpha/2}[\hat{V}_2(\theta^{(q-1)})/M^{(q)}]^{1/2}$. This provides an estimate of the Monte Carlo error in estimation of the log likelihood at the beginning of a Newton-Raphson cycle in the overall estimation procedure, under the starting assumption which will be addressed in Section 4.4.4.

Notice that the estimator $\hat{V}_1(\theta)$ involves only the computation of a sample variance of simulated values and thus does not suffer the same deficiency of being calculated as a function of $k_M(\theta)$ as does $\hat{V}_2(\theta)$. Thus, an increase in estimated values from $\hat{V}_1(\theta^{(q-1)})$ to $\hat{V}_1(\theta^{(q)})$ indicates a deterioration in precision over the course of a Newton-Raphson cycle, and the need to select a new importance distribution for an additional cycle. What is required for this use of estimated values of $V_1(\theta)$ is a meaningful scale against which to judge the magnitude of changes. This is not provided automatically, since $V_1(\theta)$ is the (asymptotic) variance of a normalizing constant alone. Now, consider a small neighborhood $\mathcal{N}(\hat{\theta})$ of the true MLE $\hat{\theta}$, so that $k(\theta) \approx k(\hat{\theta})$ for any $\theta \in \mathcal{N}(\hat{\theta})$. Then the change in $V_2(\theta)$ over two parameter
values \( \theta^*, \theta^e \in \mathcal{N}(\hat{\theta}) \) is such that

\[
\frac{|V_2(\theta^*) - V_2(\theta^e)|}{V_2(\theta^e)} \approx \frac{|V_1(\theta^*) - V_1(\theta^e)|}{V_1(\theta^e)}.
\]

Note that the assumption inherent to this expression is that \( \mathcal{N}(\hat{\theta}) \) is small enough so that the normalizing constant is approximately equavalued within this neighborhood, but not necessarily so small that the same is true of \( V_2(\theta) \). We judge the change in precision of \( L_{M(e)}(\theta) \) from evaluation at \( \theta^{(q-1)} \) to evaluation at \( \theta^{(q)} \) by estimating the above quantity as

\[
\tilde{V}_1(\theta^{(q-1)}, \theta^{(q)}) \equiv \frac{|\tilde{V}_1(\theta^{(q)}) - \tilde{V}_1(\theta^{(q-1)})|}{\tilde{V}_1(\theta^{(q-1)})}.
\]

While there is no guarantee that \( \tilde{V}_1(\theta^{(q-1)}, \theta^{(q)}) \) accurately reflects the proportional change in Monte Carlo error of \( L_M(\theta) \) at all stages of the estimation procedure, if it is 'large', (as defined in Section 4.4.3), then it cannot be that the overall strategy has resulted in convergence. In this case, it will be required to select a new importance distribution and conduct an additional cycle of Newton-Raphson iterations.

### 4.4.3 Assessing Convergence in Practice

The overall estimation strategy has been formed on the basis of constructing a sequence of approximations to the log likelihood at various points in the parameter space \( \Theta \). At iteration \( q \) of the overall procedure, we call \( \theta^{(q-1)} \) the 'point of origin' of the current approximation \( L_{M(e)}(\theta) \), and \( \theta^{(q)} \) the 'point of maximization' of the approximation. As indicated in Section 4.4.2, we continue to form new approximations so long as we have an indication, based on (4.17), that the precision of the approximation may have changed between its point of origin and its point of maximization. To make this prescription more concrete, we propose that, at the end of step 5 in the overall estimation procedure, a new importance distribution be chosen (using the approach of Section 4.4.1) if \( \tilde{V}_1(\theta^{(q-1)}, \theta^{(q)}) \geq \delta \). While arbitrary, our suggested value for the threshold is \( \delta = 0.10 \), based on the rationale that, if \( V_2(\theta^q) = V_2(\theta^{(q-1)}(1 + \delta)) \), then an interval formed as \( \pm z_{\alpha/2}[V_2(\theta^{(q)})/M^{(q-1)}]^{1/2} \) will be about 1.1 times as long as the corresponding interval formed with \( V_2(\theta^{(q-1)}) \).

If \( \tilde{V}_1(\theta^{(q-1)}, \theta^{(q)}) < \delta \), then it would appear that an additional iteration of the estimation procedure is not needed. But, because this is only indirect evidence of convergence in the overall algorithm, we deem it alone to be an insufficient criterion. Therefore, we make use of two more familiar quantities as checks on convergence, namely \( |L_{M(e)}(\theta^{(q-1)}) - L_{M(e)}(\theta^{(q)})| \), and \( ||\theta^{(q-1)} - \theta^{(q)}|| \), where \( ||a|| \) is the Euclidean norm of a vector \( a \). These quantities are used only as additional checks on convergence,
rather than as original convergence criteria, because small values may result by chance unless Monte Carlo precision has remained stable throughout an entire Newton-Raphson cycle (i.e., the maximization of a given approximation to the log likelihood).

4.4.4 Determination of Baseline Monte Carlo Error

The estimation strategy proposed in this article depends critically on the assumption that a suitable level of precision has been attained in each approximation of the log likelihood at its point of origin. We call this the 'baseline' level of precision for each approximation formed. If the baseline level of Monte Carlo precision for a given approximation \( L_M(\theta) \) is insufficient in the first place, numerical instability in estimated quantities will render even the most careful efforts to assess convergence of the overall estimation procedure futile. As indicated in Section 4.4.2, we assess the baseline level of precision for the current approximation through use of the estimated interval \( \pm z_{\alpha/2}[\hat{V}_2(\theta^{(r-1)})/M(0)]^{1/2} \). The adequacy of this interval as an estimate of the true Monte Carlo error depends on the assumption that \( M(0) \) is large enough to provide sufficient precision in estimation of \( k(\theta) \) for estimation of \( V_2(\theta) \).

Our approach to providing justification for a selected Monte Carlo sample size \( M(0) \) depends firstly on beginning the estimation procedure in step 1 of Section 4.3 with a value \( \theta^{(0)} \) which is itself a solution to a log likelihood approximation. This is discussed in Section 4.5. Then, for an initial Monte Carlo sample size, \( M_1^{(1)} \) say, an \((1 - \alpha) \times 100\%\) interval estimate of Monte Carlo error is computed. The sample size is increased to \( M_2^{(1)} \) and an additional interval computed, where the estimate of \( V_2(\theta^{(0)}) \) is based on the \( M_2^{(1)} \) samples but the interval on the initial sample size \( M_1^{(1)} \). The process is repeated until interval length shows no appreciable decrease, and that sample size is selected a \( M^{(1)} \).

4.5 Starting Value

In this section we discuss selection of a starting value \( \theta^{(0)} \), the first step of the overall procedure described in Section 4.3. Finding a suitable starting value \( \theta^{(0)} \) is important in the overall strategy to reduce the number of repetitions of the entire procedure (steps 2 through 6) required to locate the MLE of \( \theta \). Selection of the importance sampling distribution in step 2 of the procedure constitutes the major burden in our strategy, since this is the step that requires the use of a Markov chain sampler. Thus, one should be willing to expend considerable effort in selection of a starting value to reduce the total effort required in estimation. Under the approach proposed here, \( \theta^{(0)} \) is itself an approximate solution to the total estimation problem, and requires the use of two nested iterative procedures.
Consider the exact log likelihood of equation (4.5), in which

\[ k(\theta) \equiv \int_{-\infty}^{\infty} \exp\{Q(t|\theta)\} \, dt \]

is the normalizing constant of the joint density or mass function (4.4). A Laplace approximation (e.g., Tierney and Kadane 1986) of this integral is,

\[ k(\theta) \approx (2\pi)^{n/2} \left( \det \Sigma \right)^{1/2} \exp\{Q(\hat{\theta}|\theta)\}, \]

where \( \hat{\theta} \) is the mode of \( Q(y|\theta); \, y \in \Omega \), considered as a function of \( y \) for fixed \( \theta \), and \( \Sigma \) is the inverse negative Hessian of \( Q(y|\theta) \) evaluated at \( \hat{\theta} \). An approximate solution to the estimation problem then results from maximizing in \( \theta \) the quantity

\[ \bar{L}(\theta) = Q(\hat{\theta}|\theta) - \log \left( (2\pi)^{n/2} \left( \det \Sigma \right)^{1/2} \right) \exp\left\{ g(y^*|\theta) \right\}, \]  
(4.18)

rather than the exact log likelihood (4.5). Because the complexity of \( \bar{L}(\theta) \) will be determined by the particular model under consideration, there is not a generally applicable prescription for its maximization. If \( \theta \) is of low dimension, direct search algorithms appear to perform adequately. If \( \theta \) is of higher dimension, profiling methods may be applied or, possibly, an algorithm depending on evaluation of derivatives of \( Q(y|\theta) \) could be employed. In this latter case, the Laplace method of Tierney and Kadane (1986) would be applied to each of the derivatives necessary for an iterative algorithm.

Regardless of the technique employed for maximization of (4.18), a nested procedure is required for determination of \( \hat{\theta} \) for a given value of \( \theta \). That is, \( \hat{\theta} \), being the mode of \( Q(y|\theta); \, y \in \Omega \), is a function or transformation of \( \theta \). At least for exponential family models, such as the auto-models of Section 4.2, we have found standard iterative algorithms depending on second derivative information appropriate for finding this value. To determine \( \hat{\theta} \) for an exponential family auto-model, \( \theta \equiv (\{\alpha_i : i = 1, \ldots, n\}, \{\eta_{ij} : 1 \leq i < j \leq n\}) \), and one may make use of the derivatives

\[ \frac{\partial Q(y)}{\partial y(s_i)} = \alpha_i \frac{\partial B_i(y(s_i))}{\partial y(s_i)} + \frac{\partial C_i(y(s_i))}{\partial y(s_i)} + \sum_{j \in D_i} \eta_{ij} \frac{\partial B_j(y(s_j))}{\partial y(s_i)} \]

\[ \frac{\partial^2 Q(y)}{\partial y(s_i) \partial y(s_j)} = \eta_{ij} \frac{\partial B_i(y(s_i)) \partial B_j(y(s_j))}{\partial y(s_i)} \partial y(s_j) \]

and

\[ \frac{\partial^2 Q(y)}{\partial y^2(s_i)} = \frac{\partial^2 C_i(y(s_i))}{\partial y^2(s_i)} \]

where, in all of these expressions, \( \eta_{ij} = 0 \) if \( j \notin D_i \). These expressions typically are not complex, and the matrix \( \Sigma \) tends to be easily inverted despite being of large dimension \((n \times n)\). In fact, in many applications, the dependence sets \( D_i; \, i = 1, \ldots, n \), contain less than 10 index values so that many of the entries in \( \Sigma \) are zero. Thus, for a fixed value of \( \theta \), the mode of \( Q(y|\theta) \) in \( y \) is easily located.
The entire process for finding a starting value \( \theta^{(0)} \) then consists of a nested iterative algorithm. For given \( \theta \), maximization in \( y \) produces \( \hat{y}_\theta \), which is then used for evaluation of \( \hat{L}(\theta) \) as given in (4.18). We reiterate that the goal in starting value selection is to find a 'good' value \( \theta^{(0)} \) for use in the larger estimation process. A value that is 'good' is one that leads to steps 2 through 6 of the overall strategy given in Section 4.3 being repeated only a small number of times (e.g., twice).

### 4.6 Example Using a Winsorized Poisson Model

As an application of the estimation strategy proposed in this article, we analyzed a data set containing counts of a particular tree species in spatial sampling plots in the northeastern United States. The data were collected as part of the Environmental Monitoring and Assessment Program (EMAP) conducted by the U.S. Environmental Protection Agency. Each sampling location contained 4 sub-plots which were combined for this analysis. A complete description of the sampling methodology used is contained in Tallent-Halsell (1994). We selected sweet (or cherry) birch (\( \text{Betula lenta} \)) as the species of interest because this species is somewhat habitat specific, requiring moist woodland conditions to flourish. It is not a dominant member of most woodland communities, but occurs from southern Maine and southwestern Quebec to Delaware and Kentucky, also appearing in mountainous regions as far south as Georgia (Gleason and Cronquist, 1963). Seeds are contained in fruiting catkins so that, along with moderately specific habitat requirements, it would be reasonable to suppose that counts of this species within small spatial sampling plots might exhibit positive spatial dependence. A map of the sampling locations from which data were collected is presented in Figure 4.1, and the actual counts of sweet birch trees for each location are given in Table 4.1.

<table>
<thead>
<tr>
<th>Location No.</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9</th>
<th>10</th>
<th>11</th>
</tr>
</thead>
<tbody>
<tr>
<td>Tree Count</td>
<td>1</td>
<td>3</td>
<td>5</td>
<td>3</td>
<td>2</td>
<td>8</td>
<td>1</td>
<td>3</td>
<td>2</td>
<td>2</td>
<td>1</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Location No.</th>
<th>12</th>
<th>13</th>
<th>14</th>
<th>15</th>
<th>16</th>
<th>17</th>
<th>18</th>
<th>19</th>
<th>20</th>
<th>21</th>
<th>22</th>
</tr>
</thead>
<tbody>
<tr>
<td>Tree Count</td>
<td>2</td>
<td>1</td>
<td>7</td>
<td>5</td>
<td>2</td>
<td>3</td>
<td>4</td>
<td>2</td>
<td>8</td>
<td>4</td>
<td>1</td>
</tr>
</tbody>
</table>

While situations such as this are often modeled as a spatial point process, our goal was to describe the dependence structure with an appropriate lattice model. For this purpose, we utilized the Winsorized Poisson auto-model of Kaiser and Cressie (1997). In this example it was not clear how the underlying Markov random field (i.e., the neighborhood structure) should be defined. We used the following data-driven procedure. Let \( u_i \) denote the longitude and \( v_i \) the latitude of each sampling location shown in
Figure 4.1  Latitude and longitude for sampling locations used in the analysis of sweet birch tree counts
Figure 4.1, and define the location indices \( s_i \equiv (u_i, v_i)^T; \ i = 1, \ldots, n \), where in this example \( n = 22 \).

The neighborhood structure was defined from physical proximity of sampling locations as,

\[
h \in D_i \text{ if } \|s_h - s_i\| \leq \kappa,
\]

where \( \kappa \) is a selected constant, \( \|x\| \) is the Euclidean norm of \( x \) and \( D_i \) is the dependence index set of location \( s_i \), as defined in Section 4.2. With this definition of the form of neighborhoods, \( \kappa \) was essentially treated as an additional parameter in the model. A finite number of neighborhood structures (i.e., Markov random fields) are possible over a range of values of \( \kappa \), and estimation was conducted for each value of \( \kappa \) between 28 km and 56 km at which the neighborhood index sets changed. The 'final' model was selected based on the value of \( \kappa \) for which the (estimated) maximized log likelihood was the greatest. Distances of less than 28 km for \( \kappa \) resulted in essentially no neighborhood structure among the sampled locations, and distances greater than 56 km resulted in all but a few locations being considered neighbors of all other locations. The neighborhood index sets resulting from the possible choices of \( \kappa \) are given in Table 4.2.

To formulate a Winsorized Poisson auto-model, let \( \{Y(s_i) : i = 1, \ldots, n\} \) be random variables connected with the number of sweet birch trees observed at each location, and specify the conditional probability mass functions

\[
f(y(s_i) | (y(s_j) : s_j \in N_i)) = \exp\left[A_i((y(s_j) : s_j \in N_i)) - H_i((y(s_j) : s_j \in N_i)) - \log(y(s_i))\right]; \ i = 1, \ldots, n,
\]

where each of these conditional functions has the support set \( \{0, 1, \ldots, R\} \) for some arbitrary large integer \( R \). We assume the positivity condition of Besag (1974), that the support of the joint probability mass function, denoted as \( \Omega \) in equation (4.2), is the \( n \)-fold Cartesian product of the set \( \{0, 1, \ldots, R\} \).

In the conditional specifications (4.19) we take,

\[
H_i((y(s_j) : s_j \in N_i)) = \begin{cases} 
\exp\{A_i((y(s_j) : s_j \in N_i))\} & \text{if } y(s_i) \leq R - 1 \\
\exp\{A_i((y(s_j) : s_j \in N_i))\} - \psi_i & \text{if } y(s_i) = R,
\end{cases}
\]

for some \( 0 < \psi_i < \exp\{A_i((y(s_j) : s_j \in N_i))\} \), and,

\[
A_i((y(s_j) : s_j \in D_i)) = a_i + \sum_{j \in D_i} \eta_{i,j} y(s_j).
\]

The Winsorization value \( R \) must be selected in the modeling process. The average count in the data of Table 4.1 is 3.2 and, following Kaiser and Cressie (1997), we set \( R = 12 \), greater than 3 times the average count. The values \( \psi_i; i = 1, \ldots, n \) appearing in the functions \( H_i(\cdot) \) arise from an application
Table 4.2 Neighborhoods defined for sampling locations of Figure 4.1 using various values of the threshold distance $\kappa$

<table>
<thead>
<tr>
<th>Location</th>
<th>$\kappa = 28$</th>
<th>$\kappa = 29$</th>
<th>$\kappa = 47$</th>
<th>$\kappa = 48$</th>
<th>$\kappa = 49$</th>
<th>$\kappa = 55$</th>
<th>$\kappa = 56$</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>3</td>
<td>3</td>
<td>3</td>
<td>3</td>
<td>3</td>
<td>3</td>
<td>2,3</td>
</tr>
<tr>
<td>2</td>
<td>3</td>
<td>3</td>
<td>3</td>
<td>3</td>
<td>3,7</td>
<td>3</td>
<td>3</td>
</tr>
<tr>
<td>3</td>
<td>1</td>
<td>1,2</td>
<td>1,2,11</td>
<td>1,2,11</td>
<td>1,2,9,11</td>
<td>1,2,9,11</td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>5,6</td>
<td>5,6</td>
<td>5,6,8</td>
<td>5,6,8</td>
<td>5,6,7,8</td>
<td>5,6,7,8</td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>4,6,7,8</td>
<td>4,6,7,8</td>
<td>4,6,7,8</td>
<td>4,6,7,8</td>
<td>4,6,7,8</td>
<td>4,6,7,8</td>
<td>2,4,6,7,8</td>
</tr>
<tr>
<td>6</td>
<td>5,8</td>
<td>4,5,8</td>
<td>4,5,7,8</td>
<td>4,5,7,8</td>
<td>4,5,7,8</td>
<td>4,5,7,8</td>
<td>4,5,7,8</td>
</tr>
<tr>
<td>7</td>
<td>8</td>
<td>5,8</td>
<td>5,6,8,16</td>
<td>5,6,8,16</td>
<td>2,5,6,8,16</td>
<td>2,4,5,6,8,16</td>
<td>2,4,5,6,8,16</td>
</tr>
<tr>
<td>8</td>
<td>6,7</td>
<td>5,6,7</td>
<td>5,6,7</td>
<td>4,5,6,7</td>
<td>4,5,6,7,16</td>
<td>4,5,6,7,16</td>
<td></td>
</tr>
<tr>
<td>9</td>
<td>10,11</td>
<td>9,10</td>
<td>9,10</td>
<td>9,10</td>
<td>3,9,10</td>
<td>3,9,10</td>
<td>3,9,10</td>
</tr>
<tr>
<td>10</td>
<td>9,11</td>
<td>9,11</td>
<td>9,11</td>
<td>9,11</td>
<td>9,11</td>
<td>9,11</td>
<td>9,11,12</td>
</tr>
<tr>
<td>11</td>
<td>9,10</td>
<td>9,10</td>
<td>9,10</td>
<td>3,9,10</td>
<td>3,9,10,12</td>
<td>3,9,10,12</td>
<td>3,9,10,12</td>
</tr>
<tr>
<td>12</td>
<td>13,13</td>
<td>13,16</td>
<td>13,16</td>
<td>13,16</td>
<td>11,13,16</td>
<td>11,13,15,16</td>
<td>10,11,13,15,16</td>
</tr>
<tr>
<td>13</td>
<td>12,16</td>
<td>12,16</td>
<td>12,16,15,16</td>
<td>12,15,16</td>
<td>12,15,16</td>
<td>12,14,15,16</td>
<td>12,14,15,16</td>
</tr>
<tr>
<td>14</td>
<td>15</td>
<td>15</td>
<td>15,20</td>
<td>15,20</td>
<td>15,20</td>
<td>13,15,20</td>
<td>13,15,20</td>
</tr>
<tr>
<td>15</td>
<td>14</td>
<td>13,14</td>
<td>13,14</td>
<td>13,14</td>
<td>12,13,14,20</td>
<td>12,13,14,20</td>
<td>12,13,14,20</td>
</tr>
<tr>
<td>16</td>
<td>13,13</td>
<td>7,12,13</td>
<td>7,12,13</td>
<td>7,12,13</td>
<td>7,8,12,13</td>
<td>7,8,12,13</td>
<td>7,8,12,13</td>
</tr>
<tr>
<td>17</td>
<td>19</td>
<td>19</td>
<td>19</td>
<td>19</td>
<td>18,19</td>
<td>18,19</td>
<td>18,19</td>
</tr>
<tr>
<td>18</td>
<td>19</td>
<td>19</td>
<td>19</td>
<td>19</td>
<td>18,19</td>
<td>17,19</td>
<td>18,19</td>
</tr>
<tr>
<td>19</td>
<td>19</td>
<td>18</td>
<td>17,18</td>
<td>17,18</td>
<td>17,18</td>
<td>17,18</td>
<td>17,18</td>
</tr>
<tr>
<td>20</td>
<td>21</td>
<td>21</td>
<td>14,21</td>
<td>14,21</td>
<td>14,15,21</td>
<td>14,15,21</td>
<td>14,15,21</td>
</tr>
<tr>
<td>21</td>
<td>20</td>
<td>20</td>
<td>20</td>
<td>20</td>
<td>20</td>
<td>20</td>
<td>20</td>
</tr>
</tbody>
</table>
of Taylor’s formula in the Winsorization process and, in contrast to $R$, have no effect on the joint distribution for $Y$. That is, it is only the existence of the $\psi_i$ that is important, not their actual values. As shown by Kaiser and Cressie (1997), the joint probability mass function for $Y$ may be written in the form (4.2) with

$$Q(y) = \sum_{1 \leq i \leq n} [\alpha_i y(s_i) - \log(y(s_i)!)] + \sum_{1 \leq i < j \leq n} \eta_{i,j} y(s_i) y(s_j),$$

(4.20)

where $\eta_{i,j} = \eta_{j,i}$, $\eta_{i,i} = 0$, and $\eta_{i,j} = 0$ if $j \notin D_i$. In (4.20), the parameters $\alpha_i$ and $\eta_{i,j}$; $i,j = 1,\ldots,n$ may assume any real values. For analysis of the sweet birch tree counts we took $\alpha_i = \alpha_i = 1,\ldots,n$ and $\eta_{i,j} = \eta_i$; $1 \leq i < j \leq n$. Parameterized in this way, values of $\eta < 0$ correspond to negative spatial dependence, while values of $\eta > 0$ correspond to positive spatial dependence.

The maximum likelihood estimate of $\theta = (\alpha, \eta)^T$ was found for the data of Table 4.1 with each possible value of $\kappa$ from Table 4.2 by using the overall estimation strategy of Section 4.3. Importance sampling distributions were constructed as prescribed in Section 4.4.1 from a set of $n$ Poisson distributions, and questions of Monte Carlo precision and convergence were dealt with following Sections 4.4.2 through 4.4.4. The resulting estimates are summarized in Table 4.3, resulting in values for a final model of $K = 48$, $\alpha = 0.9765$ and $\eta = 0.0221$.

<table>
<thead>
<tr>
<th>$\kappa$</th>
<th>MLE $\theta = (\alpha, \eta)$</th>
<th>$L_M(\theta)$</th>
</tr>
</thead>
<tbody>
<tr>
<td>28</td>
<td>(1.2130, 0.0193)</td>
<td>-45.8894</td>
</tr>
<tr>
<td>29</td>
<td>(1.1009, 0.0114)</td>
<td>-45.9076</td>
</tr>
<tr>
<td>47</td>
<td>(1.1942, 0.0057)</td>
<td>-45.9447</td>
</tr>
<tr>
<td>48</td>
<td>(0.9765, 0.0221)</td>
<td>-45.5611</td>
</tr>
<tr>
<td>49</td>
<td>(0.8827, 0.0387)</td>
<td>-45.8583</td>
</tr>
<tr>
<td>55</td>
<td>(0.9815, 0.0163)</td>
<td>-45.6629</td>
</tr>
<tr>
<td>56</td>
<td>(1.0653, 0.0157)</td>
<td>-45.9029</td>
</tr>
</tbody>
</table>

The negative inverse Hessian matrix at convergence was

$$
\begin{pmatrix}
0.002160 & -0.000184 \\
-0.000184 & 0.000021
\end{pmatrix}
$$

Although standard maximum likelihood theory does not apply here, a rough guide as to whether the estimate $\eta = 0.0221$ provides evidence of positive spatial dependence might be the usual Wald statistic used with independent data, $0.0221/\sqrt{0.000021} = 4.829$. We would conclude from this value that positive spatial dependence is exhibited in these data.
Because the primary purpose of this example is to illustrate the estimation strategy proposed in this article, a more detailed examination of the estimation history for $\kappa = 48$ is provided here and in Table 4.4.

A starting value was determined from the Laplace approximation method of Section 4.5 as $\theta^{(0)} = (0.9172, 0.0294)^T$. Using this value, a sample of size 100,000 was generated from the model, using the conditional specifications (4.19) in a Gibbs algorithm. The $n = 22$ marginal means from this sample were used to select values for $\lambda_i; i = 1, \ldots, 22$ and a sampling distribution constructed as in equation (4.12) with each $f(y(s_i)|\lambda_i)$ being a Poisson probability mass function. Assessment of baseline Monte Carlo error as described in Section 4.4.4 resulted in an initial Monte Carlo sample size of $M^{(1)} = 30,000$, which yielded $\hat{V}_2(\theta^{(0)}) = 0.7510$ and an 95% interval estimate of Monte Carlo error, $[L_{M^{(1)}}(\theta^{(0)}) - L(\theta^{(0)})]$, of $\pm 0.0098$. Additional cycles of Newton-Raphson iterations were conducted using this same Monte Carlo sample size of 30,000. The first cycle of Newton-Raphson then began with $\hat{V}_1(\theta^{(0)}) = -45.6234$ and, after 4 iterations, produced the updated parameter value $\theta^{(1)} = (0.9742, 0.0224)^T$. The value of $\hat{V}_1(\theta^{(0)}, \theta^{(1)})$ was 28.37, indicating the need to select a new importance sampling distribution and enter a second cycle of Newton-Raphson iterations. To illustrate the need for basing this decision on the value of $\hat{V}_1(\theta^{(0)}, \theta^{(1)})$, rather than change in the estimated value of $V_2(\theta)$, an 95% interval estimate of $[L_{M^{(1)}}(\theta^{(1)}) - L(\theta^{(1)})]$, based on the value $\hat{V}_2(\theta^{(1)})$ produced at the end of the first cycle (i.e., using the original Monte Carlo sample), was $\pm 0.0081$, actually smaller than the beginning interval for this cycle of $\pm 0.0098$. Thus, if change in Monte Carlo precision were based on such intervals instead of the more reliable value of $\hat{V}_1(\theta^{(0)}, \theta^{(1)})$, we would have erroneously concluded that the precisions of $L_{M^{(1)}}(\theta^{(0)})$ and $L_{M^{(1)}}(\theta^{(1)})$ were similar. The argument of Section 4.4.2 preceding equation (4.17) shows that, with $\hat{V}_1(\theta^{(0)}, \theta^{(1)}) = 28.37$, this cannot have been the case. Thus, although $L_{M^{(1)}}(\theta^{(1)}) = -45.5680$ is the maximum value of $L_{M^{(1)}}(\theta^{(1)})$ by definition, it cannot be accepted as a reliable estimate of $L(\theta^{(1)})$. A new sampling distribution was chosen by generating another Gibbs sample.

Table 4.4 Detailed estimation summary for $\kappa = 48$

<table>
<thead>
<tr>
<th>Cycle</th>
<th>No. of NR</th>
<th>No. of MC</th>
<th>Value of $\theta$</th>
<th>Value of $L_M(\theta)$</th>
<th>Value of $\hat{V}_1$</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 Start</td>
<td>4</td>
<td>30,000</td>
<td>(0.9172, 0.0294)</td>
<td>-45.6234</td>
<td></td>
</tr>
<tr>
<td>1 End</td>
<td>4</td>
<td>30,000</td>
<td>(0.9742, 0.0224)</td>
<td>-45.5680</td>
<td>28.37</td>
</tr>
<tr>
<td>2 Start</td>
<td>2</td>
<td>30,000</td>
<td>(0.9778, 0.0220)</td>
<td>-45.5617</td>
<td></td>
</tr>
<tr>
<td>2 End</td>
<td>2</td>
<td>30,000</td>
<td>(0.9765, 0.0221)</td>
<td>-45.5611</td>
<td>0.10</td>
</tr>
</tbody>
</table>
of size 100,000 from the conditional specifications (4.19), now with the parameter $\theta^{(1)}$ and following the prescription of Section 4.4.1. Another cycle of Newton-Raphson iterations yielded the estimate $\theta^{(2)} = (0.9778, 0.0220)^T$ and $\tilde{V}_1(\theta^{(1)}, \theta^{(2)}) = 0.10$. This value was actually greater than 0.10 in extended decimal places, so one additional iteration of the overall estimation procedure was conducted, resulting in the final estimate $\theta^{(3)} = (0.9765, 0.0221)^T$ and a value $\tilde{V}_1(\theta^{(2)}, \theta^{(3)}) = 0.08$. Based on this value, we were now willing to accept $L_{M(3)}(\theta^{(2)}) = -45.5610$ and $L_{M(3)}(\theta^{(3)}) = -45.5611$ as similar enough in precision to allow assessment of $L_{M(3)}(\theta^{(3)}) - L_{M(3)}(\theta^{(2)}) = (-45.56106) - (-45.56108) = 0.00002$ as indicative of convergence; the values in Table 4.4 have been rounded to four decimal places. The Euclidean norm of the change in estimated parameter values was $||\theta^{(2)} - \theta^{(3)}|| = 0.00130$.

We conducted an additional check for developmental purposes, primarily to verify that the procedures of Sections 4.4.3 and 4.4.4 for assessment of Monte Carlo error were functioning correctly. Our basic concern was that even the use of $\tilde{V}(\theta^{(n-1)}, \theta^{(n)})$ to judge changes in Monte Carlo precision could be misleading, if the prescription of Section 4.4.4 results in an overly-optimistic assessment of baseline Monte Carlo error. The essential issue here is Monte Carlo sample size, and we repeated the estimation procedure for this example using a Monte Carlo sample size of 600,000, 20 times greater than that used in the estimation summarized in Table 4.4. With an initial Monte Carlo sample size of $M^{(1)} = 600,000$, and the same starting value of $\theta^{(0)} = (0.9172, 0.0294)^T$ as used previously, the estimated log likelihood was $L_{M(3)}(\theta^{(0)}) = -45.6182$. This differs from the corresponding estimate with $M^{(1)} = 30,000$ by 0.0052, well within the range of precision estimated for that Monte Carlo log likelihood. Similarly, an 95% interval for $L_M(\theta^{(0)}) - L(\theta^{(0)})$, calculated for a sample size of 30,000 but based on the estimated value $\tilde{V}_2(\theta^{(0)})$ from a sample of size 600,000, was ±0.0143. This compares favorably with the original value for this interval of ±0.0098. Completion of an estimation procedure with a Monte Carlo sample size of 600,000 throughout was similar to that presented in Table 4.4. Values of $\tilde{V}_1(\theta^{(n-1)}, \theta^{(n)})$ decreased from 11.70 at cycle 1 to 0.03 at cycle 3 and the final parameter estimate was $\hat{\theta} = (0.9738, 0.0224)$, differing from the final estimate of Table 4.4 by a Euclidean norm of 0.0027. The Monte Carlo log likelihood for this estimate was $L_{M(3)}(\theta^{(3)}) = -45.5585$, differing from the corresponding value of Table 4.4 by 0.0026. This use of an inflated Monte Carlo sample size verifies that the convergence and sample size selection procedures outlined in Sections 4.4.3 and 4.4.4 performed adequately in this example.

### 4.7 Concluding Remarks

In this article we have presented a method for maximum likelihood estimation of parameters appearing in conditionally specified statistical models that incorporate complex dependence structures.
In so doing, we have addressed a number of practical issues that arise in the application of Monte Carlo methods to repeated evaluations of a log likelihood. With the exception of the data-analytic issue of neighborhood selection in our spatial example, our solutions to these issues have all depended on one cornerstone idea, that being the construction of importance sampling distributions as the product of marginal densities, equivalent to forming 'dummy' independence models to parallel the behavior of the actual dependence model of interest. By matching the lower moments of marginal densities in the dummy model with estimates based on data sets simulated from the actual dependence model, we are able to produce sampling distributions that allow sufficiently precise Monte Carlo estimation and maximization of the actual log likelihood. The use of a dummy model constructed for independent random variables also allows application of theoretical results for independence case.

Matching the lower-order behavior of the dummy and actual models allows the use of Penttinen's (1984) independent marginals in cases that exhibit more than weak dependence. It is true that estimated parameter values in our example would indicate weak spatial dependence; data sets simulated from the final estimated model showed pairwise correlations for locations were generally less than 0.10. Despite this, applications to simulated data from a Winsorized Poisson auto-model have indicated that the procedure is able to provide maximum likelihood estimates for situations with much stronger dependence; pairwise correlations among locations in these simulations have been as high as 0.40.

In this article, as is often true in statistical applications of Monte Carlo methodology, importance sampling has been used to allow a successful approach to be developed in the first place, rather than as a variance reduction technique in its own right. As statistical applications of Monte Carlo techniques become more advanced, the issue of variance reduction will likely see more emphasis. Our construction of sampling distributions as a simple product of marginals would seem well suited for the pursuit of variance reduction goals in complex models, since the process of sampling from the dummy model involves only the generation of independent values. Basic Monte Carlo variance reduction techniques might be easily applied to improve performance of the estimation procedure further, or ease the burden of required Monte Carlo sample size. For example, lack of dependence among sampled values should allow antithetic sampling to be easily incorporated into the sampling procedure. The adjective 'easily' is appropriate here because sampling from our importance distributions does not require the use of Markov chain methods, although the formulation of those importance distributions does.
Acknowledgment

This research was supported by the U.S. Environmental Protection Agency under Assistance Agreement No. CR822919-01-0.
CHAPTER 5 MAXIMUM LIKELIHOOD ESTIMATION IN CONDITIONALLY SPECIFIED STATISTICAL MIXTURE MODELS

5.1 Introduction

Mixture models are widely used for modeling various statistical problems including overdispersed data, longitudinal data, missing data, etc. If there is no dependence among data, estimation of interesting parameters and inferences in mixture models are quite available. When it comes to data with dependence, not only estimation and inference, but even modeling is not easy. The difficulty in modeling is incorporating the dependence. There are some choices, but not all are feasible in terms of estimation. The difficulty in estimation is existence of unknown normalizing constants which can be somewhat handled according to the modeling methods. We will consider in this chapter the following modeling: a data model is conditionally independent given parameters, and the distribution of parameters is formulated through specification of a set of conditional distributions. That is to say, let the joint distribution of the data be \( f(z|\theta) \), where the components of \( z \) are conditionally independent given \( \theta \). And let \( \theta \) be a random vector whose joint distribution, say \( g(\theta; \tau) \), is characterized by specifying the conditional distributions \( g(\theta_i|(\theta_j, j \neq i)) \), for \( i = 1, \ldots, n \). Then the mixture distribution of \( z \) given \( \tau \) is \( f(z|\theta)g(\theta; \tau)d\theta \). In this situation, we want to consider the problem of obtaining full maximum likelihood estimates (MLEs) of the parameter vector, \( \tau \), by maximizing the log likelihood of mixture distribution, i.e., maximizing \( L(\tau) \equiv \log(\int f(z|\theta)g(\theta; \tau)d\theta) \) with respect to \( \tau \). Because the distribution of \( \theta \) is defined by a set of conditional distributions, \( \{g(\theta_i|(\theta_j, j \neq i))|i = 1, \ldots, n\} \), the joint distribution of \( g(\theta|\tau) \) contains an unknown normalizing constant, i.e., \( g(\theta; \tau) = h(\theta; \tau)/k(\tau) \), where \( k(\tau) \equiv \int h(\theta; \tau)d\tau \) is hard to evaluate. In this case, the log likelihood of mixture distribution can be expressed as the logarithm of a ratio of two intractable integrations, i.e., \( \log \frac{\int f(z|\theta)h(\theta; \tau)d\theta}{\int h(\theta; \tau)d\theta} \). Therefore, it is impossible to get the MLE by maximizing the log likelihood of mixture distribution. Although we cannot get exact MLE, we can get the sequence of random variables which converges almost surely to the true MLE if we employ Monte Carlo strategy as Geyer and Thompson (1992) and Lee and Kaiser (1997) did in non-mixture models.
in which there is only one intractable integration. Actually, the estimation of a ratio of two intractable integrations is easier than that of one intractable integration. (Geyer and Thompson (1992) also used this idea.) Therefore, many attempts have been made to estimate the ratio of two intractable integrations either by simulation methods or non-simulation methods. For the non-simulation methods, see Naylor and Smith (1982), Tierney and Kadane (1986 and 1989), and Morris (1988). For the simulation methods, see Gelfand and Carlin (1993), Newton and Raftery (1994), Chibs (1995), Meng and Wong (1995), and Ogata (1996). All those methods, except Gelfand and Carlin (1993), are designed to estimate the ratio of integrals at one point, but it may be possible to use those methods in maximization problem. The advantage of using non-simulation methods in a maximization process is fast execution, but the major problematic aspect is that nothing can be said about consistency or asymptotic behaviors of the resulting estimates. Every simulation-based method cannot make it possible to show the desired large sample properties, which is somewhat dependent on the way Monte Carlo samples are used. In the simulation-based optimization of likelihood, we must evaluate likelihood and possibly derivatives at each iteration. Using different samples from different sampling distributions at each iteration will be very inefficient. One of the ways to use only one sample all through the iterations is to use importance densities. This idea was developed by Geyer and Thompson (1992) and used in the models of missing and constrained data. These models include the mixture models by Gelfand and Carlin (1993) who did not quite make use of the fact that likelihood has the form of a ratio of two integrals as Meng and Wong (1995) did, but did make use of the fact that maximizing \( \frac{L(\tau)}{L(\psi)} \) with respect to \( \tau \) is equivalent to maximizing \( L(\tau) \) when \( \psi \) is fixed independently of \( \tau \). Their unnormalized density is \( h(z, \theta; \tau) \) with \( \theta \) missing and \( z \) observed. Then \( L(\tau) \) is

\[
\log \frac{\int h(z, \theta; \tau) d\theta}{\int h(z, \theta; \tau) d\theta dz} = \log \left[ \frac{\int h(z, \theta; \psi) d\theta}{\int h(z, \theta; \psi) d\theta dz} \right] + \log \left[ \frac{\int h(z, \theta; \psi) h(z, \theta; \psi)}{\int h(z, \theta; \psi) d\theta dz} \right] - \log \left[ \frac{\int h(z, \theta; \psi) h(z, \theta; \psi)}{\int h(z, \theta; \psi) d\theta dz} \right] = L(\psi) + \log \left[ E_{\psi} \left\{ \frac{h(Z, \theta; \tau)}{h(Z, \theta; \psi)} | Z = z \right\} \right] - \log \left[ E_{\psi} \left\{ \frac{h(Z, \theta; \tau)}{h(Z, \theta; \psi)} \right\} \right],
\]

and, dropping \( L(\psi) \) unrelated to \( \tau \), its MC estimation is

\[
\log \left[ \sum \frac{h(z, \theta; \tau)}{h(z, \theta; \psi)} \right] - \log \left[ \sum \frac{h(Z_i, \theta; \tau)}{h(Z_i, \theta; \psi)} \right],
\] (5.1)

where \( \{\theta_i\} \) is a sample from the conditional distribution of \( \theta \) given \( Z = z \) and \( \{(Z_i, \theta_i)\} \) is a sample from the unconditional distribution (both for the parameter value \( \psi \)). Gelfand and Carlin (1993)
suggested maximizing (5.1) to obtain Monte Carlo MLE. This method can be applied to our case in which \( f h(z, \theta; \psi)d\theta dz = \int h(\theta; \psi)d\theta \) so that double sampling still remains, but sampling from joint distribution is not necessary. Then (5.1) becomes

\[
\log \left( \sum \frac{h(z, \theta_i^*; \tau)}{h_z(z, \theta_i^*; \psi)} \right) = \log \left( \sum \frac{h(\theta_i; \tau)}{h(\theta_i; \psi)} \right),
\]

where \( \theta_i^* \) is a sample from the conditional distribution of \( \theta \) given \( Z = z \) and \( \{\theta_i\} \) is a sample from the mixing distribution. The key point for maximizing (5.1) is to choose \( \psi \) in order that \( h(z, \theta; \psi) \) may dominate \( h(z, \theta; \tau) \) for any \( \tau \). The best choice is true MLE, \( \hat{\tau} \) which is impossible. If \( \psi \) is far away from the true MLE, a maximization process like a Newton-Raphson algorithm usually ends up with the second derivative matrix which is singular and can not be inverted. Actually, it is impossible to have such an \( h(z, \theta; \psi) \) if we insist that we should have same function \( h \) as \( h(z, \theta; \tau) \) but different point, \( \psi \). Therefore, our suggestion is to use \( m \) which has a different function from \( h \). It would also be better if we do not have any unknown terms in \( m \). This approach was done in non-mixture model by Lee and Kaiser (1997). From their suggestion, the estimated log likelihood, \( L_M(\tau) \), looks like

\[
\log \left( \frac{1}{M} \sum_{i=1}^{M} f(z|\theta_i)h(\theta_i; \tau) \right) = \log \left( \frac{1}{M} \sum_{i=1}^{M} m_1(\theta_i; \psi) \right),
\]

where \( \theta_i \)'s are a sample from \( m_1 \), \( \theta_2j \)'s are a sample from \( m_2 \). Here, \( m_1 \) and \( m_2 \) could be the same density, but in practice, it is difficult to get one good importance density for two different integrands, even if the importance density has the form of hybrid of \( f(z|\theta)h(\theta; \tau) \) and \( h(\theta; \tau) \) as suggested by Newton and Raftery (1994).

With the difficulty to find a good \( \psi \) (\( \lambda_i \)'s in our case), Gelfand and Carlin (1993) follow the idea of constrained maximization algorithm by Geyer and Thompson (1992). They suggested the iterative scheme of a maximization process focusing on the convergence of the Monte Carlo MLE to the true MLE as follows. Starting with \( \psi = \tau_0 \), (5.1) is maximized within the neighborhood of \( \tau_0 \), i.e., within the closed ball centered at \( \tau_0 \) with a small radius, say \( \delta \). The maximizer, \( \tau_1 \) could occur either on the boundary or inside the ball. If \( \tau_1 \) is inside the ball, \( \tau_1 \) could be very close to the true MLE, so generating new samples with a large size at \( \tau_1 \), and maximizing (5.1) to get the final MC MLE. If \( \tau_1 \) lies on the boundary, let \( \psi = \tau_1 \), and maximize (5.1) with new samples within the other closed ball. Continue these steps until convergence occurs. In this process, important steps are selecting a sampling density which must dominate the integrand within some reasonable range, and deciding the need of another sampling density which has been done by comparing predetermined 'step size', \( \delta \), with the distance between the starting point and ending point at which maximum occurs. Neither Geyer and Thompson (1992) nor Gelfand and Carlin (1993) can give the nice way to choose the importance densities and to decide the
step size. Another problematic part of this process is the selection of a starting point. If a starting point is selected poorly, the maximization cycle cannot guarantee the convergence, so many cycles for the overall convergence are needed. Lee and Kaiser (1997) provided nice solutions to those problems in non-mixture models. First, the selected sampling density consists of independent components and yet captures the essential impact of dependence on the pattern of realizations from the actual model by matching the mean and variance of each component with the sample from the actual model. The way to construct the sampling density makes it possible for sampling density to dominate the integrand. Second, the Central Limit Theorems (CLT) for estimated normalizing constant and log likelihood are straightforward and it is easy to get the estimation of asymptotic variances because of the independent sample. From this, the criterion on whether to need another sampling density can be made. Along with this criterion, it is possible to decide whether the maximum of current cycle is really close to the true maximum. Finally, selecting a good starting point can be done by Laplace approximation which is fast because no simulation is used. In this chapter, we will extend the methods used in Lee and Kaiser (1997) to the mixture models. The only difference is that, in the log likelihood of mixture models, there are two integrations each of which can be approximated by different importance densities. Therefore, the extension may be straightforward.

In Section 5.2, the conditionally specified mixture models are defined and some useful classes of models are introduced with some properties. In Section 5.3, the overall estimation process will be given. Section 5.4 presents a selection method of the two importance densities for the Monte Carlo evaluation of the log likelihood, and the ways to assess the precision of estimation of the log likelihood and determine the overall convergence. In Section 5.5, the choice of starting value via Laplace approximation will be given, and, in Section 5.6, the method developed in Section 5.3 to 5.5 will be applied to the data from Environmental Protection Agency. In Section 5.7, discussion will be provided.

5.2 Conditionally Specified Mixture Models

5.2.1 Definition

Suppose the data are

\[ Z \equiv (Z(s_1), \ldots, Z(s_n))^t, \]  \hspace{1cm} (5.2)

where \( s_i \) denotes the 'location' of \( Z \) in a random field. This location may be an actual geographical location, but it may also refer to a time of occurrence in longitudinal data or a grouping mechanism in a sub-sampling or repeated measures study. For example, in a spatial problem, we might have
$s_i \equiv (x_i, y_i)$, where $x_i$ is longitude and $y_i$ is latitude, while in a repeated measure study we might take $s_i \equiv (k, j)$, where $k$ indexes subject and $j$ indexes observation number. In a multivariate time series applications, a possible definition would be $s_i \equiv (k, t(j))$ with $k$ indexing variable and $t(j)$ being the time of the $j^\text{th}$ observation.

As explained in Section 5.1, we have various situations in which we must use mixture models. When correlations among data are expected, we have many choices to incorporate the dependence structure into the mixture models. One of them is to use conditionally independent model for data and to use some mixing distribution with dependent structure for unobservable random parameters. Mixing distributions with dependent structures are very limited, but conditionally specified models can provide a variety of classes of distributions, even though we have unknown normalizing constants. When we use conditionally specified models as mixing distributions, we call the mixture models conditionally specified mixture models.

We will briefly describe on conditionally specified models. Let $\theta \equiv (\theta(s_1), \ldots, \theta(s_n))$. With $g(\cdot \mid \cdot)$ denoting a generic conditional probability density or mass function, a conditionally specified model consists of the $n$ functions

$$g(\theta(s_i) \mid \{\theta(s_j) : j \neq i\}); \ i = 1, \ldots, n$$

Dependence is modeled by defining, for each $\theta(s_i)$, a dependence set or neighborhood,

$$N_i \equiv \{s_h : g(\theta(s_i) \mid \{\theta(s_j) : j \neq i\}) \text{ depends on } \theta(s_h); h \neq i\},$$

and corresponding dependence index set $D_i \equiv \{h : s_h \in N_i\}$. Then, assuming it exists, we can identify the joint distribution corresponding to the set of specified conditionals. Kaiser and Cressie (1996) outline the constructive process involved and give conditions under which the joint density or mass function may be identified up to a normalizing constant as,

$$g(\theta) = \frac{\exp\{Q(\theta)\}}{\int \exp\{Q(\cdot)\} d\mu(\cdot)}; \ \theta \in \Theta,$$

where $\Theta$ is the support of $g(\theta)$ and $Q(\cdot)$. This is called Negpotential function, which is a linear combination of functions constructed from the conditionals in (5.3). Notice that expression (5.4) can be considered to be quite general since any distribution may be expressed in this manner.

Now, $h(z)$, the mixture distribution of $Z$, becomes

$$\int f(z) g(\theta) d\mu(\theta) = \frac{\int f(z) \exp\{Q(\theta)\} d\mu(\theta)}{\int \exp\{Q(\cdot)\} d\mu(\cdot)}.$$

(5.5)
5.2.2 Some Useful Classes of Models

To produce some useful classes of conditionally specified mixture models, let's assume that conditionally independent data model belongs to exponential family and the conditional distribution of one random parameter given all the other parameters is the form of conjugate prior, i.e., the distribution of $Z(s_i)$ given $\theta(s_i)$ is

$$f(z(s_i)|\theta(s_i)) = \exp\{\phi[z(s_i)\theta(s_i) - b(\theta(s_i))] + c(z(s_i), \phi)\}, \quad (5.6)$$

and the conditional distribution of $\theta(s_i)$ is

$$g(\theta(s_i)|{\{\theta(s_j) : j \in D_i}\}}) = \exp\{A_{i1}(D_i)\theta(s_i) - A_{i2}(D_i)b(\theta(s_i))\} + k(A_{i1}(D_i), A_{i2}(D_i)), \quad (5.7)$$

$i = 1, \cdots, n$.

It is easily seen that the conditional distribution of $\theta(s_i)$ is the form of conjugate prior for the natural exponential family distribution. Three classes of models can be defined by $A_{i1}(D_i)$ and $A_{i2}(D_i)$ in three different ways.

- Class 1:
  $$A_{ik}(D_i) = \alpha_{ik} + \sum_{j \in D_i} \eta_{ij}[\theta(s_j) - b(\theta(s_j))], \quad k = 1, 2$$
  , where $\eta_{ij} = \eta_{ji}$.

- Class 2:
  $$A_{i1}(D_i) = \alpha_{i1} + \sum_{j \in D_i} \eta_{ij}\theta(s_j)$$
  $$A_{i2}(D_i) = \alpha_{i2} + \sum_{j \in D_i} \eta_{ij2}[-b(\theta(s_j))]$$

- Class 3:
  $$A_{i1}(D_i) = \alpha_{i1} + \sum_{j \in D_i} \eta_{ij}[-b(\theta(s_j))]$$
  $$A_{i2}(D_i) = \alpha_{i2} + \sum_{j \in D_i} \eta_{ij}\theta(s_j)$$
  , where $\eta_{ij} = \eta_{ji}$.

For the argument of validness of these parametrizations for the Markov Random field, see Kaiser and Cressie (1996). One thing we have to be careful in using one of the above models is parameter restriction. For example, let (5.7) be the beta distribution, then $A_{i1}$ and $A_{i2}$ play the roles of parameters of the
beta distribution, so they must be positive. It is easily verified that all the $\alpha_{ij}$'s must be positive, and all the $\eta_{ij}$'s must be negative. Although $\eta_{ij}$'s must be negative in all the classes, class 1 can give both positive and negative correlation structures, but cover only small range of correlations. Class 2 or class 3 is recommended for data showing high correlations if we know whether the data have the negative or positive correlations. Class 2 gives only negative dependence structures and class 3 gives only positive ones.

Each class provides a distinct form for the negpotential function which was mentioned in the early part of this section.

- Class 1:
  \[
  Q(\theta) = \sum_{i=1}^{n} \{\alpha_{i1}\theta(s_i) - \alpha_{i2}b(\theta(s_i))\} + \sum_{1 \leq i < j \leq n} \{\eta_{ij}[\theta(s_i)\theta(s_j) - \theta(s_i)b(\theta(s_j)) - \theta(s_j)b(\theta(s_i)) + b(\theta(s_i))b(\theta(s_j))]]
  \]

- Class 2:
  \[
  Q(\theta) = \sum_{i=1}^{n} \{\alpha_{i1}\theta(s_i) - \alpha_{i2}b(\theta(s_i))\} + \sum_{1 \leq i < j \leq n} \eta_{ij}[\theta(s_i)\theta(s_j)] + \eta_{ij2}[b(\theta(s_i))b(\theta(s_j))]
  \]

- Class 3:
  \[
  Q(\theta) = \sum_{i=1}^{n} \{\alpha_{i1}\theta(s_i) - \alpha_{i2}b(\theta(s_i))\} + \sum_{1 \leq i < j \leq n} \eta_{ij2}[-\theta(s_i)b(\theta(s_j)) - \theta(s_j)b(\theta(s_i))]
  \]

5.2.3 Properties

The wonderful feature for these classes of models is in that the sampling schemes of $\theta$ marginally and conditionally given $z$ are identical, which is very useful for MCMC methods, especially in Bayesian analysis. This feature is coming from the conjugate form. To see this clearly, think about the situation where we need samples from both $g(\theta)$ and $g(\theta|z)$, i.e., from prior, and posterior in the Bayesian analysis.

It is easy to sample from $g(\theta)$ because we know all the conditionals, $g(\theta_i|\{\theta(s_j) : j \in D_i\})$, $i = 1, \ldots, n$.

Then we can use the same program to sample from $g(\theta|z)$ because $g(\theta_i|z, \{\theta(s_j) : j \in D_i\})$ will have the same form as $g(\theta_i|\{\theta(s_j) : j \in D_i\})$ adding $\phi(z(s_i))$ to $A_{i1}(D_i)$ and $\phi$ to $A_{i2}(D_i)$.

Another nice property is that we can get the explicit form of correlation between $z(s_i)$ and $z(s_j)$, for all $i, j = 1, \ldots, n$, with respect to marginal distribution as functions of $\theta(s_i)$ and $\theta(s_j)$.

\[
Cov_h(z(s_i), z(s_j)) = E_h(z(s_i)z(s_j)) - E_h(z(s_i))E_h(z(s_j))
\]
\[
= E_g[E_f(z(s_i)z(s_j)|\theta) - E_g[E_f(z(s_i)|\theta)]E_g[E_f(z(s_j)|\theta)]
\]
\[
= E_g[E_f(z(s_i)|\theta)E_f(z(s_j)|\theta)] - E_g[E_f(z(s_i)|\theta)]E_g[E_f(z(s_j)|\theta)]
\]
and

\[
\text{Var}_h(z(s_i)) = E_g[\text{Var}_f(z(s_i)|\theta)] + \text{Var}_g[E_f(z(s_i)|\theta)] = E_g\left[\frac{b''(\theta(s_i))}{\phi}\right] + \text{Var}_g[b'(\theta(s_i))].
\]

Therefore, we can have

\[
\text{Corr}_h(z(s_i), z(s_j)) = \frac{\text{Cov}_g[b'(\theta(s_i)), b'(\theta(s_j))]}{\sqrt{E_g\left[\frac{b''(\theta(s_i))}{\phi}\right] + \text{Var}_g[b'(\theta(s_i))]} \sqrt{E_g\left[\frac{b''(\theta(s_j))}{\phi}\right] + \text{Var}_g[b'(\theta(s_j))]}}, \tag{5.8}
\]

which is less than \(\text{Corr}_g[b'(\theta(s_i)), b'(\theta(s_j))]\) because \(E_g\left[\frac{b''(\theta(s_i))}{\phi}\right] > 0\), for all \(i = 1, \ldots, n\). In the above expression \(h\) denotes the mixture distribution of \(z\) as in (5.5), \(f\) denotes conditional distribution of \(z\) given \(\theta\), and \(g\) denotes the joint distribution of \(\theta\) obtained by the specification of (5.7).

### 5.3 The Overall Estimation Strategy

In this section, the overall estimation strategy used to find MLEs of parameters appearing in (5.5) is described. Let such parameters be \(\tau\) which is \((\tau_1, \ldots, \tau_p)^T\). Then we can write (5.5) as

\[
h(z|\tau) = \frac{\int f(z|\theta) \exp\{Q_0(\theta|\tau)\} d\theta}{\int \exp\{Q_0(\theta|\tau)\} d\theta} = \frac{\int \exp\{\log(f(z|\theta)) + Q_0(\theta|\tau)\} d\theta}{\int \exp\{Q_0(\theta|\tau)\} d\theta}
\]

\[
= \frac{\int \exp\{Q_1(\theta|\tau)\} d\theta}{\int \exp\{Q_0(\theta|\tau)\} d\theta} = k_1(\tau) k_0(\tau)^{-1}, \tag{5.9}
\]

where we suppress the dependence of \(Q_1\) and \(k_1\) upon \(z\). From (5.9) the log likelihood may be written as

\[
L(\tau) = \log(k_1(\tau)) - \log(k_0(\tau)), \tag{5.10}
\]

the first derivatives of \(L\), if exist, may be written, for \(j = 1, \ldots, p\), as

\[
\frac{\partial L(\tau)}{\partial \tau_j} = E_{fg} \left( \frac{\partial Q_0(\theta|\tau)}{\partial \tau_j} \right) - E_g \left( \frac{\partial Q_0(\theta|\tau)}{\partial \tau_j} \right), \tag{5.11}
\]

where \(E_{fg}(\cdot)\) denotes the expectation with respect to \(\exp\{Q_1(\theta|\tau)\}/k_1(\tau)\), and \(E_g(\cdot)\) with respect to \(g(\theta|\tau)\).
and the second derivatives, if exist, may be written, for \( j, k = 1, \cdots, p \), as

\[
\frac{\partial^2 L(\tau)}{\partial \tau_j \partial \tau_k} = E_{f_1}(\frac{\partial^2 Q(\theta|\tau)}{\partial \tau_j \partial \tau_k}) + E_{f_2}(\frac{\partial Q_0(\theta|\tau)}{\partial \tau_j} \frac{\partial Q_0(\theta|\tau)}{\partial \tau_k}) - E_{f_1}(\frac{\partial Q_0(\theta|\tau)}{\partial \tau_j}) E_{f_2}(\frac{\partial Q_0(\theta|\tau)}{\partial \tau_k})
\]

\[
- E_g(\frac{\partial^2 Q(\theta|\tau)}{\partial \tau_j \partial \tau_k}) - E_g(\frac{\partial Q_0(\theta|\tau)}{\partial \tau_j}) E_g(\frac{\partial Q_0(\theta|\tau)}{\partial \tau_k}).
\]  

(5.12)

Let \( \hat{\tau} \in \mathcal{T} \) denotes the maximizing value of (5.10), that is, the true MLE of \( \tau \). Our strategy is to construct a sequence of estimators \( \{ L_{M(q)}(\tau^{(q)}) : q = 1, \ldots \} \), such that \( L_{M(q)}(\tau^{(q)}) \) becomes a sufficiently precise estimator of \( L(\hat{\tau}) \) as \( q \) increases; in practice we desire \( q \) to be small (i.e., \( q = 2 \) or \( q = 3 \)). This sequence of estimators will be based on Monte Carlo estimation of \( L(\tau) \), following the concept of constrained maximization introduced by Geyer and Thompson (1992). For this purpose, let \( m_1(\theta|\lambda_1) \) and \( m_0(\theta|\lambda_0) \) be two different importance distributions (densities) such that \( \lambda_1 \) and \( \lambda_0 \) do not depend on \( \tau \), and \( m_1(\theta|\lambda_1) \) and \( m_0(\theta|\lambda_0) \) dominate \( \exp\{Q_1(\theta|\tau)/k_1(\tau) \} \) and \( \exp\{Q_0(\theta|\tau)/k_0(\tau) \} \) respectively in that \( m_1(\theta|\lambda_1) = 0 \) implies \( \exp\{Q_1(\theta|\tau)/k_1(\tau) \} = 0 \), and \( m_0(\theta|\lambda_0) = 0 \) implies \( \exp\{Q_0(\theta|\tau)/k_0(\tau) \} = 0 \) (see Geyer, 1996). For any value \( \tau \in \mathcal{T} \), a Monte Carlo estimator of \( k_1(\tau) \) is

\[
k_{iM(q)}(\tau) = \frac{1}{M(q)} \sum_{j=1}^{M(q)} \frac{\exp\{Q_i(\theta_{i}^{(q)}|\tau)\}}{m_i(\theta_{i}^{(q)}|\lambda_{i}^{(q)})},
\]

(5.13)

where \( \theta_{i}^{(q)}, \ldots, \theta_{M(q)}^{(q)} \) is a sample from \( m_i(\theta|\lambda_{i}^{(q)}) \), for \( i = 1 \) or 0. A Monte Carlo estimator of the \( 0 \) (5.10) is then

\[
L_{M(q)}(\tau) = \log \{ k_{1M(q)}(\tau) \} - \log \{ k_{0M(q)}(\tau) \},
\]

(5.14)

and an estimator of the maximized log likelihood \( L(\tau) \) is \( L_{M(q)}(\tau^{(q)}) \), where \( \tau^{(q)} \) is the value of \( \tau \in \mathcal{T} \) that maximizes \( L_{M(q)}(\tau) \).

Notice that, while (5.14) is a Monte Carlo estimator of \( L(\tau) \) over the entire parameter space, for a given Monte Carlo sample size \( M(q) \) it will be sufficiently precise for estimation purposes only for some subset of that space, \( \tau \in \mathcal{T}_{q} \) say; the index \( \lambda^{(q)} \equiv (\lambda_{1}^{(q)}, \lambda_{0}^{(q)}) \) indicates dependence of this subset on the parameters of the importance distributions \( m_1(\theta|\lambda_{1}^{(q)}) \) and \( m_0(\theta|\lambda_{0}^{(q)}) \). Thus, the value \( \tau^{(q)} \) that maximizes \( L_{M(q)}(\tau) \) will be a sufficiently precise estimator of \( \tau \) only if \( \tau \in \mathcal{T}_{q} \), and hence the need for a sequence of such estimators, indexed here by \( q \). Because identification of the appropriate subset \( \mathcal{T}_{q} \) is a difficult task a priori, it is crucial to have available a means by which to assess how well \( L_{M(q)}(\tau) \) approximates \( L(\tau) \) for various values \( \tau \in \mathcal{T} \). In addition, construction of the sequence \( \{ L_{M(q)}(\tau^{(q)}) : q = 1, \ldots \} \) requires a way to maximize the Monte Carlo log likelihood (5.14). The remainder of this section deals with this second issue, maximization of a given \( L_{M(q)}(\tau) \).
Our approach to maximization of (5.14) in \( \tau \) is simply to use a Newton-Raphson algorithm. The first and second derivatives of \( L_M(\tau) \) exist as long as those of \( Q_0 \) exist, and they turn out to be the consistent estimators of the first and second derivatives of true log likelihood \( L(\tau) \) in (5.11) and (5.12). Notice that we need the precise \( L_M(\tau) \) as a estimator of \( L(\tau) \) but we don't need to assess the accuracies of derivatives of \( L_M(\tau) \) as the estimators of (5.11) and (5.12) because we want the maximizer of \( L_M(\tau) \).

We will deal with the issue of precision in \( L_M(\tau) \) and the choice of \( m_i(\theta|\tau) \)'s which make it possible to assess the precision of Monte Carlo estimation and have domination properties in Section 5.4.

In the estimation strategy proposed, one-time sample, \( \{\theta_1^0, \ldots, \theta_M^0, \theta_1^1, \ldots, \theta_M^1\} \), from \( m_0(\cdot|\cdot) \) and \( m_1(\cdot|\cdot) \) is used to compute all quantities needed for a set of iterations, called here a 'cycle', of a standard Newton-Raphson algorithm. Upon convergence, we will decide whether changes in the parameter estimate have rendered the original sampling distribution unreliable for Monte Carlo estimation by assessing the error in Monte Carlo estimation of the maximized log likelihood \( L(\tau^{(q)}) \). The way of assessment will be given in Section 5.4. If this is the case, the new sampling distributions are used for another cycle of Newton-Raphson iterations; otherwise, look at the convergence criteria in terms of changes in parameter estimates and changes in log likelihood values to make sure the overall convergence.

The total convergence procedure is then as follows:

1. Choose a starting value \( \tau^{(0)} \).

2. For any current parameter value \( \tau^{(q-1)} \); \( q \geq 1 \), choose sampling densities \( m_1(\theta|\lambda_1^{(q)}) \) and \( m_0(\theta|\lambda_0^{(q)}) \).

3. Draw 2 samples of size \( M^{(q)} \) each from \( m_1(\theta|\lambda_1^{(q)}) \) and \( m_0(\theta|\lambda_0^{(q)}) \), assessing the error of Monte Carlo estimation of the log likelihood \( L_M(\tau^{(q-1)}) \) to ensure sufficient precision.

4. Use the samples from step 3 for numerical evaluations of needed quantities.

5. Update the parameter estimate through use of a Newton-Raphson algorithm, repeating step 4 as needed until convergence, always with the same sample from \( m_1(\theta|\lambda_1^{(q)}) \) and \( m_0(\theta|\lambda_0^{(q)}) \). Call the resulting estimate of the parameter from this Newton-Raphson cycle \( \tau^{(q)} \).

6. Assess the error in Monte Carlo estimation of the maximized log likelihood \( L(\tau^{(q)}) \). If this error has increased from that of \( L_M(\tau^{(q-1)}) \) in step 3, update \( q \) to \( q + 1 \) and repeat, beginning at step 2.

Our overall estimation strategy consists of three levels of iterative procedures. At the first level, Monte Carlo samples are used to evaluate the log likelihood and its derivatives. At the second level, a
Newton-Raphson algorithm is used to maximize the log likelihood as estimated by the current Monte Carlo samples. Finally, at the third level, the constrained maximization procedure of Geyer and Thompson (1992) is employed to insure that a global maximum of the true log likelihood has been found.

5.4 Importance Sampling

In this section, we will describe the way to select importance sampling distributions \( m_1(\cdot|\cdot) \) and \( m_0(\cdot|\cdot) \) which are easy to sample from, appropriate for use with a cycle of Newton-Raphson iterations in the overall estimation strategy with the domination properties, and allows determination of the need of new sampling distributions. First, how to select two sampling densities and sampling methods will be discussed. Next, we will provide the way to assess the precision in Monte Carlo log likelihood and the way to determine whether we need new sampling densities after one cycle of Newton-Raphson algorithm by use of the feature of selected sampling densities.

5.4.1 Selection of Sampling Distributions

First, the method to select a sampling distribution for approximating \( k_0 \), the second term in (5.10) will be described. Consider a model specified through a set of conditional probability or mass functions as in (5.3). Typically, these conditional specifications will be in parameterized form, \( \pi(\theta(s_i)|\tau_i(\{\theta(s_j) : j \in D_i\}); i = 1, \ldots, n \). We will assume the positivity condition of Besag (1974) that, for \( \theta(s_i) \in \Theta_i, \theta \in \Theta = \Theta_1 \times \ldots \times \Theta_n \). We will not make a corresponding assumption about the \( r(\cdot) \), since construction of a joint distribution in the form of expression (5.4) may require additional constraints on allowable parameter values to ensure that the normalizing constant \( k_0(\tau) \) exists. For the purposes of this Chapter, we will simply assume that the \( \tau(\cdot) \) appearing in the conditional probability density functions are compatible with \( \tau \in \mathcal{T} \) in the joint density. Given such a set of specified conditional density functions, it will be possible to produce a sample of size \( S \) from the joint density (5.4) through the use of a Gibbs Sampling algorithm. For \( i = 1, \ldots, n \), let \( \hat{\mu}_{\mathcal{D}}(s_i) \) denote the sample mean of these values at location \( s_i \), and \( \hat{\sigma}^2_{\mathcal{D}}(s_i) \) the corresponding sample variance. To construct a sampling distribution for use in the overall estimation strategy of Section 5.3, we develop a 'dummy' model for \( \theta \) using the same set of locations \( \{s_i : i = 1, \ldots, n\} \) and the same support \( \Omega \). This dummy model takes the random variables \( \theta(s_1), \ldots, \theta(s_n) \) to be independent with marginal densities of the same form as the conditional functions specified in the actual model as in (5.3). The dummy model is then given by the collection of marginals
The parameters \( \lambda_{0i} \); \( i = 1, \ldots, n \) are chosen such that
\[
E_{\lambda_{0i}}(\theta(s_i)) = \mu_0 \delta(s_i) \quad \text{and} \quad \var_{\lambda_{0i}}(\theta(s_i)) = \sigma^2_{\theta,0}(s_i).
\]
We then construct a sampling distribution as
\[
m_0(\theta|\lambda_0) = \prod_{i=1}^n g(\theta(s_i)|\lambda_{0i}); \; \theta \in \Theta,
\]
where \( \lambda_0 = (\lambda_{01}, \ldots, \lambda_{0n})' \). As pointed out by Geyer (1994) it is not necessary that \( \lambda_0 \) have the same parameter space as does \( \tau \), as long as \( m_0(\theta|\lambda_0) \) dominates \( g(\theta|\tau) \), although this will often be the case simply by default. A sample \( (\theta_1(s_1), \ldots, \theta_1(s_n))^T, \ldots, (\theta_M(s_1), \ldots, \theta_M(s_n))^T \) generated from \( m_0(\theta|\lambda_0) \) will not reflect dependence among its values relative to the dummy ‘independence’ model. However, because the mean and variance of each \( g(\theta(s_i)|\lambda_{0i}); \; i = 1, \ldots, n \) have been matched with a sample from the actual dependence model, these values will reflect dependence among the \( \{\theta(s_i) : i = 1, \ldots, n\} \) under the actual model of interest, up to the effect of that dependence on the first two moments. For the sampling distribution for the first term in (5.10), \( k_1 \), we can use the same idea as used for \( k_0 \). One thing we want to mention is that a sampling method to get \( \mu_1 \delta(s_i) \) and \( \sigma_1 \delta(s_i) \) would not be a Gibbs sampling but be a more general one like the Hastings algorithm depending on the form of \( f \) and \( Q_0 \).

For example, for the third class of models in Section 5.2, the data model and the random parameter model can be written as follows;

\[
f(z|\theta) = \exp\left\{\sum_{i=1}^n [c(z(s_i), \theta(s_i))] + c(z(s_i), \phi)\right\},
\]

and

\[
g(\theta|\tau) = \exp(Q_0)/k_0(\tau),
\]

where \( Q_0 = \sum_{i=1}^n \{[\alpha_{i1} \theta(s_i) - \alpha_{12} \theta(s_i)] + \sum_{1 \leq i < j \leq n} \eta_{ij} [-\theta(s_i) \theta(s_j) - \theta(s_j) \theta(s_i)]\}, \quad k_0(\tau) = \int \exp(Q_0)d\theta, \quad \text{and} \quad \theta = (\theta(s_1), \ldots, \theta(s_n))^T.
\]

Then
\[
f(z|\theta) \exp\{Q_0(\theta|\tau)/k_1(\tau)\} = \exp\{\sum_{i=1}^n \{[\alpha_{i1} + \phi z(s_i)] \theta(s_i) - [\alpha_{12} + \phi] b(\theta(s_i))] + \sum_{1 \leq i < j \leq n} \eta_{ij} [-\theta(s_i) b(\theta(s_j)) - \theta(s_j) b(\theta(s_i))]\}. \quad (5.16)
\]

It is easily seen that \( \exp\{Q_1(\theta|\tau)/k_1(\tau)\} \) is the joint distribution of \( \theta \) which has the following conditional;

\[
p(\theta(s_i)|\{\theta(s_j), j \in D_i\}, z) 
\equiv \exp\{[A_{i1}(D_i) + \phi z(s_i)] \theta(s_i) - [A_{i2}(D_i) + \phi] b(\theta(s_i)) + k(A_{i1}(D_i) + \phi z(s_i), A_{i2}(D_i) + \phi)\}.
\]
where

\[ A_{i1}(D_i) = \alpha_{i1} + \sum_{j \in D_i} \eta_{ij} \theta(s_j) \] and \[ A_{i2}(D_i) = \alpha_{i2} + \sum_{j \in D_i} \eta_{ij} \theta(s_j). \]

Therefore, we can also use a Gibbs sampler. What is better is that the sampling algorithm for
\[ g(\theta(s_i)|\{\theta(s_j), j \in D_i\}) \] can be used again for \( p(\theta(s_i)|\{\theta(s_j), j \in D_i\}, z) \) with just a little bit of change.

That is to say, we can sample from \( g(\theta(s_i)|\{\theta(s_j), j \in D_i\}) \) with parameters \( A_{i1}(D_i) \) and \( A_{i2}(D_i) \), and sample from \( p(\theta(s_i)|\{\theta(s_j), j \in D_i\}, z) \) with parameters \( A_{i1}(D_i) + \phi z(s_i) \) and \( A_{i2}(D_i) + \phi \). Therefore, although we use two important densities, sampling procedure can't be burden if we use some classes of models.

In this way, we have constructed importance sampling distributions for use in the overall estimation strategy of Section 5.3 that are easily sampled from, because they consist of independent components, and yet captures the essential impact of dependence on the pattern of realizations from the actual model. The independence of random variables that have a joint distribution given by the dummy model also provides the necessary theory for evaluation of the precision of the Monte Carlo log likelihood (5.14) as an estimator of the true log likelihood (5.10). Because \( m_0(\theta|\lambda_0) \) is a independent multivariate density whose marginal means and variances are same as those of \( g(\theta|\tau) \), jointly \( m_0(\theta|\lambda_0) \) has more dispersion than \( g(\theta|\tau) \), resulting in having the domination property. Similar argument holds for \( m_1(\theta|\lambda_1) \).

### 5.4.2 Precision in Monte Carlo Log Likelihood

To assess precision of the Monte Carlo log likelihood for a given value of the parameter \( \tau \), we make use of the fact that the sampling distribution \( m(\cdot|\cdot) \) in (5.15) was constructed as a product of marginal probability density functions. In the first part of this section, we suppress dependence of the Monte Carlo quantities on the current iteration of the overall estimation strategy and write \( M \equiv M^{(q)} \) and \( \lambda_i \equiv \lambda^{(q)}_i \) throughout.

For random variables \( \{\theta^r_s: r = 1, \ldots, M\} \) with density or mass function \( m_r(\theta|\lambda_i) \), define the random variables

\[ D_i(\theta^r_s, \tau, \lambda_i) \equiv \exp \left[ Q_i(\theta^r_s|\tau) - m_i(\theta^r_s|\lambda_i) \right], \]

for \( i = 1, 0 \). Then, Monte Carlo estimators (5.13) and (5.14) of the two normalizing constants and log likelihood may then be written as, for \( i = 1, 0 \),

\[ k_iM(\tau) = \frac{1}{M} \sum_{r=1}^{M} D_i(\theta^r_s, \tau, \lambda_i), \quad (5.18) \]

and

\[ L_M(\tau) = \log \left[ \frac{1}{M} \sum_{r=1}^{M} D_1(\theta^r_s, \tau, \lambda_1) \right] - \log \left[ \frac{1}{M} \sum_{r=1}^{M} D_0(\theta^r_s, \tau, \lambda_0) \right]. \quad (5.19) \]
The terms $D_i(\theta^*_r, \tau, \lambda_i)$ inside the summations of (5.18) and (5.19) are iid random variables that have expected values, with respect to $m_i(\theta^*|\lambda_i)$, of

$$E_{\lambda_i}(D_i(\theta^*_r, \tau, \lambda_i)) = \int_{\Theta^*} \exp\{Q_i(t|\tau)\} \, d\mu(t) = k_i(\tau), \text{ for } i = 0, 1.$$  

The iid structure of $\theta^*_1, \theta^*_2, \ldots$, allows application of the central limit theorem in (5.18) to give,

$$M^{1/2} [k_{iM}(\tau) - k_i(\tau)] \xrightarrow{D} N(0, V_{i1}(\tau)), \quad (5.20)$$

and an additional log transformation yields, from (5.19),

$$M^{1/2} [L_M(\tau) - L(\tau)] \xrightarrow{D} N(0, V_2(\tau)), \quad (5.21)$$

where $V_2(\tau) = (V_{11}(\tau)/k_1^2(\tau)) + (V_{01}(\tau)/k_0^2(\tau))$ because $\theta^*_1 : r = 1, \ldots, M$ and $\theta^*_2 : r = 1, \ldots, M$ are independent. In equations (5.20) and (5.21), $V_{i1}(\tau)$'s and $V_2(\tau)$ may easily be shown to exist, although their explicit form will generally be unavailable, and will require estimation in practice. We would like to make use of these results to assess (1) the precision in Monte Carlo estimates $L_M(\tau)$ and (2) the possible need to select a new importance sampling distribution at step 6 of the overall estimation strategy.

In principle, one could assess the precision of $L_M(\tau)$ through the use of (5.21) for two one-time samples drawn from two importance distributions and any value of $\tau \in \mathcal{T}$. Thus, returning to the use of the superscript $(q)$, at the beginning of a Newton-Raphson cycle in step 5 of the overall estimation procedure (with current parameter $\tau^{(q-1)}$) we could select a sample size $M^{(q)}$ based on the resulting level of precision in $L_{M^{(q)}}(\tau^{(q-1)})$, and base decisions of convergence on whether that level of precision had changed by the end of the cycle, at $L_{M^{(q)}}(\tau^{(q)})$; recall that $\tau^{(q)}$ maximizes $L_{M^{(q)}}(\tau)$. Unfortunately, two aspects of the interplay between the Monte Carlo sample size, the importance sampling distribution used, and the asymptotics of (5.21) vitiate the usefulness of this approach. First, in the actual analysis of data, $V_{2}(\tau)$ must be estimated as a function of $k_0(\tau)$ and $k_1(\tau)$, which will itself be estimated as $k_{0,M^{(q)}}(\tau)$ and $k_{1,M^{(q)}}(\tau)$, respectively. Thus, use of (5.21) to estimate Monte Carlo error for the purpose of determining whether or not $M$ is sufficiently large involves a circular argument. That is, the Monte Carlo error in $L_{M^{(q)}}(\tau)$ can only be estimated on the basis of (5.21) if we are already assured of the sufficiently precise estimates of $k_i(\tau)$'s, in which case the issue of sample size determination is moot. The second problematic aspect in practical use of (5.21) is that, for the given importance distributions $m_i(\theta|\lambda_i^{(q)})$'s, the Monte Carlo sample size $M^{(q)}$ required for this result to provide a reasonable approximation to the error in $L_{M^{(q)}}(\tau^{(q-1)})$ may no longer be adequate for a similar assessment of $L_{M^{(q)}}(\tau^{(q)})$. Nevertheless, we believe that the combined use of (5.20) and (5.21) can form the basis for a practical assessment.
of Monte Carlo error, and convergence in the overall estimation strategy, under the following heuristic argument.

Under the assumption that a Monte Carlo sample size $M^{(q)}$ is large enough for (5.21) to provide an adequate indication of the error in $L_{M^{(q)}}(\tau^{(q-1)})$, we can estimate $V_{11}(\tau^{(q-1)})$ as the sample variance of $D_i(\theta^{(i)}_r, \tau^{(q-1)}, \lambda^{(q)}_i)$, where $\theta^{(i)}_1, \ldots, \theta^{(i)}_{M^{(q)}}$ are sampled values from the importance distribution $m_i(\theta|\lambda^{(q)}_i)$. Call this estimate $\hat{V}_{11}(\tau^{(q-1)}); i = 0, 1$. Then $V_2(\tau^{(q-1)})$ is estimated as $V_2(\tau^{(q-1)}) = (\hat{V}_{11}(\tau^{(q-1)})/k_{1M^{(q)}}^{2}(\tau^{(q-1)})) + (\hat{V}_{01}(\tau^{(q-1)})/k_{0M^{(q)}}^{2}(\tau^{(q-1)}))$, and an $(1 - \alpha) \times 100\%$ interval estimate of the Monte Carlo error $L_{M^{(q)}}(\tau^{(q-1)}) - L(\tau^{(q-1)})$ formed as $\pm z_{\alpha/2}[V_2(\tau^{(q-1)})/M^{(q)}]^{1/2}$. This provides an estimate of the Monte Carlo error in estimation of the log likelihood at the beginning of a Newton-Raphson cycle in the overall estimation procedure, under the starting assumption which will be addressed in Section 5.4.4.

Notice that the estimators $V_{11}(\tau)$'s involves only the computation of sample variances of simulated values and thus do not suffer the same deficiency of being calculated as a function of $k_iM^{(q)}$'s as does $V_2(\tau)$. Thus, a maximum value of two increases in estimated values from $\hat{V}_{11}(\tau^{(q-1)})$ to $\hat{V}_{11}(\tau^{(q)})$, for $i = 0, 1$, indicates a deterioration in precision over the course of a Newton-Raphson cycle, and the need to select a new importance distribution for an additional cycle. What is required for this use of estimated values of $\hat{V}_{11}(\tau)$'s is a meaningful scale against which to judge the magnitude of changes. This is not provided automatically, since $V_{11}(\tau)$'s are the (asymptotic) variances of normalizing constants alone. Now, consider a small neighborhood $\mathcal{N}(\hat{\tau})$ of the true MLE $\hat{\tau}$, so that $k_i(\tau) \approx k_i(\hat{\tau})$ for any $\tau \in \mathcal{N}(\hat{\tau})$, and for any $i = 1, 0$. Then the change in $V_2(\tau)$ over two parameter values $\tau', \tau'' \in \mathcal{N}(\hat{\tau})$ is such that

$$\left| \frac{V_2(\tau') - V_2(\tau'')}{V_2(\tau')} \right| \leq \max \left\{ \left| \frac{V_{11}(\tau') - V_{11}(\tau'')}{V_{11}(\tau')} \right|, \left| \frac{V_{01}(\tau') - V_{01}(\tau'')}{V_{01}(\tau')} \right| \right\} \equiv \gamma(\tau', \tau'').$$

(5.22)

Note that the assumption inherent to this expression is that $\mathcal{N}(\hat{\tau})$ is small enough so that the normalizing constant is approximately equvalued within this neighborhood, but not necessarily so small that the same is true of $V_2(\tau)$. We judge the change in precision of $L_{M^{(q)}}(\tau)$ from evaluation at $\tau^{(q-1)}$ to evaluation at $\tau^{(q)}$ by estimating the above quantity as

$$\tilde{V}_i(\tau^{(q-1)}, \tau^{(q)}) \equiv \gamma(\tau^{(q-1)}, \tau^{(q)}).$$

While there is no guarantee that $\tilde{V}_i(\tau^{(q-1)}, \tau^{(q)})$ accurately reflects the proportional change in Monte Carlo error of $L_{M}(\tau)$ at all stages of the estimation procedure, if it is 'large', (as defined in Section 5.4.3), then it cannot be that the overall strategy has resulted in convergence. In this case, it will be
required to select new importance distributions and conduct an additional cycle of Newton-Raphson iterations.

5.4.3 Assessing Convergence in Practice

The overall estimation strategy has been formed on the basis of constructing a sequence of approximations to the log likelihood at various points in the parameter space \( T \). At iteration \( q \) of the overall procedure, we call \( \tau^{(q+1)} \) the 'point of origin' of the current approximation \( L_{M^{(q)}}(\tau) \), and \( \tau^{(q)} \) the 'point of maximization' of the approximation. As indicated in Section 5.4.2, we continue to form new approximations so long as we have an indication, based on (5.21), that the precision of the approximation may have changed between its point of origin and its point of maximization. To make this prescription more concrete, we propose that, at the end of step 5 in the overall estimation procedure, new importance distributions be chosen (using the approach of Section 5.4.1) if \( V_1(\tau^{(q-1)}, \tau^{(q)}) \geq \delta \). While arbitrary, if

\[
\frac{|V_{11}(\tau^*) - V_{11}(\tau^*)|}{V_{11}(\tau^*)} = \frac{|V_{01}(\tau^*) - V_{01}(\tau^*)|}{V_{01}(\tau^*)},
\]

our suggested value for the threshold is \( \delta = 0.10 \), based on the rationale that, if \( V_2(\tau^q) = V_2(\tau^{(q-1)})(1 + \delta) \), then an interval formed as \( \pm z_{0.1/2}[V_2(\tau^{(q)})/M^{(q)}]^{1/2} \) will be about 1.1 times as long as the corresponding interval formed with \( V_2(\tau^{(q-1)}) \). If the relative change in \( V_{11} \) is quite different from that in \( V_{01} \), a little bit larger value, say 0.2, for \( \delta \) is acceptable because the bound in (5.22) will not be very sharp.

If \( V_1(\tau^{(q-1)}, \tau^{(q)}) < \delta \), then it would appear that an additional iteration of the estimation procedure is not needed. But, because this is only indirect evidence of convergence in the overall algorithm, we deem it alone to be an insufficient criterion. Therefore, we make use of two more familiar quantities as checks on convergence, namely \( |L_{M^{(q)}}(\tau^{(q-1)}) - L_{M^{(q)}}(\tau^{(q)})| \) and \( ||\tau^{(q-1)} - \tau^{(q)}|| \), where \( ||a|| \) is the Euclidean norm of a vector \( a \). These quantities are used only as additional checks on convergence, rather than as original convergence criteria, because small values may result by chance unless Monte Carlo precision has remained stable throughout an entire Newton-Raphson cycle (i.e., the maximization of a given approximation to the log likelihood).

5.4.4 Determination of Baseline Monte Carlo Error

The estimation strategy proposed in this article depends critically on the assumption that a suitable level of precision has been attained in each approximation of the log likelihood at its point of origin. We call this the 'baseline' level of precision for each approximation formed. If the baseline level of Monte
Carlo precision for a given approximation $L_M(\tau)$ is insufficient in the first place, numerical instability in estimated quantities will render even the most careful efforts to assess convergence of the overall estimation procedure futile. As indicated in Section 5.4.2, we assess the baseline level of precision for the current approximation through use of the estimated interval $\pm z_{\alpha/2} \left[ \hat{V}_2(\tau^{(x-1)}) / M^{(q)} \right]^{1/2}$. The adequacy of this interval as an estimate of the true Monte Carlo error depends on the assumption that $M^{(q)}$ is large enough to provide sufficient precision in estimation of $k_i(\tau)$’s for estimation of $V_2(\tau)$. Our approach to providing justification for a selected Monte Carlo sample size $M^{(q)}$ depends firstly on beginning the estimation procedure in step 1 of Section 5.3 with a value $\tau^{(0)}$ which is itself a solution to a log likelihood approximation. This is discussed in Section 5.5. Then, for an initial Monte Carlo sample size, $M_1^{(1)}$ say, an $(1 - \alpha) \times 100\%$ interval estimate of Monte Carlo error is computed. The sample size is increased to $M_2^{(1)}$ and an additional interval computed, where the estimate of $V_2(\tau^{(0)})$ is based on the $M_2^{(1)}$ samples but the interval on the initial sample size $M_1^{(1)}$. The process is repeated until interval length shows no appreciable decrease, and that sample size is selected a $M_5^{(1)}$.

5.5 Starting Value

As briefly mentioned in Section 5.1, the starting value $\tau^{(0)}$ in the first step of the overall procedure described in Section 5.3 plays an important role in that a good starting value reduces the number of repetitions of the entire procedure (steps 2 through 6) required to locate the MLE of $\tau$. We have to expend considerable effort in reducing the number of repetitions of the entire procedure because selection of sampling distributions in step 2 of the procedure requires the use of a Markov chain sampler which could be burden and drawing samples in step 3 is the most time-consuming part of the procedure. Getting a good starting value means that it should be as close as the true MLE. Therefore, it is better for $\tau^{(0)}$ to be an approximate solution to the total estimation problem. Laplace approximation can be used to approximate the two normalizing constants, and then we can maximize the approximated log likelihood which requires the use of two nested iterative procedures. This method is the only possibility without Monte Carlo sampling for the mixture model.

Consider the exact log likelihood of equation (5.10), in which

$$k_i(\tau) = \int \exp\{Q_i(\theta | \tau)\} d\theta,$$

for $i = 0$ or 1. A Laplace approximation of $k_i(\tau)$ is

$$k_i(\tau) \approx (2\pi)^{n/2} |\text{det}(\hat{\Sigma}_i)|^{1/2} \exp[Q_i(\hat{\theta}_i | \tau)],$$
where $\hat{\theta}_i^1$ is the mode of $Q_i(\theta|\tau)$ with respect to $\theta \in \Theta^i$, considered as a function of $\theta$ for fixed $\tau$, and $\hat{\Sigma}_i$ is the minus inverse Hessian of $Q_i(\theta|\tau)$ at $\hat{\theta}_i^1$. An approximate solution to the estimation problem then results from maximizing in $\tau$ the quantity

$$\hat{L}(\tau) = 0.5 \log \left[ \text{det}(\hat{\Sigma}_1) - \text{det}(\hat{\Sigma}_0) \right] + \left[ Q_1(\hat{\theta}_1^1|\tau) - Q_1(\hat{\theta}_0^1|\tau) \right]. \quad (5.23)$$

The difficulty in incorporating this method into the maximization process is that we have to calculate two modes, $\hat{\theta}_i^1$ and $\hat{\theta}_0^1$, numerically at each iteration step if it’s impossible to get the explicit forms of $\hat{\theta}_r^1$ and $\hat{\theta}_r^0$ as a function of $\tau$. If $\tau$ is of low dimension, direct search algorithms seems to perform adequately. In this case, we first need two codes for getting two modes. Those codes are quite similar, so slight change of one could be the other. If $\tau$ is of high dimension, we can use the method like a Newton-Raphson algorithm. In this case, we need extra codes for modes of some functions related to the first derivatives and the second derivatives. Let’s take a look at the first derivatives for details. The $j^{th}$ element of the first derivatives vector of $L(\tau)$ is

$$\frac{\partial L(\tau)}{\partial \tau_j} = E_{fg} \left( \frac{\partial Q_0(\theta|\tau)}{\partial \tau_j} \right) - E_g \left( \frac{\partial Q_0(\theta|\tau)}{\partial \tau_j} \right), \quad (5.24)$$

where $E_{fg}(\cdot)$ denotes the expectation with respect to $\exp\{Q_1(\theta|\tau)/k_1(\tau)\}$, and $E_g(\cdot)$ with respect to $g(\theta|\tau)$. Because

$$E_{fg} \left( \frac{\partial Q_0(\theta|\tau)}{\partial \tau_j} \right) = \frac{\int \frac{\partial Q_0(\theta|\tau)}{\partial \tau_j} \exp\{Q_1(\theta|\tau)\} d\theta}{\int \exp\{Q_1(\theta|\tau)\} d\theta},$$

we need another code for mode of the numerator to get the Laplace approximation of the first term of (5.24). Therefore the number of modes needed for each iteration of Newton-Raphson algorithm is $1 + 2p + \frac{p(p+1)}{2}$ when the dimension of $\tau$ is $p$.

Whatever maximization methods are used, the maximization of (5.23) in $\tau$ need not be accomplished to exacting specifications since the goal is to find an appropriate starting value $\tau^{(0)}$ for use in the overall strategy of Section 5.3. If the selected starting value is poor, this fact will become apparent in the application of that strategy. The procedure of maximization of (5.23) may be as follows:

- **Step1**: With $\hat{\tau}^k$, calculate the two modes for $\hat{L}(\hat{\tau}^k)$. If needed, calculate additional modes for derivatives.

- **Step2**: Update $\hat{\tau}^k$ to $\hat{\tau}^{k+1}$ by whatever maximization method.

- **Step3**: Continue Step1 and Step2 until convergence occur.

The above procedure requires a nested procedure for determination of modes, $\hat{\theta}_r^i$'s for a given value of $\tau$. At least, for the classes we introduced in Section 5.2, we have found standard iterative algorithms...
depending on the second derivative information appropriate for finding this value. To get mode \( \theta_i^2 \) for the third class of models in Section 5.2, the following derivatives may be used.

\[
\frac{\partial Q_0(\theta; r)}{\partial \theta(s_i)} = \alpha_{11} - \alpha_{12} b'(\theta(s_i)) + \sum_{j \in D_i} \{n_{ij}[-b'(\theta(s_j)) - \theta(s_j)b'(\theta(s_i))]\} \tag{5.25}
\]

\[
\frac{\partial^2 Q_0(\theta; r)}{\partial \theta(s_i)^2} = -\alpha_{12} b''(\theta(s_i)) + \sum_{j \in D_i} \{n_{ij}[-\theta(s_j)b''(\theta(s_i))]\} \tag{5.26}
\]

\[
\frac{\partial^2 Q_0(\theta; r)}{\partial \theta(s_i) \partial \theta(s_j)} = n_{ij}[-b'(\theta(s_j)) - b'(\theta(s_i))] \tag{5.27}
\]

For the derivatives of \( Q_1 \), just replace \( \alpha_{11} \) with \( \alpha_{11} + \phi z(s_i) \) and \( \alpha_{12} \) with \( \alpha_{12} + \phi \) in (5.25)-(5.27).

If one wants to use any algorithm making use of the derivatives in step 2, one may need the modes for derivatives. For the modes associated with derivatives, for example, the mode associated with the numerator of \( E[\frac{\partial Q_0(\theta|r)}{\partial \theta(s_i)}] \), we will find a mode of

\[
\log[\frac{\partial Q_0(\theta|r)}{\partial \theta(s_i)}] + Q_0(\theta|r); \theta \in \Theta.
\]

Therefore, we need additional derivatives such as

\[
\frac{\partial}{\partial \theta(s_i)} \left( \log[\frac{\partial Q_0(\theta|r)}{\partial \theta(s_i)}] + Q_0(\theta|r) \right), \quad \frac{\partial^2}{\partial \theta(s_i)^2} \left( \log[\frac{\partial Q_0(\theta|r)}{\partial \theta(s_i)}] + Q_0(\theta|r) \right), \quad \text{and} \quad \frac{\partial^2}{\partial \theta(s_i) \partial \theta(s_j)} \left( \log[\frac{\partial Q_0(\theta|r)}{\partial \theta(s_i)}] + Q_0(\theta|r) \right). \tag{5.28}
\]

As seen, the codes for getting the modes of \( Q_0, Q_1 \) and the derivatives are quite similar, so the programming is very efficient. One thing we have to be careful of when we deal with the derivatives is to make sure that terms in logarithm must be positive.

5.6 Examples

In this section, we will illustrate how to apply the methods we developed throughout the previous sections by analyzing the data from the Environmental Protection Agency (EPA).

5.6.1 Data Description

One index measuring the degree of pollution for a tree is Crown Die back (CD) which measures as a percentage so that the range of measured values is from 0 to 100. The higher CD value a tree has, the more polluted it is. We have 36 locations which are characterized by longitudes and latitudes. In each location, \( s_i \), there are \( n_i \) birch trees, and each tree has its CD index. We calculated the 75\(^{th}\) percentile CD value for all the trees. We want to consider the trees whose CD index are greater or equal to this 75\(^{th}\) percentile as polluted. We counted the number of trees considered polluted in each location. The
data we will analyze are those 36 counted numbers, \( z(s_i) \), with \( n_i \). The first plot in Figure 5.1 shows
the locations with \( \frac{z(s_i)}{n_i} \), estimated polluted ratio, and the second plot shows site numbers. The distance
between locations is measured in kilometers. The first step is to define the dependence structure. Let
\( u_i \) denote the longitude and \( v_i \) the latitude of each sampling location shown in Figure 1, and define the
location indices 

\[ s_i = (u_i, v_i)^T; \quad i = 1, \ldots, n, \]

where in this example \( n = 36 \). The neighborhood structure
was defined from physical proximity of sampling locations as,

\[ h \in D_i \text{ if } \| s_h - s_i \| \leq \kappa \]

where \( \kappa \) is 48 kilometers, \( \| x \| \) is the Euclidean norm of \( x \) and \( D_i \) is the dependence index set of location
\( s_i \), as defined in Section 5.2. In this example, \( \kappa \) was fixed, not treated as an additional parameter in
the model. Table 5.1 provides the all information on data.

5.6.2 Modeling the Birch Tree Data

The data model is independent Binomial distribution given parameters, i.e.,

\[ Z(s_i) | \theta \sim \text{Binomial}(n_i, \theta(s_i)), \]

![Observed Probabilities and Site Numbers for Birch trees](image-url)
Table 5.1 Summary for the Birch Data

<table>
<thead>
<tr>
<th>location</th>
<th>$n_i$</th>
<th>$z(s_i)$</th>
<th>$\frac{z(s_i)}{n_i}$</th>
<th>neighbors</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>5</td>
<td>3</td>
<td>0.60</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>11</td>
<td>2</td>
<td>0.18</td>
<td>3</td>
</tr>
<tr>
<td>3</td>
<td>6</td>
<td>2</td>
<td>0.33</td>
<td>2</td>
</tr>
<tr>
<td>4</td>
<td>13</td>
<td>3</td>
<td>0.23</td>
<td>5</td>
</tr>
<tr>
<td>5</td>
<td>7</td>
<td>2</td>
<td>0.29</td>
<td>4,6,8,9</td>
</tr>
<tr>
<td>6</td>
<td>7</td>
<td>1</td>
<td>0.14</td>
<td>5,7,14</td>
</tr>
<tr>
<td>7</td>
<td>13</td>
<td>2</td>
<td>0.15</td>
<td>6,9</td>
</tr>
<tr>
<td>8</td>
<td>5</td>
<td>2</td>
<td>0.40</td>
<td>5,9</td>
</tr>
<tr>
<td>9</td>
<td>5</td>
<td>1</td>
<td>0.20</td>
<td>5,7,8</td>
</tr>
<tr>
<td>10</td>
<td>8</td>
<td>3</td>
<td>0.38</td>
<td></td>
</tr>
<tr>
<td>11</td>
<td>9</td>
<td>1</td>
<td>0.11</td>
<td></td>
</tr>
<tr>
<td>12</td>
<td>6</td>
<td>3</td>
<td>0.50</td>
<td></td>
</tr>
<tr>
<td>13</td>
<td>5</td>
<td>1</td>
<td>0.20</td>
<td></td>
</tr>
<tr>
<td>14</td>
<td>15</td>
<td>3</td>
<td>0.20</td>
<td>6</td>
</tr>
<tr>
<td>15</td>
<td>18</td>
<td>8</td>
<td>0.44</td>
<td>16,17,26</td>
</tr>
<tr>
<td>16</td>
<td>6</td>
<td>3</td>
<td>0.50</td>
<td>15,17,26</td>
</tr>
<tr>
<td>17</td>
<td>13</td>
<td>1</td>
<td>0.08</td>
<td>15,16</td>
</tr>
<tr>
<td>18</td>
<td>7</td>
<td>2</td>
<td>0.29</td>
<td>20,21</td>
</tr>
<tr>
<td>19</td>
<td>14</td>
<td>2</td>
<td>0.14</td>
<td>32</td>
</tr>
<tr>
<td>20</td>
<td>22</td>
<td>3</td>
<td>0.14</td>
<td>18,21</td>
</tr>
<tr>
<td>21</td>
<td>6</td>
<td>2</td>
<td>0.33</td>
<td>18,20,22</td>
</tr>
<tr>
<td>22</td>
<td>7</td>
<td>1</td>
<td>0.14</td>
<td>21</td>
</tr>
<tr>
<td>23</td>
<td>16</td>
<td>12</td>
<td>0.75</td>
<td>24</td>
</tr>
<tr>
<td>24</td>
<td>7</td>
<td>3</td>
<td>0.43</td>
<td>23</td>
</tr>
<tr>
<td>25</td>
<td>8</td>
<td>1</td>
<td>0.13</td>
<td></td>
</tr>
<tr>
<td>26</td>
<td>8</td>
<td>2</td>
<td>0.25</td>
<td>15,16</td>
</tr>
<tr>
<td>27</td>
<td>5</td>
<td>1</td>
<td>0.20</td>
<td>28</td>
</tr>
<tr>
<td>28</td>
<td>7</td>
<td>2</td>
<td>0.29</td>
<td>27</td>
</tr>
<tr>
<td>29</td>
<td>5</td>
<td>3</td>
<td>0.60</td>
<td>30</td>
</tr>
<tr>
<td>30</td>
<td>11</td>
<td>2</td>
<td>0.18</td>
<td>29,31</td>
</tr>
<tr>
<td>31</td>
<td>8</td>
<td>5</td>
<td>0.63</td>
<td>30</td>
</tr>
<tr>
<td>32</td>
<td>11</td>
<td>4</td>
<td>0.36</td>
<td>19</td>
</tr>
<tr>
<td>33</td>
<td>13</td>
<td>5</td>
<td>0.38</td>
<td></td>
</tr>
<tr>
<td>34</td>
<td>26</td>
<td>20</td>
<td>0.77</td>
<td></td>
</tr>
<tr>
<td>35</td>
<td>6</td>
<td>3</td>
<td>0.50</td>
<td></td>
</tr>
<tr>
<td>36</td>
<td>7</td>
<td>2</td>
<td>0.29</td>
<td></td>
</tr>
</tbody>
</table>
and random parameter model is conditionally specified Beta distribution,
\[
\theta(s_i) | \{\theta(s_j) : j \in D_i\} \sim \text{Beta}(A_{i1}(|\{\theta(s_j) : j \in D_i\}|), A_{i2}(|\{\theta(s_j) : j \in D_i\}|)),
\]
where \(D_i\) is in the last column of Table 5.1, and
\[
A_{i1}(|\{\theta(s_j) : j \in D_i\}|) = \alpha_1 + \eta \sum_{j \in D_i} \log(1 - \theta(s_j))
\]
\[
A_{i2}(|\{\theta(s_j) : j \in D_i\}|) = \alpha_2 + \eta \sum_{j \in D_i} \log(\theta(s_j))
\]
That is to say, we will use the third class for Beta distribution because the data are expected to have positive correlations. From this modeling, we have three hyper parameters to be estimated, notated by \(\tau \equiv (\alpha_1, \alpha_2, \eta)'\).
Then, letting \(\theta = (\theta(s_1), \ldots, \theta(s_n))'\),
\[
f(x|\theta) = \exp \left[ \sum \{ z(s_i) \log\left( \frac{\theta(s_i)}{1 - \theta(s_i)} \right) \} + \sum \{ n_i \log(1 - \theta(s_i)) \} + c(z, n_1, \ldots, n_n) \right].
\]
\[
h(\theta; \tau) = \exp \left[ \alpha_1 \sum \{ \log(\theta(s_i)) \} + \alpha_2 \sum \{ \log(1 - \theta(s_i)) \} + \eta \sum_{i=1}^{n} \sum_{i \neq j \in D_i} \{ \log(\theta(s_i)) \log(1 - \theta(s_j)) + \log(\theta(s_j)) \log(1 - \theta(s_i)) \} \right]
\]
so,
\[
f(x|\theta) h(\theta; \tau) = \exp \left[ \sum \{ (\alpha_1 + z(s_i)) \log(\theta(s_i)) \} + \sum \{ (n_i - z(s_i) + \alpha_2) \log(1 - \theta(s_i)) \} + \eta \sum_{i=1}^{n} \sum_{i \neq j \in D_i} \{ \log(\theta(s_i)) \log(1 - \theta(s_j)) + \log(\theta(s_j)) \log(1 - \theta(s_i)) \} \right]
\]
The form of \(Q_1(\theta|\tau)\) is same as that of \(Q_0(\theta|\tau)\), replacing \(\alpha_1\) and \(\alpha_2\) with \(\alpha_1 + z(s_i)\) and \(\alpha_2 + n_1 - z(s_i)\), respectively. Therefore the conditionally specified distributions to define \(Q_1(\theta|\tau)\) are also same as those to define \(Q_0(\theta|\tau)\). We want MLE of
\[
\log \left( \int f(x|\theta) h(\theta; \tau) d\theta \right) - \log \left( \int h(\theta; \tau) d\theta \right),
\]
with respect to \(\tau\).
5.6.3 Selection of Importance Densities

According to the selection procedure of importance distributions, \( m_1 \) and \( m_0 \) in Section 5.4, appropriate two densities for \( \exp[Q_1(\theta; \tau)]/k_1(\tau) \) and \( \exp[Q_0(\theta; \tau)]/k_0(\tau) \) have to be selected.

For \( \exp[Q_0(\theta; \tau)]/k_0(\tau) \), one of the choices for an importance density in (5.15) is

\[
m_0(\theta; \lambda_0) = \frac{\exp[\sum \{\alpha_{i1}^0 \log(\theta(s_i))\} + \sum \{\alpha_{i2}^0 \log(1 - \theta(s_i))\}]}{\kappa_0(\lambda_0)},
\]

where \( \lambda_0 = (\alpha_{i1}^0, \ldots, \alpha_{i1}^0, \alpha_{i2}^0, \ldots, \alpha_{i2}^0) \), and the normalizing constant, \( \kappa_0(\lambda_0) \), is

\[
\left( \prod_{i=1}^n \frac{\Gamma(\alpha_{i1}^0 + \alpha_{i2}^0)}{\Gamma(\alpha_{i1}^0)\Gamma(\alpha_{i2}^0)} \right)^{-1}.
\]

Now we have to find \( \alpha_{i1}^0 \)'s, and \( \alpha_{i2}^0 \)'s so that \( \exp[Q_0(\theta; \tau)]/k_0(\tau) \) and \( m_0(\theta; \lambda_0) \) have the same first two moments marginally. First, generate sample from the normalized density of (5.30) using Gibbs sampler, and get the sample means and sample variances of \( \theta(s_i) \)'s. Let \( \mu_i \) be the sample mean of \( \theta(s_i) \), and \( \sigma_i^2 \) be the sample variance of \( \theta(s_i) \). Then we can expect that \( \lambda_0 \) will give the same first two moments as \( \exp[Q_0(\theta; \tau)]/k_0(\tau) \) if \( \lambda_0 \) satisfies the followings:

\[
\alpha_{i1}^0 = \frac{\mu_i^2 - \mu_i^2 - \mu_i^2}{\sigma_i^2}, \quad \text{and} \quad \alpha_{i2}^0 = \frac{\alpha_{i2}^0(1 - \mu_i)}{\mu_i}.
\]

For the selection of \( m_1(\theta; \lambda_1) \), we can do the similar procedure as \( m_0 \). That is to say, find \( \alpha_{i1}^1 \)'s, and \( \alpha_{i2}^1 \) such that \( m_1(\theta; \lambda_1) \), where

\[
m_1(\theta; \lambda_1) = \frac{\exp[\sum \{\alpha_{i1}^1 \log(\theta(s_i))\} + \sum \{\alpha_{i2}^1 \log(1 - \theta(s_i))\}]}{\kappa_1(\lambda_1)},
\]

has the same first two marginal moments as \( \exp[Q_1(\theta; \tau)]/k_1(\tau) \). The \( \lambda_1 \) can be determined as in (5.35) using the sample means and variances obtained from the normalized density of (5.31). If we use (5.33) and (5.36) for the importance densities, the form of MC estimation of log likelihood \( L_M(\tau) \) is

\[
\log \left( \frac{\kappa_1}{\kappa_0} \right) + \log \left( \sum_{j=1}^N \exp \left[ \sum_{i=1}^n \left\{ (y(s_i) + \alpha_1 - \alpha_{i1}^1) \log(\theta_j^1) + (n_i - y(s_i) + \alpha_2 - \alpha_{i2}^1) \log(1 - \theta_j^1) \right\} + \eta S(\theta_j^1) \right] \right)
\]

\[
+ \log \left( \sum_{j=1}^N \exp \left[ \sum_{i=1}^n \left\{ (\alpha_1 - \alpha_{i1}^1) \log(\theta_j^0) + (\alpha_2 - \alpha_{i2}^0) \log(1 - \theta_j^0) \right\} + \eta S(\theta_j^0) \right] \right),
\]

where \( \theta_j^1 = (\theta_{j1}, \ldots, \theta_{j1})' \) is the \( j \)'th sample from \( m_1(\theta; \lambda_1) \), and \( \theta_j^0 = (\theta_{j1}, \ldots, \theta_{j1})' \) is the \( j \)'th sample from \( m_0(\theta; \lambda_0) \), and, for \( i = 1, 0 \),

\[
S(\theta_j^i) = \sum_{r=1}^n \sum_{r \leq k \in D} \{ \log(\theta_{ij}^r) \log(1 - \theta_{ij}^r) + \log(\theta_{ij}^r) \log(1 - \theta_{ij}^r) \}.
\]
5.6.4 Estimation Process

For the birch data, we will maximize (5.32) first using the Laplace approximation and then MC approximation of (5.32) will be maximized with the answer from the Laplace method as the starting value. The maximum likelihood estimate of $\tau \equiv (\alpha, \beta, \eta)^T$ was found by using the overall estimation strategy of Section 5.3. For a final model, we obtained $\hat{\alpha} = 4.121, \hat{\beta} = 6.524$ and $\hat{\eta} = -4.489$.

Because the primary purpose of this example is to illustrate the estimation strategy proposed in this chapter, a more detailed examination of the estimation history is provided here and in Table 5.2. A starting value was determined from the Laplace approximation method of Section 5.5 as $\tau^{(0)} = (3.583, 5.774, -3.733)$. Using this value, each sample of size 200,000 was generated from $\exp[Q_0(\theta, \tau^{(0)})]/K_0(\tau^{(0)})$ and $\exp[Q_1(\theta, \tau^{(0)})]/K_1(\tau^{(0)})$ respectively, using a Gibbs algorithm. Two sets of $n = 36$ marginal means and variances from these samples were used to select values for $\lambda^0$ and $\lambda^1$, and sampling distributions constructed as in equation (5.33) and (5.36). Assessment of baseline Monte Carlo error as described in Section 5.5.4 resulted in an initial Monte Carlo sample size of $M^{(1)} = 600,000$, which yielded $\hat{V}_2(\tau^{(0)}) = 549.6040$ and an 95% interval estimate of Monte Carlo error, $[L_{M^{(1)}}(\tau^{(0)}) - L(\tau^{(0)})]$, of $\pm 0.0514$. Additional cycles of Newton-Raphson iterations were conducted using this same Monte Carlo sample size of 600,000. The first cycle of Newton-Raphson then began with $L_{M^{(1)}}(\tau^{(0)}) = -210.6008$ and, after 4 iterations, produced the updated parameter value $\tau^{(1)} = (4.118, 6.549, -4.484)^T$. The values of $\hat{V}_1(\tau^{(0)}, \tau^{(1)})$ was $-1.0$, indicating the need to select new importance sampling distributions and enter a second cycle of Newton-Raphson iterations. To illustrate the need for basing this decision on the value of $\hat{V}_1(\tau^{(0)}, \tau^{(1)})$, rather than change in the estimated value of $V_2(\tau)$, an 95% interval estimate of $[L_{M^{(1)}}(\tau^{(1)}) - L(\tau^{(1)})]$, based on the value $\hat{V}_2(\tau^{(1)})$ produced at the end of the first cycle (i.e., using the original Monte Carlo sample), was $\pm 0.0558$. Thus, if change in Monte Carlo precision were based on such intervals instead of the more reliable value of $\hat{V}_1(\tau^{(0)}, \tau^{(1)})$, we would have erroneously concluded that the precisions of $L_{M^{(1)}}(\tau^{(0)})$ and $L_{M^{(1)}}(\tau^{(1)})$ were similar. With $\hat{V}_1(\tau^{(0)}, \tau^{(1)}) = -1$, we cannot guarantee the reliability of precision in $L_{M^{(1)}}(\tau^{(1)})$. Thus, although $L_{M^{(1)}}(\tau^{(1)}) = -210.57456$ is the maximum value of $L_{M^{(1)}}(\tau)$ by definition, it cannot be accepted as a reliable estimate of $L(\tau^{(1)})$. The new sampling distributions were chosen by generating another Gibbs sample of size 200,000 from the conditional specifications, now with the parameter $\tau^{(1)}$ and following the prescription of Section 5.4.1. Another cycle of Newton-Raphson iterations yielded the estimate $\tau^{(2)} = (4.127, 6.433, -4.548)^T$ and the relative changes in $V_{11}$ and in $V_{01}$ were 0.204 and 0.342 respectively. These values were improved but we need more accuracies, so one additional iteration of the overall estimation procedure was conducted, resulting in the final estimate $\tau^{(3)} = (4.121, 6.524, -4.489)^T$ and
values 0.083 and 0.164 for the relative changes in $V_{11}$ and in $V_{01}$. Based on these values, we were now willing to accept $L_M(\tau^{(2)}) = -210.56927$ and $L_M(\tau^{(3)}) = -210.56557$ as similar enough in precision to allow assessment of $L_M(\tau^{(3)}) - L_M(\tau^{(2)}) = 0.003699$ as indicative of convergence; the values in Table 5.2 have been rounded to fourth decimal places. The Euclidean norm of the change in estimated parameter values was $||\tau^{(3)} - \tau^{(2)}|| = 0.10785$. The asymptotic variance-covariance matrix of $\tau^{(3)} - \tilde{\tau}$, where $\tilde{\tau}$ is the true MLE, is

$$
\begin{pmatrix}
0.021 & 0.033 & -0.039 \\
0.033 & 0.055 & -0.061 \\
-0.039 & -0.061 & 0.078
\end{pmatrix}, \quad (5.38)
$$

which can be calculated as in Chapter 6.

5.6.5 Further Analysis

5.6.5.1 Dependence Test

If we want to test whether there is a positive correlation between trees which are neighbors or between the pollution probabilities, we could test whether $\eta = 0$ or not. Obviously, non-zero $\eta$ makes the correlations greater than zero. To test $H_0 : \eta = 0$, we need $\tilde{\eta}$ and its distribution with $\text{var}(\tilde{\eta})$. The asymptotic variance-covariance matrix in (5.38) is not for the error between MC MLE and true parameter as the data size increases, but for the error between MC MLE and true MLE due to the use of MC marginal log likelihood instead of true one. Therefore, we cannot use 0.078, the third element of (5.38), for the above test. Natural way is to use $\eta^{(3)}$, MC MLE, as $\tilde{\eta}$, and the third element of the inverse of $-\nabla^2 L_M(\tau^{(3)})$ divided by the size of data as $\text{var}(\tilde{\eta})$ with normality. The inverse of $-\nabla^2 L_M(\tau^{(3)})$
divided by the size of data is

\[
\begin{pmatrix}
0.143 & 0.201 & -0.189 \\
0.201 & 0.334 & -0.205 \\
-0.189 & -0.205 & 0.379
\end{pmatrix}.
\] (5.39)

From the fact that \( \eta^{(3)} = -4.489 \) and estimated asymptotic variance is 0.379, we can conclude that \( \eta \) is not zero, so there are positive correlations for pairs of trees (pollution probabilities) which are neighbors.

### 5.6.5.2 Estimation of Correlations

We know that there are positive correlations. But it is not easy to see how strong correlations are. To see this, we can generate samples at \( r^{(3)} \) and calculate the sample correlations. This correlation is the estimated correlation between \( \theta(s_i) \)'s with respect to the mixing density. What about the correlations between \( z(s_i) \) and \( z(s_j) \) if \( j \in D_i \)? For example, what is \( \text{corr}(z(s_2), z(s_3)) \)? This can be calculated using (5.8). In this example,

\[
\text{Corr}_h(z(s_2), z(s_3)) = \frac{n_2n_3 \text{Cov}_g[\theta(s_2), \theta(s_3)]}{\sqrt{n_2E_g[\theta(s_2) - \theta(s_2)^2] + n_2^2 \text{Var}_g[\theta(s_2)]} \sqrt{n_3E_g[\theta(s_3) - \theta(s_3)^2] + n_3^2 \text{Var}_g[\theta(s_3)]}}
\]

where \( n_j \) is the total number of trees in \( j^{\text{th}} \) site. The estimated correlations between pollution probabilities which are neighbors are from 0.36 to 0.55. The estimated correlations between trees which are neighbors are from 0.13 to 0.23. The details are in Table 5.3.

### 5.6.5.3 Final Comments

The model we used in this data set focuses on the correlated structure of random parameters, \( \theta \). Therefore, the correlations among trees are less than those among \( \theta_j \)'s. It is meaningful to see the correlations among \( \theta_j \)'s as the hidden process. From Table 5.3, the degree of correlation between trees does not depend on the distance between trees because we did not account for the distance factor in the dependence parameter. In this analysis, we did not consider geographical features of sampling regions. If it is reasonable that the distance is proportional to correlation, we can use other parametrizations which can provide desired results. Also, we can see that some correlations between trees, which are \( \text{not} \) neighbors, are not zeros. For example, site 6 and site 9 are not neighbors, but the correlation is not zero, which is possible because site 7 is the common neighbor of site 6 and site 9.
Table 5.3 Distances and Correlations for Birch Data

<table>
<thead>
<tr>
<th>Sites</th>
<th>Distance (in K-meter)</th>
<th>Corr. btw $\theta$'s</th>
<th>Corr. btw $z$'s</th>
</tr>
</thead>
<tbody>
<tr>
<td>2, 3</td>
<td>26.8</td>
<td>0.530</td>
<td>0.204</td>
</tr>
<tr>
<td>4, 5</td>
<td>27.2</td>
<td>0.415</td>
<td>0.137</td>
</tr>
<tr>
<td>5, 6</td>
<td>47.1</td>
<td>0.417</td>
<td>0.102</td>
</tr>
<tr>
<td>5, 7</td>
<td>54.6</td>
<td>0.307</td>
<td>0.100</td>
</tr>
<tr>
<td>5, 8</td>
<td>47.0</td>
<td>0.483</td>
<td>0.109</td>
</tr>
<tr>
<td>5, 9</td>
<td>47.9</td>
<td>0.493</td>
<td>0.107</td>
</tr>
<tr>
<td>6, 7</td>
<td>28.5</td>
<td>0.442</td>
<td>0.143</td>
</tr>
<tr>
<td>6, 9</td>
<td>48.3</td>
<td>0.306</td>
<td>0.060</td>
</tr>
<tr>
<td>7, 9</td>
<td>27.1</td>
<td>0.439</td>
<td>0.126</td>
</tr>
<tr>
<td>6, 14</td>
<td>47.1</td>
<td>0.436</td>
<td>0.156</td>
</tr>
<tr>
<td>15, 16</td>
<td>26.5</td>
<td>0.588</td>
<td>0.210</td>
</tr>
<tr>
<td>15, 17</td>
<td>27.4</td>
<td>0.546</td>
<td>0.258</td>
</tr>
<tr>
<td>15, 26</td>
<td>46.6</td>
<td>0.542</td>
<td>0.220</td>
</tr>
<tr>
<td>16, 17</td>
<td>27.1</td>
<td>0.546</td>
<td>0.183</td>
</tr>
<tr>
<td>16, 26</td>
<td>27.0</td>
<td>0.542</td>
<td>0.156</td>
</tr>
<tr>
<td>18, 20</td>
<td>27.1</td>
<td>0.556</td>
<td>0.239</td>
</tr>
<tr>
<td>18, 21</td>
<td>47.6</td>
<td>0.547</td>
<td>0.156</td>
</tr>
<tr>
<td>19, 32</td>
<td>28.5</td>
<td>0.528</td>
<td>0.260</td>
</tr>
<tr>
<td>20, 21</td>
<td>28.3</td>
<td>0.550</td>
<td>0.216</td>
</tr>
<tr>
<td>20, 22</td>
<td>55.9</td>
<td>0.258</td>
<td>0.114</td>
</tr>
<tr>
<td>21, 22</td>
<td>27.6</td>
<td>0.470</td>
<td>0.138</td>
</tr>
<tr>
<td>23, 24</td>
<td>47.7</td>
<td>0.531</td>
<td>0.237</td>
</tr>
<tr>
<td>27, 28</td>
<td>28.1</td>
<td>0.532</td>
<td>0.169</td>
</tr>
<tr>
<td>29, 30</td>
<td>26.9</td>
<td>0.498</td>
<td>0.170</td>
</tr>
<tr>
<td>30, 31</td>
<td>47.1</td>
<td>0.499</td>
<td>0.199</td>
</tr>
</tbody>
</table>
In this example, the log likelihood value could be changed by varying the dependence structures. We did not do this analysis because there are many choices of dependence structures and there is no guarantee that each choice of dependent structure with limited MC sample size gives a nice convergence. Simultaneous confidence region or individual confidence intervals of a final estimation of \( \tau \) will have a large area or width, so we need some methods to get a more precise result.

5.7 Discussion

We developed a new and efficient way to get full maximum likelihood estimates when the likelihood is in the form of a ratio of two integrations which is impossible to evaluate. The feature for our method is to use two importance densities, each of which is an independent approximation of the integrand in terms of matching the first two moments and the second moments. MC samples are used, but not MCMC samples. Independent approximation at one point could be bad for the other points. This makes it difficult to do a maximization process like N-R algorithm with fixed importance densities. We solved this inefficiency by choosing a starting value as close to true MMLE as possible by maximizing the Laplace approximation of log likelihood. The convergence of MC log likelihood to true log likelihood and the asymptotic results are very easy because our samples are i.i.d. What is better, we can easily assess the accuracy of the MC log likelihood, which is possible because the asymptotic variance-covariance matrix of the MC log likelihood can be estimated with ease. We did not drop the first term in (5.36), the ratio of normalizing constants of importance densities, because we need the full information in \( L_M(\tau) \) for \( L_M^{(q)}(\tau^{(q)}) \) and \( L_M^{(q+1)}(\tau^{(q)}) \) to be comparable. This was impossible in Gelfand and Carlin (1993).

This method can be applied to other classes of models, not conjugate forms, with more effort to make two different codes for sampling. This method can be also applied to more complex models like three or more stage hierarchical models.

In our example, we have only three hyper parameters, i.e., \( \alpha_{1j} = \alpha_1, \alpha_{2j} = \alpha_2, \) and \( \eta_j = \eta \). This parametrization is the simplest. The more complicated parametrizations, which include parametrization considering the effect of distance on the dependence and parametrization considering some covariates in Generalized Linear Model structure are possible.

As pointed out in Section 5.6.5.3, unlike the non-mixture models, mixture models need a huge AC sample size to produce accurate results. In practice, increasing the MC sample may not be a good solution to get a precise result. Instead, we have to seek other ways to reduce the MC errors in estimating the integrals. Basic Monte Carlo variance reduction techniques might be easily applied to improve performance of the estimation procedure further, or ease the burden of required Monte Carlo
sample size. For example, lack of dependence among sampled values should allow antithetic sampling to be easily incorporated into the sampling procedure.
CHAPTER 6 STRONG CONSISTENCY AND ASYMPTOTIC
NORMALITY OF MONTE CARLO MAXIMUM LIKELIHOOD
ESTIMATE

6.1 Introduction

When the log likelihood \( L \) contains unknown normalizing constant(s), the true Maximum Likelihood Estimates (MLEs) cannot be obtained. Instead, we can approximate \( L \) by Monte Carlo (MC) log likelihood, \( L_M \), where \( M \) is the sample size, and maximize \( L_M \) resulting in MC MLE. In this Chapter, we will deal with the strong consistency and asymptotic normality of MC MLE. Every MC MLE cannot be proved to have nice large sample properties. Those nice properties can be achieved depending on the maximization process to get MC MLE. One of the maximization methods to ensure the nice large sample properties is to use an adequately selected importance density so that the sample from this density can be used unchanged for evaluating \( L_M \) and possibly the first and second derivatives of \( L_M \) throughout the maximization process. Strong consistency and asymptotic normality of MC MLE from this maximization method will be shown for useful classes of models.

There are several reasons for the existence of unknown normalizing constants in \( L \). One of them is due to the dependence structure. In the spatial lattice model, the dependence structure can be well-established by a set of conditional distributions. In this chapter, we will consider the two classes of models; one for conditionally specified statistical models in which \( L \) contains only one unknown normalizing constant, and the other for conditionally specified statistical mixture models in which \( L \) contains two unknown normalizing constants. It may be easily seen that conditional specification is not used crucially in the proofs of this chapter, which means that all the theories in this Chapter can be applied to the models with unknown normalizing constants for whatever reasons. The important factor in proofs is the independence sample used in \( L_M \). We want to keep those two classes in attention in order to show the nice large sample properties for the MC MLE obtained from the models and the methods we developed in Chapter 4 and Chapter 5. From these properties, theoretical backgrounds for
assessing error in MC MLE and the accuracy in MC log likelihood can be provided.

Geyer (1994) showed more general results than ours in strong consistency, but his proof is more complex. We will simplify the proof, taking advantage of the use of independence samples both in strong consistency and asymptotic behaviors. Because we have the sum of the independent samples in $L_M$, we can use any large sample theory already established on independent data with slight modification.

In Section 6.2, we will briefly introduce the conditionally specified statistical models, show the desired properties, and provide some application in a practical situation. In Section 6.3, the results from Section 6.2 will be extended to the mixture models.

### 6.2 On MC MLE in Conditionally Specified Statistical Models

We will briefly describe conditionally specified models. Let $Y = (Y(s_1), \ldots, Y(s_n))$. With $f(\cdot \mid \cdot)$ denoting a generic conditional probability density or mass function, a conditionally specified model consists of the $n$ functions

$$f(y(s_i) \mid \{y(s_j) : j \neq i\}; i = 1, \ldots, n.$$ 

Dependence is modeled by defining, for each $y(s_i)$, a dependence set or neighborhood,

$$N_i \equiv \{s_h : g(y(s_i) \mid \{y(s_j) : j \neq i\}) \text{ depends on } y(s_h); h \neq i\},$$

and corresponding dependence index set $D_i \equiv \{h : s_h \in N_i\}$. Then, assuming the joint distribution exists, we can identify the joint distribution, corresponding to the set of specified conditionals, up to a normalizing constant as,

$$f(y) = \frac{\exp\{Q(y)\}}{\int \exp\{Q(t)\} d\mu(t)}; y \in \Omega,$$

where $\Omega$ is the support of $f(y)$ and $Q(\cdot)$, which is called Negpotential function, is a linear combination of functions constructed from the conditionals.

#### 6.2.1 Strong Consistency

Let $L(\theta)$ be the log likelihood of $f(y)$, then

$$L(\theta) = Q(y \mid \theta) - \log \left(\int \exp[Q(x \mid \theta)] dx\right).$$

Let $L_M(\theta)$ be the Monte Carlo estimate of $L(\theta)$, i.e.,

$$L_M(\theta) = Q(y \mid \theta) - \log \left(\frac{1}{M} \sum_{i=1}^{M} \frac{\exp[Q(x_i \mid \theta)]}{m(x_i \mid \lambda)}\right), \quad (6.1)$$
where \( \{x_i\}_{i=1}^M \) is the sample from a density \( m(x|\lambda) \) which satisfies \( m(x|\lambda) = 0 \) implies \( \exp[Q(x|\theta)]/K(\theta) = 0 \), for all \( \theta \in \Theta \) and for almost every \( x \). Then for any fixed \( \theta \), considering \( y \) fixed,

\[
L_M(\theta) \to L(\theta), \quad \text{almost surely } [P_\lambda]
\]

as \( M \to \infty \), by the strong law of large number if \( E_\lambda(\exp[Q(x|\theta)]/m(x|\lambda)) \) is finite for all \( \theta \in \Theta \). From this fact, we want to see

\[
\hat{\theta}_M \to \hat{\theta}, \quad \text{almost surely,}
\]

where \( \hat{\theta}_M \) is the maximizer of \( L_M(\theta) \), and \( \hat{\theta} \) is the maximizer of \( L(\theta) \). This is not always true. Geyer and Thompson (1992) showed that (6.3) holds for exponential family because of concavity of \( L_M \), and \( L \). And Geyer (1994) extended this result to non-exponential family. In his extension, he used the idea of hypoconvergence, and its properties. In this section, we propose sufficient conditions, stronger than Geyer's but weaker than Geyer and Thompson's, for (6.3) to hold in a different way. Our development will be based on the following Lemma in Wu (1981).

**Lemma 6.2.1** Assume that, for any \( \eta > 0 \),

\[
\liminf_{M \to \infty} \inf_{|\hat{\theta} - \theta| \geq \eta} [L_M(\hat{\theta}) - L_M(\theta)] > 0, \quad \text{a.s.}
\]

Then,

\[
\hat{\theta}_M \to \hat{\theta}, \quad \text{a.s. as } M \to \infty.
\]

**proof** See Wu (1981)./\

**Lemma 6.2.2** Suppose that almost sure convergence in (6.3) is uniform on \( \Theta \). And suppose that \( \hat{\theta} \) is the unique maximizer of \( L \). Then,

\[
\hat{\theta}_M \to \hat{\theta}, \quad \text{a.s. as } M \to \infty.
\]

**proof** See Gallant and White (1988,pp.18-19,36-37) for original proof, and see Fuller (1996,pp.252-253) for modified proof./\

In Gallant and White's proof, compactness of \( \Theta \), and almost sure continuity of \( L_M \)'s are needed because these assumptions guarantee the existence and measurability of \( \hat{\theta}_M \). Sometimes, without compactness or continuity, \( \hat{\theta}_M \) can exist, so we need certain criteria for checking out the almost sure convergence of \( \hat{\theta}_M \) to \( \hat{\theta} \). Therefore, Lemma 6.2.1 or Lemma 6.2.2 can be used for this purpose. For
example, if we consider exponential family case, the uniform convergence condition in Lemma 6.2.2 will be satisfied from the fact that the first derivatives are bounded uniformly, so MC MLE converge almost surely to true MLE. If we can't say anything about differentiability, we need the compactness and equicontinuity condition.

**Definition 6.2.3** A family $\mathcal{L}$ of functions is equicontinuous at the point $\theta_0 \in \Theta$ if for every $\epsilon > 0$, there is a $\delta > 0$ such that $|L(\theta) - L(\theta_0)| < \epsilon$ whenever $||\theta - \theta_0|| < \delta$, for all $L \in \mathcal{L}$. $\mathcal{L}$ is uniformly equicontinuous on $\Theta$ if for every $\epsilon > 0$, there is a $\delta > 0$ such that $|L(\theta) - L(\theta)| < \epsilon$ whenever $||\theta - \theta_0|| < \delta$, for all $\theta_0 \in \Theta$ and for all $L \in \mathcal{L}$.

Almost sure version of equicontinuity is straightforward, so we are ready for the following Lemma.

**Lemma 6.2.4** Given $(\Omega, \mathcal{F}, P)$ and a compact set $\Theta$ which is a subset of $k$-dimensional Euclidean space, $\mathbb{R}^k$, let $L_M : \Omega \times \Theta \to \mathbb{R}$ be a random function such that $\{L_M(\theta), M = 1, \cdots\}$ be a almost surely equicontinuous family of random functions. And suppose that $\hat{\theta}$ is the unique maximizer of $L$. Then,

$$\hat{\theta}_M \to \hat{\theta}, \text{ a.s. as } M \to \infty.$$  

**proof** Since equicontinuity on compact set implies uniformly equicontinuity on that set, the convergence of (6.3) is uniform on $\Theta$ (Kripke, 1968, pp.163-164). From Lemma 6.2.2, the result is followed.//

**Lemma 6.2.5** Assume that $g(x)$ is continuous function on a subset of $\mathbb{R}$. And assume that $\{L_M(\theta), M = 1, \cdots\}$ defined in Lemma 6.2.4 is almost surely equicontinuous family. Then $\{g(L_M(\theta)), M = 1, \cdots\}$ is also almost surely equicontinuous family.

**proof** From the equicontinuity of $\{L_M(\theta), M = 1, \cdots\}$, for each $\theta_0$, there exist a $F(\theta_0)$ in $\mathcal{F}$ with $P_\lambda(F) = 1$ such that, given any $\delta > 0$, for each $\omega$ in $F(\theta_0)$ there exists a $\gamma(\omega, \delta, \theta_0) > 0$ such that

$$|L_M(\omega, \theta_0) - L_M(\omega, \theta)| < \delta \text{ whenever } ||\theta - \theta_0|| < \gamma(\omega, \delta, \theta_0), \text{ for all } M. \quad (6.4)$$

For any $\omega \in F$, the image of $L_M(\omega, \cdot)$ on $\Theta$ is compact because $\Theta$ is compact and $L_M(\omega, \cdot)$ is continuous on $\Theta$. Therefore, letting $L_M(\omega, \cdot)$ be the image of $L_M(\omega, \cdot)$ on $\Theta$, $g(x)$ is uniformly continuous on compact set $L_M(\omega, \Theta)$, i.e., given $\epsilon > 0$, there exists a $\delta(\omega, \epsilon) > 0$ such that

$$|g(L_M(\omega, \theta_0)) - g(L_M(\omega, \theta))| < \epsilon \text{ whenever } |L_M(\theta) - L_M(\theta_0)| < \delta(\omega, \epsilon) \quad (6.5)$$
and for this $\delta(\omega, \epsilon)$, there exists a $\gamma(\delta, \theta_0) = \gamma(\omega, \epsilon, \theta_0) > 0$ satisfying (6.4). Now, for any $K \geq 1$,

$$|g(L_{M+K}(\omega, \theta_0)) - g(L_{M+K}(\omega, \theta))| < \epsilon$$

because $|L_{M+K}(\theta) - L_{M+K}(\theta_0)| < \delta(\omega, \epsilon)$ as long as $||\theta - \theta_0|| < \gamma(\delta, \theta_0)$. Therefore, we prove that, for any $\omega \in F$ with $P(F) = 1$, given any $\epsilon > 0$, there exists a $\gamma(\omega, \epsilon, \theta_0) > 0$ satisfying (6.5) for all $M$, which is the almost sure equicontinuity of $\{g(L_M(\theta)), M = 1, \cdots, \}$.

The continuity (or equicontinuity) of $L_M$ depends on the continuity of $Q(x|\theta)$, and the condition in terms of $Q(x|\theta)$ is much easier to check.

**Theorem 6.2.6** Assume that every condition is same under the Lemma 6.2.4 except equicontinuity and assume that $Q(x|\theta)$ is continuous at each $\theta$ almost surely, and $Q(y|\theta)$ is continuous for observed $y$. Then,

$$\hat{\theta}_M \to \hat{\theta}, \text{ a.s. as } M \to \infty.$$ 

**Proof** Let $h(x|\theta) = \frac{\exp(Q(x|\theta))}{m(x|\lambda)}$. Then $h(x|\theta)$ is almost sure continuous at each $\theta$ from the continuity of $Q(x|\theta)$, i.e., for each $\theta_0$, there exists a $F(\theta_0)$ in $A$ with $P_x(\Omega) = 1$ such that, given any $\epsilon > 0$, for each $x$ in $F(\theta_0)$ there exists a $\delta(\omega, \epsilon, \theta_0) > 0$ such that

$$|h(x(\omega), \theta_0) - h(x(\omega), \theta)| < \epsilon \text{ whenever } ||\theta - \theta_0|| < \delta(\omega, \epsilon, \theta_0).$$ \hspace{1cm} (6.6)

Now we want to show that the family, $\{\frac{\sum_{i=1}^{M} h(x_i|\theta)}{M}, M = 1, 2, \cdots \}$ is almost surely equicontinuous for each $\theta$. For any $\omega \in F$ and $\theta_0$, given $\epsilon$, there exists $\delta(\omega, \epsilon, \theta_0) > 0$ such that

$$\left|\frac{\sum_{i=1}^{K} h(x_i(\omega)|\theta_0)}{K} - \frac{\sum_{i=1}^{K} h(x_i(\omega)|\theta)}{K}\right| < \frac{\sum_{i=1}^{K} |h(x_i(\omega)|\theta_0) - h(x_i(\omega)|\theta)|}{K} < K \epsilon/K = \epsilon, \hspace{1cm} (6.7)$$

whenever $||\theta - \theta_0|| < \delta$, since $x_i$'s are identically distributed, and from (6.6). This is true for any $K$. Therefore, $\{\frac{\sum_{i=1}^{M} h(x_i|\theta)}{M}, M = 1, 2, \cdots \}$ is almost surely equicontinuous. Since $L_M = Q(y|\theta) - \log(\frac{\sum_{i=1}^{M} h(x_i|\theta)}{M})$, the family $\{L_M(\theta), M = 1, \cdots, \}$ is almost surely equicontinuous at each $\theta$ from the continuity assumption of $Q(y|\theta)$ and Lemma 6.2.5. Therefore, the result is followed from Lemma 6.2.4.//

If $\Theta$ is a finite set, it is natural that the convergence in (6.3) is uniform, so, by Lemma 6.2.2, we can get the result we want. As Wu (1981) pointed out in Remark 1, the finiteness assumption on $\Theta$ is not quite a restriction from the practical viewpoint because, in actual computation, we can only search
the maximum over the finite set, say, to the eighth decimal place. If Θ is not finite but compact, the only condition for the almost sure convergence of MC MLE to true MLE is the continuity of Q-function according to Theorem 6.2.6. And the compactness condition on Θ is feasible because we want to maximize \( L_M \) within a small neighborhood of some \( \theta \in \Omega \) close to true MLE. As we mentioned before, the compactness of \( \Theta \) and almost sure continuity of \( L_M \) ensures the existence and measurability of maximizer of \( L_M \), for all \( M \).

6.2.2 Asymptotic Normality

The asymptotic normality of \( \sqrt{M} (\hat{\theta}_M - \hat{\theta}) \) could be easily established under the usual regularity conditions because \( L_M(\theta) \) contains the term of sum of independent samples. We need 2 Lemmas for main Theorem.

**Lemma 6.2.7** \( \nabla^2 L_M(\theta) \) converges to \( \nabla^2 L(\theta) \) almost surely for each \( \theta \) in \( \Theta \).

**Proof** Let \( h_i(\theta) \) be \( \frac{\exp(Q[x_i|\theta])}{m(x_i|\lambda)} \) as in (6.1). Then, element of \( j^{th} \) row and \( k^{th} \) column of \( \nabla^2 L_M(\theta) \) is

\[
Q^*_{j,k}(\theta) = \frac{\sum_{i=1}^{M} (Q_{j,i}^x(\theta) + Q_{j,i}^e(\theta)Q_{k,i}^x(\theta))h_i(\theta)}{\sum_{i=1}^{M} h_i(\theta)} + \left( \frac{\sum_{i=1}^{M} Q_{j,i}^x(\theta)h_i(\theta)}{\sum_{i=1}^{M} h_i(\theta)} \right) \left( \frac{\sum_{i=1}^{M} Q_{k,i}^e(\theta)h_i(\theta)}{\sum_{i=1}^{M} h_i(\theta)} \right)
\]

(6.8)

where \( Q^*_{i_1,\cdots,i_n}(\theta) \) is the partial derivatives of \( Q(y|\theta) \) with respect to \( \theta_{i_1}, \cdots, \theta_{i_n} \).

From the fact that, for each \( \theta \),

\[
\sum_{i=1}^{M} h_i(\theta)/M \to \int k(\theta) \equiv \exp(Q[x|\theta])dx \quad \text{a.s.,}
\]

and

\[
\sum_{i=1}^{M} Q_{i_1,\cdots,i_n}^x(\theta)h_i(\theta)/M \to \int Q_{i_1,\cdots,i_n}^x(\theta) \exp(Q[x|\theta])dx \quad \text{a.s.,}
\]

it is straightforward to see that the element of \( j^{th} \) row and \( k^{th} \) column of \( \nabla^2 L_M(\theta) \) converge to that of \( \nabla^2 L(\theta) \) almost surely, for all \( i, j \). //
Lemma 6.2.8  \( \sqrt{M} \nabla L_M(\hat{\theta}) \xrightarrow{L} N(0, \Sigma) \), for some covariance matrix \( \Sigma \).

Proof  \( i^{th} \) element of \( \nabla L_M(\hat{\theta}) \) is

\[
\sum_{i=1}^{M} \frac{[Q^T_i(\hat{\theta}) - Q^T_i(\hat{\theta})]h_i(\hat{\theta})}{\sum_{i=1}^{M} h_i(\hat{\theta})}
\]

(6.9)

where \( Q^T_i(\hat{\theta}) \) is \( Q^T_i(\hat{\theta}) \) evaluated at \( \hat{\theta} \).

Let \( v_{ik} \equiv [Q^T_i(\hat{\theta}) - Q^T_i(\hat{\theta})]h_i(\hat{\theta}) \), and \( v_i \equiv (v_{i1}, \cdots, v_{iK})' \). Then \( \{v_i\}_{i=1}^{M} \) are i.i.d. with some common mean \( \mu \), and variance-covariance \( A \) because \( \{x_i\}_{i=1}^{M} \) are i.i.d. from \( m(x|\lambda) \). Sample mean of \( v_{ik} \)'s is the numerator of \( i^{th} \) element of \( \nabla L_M(\hat{\theta}) \) divided by \( M \). Therefore, by Central Limit Theorem and the fact that \( \frac{1}{M} \sum_{i=1}^{M} h_i(\hat{\theta}) \) converge to \( k(\hat{\theta}) \) a.s.,

\[
\sqrt{M} \nabla L_M(\hat{\theta}) \xrightarrow{L} N(\mu, A/k(\hat{\theta})^2).
\]

Now, \( \mu = E_{\lambda}(v_1) = 0 \) because \( \hat{\theta} \) is the maximizer of \( L(\theta) \), so \( \partial L(\hat{\theta})/\partial \theta = 0 \), i.e.,

\[
Q^T_\theta(\hat{\theta}) - \frac{\int Q^T_\theta(\hat{\theta}) \exp(Q[x|\theta])}{k(\hat{\theta})} = 0.
\]

The proof will be finished by letting \( \Sigma = A/k(\hat{\theta})^2 \) which is valid for variance-covariance matrix. //

Theorem 6.2.9 Assume the following conditions.

(a) \( \hat{\theta}_M \) converge to \( \hat{\theta} \) in probability.

(b) \( \hat{\theta} \) is unique.

(c) There exists a convex, compact neighborhood \( C \) of \( \hat{\theta} \) such that \( L_M(\cdot, \cdot) \) is measurable for each \( \theta \) in \( C \), \( L_M(\omega, \cdot) \) is almost surely twice continuous differentiable with respect to \( \theta \) on \( C \), \( C \) is in \( \Theta \), and \( \hat{\theta} \) is an interior point of \( C \).

(d) \( \int \exp[Q(y; \theta)]dy \) is twice differentiable under the integral sign.

(e) \( B = -\nabla^2 L(\hat{\theta}) \) is positive definite, and continuous at \( \hat{\theta} \).

(f) The convergence in Lemma 6.2.7 is uniform on \( C \).

Then

\[
\sqrt{M}(\hat{\theta}_M - \hat{\theta}) \xrightarrow{L} N(0, B^{-1} \Sigma B^{-1}).
\]

(6.10)

Proof  Since \( \hat{\theta}_M \xrightarrow{P} \hat{\theta}, \hat{\theta}_M \) eventually takes its values in \( C \). It is thus legitimate to expand \( \nabla L_M(\theta) \) in the neighborhood of \( \hat{\theta}_M \). From the mean value theorem, there exists a \( \lambda_M \in [0,1] \) such that

\[
\nabla L_M(\hat{\theta}) = \nabla L_M(\hat{\theta}_M) + \nabla^2 L_M(\theta^*_M)(\hat{\theta} - \hat{\theta}_M),
\]

(6.11)
where \( \theta^*_M = (1 - \lambda M ) \hat{\theta} + \lambda M \hat{\theta} \) is measurable. Since \( \hat{\theta}_M \) is in the interior of \( \Theta \) eventually, \( \nabla L_M(\hat{\theta}_M) = 0 \).

Now (6.11) can be rewritten as

\[
0 = \nabla L_M(\hat{\theta}) + \nabla^2 L_M(\theta^*_M)(\hat{\theta}_M - \hat{\theta}).
\]

By assumption (a), (e), and (f),

\[
\nabla^2 L_M(\theta^*_M) \overset{P}{\to} \nabla^2 L_M(\hat{\theta}) = -B. \tag{6.12}
\]

Therefore,

\[
\hat{\theta}_M - \hat{\theta} = B^{-1} \nabla L_M(\hat{\theta}) + r_M,
\]

where \( r_M \) is of smaller order in probability than \( \hat{\theta}_M - \hat{\theta} \). By Lemma 6.2.8, we can have

\[
\sqrt{M}(\hat{\theta}_M - \hat{\theta}) \overset{L}{\to} N(0, B^{-1} \Sigma B^{-1}).
\]

**Remark 1** Unfortunately, \( \Sigma \) is different from \( B \). Assuming \( Q_{j,k}^{x}(\theta) = 0 \), for any \( j \) and \( k \),

\[
B = \left( \int [Q_j^{x}(\hat{\theta}) - Q_j^{x}(\hat{\theta})][Q_k^{x}(\hat{\theta}) - Q_k^{x}(\hat{\theta})] \frac{\exp(Q[x|\hat{\theta}])}{k(\hat{\theta})} dx \right)_{j,k}, \tag{6.13}
\]

and

\[
\Sigma = \left( \int [Q_j^{x}(\hat{\theta}) - Q_j^{x}(\hat{\theta})][Q_k^{x}(\hat{\theta}) - Q_k^{x}(\hat{\theta})] \frac{h(x|\hat{\theta}) \exp(Q[x|\hat{\theta}])}{k(\hat{\theta})} dx \right)_{j,k}. \tag{6.14}
\]

**Remark 2** \( -\nabla^2 L_M(\hat{\theta}_M) \) is the consistent estimator of \( B \) by Lemma 6.2.7 and conditions (e) and (f) in Theorem 6.2.9. For the estimation of \( \Sigma \), the following quantity is reasonable from the proof of Lemma 6.2.8;

\[
\left( \sum_{i=1}^{M} [Q_i^{x}(\hat{\theta}_M) - Q_i^{x}(\hat{\theta}_M)][Q_i^{x}(\hat{\theta}_M) - Q_i^{x}(\hat{\theta}_M)]h_i(\hat{\theta}_M)^2 \right)_{j,k}.
\]

**Remark 3** The condition (f) can be checked (or replaced) by the uniform boundness of \( \nabla^3 L_M(\theta) \) on \( C \).

### 6.2.3 Application

When we use \( L_M(\theta) \) as a substitute of \( L(\theta) \) in maximization process, it's necessary to check 1) how accurate the MC MLE, \( \hat{\theta}_M \) is when it is compared with true MLE, \( \hat{\theta} \), and 2) the accuracies of
to make sure that the maximization process goes in a right way. The first check-up can be done using Theorem 6.2.9. with estimated $\Sigma$ and $B$ as in Remark 2. The second one can be done using asymptotic normality of $L_M(\theta)$. For example, the accuracy of $L_M(\theta)$ depends on

$$\log \left( \frac{1}{M} \sum_{i=1}^{M} \frac{\exp(x_i|\theta)}{m(x_i|\lambda)} \right)$$

which is the approximation of $\log(k(\theta))$. Therefore, we would like to tolerate the discrepancy between

$$\log \left( \frac{1}{M} \sum_{i=1}^{M} \frac{\exp(x_i|\theta)}{m(x_i|\lambda)} \right)$$

and $\log(k(\theta))$ if it's small enough. It is easily verified from CLT and delta method that

$$\sqrt{M} \left( \log \left[ \frac{1}{M} \sum_{i=1}^{M} \frac{\exp(x_i|\theta)}{m(x_i|\lambda)} \right] - \log[k(\theta)] \right) \overset{L}{\to} N(0, \sigma^2),$$

where

$$\sigma^2 = \text{Var}_\lambda \left( \frac{\exp(x|\theta)}{m(x|\lambda)} \right) / E_\lambda \left( \frac{\exp(x|\theta)}{m(x|\lambda)} \right)^2.$$

(6.16)

For given $M$, and $\delta$, we want to have

$$Pr \left( \left| \log \left[ \frac{1}{M} \sum_{i=1}^{M} \frac{\exp(x_i|\theta)}{m(x_i|\lambda)} \right] - \log[k(\theta)] \right| > \epsilon \right) = Pr(|L_M(\theta) - L(\theta)| > \epsilon < \delta),$$

(6.17)

for small $\epsilon$.

Now, (6.17) can be written as

$$Pr(|Z| > \sqrt{M}\epsilon/\sigma) < \delta$$

where $Z$ is the random variable whose distribution is standard Normal. Therefore, it is enough that

$$\sqrt{M}\epsilon/\sigma = z_{\delta/2},$$

i.e.,

$$\epsilon = \sigma z_{\delta/2}/\sqrt{M}.$$

When $\delta = 0.05$, $\epsilon$ is $1.96\sigma/\sqrt{M}$. And $\sigma$ can be estimated from the samples, i.e., just plug sample variance and sample mean into numerator and denominator of right-side quantity in (6.16) respectively for the estimate of $\sigma^2$.

### 6.3 On MC MLE in Conditionally Specified Statistical Mixture Models

In this section, we want to extend the results in section 6.2 to the mixture models. Suppose $Y \equiv (Y(s_1), \cdots, Y(s_n))$ has the following density;

$$f(y|\theta) = \exp(Q_1(y|\theta))/K_1(\theta),$$

(6.18)
where \( \theta \equiv (\theta(s_1), \ldots, \theta(s_n)) \), and the normalizing constant \( K_1(\theta) \) is known. And assume that

\[
g(\theta|\tau) = \exp(Q_0(\theta|\tau))/K_0(\tau),
\]

(6.19)

where the joint distribution \( g(\theta|\tau) \) is characterized by a set of conditional distributions,

\[
g(\theta(s_i)|\{\theta(s_j) : j \neq i\}; i = 1, \ldots, n).
\]

Then, the normalizing constant \( K_0(\tau) = \int \exp(Q_0(\theta|\tau)) d\theta \) is unknown. We want to estimate \( \tau \) which maximizes log likelihood \( L(\tau) \) with respect to \( \tau \), where

\[
L(\tau) \equiv \log \left( \int f(y|\theta) g(\theta|\tau) d\theta \right)
\]

\[
= \log \left( \int \exp(Q_1(y|\theta) - \log(K_1(\theta)) + Q_0(\theta|\tau)) d\theta \right) - \log \left( \int \exp(Q_0(\theta|\tau)) d\theta \right).
\]

6.3.1 Strong Consistency

Let \( L_M(\tau) \) be the Monte Carlo estimate of \( L(\tau) \), i.e.,

\[
L_M(\tau) = \log \left( \frac{1}{M} \sum_{i=1}^M \exp[Q(\theta_i^1|\tau, y)] \right) - \log \left( \frac{1}{M} \sum_{i=1}^M \exp[Q(\theta_i^0|\tau)] \right)
\]

(6.20)

where \( Q(\theta|\tau, y) = Q_1(y|\theta) - \log(K_1(\theta)) + Q_0(\theta|\tau) \), \( \{\theta_i^1\}_{i=1}^M \) is a sample from \( m_1(\theta|\lambda_1, y) \) which dominates normalized density of \( Q(\theta|\tau, y) \), and \( \{\theta_i^0\}_{i=1}^M \) is a sample from \( m_0(\theta|\lambda_0) \) which also dominates normalized density of \( Q_0(\theta|\tau) \). Let \( P_\lambda \) be the product measure of \( P_{\lambda_1} \) and \( P_{\lambda_0} \), then for any fixed \( \tau \), considering \( y \) fixed,

\[
L_M(\tau) \rightarrow L(\tau), \text{ almost surely } [P_\lambda]
\]

(6.21)
as \( M \rightarrow \infty \), by the strong law of large number if \( E_{\lambda_1}(\exp[Q(\theta|\tau, y)]/m_1(\theta|\lambda_1, y)) \) and \( E_{\lambda_0}(\exp[Q_0(\theta|\tau)]/m_0(\theta|\lambda_0)) \) are finite for all \( \tau \in \mathcal{T} \).

Every Lemma in Section 6.2 can be applied to check out if

\[
\hat{\tau}_M \rightarrow \hat{\tau}, \text{ almost surely } [P_\lambda],
\]

(6.22)

where \( \hat{\tau}_M \) is the maximizer of \( L_M, \ M = 1, \ldots, \), and \( \hat{\tau} \) is the maximizer of \( L \). Therefore, only the main Theorem will be stated with brief proof.

Theorem 6.3.1 Given \( (\Omega, A, P_\lambda) \) which is the product measure space of \( (\Omega_0, A_0, P_{\lambda_0}) \) and \( (\Omega_1, A_1, P_{\lambda_1}) \), and a compact set \( \mathcal{T} \) which is a subset of \( k \)-dimensional Euclidean space, \( \mathbb{R}^k \), let \( \{L_M(\tau), M = 1, \ldots, \} \) be a family of random functions from \( \Omega \times \mathcal{T} \) to \( \mathbb{R} \). And assume that \( Q_0(\theta|\tau) \) is
almost sure continuous at each \( r \) with respect to both \( P_{\lambda_0} \) and \( P_{\lambda_1} \). If \( \hat{\tau} \) is the unique maximizer of \( L \). Then,

\[ \hat{\tau}_M \to \hat{\tau}, \text{ a.s. } [P_\lambda] \text{ as } M \to \infty. \]

**Proof** Let \( h_1(r) = \exp[Q(\theta|r,y)]/m_1(\theta|\lambda_1,y) \), and \( h_0(r) = \exp[Q_0(\theta|r)]/m_0(\theta|\lambda_0) \). Then \( h_1(r) \) is almost sure continuous at each \( r \) with respect to \( P_{\lambda_1} \) because of almost sure continuity of \( Q_0(\theta|r) \). For the same reason, \( h_0(r) \) is almost sure continuous at each \( r \) with respect to \( P_{\lambda_0} \). Therefore, on compact set \( T \), \( \{\log(\sum_{i=1}^M h_1(\tau_i^M)/M)\} \) \( \{\log(\sum_{i=1}^M h_0(\tau_i^0)/M)\} \) is the almost sure equicontinuous family with respect to \( P_{\lambda_1} \) \( (P_{\lambda_0}) \) because of Lemma 6.2.5. It is easily seen that the family consisting of \( L_M(r) \)'s defined in (6.20) is almost sure equicontinuous with respect to product measure \( P_\lambda \). Then we can get the result by Lemma 6.2.4. //

**6.3.2 Asymptotic Normality**

The asymptotic normality of \( \sqrt{M}(\hat{\tau}_M - \hat{\tau}) \) can be established in this section from which we can assess Monte Carlo error in using \( \hat{\tau}_M \) as a MLE instead of true one, \( \hat{\tau} \). The similar Lemmas in Section 6.2.2 can be obtained obviously, but the proof will not be omitted for the hint of estimates of \( \Sigma \) and \( B \) in (6.33). We need some notations. Let

\[ h_0(\theta^0|r) = \frac{\exp[Q_0(\theta^0|r)]}{m_0(\theta^0|\lambda_0)}, \]  

(6.23)

where \( \theta^0 = \theta^0(\omega_0) \) is the random vector defined on \( (\Omega_0,A_0,P_{\lambda_0}) \),

\[ h_1(\theta^1|r) = \frac{\exp[Q(\theta^1|r)]}{m_1(\theta^1|\lambda_1)}, \]  

(6.24)

where \( \theta^1 = \theta^1(\omega_1) \) is the random vector defined on \( (\Omega_1,A_1,P_{\lambda_1}) \),

\[ K_M^k(\tau) = \sum_{j=1}^M h_k(\theta^k_j|r), \ k = 0,1, \]  

(6.25)

and

\[ Q_{0,i_1,\ldots,i_n}(\tau) = \frac{\partial^n Q_0(\theta^k|r)}{\partial \tau_{i_1} \cdots \partial \tau_{i_n}}, \ k = 0,1. \]  

(6.26)

**Lemma 6.3.2** \( \nabla^2 L_M(\tau) \) converges to \( \nabla^2 L(\tau) \) almost surely \([P_\lambda]\) for each \( \tau \) in \( T \).

**Proof** The element of \( j^{th} \) row and \( k^{th} \) column of \( \nabla^2 L_M(\tau) \) is

\[ \frac{\sum_{i=1}^M Q_{0,i,k}^j(\tau) h_1(\theta^1_i|r)}{K_M^k(\tau)} + \frac{\sum_{i=1}^M Q_{0,i,k}^j(\tau) Q_{0,i,k}^{j*}(\tau) h_1(\theta^1_i|r)}{K_M^k(\tau)} \]
From the fact that, for each $r$,  
\[ K_M^0 (r) / M \to \int \exp(Q_0[\theta^0 | r]) d\theta^0 \text{ a.s. } [P_{\lambda_0}], \]
\[ K_M^1 (r) / M \to \int \exp(Q[\theta^1 | r]) d\theta^1 \text{ a.s. } [P_{\lambda}], \]
\[ \sum_{i=1}^M Q_{0,i,...,i,n}^\delta (r) h_0(\theta_i^0 | r) / M \to \int Q_{0,i,...,i,n}^\delta (r) \exp(Q_0[\theta^0 | r]) d\theta^0 \text{ a.s. } [P_{\lambda_0}], \]
and
\[ \sum_{i=1}^M Q_{0,i,...,i,n}^\delta (r) h_1(\theta_i^1 | r) / M \to \int Q_{0,i,...,i,n}^\delta (r) \exp(Q[\theta^1 | r]) d\theta^1 \text{ a.s. } [P_{\lambda}]. \]

It is straightforward to see that the element of $j^{th}$ row and $k^{th}$ column of $\nabla^2 L_M (r)$ converge to that of $j^{th}$ row and $k^{th}$ column of $\nabla^2 L (r)$, a.s. $[P_{\lambda}]$. Therefore, for each $r$, $\nabla^2 L_M (r)$ converges to $\nabla^2 L (r)$ almost surely $[P_{\lambda}]. //$

**Lemma 6.3.3** $\sqrt{M} \nabla L_M (r) \xrightarrow{L} N(0, \Sigma)$, for some covariance matrix $\Sigma$.

**Proof** Let's think about the following i.i.d. random vectors;

\[ u_i (r) \equiv (u_{i1} (r) | u_{i2} (r))' \]
\[ \equiv (Q_{0,1}^\delta (r) h_1(\theta_i^1 | r), \cdots, Q_{0,k}^\delta (r) h_1(\theta_i^1 | r), h_1(\theta_i^1 | r), Q_{0,1}^\delta (r) h_0(\theta_i^0 | r), \cdots, Q_{0,k}^\delta (r) h_0(\theta_i^0 | r), h_0(\theta_i^0 | r))', \]

\[ i = 1, \cdots, M. \] It is easily seen that

\[ E(u_i (r)) \]
\[ = (\int Q_{0,1}^\delta (r) \exp(Q[\theta^1 | r]) d\theta^1, \cdots, \int Q_{0,k}^\delta (r) \exp(Q[\theta^1 | r]) d\theta^1, \int \exp(Q[\theta^1 | r]) d\theta^1). \]
\[
\int Q_{0,1}^{\theta_0}(\tau) \exp(Q_{0}[\theta_0 | \tau])d\theta_0, \ldots, \int Q_{0,k}^{\theta_0}(\tau) \exp(Q_{0}[\theta_0 | \tau])d\theta_0, \int \exp(Q_{0}[\theta_0 | \tau])d\theta_0 \right)^t (6.29)
\]

and

\[
\begin{align*}
\text{Var}(v_1(\tau)) &= E(v_1(\tau)V_1(\tau)) - E(v_1(\tau))E(V_1(\tau))^t \\
&= \begin{pmatrix} E(v_{11}(\tau)v_{11}(\tau)^t) - \mu_1(\tau)\mu_1(\tau)^t & 0 \\
0 & E(v_{12}(\tau)v_{12}(\tau)^t) - \mu_2(\tau)\mu_2(\tau)^t \end{pmatrix}.
\end{align*}
\]

Then, by CLT, letting
\[
\sqrt{M}(\bar{v}(\tau) - E(v_1(\tau))) \xrightarrow{L} N(0, \text{Var}(v_1(\tau))).
\]

Now, define a function, \(G(\cdot) : \mathbb{R}^{2k+2} \to \mathbb{R}^k\), such that
\[
G(x) = (x_1x_{k+1}^{-1} - x_{k+2}x_{2k+2}^{-1}, \ldots, x_kx_{k+1}^{-1} - x_{2k+1}x_{2k+2}^{-1})^t.
\]
then \(G\) has continuous first derivatives as long as \(x_{k+1}\) and \(x_{2k+2}\) are not zero. By delta method,
\[
\sqrt{M}(G(\bar{v}(\tau)) - G(E(v_1(\tau)))) \xrightarrow{L} N(0, [G(E(v_1(\tau))] \text{Var}(v_1(\tau))[G(E(v_1(\tau))]^t). (6.31)
\]

which is same as
\[
\sqrt{M}(\nabla L_M(\tau) - \nabla L(\tau)) \xrightarrow{L} N(0, [G(E(v_1(\tau))] \text{Var}(v_1(\tau))[G(E(v_1(\tau))]^t). (6.32)
\]

When \(\tau = \hat{\tau}\), \(\nabla L(\hat{\tau}) = 0\) and the desired matrix \(\Sigma = [G(E(v_1(\hat{\tau}))) \text{Var}(v_1(\hat{\tau}))[G(E(v_1(\hat{\tau})))^t]\). Therefore, the Lemma is proved. //

**Theorem 6.3.4** Assume the following conditions.

(a) \(\tau_M\) converge to \(\hat{\tau}\) in probability.

(b) \(\hat{\tau}\) is unique.

(c) There exists a convex, compact neighborhood \(C\) of \(\hat{\tau}\) such that \(L_M(\cdot, \tau)\) is measurable for each \(\tau\) in \(C\), \(L_M(\cdot, \cdot)\) is almost surely twice continuous differentiable with respect to \(\tau\) on \(C\), \(C\) is in \(T\), and \(\hat{\tau}\) is an interior point of \(C\).

(d) \(\int \exp(Q_1(\theta | \theta) - \log(K_1(\theta)) + Q_0(\theta | \tau))d\theta\) and \(\int \exp(Q_0(\theta | \tau))d\theta\) are twice differentiable under the integral sign.

(e) \(B = -\nabla^2 L(\hat{\tau})\) is positive definite, and continuous at \(\hat{\tau}\).

(f) The convergence in Lemma 6.3.2 is uniform on \(C\).
Then
\[ \sqrt{M}(\hat{T}_M - \tau) \overset{D}{\to} N(0, B^{-1}\Sigma B^{-1}). \] (6.33)

**Remark 1** When we analyze real data, we must know \( B^{-1}\Sigma B^{-1} \) in order to assess the Monte Carlo error in \( \hat{T}_M \) which is impossible. Therefore, we need to estimate \( B^{-1}\Sigma B^{-1} \). The estimate of \( B \) is obviously \( \nabla^2 L_M(\hat{T}_M) \) from Lemma 6.2.2, and conditions (e) and (f) in Theorem 6.2.4. For the estimate of \( \Sigma \), it is reasonable to use \([\nabla G(\bar{v}(\hat{T}_M))]S(\bar{v}(\hat{T}_M))[\nabla G(\bar{v}(\hat{T}_M))]'\) from the proof of Lemma 6.2.3, where

\[ S(\bar{v}(\hat{T}_M)) = \sum_{i=1}^{M}(\bar{v}_i(\hat{T}_M)\bar{v}_i(\hat{T}_M)' - \bar{v}(\hat{T}_M)\bar{v}(\hat{T}_M)')/(M - 1). \]

### 6.3.3 Application

As we did in Section 6.2.3, the asymptotic normality of \( L_M(\tau) \) can be used for the assessment of the accuracy of \( L_M(\tau) \) during the maximization process. It is easily seen that

\[ \sqrt{M}(L_M(\tau) - L(\tau)) \overset{D}{\to} N(0, \sigma^2), \]

where

\[ \sigma^2 = \text{Var}_\lambda\left(\frac{\exp[Q(\theta|\tau, y)]}{m_1(\theta|y, \lambda_1)}\right)/E_\lambda\left(\frac{\exp[Q(\theta|\tau, y)]}{m_1(\theta|y, \lambda_1)}\right)^2 + \text{Var}_\lambda_0\left(\frac{\exp[Q_0(\theta|\tau)]}{m_0(\theta|\lambda_0)}\right)/E_\lambda_0\left(\frac{\exp[Q_0(\theta|\tau)]}{m_0(\theta|\lambda_0)}\right)^2. \]

Therefore, given \( \delta \) and \( M \), an asymptotic \((1 - \delta)\%\) confidence interval of \( L_M(\tau) - L(\tau) \) is \( z_{\delta/2}\sigma/\sqrt{M} \). where \( \sigma \) may be estimated by the function of sample means and sample variances.
CHAPTER 7  GENERAL CONCLUSION

Conditionally specified statistical models are becoming increasingly important in the statistical analysis of multivariate and spatial problems. Chapter 3 deals with the modeling issue of conditionally specified statistical models when the conditional distributions belong to a one-parameter exponential family. Chapter 4 to Chapter 6 focus on the estimation issues for the conditionally specified statistical models with some brief review of modeling. The results given in Chapter 3 are extensions of those presented by Besag (1974) and show that the dependence present must be embodied in a finite set of dependence parameters, more general than those contained in the auto-models of Besag (1974). Thus, having all dependence parameters equal to zero implies an independence model that, in principle, could be tested for during model development. We show that pairwise-only dependence is a special case that, although more general than the independence model, is nested within a class of models exhibiting multiway dependence. With multi-parameter exponential auto-models, the results in Chapter 3 could provide a useful class of models which can be compatible with those of Arnold, Castillo, and Sarabia (1992). Inference for the parameters in the multiway dependence models is complicated by an intractable normalizing constant. These models are well suited to inference based on Markov Chain Monte Carlo methods.

In Chapter 4, we have presented a method for maximum likelihood estimation of parameters appearing in conditionally specified statistical models that incorporate complex dependence structures. In so doing, we have addressed a number of practical issues that arise in the application of Monte Carlo methods to repeated evaluations of a log likelihood. One issue is to select an importance sampling distribution which must play a good role in importance weights to provide stable estimation of intractable integrals throughout the maximization cycle. Our solution to this issue is based on the idea that the importance sampling distribution is constructed as the product of marginal densities, equivalent to forming 'dummy' independence models to parallel the behavior of the actual dependence model of interest. By matching the lower moments of marginal densities in the dummy model with estimates based on data sets simulated from the actual dependence model, we are able to produce sampling distributions with
which we can provide the solutions to other issues. Although we use the dummy model constructed for independent random variables, we can cover the high correlated data, which is thought to be impossible by some authors, because the dummy independent model will behave like the actual dependence model up to the effect of that dependence on the first two moments. The use of a dummy model constructed for independent random variables also allows application of theoretical results for independence case.

In Chapter 5, the basic idea developed in Chapter 4 is used to get the maximum likelihood estimates in conditionally specified statistical mixture models where the likelihood is the form of a ratio of two integrals resulting in being the two different conditionally specified statistical models in one log likelihood. The feature for our method is to use two importance densities, each of which is a dummy independence model performing like each actual dependence model in terms of the first two moments. Two MC samples can be used similarly in Chapter 4 for doing the estimation process. Chapter 5 focuses on the exponential family data models with the mixing distributions being of conjugate forms. But every method in Chapter 5 can be applied to other classes of models with more effort to make two different codes for sampling.

The convergence of MC estimation of log-likelihood to true log-likelihood and the asymptotic results are very easy because our samples are i.i.d. In Chapter 6, in addition to the large sample theory of MC estimation of log-likelihood, that of the MC MLE is tackled for both conditionally specified statistical models and mixture models. As a result, the sufficient condition for the strong consistency of MC MLE is the continuity of negpotential function, which is identified by a set of conditional distributions, over the compact set. Using the asymptotic results in Chapter 6, assessment of the accuracy in Monte Carlo log likelihood is possible which, in turn, gives solutions for the issues arising from the practical use of estimation process developed in Chapter 4 and 5.

Despite the efforts made in this thesis for the complete estimation process in conditionally specified statistical models that incorporate complex dependence structures, there are still many open problems in this field for further research. One such problem concerns selection of appropriate parametrizations. In the examples in Chapter 4 and 5, the simplest parametrization is used in conditionally specified statistical models, i.e., \( \alpha_J = \alpha \) and \( \eta_J = \eta \). The more complicated parametrizations are possible with the incorporation of the Generalized Linear Model structure. If we have some explanatory variables in each location, where should we put the term like \( g^{-1}(x' \beta) \) for parameterization? The second open question concerns inference for estimated parameters and the related issue of model selection. As shown in Chapter 6, the MC error in MLE due to use of estimated log likelihood instead of true log likelihood in maximization process can be assessed by looking at the asymptotic variance-covariance matrix of
MLE from the estimated log likelihood, say MC MLE, as the MC sample size increases. But it is still difficult to see the asymptotic variance-covariance matrix of true MLE even if the MC error in MC MLE is small enough to consider the MC MLE as the true MLE. Therefore, we cannot perform interesting hypothesis tests exactly nor asymptotically. The third open question is how to use the estimation process developed in Chapters 4 and 5 in different models including three or more stage hierarchical models and in conditionally specified statistical mixture models where not only mixing distributions but also data models are identified by a set of conditional distributions. Finally, the possibility of incorporation of variance reduction techniques into the sampling procedure should be considered. As is often true in statistical applications of Monte Carlo methodology, importance sampling has been used to allow a successful approach to be developed in the first place, rather than as a variance reduction technique in its own right. As statistical applications of Monte Carlo techniques become more advanced, the issue of variance reduction will likely see more emphasis. Our construction of sampling distributions as a simple product of marginals would seem well suited for the pursuit of variance reduction goals in complex models since the process of sampling from the dummy model involves only the generation of independent values. Basic Monte Carlo variance reduction techniques might be easily applied to improve performance of the estimation procedure further, or ease the burden of required Monte Carlo sample size. For example, lack of dependence among sampled values should allow antithetic sampling to be easily incorporated into the sampling procedure. The adjective ‘easily’ is appropriate here because sampling from our importance distributions does not require the use of Markov chain methods, although the formulation of those importance distributions does.
BIBLIOGRAPHY


ACKNOWLEDGMENTS

First of all, I thank the God who has always been with me. Without the help and guidance of the God, I could not get what I have and could not become who I am.

I am indebted to my family, Mikyung, Inyoung, and Dongho. They always pray for me wherever I am. They always believe me whatever I do. They always agree with me whatever I insist.

I am especially grateful to the lots of people to support me financially and spiritually, my parents, brothers, mother-in-law, brothers-in-law, Mr. and Mrs. Song, and friends in Korea.

I am extremely grateful to my advisors, Professor Mark Kaiser and Professor Noel Cressie for their expert guidance and valuable suggestions in my research. They provided me with one of the best environments a graduate student could ask for so that I can feel stable, comfortable, and confident in my works.

I would like to thank the members of my committee: Dr. Meeker, Dr. Song, and Dr. Stern for their suggestions.

I also wish to be thankful to Youngwoo Park and Samyeong Kim for sacrificing them to show me how to live for the person in need. Their ways of living and thoughts greatly affect me in improving my personality.

I would like to thank Jaekwang Kim with whom I had many useful and enlightening discussions in theoretical aspects of my thesis and thank Uk Yang who has been with me giving laughs and smiles when I had a hard time.

I am grateful to Chang-kwan Lee, Hyun-soo Lee, Sun-min Park, Won-seong Oh, and Yeon-soo Kim for allowing me to use their computer accounts.