DISSERTATION

RANDOM EFFECTS GRAPHICAL MODELS FOR DISCRETE COMPOSITIONAL DATA

Submitted by

Devin S. Johnson

Department of Statistics

In partial fulfillment of the requirements

for the Degree of Doctor of Philosophy

Colorado State University

Fort Collins, Colorado

Fall 2003

COLORADO STATE UNIVERSITY

September 5, 2003

WE HEREBY RECOMMEND THAT THE DISSERTATION RANDOM EFFECTS GRAPHICAL MODELS FOR DISCRETE COMPOSITIONAL DATA PREPARED UNDER OUR SUPERVISION BY DEVIN S. JOHNSON BE ACCEPTED AS FULFILLING IN PART REQUIREMENTS FOR THE DEGREE OF DOCTOR OF PHILOSOPHY.

Committee on Graduate Work

Committee Member

Committee Member

Committee Member

Adviser

Department Head

ABSTRACT

RANDOM EFFECTS GRAPHICAL MODELS FOR DISCRETE COMPOSITIONAL DATA

In this dissertation, we consider state-space models for the analysis of discrete compositional data. Compositional data are non-negative multivariate vectors that lay on the simplex defined by the sum-to-one constraint. The sum-to-one constraint simply implies that the vector elements sum to one (or some other scaler constant) for every element of the multivariate sample space. Discrete compositional data are multivariate vectors of integer counts that have been normalized to give the relative abundance of each element of the multivariate count vector. The logistic normal (LN) distribution and the associated perturbation operator provide a flexible model for compositional data. However, the LN distribution may be a poor model for discrete compositional data due to the extra sampling variability of integer counts and the possible presence of zeros in the compositional observation. Here, we propose a class of state-space models for compositional data based on traditional graphical models.

Graphical models are distributions for analyzing the conditional relationships of a Markov random field. We propose a two component graphical chain model, the discrete regression distribution, in which a set of categorical (or discrete) random variables is modeled as a response to a set of categorical and continuous covariates. This new graphical model, for a single observation of a multivariate count vector, serves as the basis for a state-space model for compositional data. We examine necessary and sufficient conditions for a discrete regression distribution to be described by the graph of a Markov random field.

The discrete regression formulation is extended to a state-space representation for the analysis of many discrete compositional observations. Models are constructed for compositions defined by a single classification criteria and, also, those defined by multiple classification criteria. We define an extended chain graph which possesses an extra vertex associated with the random state. Necessary and sufficient conditions are given for a random effects discrete regression to be Markovian with respect to this extended graph. We also give sufficient conditions for the Markov properties of the marginal distribution of the covariates and categorical response. A Bayesian approach to parameter inference is adopted. Markov chain Monte Carlo (MCMC) methods are used for estimated for data sets concerning feeding type composition of stream invertebrates in Oregon and fish species richness in the Mid-Atlantic Highlands.

Following the analysis of traditional discrete compositional data, we examine a state-space representation of another type of ecological composition analysis, capture-recapture models. Capture-recapture models provide inference to survival in wild animal populations. In state-space capture-recapture models the survival rate for each time period represents an unobserved composition. We illustrate why capture-recapture models are nearly identical to traditional multi-way discrete composition models. An autoregressive random survival state is incorporated into traditional capture-recapture models and an MCMC methodology presented for inference. The methodology is demonstrated on a long term data set of marked Pintail ducks.

Devin S. Johnson
Department of Statistics
Colorado State University
Fort Collins, Colorado 80523
Fall 2003

ACKNOWLEDGEMENTS

After five years in the Statistics department at Colorado State University it's hard to summarize all of the individuals who helped form my education. But, none the less, I will try to do just that in this brief space. First and foremost, I would like to thank my wife, Shea, for her tireless support during grad school. The past five years would not have been possible without her. Secondly, thanks to my adviser Dr. Jennifer Hoeting. From my Master's project through my dissertation, she was all that a student could hope for in an adviser. She allowed me the freedom to explore and make my own mistakes, but was always there to ask me to justify my convictions. I would like to acknowledge both Dr. Hari Iyer and Dr. Richard Davis for the stimulating discussions on statistical matters as well as other subjects. Also, thanks to my other committee members. Dr. F. Jay Breidt is the best proofreader I have ever met, and a great resource on a wide ranging set of statistical topics. Dr. David Anderson and Dr. Kenneth Burnham provided a biologists perspective and were always ready to ask very practical questions. I would like to thank STARMAP, and especially Dr. Scott Urquhart, for providing funding for my project and allowing me to work in a subject area close to my heart. Along those lines I would like to acknowledge Dr. LeRoy Poff, Dr. Brian Bledsoe, and Dr. Alan Herlihy for helping me with concepts of stream ecology and EPA datasets. Finally, I would like to thank all of my fellow graduate students for all those lively homework sessions at the white board in the middle of the night.

DEDICATION

This work is dedicated to my wonderful wife Shea. Without her none of this work would have been possible. She was the inexhaustible voice that kept me going on a very long journey from Fairbanks, AK to Fairbanks, AK.

CONTENTS

1 Introduction and Preliminaries 1
1.1 The Logistic Normal Distribution and the ALR Transformation 3
1.2 Modeling the LN Location Parameter with Covariates 4
1.3 Discrete Compositions: Statistical Concerns and the State-Space Model . 7
1.3.1 Statistical Concerns 8
1.3.2 A State-Space Model for Discrete Compositional Data 8
1.4 Graphical Models and the Analysis of Discrete Compositional Data . . . 11
1.5 Markov Chain Monte Carlo Algorithms for Bayesian Inference 13
1.5.1 Metropolis-Hastings Algorithm 14
1.5.2 Gibbs Sampler 15

2 Graphical Models and the Discrete Regression Distribution 18
2.1 Graph Theory Notation 20
2.2 Undirected Graphical Models 22
2.2.1 Markov Properties 23
2.2.2 Models for Discrete, Gaussian, and Mixed Variables 30
2.3 Chain Independence Graphs 36
2.3.1 Chain Graph Notation and Requirements 37
2.3.2 Markov Properties 39
2.3.3 Comments on Chain Graph Markov Properties 42
2.4 Discrete Regression Models 42
2.4.1 Model Formulation 43
2.4.2 Markov Properties of the DR Distribution 46

3 Bayesian Analysis of Discrete Compositional Data: A Graphical Model Approach 49
3.1 Introduction 49
3.2 Model Formulation 52
3.2.1 Single Individual and Single Site Models 52
3.2.2 Random Effects Discrete Regression 57
3.3 Markov Properties of the Random Effects DR Model 61
3.4 Parameter Inference 65
3.4.1 Hierarchical Centering Parameterization 67
3.4.2 Implementing the Gibbs Sampler 70
3.5 Graphical Analysis of Benthic Invertebrate Functional Groups 76

3.5.1 Data Description 77
3.5.2 Model Description and Analysis 80
3.5.3 Results and Discussion 84

4 State-Space Models for the Analysis of Multi-way Discrete Compositional Data 91
4.1 Introduction 91
4.2 Model Formulation 93
4.2.1 Models for a Single Individual at a Single Site 93
4.2.2 Single Site Models 96
4.2.3 Random Effects Discrete Regression 99
4.3 Markov Properties of Multi-Way Composition Models 104
4.3.1 Preservative REDR Models 107
4.3.2 Markov Properties of Preservative REDR Models 108
4.3.3 Correlated Random Effects and Preservative Models 110
4.4 Parameter Inference 110
4.4.1 Hierarchical Centering Parameterization 113
4.4.2 Implementing the Gibbs Sampler 116
4.5 Graphical Analysis of Fish Species Richness 122
4.5.1 Data Description 124
4.5.2 Model Description and Inference 127
4.5.3 Results and Discussion 131

5 Autoregressive Models for Capture-Recapture Data 136
5.1 Introduction 136
5.2 Likelihood for Capture-Recapture Data 138
5.2.1 Open population mark-recapture likelihood 139
5.2.2 Band recovery likelihood 139
5.2.3 Similarity with Random Effects Discrete Regression Models 140
5.3 A Bayesian Approach for AR(m) Survival Models 141
5.3.1 Model specification 142
5.3.2 Bayesian parameter estimation 143
5.4 Example: Northern Pintails 149
5.5 Discussion 153

6 Conclusions and Future Work 160
6.1 Discussion 160
6.2 Future Work 163
6.2.1 Model Selection 163
6.2.2 Spatio-Temporal Models 165

A WinBUGS Code for Compositional Data and AR(2) Capture-Recapture Models **166**
A.1 WinBUGS Code for a Single Composition Model 166
A.2 WinBUGS Code for a Multi-Way Composition Model 168
A.3 WinBUGS Code for an AR(2) Capture-Recapture Model 171

LIST OF FIGURES

2.1 Example of a graph 21
2.2 Example of an undirected independence graph 25
2.3 Illustration of chain graph concepts 38

3.1 Example of an extended graph for REDR models with a single response variable 62
3.2 Distribution of feeding type relative abundance. 79
3.3 Data suggested chain graph for feeding type composition. 89

4.1 Example of an extended graph for random effects Discrete Regression models 105
4.2 Distribution of cell counts for fish species tolerance and habit. 126
4.3 Data suggested chain graph for the multi-way composition of habit and tolerance. 135

5.1 Marginal Posterior densities for the parameters of an AR(2) band recovery model 158
5.2 Plot of yearly survival estimates for Northern Pintail dataset with no linear time trend 159

LIST OF TABLES

3.1 Summary of stream invertebrate feeding groups 78
3.2 Summary of environmental covariates for Oregon REMAP streams. . . . 80
3.3 HPD intervals for feeding type covariate interactions 87
3.4 HPD intervals for the off-diagonal elements of $\Psi_\emptyset$ 88
3.5 HPD intervals for feeding type predictions 90

4.1 Summary of environmental covariates for 1994 MAHA streams. 127
4.2 DIC and model complexity for multi-way fish species richness models . . 131
4.3 95% HPD intervals for covariate interaction parameters in the analysis of fish species richness. 133
4.4 HPD intervals for the off-diagonal elements of $\Psi_\emptyset$ for the MAHA environmental variables 134

5.1 Northern Pintail recovery data for banding years 1955 - 1983 156
5.2 Posterior means, standard deviations, and 90% highest probability density (HPD) intervals for the AR(2) model parameters 157

Chapter 1

INTRODUCTION AND PRELIMINARIES

Compositional data are D dimensional multivariate observations whose probability space consists of the $d = D - 1$ dimensional simplex ∇^d. In other words, a compositional observation $\mathbf{P} = (P_1, \ldots, P_D)$ possess the two constraints: $\sum_j P_j = 1$ and $P_j \geq 0$ for $j = 1, \ldots, D$. Compositional data are preferred to the unconstrained positive multivariate observations if relative size is more important than the absolute values of each vector element.

In a detailed description of compositional data Aitchison (1986, pg. 48) describes three challenges in modeling compositional vectors. First, there is a necessary correlation structure due to the sum-to-one constraint. In fact, Pearson (1897) used compositional data as an example of spurious correlation. Secondly, there is an absence of an interpretable correlation structure. Not all positive definite matrices are valid covariance matrices for compositional random vectors. Finally, many existing models impose a rigid correlation structure that is due solely to the sum-to-one constraint. The Dirichlet distribution possesses this inflexibility.

Compositional data are often encountered in disciplines such as geology, biology, ecology, economics, and chemistry. In this dissertation, we will focus on compositional data arising from ecological monitoring programs. Direct observation of organisms in the environment is often a key component of determining ecosystem health. Often, sampling schemes for collecting individuals preclude use of the total number of individuals to make inference to ecosystem health. The total number of individual organisms observed from various taxa at a sampling site may not be

representative of the actual number of individuals inhabiting the site. Another possibility is that the total number of individual organisms at a site may simply reflect overall site productivity. The relative proportions of individuals in various taxa are often what is of interest in this case.

An emerging area of research in ecology is the analysis of functional species assemblages (Poff and Allen, 1995). In essence, the analysis of functional assemblages is concerned with determining and predicting the composition of individuals categorized according to different life history traits instead of strict taxa names (Poff, 1997). This provides an inference that is portable to other ecosystems. With this goal in mind, the remainder of this dissertation will be devoted to the development of models that allow an examination of environmental variable influence and permit prediction of functional trait compositions. Although the models are developed with this end goal in mind, they are completely general in that they need not be limited to ecological data of species abundance.

The dissertation is organized in the following manner. Chapter 1 provides an introduction to the analysis of compositional data and Bayesian Markov chain Monte Carlo (MCMC) methodology, which we use for parameter estimation. Chapter 2 examines graphical models, which we use as a basis for building compositional data models. Graphical modeling provides a method for examining complicated relationships between variables, both continuous and categorical, in a multivariate vector. Therefore, graphical models serve as a strong basis for exploring relationships between a categorical life history trait and environmental covariates. In Chapter 2, we propose a graphical model, the discrete regression distribution, which will serve as the basis of our analysis of compositional data. In Chapters 3 and 4 we extend the discrete regression model for single composition and multi-way composition analysis, respectively. In Chapter 5, we examine composition models for capture-recapture data. Finally, in Chapter 6 we present our conclusions along with some possible extensions for future work.

1.1 The Logistic Normal Distribution and the ALR Transformation

Aitchison (1982) proposes the logistic-normal (LN) distribution in order to improve on the inflexibility of previous models for compositional data. The LN distribution is built on the foundation of a Multivariate Normal (MVN) distribution. First, draw a d dimensional MVN random vector $\mathbf{X} = (X_1, \ldots, X_d)$ with mean $\boldsymbol{\mu}$ and covariance $\boldsymbol{\Sigma}$. Then, perform the following one-to-one transformation,

$$\mathbf{P} = \mathbf{g}(\mathbf{x}) = \left[\frac{e^{X_1}}{1 + \sum_{k=1}^{d} e^{X_k}}, \ldots, \frac{e^{X_d}}{1 + \sum_{k=1}^{d} e^{X_k}}, \frac{1}{1 + \sum_{k=1}^{d} e^{X_k}} \right] \tag{1.1}$$

Then, the vector $\mathbf{P} = (P_1, \ldots, P_D)$, where $D = d+1$ will have a LN distribution. It can be clearly observed that $\mathbf{P}$ is contained within ∇^d with probability 1. Conversely, one can perform the following transformation of a composition vector $\mathbf{P}$,

$$X_j = \log(P_j/P_D), \tag{1.2}$$

for $j = 1, \ldots, d$ and $D = d + 1$. The transformation in (1.2) is known as the *additive log ratio* (ALR) transformation and is a one-to-one transformation from ∇^d to $\mathbb{R}^d$. If the vector $\mathbf{X} = (X_1, \ldots, X_d)$, from (1.2), has a MVN distribution with mean $\boldsymbol{\mu}$ and covariance $\boldsymbol{\Sigma}$, then $\mathbf{P}$ has a LN($\boldsymbol{\mu}$, $\boldsymbol{\Sigma}$) distribution. Aitchison and Shen (1980) introduce the LN distribution and describe various properties. The LN($\boldsymbol{\mu}, \boldsymbol{\Sigma}$) density f_{LN} for a vector of compositional random variables $\mathbf{P}$ is given by,

$$\begin{aligned} f_{LN}(\mathbf{p}) = (2\pi)^{-d/2} |\boldsymbol{\Sigma}|^{-1/2} (p_1 \ldots p_D)^{-1} \\ \times \exp\left[-\frac{1}{2} \{\mathbf{h}(\mathbf{p}) - \boldsymbol{\mu}\}' \boldsymbol{\Sigma}^{-1} \{\mathbf{h}(\mathbf{p}) - \boldsymbol{\mu}\} \right], \end{aligned} \tag{1.3}$$

where

$$\mathbf{h}(\mathbf{p}) = \{\log(p_1/p_D), \ldots, \log(p_d/p_D)\}'. \tag{1.4}$$

The location parameter $\boldsymbol{\mu}$ and scale parameter $\boldsymbol{\Sigma}$ provide greatly improved flexibility over previous distributions when modeling compositional data. For example, the LN distribution can be constructed from *correlated* normal variables,

while the Dirichlet distribution, a previously common compositional data model, is constructed from normalizing *independent* gamma variables (Aitchison, 1986). This composition of independent variables induces a rigid correlation structure for the compositional model.

Other models have been proposed for compositional data. Most recently, researchers have investigated the use of the Liouville distribution (Smith and Rayens, 2002; Iyenger and Dey, 2002). Gupta and Richards (2001) give a detailed description of the Liouville and Dirichlet distributions. Barndorff-Nielsen and Jørgensen (1991) proposed a distribution which they titled the S^- distribution. This distribution, however, suffers, the same inflexibility as the Dirichlet distribution. It does, however, posses nice mathematical properties compared to the LN distribution. For example, the S^- class of distributions is closed under marginalization, unlike the LN distribution. Stephens (1982) proposes the von Mises distribution for the square root transformed composition components. This model seems to have received little attention, perhaps, due to the complexity of the von Mises distribution or lack of real world interpretability of the process.

1.2 Modeling the LN Location Parameter with Covariates

In this section, we provide an introduction to log-ratio linear modeling. Let $\mathbf{P} \sim \text{LN}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, where $\mathbf{P}$ is given in (1.1). Log-ratio linear modeling essentially makes use of a linear model to parameterize the location vector $\boldsymbol{\mu}$. We give a brief introduction here, by only considering a model with one continuous covariate. The extension to more covariates, as well as categorical covariates, is straight forward. Aitchison (1986, pg. 158) gives a through description and general formulation of the log-ratio model.

Developing a linear model to allow covariate information to explain or predict a compositional vector is extremely difficult. The unusual shape of the simplex

∇^d, due to the sum-to-one constraint, prevents the use of a standard linear model to model a composition observation. Linear model theory and applications have been explored for models in the unconstrained $\mathbb{R}^d$ space, however. Therefore, the log-ratio linear model proceeds by applying the ALR transform to a composition observation and using a linear model to parameterize the normal distribution of the transformed composition. For a single composition vector $\mathbf{P}$ and a covariate u, the log-ratio linear model is defined as

$$\mathbf{h}(\mathbf{P}) = \boldsymbol{\beta}_0 + \boldsymbol{\beta}_1 u + \boldsymbol{\epsilon}, \tag{1.5}$$

where $\mathbf{h}(\mathbf{P})$ is given by (1.4), $\boldsymbol{\beta}_0$ is a d vector of intercept parameters, $\boldsymbol{\beta}_1$ is a d vector of regression coefficients, and $\boldsymbol{\epsilon}$ is a d vector of error terms. As in the usual multivariate regression model, $\boldsymbol{\epsilon}$ has a MVN distribution with mean $\mathbf{0}$ and covariance matrix $\boldsymbol{\Sigma}$. The resulting distribution of $\mathbf{P}$ is $\mathrm{LN}(\boldsymbol{\beta}_0 + \boldsymbol{\beta}_1 u, \boldsymbol{\Sigma})$. The model in (1.5) can be written in a form that is identical to standard multivariate regression when many composition observations, $\mathbf{P}_i,\ i = 1, \ldots, S$, are available,

$$\mathbf{H} = \mathbf{X}\mathbf{B} + \mathbf{E}, \tag{1.6}$$

where the rows of $\mathbf{H} = \mathbf{h}(\mathbf{P}_i)$, the rows of $\mathbf{X} = (1,\ u_i)$, $\mathbf{B}$ is a $2 \times d$ matrix with the first row given by $\boldsymbol{\beta}_0$ and the second row by $\boldsymbol{\beta}_1$, and the rows of $\mathbf{E}$ are the error vectors $\boldsymbol{\epsilon}_i$. A standard assumption for inference of the parameters in the log-ratio linear model is that the error terms $\boldsymbol{\epsilon}$ in (1.5) are independent. Several authors have explored extensions to the standard log-ratio model. Billheimer and Guttorp (1997) considered a lattice spatial process for the error terms in order to include spatial correlation for compositions with a spatial index. Tjelmeland and Lund (2003) recently examined a model for compositions observed in a continuous spatial domain. Essentially, instead of a lattice spatial process for the error term, Tjelmeland and

Lund (2003) use a multivariate geostatistical model with the following covariance function

$$c_{\theta,\Psi}(u,u') = \alpha_\theta(u,u')\Psi, \tag{1.7}$$

where u and u' are spatial locations, $\alpha_\theta(u,u')$ is a scalar covariance function, and Ψ is a $d \times d$ matrix. Tjelmeland and Lund (2003) term the spatial composition process a *logistic Gaussian field.* In a similar manner, Brunsdon and Smith (1998) model compositional time series by modeling the error term with a vector ARMA process to induce serial correlation in a series of compositions from repeated surveys. Billheimer (1995) also examines an ARMA representation for composition time series in terms of the perturbation operator. Abbitt and Breidt (2001) consider a slightly different type of model, a "measurement error" model, for soil composition. These authors use a hierarchical model to estimate the actual soil composition in the field when the data available are laboratory measurements, which are suspected to be biased and have highly variable measurement error.

With the exception of Brunsdon and Smith (1998), all of the dependent error models presented were studied under a Bayesian inference paradigm. The non-linear inverse ALR transformation (1.1) often makes Bayesian methods more attractive, as likelihood methods become impractical. In our investigation of compositional data, we also pursue model inference in a Bayesian setting.

Another interesting property of the LN distribution and log-ratio linear models is that the location parameter $\boldsymbol{\mu}$, can be represented as a composition. This provides the benefit that $\boldsymbol{\mu}$ can be interpreted as an element from ∇^d. For an explanation, we must begin with the concept of the *perturbation* operator.

The *perturbation operator* provides a method of "addition" in ∇^d. For a composition element $\mathbf{p} \in \nabla^d$ and a vector $\boldsymbol{\epsilon} \in \mathbb{R}^D$, we define the "perturbation" $\mathbf{p} \circ \boldsymbol{\epsilon}$ of $\mathbf{p}$ by $\boldsymbol{\epsilon}$ as

$$\mathbf{p} \circ \boldsymbol{\epsilon} = \left[\frac{p_1\epsilon_1}{\sum_{j=1}^D p_j\epsilon_j}, \ldots, \frac{p_D\epsilon_D}{\sum_{j=1}^D p_j\epsilon_j} \right] \tag{1.8}$$

Billheimer (1995) details the properties of the perturbation operator and uses it as a basis for proving that ∇^d is a Hilbert space.

Now, to give the representation of $\boldsymbol{\mu}$ as a composition, we assume that $\mathbf{P}$ has a $\text{LN}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ distribution, then taking the ALR transform we can formulate the linear model

$$\mathbf{h}(\mathbf{P}) = \boldsymbol{\mu} + \boldsymbol{\epsilon}, \tag{1.9}$$

where $\boldsymbol{\epsilon}$ has a zero mean MVN distribution with covariance $\boldsymbol{\Sigma}$. We now can apply the inverse ALR transformation, $\mathbf{g}$ in (1.1) to obtain

$$\begin{aligned} \mathbf{P} &= \left[\frac{e^{\mu_j+\epsilon_j}}{1+\sum_{k=1}^{d} e^{\mu_k+\epsilon_k}}, \ldots, \frac{e^{\mu_j+\epsilon_j}}{1+\sum_{k=1}^{d} e^{\mu_k+\epsilon_k}}, \frac{1}{1+\sum_{k=1}^{d} e^{\mu_k+\epsilon_k}} \right] \\ &= \mathbf{g}(\boldsymbol{\mu}) \circ \mathbf{g}(\boldsymbol{\epsilon}). \end{aligned} \tag{1.10}$$

So, an LN distributed composition can be represented as a fixed composition $\mathbf{g}(\boldsymbol{\mu})$ that is perturbed by a random "zero mean" composition $\mathbf{g}(\boldsymbol{\epsilon})$.

1.3 Discrete Compositions: Statistical Concerns and the State-Space Model

The inverse ALR transformation (1.1) illustrates that the LN distribution is designed to model continuous compositions. By continuous composition, we mean that the composition measures the relative size of each element of a continuous multivariate vector. *Discrete compositions* arise by examining the relative size of each element of a vector containing integer values. Such a vector observation might arise with count data. For example, with the ecological data considered in this dissertation we are interested in the relative abundance of stream invertebrate traits at sampled stream sites or the relative abundance of fish species with certain life history traits. For these types of data, the LN distribution may not be a good model.

1.3.1 Statistical Concerns

The LN model is often a poor choice for discrete compositions. There is often an inherent increase is compositional variability due to sampling error with discrete compositional data. If, for a given compositional observation, a small sample of individuals is selected, the calculated relative abundances may have substantial error compared to the true composition of individuals. This is the case with the fish species data set examined in Chapter 4. The total number of fish species observed at a sampled stream may be small, so, there may be substantial variability in the observed abundances and the actual abundance due to the integer nature of the data. If the number of individuals observed at a site is large, this may not be as much of a problem. This is the case with the data considered in Chapter 3. In Chapter 3, we examine relative abundance of stream invertebrates. In each sample, the relative abundance of individual invertebrates is examined. Unlike the fish data, large counts are observed at each sampled stream site.

The second drawback when using the LN density is that of zero counts for a given category. The LN density (1.3) is not defined for a relative abundance of zero for any category. This can be a serious problem with abundance counts, as there is often a non-trivial probability that one will observe an abundance of zero for at least one of the categories. Indeed, this is a real problem for data like those examined in Chapters 3 and 4. Aitchison (1986, pg. 271) proposes a solution to this with a mixture of LN distributions. The solution he proposes, however, is *ad hoc* and can become overly complicated if the number of categories with zero entries is large. A more practical solution using a state-space model was proposed by Billheimer (1995).

1.3.2 A State-Space Model for Discrete Compositional Data

State-space models provide a conceptual approach for complex modeling situations by linking hypothetical (or real) unobserved variables, which represent the

state of nature, and observed data. The states are allowed to vary over space, time, or some other index in response to covariates or possibly each other. State-space models provide an intuitive way to model complex data by building a large model from two smaller models. The two pieces in a state-space model are a model for the state of nature and a model for observed data given the state of nature at the location and time the data are observed. It is often easier to build a large model, with the desired properties, in this hierarchical fashion.

State-space models have often been used in time series analysis. The hierarchical structure of their design allows one to build a model with complex serial correlation structure though several small models. This is usually necessary with serially correlated data that is non-Gaussian such as a time series of count or categorical data. Durbin and Koopman (2000) provide an analysis of non-Gaussian time series models from both a Bayesian perspective and a "classical" likelihood perspective. For an observed time series $\{y_t\}$ the Durbin-Koopman state-space model is of the form

$$f(y_t|\alpha_1, \ldots, \alpha_t, y_1, \ldots, y_{t-1}) = f(y_t|\alpha_t) \tag{1.11}$$

for the observed data y_t and the current state α_t, where f is an exponential family distribution. The states $\{\alpha_t\}$ are modeled as

$$\alpha_t = T_t\alpha_{t-1} + R_t\eta_t, \qquad \eta_t \sim \text{ indep. } p(\eta_t). \tag{1.12}$$

So, for example, one could allow f to be a Poisson or multinomial likelihood in order to build a model for count data that is marginally correlated in time. In general, the states may or may not be of interest. The states can simply serve as a conceptual tool for modeling complex data, such as serially correlated count data, or, the states may be the primary interest. In the former case, the observations may represent data with "measurement error" where the true measurement is unobservable. The

second situation is the inspiration that Billheimer (1995) use to develop a state-space model for discrete compositional data. In this dissertation, however, we use both situations for inspiration.

Billheimer (1995) proposed the following state-space model to eliminate the problems with the LN model caused by zeros and low abundance counts when modeling species composition. First, the observed abundance counts, which represent the observed data, are modeled with the multinomial likelihood

$$f_M(\mathbf{c}|N,\mathbf{p}) = \frac{N!}{\prod_{j=1}^{D} c_j!} \prod_{j=1}^{D} p_j^{c_j}, \tag{1.13}$$

where N represents the total number of individuals at a site, c_j represents the number of individuals in category $j = 1,\ldots,D$, and $\mathbf{p} = (p_1,\ldots,p_D)$ represents the true composition (the state) of individuals. The true composition (or state) is then modeled with a LN$(\boldsymbol{\mu},\boldsymbol{\Sigma})$ distribution (1.3), where $\boldsymbol{\mu}$ may be parameterized with the linear model (1.5). Here, the unobserved states are the objects of interest and the counts represent a "measurement error" observation. In addition, the parameters of the log-ratio linear model may also be of interest, or they may represent an "adjustment" for an observed covariate (Billheimer, 1995, pg. 125).

Billheimer's state-space model eliminates the requirement of a positive observed abundance enforced by the LN distribution by incorporating a sampling error process as part of the model. The true compositions are modeled as a positive vector, however, the multinomial sampling of individuals allows the possibility of observing an abundance of zero even though the true relative abundance is positive. One drawback to this model is that, while this model eliminates the need to observe only non-zero counts, it still makes the assumption that the true relative abundance is positive. So, this model is still inappropriate for situations where one would like to make inference to the absence of a certain category. If, however, a positive value for all compositional elements makes sense, then this model provides a way to

incorporate the extra multinomial sampling variability into estimation of the true composition.

In this dissertation, we extend the approach of Billheimer (1995) in several ways. First, we examine a model with many covariates and consider the multivariate vector, composed of covariates and categorical responses, as a realization of a Markov random field. Conditions concerning the parameters of the model are given which determine the conditional independence properties of the Markov random field. The second extension is a model which allows multi-way compositional data to be modeled. Multi-way compositions arise when individuals are cross-classified according to several traits. Finally, we adopt a graphical model framework which allows for easily interpretable visual representations of the relationships between the variables of interest.

1.4 Graphical Models and the Analysis of Discrete Compositional Data

A graphical model is a probability density function for a multivariate vector that is parameterized in such a way that a complex independence structure can be characterized by a mathematical graph. Graphical models for multivariate categorical variables have been studied for years in the form of log-linear models for contingency tables (Whittaker, 1990). One cannot help but notice the similarity in data structure between a contingency table and discrete compositional data, the main difference being that a contingency table represents a single sample from one population. Discrete compositional data, however, consist of many samples (one for each compositional observation) from many different populations of individuals.

In the contingency table scenario, there exists one set of probabilities for each category of the table. This set of unknown probabilities represents the one unknown composition parameter. This one set of probabilities (composition) is usually modeled with a log-linear model (Christensen, 1990) for testing hypotheses of conditional

independence of the categorial variables that compose the classifications for the categories. Discrete compositional data arise from collection of "contingency table data" from many sites, times, or other such index. Therefore, instead of one category composition of interest, as with contingency tables, there is one composition for every sampled observation. Aitchison (1986, pg. 327) actually describes this relationship in one brief sentence. He notes that compositions originating from discrete counts could be thought of as "random effects" for standard categorical data analysis techniques such as log-linear modeling. Aitchison's brief comment concerning the link between categorical data analysis and discrete compositional data serves as the driving force behind this dissertation.

Research in graphical models has grown considerably over recent years. Whittaker (1990) and Lauritzen (1996) offer comprehensive overviews of the field. This growth in research effort is undoubtedly due to wide applicability of graphical models in many areas of statistics. These models provide methods for examining complex relationships present in multivariate distributional models. These complex relationships are described by the Markov properties of the distribution.

Lauritzen and Wermuth (1989) extended the class of graphical log-linear models to allow joint modeling of continuous and discrete variables. These graphical models will form the basis for examining relationships between species abundance in discrete categories and environmental covariates. We will provide a random effects formulation for the models and examine the Markov properties of these graphical models for compositional data. The benefit of forming compositional data models as random effects graphical models is that it allows a visual representation of the relationships between variables in the model. For example, in the analysis of fish species richness, we can visually examine the relationships between, possibly, several life history traits and several environmental covariates. Therefore, we can visualize the system as a whole.

1.5 Markov Chain Monte Carlo Algorithms for Bayesian Inference

Parameter inference is based on Bayesian methodology for the models proposed in this dissertation. Under the Bayesian paradigm of inference one first formulates a plausible model to generate a set of data $\mathbf{x}$; a likelihood model $\mathcal{L}(\mathbf{x}|\boldsymbol{\theta})$, where $\boldsymbol{\theta}$ represents a set of parameters. Then, one formulates a probability model that reflects beliefs about the parameters prior to the collection of data. This prior belief model is the *prior distribution*, $\pi(\boldsymbol{\theta})$, $\boldsymbol{\theta} \in \Theta$, of the parameters. Under the Bayesian paradigm of estimation, once the data are collected, they are considered fixed constants. After data collection, therefore, the inference concerning model parameters is based on the conditional distribution of the parameters given the observed data. This distribution is the *posterior distribution* of the parameters

$$\pi(\boldsymbol{\theta}|\mathbf{x}) = \frac{\mathcal{L}(\mathbf{x}|\boldsymbol{\theta})\pi(\boldsymbol{\theta})}{C(\mathbf{x})}, \tag{1.14}$$

where $C(\mathbf{x}) = \int_\Theta \mathcal{L}(\mathbf{x}|\boldsymbol{\theta})\pi(\boldsymbol{\theta})d\boldsymbol{\theta}$. Functionals of the distribution $T\{\pi(\boldsymbol{\theta}|\mathbf{x})\}$ such as the expected value, $Eh(\boldsymbol{\theta}) = \int_\Theta h(\boldsymbol{\theta})\pi(\boldsymbol{\theta}|\mathbf{x})d\boldsymbol{\theta}$, for an integrable function h are the primary objects of interest for parameter inference. Choices of h often used are the identity function, to obtain the posterior expected value of the parameters, the squared deviation $(\boldsymbol{\theta} - E\boldsymbol{\theta})^2$ to obtain the posterior variance, or the indicator function $I_{\boldsymbol{\theta}<\alpha}(\boldsymbol{\theta})$, to obtain a posterior α probability.

Unfortunately, closed forms of the summary functionals $T\{\pi(\boldsymbol{\theta}|\mathbf{x})\}$ are rarely available in practice. Instead, a sample, $\boldsymbol{\theta}^{(1)}, \ldots, \boldsymbol{\theta}^{(N)}$, is drawn from (1.14) and used to approximate functionals such as expected values and quantiles with the sample version

$$\begin{aligned} T\{\pi(\boldsymbol{\theta}|\mathbf{x})\} &\approx T\{\hat{\pi}_N(\boldsymbol{\theta}|\mathbf{x})\} \\ &= \frac{1}{N}\sum_{i=1}^{N} h(\boldsymbol{\theta}^{(i)}), \end{aligned} \tag{1.15}$$

where $\hat{\pi}_N(\boldsymbol{\theta}|\mathbf{x})$ is the empirical posterior distribution. Markov Chain Monte Carlo (MCMC) is a collection of methods for drawing the sample values $\boldsymbol{\theta}^{(i)}$, $i = 1, \ldots, N$.

The sample values are drawn as a realization of an ergodic Markov chain with stationary distribution $\pi(\boldsymbol{\theta}|\mathbf{x})$. The ergodicity of the Markov chain ensures that

$$\frac{1}{N}\sum_{i=1}^{N} h(\boldsymbol{\theta}^{(i)}) \longrightarrow \int_{\Phi} h(\boldsymbol{\theta})\pi(\boldsymbol{\theta}|\mathbf{x})d\boldsymbol{\theta} \qquad \text{a. s.,} \tag{1.16}$$

as $N \rightarrow \infty$ and for sufficiently large N, the initial value $\boldsymbol{\theta}^{(0)}$ becomes irrelevant (Tierney, 1994). Therefore, the approximation, theoretically, can be as accurate as desired. There are many methods for constructing transition kernels for MCMC procedures. Chen et al. (2000) provide an overview of MCMC methods. Robert and Casella (1999) provide a mathematically rigorous exposition on continuous state Markov chains and various MCMC procedures. We will give a brief description, here, of two general MCMC algorithms, the Metropolis-Hastings algorithm and the Gibbs sampler.

1.5.1 Metropolis-Hastings Algorithm

The Metropolis-Hastings (MH) algorithm (Metropolis et al., 1953; Hastings, 1970) is the most general of the MCMC algorithms. The MH algorithm starts with a target density. Since this is a general description, we shall use the generic notation π for this density, with the understanding that for Bayesian analysis $\pi = \pi(\boldsymbol{\theta}|\mathbf{x})$ in (1.14). Next, a *proposal density* q is defined on the same support as π. The user must choose q such that the resulting chain will be irreducible and aperiodic. This is usually satisfied if the support of q is connected and q is continuous. Now, a Markov chain generated from the follow algorithm will have stationary distribution π.

Metropolis-Hastings Algorithm

Given $\boldsymbol{\theta}^{(t)}$,

1. Generate $\boldsymbol{\theta}'$ from $q(\boldsymbol{\theta}|\boldsymbol{\theta}^{(t)})$.

2. Take

$$\boldsymbol{\theta}^{(t+1)} = \begin{cases} \boldsymbol{\theta}' \text{ with probability } \alpha(\boldsymbol{\theta}^{(t)}, \boldsymbol{\theta}'), \\ \boldsymbol{\theta}^{(t)} \text{ with probability } 1 - \alpha(\boldsymbol{\theta}^{(t)}, \boldsymbol{\theta}'), \end{cases}$$

where

$$\alpha(\boldsymbol{\theta}^{(t)}, \boldsymbol{\theta}') = \min \left\{ \frac{\pi(\boldsymbol{\theta}')}{\pi(\boldsymbol{\theta}^{(t)})} \frac{q(\boldsymbol{\theta}^{(t)}|\boldsymbol{\theta}')}{q(\boldsymbol{\theta}'|\boldsymbol{\theta}^{(t)})}, \ 1 \right\}$$

Given a starting value $\boldsymbol{\theta}^{(0)}$, the Markov chain $\boldsymbol{\theta}^{(0)}, \boldsymbol{\theta}^{(1)}, \boldsymbol{\theta}^{(2)}, \ldots$ defined by the MH algorithm will have π as a stationary distribution. Therefore, after a suitable number of iterations N, $\boldsymbol{\theta}^{(N)}, \boldsymbol{\theta}^{(N+1)}, \boldsymbol{\theta}^{(N+2)}, \ldots$ is a sample with distribution that is approximately π. Notice, the transition kernel of the Markov chain only depends on the ratio $\pi(\boldsymbol{\theta}')/\pi(\boldsymbol{\theta}^{(t)})$, therefore if π is the posterior distribution (1.14), there is no need to calculate the normalizing coefficient $C(\mathbf{x})$. It is for this reason that MCMC methods are so popular for Bayesian applications.

1.5.2 Gibbs Sampler

The Gibbs sampler is actually a special case of the MH algorithm, although, it was developed separately. For the Gibbs sampler, we will need some extra notation. Let $\boldsymbol{\theta} = (\theta_1, \ldots \theta_p)$ and have distribution π as in the MH algorithm. Again, this distribution could be a posterior distribution (1.14). Moreover, suppose it is possible to simulate

$$\tilde{\theta}_i|\theta_1, \ldots, \theta_{i-1}, \theta_{i+1}, \ldots, \theta_p \sim f_i(\theta_i|\theta_1, \ldots, \theta_{i-1}, \theta_{i+1}, \ldots, \theta_p)$$

for $i = 1, \ldots, p$. The Gibbs sampling algorithm is then defined by the following transitions from $\boldsymbol{\theta}^{(t)}$ to $\boldsymbol{\theta}^{(t+1)}$.

Gibbs Sampler

Given $\boldsymbol{\theta}^{(t)}$, generate

1. $\theta_1^{(t+1)} \sim f_1(\theta_1|\theta_2^{(t)}, \ldots, \theta_p^{(t)})$

2. $\theta_2^{(t+1)} \sim f_2(\theta_2|\theta_1^{(t+1)}, \theta_3^{(t)} \ldots, \theta_p^{(t)})$

$\vdots$

p. $\theta_p^{(t+1)} \sim f_p(\theta_p|\theta_1^{(t+1)}, \ldots, \theta_{p-1}^{(t)})$

The densities $f_1, \ldots, f_p$ are called *full conditional* densities. A nice feature of the Gibbs sampler that they are the only densities used for sampling, so, all updates can be univariate if desired. One does not not need to propose an entire vector of parameters, as with the MH algorithm. Another feature is that there is no "accept or reject" step in the algorithm. Robert and Casella (1999, Chap. 7) describe several conditions on π that are necessary for the complete set of full conditional densities to describe fully the joint density.

It may appear, at first, that the Gibbs sampler is always a more efficient sampler than the MH algorithm since it makes use of the actual distributions of the simulated variables. There are, however, some issues that can arise when using the Gibbs sampler. The "one-at-a-time" updating of parameters can be thought of as analogous to the maximization of a multivariate function one element at a time. The Gibbs sampler has a higher tendency than the MH algorithm to get "stuck" when π is multi-modal. The second problem is that of non-standard full conditionals.

In the definition of the Gibbs sampler, we assumed that the full conditionals could be easily sampled. If, however, f_i is a non-standard or un-normalized density, then the Gibbs sampling algorithm cannot be followed exactly. There is a solution that allows the Gibbs approach to be used in this case. The solution is the *Metropolis-within-Gibbs* (MWG) sampler. The MWG sampler is defined by first choosing a proposal density $q_i(\theta_i|\theta_1, \ldots, \theta_{i-1}, \theta_{i+1}, \ldots, \theta_p)$ for $f_i(\theta_i|\theta_1, \ldots, \theta_{i-1}, \theta_{i+1}, \ldots, \theta_p)$ for any i where f_i is not available in closed form. The MWG algorithm proceeds by replacing $\theta_i^{(t+1)} \sim f_i(\theta_i|\theta_1^{(t+1)}, \ldots, \theta_{i-1}^{(t+1)}, \theta_{i+1}(t), \ldots, \theta_p^{(t)})$ with the steps described below.

Metropolis-Within-Gibbs Sampler

Given $(\theta_1^{(t+1)}, \ldots, \theta_{i-1}^{(t+1)}, \theta_i^{(t)}, \ldots, \theta_p^{(t)})$:

1. Simulate

$$\tilde{\theta}_i \sim q_i(\theta_i | \theta_1^{(t+1)}, \ldots, \theta_{i-1}^{(t+1)}, \theta_i^{(t)}, \ldots, \theta_p^{(t)}).$$

2. Take

$$\theta_i^{(t+1)} = \begin{cases} \tilde{\theta} \text{ with probability } \alpha(\theta_i^{(t)}, \tilde{\theta}_i), \\ \tilde{\theta}_i^{(t)} \text{ with probability } 1 - \alpha(\theta_i^{(t)}, \tilde{\theta}_i), \end{cases}$$

where

$$\alpha = 1 \bigwedge \left\{ \frac{\left(\frac{f_i(\tilde{\theta}_i | \theta_1^{(t+1)}, \ldots, \theta_{i-1}^{(t+1)}, \theta_{i+1}^{(t)}, \ldots, \theta_p^{(t)})}{q_i(\tilde{\theta}_i | \theta_1^{(t+1)}, \ldots, \theta_{i-1}^{(t+1)}, \theta_i^{(t)}, \ldots, \theta_p^{(t)})} \right)}{\left(\frac{f_i(\theta_i^{(t)} | \theta_1^{(t+1)}, \ldots, \theta_{i-1}^{(t+1)}, \theta_{i+1}^{(t)}, \ldots, \theta_p^{(t)})}{q_i(\theta_i^{(t)} | \theta_1^{(t+1)}, \ldots, \theta_{i-1}^{(t+1)}, \tilde{\theta}_i, \ldots, \theta_p^{(t)})} \right)} \right\}$$

One can see that, essentially, a Metropolis "accept" step is added to the corresponding step in the Gibbs sampler.

Chapter 2

GRAPHICAL MODELS AND THE DISCRETE REGRESSION DISTRIBUTION

Graphical models are essentially probability models for multivariate observations that can be represented with a mathematical graph. The graph characterizes the conditional independence structure of the elements in the random vector. A graph is composed of vertices, with each vertex representing a component of a random vector. Edges are drawn between certain vertices depending on the conditional independence relationships.

Graphical models have become very popular due to their ability to easily portray independence structure in high dimensional random variables. These models were first introduced in the discrete variable case by Darroch et al. (1980), who were concerned with describing the complicated independence structure of a high dimensional contingency table.

Following the introduction of graphical models in the discrete variable case the study of these models rapidly expanded to other situations. Graphical models for multivariate Gaussian random variables were developed by Speed and Kiiveri (1986) using previous work by Dempster (1972) and Wermuth (1976). It was also noted that often researchers were interested in relationships which exist in one direction or are "causal" in some respect. Kiiveri et al. (1984) introduced the directed acyclic graphical model (DAG) for "causal" relationships as opposed to the "undirected" graph models (UG) studied previously. Finally, Frydenburg (1990) conducted a

detailed study of chain graph models, which combine DAG and UG models into one unified model.

In this chapter, we will examine graphical models in some detail. First, we will present the notation of graphical models. In many instances the notation is not part of a standard set of statistical notation and terminology. Graphical modeling combines notation from mathematical graph theory with that of probability and statistics. Therefore, some of the notation and terminology is often unknown in the statistics realm. We then focus on the *Markov properties* of graphical models. The Markov properties of a graphical model are the link between the probability distribution of a graphical model and a mathematical graph that describes the conditional independence relationships of the distribution. Therefore, we devote most of the chapter to describing various forms of Markov properties for different types of graphical models. The Markov properties of a graphical model provide the ability to describe the relationships within a multivariate distribution. Whittaker (1990) gives a thorough introduction to graphical modeling in an applied setting. If a more mathematically rigorous exposition is desired, Lauritzen (1996) gives a thorough description of graphical models. Much of the introductory material presented here follows these two sources.

Following a description of various Markov properties, we will give examples of some commonly used probability distributions for graphical models. We will also describe conditions on the parameters of the distributions which are needed for a given graph to describe the relationships among the variables being modeled. In the final section, we describe a new probability distribution for modeling a set of categorical variables using a set of continuous and categorical "covariate" variables. Hence, we term this new distribution the *discrete regression distribution.* We will provide necessary and sufficient conditions on the parameters of the discrete regression distribution to determine if a certain graph describes the relationships between and within the set of categorical response variables and the covariate variables.

2.1 Graph Theory Notation

Throughout the remaining chapters, we follow the notation given by Lauritzen (1996). A graph, $\mathcal{G} = (V, E)$, is a pair of sets, where V is a finite set of vertices and $E \subseteq V \times V$ is a set of edges. The graphs here are simple in that they contain no loops (an edge with the same beginning and end vertex) or multiple edges (all members of E are unique).

Let $(\alpha, \beta) \in E$. If $(\beta, \alpha) \in E$, the edge is referred to as *undirected.* If $(\beta, \alpha) \notin E$, the edge is *directed.*

In subsequent chapters, we will be dealing with a mixed set of vertices, some of which represent discrete variables and some of which are continuous. Therefore, we will partition the set V into two parts, Δ will contain vertices associated with discrete variables and Γ will contain the vertices associated with the continuous variables. Therefore, we have $V = \Delta \cup \Gamma$ and $\Delta \cap \Gamma = \emptyset$.

The usefulness of graphical models is due to the fact that a graph is a visual object and can be represented by a picture. Figure 2.1 provides an illustration of the concepts that follow. Vertices are represented by dots and edges are represented by lines between the dots. For $\beta \in V$ and $\gamma \in V$, an undirected edge is represented by a line between the dots associated with β and γ. For undirected edges, both (β, γ) and $(\gamma, \beta) \in E$. In text, an undirected edge is denoted by $\beta \sim \gamma$. If α is also an element of V and $(\alpha, \beta) \in E$ but $(\beta, \alpha) \notin E$, then, an arrow is drawn from α to β on the graph. In text this is denoted as $\alpha \rightarrow \beta$. If $(\alpha, \delta) \notin E$ and $(\delta, \alpha) \notin E$, for $\delta \in V$, then there is neither an arrow nor a line between α and δ. Absence of a line is denoted by $\alpha \nsim \delta$.

If a graph contains only undirected edges it is termed an *undirected graph,* conversely, if all edges are directed, the graph is termed a *directed graph.* There is no specific terminology for a graph that contains a mixture of edge types. If certain conditions are met (discussed in Section 2.3.1) a graph with a mixture of edge types

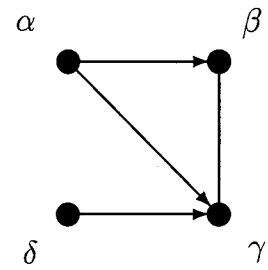

Figure 2.1: Example of a graph. Here, $V = (\alpha, \beta, \gamma, \delta)$ and $E = \{(\alpha,\beta), (\alpha,\gamma), (\delta,\gamma), (\beta,\gamma), (\gamma,\beta)\}$.

may be a *chain graph*. The undirected version, $\mathcal{G}^{\sim}$, of graph $\mathcal{G}$ is created from $\mathcal{G}$ by converting all arrows to lines.

Suppose $A \subseteq V$ is a subset of vertices. Then A induces the subgraph $\mathcal{G}_A = (A, E_A)$ from $\mathcal{G}$, where $E_A = E \cap (A \times A)$. All edges with starting and ending points in A remain in the new graph, every edge with the vertices in that are not in A, $V \setminus A$ are removed. Throughout this chapter, we use the set notation $A \setminus B$ to refer to the elements in A that are not in B. We will also use the notation $|A|$ to refer to the cardinality of the set A.

A graph is said to be *complete* if all vertices are joined by a line or arrow. So, the graph in Figure 2.1 is not complete. A subset of vertices is complete if it induces a complete subgraph. The term *maximally compete* refers to a subset A that is complete, but, if one more vertex $V \setminus A$ is added to the subset A, the subgraph loses completeness. A maximally complete subset is called a *clique*. In Figure 2.1, $\{\alpha,\beta,\gamma\}$ and $\{\delta,\gamma\}$ are the cliques of the graph.

If an arrow is present from α to β , then α is said to be a *parent* of β . Conversely, β is said to be a *child* of α . The set of parents of β is denoted by $pa(\beta)$ and the set of children of α as $ch(\alpha)$. If $\alpha \sim \beta$, then α and β are said to be *neighbors* and $ne(\alpha)$ represents the set of neighbors of α . For a subset $A \subseteq V$, $pa(A)$, $ch(A)$, and $ne(A)$ denote the vertices that are not themselves members of A, but are parents,

children, and neighbors of the vertices in A:

$$\begin{aligned} pa(A) &= \bigcup_{\alpha \in A} pa(\alpha) \backslash A \\ ch(A) &= \bigcup_{\alpha \in A} ch(\alpha) \backslash A \\ ne(A) &= \bigcup_{\alpha \in A} ne(\alpha) \backslash A. \end{aligned}$$

For example, in Figure 2.1, $pa(\{\beta, \gamma\}) = \{\alpha, \delta\}$, $ch(\alpha) = \{\beta, \gamma\}$, and $ne(\beta) = \{\gamma\}$.

The *boundary* of the subset of vertices, A, is the set of all neighbors and parents of A, $bd(A) = pa(A) \cup ne(A)$. The *closure* of A is given by $A \cup bd(A)$. In Figure 2.1, $bd(\beta) = \{\alpha, \gamma\}$ and $cl(\beta) = \{\beta, \gamma, \alpha\}$.

A *path* of length n from α to β is a sequence $\alpha = \alpha_0, \alpha_1, \ldots, \alpha_{n-1}, \alpha_n = \beta$. If such a sequence exists, it is said that α *leads* to β. For example in Figure 2.1, δ leads to β, but, δ does not lead to α because $(\beta, \alpha) \notin E$. A subset $S \subseteq V$ is said to be an $(\alpha, \beta) - separator$ if every path from α to β intersects S. Consequently, for subsets A, B, and S of V, S is said to *separate* A from B if it is an (α, β)-separator for each $\alpha \in A$ and $\beta \in B$. In Figure 2.1, $\{\gamma\}$ separates $\{\delta\}$ from $\{\alpha, \beta\}$.

2.2 Undirected Graphical Models

Here we give a description of the association between mathematical graphs and independence relationships between components of random vectors. We begin with a review of independence properties, termed *Markov properties*, in the case of an undirected graphical model. In following sections, we will extend these results to account for the asymmetric relationships in chain graph models.

Let $\mathbf{X} = (X_1, \ldots, X_v)$ denote a random vector and $V = (1, \ldots, v)$ denote the collection of vertices which are associated with the correspondingly labelled random vector component. Then the independence graph, $\mathcal{G}$, for the random vector is composed of the vertices in V. Also, all vertices are connected by an undirected edge except for those vertices, say α and β, for which

X_α is conditionally independent of X_β given the remaining variables $\mathbf{X}_{V\setminus\{\alpha,\beta\}} = (X_1, \ldots, X_{\alpha-1}, X_{\alpha+1}, \ldots, X_{\beta-1}, X_{\beta+1}, \ldots, X_v)$. We use the shorthand notation $\alpha \perp \beta \mid V \setminus \{\alpha, \beta\}$ to represent this independence statement.

2.2.1 Markov Properties

Here we describe the three Markov properties, pairwise, local, and global, for graphical model distributions. These properties are used to describe conditional independence relationships that can be deduced from a graph that is associated with a multivariate distribution. The graph provides an easy method for summarizing these relationships that may not be readily apparent from the distribution.

Before we begin the description of the Markov properties, we present some useful preliminary facts, given by Lauritzen (1996, pg 29), concerning conditional independence of random variables. Using the notation $X \perp Y \mid Z$ to mean X is conditionally independent of Y given Z and taking h to be any measurable function we have

$$(C1) \quad X \perp Y \mid Z \Rightarrow Y \perp X \mid Z$$

$$(C2) \quad X \perp Y \mid Z \text{ and } U = h(X) \Rightarrow U \perp Y \mid Z$$

$$(C3) \quad X \perp Y \mid Z \text{ and } U = h(X) \Rightarrow X \perp Y \mid (U, Z)$$

$$(C4) \quad X \perp Y \mid Z \text{ and } X \perp W \mid (Y, Z) \Rightarrow X \perp (W, Y) \mid Z$$

Another useful conditional independence property is

$$\text{(C5)} \quad X \perp Y \mid Z \text{ and } X \perp Z \mid Y \Rightarrow X \perp (Z, Y)$$

However, (C5) does not hold universally for all probability distributions. Lauritzen (1996, pg 29) provides a sufficient condition for (C5) to hold in the following proposition.

Proposition 2.1. *If the joint density of all variables is positive and continuous with respect to a product measure, then (C5) will hold.*

Another useful set of facts for determining conditional independence of the variables X and Y given a third variable Z, are based on the joint conditional density of X and Y, specifically, the factorization of the density (Lauritzen, 1996, pg 29). If we let f represent a generic probability density we obtain the following equivalence statements

$$X \perp Y \mid Z \Longleftrightarrow f(x,y,z) = f(x,z)f(y,z)/f(z)$$

$$X \perp Y \mid Z \Longleftrightarrow f(x|y,z) = f(x|z)$$

$$X \perp Y \mid Z \Longleftrightarrow f(x,z|y) = f(x|z)f(z|y)$$

$$X \perp Y \mid Z \Longleftrightarrow f(x,y,z) = h(x,z)k(y,z) \text{ for some } h,\ k$$

$$X \perp Y \mid Z \Longleftrightarrow f(x,y,z) = f(x|z)f(y,z)$$

These five equivalences hold apart from a set of triples (x,y,z) with probability zero. Factorization of the joint probability density will form the primary method for assessing conditional independence when constructing criteria for determining whether a graph describes the conditional independence structure of a multivariate probability distribution.

Now, we will examine the graphical Markov properties. If we let P be the distribution of X, then we can define the first Markov property for undirected graphical models, the pairwise property.

Undirected Pairwise Markov Property (UP) *For any $\alpha \in V$ and $\beta \in V$, if the distribution, P, of $\mathbf{X}$, satisfies $\alpha \perp \beta \mid V \setminus \{\alpha,\beta\}$ whenever the pair of vertices $\{\alpha,\beta\}$ is not complete in $\mathcal{G}$, then it is said that P possesses the pairwise Markov property with respect to $\mathcal{G}$.*

Figure 2.2 shows the undirected version of the previous graph example presented in Section 2.1. From Figure 2.2, we can deduce that if the P possesses property (UP) with respect to the graph, then, $\delta \perp \alpha \mid \{\beta,\gamma\}$.

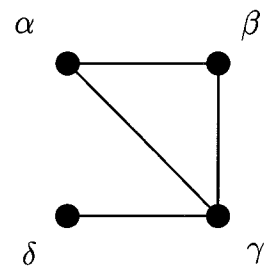

Figure 2.2: Example of an undirected independence graph. Here, $V = (\alpha, \beta, \gamma, \delta)$.

Upon examination of Figure 2.2 it would be tempting to conclude that $\delta \perp \{\alpha, \beta\} \mid \gamma$. In fact, if P possesses the local Markov property we can make that conclusion.

Undirected Local Markov Property (UL) *For any* $\alpha \in V$, *if* $\alpha \perp V \backslash cl(\alpha) \mid bd(\alpha)$, *then,* P *is said to possess the local Markov property with respect to* $\mathcal{G}$.

It would also seem logical, after examining Figure 2.2, that $\delta \perp \alpha \mid \gamma$ since any path between the two vertices intersects γ. If P possesses the final Markov property this will be a valid conclusion as well.

Undirected Global Markov Property (UG) *If* $A \perp B \mid S$ *for any disjoint sets,* A, B, *and* S, *of vertices in* $\mathcal{G}$, *such that* S *separates* A *from* B, *then* P *possesses the global Markov property with respect to* $\mathcal{G}$.

Remarkably, it can be shown that under certain conditions on the distribution P, over the sample space $\mathcal{X}$ of X, that the three Markov properties are equivalent. To show this to be true we begin with a proposition presented by Lauritzen (1996, pg 33).

Proposition 2.2. *For any undirected graph $\mathcal{G}$ and any probability distribution on $\mathcal{X}$, it holds that*

$$(UG) \Rightarrow (UL) \Rightarrow (UP)$$

The proof of Proposition 2.2 only involves properties (C1) - (C4), which implies that the proposition holds for all probability distributions P. Now in order to show equivalence of the Markov properties we need only show (UP) $\Rightarrow$ (UG). Pearl and Paz (1987) give a sufficient condition to prove the equivalence result. Suppose A, B, C, and D are sets of variables such that the following holds

$$\text{if } A \perp B \mid (C \cup D) \text{ and } A \perp C \mid (B \cup D) \text{ then } A \perp (B \cup C) \mid D, \tag{2.1}$$

then the following theorem provides the equality result.

Theorem 2.1 (Pearl and Paz). *If a probability distribution on $\mathcal{X}$ is such that (2.1) holds for disjoint sets A, B, C, D, then*

$$(UG) \Longleftrightarrow (UL) \Longleftrightarrow (UP)$$

Note that (2.1) is analogous to (C5) (Lauritzen, 1996, pg. 34). Therefore, referring to Proposition 2.1, the following corollary is readily observable.

Corollary 2.1. *If the joint probability distribution of all variables is positive and continuous with respect to a product measure, then*

$$(UG) \Longleftrightarrow (UL) \Longleftrightarrow (UP).$$

Corollary 2.1 is an extremely powerful method for checking Markov equivalence because is it often easy to determine if a model distribution is positive and continuous with respect to, say, Lebesgue or counting measure. If all three Markov properties are equivalent for a distribution P, it is said that P is Markov with respect to $\mathcal{G}$, or, P is $\mathcal{G}$ Markovian.

There is another very useful property for graphical models. The definition of the independence of two random variables is directly associated with the ability to factorize the joint probability density into the product of two separate densities, so, one can imagine that factorization must be associated with the Markov properties as well. In fact, we have the following additional property for graphical models,

Gibbs Factorization (F) *A probability measure, P, is said to have Gibbs-factorization according to graph $\mathcal{G}$ if for all complete subsets $a \subseteq V$ there exist non-negative functions ψ_a that depend on $\mathbf{x}$ though $\mathbf{x}_a = \{x_\gamma : \gamma \in a\}$ only, and there exists a product measure μ on $\mathcal{X}$ such that P has density f with respect to μ, where f has the form*

$$f(\mathbf{x}) = \prod_{a \text{ complete}} \psi_a(\mathbf{x}). \tag{2.2}$$

We offer the following equivalent condition,

Proposition 2.3. *A probability distribution P factorizes according to complete sets in $\mathcal{G}$ (2.2) if and only if P factorizes according to the set of cliques in $\mathcal{G}$ as*

$$f(\mathbf{x}) = \prod_{c \in \mathcal{C}} \psi_c(\mathbf{x}), \tag{2.3}$$

where $\mathcal{C}$ is the set of cliques of $\mathcal{G}$.

Proof. Suppose P factorizes according to complete sets a in $\mathcal{G}$, then, we can rewrite (2.2) as $f(\mathbf{x}) = \prod_{c \in \mathcal{C}} \psi_c(\mathbf{x})$ by letting $\psi_c(\mathbf{x}) = \prod_{a \subseteq c} \psi_a(\mathbf{x})$ for all $c \in \mathcal{C}$. Now, if P factorizes on the cliques of $\mathcal{G}$ then it factorizes according to complete sets since cliques are maximally complete. □

Now we examine the relationship between (F) and the other Markov properties. First, Lauritzen (1996, pg 35) offers the following proposition,

Proposition 2.4. *For an undirected graph $\mathcal{G}$ and any probability distribution on $\mathcal{X}$ it holds that*

$$(F) \Rightarrow (UG) \Rightarrow (UL) \Rightarrow (UP)$$

If P has a positive and continuous density, it only needs to be shown that (UP) $\Rightarrow$ (F) to deduce equivalence of all the presented properties.

The proof of graphical property equivalence is usually attributed to Hammersley and Clifford (1971), who first proved the result for discrete random variables. The equivalence results from the following theorem. We also give a sketch of the proof as we will be using a version of the proof in later sections.

Before giving the Hammersley-Clifford Theorem, however, we provide a lemma that is used in the proof of the theorem

Lemma 2.1 (Möbius inversion). *For two real valued functions, g and h, defined on the set of all subsets of a finite set V we have that the following two statements are equivalent*

(1) for all $a \subseteq V$: $g(a) = \sum_{b \subseteq a} h(b)$;

(2) for all $a \subseteq V$: $h(a) = \sum_{b \subseteq a} (-1)^{|a \setminus b|} g(b)$,

where $|a|$ represents the cardinality of the set a.

Now, we can present the Hammersley-Clifford Theorem, as well as a sketch of its proof.

Theorem 2.2 (Hammersley-Clifford). *A probability distribution P with positive and continuous density f with respect to a product measure μ satisfies the pairwise Markov property (UP) with respect to an undirected graph $\mathcal{G}$ if and only if it has Gibbs factorization (F) according to $\mathcal{G}$.*

Proof. Here we present a sketch of the proof given by Lauritzen (1996, pg. 36). First, using Proposition 2.4, (F) $\Rightarrow$ (UP). Now, assume that P is pairwise Markov with respect to $\mathcal{G}$. Since f is assumed positive, we will work with the log density. The definition of (F) (2.2) can be rewritten according to the log density as

$$\log f(\mathbf{x}) = \sum_{a \subseteq V} \phi_a(\mathbf{x}),$$

where $\phi_a(\mathbf{x}) = \log \psi_a(\mathbf{x})$ and $\phi_a(\mathbf{x}) \equiv 0$ unless a is complete is a compete subset of V.

First, assume that P possesses (UP) and choose a fixed but arbitrary element $\mathbf{x}^*$ of $\mathcal{X}$. For all $a \subseteq V$ define,

$$H_a(\mathbf{x}) = \log f(\mathbf{x}_a, \mathbf{x}^*_{a^c}),$$

where $(\mathbf{x}_a, \mathbf{x}^*_{a^c})$ is the element $\mathbf{y}$ of $\mathcal{X}$ such that $y_\gamma = x_\gamma$ for $\gamma \in a$ and $y_\gamma = x^*_\gamma$ for $\gamma \notin a$. Since $\mathbf{x}^*$ is fixed, $H_a(\mathbf{x})$ depends on $\mathbf{x}$ only through $\mathbf{x}_a$. Now, for all $a \subseteq V$, define the interaction term

$$\phi_a(\mathbf{x}) = \sum_{b \subseteq a} (-1)^{|a \setminus b|} H_b(\mathbf{x}), \tag{2.4}$$

where $|a|$ represents the cardinality of the set a. Next, using Möbius inversion (Lemma 2.1) the log density of P can be rewritten as

$$\log f(\mathbf{x}) = H_V(\mathbf{x}) = \sum_{a \subseteq V} \phi_a(\mathbf{x}),$$

It can be observed that $\phi_a(\mathbf{x})$ depends only on the components in $\mathbf{x}$ denoted by the subset a. Thus, it only needs to be shown that $\phi_a(\mathbf{x}) = 0$ whenever a is not a complete subset of V.

Choose $\alpha \in a$ and $\beta \in a$ such that $\alpha \not\sim \beta$. Now, letting $c = V \setminus \{\alpha, \beta\}$ and using the shorthand notation $H_a = H_a(\mathbf{x})$, we have

$$\phi_a(\mathbf{x}) = \sum_{b \subseteq c} (-1)^{|c \setminus b|} \left\{ H_b - H_{b \cup \{\alpha\}} - H_{b \cup \{\beta\}} + H_{b \cup \{\alpha, \beta\}} \right\}$$

Using the definition of $H_a(\mathbf{x})$ it can be shown that all of the terms in the curly brackets add to zero, hence $\phi_a(\mathbf{x}) = 0$ if there are members of a that are not complete in $\mathcal{G}$ and P possesses (UP). Therefore, (UP) $\Longrightarrow$ (F). □

2.2.2 Models for Discrete, Gaussian, and Mixed Variables

In this section, we describe some multivariate probability distributions that are often used for graphical modeling. We present these models with two goals in mind. First, the models provide an example of the Markov properties and Gibbs factorization. Secondly, these basic models provide a foundation from which we will build more complicated models for compositional data in the following chapters.

The first class of models we will examine are log-linear models. Log-linear models are used to model the joint distribution for a set of discrete, or categorical, variables. Next, we will examine the class of Gaussian graphical models, which are used to model the joint distribution of several continuous variables. Finally, the two models will be merged in the class of conditional Gaussian (CG) models. The class of CG models is used to model so called "mixed" variables. A set of mixed variables is a random vector which contains both continuous and discrete random variables.

Log-Linear Graphical Models

We begin with some notation for defining the discrete random vector and corresponding log-linear model. We will use Δ to represent a finite set of categorical variables, or *classification criteria.* For each $\delta \in \Delta$, $\mathcal{J}_\delta$ represents the possible levels, from 1 to the maximum number of levels, of the variable X_δ. The sample space for a discrete random vector $\mathbf{X}_\Delta = (X_1, \ldots, X_{|\Delta|})$ is given by $\mathcal{J} = \times_{\delta \in \Delta} \mathcal{J}_d$. Each of the cross-classifications $\mathbf{x}_\Delta \in \mathcal{J}$ is termed a *cell* in $\mathcal{J}$.

Now that we have defined the sample space for a random categorical vector $\mathbf{X}_\Delta$, we will define a log-linear model to represent the probability density of $\mathbf{X}_\Delta$. A log-linear model gives the probability that an individual is cross-classified to cell $\mathbf{x}_\Delta$

for a set Δ of categorical variables to be

$$f(\mathbf{x}_\Delta) = \exp\left\{\sum_{d\subseteq\Delta} \lambda_d(\mathbf{x}_\Delta)\right\}, \tag{2.5}$$

where $\lambda_\emptyset(\mathbf{x}_\Delta)$ represents the normalizing constant

$$\lambda_\emptyset(\mathbf{x}_\Delta) = -\log\left[\sum_{\mathbf{x}_\Delta} \exp\left\{\sum_{d\subseteq\Delta} \lambda_d(\mathbf{x}_\Delta)\right\}\right], \tag{2.6}$$

$\lambda_d(\mathbf{x}_\Delta)$ depends on x_Δ only through $\mathbf{x}_d = \{x_\delta : \delta \in d\}$, and $\lambda_d(\mathbf{x}_\Delta) = 0$ if $x_\delta = 1$ for any $\delta \in d$ (Whittaker, 1990). The last condition represents an identifiability condition for the model.

The parameters $\lambda_d(\mathbf{x}_\Delta)$ in (2.5) are named *interaction terms.* This is due to the fact that application of the calculations in the proof of the Hammersley-Clifford Theorem (Theorem 2.2) will show that for $\delta,\ \gamma \in \Delta$, $X_\delta \not\perp X_\gamma | \mathbf{X}_{\Delta\setminus\{\delta,\gamma\}}$ unless $\lambda_d(\mathbf{x}_\Delta) = 0$ for every d that contains the pair $\{\delta,\gamma\}$ and all cells $\mathbf{x}_\Delta$ (Whittaker, 1990, pg. 207). Following calculation of the Hammersley-Clifford interactions (2.4), one finds that $\phi_d(\mathbf{x}_\Delta) = \lambda_d(\mathbf{x}_\Delta)$, therefore, factorization cannot occur if $\lambda_d(\mathbf{x}_\Delta) \neq 0$ for all $\mathbf{x}_\Delta$ and d is not a complete subset of V. The term *interaction* is used because the joint conditional distribution of X_δ and X_γ given the remaining variables is not simply the product of the marginal conditional distributions. There is a synergistic effect that will produce probability mass above (for positive values of the interaction terms) or below (for negative values of the interaction terms) the product of the marginal distributions.

A *graphical model* for a given graph $\mathcal{G}$ is defined from (2.5) by adding the constraint that $\lambda_d(x) = 0$ if $\{\alpha,\beta\} \in d$ and $\alpha \not\sim \beta$ in $\mathcal{G}$ (Whittaker, 1990). In other words, the constraint is added that $\lambda_d(x) = 0$ unless d is complete with respect to $\mathcal{G}$. This constraint will produce a probability distribution for $\mathbf{X}_\Delta$ that is $\mathcal{G}$ Markovian. Therefore, the variables in $X_\delta,\ \delta \in \Delta$, will satisfy all three Markov properties (UP), (UL), (UG), and since the density is positive and continuous, (F), as well. One can

construct a graphical representation of a fitted model by defining a graph that has complete sets associated with the non-zero interaction terms.

It should be noted that the correspondence between a log-linear model and a graph is not one-to-one. There are many log-linear models that are Markov with respect to the same graph. In fact, for a given graph $\mathcal{G}$, it is possible to set all of the interaction terms $\lambda_d(\mathbf{x}_\Delta) = 0$ for $|d| > 2$ and obtain a distribution that is also Markov over $\mathcal{G}$ (Whittaker, 1990, pg. 208).

Log-linear models have long been used for modeling discrete random variables as well as testing hypotheses concerning conditional independence (Birch, 1963). Darroch et al. (1980) first noticed the connection between Markov random fields and log-linear models. This led to an improved method for interpreting complicated conditional structures. Many modern texts on log-linear models include a section on graphical model theory as a way to interpret large fitted models (Christensen, 1990).

Gaussian Graphical Models

We now look at defining graphical models for multivariate normal (MVN) random vectors. Speed and Kiiveri (1986) were the first to rigorously describe models for Gaussian variables in terms of independence graphs and graph theory notation. Modeling Gaussian vectors in terms of conditional independence statements, however, was not new. Dempster (1972) proposed a parsimonious model for the covariance matrix of a MVN distribution by setting certain off-diagonal elements of the inverse covariance matrix to zero. Since, an off-diagonal element of the inverse covariance matrix is, in fact, the covariance of the two associated variables given the rest, Dempster was modeling conditional independence. Therefore, Gaussian graphical models are often given Dempster's terminology *covariance selection* models. Following Dempster, Wermuth (1976) described the similarities between log-linear models and covariance selection models.

Graphical models for MVN variables are in general more straightforward than models for discrete variables. This is due to the fact that the joint distribution of a MVN vector is completely determined by the pairwise relationships. Since the elements of the inverse covariance matrix are conditional covariances and a zero covariance for a pair of Gaussian random variables implies independence, a graphical model can be based solely on the inverse covariance matrix T. Speed and Kiiveri (1986) prove that a MVN distribution, with variables indexed by the set of vertices in a graph $\mathcal{G}$, is Markov with respect to $\mathcal{G}$ if and only if

$$T(\alpha, \beta) = 0 \text{ if } (\alpha, \beta) \notin E \text{ and } \alpha \neq \beta, \tag{2.7}$$

where $T(\alpha, \beta)$ is the element of the inverse covariance matrix associated with variables α and β. A graph can be constructed for a specific MVN distribution by excluding edges if the corresponding element of T is zero.

Mixed Variable Graphical Models

Graphical models for random vectors containing both discrete and Gaussian elements were developed by Lauritzen and Wermuth (1989) using the *Conditional Gaussian Distribution* (CG). The goal is to create a coherent distribution that can describe covariation between discrete and continuous variables. The CG distribution is defined by first giving a probability density for the discrete components. Then, given the discrete components a MVN density is used to model the continuous components. Recall, that we can partition the set of vertices V into discrete and continuous components $V = \Gamma \cup \Delta$ (Section 2.1). Subsequently, we partition the random vector $\mathbf{X} = (\mathbf{X}_\Gamma, \mathbf{X}_\Delta)$ where $\mathbf{X}_\Gamma$ are the continuous variables in $\mathbf{X}$ and $\mathbf{X}_\Delta$ are the discrete variables. The vector $\mathbf{x}_\Gamma$ is a vector in the sample space $\mathbb{R}^{|\Gamma|}$ and $\mathbf{x}_\Delta$ is a cell in the sample space $\mathcal{J}$ (see previous section on log-linear graphical models). The CG density for a mixed vector $\mathbf{X}$ is given by

$$f(\mathbf{x}) = \exp\left\{ g(\mathbf{x}_\Delta) + \mathbf{h}(\mathbf{x}_\Delta)'\mathbf{x}_\Gamma - \frac{1}{2}\mathbf{x}_\Gamma'\mathbf{T}(\mathbf{x}_\Delta)\mathbf{x}_\Gamma \right\}, \tag{2.8}$$

where, for each $\mathbf{x}_\Delta \in \mathcal{J}$, g is a real number, $\mathbf{h}$ is a real valued vector in $\mathbb{R}^{|\Gamma|}$ and $\mathbf{T}$ is a real valued, positive definite, $|\Gamma| \times |\Gamma|$ matrix. It can be shown that $\mathbf{X}_\Gamma|\mathbf{x}_\Delta$ is distributed as a MVN random vector with mean $\mathbf{T}(\mathbf{x}_\Delta)^{-1}\mathbf{h}(\mathbf{x}_\Delta)$ and covariance matrix $\mathbf{T}(\mathbf{x}_\Delta)^{-1}$.

In order to determine Markov relationships, (2.8) can be re-parameterized using interaction parameters similar to the log-linear model (Lauritzen and Wermuth, 1989). In order to reparameterize the density, we construct interaction terms as in the proof of the Hammersley-Clifford Theorem (Theorem 2.2). First, let $\mathbf{x}^*_\Delta \in \mathcal{J}$ be an arbitrary but fixed cell and for every $d \subseteq \Delta$, define the interaction terms

$$\begin{aligned}
\lambda_d(\mathbf{x}_\Delta) &= \sum_{a \subseteq d} (-1)^{|d \setminus a|} g(\mathbf{x}_a, \mathbf{x}^*_{\Delta \setminus a}), \\
\boldsymbol{\eta}_d(\mathbf{x}_\Delta) &= \sum_{a \subseteq d} (-1)^{|d \setminus a|} \mathbf{h}(\mathbf{x}_a, \mathbf{x}^*_{\Delta \setminus a}), \\
\boldsymbol{\Psi}_d(\mathbf{x}_\Delta) &= \sum_{a \subseteq d} (-1)^{|d \setminus a|} \mathbf{T}(\mathbf{x}_a, \mathbf{x}^*_{\Delta \setminus a}),
\end{aligned}$$

where, $(\mathbf{x}_a, \mathbf{x}^*_{\Delta \setminus a})$ is the cell $\mathbf{y}_\Delta \in \mathcal{J}$ such that $y_\delta = x_\delta$ for $\delta \in a$ and $y_\delta = x^*_\delta$ for $\delta \notin a$. Now, using Möbius inversion (Lemma 2.1) we obtain

$$\begin{aligned}
g(\mathbf{x}_\Delta) &= \sum_{d \subseteq \Delta} \lambda_d(\mathbf{x}_\Delta), \\
\mathbf{h}(\mathbf{x}_\Delta) &= \sum_{d \subseteq \Delta} \boldsymbol{\eta}_d(\mathbf{x}_\Delta), \\
\mathbf{T}(\mathbf{x}_\Delta) &= \sum_{d \subseteq \Delta} \boldsymbol{\Psi}_d(\mathbf{x}_\Delta).
\end{aligned} \tag{2.9}$$

Using the re-parameterization of g, $\mathbf{h}$, and $\mathbf{T}$ in (2.9), we can now rewrite the CG density using interaction terms,

$$\begin{aligned}
f(\mathbf{x}) = \exp\Bigg\{ &\sum_{d \subseteq \Delta} \lambda_d(\mathbf{x}_\Delta) + \sum_{d \subseteq \Delta} \sum_{\gamma \in \Gamma} \eta_{d\gamma}(\mathbf{x}_\Delta) x_\gamma \\
&- \frac{1}{2} \sum_{d \subseteq \Delta} \sum_{\gamma, \mu \in \Gamma} \psi_{d\mu\gamma}(\mathbf{x}_\Delta) x_\gamma x_\mu \Bigg\},
\end{aligned} \tag{2.10}$$

where $\eta_{d\gamma}(\mathbf{x}_\Delta)$ is the $\gamma = 1, \ldots, |\Gamma|$ element of $\boldsymbol{\eta}_d(\mathbf{x}_\Delta)$, and $\psi_{d\gamma\mu}(\mathbf{x}_\Delta)$ is the (γ, μ) element of the matrix $\boldsymbol{\Psi}_d(\mathbf{x}_\Delta)$ for $\gamma,\ \mu = 1, \ldots, |\Gamma|$.

The interaction terms in (2.10) have a slightly more complicated interpretation than the interaction terms of either the log-linear or Gaussian graphical models due to the mixture of discrete and continuous types of variables. First, the $\lambda_d(\mathbf{x}_\Delta)$ interaction terms have the same interpretation as in the log-linear graphical models, since the marginal distribution of $\mathbf{X}_\Delta$ is a log-linear model (2.5) (Lauritzen, 1996, pg. 161). The $\eta_{d\gamma}(\mathbf{x}_\Delta)$ are *linear interactions* between the continuous variable X_γ and those discrete variables in d. The linear interaction terms function in the same manner as the effects parameters of an ANOVA model for the mean of X_γ. The $\psi_{d\gamma\mu}(\mathbf{x}_\Delta)$ parameters are termed *quadratic interactions.* In essence, the quadratic interactions are parameterizing an ANOVA-like model for the covariance matrix of $\mathbf{X}_\Gamma|\mathbf{x}_\Delta$.

Lauritzen and Wermuth (1989) give conditions for a CG distribution to be Markov with respect to a graph $\mathcal{G}$ in the following proposition which is based on Theorem 2.2.

Proposition 2.5. *A CG distribution is Markovian with respect to a graph $\mathcal{G}$ if and only if the interaction terms in (2.10) satisfy*

$$\begin{aligned} \lambda_d &\equiv 0 \text{ unless } d \text{ is complete in } \mathcal{G}, \\ \eta_{d\gamma} &\equiv 0 \text{ unless } d \cup \{\gamma\} \text{ is complete in } \mathcal{G}, \\ \psi_{d\gamma\mu} &\equiv 0 \text{ unless } d \cup \{\gamma, \mu\} \text{ is complete in } \mathcal{G}. \end{aligned}$$

The condition pertaining to $\lambda_d(\mathbf{x}_\Delta)$ is the same condition for a log-linear model and results from the fact that the marginal density of $\mathbf{X}_\Delta$ is a log-linear model. The condition placed on the $\eta_{d\gamma}(\mathbf{x}_\Delta)$ states that there can be no joint, or synergistic, effect on the mean of X_γ by all of the categorical variables $\{X_\delta : \delta \in d\}$ if the entire set $d \cup \{\gamma\}$ is not a complete subset in $\mathcal{G}$. Similarly, there can be no joint effect of the categorical variables $\{X_\delta : \delta \in d\}$ on the covariance between X_γ and X_μ if the entire set $d \cup \{\gamma, \mu\}$ is not a complete subset.

Proof. We give a proof of Proposition 2.5, presented by Lauritzen (1996), to illustrate the connection between Proposition 2.5 and the Hammersley-Clifford Theorem. One can observe the link by calculating the Hammersley-Clifford interaction terms (2.4) for the CG density. For $d \subseteq \Delta$, $c \subseteq \Gamma$, and γ, $\mu \in \Gamma$, with $\gamma \neq \mu$, these interaction terms are given by

$$\begin{aligned} \phi_d(\mathbf{x}) &= \lambda_d(\mathbf{x}_\Delta), \\ \phi_{d\cup\{\gamma\}}(\mathbf{x}) &= \eta_{d\gamma}(\mathbf{x}_\Delta)x_\gamma - \psi_{d\gamma\gamma}(\mathbf{x}_\Delta)x_\gamma^2/2, \\ \phi_{d\cup\{\gamma,\mu\}}(\mathbf{x}) &= -\psi_{d\gamma\mu}(\mathbf{x}_\Delta)x_\gamma x_\mu, \\ \phi_{d\cup c}(\mathbf{x}) &= 0 \text{ for } |c| > 2. \end{aligned} \tag{2.11}$$

Therefore, if the conditions of Proposition 2.5 are met then $\phi_{d\cup c}(\mathbf{x}) = 0$ for the set $d \cup c$ that are not complete subsets of V. If the CG distribution is $\mathcal{G}$ Markovian then the left-hand sides of each of the terms in (2.11) must be identically zero and the conclusions of the proposition follow. □

It should be noted that the CG distribution contains both the log-linear and Gaussian graphical models as special cases. Therefore, Proposition 2.5 generalizes the previous Markov conditions for both models. In addition, similar to the log-linear case, there is not a unique model for a given graph. Edwards (1990) has explored the class of *hierarchical models.* Hierarchical models are those in which interaction terms are constrained to zero if any other interaction term indexing a set of variables contained in the index of the first term is equal to zero. For example, in a hierarchical model $\eta_d(\mathbf{x}_\Delta) = 0$ if $\eta_{d^*}(\mathbf{x}_\Delta) = 0$ and $d^* \subset d$.

2.3 Chain Independence Graphs

We now extend the results in the previous section to *chain graph models.* Chain graph models allow one to construct a complex joint distribution though specification of conditional distributions. The chain graphs are based on an *a priori* ordering

of variables in terms of a known causal or influential structure. In essence, chain graph models describe a large joint distribution though a series of covariate / response models where the association between variables can be asymmetric.

Investigation of chain model properties began with the observation that the path models of Wright (1934) are very closely related to covariance selection models (Wermuth, 1980). Lauritzen and Wermuth (1989) then introduced the concept of a chain graph model and investigated some Markov properties in a graph theoretic setting. A more detailed account and application of chain models is given by Wermuth and Lauritzen (1990).

2.3.1 Chain Graph Notation and Requirements

We introduce some additional notation and requirements for chain graphs. Frydenburg (1990) was the first to describe Markov properties for chain graphs. Subsequently, most of the definitions and terminology in this section are attributable to Frydenburg (1990). Figure 2.3 gives examples for the concepts and definitions in this section. In a chain graph, there exists a known ordering of the vertices. If it is possible for α to influence or "cause" β , then α is said to precede β in this ordering. In an independence graph, an arrow may be present from α to β (using the previous notation of Section 2.1, $\alpha \rightarrow \beta$). Now, since there are both undirected and directed edges in a chain graph, there are also *undirected* and *directed paths.* When following a path from one vertex to the next, as with directed graphs, one must obey the direction of the arrow, if present. A *cycle* is a path for which the end vertex is also the starting vertex. If there is a directed path from α to β , then $\alpha \leq \beta$. If $\alpha \leq \beta$ and $\alpha \geq \beta$ then $\alpha \sim \beta$, or, there is an undirected edge between α and β. A requirement of chain graphs is that they have no directed cycles. The set of *descendants*, $de(\alpha)$, of a vertex $\alpha \in V$, is the set of vertices, β, such that there exists a path from β to α, but not from α to β. The non-descendants are

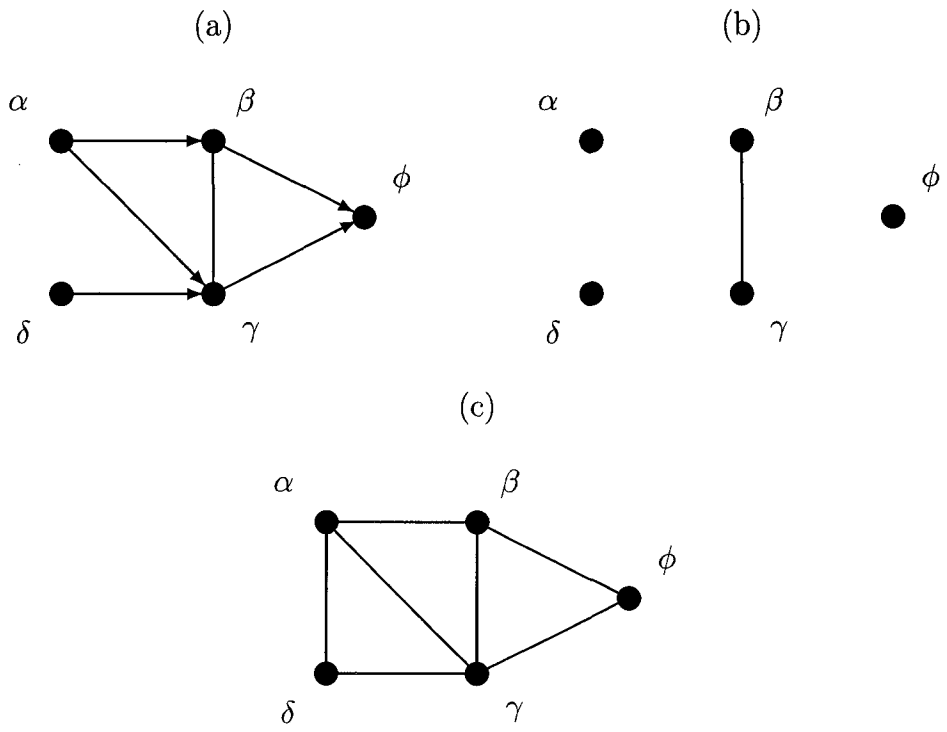

Figure 2.3: Here we present an illustration of chain graph concepts. In (a) we show a chain graph where $\{\alpha, \delta\}$ precedes $\{\beta, \gamma\}$ which precedes $\{\phi\}$ in a causal ordering, The chain components, $\{\alpha\}$, $\{\delta\}$, $\{\beta, \gamma\}$, and $\{\phi\}$, of (a) are illustrated in (b). The only terminal chain component is $\{\phi\}$. The moral graph of (a) is shown in (c).

$nd(\alpha) = V \setminus [de(\alpha) \cup \{\alpha\}]$. The non-descendants of a vertex are those vertices in the "causal" present and past.

Chain graph models are constructed by conditioning random variables based on previous elements of the chain graph. These previous random variables are designated in a chain graph by the term *chain components*. These components are defined by noting that $\sim$ induces an equivalence class $\mathcal{T}$ in V. For every $\tau \in \mathcal{T}$ and $\alpha, \beta \in \tau$ we have $\alpha \sim \beta$. The chain components can be seen in a chain graph by removing all of the directed edges. A chain component τ is called *terminal* if for every $\alpha \in \tau$, $ch(\alpha) = \emptyset$. A subset of vertices is called an *anterior set* if it can be generated by successive removal of terminal chain components.

Finally, we define the *moral graph*, $\mathcal{G}^m = (V, E^m)$, generated from $\mathcal{G}$. First, recall notation for the undirected version $\mathcal{G}^\sim$ of a graph $\mathcal{G} = (V, E)$. The undirected version $\mathcal{G}^\sim = (V, E^\sim)$, where $E^\sim$ is the same as E except that it is augmented so that all of the edges are undirected. The moral graph is generated from $\mathcal{G}$ by $E^m = E^\sim \bigcup_{\tau \in \mathcal{T}} E^*\{bd(\tau)\}$, where $E^*\{A\}$ is a complete collection of undirected edges for vertices in A. In other words, $\mathcal{G}^m$ is the undirected version of $\mathcal{G}$ with the addition that the parents of each chain component τ are made complete.

2.3.2 Markov Properties

With the previous collection of tools, we are now able to describe the Markov properties of chain graphs. Frydenburg (1990) gives extensions to the definitions of the undirected Markov properties for use with chain graph models.

Chain Markov Properties *Let $\mathcal{G}$ be a chain graph that indexes a set of variables with probability measure P. The probability measure P is then said to be*

(CP) *Pairwise Markovian with respect to $\mathcal{G}$ if, for any pair (α, β) of non-adjacent vertices with $\beta \in nd(\alpha)$*

$$\alpha \perp \beta \mid nd(\alpha) \setminus \{\beta\};$$

(L) *Local Markovian with respect to $\mathcal{G}$ if if for any vertex $\alpha \in V$*

$$\alpha \perp nd(\alpha) \mid bd(\alpha);$$

(G) *Global Markovian with respect to $\mathcal{G}$ if for any triple (A, B, S) of disjoint subsets of V such that S separates A from B in $(\mathcal{G}_{an(A \cup B \cup S)})^m$, the moral graph of the smallest ancestral set containing $A \cup B \cup S$, we have*

$$A \perp B \mid S.$$

If P possesses all three of the properties, it is said that P is Markovian with respect to $\mathcal{G}$. These properties are in fact, generalizations of the undirected Markov properties in that if $\mathcal{G}$ is an undirected graph, then the chain Markov properties are equivalent to the undirected Markov properties that were given in Section 2.2.1 (Frydenburg, 1990).

Now, as in the case of undirected graphical models, we would like to determine how the Markov properties relate to one another. Frydenburg (1990) provides a theorem which shows that the three chain Markov properties are equivalent provided that P possesses a positive and continuous density, which implies that the density has Gibbs factorization.

Theorem 2.3. *For any distribution P which has a positive and continuous density $f(\mathbf{x})$ with respect to a product measure, the following four statements are equivalent for any chain graph $\mathcal{G}$:*

(1) P is Markovian

(2) $f(\mathbf{x}) = \prod_{\tau \in \mathcal{T}} f_{\tau|bd(\tau)}(\mathbf{x}_\tau|\mathbf{x}_{bd(\tau)})$ *where $P_{cl(\tau)}$ is Markovian w.r.t $(\mathcal{G}_{cl(\tau)})^m$ for all $\tau \in \mathcal{T}$*

(3) f can be factorized

$$f(\mathbf{x}) = \prod_{\tau \in \mathcal{T}} \prod_{C \in \mathcal{C}_\tau} \psi_{C,\tau}(\mathbf{x}_{C,\tau})$$

such that

$$\int_{\mathbf{x}_\tau} \prod_{C \in \mathcal{C}_\tau} \psi_{C,\tau}(\mathbf{x}_{C,\tau}) \mu_\tau(d\mathbf{x}_\tau) \equiv 1$$

for all $\tau \in \mathcal{T}$ where $\mathcal{T}$ denotes the set of chain components in $\mathcal{G}$ and $\mathcal{C}_\tau$ denotes the collection of cliques in $(\mathcal{G}_{cl(\tau)})^m$

(4) If A is an anterior set then f_A has a Gibbs factorization w.r.t. $(\mathcal{G}_A)^m$.

We derive a 5th condition, which is not mentioned by Frydenburg (1990), but is also equivalent to the statements in Theorem 2.3

Proposition 2.6. *Statements (1) - (4) in Theorem 2.3 are equivalent to the statement*

(5) $f(\mathbf{x})$ *can be factorized*

$$f(\mathbf{x}) = \prod_{\tau \in \mathcal{T}} \prod_{B \in \mathcal{B}_\tau} \psi_{B,\tau}(\mathbf{x}_{B,\tau}) \tag{2.12}$$

such that

$$\int_{\mathbf{X}_\tau} \prod_{B \in \mathcal{B}_\tau} \psi_{B,\tau}(\mathbf{x}_{B,\tau}) \mu_\tau(d\mathbf{x}_\tau) \equiv 1$$

for all $\tau \in \mathcal{T}$ *where* $\mathcal{T}$ *denotes the set of chain components in* $\mathcal{G}$ *and* $\mathcal{B}_\tau$ *denotes a collection of complete subsets in* $(\mathcal{G}_{cl(\tau)})^m$.

Proof. Suppose that f factorizes as (2.12), then one can rewrite $f(\mathbf{x})$ as in Theorem 2.3(3) by defining

$$\psi_{C,\tau}(\mathbf{x}_{C,\tau}) = \prod_{B \subseteq C} \psi_{B,\tau}(\mathbf{x}_{B,\tau}).$$

Now, suppose f factorizes according to Theorem 2.3(3), then, it factorizes according to (2.12), since, $\mathcal{C}_\tau$ is a collection of complete subsets in $(\mathcal{G}_{cl(\tau)})^m$. □

Proposition 2.6 essentially weakens statement (3) of Theorem 2.3. The density f for probability distribution P does not have to factorize according to the cliques of a graph in order to be Markovian, it is enough for f to factorize according to complete sets which are not necessarily maximally complete. The condition in Proposition 2.6 is often easier to check than Theorem 2.3(3) when trying to ascertain whether or not a distribution is Markovian with respect to a given graph because there is no need to find the maximally complete subsets. In complicated graphs, the cliques may be hard to determine.

2.3.3 Comments on Chain Graph Markov Properties

We would now like to make some comments concerning chain graphs and chain graph Markov properties. The first comment is that two seemingly different chain graphs can have the same moral graph (Frydenburg, 1990; Andersson et al., 1997), which, implies they have the same Markov properties. Similar equivalencies were also noted by Cox and Wermuth (1993). This can be an issue of concern for selecting a graphical model from data, since there may be essentially duplicate models in the group of models from which a researcher would like to choose. These duplicates arise if there is uncertainty about the causality order.

The second comment concerns *separation criteria.* When determining conditional relationships from a large chain graph, it may be extremely hard to construct the appropriate moral graphs necessary for determining conditional independencies of the associated random variables. Therefore, several researchers have devised separation criteria so that these relationships can be directly read from the chain graph. Studený and Bouckaert (1998) and Levitz et al. (2001) both provide separation criteria which are equivalent to Theorem 2.3.

2.4 Discrete Regression Models

In this section, we propose a new chain model distribution for random vectors containing both continuous and discrete components. We title this distribution the *discrete regression distribution* (DR) due to its similarity to the conditional Gaussian (CG) regression distribution of Lauritzen and Wermuth (1989). The DR distribution is constructed by assuming that there exists a set $\Gamma \cup \Delta$ of continuous and discrete *predictor* variables which follow a CG distribution. A set of discrete *response* variables Φ is then distributed according to a log-linear model based on the predictor variables. Our desire to model the variables in Φ as a response to the variables in $\Gamma \cup \Delta$ precludes the use of a CG distribution for the entire set of variables.

The problem with the CG distribution is that it is not closed under conditioning (Lauritzen and Wermuth, 1989). Therefore, one would need to restrict the types of graphs used so that the Markov properties of the CG distribution for $V = \Phi \cup \Gamma \cup \Delta$ will match those of a chain graph with Φ as the terminal component. The discrete regression distribution eliminates the need for this restriction by building a model based on the desired conditioning.

2.4.1 Model Formulation

The full joint distribution of the predictor variables $\mathbf{X} = (\mathbf{X}_\Gamma, \mathbf{X}_\Delta)$ and discrete response variables $\mathbf{Y}_\Phi$ is given as the product density $f(\mathbf{y}_\Phi, \mathbf{x}) = f(\mathbf{y}_\Phi|\mathbf{x})f(\mathbf{x})$. More specifically, we begin by considering the conditional density of $\mathbf{Y}_\Phi|\mathbf{x}$ as a log-linear model. First, without worrying about constraining the cell probabilities to the interval [0, 1] or the sum of the probabilities to 1, we model the log probability for each cell $\mathbf{y}_\Phi$ of the response set with the linear model

$$l(\mathbf{y}_\Phi|\mathbf{x}) = \sum_{c \subseteq \Gamma} g_c(\mathbf{y}_\Phi|\mathbf{x}_\Delta) \prod_{\gamma \in c} x_\gamma + \sum_{m=2}^{M} \mathbf{h}_m(\mathbf{y}_\Phi|\mathbf{x}_\Delta)' \mathbf{x}_\Gamma^m, \tag{2.13}$$

where for every $\mathbf{y}_\Phi$ and $\mathbf{x}_\Delta$, g_c, $c \subseteq \Gamma$ is a real number, $\mathbf{h}_m$, $m = 1, \ldots, M$, is a vector in $\mathbb{R}^{|\Gamma|}$, and $\mathbf{x}_\Gamma^m = (x_1^m, \ldots, x_{|\Gamma|}^m)$. The set notation may appear unusual at first, but, $l(\mathbf{y}_\Phi|\mathbf{x})$ has the same structural formulation as a regression model that includes continuous and categorical covariates. The set notation provides a straightforward method for describing a general regression model with all levels of interaction between continuous and categorical variables. In addition, the model also includes polynomial terms, up to some finite power M, of the continuous variables. Here, we have described a log-linear model for each response cell conditioned on a realized covariate cell and set of continuous variables.

Now, if we exponentiate $l(\mathbf{y}_\Phi|\mathbf{x})$, normalize the response cell probabilities, and assume the marginal predictor density has the CG form, we obtain the DR joint

density

$$\begin{aligned} f(\mathbf{y}_\Phi, \mathbf{x}) =& f(\mathbf{y}_\Phi|\mathbf{x}) f_{CG}(\mathbf{x}) \\ =& \exp\left\{ \alpha_\Phi(\mathbf{x}) + \sum_{c\subseteq\Gamma} g_c(\mathbf{y}_\Phi|\mathbf{x}_\Delta) \prod_{\gamma\in C} x_\gamma + \sum_{m=2}^{M} \mathbf{h}_m(\mathbf{y}_\Phi|\mathbf{x}_\Delta)' \mathbf{x}_\Gamma^m \right\} \\ & \times \exp\left\{ g(\mathbf{x}_\Delta) + \mathbf{h}(\mathbf{x}_\Delta)'\mathbf{x}_\Gamma - \frac{1}{2}\mathbf{x}_\Gamma' \mathbf{T}(\mathbf{x}_\Delta)\mathbf{x}_\Gamma \right\} \end{aligned} \tag{2.14}$$

where $\alpha_\Phi(\mathbf{x})$ is a normalizing constant with respect to the response cells $\mathbf{y}_\Phi$, for all covariate cells $\mathbf{x}_\Delta$, g is a real number, h is a real vector in $\mathbb{R}^{|\Gamma|}$, and $\mathbf{T}$ is a real, positive definite matrix. The functions $g(\mathbf{x}_\Delta)$, $\mathbf{h}(\mathbf{x}_\Delta)$, $g_c(\mathbf{y}_\Phi|\mathbf{x}_\Delta)$, and $\mathbf{h}_m(\mathbf{y}_\Phi|\mathbf{x}_\Delta)$ are independent from one another for all $\mathbf{y}_\Phi$, $\mathbf{x}_\Delta$, $c \subseteq \Gamma$, and $m = 1, \ldots, M$ in the sense that they are functionally unrelated to each other.

Now, in the same manner as (2.10), we will reparameterize the density in terms of interaction effects. As in the proof of the Hammersley-Clifford Theorem 2.2, we will define interactions terms relative to an arbitrary but fixed value $(\mathbf{y}_\Phi^*, \mathbf{x}^*) = (\mathbf{y}_\Phi^*, \mathbf{0}_\Gamma, \mathbf{x}_\Delta^*)$ where $\mathbf{0}_\Gamma$ is a $|\Gamma|$ vector of zeros (see Lauritzen, 1996, pg 173). For $f \subseteq \Phi$, $c \subseteq \Gamma$, and $d \subseteq \Delta$, define the interactions

$$\beta_{fcd}(\mathbf{y}_\Phi|\mathbf{x}_\Delta) = \sum_{a\subseteq d} \sum_{e\subseteq f} (-1)^{|d\setminus a|+|f\setminus e|} g_c(\mathbf{y}_e, \mathbf{y}^*_{\Phi\setminus e}|\mathbf{x}_a, \mathbf{x}^*_{\Delta\setminus a}) \tag{2.15}$$

and

$$\boldsymbol{\omega}_{fdm}(\mathbf{y}_\Phi|\mathbf{x}_\Delta) = \sum_{a\subseteq d} \sum_{e\subseteq f} (-1)^{|d\setminus a|+|f\setminus e|} \mathbf{h}_m(\mathbf{y}_e, \mathbf{y}^*_{\Phi\setminus e}|\mathbf{x}_a, \mathbf{x}^*_{\Delta\setminus a}) \tag{2.16}$$

The following lemma shows that the classic identifiability constraint for interaction effects in ANOVA or log-linear models is satisfied by (2.15) and (2.16).

Lemma 2.2. *The interaction terms $\beta_{fcd}(\mathbf{y}_\Phi|\mathbf{x}_\Delta)$ and $\boldsymbol{\omega}_{fdm}(\mathbf{y}_\Phi|\mathbf{x}_\Delta)$ defined by (2.15) and (2.16), respectively, satisfy the two identifiability constraints*

(1) $\beta_{fcd}(\mathbf{y}_\Phi|\mathbf{x}_\Delta) = 0$ if $y_\phi = y_\phi^*$ or $x_\delta = x_\delta^*$ for any $\phi \in f$ or $\delta \in d$;

(2) $\boldsymbol{\omega}_{fdm}(\mathbf{y}_\Phi|\mathbf{x}_\Delta) = \mathbf{0}$ if $y_\phi = y_\phi^*$ or $x_\delta = x_\delta^*$ for any $\phi \in f$ or $\delta \in d$.

Proof. In order to prove the proposition, we only need calculate the interaction terms under the assumption that either $y_\phi = y_\phi^*$ for some $\phi \in f \subseteq \Phi$, or $x_\delta = x_\delta^*$ for some $\delta \in d \subseteq \Delta$. Therefore, first assume $y_\phi = y_{\phi^*}$ for some $\phi \in f \subseteq \Phi$. Then, the first interaction term in the proposition is

$$\begin{aligned}\beta_{fcd}(\mathbf{y}_\Phi|\mathbf{x}_\Delta) &= \sum_{a\subseteq d}\sum_{e\subseteq f}(-1)^{|d\setminus a|+|f\setminus e|} g_c(\mathbf{y}_e, \mathbf{y}^*_{\Phi\setminus e}|\mathbf{x}_a, \mathbf{x}^*_{\Delta\setminus a}) \\ &= \sum_{a\subseteq d}(-1)^{|d\setminus a|}\sum_{e\subseteq f\setminus\{\phi\}}(-1)^{|f\setminus\{e\cup\phi\}|}\{g_c(\mathbf{y}_e, y^*_\phi, \mathbf{y}^*_{\Phi\setminus\{e\cup\phi\}}|\mathbf{x}_a, \mathbf{x}^*_{\Delta\setminus a}) \\ &\qquad - g_c(\mathbf{y}_e, y^*_\phi, \mathbf{y}^*_{\Phi\setminus\{e\cup\phi\}}|\mathbf{x}_a, \mathbf{x}^*_{\Delta\setminus a})\} \\ &= 0.\end{aligned}$$

A completely analogous calculation holds by assuming $x_\delta = x_\delta^*$ for some $\delta \in d \subseteq \Delta$. Then one can repeat the calculations, replacing g_c with $\mathbf{h}_m$ to prove statement (2) of the lemma. □

Using Möbius inversion (Lemma 2.1), we can rewrite (2.14) as a function of interaction terms. The DR density can now be rewritten using (2.15) and (2.16) and reparameterizing the CG density for the predictor variables as in (2.10) to give,

$$\begin{aligned}f(\mathbf{y}_\Phi, \mathbf{x}) = \exp&\left[\alpha_\Phi(\mathbf{x}) + \sum_{f\subseteq\Phi}\sum_{c\subseteq\Gamma}\sum_{d\subseteq\Delta}\left\{\beta_{fcd}(\mathbf{y}_\Phi|\mathbf{x}_\Delta)\prod_{\gamma\in c}x_\gamma\right\}\right. \\ &\left. + \sum_{f\subseteq\Phi}\sum_{\gamma\in\Gamma}\sum_{d\subseteq\Delta}\sum_{m=2}^{M}\omega_{f\gamma dm}(\mathbf{y}_\Phi|\mathbf{x}_\Delta)x_\gamma^m\right] \\ &\times\exp\left\{\sum_{d\subseteq\Delta}\lambda_d(\mathbf{x}_\Delta) + \sum_{d\subseteq\Delta}\sum_{\gamma\in\Gamma}\eta_{d\gamma}(\mathbf{x}_\Delta)x_\gamma\right. \\ &\left. -\frac{1}{2}\sum_{d\subseteq\Delta}\sum_{\gamma,\mu\in\Gamma}\psi_{d\mu\gamma}(\mathbf{x}_\Delta)x_\gamma x_\mu\right\},\end{aligned} \tag{2.17}$$

where $\omega_{f\gamma dm}(\mathbf{y}_\Phi|\mathbf{x}_\Delta)$ is the γ element of the vector $\boldsymbol{\omega}_{fdm}(\mathbf{y}_\Phi|\mathbf{x}_\Delta)$ and the CG reparameterization is described by (2.10). Without loss of generality we can assume that $\beta_{fcd}(\mathbf{y}_\Phi|\mathbf{x}_\Delta) = \omega_{f\gamma dm}(\mathbf{y}_\Phi|\mathbf{x}_\Delta) = 0$ if $f = \emptyset$. Any interaction term that is not a function of $\mathbf{y}_\Phi$ will cancel with the normalizing function $\alpha_\Phi(\mathbf{x})$.

The DR distribution intersects the class of CG regression distributions given by Lauritzen and Wermuth (1989). That is to say, if we restrict the DR distribution to contain only quadratic power terms as well as only first and second order interaction terms for the continuous variables, we obtain a CG regression for a purely discrete response. While these restrictions include a wide range of useful models, we find the restriction unnecessarily confining. So, we propose the DR model as a more flexible model for purely discrete response variables.

2.4.2 Markov Properties of the DR Distribution

In order to make inference concerning the conditional independence of variables in discrete regressions, we need to determine the Markov properties of the DR chain model. So, we give the following proposition.

Proposition 2.7. *A DR distribution is Markovian with respect to a chain graph $\mathcal{G}$, with terminal chain component Δ and initial component $\Gamma \cup \Phi$, if and only if the interaction terms in (2.17) satisfy*

$$\begin{aligned} \beta_{fcd}(\mathbf{y}_\Phi|\mathbf{x}_\Delta) &\equiv 0 \text{ unless } f \cup c \cup d \text{ is complete in } \mathcal{G}, \\ \omega_{f\gamma dm}(\mathbf{y}_\Phi|\mathbf{x}_\Delta) &\equiv 0 \text{ unless } f \cup \{\gamma\} \cup d \text{ is complete in } \mathcal{G}, \end{aligned}$$

and

$$\begin{aligned} \lambda_d(\mathbf{x}_\Delta) &\equiv 0 \text{ unless } d \text{ is complete in } \mathcal{G}, \\ \eta_{d\gamma}(\mathbf{x}_\Delta) &\equiv 0 \text{ unless } d \cup \{\gamma\} \text{ is complete in } \mathcal{G}, \\ \psi_{d\mu\gamma}(\mathbf{x}_\Delta) &\equiv 0 \text{ unless } d \cup \{\mu, \gamma\} \text{ is complete in } \mathcal{G}. \end{aligned}$$

Proof. In order to prove Proposition 2.7 we will give a specialized version of the proof for the Hammersley-Clifford Theorem (Theorem 2.2) for the factorization of each of the chain components so that Theorem 2.3 (3) is satisfied.

We need only be concerned with the terminal chain component Φ. Proposition 2.5 shows that the conditions concerning the interaction terms of the CG density are necessary and sufficient for the initial chain component to factorize according to $\mathcal{G}_{\Gamma\cup\Delta} = (\mathcal{G}_{\Gamma\cup\Delta})^m$.

Suppose that the interaction terms $\beta_{fcd}(\mathbf{y}_\Phi|\mathbf{x}_\Delta)$ and $\omega_{f\gamma dm}(\mathbf{y}_\Phi|\mathbf{x}_\Delta)$, $m = 1,\ldots,M$ are equal to zero for all subsets $f\cup c\cup d$ and $f\cup\{\gamma\}\cup d$ that are not complete. Then, it is easy to observe that $f(\mathbf{y}_\Phi|\mathbf{x})$ factorizes according to complete sets in $\mathcal{G}^m_{cl(\Phi)}$, since $f(\mathbf{y}_\Phi|\mathbf{x})$ is a function only of complete factors in $\mathcal{G}^m_{cl(\Phi)}$. Since the density $f(\mathbf{y}_\Phi|\mathbf{x})$ factorizes according to complete sets in $\mathcal{G}^m_{cl(\phi)}$, it factorizes according to the cliques by Proposition 2.6.

Now, suppose that the DR distribution is Markovian with respect to $\mathcal{G}$. Then, as in the proof of the Hammersley-Clifford Theorem, for $f\subseteq\Phi$, $c\subseteq\Gamma$, and $d\subseteq\Delta$, the interaction term

$$\begin{aligned}\phi_{f\cup c\cup d}(\mathbf{y}_\Phi,\mathbf{x}) &= \sum_{e\subseteq f}\sum_{b\subseteq c}\sum_{a\subseteq d}(-1)^{|e\setminus f|+|c\setminus b|+|d\setminus a|}\log f(\mathbf{y}_e,\mathbf{y}^*_{\Phi\setminus e},\mathbf{x}_b,\mathbf{0}^*_{\Gamma\setminus b},\mathbf{x}_a,x^*_{\Delta\setminus a})\\ &= 0\end{aligned} \tag{2.18}$$

if the DR distribution is Markov and $d\cup c\cup f$ is not complete in $\mathcal{G}$. Therefore, we only need to show that $\phi_{f\cup c\cup d}(\mathbf{y}_\Phi,\mathbf{x})\equiv 0\Rightarrow\beta_{fcd}(\mathbf{y}_\Phi|\mathbf{x}_\Delta)=\omega_{f\gamma dm}(\mathbf{y}_\Phi|\mathbf{x}_\Delta)=0$ for $m=1,\ldots,M$.

Through use of the Möbius inversion theorem (Lemma 2.1) and Lemma 2.2, calculation of the Hammersley-Clifford interaction terms are as follows for $d \neq \emptyset$,

$$
\begin{aligned}
&\phi_{f\cup c\cup d}(\mathbf{y}_\Phi, \mathbf{x}) \\
&\quad = \sum_{e\subseteq f}\sum_{b\subseteq c}\sum_{a\subseteq d}(-1)^{|e\setminus f|+|c\setminus b|+|d\setminus a|}\log f(\mathbf{y}_e, \mathbf{y}^*_{\Phi\setminus e}, \mathbf{x}_b, \mathbf{0}^*_{\Gamma\setminus b}, \mathbf{x}_a, x^*_{\Delta\setminus a}) \\
&\quad = \sum_{b\subseteq c}(-1)^{|c\setminus b|}\sum_{e\subseteq f}\sum_{a\subseteq d}(-1)^{|d\setminus a|+|f\setminus e|}\left[\sum_{a\subseteq d}\sum_{b\subseteq c}\sum_{e\subseteq f}\beta_{eba}(\mathbf{y}_\Phi|\mathbf{x}_\Delta)\prod_{\gamma\in b}x_\gamma\right. \\
&\qquad\qquad \left. + \sum_{a\subseteq d}\sum_{e\subseteq f}\sum_{\gamma\in b}\sum_{j=2}^{m}\omega_{e\gamma am}(\mathbf{y}_\Phi|\mathbf{x}_\Delta)x_\gamma^m\right] \\
&\quad = \sum_{b\subseteq c}(-1)^{|c\setminus b|}\left[\beta_{dbf}(\mathbf{y}_\Phi|\mathbf{x}_\Delta)\prod_{\gamma\in b}x_\gamma + \sum_{\gamma\in b}\sum_{m=2}^{M}\omega_{f\gamma dm}(\mathbf{y}_\Phi|\mathbf{x}_\Delta)x_\gamma^j\right] \qquad (2.19) \\
&\quad = \beta_{fcd}(\mathbf{y}_\Phi|\mathbf{x}_\Delta)\prod_{\gamma\in c}x_\gamma + \sum_{\gamma\in c}\sum_{m=2}^{M}\omega_{f\gamma dm}(\mathbf{y}_\Phi|\mathbf{x}_\Delta)x_\gamma\sum_{b\subseteq c}(-1)^{|c\setminus b|}1_{[\gamma\in b]} \\
&\quad = \beta_{fcd}(\mathbf{y}_\Phi|\mathbf{x}_\Delta)\prod_{\gamma\in c}x_\gamma + \sum_{\gamma\in c}\sum_{m=2}^{M}\omega_{f\gamma dm}(\mathbf{y}_\Phi|\mathbf{x}_\Delta)x_\gamma\sum_{b\subseteq c\setminus\gamma}(-1)^{|[c\setminus\gamma]\setminus b|} \\
&\quad = \beta_{fcd}(\mathbf{y}_\Phi|\mathbf{x}_\Delta)\prod_{\gamma\in c}x_\gamma + \sum_{j=2}^{m}\omega_{f\gamma dm}(\mathbf{y}_\Phi|\mathbf{x}_\Delta)x_\gamma 1_{[c=\{\gamma\}]}
\end{aligned}
$$

If $c = \{\gamma\}$ then the interaction term $\phi_{d\cup c\cup f}(\mathbf{y}_\Phi, \mathbf{x})$ is a polynomial of order M in x_γ with coefficients $\omega_{f\gamma d2}(\mathbf{y}_\Phi|\mathbf{x}_\Delta), \ldots, \omega_{f\gamma dM}(\mathbf{y}_\Phi|\mathbf{x}_\Delta)$, and $\beta_{fcd}(\mathbf{y}_\Phi|\mathbf{x}_\Delta)$. Therefore, we have that $\phi_{d\cup c\cup f}(\mathbf{y}_\Phi, \mathbf{x}) = 0$ for any $\mathbf{x}$ implies that $\beta_{fcd}(\mathbf{y}_\Phi|\mathbf{x}_\Delta)$ and $\omega_{f\gamma dm}(\mathbf{y}_\Phi|\mathbf{x}_\Delta)$, $m = 1, \ldots, M$, must be zero. If $|c| \geq 2$, then we have the single interaction $\beta_{fcd}(\mathbf{y}_\Phi|\mathbf{x}_\Delta)\prod_{\gamma\in c}x_\gamma$ which equals zero if $\phi_{d\cup c\cup f}(\mathbf{y}_\Phi, \mathbf{x}) = 0$. Therefore, we have proven the proposition. □

Chapter 3

BAYESIAN ANALYSIS OF DISCRETE COMPOSITIONAL DATA: A GRAPHICAL MODEL APPROACH

3.1 Introduction

This chapter introduces a graphical model approach for analyzing discrete compositional data. Discrete compositional data arise from multivariate counts instead of a continuous multivariate vector. Therefore, with discrete compositional data, one is interested in the proportion of counts in a particular category of possible outcomes, or equivalently, the probability that a randomly selected individual will belong to a certain category. Here, we present models for this probability that allow for estimates of the association between the event that a new individual is placed in a given category and a set of explanatory variables of interest.

Analysis of compositional data arising from counts illustrates the main shortcoming of the LN model. As can be clearly seen from (1.2), the LN model is not defined for observations that are located on the boundary of the simplex ∇^d. In other words, the LN model is only defined if all compositional elements are positive. When analyzing counts from several categories, there is usually a non-trivial probability of observing zero individuals in a particular category.

To correct for the fact that the LN model is undefined if zero values are present in the composition, Aitchison (1986) proposes a mixture model. The mixture consists of a LN model for the entire composition mixed with a LN distribution for categories that are known to have only positive elements and a distribution degenerate at zero for categories where one might observe a zero value. This mixture

model, however, is an *ad hoc* solution to the problem. It adds unnecessary complication if there are many categories where one might observe a zero value.

In order to improve on the mixture model solution, Billheimer and Guttorp (1997) introduced a state-space model for the analysis of discrete compositional data. The model proposed by Billheimer and Guttorp has the following hierarchical structure,

$$\begin{aligned}(c_{i1},\ldots,c_{iD}) &\sim \text{multinomial}(N_i, p_{i1},\ldots,p_{iD}) \qquad i=1,\ldots,S \\ (p_{i1},\ldots,p_{iD}) &\sim \text{LN}(\boldsymbol{\mu}_i, \boldsymbol{\Sigma}),\end{aligned} \tag{3.1}$$

where c_{ij} is the observed number of individuals in category $j = 1,\ldots,D$ at site $i = 1,\ldots,S$, N_i is the total number of individuals counted at site $i = 1,\ldots,S$, and p_{ij} represents the true proportion of individuals in category j of the ith site. The location parameter $\boldsymbol{\mu}_i$ is often modeled to include covariate information as $\mu_{ij} = \beta_0 + \beta_j(x_i - \bar{x})$, $j = 1,\ldots,D$. The covariate is centered in order to improve the convergence of the MCMC methods that Billheimer and Guttorp use for inference. Model (3.1) allows the possibility of observing zero individuals in category j for the ith observation if it is believed that the *true* proportion of individuals in the ith population that belong to category j is strictly positive. This hierarchical model has been specialized for observations in a lattice spatial setting (Billheimer and Guttorp, 1997), in a time series setting (Billheimer, 1995), and in an experimental setting (Billheimer et al., 2001). In each of these models, an MCMC procedure is used to estimate $(p_{i1},\ldots,p_{iD})$ and a composition representation of $\boldsymbol{\mu}_i$ (see Section 1.2.2) for every observation. The scale parameter $\boldsymbol{\Sigma}$ is often regarded as a nuisance parameter and is included in the MCMC procedure, but usually ignored in estimation.

Billheimer and Guttorp make use of the LN distribution to model true compositions of random counts. To obtain discrete compositional data, a sample is drawn from each randomly selected site and individuals are categorized. So, in order to model the random site selection, a LN distributed composition is used to represent

the probabilities of an individual being placed into 1 of D categories for a random site. For the remainder of this dissertation, we will use the term "site" to represent a sample of individuals that is used to construct a single composition observation. The term "site" is used because discrete compositional data are often constructed by sampling a group of individuals at a geographic location. Even though we use the term site, however, these sampled populations could be indexed by other means such as time or experimental design levels.

In the case of a single sample from a fixed population, graphical models have been used for many years to model dependencies between categorical and continuous variables (Lauritzen and Wermuth, 1989). Graphical models were, however, designed for a single sample. This single sample would be equivalent to one compositional observation, or "site".

When analyzing discrete compositional data using many covariates, it would be beneficial to represent graphically the dependence between the event that a sampled individual is placed in a certain category and selected covariates. Therefore, in this chapter, we extend the class of discrete regression graphical chain models of Section 2.4 to allow multiple site sampling and hence construct graphical models for the analysis of compositional data. This is accomplished though the use of random effects. In addition, construction of a graphical model for compositional data extends the interpretation of covariate effects in discrete compositional models. With a graphical model, we can model the joint distribution of a composition variable and a set of covariates, therefore, the pairwise, local, and global Markov properties can be used to describe complex conditional dependence relationships between the composition variable and covariates. By applying a graphical model approach to the analysis of compositional data, we merge the two areas of statistical theory together.

In order to illustrate the graphical model approach for analysis of discrete compositional data, we use a multiple site graphical model to analyze data concerning

stream invertebrate functional groups. Stream invertebrates are often categorized as to how they function in the environment. In this case, we will examine invertebrate feeding groups. Feeding group classification is a major factor in determining how individual stream invertebrates interact with their environment (Cummins and Klug, 1979). Poff (1997) notes that understanding how environmental factors influence compositions of biological communities across the landscape is one of the major challenges facing ecologists today. In addition, determining how environmental factors at different spatial scales affect biological compositions is of major importance. In this chapter, we examine composition of stream invertebrate feeding groups and its relation to several environmental covariates measured at a local stream scale and at the watershed scale.

3.2 Model Formulation

3.2.1 Single Individual and Single Site Models

To begin the multiple site graphical model formulation, we first consider a model for a single individual at a randomly selected site. For each site, we are interested in the dependence relationships between covariates that will be measured at the site and the event that a randomly selected individual is placed into one of D categories. For example, in the analysis of stream invertebrate functional groups, we are interested in the event that a randomly selected invertebrate at a site is placed into one of six feeding types. Let Y denote a categorical variable that is associated with this event and is defined on the integers $1, \ldots, D$. Let y represent a realization of Y, and take an integer value from $\{1, \ldots, D\}$. In addition, let $\mathbf{X} = (X_1, \ldots, X_p)$ denote a vector of observed covariates at the randomly selected site. In terms of the vertices of a chain graph, we have one variable in the terminal chain component $\Phi = \{Y\}$ and p variables in the explanatory or initial component $\Gamma \cup \Delta$, where Γ represents the set of continuous covariates and Δ represents the set of categorical

covariates. We will use the notation $\mathbf{x}_\Gamma$ to represent a realization of the vector $\mathbf{X}_\Gamma$, and similar to Y, we will use $\mathbf{x}_\Delta$ to represent a realization of the random categorical vector $\mathbf{X}_\Delta$. The vector $\mathbf{x}_\Delta$ will contain an entry for each variable in Δ and each entry takes an integer value from one to the maximum number of levels for the associated categorical variable.

We can now use the discrete regression (DR) model of Section 2.4 to specify a joint density for $(Y, \mathbf{X})$ as $f(y|\mathbf{x})f_{CG}(\mathbf{x})$, where

$$f(y|\mathbf{x}) = \exp\left\{\alpha_\Phi(\mathbf{x}) + \sum_{c\subseteq\Gamma}\sum_{d\subseteq\Delta}\beta_{cd}(y,\mathbf{x}_\Delta)\prod_{\gamma\in c}x_\gamma + \sum_{\gamma\in\Gamma}\sum_{d\subseteq\Delta}\sum_{m=2}^{M}\omega_{\gamma dm}(y,\mathbf{x}_\Delta)x_\gamma^m\right\}, \qquad y = 1,\ldots,D, \tag{3.2}$$

and

$$\begin{aligned} f_{CG}(\mathbf{x}) &= f(\mathbf{x}_\Gamma|\mathbf{x}_\Delta)f(\mathbf{x}_\Delta) \\ &= \exp\left[\sum_{d\subseteq\Delta}\left\{\lambda_d(\mathbf{x}_\Delta) + \boldsymbol{\eta}_d(\mathbf{x}_\Delta)'\mathbf{x}_\Gamma - \frac{1}{2}\mathbf{x}_\Gamma'\boldsymbol{\Psi}_d(\mathbf{x}_\Delta)\mathbf{x}_\Gamma\right\}\right]. \end{aligned} \tag{3.3}$$

In (3.2), $\alpha_\Phi(\mathbf{x})$ is a normalizing constant with respect to $Y|\mathbf{X} = \mathbf{x}$ and $\beta_{cd}(y,\mathbf{x}_\Delta)$ and $\omega_{\gamma dm}(y,\mathbf{x}_\Delta)$ are interaction terms which depend on $\mathbf{x}_\Delta$ only through the variables contained in the set $d \subseteq \Delta$. The CG density (3.3) is given in the matrix form in order to facilitate re-parameterization in our proposed estimation procedure, as described in Section 3.4. As with the interaction terms in the response model, $\lambda_d(\mathbf{x}_\Delta)$, $\boldsymbol{\eta}_d(\mathbf{x}_\Delta)$, and $\boldsymbol{\Psi}_d(\mathbf{x}_\Delta)$ depend on $\mathbf{x}_\Delta$ only through the subset of variables in d.

To complete the DR model we must impose some constraints to ensure identifiability of the model parameters. To accomplish this, first select a reference category of Y, say y^*, and a reference cell for the categorical covariates, say $\mathbf{x}_\Delta^*$. Without loss of generality, henceforth, we assume that $y^* = 1$ and $\mathbf{x}_\Delta^*$ is an appropriately sized vector of ones, indicating the reference cell is that which is indexed by the first level of all the variables associated with Δ. Now that the reference category and cell are defined, set all interaction terms in (3.2) and (3.3) equal to zero if $y = 1$ or

$x_\delta = 1$ for any $\delta \in d$. These zero constraints are analogous to the zero constraints of interaction terms in classic ANOVA models. By using these constraints we can interpret the interaction terms as measuring interactions relative to the selected values y^* and $\mathbf{x}^*_\Delta$. For example, given any two covariates $\gamma \in \Gamma$, and $\delta \in \Delta$ and $y \neq 1$, a positive value for the interaction term $\beta_{\gamma\delta}(y, \mathbf{x}_\Delta)x_\gamma$ implies that an increase in x_γ increases the probability that a randomly selected individual will be classified according to level y over the first level of Y, and the amount of increase depends on the categorical covariate X_δ.

Instead of using a CG distribution for the marginal distribution of the covariates one could use an "iterated discrete regression" (IDR) model. The CG distribution (3.3) is based on the premise that, marginally, the categorical covariates follow a log-linear model, while conditioned on the categorical variables, the continuous variables follow a Multivariate Normal (MVN) distribution with mean and variance determined by the realized categorical variables. Philosophically, in terms of a graphical chain model, this implies that the categorical covariates precede the continuous ones in a "causal ordering". If it is theoretically more plausible to model the continuous variables as "causing" or influencing the categorical variables, then one could reverse the order of conditioning by modeling the joint distribution of the covariates as a DR distribution where the categorical covariates take the "response" role and the continuous covariates have a MVN distribution. This leads to the "iterated discrete regression" model. So, for the IDR model we have

$$\begin{aligned} f_{IDR}(\mathbf{x}) &= f(\mathbf{x}_\Delta|\mathbf{x}_\Gamma) f_N(\mathbf{x}_\Gamma) \\ &= \exp\left\{ \alpha_\Phi(\mathbf{x}_\Gamma) + \sum_{c\subseteq\Gamma}\sum_{d\subseteq\Delta} \zeta_{cd}(\mathbf{x}_\Delta) \prod_{\gamma\in c} x_\gamma \right. \\ &\qquad \left. + \sum_{\gamma\in\Gamma}\sum_{d\subseteq\Delta}\sum_{m=2}^{M} \nu_{\gamma dm}(\mathbf{x}_\Delta) x_\gamma^m \right\} \\ &\quad \times f_N(\mathbf{x}_\Gamma;\ \boldsymbol{\mu}_\Gamma,\ \boldsymbol{\Sigma}_\Gamma), \end{aligned} \tag{3.4}$$

where $f_N(\cdot; \boldsymbol{\mu}, \boldsymbol{\Sigma})$ represents a MVN density with mean $\boldsymbol{\mu}$ and covariance $\boldsymbol{\Sigma}$. Interaction terms are set to zero in the same fashion as the DR model to ensure an identifiable model. If neither model, CG or IDR, makes sense in terms of causal ordering, one can construct, in a similar manner, models for which there are some discrete and some continuous variables in each chain component by using a CG model for each explanatory chain component.

Equations (3.2) and (3.3) describe a graphical chain model for sampling a single individual at a randomly selected site. Now, we need to extend this model to account for repeated sampling of individuals at one site. We are still interested in inference for the interaction parameters of the individual model, (3.2) and (3.3), in order to construct a chain graph to depict dependence relationships, however, sampling multiple individuals is necessary to estimate the interaction terms. Therefore, we will now need to construct a likelihood model for sampling N individuals at a site if (3.2) and (3.3) represent the joint density of an individual classified to category y and the covariates observed at the site where the individual is observed.

To construct a model for repeated sampling at a single site, we will first condition on the realization of a site, which amounts to conditioning on the covariates. Now, sampling N individuals at a site provides N realizations of the variable Y. These N realizations can be summarized into a D vector of counts $\mathbf{c} = \{c(y) : y = 1, \ldots, D\}$, where $c(y)$ represents the number of individuals that were cross-classified into category y. The count vector $\mathbf{c}$ represents a complete and sufficient summarization of the N individual responses, so we can model the counts in order to make inference to (3.2). We model the count vector $\mathbf{c}$ with a multinomial distribution. For a fixed site sample size N, the joint distribution for the counts and covariates is

$$\begin{aligned} f(\mathbf{c}, \mathbf{x}) &= f_M(\mathbf{c}|\mathbf{x}) f_{CG}(\mathbf{x}) \\ &= \frac{N!}{\prod_{y=1}^{D} c(y)!} \left\{ \prod_{y=1}^{D} f(y|\mathbf{x})^{c(y)} \right\} \times f_{CG}(\mathbf{x}), \end{aligned} \tag{3.5}$$

where $f(y|\mathbf{x})$ is given by (3.2). If, however, sampling of individuals is carried out in a way where N is random and could be modeled with a Poisson distribution with mean κ, then each category count will have independent Poisson distributions with mean $\kappa f(y|\mathbf{x})$, $y = 1, \ldots, D$, (Rohatgi, 1976, pg. 200). For a Poisson(κ) random sample size, the joint density of $\mathbf{c}$ and $\mathbf{x}$ is given by

$$\begin{aligned} f(\mathbf{c}, \mathbf{x}) &= f_P(\mathbf{c}|\mathbf{x}) f_{CG}(\mathbf{x}) \\ &= \left[\prod_{y=1}^{D} \frac{\{\kappa f(y|\mathbf{x})\}^{c(y)} e^{\kappa f(y|\mathbf{x})}}{c(y)!} \right] \times f_{CG}(\mathbf{x}). \end{aligned} \tag{3.6}$$

If it is desired, the IDR model could be used for the covariate distribution in (3.6) instead of the CG model. The joint models in (3.5) and (3.6) for counts and covariates are different than the standard sampling scheme for a mixed variable graphical model. Usually, every individual sampled generates a multivariate observation of categorical and continuous variables. Here, however, there is only one observation of the covariate vector for all of the individuals observed at a particular site. The present sampling scheme is analogous to replication of an experiment at the same factor levels at each site.

The multinomial model is quite self-explanatory in terms of what we would like to model. We are interested in modeling the true composition of individuals at a site. This composition is represented by the D probabilities $f(y|\mathbf{x})$. In the Poisson model, however, another parameter, κ, has been added. This parameter is of no interest as far as determining conditional dependence relationships, but, it has to be estimated well for each site in order to have a model that fits the data adequately. Therefore, we propose two parsimonious modifications to the Poisson model that will prove useful when many sites are sampled.

First, the total number of individuals observed at a site may depend on the same covariates which are being used in the graphical model. So, we can model the expected number of individuals observed at a site, κ, using the linear model

$$\log \kappa = \sum_{c \subseteq \Gamma} \sum_{d \subseteq \Delta} \beta_{cd}(\mathbf{x}_\Delta) \prod_{\gamma \in c} x_\Gamma + \sum_{\gamma \in \Gamma} \sum_{d \subseteq \Delta} \sum_{m=2}^{M} \omega_{\gamma dm}(\mathbf{x}_\Delta) x_\gamma^m. \tag{3.7}$$

This produces a log-mean model for the category y count given by

$$\log\{\kappa f(y|\mathbf{x})\} = \sum_{f\subseteq\Phi}\sum_{c\subseteq\Gamma}\sum_{d\subseteq\Delta}\beta_{fcd}(y,\mathbf{x}_\Delta)\prod_{\gamma\in c}x_\gamma + \sum_{f\subseteq\Phi}\sum_{\gamma\in\Gamma}\sum_{d\subseteq\Delta}\sum_{m=2}^{M}\omega_{a\gamma dm}(y,\mathbf{x}_\Delta)x_\gamma^m, \tag{3.8}$$

where $\Phi = \{Y\}$. Once again, the interaction terms depend on $\mathbf{x}_\Delta$ only through the subset of variables in d. In addition, the interaction terms also depend on the variable Y through the index f. If $f = \emptyset$ then the interaction term does not depend on the category. In order to ensure identifiability, interaction terms are again set to zero if $y = 1$ or $x_\delta = 1$ for any δ in d. The normalization constant $\alpha_\Phi(\mathbf{x})$ was dropped because it is independent of the response categories and can therefore be absorbed into the parameters for the κ model. A similar approach is used for contingency table log-linear models when a random sample size is assumed (Christensen, 1990).

Another approach to modeling the κ parameter is to use covariates related to the sampling protocol. For example, covariates such as sampling effort at each site may be used to model the total site counts. These types of variables are usually not scientifically interesting as far as modeling compositions; however, they might prove valuable for modeling the total number of individuals observed at a site. This type of Poisson model is not used in graphical log-linear modeling due to the fact that there is only one sample of individuals . Compositional data will have multiple samples that might be modeled using external sample design covariates.

3.2.2 Random Effects Discrete Regression

In Section 3.2.1 we describe a model for a single randomly sampled site. Now, we will extend this model to account for possibly hundreds of randomly selected sites. For each site, a separate graphical model could be constructed, but this would increase the number of parameters to be estimated to an unmanageable level. In addition, the differences in non-zero parameter values are not of primary interest

in determining Markov relationships for a graphical model. Therefore, we propose a global graphical model for all sites that allows site-to-site flexibility in some of the non-zero parameter values. With the added flexibility, the model can adjust to fit the data more adequately. In order to add this flexibility, as well as model the randomness in site selection, we introduce a random error term to the response model (3.2).

The addition of a random effect to the response model (3.2) produces a full model for Y, $\mathbf{X}$, and the random effects $\boldsymbol{\epsilon}$ of the form

$$f(y, \mathbf{x}, \boldsymbol{\epsilon}) = f_{RE}(y|\mathbf{x}, \boldsymbol{\epsilon}) f_{CG}(\mathbf{x}) f(\boldsymbol{\epsilon}). \tag{3.9}$$

Since there is only one observation of the explanatory variables we will leave the model for the covariates, $f_{CG}(\mathbf{x})$, as it is given in (3.3). The response portion $f_{RE}(y|\mathbf{x}, \boldsymbol{\epsilon})$ of the random effects discrete regression (REDR) model is modified by the addition of a random intercept term to give,

$$\begin{aligned} f_{RE}(y|\mathbf{x}, \boldsymbol{\epsilon}) = \exp \Bigg\{ & \alpha_{\Phi}(\mathbf{x}, \boldsymbol{\epsilon}) + \sum_{c \subseteq \Gamma} \sum_{d \subseteq \Delta} \beta_{cd}(y, \mathbf{x}_{\Delta}) \prod_{\gamma \in c} x_{\gamma} \\ & + \sum_{\gamma \in \Gamma} \sum_{d \subseteq \Delta} \sum_{m=2}^{M} \omega_{\gamma dm}(y, \mathbf{x}_{\Delta}) x_{\gamma}^{m} + \epsilon(y) \Bigg\}, \end{aligned} \tag{3.10}$$

for $y = 1, \ldots, D$, where $\epsilon(y) = 0$ if $y = 1$ to ensure identifiability. The remaining random "interaction" terms, $\boldsymbol{\epsilon} = \{\epsilon(y) : y \neq 1\}$, are given a multivariate distribution with mean $\mathbf{0}$ and covariance (or scale parameter) $\boldsymbol{\Sigma}$.

The introduction of a random error term into the response model (3.2) has two benefits. First, the model can adjust for site-to-site variability. Secondly, the model will automatically add some level of overdispersion to category counts. In the model description we have left the error distribution vague. We believe that different situations may necessitate different error structures. If it is reasonable to assume that the error structure is symmetric with few outliers, then a MVN

distribution may be reasonable. In this case, the cell compositions will have a LN distribution, however, other distributions could be used. For example a multivariate t distribution with k degrees of freedom could be used if it is desirable to have an error with heavier tails. If that is the case, the category compositions will have what we term a Logistic-t (LT) distribution. The general form of a LT(k, $\boldsymbol{\mu}$, $\boldsymbol{\Sigma}$) distribution for a composition vector $\mathbf{p} = (p_1, \ldots, p_D)$ is given by

$$f_{LT}(\mathbf{p}) = \frac{\Gamma(\frac{1}{2}(k+D-1))}{\Gamma(\frac{1}{2}k)(k\pi)^{(D-1)/2}}|\boldsymbol{\Sigma}|^{-1/2}\prod_{j=1}^{D} p_j^{-1} \times \left[1 + \frac{1}{k}\{h(\mathbf{p}) - \boldsymbol{\mu}\}'\boldsymbol{\Sigma}^{-1}\{h(\mathbf{p}) - \boldsymbol{\mu}\}\right]^{-(k+D-1)/2}, \tag{3.11}$$

where $h(\mathbf{p}) = (\log(p_1/p_d), \ldots, \log(p_{D-1}/p_D))'$ and $k \geq 2$. This distribution is constructed in a manner completely analogous to the LN construction in Chapter 1. It may be desirable to use the LT distribution instead of the LN if there is a high level of overdispersion in the category counts. For the remaining discussion of the random effects DR model we will assume a MVN distribution for the random effects; however, the theoretical results will remain the same for the LT model.

Now that we have added random effects to the response portion of the model, the likelihood for the response variable cell counts given the covariates $\mathbf{x}$ changes slightly from (3.5) and (3.6). The likelihood model for the response variable cell counts $\mathbf{c}$ given the covariates $\mathbf{x}$, the total number of individuals observed at a site, and the random effects is

$$f_M(\mathbf{c}|\mathbf{x}, \boldsymbol{\epsilon}) = \frac{N!}{\prod_{y=1}^{D} c(y)!}\prod_{y=1}^{D} f_{RE}(y|\mathbf{x}, \boldsymbol{\epsilon})^{c(y)}, \tag{3.12}$$

where $f_{RE}(y|\mathbf{x}, \boldsymbol{\epsilon})$ is given by (3.10). Similarly, for a random total number of individuals, the Poisson model including random effects simply replaces the DR response $f(y|\mathbf{x})$ with the REDR response model $f_{RE}(y|\mathbf{x}, \boldsymbol{\epsilon})$ in (3.6).

Now, we focus on the multiple site likelihood for the explanatory variables. Assuming that the covariate observations are independently distributed and follow

a homogeneous CG distribution, we obtain the multiple site explanatory density

$$f(\mathbf{x}_1, \ldots, \mathbf{x}_S) = \prod_{i=1}^{S} f_{CG}(\mathbf{x}_i | \boldsymbol{\lambda}, \boldsymbol{\eta}, \boldsymbol{\Psi}_\emptyset), \tag{3.13}$$

where $\boldsymbol{\lambda}$ and $\boldsymbol{\eta}$ represent the collected parameter sets $\{\lambda_d(\mathbf{x}_\Delta) : d \subseteq \Delta\}$ and $\{\boldsymbol{\eta}_d(\mathbf{x}_\Delta) : d \subseteq \Delta\}$. Extensions to this model, such as accounting for correlation of covariates over space, are discussed in Section 6.2.2.

We now re-parameterize the homogeneous CG density in (3.13) into a more useful form. First, we break the CG density into a marginal model for the categorical components of the explanatory variable set and a conditional model for the continuous components. We then re-parameterize the conditional Gaussian distribution into an ANOVA like form. This re-parameterization gives the following form for the homogeneous CG density,

$$\begin{aligned} f_{CG}(\mathbf{x}) &= f(\mathbf{x}_\Delta) f(\mathbf{x}_\Gamma | \mathbf{x}_\Delta) \\ &= \exp\left\{ \sum_{d \subseteq \Delta} \lambda_d(\mathbf{x}_\Delta) \right\} \times \frac{1}{\sqrt{2\pi}} |\boldsymbol{\Psi}_\emptyset|^{1/2} \\ &\quad \times \exp\left\{ \frac{1}{2} \left(\mathbf{x}_\Gamma - \sum_{d \subseteq \Delta} \boldsymbol{\tau}_d(\mathbf{x}_\Delta) \right)' \boldsymbol{\Psi}_\emptyset \left(\mathbf{x}_\Gamma - \sum_{d \subseteq \Delta} \boldsymbol{\tau}_d(\mathbf{x}_\Delta) \right) \right\}, \end{aligned} \tag{3.14}$$

where $\boldsymbol{\Psi}_\emptyset$ represents the inverse covariance matrix for the continuous variables, which have a MVN distribution, $\boldsymbol{\tau}_d(\mathbf{x}_\Delta) = \boldsymbol{\Psi}_\emptyset^{-1} \boldsymbol{\eta}_d(\mathbf{x}_\Delta)$, and $\lambda_\emptyset(\mathbf{x}_\Delta)$ represents a normalizing constant in the log-linear model for $\mathbf{x}_\Delta$.

Define the vector of cell counts $\mathbf{c}_\Delta = [c(\mathbf{x}_\Delta)]$, where $c(\mathbf{x}_\Delta)$ is the number of sites for which the categorical covariates $\mathbf{X}_\Delta = \mathbf{x}_\Delta$. Using the re-parameterization of the CG density in (3.14) we can write the joint density of the covariates over all sites (3.13) as

$$\begin{aligned} f(\mathbf{x}_1, \ldots, \mathbf{x}_S) &= \left\{ \prod_{i=1}^{S} f(\mathbf{x}_{\Delta i}) \right\} \times \left\{ \prod_{i=1}^{S} f_N(\mathbf{x}_{\Gamma i} | \mathbf{x}_{\Delta i}) \right\} \\ &= \left(\frac{S!}{\prod_{\mathbf{x}_\Delta} c(\mathbf{x}_\Delta)!} \right)^{-1} f_M(\mathbf{c}_\Delta | \boldsymbol{\lambda}) \\ &\quad \times \prod_{i=1}^{S} f_N\left(\mathbf{x}_{\Gamma i};\ \sum_{d \subseteq \Delta} \tau_d(\mathbf{x}_{\Delta i}),\ \boldsymbol{\Psi}_\emptyset \right), \end{aligned} \tag{3.15}$$

where the explanatory observation at the ith site is given by $\mathbf{x}_i = (\mathbf{x}_{\Delta i}, \mathbf{x}_{\Gamma i})$ and $f_M(\mathbf{c}_\Delta|\boldsymbol{\lambda})$ is the multinomial density

$$f_M(\mathbf{c}_\Delta|\boldsymbol{\lambda}) = \frac{S!}{\prod_{\mathbf{x}_\Delta} c(\mathbf{x}_\Delta)!} \prod_{\mathbf{x}_\Delta} \exp\left\{\sum_{d\subseteq\Delta} \lambda_d(\mathbf{x}_\Delta)\right\}^{c(\mathbf{x}_\Delta)}. \tag{3.16}$$

The full likelihood for parameter estimation in the REDR model (3.9) is obtained by combining the likelihood for response variable cell counts at each site (3.12), the random effects density, and the explanatory variable likelihood (3.13).

$$\begin{aligned} f(\{\mathbf{c}\},\{\mathbf{x}\},\{\boldsymbol{\epsilon}\}) &= \prod_{i=1}^{S} f_M(\mathbf{c}_i|\mathbf{x}_i,\boldsymbol{\epsilon}_i) f_{CG}(\mathbf{x}_i) f_N(\boldsymbol{\epsilon}_i) \\ &= \prod_{i=1}^{S} f_M(\mathbf{c}_i|\mathbf{x}_i) \times K(\mathbf{c}_\Delta) f_M(\mathbf{c}_\Delta|\boldsymbol{\lambda}) \times \prod_{i=1}^{S} f_N\left(\mathbf{x}_{\Gamma i};\ \sum_{d\subseteq\Delta} \tau_d(\mathbf{x}_{\Delta i}),\ \boldsymbol{\Psi}_\emptyset\right) \\ &\quad \times \prod_{i=1}^{S} f_N\left(\boldsymbol{\epsilon}_i;\ \mathbf{0},\boldsymbol{\Sigma}\right), \end{aligned} \tag{3.17}$$

where $\{\mathbf{c}\} = \{\mathbf{c}_i : i = 1,\ldots,S\}$, $\mathbf{c}_i$ is a D vector of response variable cell counts for site i, $\{\mathbf{x}\} = \{\mathbf{x}_i : i = 1,\ldots,S\}$, $\mathbf{x}_i$ is a vector of observed covariates, c_Δ is a vector of cell counts for the categorical covariates, $K(\mathbf{c}_\Delta)$ is the inverse of the multinomial coefficient in $f_M(\mathbf{c}_\Delta|\boldsymbol{\lambda})$, $\{\boldsymbol{\epsilon}\} = \{\boldsymbol{\epsilon}_i : i = 1,\ldots,S\}$, and $\boldsymbol{\epsilon}_i$ represents the random effects vector for the ith site. If the number of individuals sampled at a site should be considered random, then the multinomial distribution $f_M(\mathbf{c}_i|\mathbf{x}_i)$ could be replaced with the Poisson model as in (3.6).

3.3 Markov Properties of the Random Effects DR Model

Now that we have defined the Random Effects DR (REDR) model (3.9), it is of interest to know what conditions determine the Markov properties of this distribution. In addition, it is also of interest to determine how these properties change when the distribution is marginalized over the random effects.

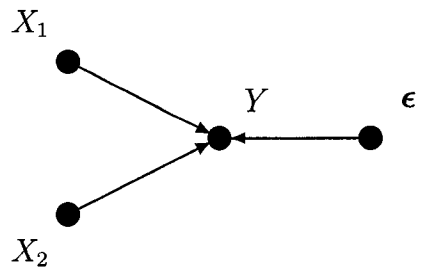

Figure 3.1: Example of an extended graph $\mathcal{G}^\epsilon$ for REDR models with a single response.

So, we begin with the first question. One can observe from (3.10) that the random effects have exactly the same mathematical effect on the response as the main effect terms of the observed covariates. Therefore, we define an *extended chain graph* $\mathcal{G}^\epsilon$ where the random effects are included as a parent of the response variable and are marginally independent of the observed covariates. The graph is represented by a set of vertices for the observable covariates and a vertex for the random effect. There is an directed edge from ϵ to Y. Since the random effects are marginally independent of the observed covariates, there is no edge between the covariates and the random effects. Figure 3.1 gives an illustration of an extended graph for a response with two covariates.

We now provide a proposition that describes necessary and sufficient conditions for a REDR distribution to be Markovian with respect to a given extended graph $\mathcal{G}^\epsilon$.

Proposition 3.1. *A REDR distribution P for a single response variable, given by (3.9) is $\mathcal{G}^\epsilon$ Markovian for a given extended chain graph $\mathcal{G}^\epsilon$, if and only if the interaction terms in (3.9) satisfy the following conditions where $c \subseteq \Gamma$, and $d \subseteq \Delta$:*

1. *(a) $\beta_{cd}(y, \mathbf{x}_\Delta) \equiv 0$ in (3.10) unless $\Phi \cup c \cup d$ is complete in $\mathcal{G}^\epsilon$,*

   (b) $\omega_{\gamma dm}(y, \mathbf{x}_\Delta) \equiv 0$ in (3.10), for $m = 1, \ldots, M$, unless $\Phi \cup \{\gamma\} \cup d$ is complete in $\mathcal{G}^\epsilon$,

2. *(a)* $\lambda_d(\mathbf{x}_\Delta) \equiv 0$ *in (3.14) unless* d *is complete in* $\mathcal{G}^\epsilon$,

   (b) $\tau_{d\gamma}(\mathbf{x}_\Delta) \equiv 0$ *unless* $d \cup \{\gamma\}$ *is complete in* $\mathcal{G}^\epsilon$, *where* $\tau_{d\gamma}(\mathbf{x}_\Delta)$ *is the element of the vector* $\boldsymbol{\tau}_d(\mathbf{x}_\Delta)$ *in (3.14) associated with the variable* X_γ *and* $\gamma \in \Gamma$,

   (c) $\psi_{\{\gamma,\mu\}} \equiv 0$ *unless* $\{\mu, \gamma\}$ *is complete in* $\mathcal{G}^\epsilon$, *where* $\psi_{\{\gamma,\mu\}}$ *is the* (μ, γ) *off-diagonal element of* $\boldsymbol{\Psi}_\emptyset$ *in (3.14) and* $\mu \in \Gamma$.

Proof. To prove Proposition 3.1 we will show that the conditions presented are necessary and sufficient for a REDR model P (3.9) to have Gibbs factorization with respect to $\mathcal{G}^\epsilon$ according to Proposition 2.6 and hence show that P is $\mathcal{G}^\epsilon$ Markovian.

Upon examination of the second set of conditions 2(a) through 2(c), one can observe that they are necessary and sufficient for factorization of the marginal density of $\mathbf{X}$ and $\boldsymbol{\epsilon}$ on the subgraph $(\mathcal{G}^\epsilon)_{\{\Gamma \cup \Delta \cup \boldsymbol{\epsilon}\}}$. The second set of conditions are essentially a re-parameterization of the necessary and sufficient factorization criteria for the CG density parameterized by (3.3) (Lauritzen and Wermuth, 1989, see Proposition 2.5). Condition 2(b) reflects the change in parameterization from (3.3) to (3.14). Conditions 2(a) - 2(c) satisfy Proposition 2.6 for the initial chain component $\Gamma \cup \Delta \cup \boldsymbol{\epsilon}$. Since, the random effects $\boldsymbol{\epsilon}$ are marginally independent from $\mathbf{X}$ by construction (see Section 3.2.2), $f(\mathbf{x}, \boldsymbol{\epsilon}) = f_{CG}(\mathbf{x}) f(\boldsymbol{\epsilon})$ factors according to $(\mathcal{G}^\epsilon)_{\{\Gamma \cup \Delta \cup \boldsymbol{\epsilon}\}}$ if and only if $f_{CG}(\mathbf{x})$ factorizes according to $(\mathcal{G}^\epsilon)_{\{\Gamma \cup \Delta\}}$.

Now, all we need show is that $f_{RE}(\mathbf{y}_\Phi | \mathbf{x}, \boldsymbol{\epsilon})$ factorizes according to complete sets in $\{(\mathcal{G}^\epsilon)_{cl(\Phi)}\}^m$ to complete the factorization of the REDR distribution P according to Proposition 2.6. In order to show this we will follow in the footsteps of the proof of Proposition 2.7. First, note that if conditions 1(a) though 1(b) hold, then P is a function only of complete sets in $\{(\mathcal{G}^\epsilon)_{cl(\Phi)}\}^m$, since the parents of Φ are complete. Now, if we calculate the Hammersley-Clifford interaction terms for the REDR model P in the same manner as in the proof of Proposition 2.7 for the DR model, we see that the interaction terms are identical except for $\phi_{\emptyset \cup \emptyset}(y, \boldsymbol{\epsilon}) = \beta_{\emptyset\emptyset}(y) + \epsilon(y)$. Now,

Φ and ϵ are always connected for any $\{(\mathcal{G}^{\epsilon})_{cl(\Phi)}\}^m$, therefore, P factorizes according to Proposition 2.6 if and only if the conditions of Proposition 3.1 hold. □

The second question, how do the Markov properties change by marginalizing over the random effects, is a more challenging question due to the fact that the model form prohibits analytical integration over the random effects. We will, however, demonstrate that if the conditions of Proposition 3.1 are satisfied for a REDR model with respect to an extended graph $\mathcal{G}^{\epsilon}$, then, the marginal distribution $(Y,\ \mathbf{X})$ is $\mathcal{G} = (\mathcal{G}^{\epsilon})_{V\setminus\epsilon}$ Markovian. Therefore, when interest lies only in the inference of dependence relationships between the covariates and response, the random effects can simply be ignored in the graphical representation.

Proposition 3.2. *If P is a REDR model with a single response variable as described by (3.9), and P is Markovian with respect to the extended graph $\mathcal{G}^{\epsilon}$, then the marginal distribution of the covariates and response, $P_{\Phi\cup\Gamma\cup\Delta}$, is $\mathcal{G} = (\mathcal{G}^{\epsilon})_{V\setminus\epsilon}$ Markovian.*

Proof. If P is a single response REDR model and the conditions of Proposition 3.1 are satisfied for an extended chain graph $\mathcal{G}^{\epsilon}$, then we only need check that the conditional density of the response variable Y factorizes on $(\mathcal{G}_{cl(\Phi)})^m$ according to Theorem 2.3(3) when the density is integrated over the random effects. The conditions are necessary and sufficient for the factorization of the covariate density since the random effects are independent of the covariates. Since the conditions of Proposition 3.1 are satisfied, we can write the conditional density of Y given $(\mathbf{x},\ \boldsymbol{\epsilon})$ as

$$\begin{aligned} f_{RE}(y|\ \mathbf{x},\ \boldsymbol{\epsilon}) &= \prod_{f\subseteq\Phi}\prod_{c\subseteq\Gamma}\prod_{d\subseteq\Delta}\psi_{fcd}(y,\mathbf{x},\boldsymbol{\epsilon}) \\ &= \prod_{f\subseteq\Phi}\prod_{p\subseteq pa(\Phi)}\psi_{fp}(y,\mathbf{x}_p,\boldsymbol{\epsilon}), \end{aligned} \tag{3.18}$$

where $\psi_{fcd}(\cdot)$ is an exponentiated Hammersley-Clifford interaction term calculated as in the proof of the Hammersley-Clifford Theorem (Theorem 2.2) and $pa(\Phi) \subseteq$

$\Gamma \cup \Delta$ represents the set of covariates that are parents of the response vertex Φ. The interaction terms $\psi_{fcd}(y, \mathbf{x}, \boldsymbol{\epsilon})$ reduce to functions ψ_{fp} of only the parents of Φ since those interaction terms including covariates that are not parents of Φ will be equal to 1.

Now, we can integrate out all of the random effects to obtain the conditional density of Y given $\mathbf{X}$,

$$\begin{aligned} f(y|\mathbf{x}) &= \prod_{p \subseteq pa(\Phi)} \int_{\boldsymbol{\epsilon}} \psi_{fp}(y, \mathbf{x}_p, \boldsymbol{\epsilon}) f(\boldsymbol{\epsilon}) d\boldsymbol{\epsilon} \\ &= \prod_{p \subseteq pa(\Phi)} h_{fp}(y, \mathbf{x}_p). \end{aligned} \tag{3.19}$$

Every set $f \cup p$ is complete in the graph $\{(\mathcal{G}^{\epsilon})_{V \setminus \epsilon}\}^m$. Therefore, the conditional density $f(y|\mathbf{x})$ in (3.19) factorizes on $\{(\mathcal{G}^{\epsilon})_{V \setminus \epsilon}\}^m$, which, together with the fact that $f(\mathbf{x})$ factorizes on the subgraph $\mathcal{G}_X$ of the covariate chain component, proves that the joint density factorizes according to Proposition 2.6 and hence, according to Theorem 2.3(3). □

3.4 Parameter Inference

In order to make inference about the parameters in the REDR model in (3.9), we adopt a Bayesian approach for parameter estimation. The hierarchical structure of the REDR model makes Bayesian procedures particularly attractive. There are a large number of unobserved random effects which may or may not be considered nuisance parameters. If the goal of the analysis is solely to make inference concerning the conditional dependencies between the response variable and the observed covariates, then the random effects are considered nuisance parameters due to the fact that the dependence relationships between the observable covariates and the response variable remain unchanged when marginalizing over the random effects. If, however, the unobserved compositions at all, or some, of the sites are of interest then estimates of the random effects are necessary for each site in order to calculate an

estimate of the true site composition. Modern Bayesian computational techniques can handle either of these needs with little modification. Therefore, the Bayesian approach provides a methodology with both goals in mind.

Bayesian inference for graphical composition models proceeds by first defining a prior distribution $\pi(\boldsymbol{\beta}, \boldsymbol{\omega}, \boldsymbol{\lambda}, \boldsymbol{\tau}, \boldsymbol{\Psi}_{\emptyset}, \boldsymbol{\Sigma})$ for the parameters of the model. Here, we have removed subscripts in order to ease notational burden. For example, $\boldsymbol{\beta}$ refers to the entire set of parameters $\{\beta_{cd}(y, \mathbf{x}_{\Delta}) : \ c \subseteq \Gamma, \text{ and } d \subseteq \Delta\}$. We will frequently use this shorthand notation when referring to parameters of the same type in the remainder of the chapter. Assuming that the observations at each site are independent, the posterior distribution of the parameters and the random effects is given by

$$\begin{aligned} f_{\text{post}}(\boldsymbol{\beta}, \boldsymbol{\omega}, \boldsymbol{\lambda}, \boldsymbol{\tau}, \boldsymbol{\Psi}_{\emptyset}, \boldsymbol{\Sigma}, \{\boldsymbol{\epsilon}\} \mid \{\mathbf{c}\}, \{\mathbf{x}\}) \propto & \prod_{i=1}^{S} f_M(\mathbf{c}_i|\boldsymbol{\beta}, \boldsymbol{\omega}, \mathbf{x}_i, \boldsymbol{\epsilon}_i) \\ & \times f_M(\mathbf{c}_{\Delta}|S, \boldsymbol{\lambda}) \\ & \times \prod_{i=1}^{S} f_N(\mathbf{x}_{\Gamma i}|\mathbf{x}_{\Delta i}, \boldsymbol{\tau}, \boldsymbol{\Psi}_{\emptyset}) \\ & \times \prod_{i=1}^{S} f_N(\boldsymbol{\epsilon}_i|\boldsymbol{\Sigma}) \\ & \times \pi(\boldsymbol{\beta}, \boldsymbol{\omega}, \boldsymbol{\lambda}, \boldsymbol{\tau}, \boldsymbol{\Psi}_{\emptyset}, \boldsymbol{\Sigma}). \end{aligned} \tag{3.20}$$

In (3.20), the conditional distribution of the category counts at each site, $f_M(\mathbf{c}_i|\boldsymbol{\beta}, \boldsymbol{\omega}, \mathbf{x}_i, \boldsymbol{\epsilon}_i)$ can be modeled with the multinomial likelihood formulation as shown or with the Poisson formulation (3.6) if the total site count should be considered random.

The posterior distribution (3.20) is a non-standard distribution, therefore, analytical inference for posterior objects of interest such as expected values and credible intervals is not possible. We will draw a sample from this distribution using Markov Chain Monte Carlo (MCMC) techniques. Robert and Casella (1999) provide an in-depth overview of MCMC techniques. Specifically, we will be using a Gibbs sampling approach (see Chapter 7 of Robert and Casella (1999)).

We can obtain simplification in the analysis of posterior distribution by noting that if we impose the *a priori* independence of the CG parameters with the remaining parameters, $\pi(\boldsymbol{\beta}, \boldsymbol{\omega}, \boldsymbol{\lambda}, \boldsymbol{\tau}, \boldsymbol{\Psi}_\emptyset, \boldsymbol{\Sigma}) = \pi(\boldsymbol{\beta}, \boldsymbol{\omega}, \boldsymbol{\Sigma}) \times \pi(\boldsymbol{\lambda}, \boldsymbol{\tau}, \boldsymbol{\Psi}_\emptyset)$, then, the posterior distribution becomes

$$\begin{aligned} & f_{\text{post}}\left(\boldsymbol{\beta}, \boldsymbol{\omega}, \boldsymbol{\lambda}, \boldsymbol{\tau}, \boldsymbol{\Psi}_\emptyset, \boldsymbol{\Sigma}, \{\boldsymbol{\epsilon}\} \mid \{\mathbf{c}\}, \{\mathbf{x}\}\right) \\ & \quad \propto \left\{ \prod_{i=1}^{S} f_M(\mathbf{c}_i \mid \boldsymbol{\beta}, \boldsymbol{\omega}, \mathbf{x}_i, \boldsymbol{\epsilon}_i) f(\boldsymbol{\epsilon}_i | \boldsymbol{\Sigma}_f) \right\} \pi(\boldsymbol{\beta}, \boldsymbol{\omega}, \boldsymbol{\Sigma}) \\ & \qquad \times \left\{ \prod_{i=1}^{S} f_N\left(\mathbf{x}_{\Gamma i} \mid \mathbf{x}_{\Delta i}, \boldsymbol{\tau}, \boldsymbol{\Psi}_\emptyset\right) \right\} f_M(\mathbf{c}_\Delta | S, \boldsymbol{\lambda}) \pi(\boldsymbol{\lambda}, \boldsymbol{\tau}, \boldsymbol{\Psi}_\emptyset) \\ & \quad \propto f_{\text{post}}(\boldsymbol{\beta}, \boldsymbol{\omega}, \boldsymbol{\Sigma}, \{\boldsymbol{\epsilon}\} \mid \{\mathbf{c}\}, \{\mathbf{x}\}) \times f_{\text{post}}(\boldsymbol{\lambda}, \boldsymbol{\tau}, \boldsymbol{\Psi}_\emptyset \mid \{\mathbf{x}\}). \end{aligned} \tag{3.21}$$

Therefore, the parameters of the explanatory portion of the graphical model and the parameters of the response portion of the graphical model are *a posteriori* independent. This simplifies the analysis of the posterior distribution because two separate MCMC analyses can be performed, one for each chain component. In fact, this type of sequential estimation is how chain model parameters are often estimated with maximum likelihood procedures as well (Whittaker, 1990, pg. 310).

Another benefit of posterior independence is that the explanatory portion of the model will remain the same even if other response variables are analyzed. One need only examine the explanatory model once. Then, models for different sets of responses can be fit by analyzing each response portion separately. The full chain graph for each model can be constructed using the same subgraph for the explanatory variable chain component and then simply adding the directed edges and response components based on the set of responses being analyzed.

3.4.1 Hierarchical Centering Parameterization

Here we present a modification to the previous response model parameterizations presented in (3.10). When using a Gibbs MCMC procedure, the Markov chains for regression coefficient parameters, such as $\boldsymbol{\beta}$ and $\boldsymbol{\omega}$, in random effects generalized linear models are often slow to converge to the marginal posterior distribution

(Chen et al., 2000, pg. 40). It has been our experience that this is also the case for graphical composition models. Therefore, we will use a hierarchical centering parameterization, as suggested by Chen et al. (2000), to help reduce the problem of poor mixing chains for the regression coefficients $\boldsymbol{\beta}$ and $\boldsymbol{\omega}$.

In order to describe the hierarchical centering parameterization, first recall the response portion of the REDR distribution (3.10). Now, we introduce a shortened notation for the "fixed" effects portion of the response model,

$$\mu_{\varphi}(y) = \sum_{c \subseteq \Gamma} \sum_{d \subseteq \Delta} \beta_{cd}(y, \mathbf{x}_{\Delta}) \prod_{\gamma \in c} x_{\gamma} + \sum_{\gamma \in \Gamma} \sum_{d \subseteq \Delta} \sum_{m=2}^{M} \omega_{\gamma dm}(y, \mathbf{x}_{\Delta}) x_{\gamma}^{m}. \tag{3.22}$$

Then, we propose the following hierarchically centered re-parameterization of (3.10),

$$f_{RE}^{(h)}(y|\boldsymbol{\varphi}) = \frac{\exp\{\varphi(y)\}}{\sum_{y=1}^{D} \exp\{\varphi(y)\}}, \tag{3.23}$$

where $\varphi(y) = \mu_{\varphi}(y) + \epsilon(y)$. If the assumption is made that $\boldsymbol{\epsilon} \sim f_N(\mathbf{0}, \boldsymbol{\Sigma})$, then $\boldsymbol{\varphi} = \{\varphi(y) : y \neq 1\} \sim f_N(\boldsymbol{\mu}_{\varphi}, \boldsymbol{\Sigma})$, where $\boldsymbol{\mu}_{\varphi}$ is the vector $\{\mu_{\varphi}(y) : \ y = 2, \ldots, D\}$. Here we have simply changed the random effects $\boldsymbol{\epsilon}$ from a zero mean process to a process, $\boldsymbol{\varphi}$, centered at the fixed effects $\boldsymbol{\mu}_{\varphi}$. While there is no theoretical result in the case of generalized linear mixed models to show that this will improve convergence of the MCMC procedure, it has been our experience that the re-parameterization often greatly improves convergence for these models.

We now provide a general parameterization of the full posterior distribution that we recommend using when employing MCMC techniques to make inference concerning the model parameters,

$$\begin{aligned} & f_{\text{post}}^{(h)}(\boldsymbol{\beta}, \boldsymbol{\omega}, \boldsymbol{\lambda}, \boldsymbol{\Psi}_{\emptyset}, \{\boldsymbol{\varphi}\} \mid \{\mathbf{c}\}, \{\mathbf{x}\}) \\ & \quad \propto \left\{ \prod_{i=1}^{S} f^{(h)}(\mathbf{c}_i|\boldsymbol{\varphi}_i) f_N(\boldsymbol{\varphi}_i|\boldsymbol{\beta}, \boldsymbol{\omega}, \boldsymbol{\Sigma}, \mathbf{x}_i) \right\} \pi(\boldsymbol{\beta}, \boldsymbol{\omega}, \boldsymbol{\Sigma}) \\ & \quad \times \left[\prod_{i=1}^{S} f_N(\mathbf{x}_{\Gamma i}|\mathbf{x}_{\Delta i}, \boldsymbol{\tau}, \boldsymbol{\Psi}_{\emptyset}) \right] f_M(\mathbf{c}_{\Delta}|\boldsymbol{\lambda}) \pi(\boldsymbol{\lambda}, \boldsymbol{\tau}, \boldsymbol{\Psi}). \end{aligned} \tag{3.24}$$

Once again, $f^{(h)}(\mathbf{c}_i|\boldsymbol{\varphi}_i)$ in (3.24) can take one of two forms. If the total number of individuals observed at a site is fixed, or should be considered fixed, then the multinomial density,

$$f_M^{(h)}(\mathbf{c}_i|N_i,\mathbf{p}_i) = \frac{N_i}{\prod_{y=1}^{D} c(y)_i!}\prod_{y=1}^{D} f_{RE}^{(h)}(y)^{c(y)_i} \tag{3.25}$$

is used. If, however, the site total is random, then the independent Poisson model,

$$f_P^{(h)}(\mathbf{c}_i|\boldsymbol{\varphi}_i) = \prod_{y=1}^{D} \frac{\varphi(y)^{c(y)_i}e^{\varphi(y)}}{c(y)_i!} \tag{3.26}$$

is used. A slight modification must be made to $\boldsymbol{\mu}_{\varphi}(y)$ for the Poisson model. In order to include a model for the total number of individuals at a site, (3.8) is used for $\boldsymbol{\mu}_{\varphi}$ in place of (3.22).

A few comments concerning this re-parameterized model are in order. First, under either parameterization, the full count likelihood, as well as the REDR density, remain unchanged when integrated over the random effects. One can see this by observing that the transformation from $(\boldsymbol{\beta},\boldsymbol{\omega},\boldsymbol{\epsilon})$ to $(\boldsymbol{\beta},\boldsymbol{\omega},\boldsymbol{\varphi})$ has a Jacobian of 1 since $\boldsymbol{\epsilon} = \boldsymbol{\varphi} - \boldsymbol{\mu}_{\varphi}$. Therefore, we obtain

$$f(\mathbf{c}_i|\mathbf{x}_i,\boldsymbol{\beta},\boldsymbol{\omega}) = \int\cdots\int f^{(h)}(\mathbf{c}_i|\boldsymbol{\varphi})f(\boldsymbol{\varphi}|\boldsymbol{\beta},\boldsymbol{\omega},\boldsymbol{\Sigma},\mathbf{x}_i)d\boldsymbol{\varphi} \tag{3.27}$$

$$= \int\cdots\int f(\mathbf{c}_i|\boldsymbol{\beta},\boldsymbol{\omega},\boldsymbol{\epsilon})f(\boldsymbol{\epsilon}|\mathbf{0},\boldsymbol{\Sigma})d\boldsymbol{\epsilon}, \tag{3.28}$$

and

$$f(y|\mathbf{x}_i) = \int\cdots\int f_{RE}^{(h)}(y|\boldsymbol{\varphi})f(\boldsymbol{\varphi}|\boldsymbol{\beta},\boldsymbol{\omega},\boldsymbol{\Sigma},\mathbf{x}_i)d\boldsymbol{\varphi} \tag{3.29}$$

$$= \int\cdots\int f_{RE}(y|\mathbf{x}_i,\boldsymbol{\epsilon})f(\boldsymbol{\epsilon}|\mathbf{0},\boldsymbol{\Sigma})d\boldsymbol{\epsilon}. \tag{3.30}$$

These equivalences possess two accompanying results. The first result is that the marginal re-parameterized model model (3.29) possesses the same Markov properties as as the integrated REDR model (3.30) (see Section 3.3). Prior to integration,

however, the Markov properties for the two parameterizations are different. For the original parameterization, the model is Markovian with respect to a chain graph where the explanatory variables and the random effects are unconnected parents of the response. For the hierarchically centered parameterization, the model is Markovian with respect to a chain graph where the explanatory variables are parents of the random effects which are in turn parents of the response variable. The second result is that since the integrated likelihoods in (3.27) and (3.28) are equivalent, the marginal posterior distributions of $\boldsymbol{\beta}$ and $\boldsymbol{\omega}$ will be equivalent under either parameterization. This can also be observed empirically by the fact that upon drawing a sample from the posterior distribution of $(\boldsymbol{\beta}, \boldsymbol{\omega}, \boldsymbol{\varphi})$ one can easily transform the sample values to $(\boldsymbol{\beta}, \boldsymbol{\omega}, \boldsymbol{\epsilon})$, thus obtaining a sample from its posterior distribution. The $\boldsymbol{\beta}$ and $\boldsymbol{\omega}$ values remain unchanged, therefore, any marginal posterior quantities calculated also remain unchanged.

3.4.2 Implementing the Gibbs Sampler

In order to implement the Gibbs sampler to draw a sample from (3.24) we need to obtain the full conditional distribution for each parameter. The full conditional distribution is the conditional distribution of the parameter in question given all remaining parameters as well as the observed data. A (non-independent) sample from the posterior is then drawn by iteratively drawing from each full conditional distribution. We derive the full conditional densities in two separate groups due to the fact that the full conditional densities for the response model parameters, $\boldsymbol{\beta}$, $\boldsymbol{\omega}$, $\{\boldsymbol{\varphi}\}$, and $\boldsymbol{\Sigma}$, will not be functions of the explanatory model parameters $\boldsymbol{\lambda}$, $\boldsymbol{\tau}$, and $\boldsymbol{\Psi}_{\emptyset}$, as can be observed from (3.24). Therefore, we can make inference to the response model parameters using only the first factor on the left hand side of the proportionality, while inference to the explanatory portion of the chain graph uses only the second factor.

Response Model Conditional Densities

Before deriving the full conditional densities we introduce some notation to ease the calculations. First, we will use the notation $\mathbf{E}$ to refer to a matrix that has S rows and has the following: a column of ones, a column corresponding to each of the explanatory variables, a column for each interaction and powers of the continuous covariates as given in (3.10). If a covariate X_δ, $\delta \subseteq \Delta$, is a categorical variable with b levels, then it will be represented by $b-1$ columns of indicator variables in $\mathbf{E}$, where each column indicates, with a one or zero, if X_δ takes the associated level at site i. The column associated with the reference level $X_\delta = 1$ is not included. We will denote the number of columns in $\mathbf{E}$ by r. In linear regression terminology, $\mathbf{E}$ represents the design matrix. The vector $\mathbf{E}_i$, $i = 1, \ldots, S$ will denote an r vector formed from the ith row of $\mathbf{E}$. In addition, let $\mathbf{B}$ represent an $r \times (D-1)$ matrix of all the interaction coefficients $\{\beta_{cd}(y) : c \subseteq \Gamma,\ d \subseteq \Delta\}$ and $\{\omega_{\gamma dm}(y, \mathbf{x}_\Delta) : \gamma \in \Gamma,\ d \subseteq \Delta,\ m = 1, \ldots, M\}$ such that the expected value (3.22) of the site i random effect $\boldsymbol{\varphi}_i$ is given by $\boldsymbol{\mu}_\varphi = \mathbf{B}'\mathbf{E}_i$. The "stacked" version of $\mathbf{B}$ will be represented by $\mathbf{B}_s$. The stacked version is a $r(D-1) \times 1$ vector where the columns of $\mathbf{B}$ have been concatenated in order. Although, previously only described as the collection of all random effects, $\boldsymbol{\varphi}$ will now specifically represent a $S \times (D-1)$ matrix of these random effects and $\boldsymbol{\varphi}_i$ is a $D-1$ vector formed from the ith row of $\boldsymbol{\varphi}$. Finally, we will make use of the inverse of the random effects covariance matrix $\mathbf{T} = \boldsymbol{\Sigma}^{-1}$.

Now we can begin to derive full conditional distributions for the parameters of the response model, the coefficients in $\mathbf{B}$, the site random effects $\boldsymbol{\varphi}$, and the random effects inverse covariance matrix $\mathbf{T}$. The reader should note, as we derive the full conditional densities, that in addition to the increased rate of convergence, the hierarchical centering provides a Gibbs sampler that is easier to implement due to the fact that the interaction coefficients as well as the random effects covariance

matrix will have standard densities. In the non-centered parameterization only the covariance matrix has a standard full conditional density. Here, we will also make the assumption that the coefficient parameters are *a priori* independent from the random effects covariance matrix. In other words, $\pi(\boldsymbol{\beta}, \boldsymbol{\omega}, \boldsymbol{\Sigma}) = \pi(\boldsymbol{\beta}, \boldsymbol{\omega})\pi(\boldsymbol{\Sigma})$

We begin with the interaction coefficients in $\mathbf{B}$. First, note that due to the centering parameterization, given the random effects at each site and the random effect covariance matrix, the interaction coefficients are independent of the category counts. If we let $\pi(\boldsymbol{\beta}, \boldsymbol{\omega}) = \pi(\mathbf{B}_s) = f_N(\boldsymbol{\mu}_{B_s}, \mathbf{V}_{B_s}^{-1})$ and $\hat{\mathbf{B}} = (\mathbf{E}'\mathbf{E})^{-1}\mathbf{E}'\boldsymbol{\varphi}$ (correspondingly $\hat{\mathbf{B}}_s$ represents the stacked version) then the full conditional distribution of the interaction coefficients is given as

$$\begin{aligned} f(\mathbf{B}_s| \ \dots \ ) &\propto \exp\left\{-\frac{1}{2}\sum_{i=1}^{S}\left(\boldsymbol{\varphi}_i - \mathbf{B}'\mathbf{E}_i\right)'\mathbf{T}\left(\boldsymbol{\varphi}_i - \mathbf{B}'\mathbf{E}_i\right)\right\} \\ &\qquad \times \exp\left\{-\frac{1}{2}(\mathbf{B}_s - \boldsymbol{\mu}_{B_s})'\mathbf{V}_{B_s}(\mathbf{B}_s - \boldsymbol{\mu}_{B_s})\right\} \\ &\propto \exp\left\{-\frac{1}{2}tr\left[\mathbf{T}(\mathbf{B} - \hat{\mathbf{B}})'\mathbf{E}'\mathbf{E}(\mathbf{B} - \hat{\mathbf{B}})\right]\right\} \\ &\qquad \times \exp\left\{-\frac{1}{2}(\mathbf{B}_s - \boldsymbol{\mu}_{B_s})'\mathbf{V}_{B_s}(\mathbf{B}_s - \boldsymbol{\mu}_{B_s})\right\} \\ &= \exp\left\{-\frac{1}{2}(\mathbf{B}_s - \hat{\mathbf{B}}_s)'(\mathbf{T}\otimes\mathbf{E}'\mathbf{E})(\mathbf{B}_s - \hat{\mathbf{B}}_s)\right\} \\ &\qquad \times \exp\left\{-\frac{1}{2}(\mathbf{B}_s - \boldsymbol{\mu}_{B_s})'\mathbf{V}_{B_s}(\mathbf{B}_s - \boldsymbol{\mu}_{B_s})\right\}, \end{aligned} \tag{3.31}$$

where $\otimes$ represents the Kronecker product. The second proportionality statement for the random effects likelihood is due to Johnson and Wichern (1992, pg. 322). One can now complete the square to show that

$$f(\mathbf{B}_s| \ \dots \ ) = f_N(\mathbf{B}_s;\ \boldsymbol{\mu}_1, \mathbf{V}_1^{-1}), \tag{3.32}$$

where the mean and covariance are given by

$$\begin{aligned} \boldsymbol{\mu}_1 &= \left[(\mathbf{T}\otimes\mathbf{E}'\mathbf{E}) + \mathbf{V}_{B_s}\right]^{-1}\left[(\mathbf{T}\otimes\mathbf{E}'\mathbf{E})\hat{\mathbf{B}}_s + \mathbf{V}_{B_s}\boldsymbol{\mu}_{B_s}\right] \\ &\text{and} \\ \mathbf{V}_1 &= (\mathbf{T}\otimes\mathbf{E}'\mathbf{E}) + \mathbf{V}_{B_s}. \end{aligned} \tag{3.33}$$

Therefore, in the Gibbs sampler, drawing samples of the interaction coefficients is relatively simple. Updating can be done for a single parameter at a time as well, each one will have a univariate Normal full conditional density.

We now derive the full conditional distribution for the inverse covariance matrix $\mathbf{T}$ of the random effects $\boldsymbol{\varphi}$. We assume, *a priori*, that $\mathbf{T}$ has a Wishart distribution, $f_W(\mathbf{T};\ a, \mathbf{K})$, with prior parameters $a > D-1$, $(D-1)\times(D-1)$ positive definite matrix $\mathbf{K}$, and density

$$\begin{aligned}\pi(\mathbf{T}) &= f_W(\mathbf{T};\ a, \mathbf{K}) \\ &\propto |\mathbf{T}|^{(a-D-2)/2} \exp\left\{-\frac{1}{2} tr\left[\mathbf{KT}\right]\right\}.\end{aligned} \tag{3.34}$$

This is equivalent to specifying an inverse Wishart prior distribution for $\boldsymbol{\Sigma}$. Now, $\mathbf{T}$ only depends on $\boldsymbol{\varphi}$ and $\mathbf{B}$ though the random effects distribution, which is a MVN distribution. Therefore, we obtain the following full conditional distribution,

$$\begin{aligned} f(\mathbf{T}|\ \ldots\ ) &\propto |\mathbf{T}|^{S/2} \exp\left\{-\frac{1}{2}\sum_{i=1}^{S}(\boldsymbol{\varphi}_i - \mathbf{B}'\mathbf{E}_i)'\mathbf{T}(\boldsymbol{\varphi}_i - \mathbf{B}'\mathbf{E}_i)\right\} \\ &\quad \times |\mathbf{T}|^{(a-D-2)/2} \exp\left\{-\frac{1}{2} tr\left[\mathbf{KT}\right]\right\} \\ &= |\mathbf{T}|^{(a+S-D-2)/2} \\ &\quad \times \exp\left\{-\frac{1}{2} tr\left[\mathbf{T}\left\{\mathbf{K} + \sum_{i=1}^{S}(\boldsymbol{\varphi}_i - \mathbf{B}'\mathbf{E}_i)'(\boldsymbol{\varphi}_i - \mathbf{B}'\mathbf{E}_i)\right\}\right]\right\}. \end{aligned} \tag{3.35}$$

It follows, then, upon examination of (3.34), the full conditional distribution of $\mathbf{T}$ is given by

$$f(\mathbf{T}|\ \ldots\ ) = f_W(\mathbf{T};\ a_1, \mathbf{K}_1) \tag{3.36}$$

where the full conditional parameters are

$$a_1 = a + S$$

and (3.37)

$$\mathbf{K}_1 = \mathbf{K} + \textstyle\sum_{i=1}^{S}(\boldsymbol{\varphi}_i - \mathbf{B}'\mathbf{E}_i)'(\boldsymbol{\varphi}_i - \mathbf{B}'\mathbf{E}_i).$$

Therefore, just like the interaction coefficients, the inverse covariance matrix $\mathbf{T}$ is relatively straightforward to sample from in the Gibbs algorithm.

We now turn our attention to the final parameter that needs to be updated during the Gibbs sampling algorithm, the vector of site random effects,$\boldsymbol{\varphi}_i$. Unfortunately, the random effects do not have a standard full conditional distribution. The full conditional density is given by

$$f(\boldsymbol{\varphi}_i| \ \ldots \ ) \propto f^{(h)}(\mathbf{c}_i|\boldsymbol{\varphi}_i) f_N(\boldsymbol{\varphi}_i;\ \mathbf{B}'\mathbf{E}_i, \mathbf{T}), \tag{3.38}$$

where $f^{(h)}(\mathbf{c}_i|\boldsymbol{\varphi}_i)$ is either a multinomial density, as in (3.5), or a product Poisson density, as in (3.6). One can see from the full conditional density that for either category count likelihood model, Poisson or Multinomial, the full conditional density is non-standard. Therefore, we can employ a Metropolis-within-Gibbs step, as described in Section 1.5.2, to sample from this full conditional distribution. Another option exists, however, for the Poisson model. The full conditional density for the random effect ϕ_{ij} is log-concave if the Poisson model is used, therefore, one can make use of the adaptive-rejection sampler of Gilks and Wild (1992) instead of the Metropolis-within-Gibbs step. Often the adaptive rejection sampler is more efficient than the Metropolis-within-Gibbs sampler (Chen et al., 2000).

Explanatory Model Conditional Distributions

In order to derive the full conditional distributions for the parameters in the explanatory variable CG model (3.14), first note that the categorical and continuous explanatory variable parameters are functionally independent. Therefore, as with the initial separation of the response variable model and the explanatory variable model we can again perform separate posterior analyses for each portion, discrete and continuous, of the explanatory model.

We begin with the analysis of the continuous variable model for $\mathbf{X}_\Gamma$ given $\mathbf{x}_\Delta$. We will follow in the same footsteps as with the interaction coefficients in the response model. Let $\mathbf{E}_\Delta$ represent a matrix with S rows and columns containing indicator variables for each categorical variable in Δ and all of the interactions $d \subseteq \Delta$.

Again, this is identical to the classic ANOVA design matrix. The notation $\mathbf{E}_{\Delta_i}$ will represent a column vector obtained from the ith row of $\mathbf{E}_\Delta$. As with the interaction coefficients of the response model, we will place the τ parameters in a matrix $\mathbf{B}_\tau$ such that $\mathbf{B}_\tau'\mathbf{E}_{\Delta_i} = \sum_{d\subseteq\Delta} \boldsymbol{\tau}_d(\mathbf{x}_\Delta)$. Again, we also denote the "stacked" version of $\mathbf{B}_\tau$ as $\mathbf{B}_{\tau s}$. Now, using the prior distributions $\pi(\mathbf{B}_{\tau s}) = f_N(\mathbf{B}_{\tau s};\ \boldsymbol{\mu}_{\tau s}, \mathbf{V}_{\tau s}^{-1})$ and $\pi(\boldsymbol{\Psi}_\emptyset) = f_W(\boldsymbol{\Psi}_\emptyset;\ a_\psi, \mathbf{K}_\psi)$ and a set of proportionalities and equalities similar to those in (3.31) and (3.35), we obtain the full conditional distributions

$$f(\mathbf{B}_{\tau s}|\ \dots\ ) = f_N(\mathbf{B}_{\tau s};\ \boldsymbol{\mu}_{\tau s,1}, V_{\tau s,1}^{-1}) \tag{3.39}$$

and

$$f(\boldsymbol{\Psi}_\emptyset|\ \dots\ ) = f_W(\boldsymbol{\Psi}_\emptyset;\ a_{\psi,1}, \mathbf{K}_{\psi,1}), \tag{3.40}$$

where the parameters are given by

$$\begin{aligned}
\boldsymbol{\mu}_{\tau s,1} &= \left[(\boldsymbol{\Psi}_\emptyset \otimes \mathbf{E}_\Delta'\mathbf{E}_\Delta) + \mathbf{V}_{\tau s}\right]^{-1}\left[(\boldsymbol{\Psi}_\emptyset \otimes \mathbf{E}_\Delta'\mathbf{E}_\Delta)\hat{\mathbf{B}}_{\tau s} + \mathbf{V}_{\tau s}\boldsymbol{\mu}_{\tau s}\right], \\
\mathbf{V}_{\tau s,1} &= (\boldsymbol{\Psi}_\emptyset \otimes \mathbf{E}_\Delta'\mathbf{E}_\Delta) + \mathbf{V}_{\tau s}, \\
a_{\psi,1} &= a_\psi + S, \qquad \text{and} \\
\mathbf{K}_{\psi,1} &= \mathbf{K}_\psi + \textstyle\sum_{i=1}^{S}\left\{\mathbf{x}_{\Gamma i} - \sum_{d\subseteq\Delta}\boldsymbol{\tau}_d(\mathbf{x}_{\Delta i})\right\}'\left\{\mathbf{x}_{\Gamma i} - \sum_{d\subseteq\Delta}\boldsymbol{\tau}_d(\mathbf{x}_{\Delta i})\right\}.
\end{aligned} \tag{3.41}$$

The final set of parameters for the explanatory variable model is $\boldsymbol{\lambda}$. Unfortunately, as with the random effects, $\lambda_d,\ d \subseteq \Delta$ does not have a standard full conditional distribution. Here, we assume an independent multivariate normal distribution for each of the non-zero elements of $\boldsymbol{\lambda}_d$, so, the full conditional density is given by

$$f(\lambda_d(\mathbf{x}_\Delta)|\ \dots\ ) \propto f_M(\mathbf{c}_\Delta|\boldsymbol{\lambda}) f_N(\boldsymbol{\lambda}_d;\ \boldsymbol{\mu}_{\boldsymbol{\lambda}_d}, \mathbf{T}_{\boldsymbol{\lambda}_d}^{-1}), \qquad d \subseteq \Delta, \tag{3.42}$$

where $f(\mathbf{c}_\Delta|\boldsymbol{\lambda})$ is the multinomial density (3.25). One can see from the full conditional density that the full conditional density is non-standard. Therefore, we can again employ a Metropolis-within-Gibbs step.

3.5 Graphical Analysis of Benthic Invertebrate Functional Groups

Relating behavioral characteristics of organisms to environmental conditions at locations in which they are found has been a challenging problem for ecologists (Legendre et al., 1997). This problem was initially motivated by the n dimensional niche hypothesis. The n dimensional niche hypothesis states there is an n dimensional space upon which species unimodally distribute themselves according to environmental adaptations (Ricklefs, 1990). The unimodal distribution of a given species along an environmental axis implies the existence of an "environmental optimum" for that species. The problem of relating species traits to environmental conditions emerged from this hypothesis because distributions of specific species are usually not of interest as they are often biogeographically constrained, thereby limiting ecological inference to one geographic range. Analysis of functional traits and behaviors rather than taxonomy provides a more portable inference to community structure (Poff and Allen, 1995).

The initial attempts to describe these relationships involved the use of Canonical Correspondence Analysis (CCA) (ter Braak, 1985). The CCA approach attempts to ordinate each species along a set of environmental axes. Dolédec et al. (1996) continued the ordination approach by developing methods for marginally and jointly analyzing so called R, L, and Q tables, where R is a table with data on environmental variables at each sampling site, L is a table of species occurrences at each site, and Q is a table of trait classifications for each species.

A more direct approach was introduced by Legendre et al. (1997). Legendre et al. (1997) termed their methodology a "solution to the fourth corner problem." For a single trait with multiple levels, the "four corners" represent four matrices: (1) a matrix of continuous (or discrete, but can be modeled as continuous) environmental variables by site, (2) an indicator matrix of species presence by site, (3) an

indicator matrix of functional trait levels by species, and (4) a matrix of parameters relating environmental variables to the trait. The parameters in matrix (4) are product moment correlations between the trait counts and environmental variables and are estimated by a method of moments approach.

There are two main problems with the previous methodologies. Firstly, the previous approaches both measure marginal association between the environment and traits in question. The conditional relationships of a Markov random field give a more detailed measure of association between variables. For example, variables that are marginally correlated may in fact be independent upon conditioning on a third variable. This may provide evidence of possible mitigation by the third variable. Secondly, the previous methods provide no predictive ability. If a researcher desires to predict community structure at a site with remotely sensed environmental measurements, the previous methods provide no means to accomplish this task.

Billheimer and Guttorp make use of the response portion of the REDR model, with MVN error terms, to estimate species composition of invertebrate traits in both an experimental setting (Billheimer et al., 2001) as well as an observational setting (Billheimer and Guttorp, 1997). In both cases, only one covariate was used in the model. In both studies, inference was made by examining the deviance of the estimated covariance parameter compositions from a "neutral" composition.

In this section we will use the full REDR distribution to model the relationship between stream invertebrate feeding type and several environmental variables for several stream sites in Oregon. By using the full REDR model we will be able to examine the complex conditional relationships of the environmental covariates and feeding type trait as a whole system through inference from a graphical chain model.

3.5.1 Data Description

Various stream sites in Oregon were visited as part of the U. S. Environmental Protection Agency's (EPA) Regional Environmental Monitoring and Assessment

Table 3.1: Summary of stream invertebrate feeding groups. Abbreviation codes as well as a brief description of the ecological role for each feeding type are given.

Code	Feeding Type	Ecological Role
CF	Collector (Filterer)	Process fine particulate organic material
CG	Collector (Gatherer)	Process fine particulate organic material
GZ	Grazer	Ingest algae located on substrate
SH	Shedder	Process coarse organic material
EP	Engulfing Predator	Ingest other benthic invertebrates
OT	Other	Mainly herbivorous

Program (REMAP). A total of $S = 94$ sites were visited from which data was collected on stream invertebrate species abundance, as well as various local environmental variables. A set of watershed scale environmental variables was also calculated for these sites from a GIS (Geographic Information System) model (Pizzi, 2002).

In order to perform a functional trait analysis, each species was categorized into one of the following types according to its feeding habit (Table 3.1). Collector organisms remain relatively stationary and, as their name suggests, collect food in their proximity. Gatherers find food that is located on the stream bed, while Filterers catch food that flows downstream suspended in the water. Grazers feed primarily on algae located on the substrate. Engulfing Predators feed on other invertebrates. Shredders feed primarily on coarse organic material, such as leaves, as it flows downstream. The Other category includes piercing predators, as well as invertebrate herbivores, groups which are very rare inclusions. Most benthic invertebrates are detritivores, meaning they ingest dead organic material. Figure 3.2 illustrates the distribution of relative abundances for each feeding type.

In our analysis, we are interested in the associations between feeding type and the following covariates from Pizzi (2002): Percent of the substrate composed of woody material, Alkalinity of the water at the sampling site (μeq/L), percent of

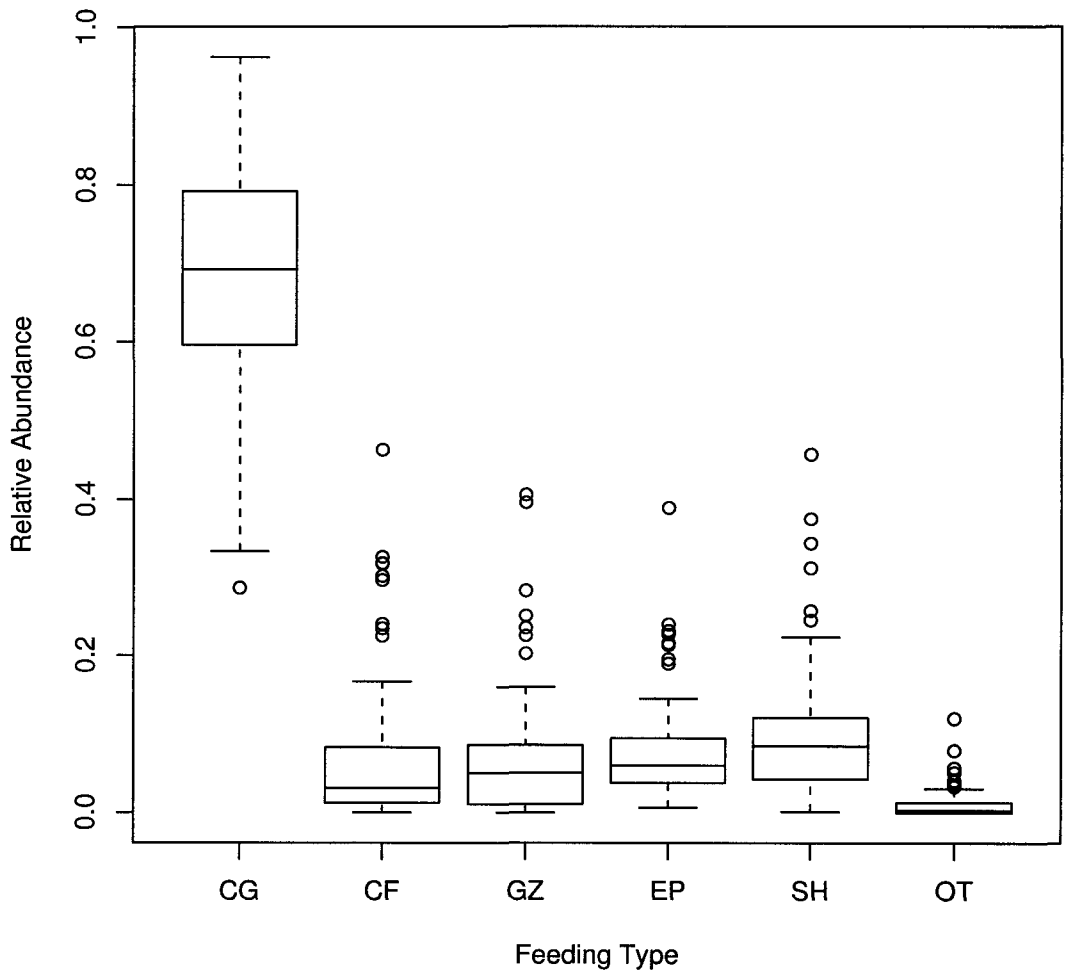

Figure 3.2: Distribution of feeding type relative abundance. Feeding type notation is given in Table 3.1.

Table 3.2: Summary of environmental covariates for Oregon REMAP streams. Alkalinity, watershed area, elevation, precipitation, and the number of road crossings were transformed via $t(x) = \log(x + 1)$ to improve normal approximation.

Covariate	Mean	St. Dev.	Min.	Max.
% Wood	8.32	7.84	0.00	32.62
Alkalinity ($\ln(\mu eq/L + 1)$)	2.70	0.28	2.14	3.73
% Barren	7.17	8.61	0.00	45.04
Area ($\ln(mi^2 + 1)$)	0.90	0.59	0.02	2.43
Elevation ($\ln(m + 1)$)	2.27	0.61	0.78	3.27
Precipitation ($\ln(in + 1)$)	1.87	0.20	1.13	2.26
Crossing ($\ln(no. + 1)$)	0.89	0.72	0.00	2.52

the watershed that is classified as barren land, the watershed area (mi^2), minimum basin elevation in the watershed (m), the mean annual amount of precipitation in watershed (in), and the number of road crossings in the watershed. Alkalinity, area, precipitation, elevation, and number of road crossings were transformed via $t(x) = \log(x + 1)$ to give them a more normal appearance. The watershed scale variables, percent barren land, watershed area, elevation, precipitation, road crossings, were determined with a GIS model, while the local scale variables alkalinity and percent woody material were measured at each sampling location (Pizzi, 2002). Table 3.2 provides a numeric summary of the environmental covariates.

3.5.2 Model Description and Analysis

Model Specification

We consider a main effects only model for the analysis of feeding type composition. In addition, only first order linear interaction terms are included. In the analysis of stream invertebrate data, the Multinomial likelihood must be used due to the field sampling technique. The observed counts are the result of a censored sample. At any given site thousands of individuals are usually collected and only a

fixed number are classified to taxonomic classes. Processing of the samples requires a secondary random sample of a fixed number of individuals. If the primary sample from the stream site contains less than this fixed number, then all of the individuals are classified. In the case of the Oregon REMAP samples, 300 of the total number of sampled individuals were classified. The Poisson likelihood model would be inappropriate due to the sample censoring. Therefore, we use the following REDR model for the category counts,

$$f_M(\mathbf{c}_i|\mathbf{x}_i,\boldsymbol{\epsilon}) = \frac{N!}{\prod_{y=1}^{D} c(y)_i!}\prod_{y=1}^{D} f_{RE}(y|\mathbf{x}_i,\boldsymbol{\epsilon}_i)^{c(y)_i}, \qquad i = 1,\ldots,89, \tag{3.43}$$

where

$$f_{RE}(y|\mathbf{x}_i,\boldsymbol{\epsilon}_i) = \exp\left\{\alpha_\Phi(\mathbf{x}_i,\boldsymbol{\epsilon}_i) + \sum_{\gamma=0}^{7}\beta_\gamma(y)(x_{i\gamma}-\bar{x}_\gamma)s_\gamma^{-1} + \epsilon_i(y)\right\}. \tag{3.44}$$

Each of the environmental covariates was centered by subtracting its mean $\bar{x}_\gamma$ and dividing by its standard deviation s_γ. This was done to improve Markov chain convergence to the posterior distribution. We chose the reference category y^* to be the CG feeding type ($y = 1$). Therefore, in equation (3.44) $\beta_\gamma(y) \equiv \epsilon(y) \equiv 0$ for $y = 1$ to ensure identifiability of the model. The CG feeding type was the most common for the Oregon data and in our experience, fixing the parameters of the most common type also tends to lead to faster convergence of the Markov chains for $\boldsymbol{\beta}$ to their posterior distributions.

Since all of the included environmental covariates are continuous, the homogeneous CG model reduces to a MVN distribution. Therefore, we modeled the mean centered covariates as $f_{CG}(\mathbf{x}_i - \bar{\mathbf{x}}) = MVN(\mathbf{x}_i - \bar{\mathbf{x}};\ \mathbf{0}, \boldsymbol{\Psi}_\emptyset)$. The centering here allows the elimination of the nuisance parameter $\boldsymbol{\tau}_\emptyset$ in (3.14), which is irrelevant for determining conditional independencies. Note, that in this case one can analytically determine the posterior distribution of $\boldsymbol{\Psi}_\emptyset$ to be a Wishart distribution. However, since we are interested in the off-diagonal elements of $\boldsymbol{\Psi}_\emptyset$, using an MCMC sampling technique allows straightforward inference of these elements.

Model Estimation and Performance

For estimation purposes, the model was re-parameterized with the hierarchical centering approach of Section 3.4.1. MCMC procedures were performed with the hierarchically centered version of (3.44) as well as the normal REDR parameterization in (3.44). As hypothesized, the hierarchically centered version reached satisfactory convergence with substantially fewer MCMC iterations than (3.44).

In order to assess convergence of the Markov chains to the posterior distribution, the diagnostic procedure of Raftery and Lewis (1992) was employed. This diagnostic procedure estimates the number of iterations necessary to estimate a specified quantile of a parameter's posterior distribution within a given degree of accuracy. Since we are primarily interested in credible intervals for the β coefficients and the off-diagonal entries of $\mathbf{\Psi}_{\emptyset}$, we felt that this procedure provided the appropriate measure of convergence. For this analysis we specified the 0.025 quantile for estimation with an error no more than ± 0.005 with 95% probability.

The program WinBUGS was used to run the Gibbs sampler (Spiegelhalter et al., 2000). The Gibbs sampling algorithm was run for an initial 4000 iterations in which a MVN proposal density was tuned so that the Metropolis-within-Gibbs step for the $\boldsymbol{\varphi}_i$ parameters would have an acceptance rate of around 30%. The first 4000 iterations were then discarded, after which the sampler was run for an additional 40,000 iterations. The Raftery and Lewis convergence diagnostic procedure estimated the run length necessary for accurate quantile estimation to be less than 40,000 iterations for all parameters.

Performance of the model was assessed in two ways. First, the Bayesian posterior predictive p-value method of Gelman et al. (1996) was used to assess model fitness. For a "goodness-of-fit" statistic $D(y)$, which can be a function of the observed data y and model parameters, the Bayesian predictive p-value is defined as $P_b = Pr\{D(y^{rep}) > D(y) \mid y\}$. The data y^{rep} represents a hypothesized replicate

data set that could have resulted from the model. So, the interpretation is essentially the same as the classic p-value with the addition that the "null" distribution is the distribution of the statistic given only the observed data y. In an MCMC setting P_b is particularly easy to approximate. One simply performs the usual Gibbs sampler for parameter estimation with the addition that at each iteration a replicate data set is generated from the sampled parameter values. At each iteration one calculates $D(y)$ and $D(y^{rep})$ and records the proportion of iterations in which $D(y^{rep}) > D(y)$ (Gelman et al., 1996). We used this approach in our Gibbs algorithm for the feeding type REDR model. The goodness-of-fit statistic used for this analysis was the Freeman-Tukey statistic (Freeman and Tukey, 1950)

$$D(\mathbf{c}_1, \ldots, \mathbf{c}_S) = \sum_{i=1}^{S} \sum_{y=1}^{D} \left(\sqrt{c(y)_i} - \sqrt{N_i f_{RE}(y|\mathbf{x}_i, \boldsymbol{\epsilon}_i)} \right)^2, \tag{3.45}$$

where N_i is the total number of invertebrates observed at site i, $c(y)_i$ is number of invertebrates at site i that are feeding type $y = 1, \ldots, 6$, and $f_{RE}(y|\mathbf{x}_i, \boldsymbol{\epsilon}_i)$ is given by (3.44) and represents the site composition. Other statistics could be used, such as Pearson's χ^2, however, there are many cells with small counts and the Freeman-Tukey statistic eliminates the problem of over-weighting by those cells (Brooks et al., 2000a).

The second measure of model performance we employed is prediction of the feeding type composition at unmeasured sites. To accomplish this the category counts for five of the steam sites were removed from the estimation procedure, leaving $S = 89$ sites for parameter inference. Predictions of the feeding type compositions at each site were then based on the posterior predictive distribution of the latent composition elements

$$f(p(y)_i|\mathbf{x}_i, \mathbf{c}_1, \ldots, \mathbf{c}_{89}) = \int f_{RE}(y|\mathbf{x}_i, \boldsymbol{\epsilon}) f_{\text{post}}(\boldsymbol{\beta}, \boldsymbol{\epsilon}_i, \boldsymbol{\Psi}_\emptyset)\, d\boldsymbol{\beta}\, d\boldsymbol{\epsilon}_i\, d\boldsymbol{\Psi}_\emptyset, \quad y = 1, \ldots, 6, \tag{3.46}$$

where $p(y)_i$ represents the compositional element associated with feeding type y at site i. The predictive distribution can be approximated with Gibbs sampler by simply updating the missing counts as additional parameters in the Gibbs algorithm (Besag et al., 1995). This is essentially the same procedure as generating the y^{rep} data for the Bayesian p-value. The "unobserved" counts are multinomially distributed with probabilities $f_{RE}(y|\mathbf{x}_i, \boldsymbol{\epsilon})$. So, the counts can be updated with little effort, the $\boldsymbol{\beta}$ and $\boldsymbol{\epsilon}$ parameters in the probabilities of interest are updated as in the Gibbs sampler description. Credible intervals for (3.46) can then be approximated using empirical quantiles of the Gibbs sample. Point estimates can be made from the sample means of the MCMC output, however, we propose an alternative estimate that Billheimer (1995) uses. We suggest a parametric mode estimate calculated by treating the sampled $\mathbf{p}_i = \{p(y)_i : \ y = 1, \ldots, 6\}$ compositions as a sample from a LN distribution. The maximum likelihood estimates of the LN parameters are easily calculated using the ALR transformation (Aitchison, 1986). Then, a point estimate can be obtained by selecting the sampled $\mathbf{p}_i$ that maximizes the empirical LN distribution. This summary provides a better point estimate due to the fact that the skewed predictive distributions.

3.5.3 Results and Discussion

The 95% Highest Probability Density (HPD) intervals for the β interaction coefficients are given in Table 3.1. Proposition 2.7 says that an environmental covariate must be a parent of the feeding type response if at least one of the interaction coefficients associated with that variable is not equal to zero. Upon examination of the HPD intervals one can see that there is ample evidence that % Wood, Precipitation, Elevation, % Barren land, and Watershed area are parents of the feeding type response. The HPD intervals for the Alkalinity and Road Crossing interaction terms contain zero for each feeding type, therefore, the analysis does not provide

strong evidence that they are parents of feeding type. The 95% HPD intervals for the off-diagonal elements of $\Psi_\emptyset$ are given in Table 3.2. Again, by examining Table 3.2, the intervals that do not contain zero provide strong evidence that there exists an undirected edge between the two associated variables in the marginal graph for the covariates.

Figure 3.1 illustrates the chain graph that is suggested by the HPD intervals that do not contain zero. By applying the Markov properties of chain graphs to this model one can observe that there seems to be little evidence that the disturbance variables Alkalinity and Number of Road Crossing directly affect the Feeding Type composition. This implies that Feeding Type composition is not a useful index for measuring the degradation of the stream due to these two disturbances. Their effects are mediated through the parents of Feeding Type in Figure 3.1. Another result that can be observed in Figure 3.1 is that variables from both the local scale and watershed scale are parents of Feeding Type, supporting the opinion of Poff (1997) that measuring the environment at multiple spatial scales can provide a more accurate picture of how the environment influences functional traits. If only local scale variables were included in this analysis, it would appear that Alkalinity directly influences Feeding Type. Marginalizing the chain graph of Figure 3.1 over the watershed scale variables would produce a distribution that is Markovian with respect to a chain graph in which Alkalinity is a parent of Feeding Type.

The results of the two model goodness-of-fit procedures show that there is little evidence for a systematic lack of fit. The Bayesian p-value was $P_b = 0.31$. This states that around 1/3 of the replicated data sets produced a test statistic larger than the value for the observed data. We also fit a model with a Multivariate t distribution for the errors as described by (3.11). This model was expected to increase the fitness by allowing larger outliers for the random effects. Only a slight gain, $P_b = 0.37$, was realized, however. The results of the left out site predictions are given in Table 3.3.

Note, however, that the observed compositions represent a sample of individuals from a population described by a true composition. The predictions are targeted at these true compositions. So, the prediction is not trying to estimate the observed value, but the true unobserved composition. We are merely trying to get a sense of accuracy by comparing the prediction to the observation. In general, however, the point estimates reflected the overall percentages observed at the sites and all of the observed values were contained in the HPD intervals.

Table 3.3: 95% HPD intervals for interaction coefficients for stream invertebrate feeding type analysis. Feeding types are given in Table 3.1.

	Feeding Type					
Covariate	CG*	CF	GZ	EP	SH	OT
% Wood	–	(-0.805, -0.057)	(-0.594, 0.067)	(-0.076, 0.273)	(-0.226, 0.184)	(-0.650, 0.138)
Alkalinity	–	(-0.667, 0.270)	(-0.590, 0.240)	(-0.370, 0.072)	(-0.467, 0.053)	(-0.345, 0.615)
% Barren	–	(-0.254, 0.457)	(-0.206, 0.428)	(-0.186, 0.151)	(-0.220, 0.176)	(-0.888, -0.048)
Area	–	(-0.178, 1.154)	(-0.549, 0.653)	(-0.379, 0.250)	(-0.614, 0.121)	(0.041, 1.394)
Elevation	–	(-0.377, 0.418)	(0.120, 0.845)	(0.034, 0.419)	(-0.071, 0.371)	(-0.534, 0.291)
Precipitation	–	(-0.501, 0.420)	(0.199, 1.016)	(0.001, 0.444)	(-0.072, 0.444)	(-0.392, 0.546)
Crossings	–	(-1.230, 0.145)	(-0.882, 0.342)	(-0.428, 0.217)	(-0.609, 0.142)	(-1.079, 0.283)

*In this analysis, the CG feeding type was used as the reference category, therefore, the interaction coefficients are set to zero for all covariates.

Table 3.4: HPD intervals for the off-diagonal elements of the inverse covariance matrix $\Psi_\emptyset$.

	Alkalinity	% Barren	Area	Elevation	Precipitation	Crossings
% Wood	(0.007, 0.243)	(-0.005, 0.001)	(-0.048, 0.130)	(-0.029, 0.076)	(-0.059, 0.256)	(-0.062, 0.087)
Alkalinity		(-0.060, 0.148)	(-2.867, 2.339)	(-0.976, 2.123)	(6.141, 16.300)	(-3.077, 1.379)
% Barren			(-0.103, 0.054)	(-0.101, -0.006)	(-0.164, 0.111)	(-0.057, 0.076)
Area				(-1.086, 1.274)	(-3.293, 3.738)	(-9.191, -4.928)
Elevation					(1.091, 5.451)	(-0.129, 1.911)
Precipitation						(-1.828, 4.123)

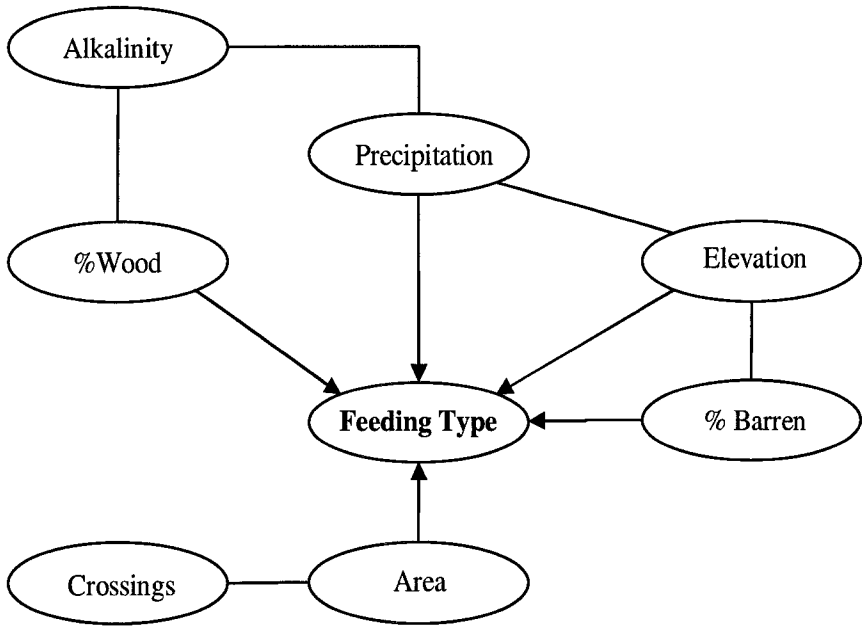

Figure 3.3: Data suggested chain graph for feeding type composition.

Table 3.5: 95% HPD intervals for feeding type prediction at the five withheld Oregon stream sites. The results are given in percentage form with the LN mode point estimates for each compositional elements located outside of the interval parentheses. The observed count percentage for each category is given in the second row for each site.

	Feeding Type[a]					
Site[c]	CG	CF	GZ	EP	SH	OT
454	78.0 (32.8, 93.0)	2.9 (0.0[b], 28.8)	5.1 (0.0[b], 33.9)	5.9 (0.6, 15.7)	7.6 (0.2, 26.2)	0.5 (0.0[b], 5.6)
	73.6	2.7	5.0	3.6	15.1	0.0
456	78.1 (32.1, 90.1)	2.9 (0.0[b], 22.8)	5.0 (0.0[b], 27.2)	6.0 (0.9, 19.2)	7.7 (0.6, 37.6)	0.4 (0.0[b], 3.4)
	74.0	2.1	2.5	4.0	16.9	0.6
457	75.3 (38.4, 92.9)	3.1 (0.0[b], 24.1)	5.6 (0.0[b], 19.1)	6.6 (0.9, 19.7)	9.0 (0.4, 29.3)	0.5 (0.0[b], 4.1)
	84.0	0.0	0.9	4.7	6.9	3.6
461	75.5 (36.6, 91.7)	3.9 (0.0[b], 15.3)	4.9 (0.0[b], 23.9)	6.0 (1.1, 22.9)	9.2 (0.7, 34.4)	0.4 (0.0[b], 2.9)
	51.1	3.1	13.3	9.6	22.0	1.0
465	78.4 (51.1, 96.1)	2.7 (0.0[b], 15.5)	5.2 (0.0[b], 8.4)	5.8 (0.6, 18.7)	7.4 (0.2, 24.1)	0.6 (0.0[b], 3.5)
	91.6	0.9	0.9	3.3	3.1	0.1

[a] Notation for feeding type categories is given in Table 3.1

[b] HPD intervals shown with a lower bound of zero have been rounded to three decimal places. Actual values are greater than zero.

[c] The full REMAP names for these site are preceded by ORST97.

Chapter 4

STATE-SPACE MODELS FOR THE ANALYSIS OF MULTI-WAY DISCRETE COMPOSITIONAL DATA

4.1 Introduction

In this chapter, we introduce an extension of the state-space model used previously for analysis of a single composition variable to a model suitable for the analysis of multi-way compositions. As with the "univariate" discrete compositional data of Chapter 3, multi-way discrete compositional data arise from multivariate counts instead of a continuous multivariate vector. For multi-way compositional data sampled individuals are classified according to two or more different categorical variables. Therefore, with multi-way discrete compositional data, one is interested in the proportion of counts of a particular category of possible *joint* outcomes, or equivalently, the probability that a randomly selected individual will belong to a certain cross-classification, or *cell*.

Another interesting question that can be asked with multi-way discrete compositional data concerns the independence structure of the discrete variables used for cross-classification. For example, if we have a two-way cross-classification of individuals according to the discrete variables I and J, as with a single two-way contingency table we are often interested in whether the classification of a random individual to category i of I is independent of the event that the individual is classified to category j of J. In other words, we may be interested in whether separate probabilities should be modeled for each cell, or whether a model of the form $p_{ij} = p_i p_j$ would

be more appropriate, where p_{ij}=Pr[indiv. in cell (i,j)] and p_i (p_j)=Pr[indiv. in category i (j) of I (J)]. Even if independence of the classification variables is not a primary research concern, it may still be appropriate to include some independence structure to reduce the number of parameters and provide a more parsimonious model.

As in Chapter 3 we will use the term *site* to index each multi-way discrete compositional observation. If one is interested in using a fully saturated model (complete dependence among all classification variables) to estimate the true, unobserved cell proportions or probabilities at each site, then, one can simply use the state-space model developed in Chapter 3. This is accomplished by simply modeling all of the cells as one large single composition. This is the approach used by Dominici (2000) to combine several contingency tables, some with missing dimensions. In fact, the saturated version of the model presented here can be re-parameterized to give the model of Chapter 3. However, if one is interested in introducing some independence structure to the cell probabilities, then an extension for the state-space model of Chapter 3 must be constructed.

Aitchison (1986, pg. 324) introduces the analysis of independence for two-way compositional data through a transformation of the composition similar to the log-ratio transform (1.2). A Hotelling's T test is then be performed on the mean of the transformed compositions to test for independence. Aitchison (1986, pg. 326) also comments that more complicated forms of dependence have a parametric expression in the covariance matrix of the logistic-normal (LN) distribution. Aitchison introduces the notion that the LN model essentially represents a "random effects" formulation of traditional categorical data analysis.

One major concern with Aitchison's approach is that one is testing whether the "average" composition possesses certain independence constraints. Each realized composition has zero probability of actually possessing these independence

constraints. Therefore, we propose a random effects formulation of the Discrete Regression (DR) model of Section 2.4. Our random effects formulation will allow independency to be incorporated into a multi-way model through a graphical model interpretation. This results in a model in which each realized composition will follow any specified independence constraint. In addition, with the inclusion of covariates for each site, the random effects extension will also possess Markov properties with respect to a chain graph of the covariates, random effects, and classification variables.

In order to illustrate the graphical model approach for analysis of discrete multi-way compositional data, we use a graphical model to analyze data concerning fish species richness in the Mid-Atlantic Highlands region of the United States. The U.S. Environmental Protection Agency (EPA) measures fish species richness as part of the stream surveys conducted for the Environmental Monitoring and Assessment Program (EMAP). The species of fish are categorized according to several traits. Two very important traits for assessing stream impairment are habit and pollution tolerance (McCormack et al., 2001). The habit classification has two levels, *benthic* and *column.* Benthic species comprise those species living at, or near, the stream bottom, and column species inhabit the vertical column of water. The pollution tolerance classification has three levels, *intolerant*, *intermediate*, and *tolerant.* In this chapter, we examine the multi-way composition of species habit and pollution tolerance and its relation to several disturbance variables as well as some climate oriented variables.

4.2 Model Formulation

4.2.1 Models for a Single Individual at a Single Site

In order to begin development of the multi-way composition models some notation is needed. As opposed to the graphical models developed in Chapter 3, the

discrete response variable is now truly a multivariate response. The notation Φ will denote the set of categorical *response variables* of which there are D possible cross-classifications on the product space of the response variable levels. We use the notation $\mathbf{y}_\Phi$ to index a specific cell in the product space of response variable levels. The index $\mathbf{y}_\Phi$ can be thought of as a vector $\{y_\phi : \phi \in \Phi\}$ in which each element y_ϕ corresponds to an element in Φ and takes integer values from one to the maximum number of categories for the discrete variable Y_ϕ.

To begin the multi-way composition model formulation, we first consider a model for a single individual at a randomly selected site. For each site, we are interested in the dependence relationships between covariates that will be measured at the site and the event that a randomly selected individual is cross-classified into one of the D cells $\mathbf{y}_\Phi$. In the analysis of fish species richness of Section 4.5, we are interested in the event that a randomly selected fish species belongs to a certain cross-classification of life-history traits. Let $\mathbf{Y}_\Phi = \{Y_\phi : \phi \in \Phi\}$ denote the response vector for a single individual, which takes one of the D vectors, say $\mathbf{y}_\Phi$. In addition, let $\mathbf{X} = (X_1, \ldots, X_p)$ denote a vector of observable covariates at the randomly selected site. The vector $\mathbf{X}$ can also be written $(\mathbf{X}_\Gamma, \mathbf{X}_\Delta)$, where Γ indexes the continuous covariates and Δ indexes the discrete covariates. In terms of the vertices of a chain graph, we have a chain graph with terminal chain components composed of those variables in Φ and an anterior set composed of the p explanatory variables.

We can now use the Discrete Regression (DR) model of Section 2.4 to specify a joint density for $(\mathbf{Y}_\Phi, \mathbf{X})$ as $f(\mathbf{y}_\Phi|\mathbf{x})f_{CG}(\mathbf{x})$, where

$$f(\mathbf{y}_\Phi|\mathbf{x}) = \exp\left\{\alpha_\Phi(\mathbf{x}) + \sum_{f\subseteq\Phi}\sum_{c\subseteq\Gamma}\sum_{d\subseteq\Delta}\beta_{fcd}(\mathbf{y}_\Phi, \mathbf{x}_\Delta)\prod_{\gamma\in c}x_\gamma + \sum_{\gamma\in\Gamma}\sum_{d\subseteq\Delta}\sum_{m=2}^{M}\omega_{\gamma dm}(\mathbf{y}_\Phi, \mathbf{x}_\Delta)x_\gamma^m\right\} \tag{4.1}$$

and

$$f_{CG}(\mathbf{x}) = \exp\left[\sum_{d\subseteq\Delta}\left\{\lambda_d(\mathbf{x}_\Delta) + \boldsymbol{\eta}_d(\mathbf{x}_\Delta)'\mathbf{x}_\Gamma - \frac{1}{2}\mathbf{x}_\Gamma'\boldsymbol{\Psi}_d(\mathbf{x}_\Delta)\mathbf{x}_\Gamma\right\}\right]. \tag{4.2}$$

In (4.1), $\alpha_\Phi(\mathbf{x})$ is a normalizing constant with respect to $\mathbf{Y}_\Phi|\mathbf{x}$ and $\beta_{fcd}(\mathbf{y}_\Phi, \mathbf{x}_\Delta)$ and $\omega_{\gamma dm}(\mathbf{y}_\Phi, \mathbf{x}_\Delta)$ are interaction terms which depend on $\mathbf{y}_\Phi$ and $\mathbf{x}_\Delta$ only through the variables associated with the sets $f \subseteq \Phi$ and $d \subseteq \Delta$, respectively. In the response portion of the model (4.1), interaction terms for which $f = \emptyset$ can, without loss of generality, be set to zero (e.g. $\beta_{\emptyset cd} \equiv 0$ for any $c \subseteq \Gamma$ and $d \subseteq \Delta$) as they do not depend on $\mathbf{y}_\Phi$ and will cancel with the normalizing term $\alpha_\Phi(\mathbf{x})$. The CG density (4.2) is given in the matrix form in order to facilitate re-parameterization in our proposed estimation procedure described in Section 4.4. As with the interaction terms in the response model, $\lambda_d(\mathbf{x}_\Delta)$, $\boldsymbol{\eta}_d(\mathbf{x}_\Delta)$, and $\boldsymbol{\Psi}_d(\mathbf{x}_\Delta)$ depend on $\mathbf{x}_\Delta$ only through the subset of variables associated with the set $d \subseteq \Delta$.

To complete the DR model we must impose some constraints to ensure identifiability of the model parameters. To accomplish this, first select a reference cell of $\mathbf{Y}$, say $\mathbf{y}^*_\Phi$, and a reference cell for the categorical covariates, say $\mathbf{x}^*_\Delta$. Without loss of generality, henceforth, we assume that $\mathbf{y}^*_\Phi$ and $\mathbf{x}^*_\Delta$ are appropriately sized vectors of ones, indicating the reference cells are those indexed by the first level of all the variables associated with Φ and Δ. Now that the reference cells are defined, set all interaction terms in (4.1) and (4.2) equal to zero if $y_\phi = 1$ for any $\phi \in f$ or $x_\delta = 1$ for any $\delta \in d$. These zero constraints are analogous to the zero constraints of interaction terms in classic ANOVA models. By using these constraints we can interpret the interaction terms as measuring interactions relative to the selected values $\mathbf{y}^*_\Phi$ and $\mathbf{x}^*_\Delta$. For example, given any response variable $\phi \in \Phi$ and any two covariates $\gamma \in \Gamma$, and $\delta \in \Delta$, a positive value for the interaction term $\beta_{\phi\gamma\delta}(\mathbf{y}_\Phi, \mathbf{x}_\Delta)x_\gamma$ implies that an increase in x_γ increases the probability that a randomly selected individual will be cross-classified according to a cell where $Y_\phi = y_\phi$ over a cell for which $Y_\phi = 1$ and the amount of increase depends on the categorical covariate X_δ.

For the remainder of this chapter, we will consider only the homogeneous CG distribution, where $\boldsymbol{\Psi}_d(\mathbf{x}_\Delta) = \mathbf{0}$ for $d \neq \emptyset$. This restriction is identical to the

assumption that the covariance matrix of the continuous variables is constant over all of the cells. The model can be extended, however, to be non-homogeneous if desired.

The CG distribution (4.2) is based on the premise that, marginally, the categorical covariates follow a log-linear model, while conditioned on the categorical variables, the continuous variables follow a Multivariate Normal (MVN) distribution with mean and variance determined by the realized categorical variables. Philosophically, in terms of a graphical chain model, this implies that the categorical covariates precede the continuous ones in "causal ordering". If it is theoretically more plausible to model the continuous variables as "causing" or influencing the categorical variables, then one could reverse the order of conditioning by modeling the joint distribution of the covariates as a DR distribution where the categorical covariates take the "response" role and the continuous covariates have a MVN distribution. This leads to the "Iterated Discrete Regression" (IDR) model

$$\begin{aligned} f_{IDR}(\mathbf{x}) &= f(\mathbf{x}_\Delta|\mathbf{x}_\Gamma)f(\mathbf{x}_\Gamma) \\ &= \exp\left\{\alpha_\Delta(\mathbf{x}_\Gamma) + \sum_{c\subseteq\Gamma}\sum_{d\subseteq\Delta}\zeta_{cd}(\mathbf{x}_\Delta)\prod_{\gamma\in c}x_\gamma \right. \\ &\qquad \left. + \sum_{\gamma\in\Gamma}\sum_{d\subseteq\Delta}\sum_{m=2}^{M}\nu_{\gamma dm}(\mathbf{x}_\Delta)x_\gamma^m\right\} \\ &\quad \times f_N(\mathbf{x}_\Gamma;\ \boldsymbol{\mu}_\Gamma,\ \boldsymbol{\Sigma}_\Gamma), \end{aligned} \tag{4.3}$$

where $f_N(\cdot;\boldsymbol{\mu},\ \boldsymbol{\Sigma})$ represents a MVN density with mean $\boldsymbol{\mu}$ and covariance $\boldsymbol{\Sigma}$. Interaction terms are set to zero as before to ensure an identifiable model. The joint density of $(\mathbf{Y}_\Phi, \mathbf{X})$ can be written as $f(\mathbf{y}_\Phi|\mathbf{x})f_{IDR}(\mathbf{x})$.

4.2.2 Single Site Models

Equations (4.1) and (4.2) describe a graphical chain model for sampling a single individual at a randomly selected site. Now, we need to extend this model to account for repeated sampling of individuals at one site. In order to construct a chain graph

to depict dependence relationships we are interested in inference for the interaction parameters of the model for one individual, (4.1) and (4.2); however, sampling multiple individuals is necessary to estimate the interaction terms. Therefore, we will now need to construct a likelihood model for sampling N individuals at a site if (4.1) and (4.2) represent the joint density of an individual classified to cell $\mathbf{y}_\Phi$ and covariates observed at the site where the individual is observed.

To construct a model for repeated sampling at a single site, we will first condition on the realization of a site, which amounts to conditioning on the covariates. Now, sampling N individuals at a site provides N realizations of the variable $\mathbf{Y}_\Phi$. These N realizations can be summarized into a D vector of counts $\mathbf{c} = \{c(\mathbf{y}_\Phi)\}$, where $c(\mathbf{y}_\Phi)$ represents the number of individuals that were cross-classified into cell $\mathbf{y}_\Phi$. The count vector $\mathbf{c}$ represents a complete and sufficient summarization of the N individual responses, so we can model the counts in order to make inference to (4.1). We model the count vector $\mathbf{c}$ with a multinomial distribution. For a fixed site sample size N, the joint distribution for the counts and covariates is

$$\begin{aligned} f(\mathbf{c},\mathbf{x}) &= f_M(\mathbf{c}|\mathbf{x}) f_{CG}(\mathbf{x}) \\ &= \frac{N!}{\prod_{\mathbf{y}_\Phi} c(\mathbf{y}_\Phi)!} \left\{ \prod_{\mathbf{y}_\Phi} f(\mathbf{y}_\Phi|\mathbf{x})^{c(\mathbf{y}_\Phi)} \right\} \times f_{CG}(\mathbf{x}), \end{aligned} \tag{4.4}$$

where $f(\mathbf{y}_\Phi|\mathbf{x})$ is given by (4.1). If, however, sampling of individuals is carried out in a way where N is random and could be modeled with a Poisson distribution with mean κ, then each category count will have independent Poisson distributions with mean $\kappa f(\mathbf{y}_\Phi|\mathbf{x})$ (Rohatgi, 1976, pg. 200). For a Poisson(κ) random sample size, the joint density of $\mathbf{c}$ and $\mathbf{x}$ is given by

$$\begin{aligned} f(\mathbf{c},\mathbf{x}) &= f_P(\mathbf{c}|\mathbf{x}) f_{CG}(\mathbf{x}) \\ &= \left[\prod_{\mathbf{y}_\Phi} \frac{\{\kappa f(\mathbf{y}_\Phi|\mathbf{x})\}^{c(\mathbf{y}_\Phi)} e^{\kappa f(\mathbf{y}_\Phi|\mathbf{x})}}{c(\mathbf{y}_\Phi)!} \right] \times f_{CG}(\mathbf{x}). \end{aligned} \tag{4.5}$$

If it is desired, the IDR model could be used for the covariate distribution in (4.5) instead of the CG model.

The joint models in (4.4) and (4.5) for counts and covariates are different than the standard sampling scheme for a mixed variable graphical model. Usually, every individual sampled generates an multivariate observation of categorical and continuous variables. Here, however, there is only one observation of the covariate vector for all of the individuals observed at a particular site. The present sampling scheme is analogous to replication of an experiment at the same factor levels at each site.

The multinomial model (4.4) is quite self-explanatory in terms of what we would like to model. We are interested in modeling the true composition of individuals at a site. This composition is represented by the D probabilities $f(\mathbf{y}_\Phi|\mathbf{x})$. In the Poisson model, however, another parameter, κ, has been added. This parameter is of no interest as far as determining conditional dependence relationships, but, it has to be estimated well for each site in order to have a model that fits the data adequately. Therefore, we propose two parsimonious modifications to the Poisson model that will prove useful when many sites are sampled.

First, the total number of individuals observed at a site may depend on the same covariates which are being used in the graphical model. So, we can model the mean number of total individuals observed, κ, using the linear model

$$\log \kappa = \sum_{c\subseteq\Gamma}\sum_{d\subseteq\Delta} \beta_{cd}(\mathbf{x}_\Delta)\prod_{\gamma\in c} x_\Gamma + \sum_{\gamma\in\Gamma}\sum_{d\subseteq\Delta}\sum_{m=2}^{M}\omega_{\gamma dm}(\mathbf{x}_\Delta)x_\gamma^m. \tag{4.6}$$

This produces a log-mean model for the cell count $c(\mathbf{y}_\Phi)$ given by

$$\begin{aligned}\log\{\kappa f(\mathbf{y}_\Phi|\mathbf{x})\} = &\sum_{f\subseteq\Phi}\sum_{c\subseteq\Gamma}\sum_{d\subseteq\Delta}\beta_{fcd}(\mathbf{y}_\Phi,\mathbf{x}_\Delta)\prod_{\gamma\in c}x_\gamma \\ &+\sum_{f\subseteq\Phi}\sum_{\gamma\in\Gamma}\sum_{d\subseteq\Delta}\sum_{m=2}^{M}\omega_{f\gamma dm}(y_j,\mathbf{x}_\Delta)x_\gamma^m.\end{aligned} \tag{4.7}$$

Once again, the interaction terms depend on $\mathbf{y}_\Phi$ and $\mathbf{x}_\Delta$ only though the subset of variables in f and d respectively. If $f=\emptyset$ then the interaction term does not depend on the cell $\mathbf{y}_\Phi$, but it is not necessarily assumed to be zero as first defined for (4.1). In order to ensure identifiability, again, interaction terms are set to zero if $y_\phi = 1$ for

any $\phi \in f$ or $x_\delta = 1$ for any δ in d. The normalization constant $\alpha_\Phi(\mathbf{x})$ was dropped because it is independent of the response categories and can therefore be absorbed into the parameters for the κ model. A similar approach is used for contingency table log-linear models when a random sample size is assumed (Christensen, 1990).

Another approach to modeling the κ parameter is to use covariates related to the sampling protocol. For example, covariates such as sampling effort at each site may be used to model the total site counts. These types of variables are usually not scientifically interesting as far as modeling compositions, however, they might prove valuable for modeling the total number of individuals observed at a site. This type of Poisson model is not used in graphical log-linear modeling due to the fact that there is only one sample of individuals. Compositional data will have multiple samples that might be modeled using external sample design covariates.

4.2.3 Random Effects Discrete Regression

In Section 4.2.2 we describe a model for a single randomly sampled site. Now, we will extend this model to account for possibly hundreds of randomly selected sites. For each site, a separate graphical model could be constructed, but this would increase the number of parameters to be estimated to an unmanageable level. In addition, the differences in non-zero parameter values are not of primary interest in determining Markov relationships for a graphical model. Therefore, we propose a global graphical model for all sites that allows site-to-site flexibility in some of the non-zero parameter values. With the added flexibility, the model can adjust to fit the data more adequately. In order to add this flexibility, as well as model the randomness in site selection, we introduce a random error term to the response model (4.1).

The addition of a random effect to the response model (4.1) produces a full model for $\mathbf{Y}_\Phi$, $\mathbf{X}$, and the random effects $\boldsymbol{\epsilon}$ of the form

$$f(\mathbf{y}_\Phi, \mathbf{x}, \boldsymbol{\epsilon}) = f_{RE}(\mathbf{y}_\Phi|\mathbf{x}, \boldsymbol{\epsilon}) f_{CG}(\mathbf{x}) f(\boldsymbol{\epsilon}). \tag{4.8}$$

Since there is only one observation of the explanatory variables we will leave the model for the covariates, $f_{CG}(\mathbf{x})$, as it is given in (4.2). The response portion $f_{RE}(\mathbf{y}_\Phi|\mathbf{x},\boldsymbol{\epsilon})$ of the random effects Discrete Regression (REDR) model is modified by the addition of a random intercept term to give,

$$f_{RE}(\mathbf{y}_\Phi|\mathbf{x},\boldsymbol{\epsilon}) = \exp\left\{\alpha_\Phi(\mathbf{x}) + \sum_{f\subseteq\Phi}\sum_{c\subseteq\Gamma}\sum_{d\subseteq\Delta}\beta_{fcd}(\mathbf{y}_\Phi,\mathbf{x}_\Delta)\prod_{\gamma\in c}x_\gamma + \sum_{f\subseteq\Phi}\sum_{\gamma\in\Gamma}\sum_{d\subseteq\Delta}\sum_{m=2}^{M}\omega_{f\gamma dm}(\mathbf{y}_\Phi,\mathbf{x}_\Delta)x_\gamma^m + \sum_{f\subseteq\Phi}\epsilon_f(\mathbf{y}_\Phi)\right\}, \tag{4.9}$$

where $\epsilon_f(y_\Phi) = 0$, if $y_\phi = 1$ for any $\phi \subseteq f$, to ensure identifiability. In order to allow modeling of a given independence structure for the multi-way response, we also introduce one other constraint on the random effects. If f is not complete (see Section 2.1) in the graphical representation of the desired independence structure, then all of the random effect terms in $\epsilon_f(\mathbf{y}_\Phi)$ are defined to be 0 for all cells $\mathbf{y}_\Phi$. The remaining random interactions $\boldsymbol{\epsilon}_f = \{\epsilon_f(\mathbf{y}_\Phi) : y_\phi \neq 1 \text{ for any } \phi \subseteq f\}$, for f complete in the graphical representation of the desired independence structure, are given a multivariate distribution with mean $\mathbf{0}$ and covariance (or scale parameter) $\boldsymbol{\Sigma}_f$. For now, we will consider each of the random effects vectors as being independently distributed. In Section 4.3.3, however, we will show that this restriction can be relaxed to some degree.

The introduction of random error terms in the manner given in (4.9) has three benefits. First, the model can adjust for site-to-site variability. Secondly, the model will automatically add some level of overdispersion to cell counts. Finally, every realization of the random effects provides cell probabilities that maintain the desired independence relationships among the response variable, as well as between the response variables and the site covariates. To see this one can simply calculate the Hammersley-Clifford interaction terms described in the proof of Theorem 2.2 and illustrated for DR models in the proof of Proposition 2.7. For each realization

of the random effects in (4.9), all of the interaction terms will remain the same as those calculated in Proposition 2.7, except $\phi_{f\cup\emptyset\cup\emptyset}(\mathbf{y}_\Phi, \mathbf{x}_\Delta) = \beta_{f\emptyset\emptyset}(\mathbf{y}_\Phi) + \epsilon_f(\mathbf{y}_\Phi)$. Therefore, since $\epsilon_f(\mathbf{y}_\Phi)$ is set to zero for sets f that are not complete in the graphical representation of the desired independence structure, the cell probabilities will factor according to that structure with probability 1. This provides an improvement over the approach of Aitchison (1986), since each site can be fit with a model that possesses a given independence structure, one is not just examining the "average" composition independence structure. Using the Aitchison approach, a fully dependent structure could best fit the data at each site, while the "average" model may provide some evidence for independence, giving a misleading inference.

In the REDR model description we have left the error distribution vague. We believe that different situations may necessitate different error structures. If it is reasonable to assume that the error structure is symmetric with few outliers, then a MVN distribution may be reasonable. In this case, the cell compositions will have a LN distribution. However, other distributions could be used. For example a multivariate t distribution with k degrees of freedom could be used if it is desirable to have an error with heavier tails. It may be desirable to use the t errors instead of normal errors if there is a high level of overdispersion in the cell counts. For the remaining discussion of the random effects DR model we will assume a MVN distribution for the random effects (i.e., $f(\boldsymbol{\epsilon}_f) = f_N(\boldsymbol{\epsilon}_f;\ \mathbf{0}, \boldsymbol{\Sigma}_f)$); however, the theoretical results will remain the same for the t error model.

Now that we have added random effects to the response portion of the model, the likelihood for the response variable cell counts given the covariates $\mathbf{x}$ changes slightly from that given in (4.4) and (4.5). The likelihood model for the response variable cell counts $\mathbf{c}$ given the covariates $\mathbf{x}$, the total number of individuals observed at a site, and the random effects is

$$f_M(\mathbf{c}|\mathbf{x}, \boldsymbol{\epsilon}) = \frac{N!}{\prod_{\mathbf{y}_\Phi} c(\mathbf{y}_\Phi)!} \prod_{\mathbf{y}_\Phi} f_{RE}(\mathbf{y}_\Phi|\mathbf{x}, \boldsymbol{\epsilon})^{c(\mathbf{y}_\Phi)}, \tag{4.10}$$

where $f_{RE}(\mathbf{y}_\Phi|\mathbf{x},\epsilon)$ is given by (4.9). Similarly, for a random total number of individuals, the Poisson model including random effects simply replaces the DR response $f(\mathbf{y}_\Phi|\mathbf{x})$ with the REDR response model $f_{RE}(\mathbf{y}_\Phi|\mathbf{x},\epsilon)$ in (4.5).

Now, we focus on the multiple site likelihood for the explanatory variables. Assuming that the covariate observations are independently distributed and follow a homogeneous CG distribution, we obtain the multiple site explanatory density

$$f(\mathbf{x}_1,\ldots,\mathbf{x}_S) = \prod_{i=1}^{S} f_{CG}(\mathbf{x}_i|\boldsymbol{\lambda},\boldsymbol{\eta},\boldsymbol{\Psi}_\emptyset), \tag{4.11}$$

where $\mathbf{x}_i$ denotes the set of observed covariates for $i = 1,\ldots,S$ and $\boldsymbol{\lambda}$ and $\boldsymbol{\eta}$ represent the collected parameter sets $\{\lambda_d(\mathbf{x}_\Delta) : d \subseteq \Delta\}$ and $\{\boldsymbol{\eta}_d(\mathbf{x}_\Delta) : d \subseteq \Delta\}$. Extensions to this model, such as accounting for correlation of covariates over space, are discussed in Section 6.2.2.

We now re-parameterize the homogeneous CG density in (4.11) into a more useful form. First, we break the CG density into a marginal model for the categorical components of the explanatory variable set and a conditional model for the continuous components. We then re-parameterize the conditional Gaussian distribution into an ANOVA like form. This re-parameterization gives the following form for the homogeneous CG density,

$$\begin{aligned} f_{CG}(\mathbf{x}) &= f(\mathbf{x}_\Delta) f(\mathbf{x}_\Gamma|\mathbf{x}_\Delta) \\ &= \exp\left\{\sum_{d\subseteq\Delta} \lambda_d(\mathbf{x}_\Delta)\right\} \times \frac{1}{\sqrt{2\pi}}|\boldsymbol{\Psi}_\emptyset|^{1/2} \\ &\quad \times \exp\left\{\frac{1}{2}\left(\mathbf{x}_\Gamma - \sum_{d\subseteq\Delta}\boldsymbol{\tau}_d(\mathbf{x}_\Delta)\right)' \boldsymbol{\Psi}_\emptyset \left(\mathbf{x}_\Gamma - \sum_{d\subseteq\Delta}\boldsymbol{\tau}_d(\mathbf{x}_\Delta)\right)\right\}, \end{aligned} \tag{4.12}$$

where $\boldsymbol{\Psi}_\emptyset$ represents the inverse covariance matrix for the continuous variables, which have a MVN distribution, $\boldsymbol{\tau}_d(\mathbf{x}_\Delta) = \boldsymbol{\Psi}_\emptyset^{-1}\boldsymbol{\eta}_d(\mathbf{x}_\Delta)$, and $\lambda_\emptyset(\mathbf{x}_\Delta)$ represents a normalizing constant in the log-linear model for $\mathbf{x}_\Delta$.

Define the vector of cell counts $\mathbf{c}_\Delta = [c(\mathbf{x}_\Delta)]$, where $c(\mathbf{x}_\Delta)$ is the number of sites for which the categorical covariates $\mathbf{X}_\Delta = \mathbf{x}_\Delta$. Using the re-parameterization

of the CG density in (4.12) we can write the joint density of the covariates over all sites (4.11) as

$$\begin{aligned} f(\mathbf{x}_1, \ldots, \mathbf{x}_S) &= \left\{ \prod_{i=1}^{S} f(\mathbf{x}_{\Delta i}) \right\} \times \left\{ \prod_{i=1}^{S} f_N(\mathbf{x}_{\Gamma i} | \mathbf{x}_{\Delta i}) \right\} \\ &= \left(\frac{S!}{\prod_{\mathbf{x}_\Delta} c(\mathbf{x}_\Delta)!} \right)^{-1} f_M(\mathbf{c}_\Delta | \boldsymbol{\lambda}) \\ &\quad \times \prod_{i=1}^{S} f_N \left(\mathbf{x}_{\Gamma i};\ \sum_{d \subseteq \Delta} \tau_d(\mathbf{x}_{\Delta i}),\ \boldsymbol{\Psi}_\emptyset \right), \end{aligned} \tag{4.13}$$

where the explanatory observation at the ith site is given by $\mathbf{x}_i = (\mathbf{x}_{\Delta i}, \mathbf{x}_{\Gamma i})$ and $f_M(\mathbf{c}_\Delta | \boldsymbol{\lambda})$ is the multinomial density

$$f_M(\mathbf{c}_\Delta | \boldsymbol{\lambda}) = \frac{S!}{\prod_{\mathbf{x}_\Delta} c(\mathbf{x}_\Delta)!} \prod_{\mathbf{x}_\Delta} \exp \left\{ \sum_{d \subseteq \Delta} \lambda_d(\mathbf{x}_\Delta) \right\}^{c(\mathbf{x}_\Delta)}. \tag{4.14}$$

The full likelihood for parameter estimation in the REDR model (4.8) is obtained by combining the likelihood for response variable cell counts at each site (4.10), the random effects density, and the explanatory variable likelihood (4.11).

$$\begin{aligned} f\left(\{\mathbf{c}_i\}, \{\mathbf{x}_i\}, \{\boldsymbol{\epsilon}_i\}\right) &= \prod_{i=1}^{S} f_M(\mathbf{c}_i | \mathbf{x}_i, \boldsymbol{\epsilon}_i) f_{CG}(\mathbf{x}_i) f_N(\boldsymbol{\epsilon}_i) \\ &= \prod_{i=1}^{S} f_M(\mathbf{c}_i | \mathbf{x}_i) \times K(\mathbf{c}_\Delta) f_M(\mathbf{c}_\Delta | \boldsymbol{\lambda}) \times \prod_{i=1}^{S} f_N \left(\mathbf{x}_{\Gamma i};\ \sum_{d \subseteq \Delta} \tau_d(\mathbf{x}_{\Delta i}),\ \boldsymbol{\Psi}_\emptyset \right) \\ &\quad \times \prod_{i=1}^{S} \prod_{f \subseteq \Phi} f_N \left(\boldsymbol{\epsilon}_{f,i};\ \mathbf{0}, \boldsymbol{\Sigma}_f \right), \end{aligned} \tag{4.15}$$

where $\mathbf{c}_i$ is a D vector of response variable cell counts for site i, $\mathbf{x}_i$ is a vector of observed covariates, c_Δ is a vector of cell counts for the categorical covariates, $K(\mathbf{c}_\Delta)$ is the inverse of the multinomial coefficient in $f_M(\mathbf{c}_\Delta | \boldsymbol{\lambda})$, and $\boldsymbol{\epsilon}_i$ represents the collection of random effects vectors $\{\boldsymbol{\epsilon}_{f,i} : f \subseteq \Phi \text{ and } f \text{ complete}\}$ for the ith site. If the number of individuals sampled at a site should be considered random, then the multinomial distribution $f_M(\mathbf{c}_i | \mathbf{x}_i)$ could be replaced with the Poisson model as

in (4.5). It should be noted that all of the random effects vectors for a given site $\epsilon_{f,i}$ are modeled as independent random variables in (4.15). This can be relaxed to some degree and will be discussed in Section 4.3.3.

4.3 Markov Properties of Multi-Way Composition Models

Now that we have defined the Random Effects DR (REDR) model (4.8), it is of interest to know what conditions determine the Markov properties of this distribution. In addition, it is also of interest to determine how these properties change when the distribution is marginalized over the random effects.

So, we begin with the first question. One can observe from (4.9) that the random effects have exactly the same mathematical effect on the response as the interaction terms of the observed covariates. Therefore, we define an *extended chain graph* $\mathcal{G}^{\epsilon}$ where the random effects are included as parents of the response variables and are marginally independent of the observed covariates. The graph is represented by a set of vertices for the observable covariates and a set of vertices for the random effects and a set of vertices for the response variables. There are directed edges from observable covariates to the response. There are undirected edges between the response variables as well, depending on their conditional dependence structure. Also, for each $f \in \Phi$ there is a directed edge from ϵ_f to y_ϕ for every $\phi \in f$. Since the random effects are marginally independent of the observed covariates, there are no edges between the covariates and the random effects. Figure 4.1 gives an illustration of two possible extended graphs for a two variable response vector, (a) one where both components of Φ are connected, and (b) one where the components of Φ are not connected. In addition, for now we will consider each ϵ_f, $f \subseteq \Phi$ to be independently distributed, so, there are no edges between the random effects. As mentioned earlier this will be relaxed to some degree at a later point in Section 4.3.3.

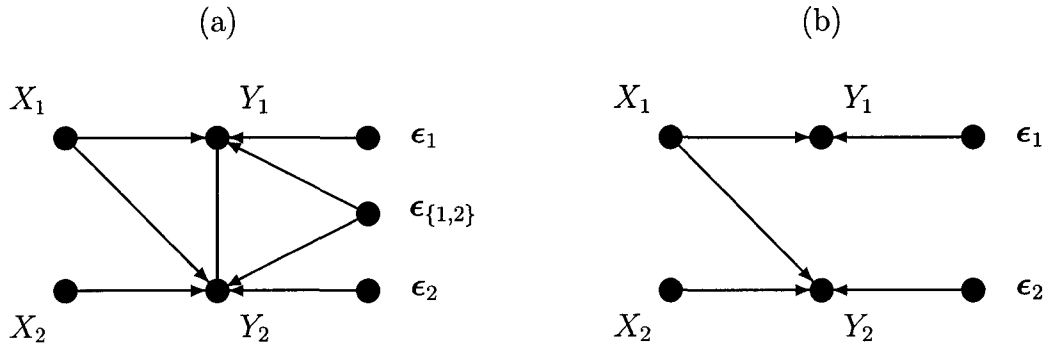

Figure 4.1: Example of an extended graph $\mathcal{G}^\epsilon$ for random effects Discrete Regression models. Here, (a) represents an extended graph for two response vertices that are connected, therefore, ϵ_1, ϵ_2, and $\epsilon_{1,2}$ are all parents of Φ and (b) is an example of an extended graph for vertices in Φ that are unconnected, therefore, only ϵ_1 and ϵ_2 are parents of Φ. In, (b), by definition, $\epsilon_{\{1,2\}}$ is defined to be a zero vector.

We now provide a proposition that describes necessary and sufficient conditions for a REDR distribution to be Markovian with respect to a given extended graph $\mathcal{G}^\epsilon$.

Proposition 4.1. *A REDR distribution P, given by (4.8) is $\mathcal{G}^\epsilon$ Markovian for a given extended chain graph $\mathcal{G}^\epsilon$, if and only if the interaction terms and random effects in (4.8) satisfy the following conditions where $f \subseteq \Phi$, $c \subseteq \Gamma$, and $d \subseteq \Delta$:*

1. *(a) $\beta_{fcd}(\mathbf{y}_\Phi, \mathbf{x}_\Delta) \equiv 0$ in (4.9) unless $f \cup c \cup d$ is complete in $\mathcal{G}^\epsilon$ for c, $d \neq \emptyset$,,*

   (b) $\omega_{f\gamma dm}(\mathbf{y}_\Phi, \mathbf{x}_\Delta) \equiv 0$ in (4.9), for $m = 1, \ldots, M$, unless $f \cup \{\gamma\} \cup d$ is complete in $\mathcal{G}^\epsilon$,

   (c) $\epsilon_f(\mathbf{y}_\Phi) = -\beta_{f\emptyset\emptyset}(\mathbf{y}_\Phi)$ in (4.9), with probability 1, for all cells $\mathbf{y}_\Phi$ if f is not complete in $\mathcal{G}^\epsilon$.

2. *(a) $\lambda_d(\mathbf{x}_\Delta) \equiv 0$ in (4.12) unless d is complete in $\mathcal{G}^\epsilon$,*

   (b) $\tau_{d\gamma}(\mathbf{x}_\Delta) \equiv 0$ unless $d \cup \{\gamma\}$ is complete in $\mathcal{G}^\epsilon$, where $\tau_{d\gamma}(\mathbf{x}_\Delta)$ is the element of the vector $\boldsymbol{\tau}_d(\mathbf{x}_\Delta)$ in (4.12) associated with X_γ and $\gamma \in \Gamma$,

   (c) $\psi_{\{\gamma,\mu\}} \equiv 0$ unless $\{\mu, \gamma\}$ is complete in $\mathcal{G}^\epsilon$, where $\psi_{\{\gamma,\mu\}}$ is the (μ, γ) off-diagonal element of $\boldsymbol{\Psi}_\emptyset$ in (4.12) and $\mu \in \Gamma$.

Proof. To prove Proposition 4.1 we will show that the conditions presented are necessary and sufficient for a REDR model P (4.8) to have Gibbs factorization with respect to $\mathcal{G}^{\epsilon}$ according to Proposition 2.6 and hence show that P is $\mathcal{G}^{\epsilon}$ Markovian.

Upon examination of the second set of conditions 2(a) through 2(c), one can observe that they are necessary and sufficient for factorization of the marginal density of $\mathbf{X}$ and $\boldsymbol{\epsilon}$ on the subgraph $(\mathcal{G}^{\epsilon})_{\{\Gamma\cup\Delta\cup\boldsymbol{\epsilon}\}}$. The second set of conditions are essentially a re-parameterization of the necessary and sufficient factorization criteria for the CG density parameterized by (4.2) (Lauritzen and Wermuth, 1989, see Proposition 2.5). Condition 2(b) reflects the change in parameterization from (4.2) to (4.12). Conditions 2(a) - 2(c) satisfy Proposition 2.6 for the initial chain component $\Gamma \cup \Delta \cup \boldsymbol{\epsilon}$. Since, the random effects $\{\boldsymbol{\epsilon}_f : f \subseteq \Phi\}$ are marginally independent from each other and $\mathbf{X}$ by construction (see Section 4.2.3), $f(\mathbf{x}, \boldsymbol{\epsilon}) = f_{CG}(\mathbf{x}) \prod_{f\subseteq\Phi} f(\boldsymbol{\epsilon}_f)$ factors according to $(\mathcal{G}^{\epsilon})_{\{\Gamma\cup\Delta\cup\boldsymbol{\epsilon}\}}$ if and only if $f_{CG}(\mathbf{x})$ factorizes according to $(\mathcal{G}^{\epsilon})_{\{\Gamma\cup\Delta\}}$.

Now, all we need to show is that $f_{RE}(\mathbf{y}_{\Phi}|\mathbf{x}, \boldsymbol{\epsilon})$ factorizes according to complete sets in $\{(\mathcal{G}^{\epsilon})_{cl(\Phi)}\}^m$ to complete the factorization of the REDR distribution P according to Proposition 2.6. In order to show this we will follow a similar approach as in the proof of Proposition 2.7. First, note that if conditions 1(a) though 1(c) hold, then P is a function only of complete sets in $\{(\mathcal{G}^{\epsilon})_{cl(\Phi)}\}^m$, since the parents of Φ are complete. Now, if we calculate the Hammersley-Clifford interaction terms for the REDR model P in the same manner as in the proof of Proposition 2.7 for the DR model, we see that the interaction terms are identical except for $\phi_{f\cup\emptyset\cup\emptyset}(\mathbf{y}_{\Phi}, \boldsymbol{\epsilon}) = \beta_{f\emptyset\emptyset}(\mathbf{y}_{\Phi}) + \epsilon_f(\mathbf{y}_{\Phi})$. Now, if it is assumed that P is $\mathcal{G}^{\epsilon}$ Markovian, then $\phi_{f\cup\emptyset\cup\emptyset}(\mathbf{y}_{\Phi}, \boldsymbol{\epsilon}) = 0$ for all values of $\mathbf{y}_{\Phi}$ and $\epsilon_f(\mathbf{y}_{\Phi})$. Therefore, $\epsilon_f(\mathbf{y}_{\Phi}) = -\beta_{f\emptyset\emptyset}(\mathbf{y}_{\Phi})$ with probability 1. □

The second question, how do the Markov properties change by marginalizing over the random effects, is a more challenging question due to the fact that the

model form prohibits analytical integration. Unfortunately, independence relationships between the response variables are not generally preserved when marginalizing over all of the random effects. In certain instances, the random effects can act as a mixing distribution. Marginalizing over the random effects has the potential to destroy conditional independencies between response variables. Some model structures, however, are "preservative" in the sense that when one marginalizes over the random effects, the independence relationships between and within the covariates in $\Gamma \cup \Delta$ and responses in Φ are preserved. These model structures are explored further in the next section.

4.3.1 Preservative REDR Models

There is a sizable class of models for which a specified independence structure is guaranteed to be preserved when integrating the multi-way REDR density (4.8) over the random effects in (4.9). We term this class of models *preservative* due to this property. This class of preservative REDR models is defined by the following two conditions,

(1) All connected components a_q, $q = 1, \ldots, Q$, of Φ in $\mathcal{G}^{\epsilon}$ are complete, where Q represents the number of connected components in Φ,

and

(2) Any $\delta \in \Gamma \cup \Delta$ that is a parent of $\phi \in a_q$ is also a parent of every other $\phi \in a_q$, $q = 1, \ldots, Q$.

Formulation of REDR models in this fashion essentially allows the vector variables $\mathbf{Y}_{a_q}$, $q = 1, \ldots, Q$, to function as a single unit when examining independence relationships.

Some useful independence models are members of the class of preservative REDR models. The first, obviously, is the completely saturated model, where all of the members of Φ are connected and complete and any parents of Φ are parents

of every member. This does not mean that all covariates must be parents of every response variable, only that those covariates that *are* parents must be parents to every response variable. Another useful set of models is the completely independent response model. If all responses are conditionally independent of one another, then there is no restriction on the covariates as to whom they must be parents of in order to preserve response conditional independence.

4.3.2 Markov Properties of Preservative REDR Models

Now, we will demonstrate that if the conditions of Proposition 4.1 are satisfied for a REDR model with respect to a preservative extended graph $\mathcal{G}^\epsilon$, then the marginal distribution $(\mathbf{Y}_\Phi,\ \mathbf{X})$ is $\mathcal{G} = (\mathcal{G}^\epsilon)_{V\setminus\epsilon}$ Markovian. Therefore, when interest lies only in the inference of dependence relationships between and within the covariates and response, the random effects can simply be ignored in the graphical representation.

Proposition 4.2. *If P is a preservative REDR model as described in Section 4.3.1, and P is Markovian with respect to the extended graph $\mathcal{G}^\epsilon$, then the marginal distribution of the covariates and responses, $P_{\Phi\cup\Gamma\cup\Delta}$, is $\mathcal{G} = (\mathcal{G}^\epsilon)_{V\setminus\epsilon}$ Markovian.*

Proof. If P is a preservative REDR model and the conditions of Proposition 4.1 are satisfied for an extended chain graph $\mathcal{G}^\epsilon$, then we only need check that the conditional density of the response variable $\mathbf{Y}_\Phi$ factorizes on $(\mathcal{G}_{cl(\Phi)})^m$ according to Theorem 2.3(3) when the density is integrated over the random effects. The conditions are necessary and sufficient for the factorization of the covariate density since the random effects are independent of the covariates. Since the conditions of Proposition 4.1 are satisfied, we can write the conditional density of $(\mathbf{Y}_\Phi|\ \mathbf{X},\ \boldsymbol{\epsilon})$ as

$$\begin{aligned} f_{RE}(\mathbf{y}_\Phi|\ \mathbf{x},\ \boldsymbol{\epsilon}) &= \prod_{f\subseteq\Phi}\prod_{c\subseteq\Gamma}\prod_{d\subseteq\Delta}\psi_{fcd}(\mathbf{y}_\Phi,\mathbf{x},\boldsymbol{\epsilon}_f) \\ &= \prod_{q=1}^{Q}\prod_{p_q\subseteq pa(a_q)}\psi_{a_qp_q}(\mathbf{y}_{a_q},\mathbf{x}_{p_q},\boldsymbol{\epsilon}_{a_q}), \end{aligned} \tag{4.16}$$

where $\psi_{fcd}(\cdot)$ is a Hammersley-Clifford interaction term calculated as in the proof of the Hammersley-Clifford Theorem (Theorem 2.2),

$$\psi_{a_q p_q}(\mathbf{y}_{a_q}, \mathbf{x}_{p_q}, \boldsymbol{\epsilon}_{a_q}) = \prod_{f \subseteq a_q} \psi_{fcd}(\mathbf{y}_\Phi, \mathbf{x}, \boldsymbol{\epsilon}_f), \tag{4.17}$$

and $pa(a_q) \subseteq \Gamma \cup \Delta$ represents the set of covariates that are parents of the response variables $\mathbf{y}_{a_q}$. The interaction terms $\psi_{a_q p_q}$ reduce to functions of only the parents of a_q since those interaction terms including covariates that are not parents of a_q will be equal to 1.

Now, using the independence of the random effects, we can integrate out all of the random effects to obtain the conditional density of $\mathbf{Y}_\Phi$ given $\mathbf{X}$,

$$\begin{aligned} f(\mathbf{y}_\Phi|\mathbf{x}) &= \prod_{q=1}^{Q} \prod_{p_q \subseteq pa(a_q)} \int_{\boldsymbol{\epsilon}_{a_q}} \psi_{a_q p_q}(\mathbf{y}_{a_q}, \mathbf{x}_{p_q}, \boldsymbol{\epsilon}_{a_q}) f(\boldsymbol{\epsilon}_{a_q}) d\boldsymbol{\epsilon}_{a_q} \\ &= \prod_{q=1}^{Q} \prod_{p_q \subseteq pa(a_q)} h_{a_q p_q}(\mathbf{y}_{a_q}, \mathbf{x}_{p_q}). \end{aligned} \tag{4.18}$$

Since P is a preservative REDR model, every set $a_q \cup p_q$, $q = 1, \ldots, Q$, is complete in the graph $\{(\mathcal{G}^\epsilon)_{V \setminus \epsilon}\}^m$. Therefore, the conditional density $f(\mathbf{y}_\Phi|\mathbf{x})$ in (4.18) factorizes on $\{(\mathcal{G}^\epsilon)_{V \setminus \epsilon}\}^m$, which, together with the fact that $f(\mathbf{x})$ factorizes on the subgraph $\mathcal{G}_X$ of the covariate chain component, proves that the joint density factorizes according to Proposition 2.6 and hence, according to Theorem 2.3(3). □

In the situation where the model of interest is not a preservative model, integration over the random effects can still be carried out. Any REDR model is Markovian with respect to a "preservative" graph. All that needs to be done is to create a graph from the graph $\mathcal{G}^\epsilon$ for which the non-preservative model is Markovian by completing all connected response variable components and adding a directed edge from every parent of a connected response component to every other member of that component. Since we are adding edges, the original REDR model will be Markovian with respect to this new supplemental graph, since it will still factorize according to complete vertex sets. We can then proceed as shown in the proof.

4.3.3 Correlated Random Effects and Preservative Models

It is possible to generalize the REDR model to allow some degree of association between the random effects for preservative models. In the proof of Proposition 4.2 we used the fact that the random effects vectors were independent from one another. Upon examination of the proof, however, it can be observed that if elements of $\boldsymbol{\epsilon}_f$ and $\boldsymbol{\epsilon}_{f'}$ are correlated for each f and $f' \subseteq a_q$, then the results will remain the same. We can therefore marginalize over random effects that are correlated, and still preserve the associations between the responses and covariates, if all of the correlation occurs between random effects associated with the same complete response component.

In Chapter 3, we proved that the single composition models are preservative in that one can marginalize over the random effects without destroying the conditional independencies. It is also a preservative model by the definition in Section 4.3.1 since it has only one response. It is tempting to conclude at first that the saturated multi-way model is, aside from a combination of parameters, just a single composition model. This is not, however, true in general. A saturated multi-way model is actually more restrictive than a single composition model applied to all of the cells because it has fewer parameters. Actually, one needs to allow dependence among the random effects to be able to obtain equivalence of the single composition model or all the cells and a saturated multi-way model. Therefore, we can now allow correlation between random effects of connected response components. If the model is a preservative REDR model, the independence relationships will be maintained. In addition, the saturated preservative model will correspond to a single composition model for all of the cells.

4.4 Parameter Inference

In order to make inference about the parameters in the multi-way REDR model (4.8), we adopt a Bayesian approach for parameter estimation. The hierarchical

structure of the multi-way REDR model makes Bayesian procedures particularly attractive. There are a large number of unobserved random effects that may or may not be considered nuisance parameters. If one is interested in dependence relationships between the observable covariates and the response variables, or dependence relationships between the response variables given the covariates, the random effects are usually considered nuisance parameters. As presented in the previous section, these random effects can be marginalized over in some cases without affecting the Markov properties of the joint distribution of the response and covariates. If, however, the unobserved compositions at all, or some, sites are of interest then estimates of the random effects are necessary for each site to calculate an estimate of the true site composition. In some cases, however, interest lies in predicting compositions at sites that are not measured for response, but only covariates are obtained. Estimates of the random effects for that site are necessary for estimation of the site composition. Modern Bayesian computational techniques can handle all of these goals with little modification.

Bayesian inference for graphical composition models proceeds by first defining a prior distribution for the parameters of the model $\pi(\boldsymbol{\beta}, \boldsymbol{\omega}, \boldsymbol{\lambda}, \boldsymbol{\tau}, \boldsymbol{\Psi}_{\emptyset}, \boldsymbol{\Sigma})$. Here, we have removed subscripts in order to ease notational burden. For example, $\boldsymbol{\beta}$ refers to the entire set of parameters $\{\beta_{fcd}(\mathbf{y}_{\Phi}, \mathbf{x}_{\Delta}) : f \subseteq \Phi, c \subseteq \Gamma, \text{ and } d \subseteq \Delta\}$. We will frequently use this shorthand notation when referring to parameters of the same type in the remainder of the chapter. Assuming that the observations at each site are independent, the posterior distribution of the parameters and the random effects

is given by

$$
\begin{aligned}
f_{\text{post}}(\boldsymbol{\beta}, \boldsymbol{\omega}, \boldsymbol{\lambda}, \boldsymbol{\tau}, \boldsymbol{\Psi}_{\emptyset}, \boldsymbol{\Sigma}, \{\boldsymbol{\epsilon}\} \mid \{\mathbf{c}\}, \{\mathbf{x}\}) \propto & \prod_{i=1}^{S} f_M(\mathbf{c}_i|\boldsymbol{\beta}, \boldsymbol{\omega}, \mathbf{x}_i, \boldsymbol{\epsilon}_i) \\
& \times f_M(\mathbf{c}_\Delta|S, \boldsymbol{\lambda}) \\
& \times \prod_{i=1}^{S} f_N(\mathbf{x}_{\Gamma i}|\mathbf{x}_{\Delta i}, \boldsymbol{\tau}, \boldsymbol{\Psi}_{\emptyset}) \qquad (4.19) \\
& \times \prod_{i=1}^{S} \prod_{f \subseteq \Phi} f_N(\boldsymbol{\epsilon}_{f,i}|\boldsymbol{\Sigma}_f) \\
& \times \pi(\boldsymbol{\beta}, \boldsymbol{\omega}, \boldsymbol{\lambda}, \boldsymbol{\tau}, \boldsymbol{\Psi}_{\emptyset}, \boldsymbol{\Sigma}),
\end{aligned}
$$

where $\{\boldsymbol{\epsilon}\} = \{\boldsymbol{\epsilon}_i : \ i = 1, \ldots, S\}$, $\{\mathbf{c}\} = \{\mathbf{c}_i : \ i = 1, \ldots, S\}$, and $\{\mathbf{x}\} = \{\mathbf{x}_i : i = 1, \ldots, S\}$. In (4.19), the conditional distribution of the cell counts at each site, $f_M(\mathbf{c}_i|\boldsymbol{\beta}, \boldsymbol{\omega}, \mathbf{x}_i, \boldsymbol{\epsilon}_i)$ can be modeled with the multinomial likelihood formulation as shown or with the Poisson formulation (4.5) if the total site count should be considered random.

The posterior distribution (4.19) is a non-standard distribution, therefore, analytical inference for posterior objects of interest such as expected values and credible intervals is not possible. We will draw a sample from this distribution using Markov Chain Monte Carlo (MCMC) techniques. Robert and Casella (1999) provide an in-depth overview of MCMC techniques. Specifically, we will be using a Gibbs sampling approach (see Chapter 7 of Robert and Casella (1999)).

We can obtain simplification in the analysis of the posterior distribution by noting that if we impose the *a priori* independence of the CG parameters with the remaining parameters, $\pi(\boldsymbol{\beta}, \boldsymbol{\omega}, \boldsymbol{\lambda}, \boldsymbol{\tau}, \boldsymbol{\Psi}_{\emptyset}, \boldsymbol{\Sigma}) = \pi(\boldsymbol{\beta}, \boldsymbol{\omega}, \boldsymbol{\Sigma}) \times \pi(\boldsymbol{\lambda}, \boldsymbol{\tau}, \boldsymbol{\Psi}_{\emptyset})$, then the posterior distribution becomes

$$
\begin{aligned}
& f_{\text{post}}(\boldsymbol{\beta}, \boldsymbol{\omega}, \boldsymbol{\lambda}, \boldsymbol{\tau}, \boldsymbol{\Psi}_{\emptyset}, \boldsymbol{\Sigma}, \{\boldsymbol{\epsilon}\} \mid \{\mathbf{c}\}, \{\mathbf{x}\}) \\
& \quad \propto \left[\prod_{i=1}^{S} f_M(\mathbf{c}_i \mid \boldsymbol{\beta}, \boldsymbol{\omega}, \mathbf{x}_i, \boldsymbol{\epsilon}_i) \left\{ \prod_{f \subseteq \Phi} f(\boldsymbol{\epsilon}_{fi}|\boldsymbol{\Sigma}_f) \right\} \right] \pi(\boldsymbol{\beta}, \boldsymbol{\omega}, \boldsymbol{\Sigma}) \\
& \qquad \times \left[\prod_{i=1}^{S} f_N(\mathbf{x}_{\Gamma i} \mid \mathbf{x}_{\Delta i}, \boldsymbol{\tau}, \boldsymbol{\Psi}_{\emptyset}) \right] f_M(\mathbf{c}_\Delta|S, \boldsymbol{\lambda}) \pi(\boldsymbol{\lambda}, \boldsymbol{\tau}, \boldsymbol{\Psi}_{\emptyset}) \qquad (4.20) \\
& \quad \propto f_{\text{post}}(\boldsymbol{\beta}, \boldsymbol{\omega}, \boldsymbol{\Sigma}, \{\boldsymbol{\epsilon}\} \mid \{\mathbf{c}\}, \{\mathbf{x}\}) \times f_{\text{post}}(\boldsymbol{\lambda}, \boldsymbol{\tau}, \boldsymbol{\Psi}_{\emptyset} \mid \{\mathbf{x}\}).
\end{aligned}
$$

Therefore, the parameters of the explanatory portion of the graphical model and the parameters of the response portion of the graphical model are *a posteriori* independent. This simplifies the analysis of the posterior distribution because two separate MCMC analyses can be performed, one for each chain component. In fact, this type of sequential estimation is often used to estimate chain model parameters (Whittaker, 1990, pg. 310).

Another benefit of posterior independence is that the explanatory portion of the model will remain the same even if other response variables are analyzed. One need only examine the explanatory model once. Then, models for different sets of responses can be fit by analyzing each response portion separately. The full chain graph for each model can be constructed using the same subgraph for the explanatory variable chain component and then simply adding the directed edges and response components based on the set of responses being analyzed.

4.4.1 Hierarchical Centering Parameterization

Here we present a modification to the response model parameterizations presented in (4.9). When using a Gibbs MCMC procedure, the Markov chains for regression coefficient parameters, such as $\boldsymbol{\beta}$ and $\boldsymbol{\omega}$, in random effects generalized linear models are often slow to converge to the marginal posterior distribution (Chen et al., 2000, pg. 40). It has been our experience that this is also the case for graphical composition models. Therefore, we will use a hierarchical centering parameterization, as suggested by Chen et al. (2000), to help reduce the problem of poorly mixing chains for the regression coefficients $\boldsymbol{\beta}$ and $\boldsymbol{\omega}$.

In order to describe the hierarchical centering parameterization, first recall the response portion of the REDR distribution (4.9). Now, we introduce a shortened notation for the "fixed" effects portion of the response model for each $f \subseteq \Phi$,

$$\mu_f(\mathbf{y}_\Phi) = \sum_{c \subseteq \Gamma} \sum_{d \subseteq \Delta} \beta_{fcd}(\mathbf{y}_\Phi, \mathbf{x}_\Delta) \prod_{\gamma \in c} x_\gamma + \sum_{\gamma \in \Gamma} \sum_{d \subseteq \Delta} \sum_{m=2}^{M} \omega_{f\gamma dm}(\mathbf{y}_\Phi, \mathbf{x}_\Delta) x_\gamma^m. \qquad (4.21)$$

Then, we propose the following hierarchically centered re-parameterization of (4.9),

$$f_{RE}^{(h)}(\mathbf{y}_\Phi|\boldsymbol{\varphi}) = \exp\left\{\sum_{f\subseteq\Phi}\varphi_f(\mathbf{y}_\Phi)\right\}, \tag{4.22}$$

where $\varphi_f(\mathbf{y}_\Phi) = \mu_f(\mathbf{y}_\Phi) + \epsilon_f(\mathbf{y}_\Phi)$, $f \neq \emptyset$ and $\varphi_\emptyset$ represents the log normalizing constant with respect to $\mathbf{Y}_\Phi$ given $\boldsymbol{\varphi}$,

$$\varphi_\emptyset = -\log\left[\sum_{\mathbf{y}_\Phi}\exp\left\{\sum_{f\subseteq\Phi}\varphi_f(\mathbf{y}_\Phi)\right\}\right], \quad f \neq \emptyset. \tag{4.23}$$

If the assumption is made that $\boldsymbol{\epsilon}_f \sim f_N(\mathbf{0},\ \boldsymbol{\Sigma}_f)$ for $f \subseteq \Phi$ that are complete, as portrayed in (4.15), then $\boldsymbol{\varphi}_f = \{\varphi_f(\mathbf{y}_\Phi) : y_\phi \neq 1 \text{ for any } \phi \subseteq f\} \sim f_N(\boldsymbol{\mu}_f, \boldsymbol{\Sigma}_f)$, where $\boldsymbol{\mu}_f$ is the vector $[\mu_f(\mathbf{y}_\Phi)]$. Here we have simply changed the random effects $\boldsymbol{\epsilon}_f$ from a zero mean process to a process, $\boldsymbol{\varphi}_f$, centered at the fixed effects $\boldsymbol{\mu}_f$. In the case of generalized linear mixed models there is no theoretical result to show that this will improve convergence of the MCMC procedure. It has been our experience, however, that the re-parameterization often greatly improves convergence for these models.

We now provide a general parameterization of the full posterior distribution which we recommend using when employing MCMC techniques to make inference concerning the model parameters,

$$\begin{aligned} &f_{\text{post}}^{(h)}(\boldsymbol{\beta},\boldsymbol{\omega},\boldsymbol{\lambda},\boldsymbol{\Psi}_\emptyset,\{\boldsymbol{\varphi}_i\} \mid \{\mathbf{c}_i\},\{\mathbf{x}_i\}) \\ &\quad\propto \left[\prod_{i=1}^{S} f^{(h)}(\mathbf{c}_i|\boldsymbol{\varphi}_i)\left\{\prod_{f\subseteq\Phi} f_N(\boldsymbol{\varphi}_{fi}|\boldsymbol{\beta}_f,\boldsymbol{\omega}_f,\boldsymbol{\Sigma}_f,\mathbf{x}_i)\right\}\right]\pi(\boldsymbol{\beta},\boldsymbol{\omega},\boldsymbol{\Sigma}) \\ &\qquad\times\left[\prod_{i=1}^{S} f_N(\mathbf{x}_{\Gamma i}|\mathbf{x}_{\Delta i},\boldsymbol{\tau},\boldsymbol{\Psi}_\emptyset)\right] f_M(\mathbf{c}_\Delta|\boldsymbol{\lambda})\pi(\boldsymbol{\lambda},\boldsymbol{\tau},\boldsymbol{\Psi}). \end{aligned} \tag{4.24}$$

Once again, $f^{(h)}(\mathbf{c}_i|\boldsymbol{\varphi}_i)$ in (4.24) can take one of two forms. If the total number of individuals observed at a site is fixed, or should be considered fixed, then the multinomial density,

$$f_M^{(h)}(\mathbf{c}_i|N_i,\mathbf{p}_i) = \frac{N_i}{\prod_{\mathbf{y}_\Phi} c(\mathbf{y}_\Phi)_i!}\prod_{\mathbf{y}_\Phi} f_{RE}^{(h)}(\mathbf{y}_\Phi)^{c(\mathbf{y}_\Phi)_i} \tag{4.25}$$

is used, where $c(\mathbf{y}_\Phi)_i$ is the count for cell $\mathbf{y}_\Phi$ at the ith site. If, however, the site total is random, then the independent Poisson model,

$$f_P^{(h)}(\mathbf{c}_i|\boldsymbol{\varphi}_i) = \prod_{\mathbf{y}_\Phi} \frac{\left\{\sum_{f\subseteq\Phi}\varphi_f(\mathbf{y}_\Phi)\right\}^{c(\mathbf{y}_\Phi)_i} \exp\left\{\sum_{f\subseteq\Phi}\varphi_f(\mathbf{y}_\Phi)\right\}}{c(\mathbf{y}_\Phi)_i!} \tag{4.26}$$

is used where $\varphi_\emptyset(\mathbf{y}_\Phi) = 0$. A slight modification must be made to $\boldsymbol{\mu}_f(\mathbf{y}_\Phi)$ for the Poisson model. In order to include a model for the total number of individuals at a site, an intercept $\mu_\emptyset(\mathbf{y}_\Phi)$ term, possibly of the form (4.6), must be added to all $\mu_f(\mathbf{y}_\Phi)$ in (4.21).

A few comments concerning this re-parameterized model are in order. First, under either parameterization, the full count likelihood, as well as the REDR density, remain unchanged when integrated over the random effects. One can see this by observing that the transformation from $(\boldsymbol{\beta}, \boldsymbol{\omega}, \boldsymbol{\epsilon})$ to $(\boldsymbol{\beta}, \boldsymbol{\omega}, \boldsymbol{\varphi})$ has a Jacobian of 1 since $\boldsymbol{\epsilon}_f = \boldsymbol{\varphi}_f - \boldsymbol{\mu}_f$. Therefore, we obtain

$$f(\mathbf{c}_i|\mathbf{x}_i, \boldsymbol{\beta}, \boldsymbol{\omega}) = \int\cdots\int f^{(h)}(\mathbf{c}_i|\boldsymbol{\varphi}) \prod_{f\subseteq\Phi} f(\boldsymbol{\varphi}_f|\boldsymbol{\beta}_f, \boldsymbol{\omega}_f, \boldsymbol{\Sigma}_f, \mathbf{x}_i) d\boldsymbol{\varphi}_f \tag{4.27}$$

$$= \int\cdots\int f(\mathbf{c}_i|\boldsymbol{\beta}, \boldsymbol{\omega}, \boldsymbol{\epsilon}) \prod_{f\subseteq\Phi} f(\boldsymbol{\epsilon}|\mathbf{0}, \boldsymbol{\Sigma}_f) d\boldsymbol{\epsilon}_f, \tag{4.28}$$

and

$$f(\mathbf{y}_\Phi|\mathbf{x}_i) = \int\cdots\int f_{RE}^{(h)}(\mathbf{y}_\Phi|\boldsymbol{\varphi}) \prod_{f\subseteq\Phi} f(\boldsymbol{\varphi}_f|\boldsymbol{\beta}_f, \boldsymbol{\omega}_f, \boldsymbol{\Sigma}_f, \mathbf{x}_i) d\boldsymbol{\varphi}_f \tag{4.29}$$

$$= \int\cdots\int f_{RE}(\mathbf{y}_\Phi|\mathbf{x}_i, \boldsymbol{\epsilon}) \prod_{f\subseteq\Phi} f(\boldsymbol{\epsilon}_f|\mathbf{0}, \boldsymbol{\Sigma}_f) d\boldsymbol{\epsilon}_f. \tag{4.30}$$

These equivalences suggest two additional results. The first result is that the marginal re-parameterized model model (4.29) possesses the same Markov properties as as the integrated REDR model (4.30) (see Section 4.3). Prior to integration, however, the Markov properties for the two parameterizations are different. For the original parameterization, the model is Markovian with respect to a chain graph

where the explanatory variables and the random effects are unconnected parents of the response. For the hierarchically centered parameterization, the model is Markovian with respect to a chain graph where the explanatory variables are parents of the random effects which are in turn parents of the response variable. The second result is that since the integrated likelihoods in (4.27) and (4.28) are equivalent, the marginal posterior distributions of $\boldsymbol{\beta}$ and $\boldsymbol{\omega}$ will be equivalent under either parameterization. This can also be observed empirically by the fact that upon drawing a sample from the posterior distribution of $(\boldsymbol{\beta}, \boldsymbol{\omega}, \boldsymbol{\varphi})$ one can easily transform the sample values to $(\boldsymbol{\beta}, \boldsymbol{\omega}, \boldsymbol{\epsilon})$, thus obtaining a sample from its posterior distribution. The $\boldsymbol{\beta}$ and $\boldsymbol{\omega}$ values remain unchanged, therefore, any marginal posterior quantities calculated also remain unchanged.

4.4.2 Implementing the Gibbs Sampler

In order to implement the Gibbs sampler to draw a sample from (4.24) we need to obtain the full conditional distribution for each parameter. The full conditional distribution is the conditional distribution of the parameter in question given all remaining parameters as well as the observed data. A (non-independent) sample from the posterior is then drawn by iteratively drawing from each full conditional distribution. We derive the full conditional densities in two separate groups due to the fact that the full conditional densities for the response model parameters, $\boldsymbol{\beta}$, $\boldsymbol{\omega}$, $\{\boldsymbol{\varphi}_i\}$, and $\boldsymbol{\Sigma}$, will not be functions of the explanatory model parameters $\boldsymbol{\lambda}$, $\boldsymbol{\tau}$, and $\boldsymbol{\Psi}_{\emptyset}$, as can be observed from (4.24). Therefore, we can make inference to the response model parameters using only the first factor on the left hand side of the proportionality, while inference to the explanatory portion of the chain graph uses only the second factor.

Response Model Conditional Densities

Before deriving the full conditional densities we introduce some notation to ease the calculations. First, we will use the notation $\mathbf{E}$ to refer to a matrix that has S rows and has the following: a column of ones, a column corresponding to each of the explanatory variables, a column for each interaction and powers of the continuous covariates as given in (4.9). If a covariate X_δ, $\delta \subseteq \Delta$, is a categorical variable with b levels, then it will be represented by $b-1$ columns of indicator variables in $\mathbf{E}$, where each column indicates, with a one or zero, if X_δ takes the associated level at site i. The column associated with the reference level $X_\delta = 1$ is not included. We will denote the number of columns in $\mathbf{E}$ by r. In linear regression terminology, $\mathbf{E}$ represents the design matrix. The vector $\mathbf{E}_i$, $i = 1, \ldots, S$ will denote an r vector formed from the ith row of $\mathbf{E}$. In addition, let D_f denote the length of $\boldsymbol{\varphi}_f(\mathbf{y}_\Phi)$ and let $\mathbf{B}_f$ represent an $r \times D_f$ matrix of all the interaction coefficients $\{\beta_{fcd}(\mathbf{y}_\Phi) : c \subseteq \Gamma,\ d \subseteq \Delta\}$ and $\{\omega_{f\gamma dm}(\mathbf{y}_\Phi, \mathbf{x}_\Delta) : \gamma \in \Gamma,\ d \subseteq \Delta,\ m = 1, \ldots, M\}$ such that the expected value (4.21) of the site i random effect $\boldsymbol{\varphi}_{f,i}$ is given by $\boldsymbol{\mu}_f = \mathbf{B}_f' \mathbf{E}_i$. The "stacked" version of $\mathbf{B}_f$ will be represented by $\mathbf{B}_{fs}$. The stacked version is a $rD_f \times 1$ vector where the columns of $\mathbf{B}_f$ have been concatenated in order. Although previously described as the collection of all random effects, $\boldsymbol{\varphi}_f$ will now specifically represent a $S \times D_f$ matrix of these random effects and $\boldsymbol{\varphi}_{f,i}$ is a D_f vector formed from the ith row of $\boldsymbol{\varphi}_f$. Finally, we will make use of the inverse of the $D_f \times D_f$ random effects covariance matrix $\mathbf{T}_f = \boldsymbol{\Sigma}_f^{-1}$.

Now we can begin to derive full conditional distributions for the parameters of the response model, the coefficients in the matrices $\{\mathbf{B}_f : f \subseteq \Phi\}$, the site random effect matrices $\{\boldsymbol{\varphi}_f : f \subseteq \Phi\}$, and the random effects inverse covariance matrices $\{\mathbf{T}_f : f \subseteq \Phi\}$. The reader should note, as we derive the full conditional densities, that in addition to the increased rate of convergence, the hierarchical centering provides a Gibbs sampler that is easier to implement due to the fact that the interaction coefficients as well as the random effects covariance matrices

will have standard full conditional densities. In the non-centered parameterization only the covariance matrices have standard full conditional densities. Here, we will also make the assumption that for each $f \subseteq \Phi$, the interaction coefficients and the covariance matrices are *a priori* mutually independent across all f. In other words, $\pi(\boldsymbol{\beta}, \boldsymbol{\omega}, \boldsymbol{\Sigma}) = \prod_{f \subseteq \Phi} \pi(\boldsymbol{\beta}_f, \boldsymbol{\omega}_f)\pi(\boldsymbol{\Sigma}_f)$. Although, we have made this independence assumption, it is not critical for the remaining derivations in this section. In fact, the notational subscript f could just as easily be used to represent sets in Φ for which the parameters are independent. So, the results of this section will remain virtually the same in appearance even if more parameter dependence is added.

We begin with the interaction coefficients in $\mathbf{B}_f$ for any $f \subseteq \Phi$. First, note that due to the centering parameterization, given the random effects $\boldsymbol{\varphi}_{fi}$ at each site and the random effect inverse covariance matrix T_f, the interaction coefficients are independent of the cell counts. If we let $\pi(\boldsymbol{\beta}_f, \boldsymbol{\omega}_f) = \pi(\mathbf{B}_{fs}) = f_N(\mathbf{B}_{fs};\ \boldsymbol{\mu}_{B_{fs}}, \mathbf{V}_{B_{fs}}^{-1})$ and $\hat{\mathbf{B}}_f = (\mathbf{E}'\mathbf{E})^{-1}\mathbf{E}'\boldsymbol{\varphi}_f$ (correspondingly $\hat{\mathbf{B}}_{fs}$ represents the stacked version) then the full conditional distribution of the interaction coefficients is given as

$$
\begin{aligned}
f(\mathbf{B}_{fs}| \ \ldots\ ) &\propto \exp\left\{-\frac{1}{2}\sum_{i=1}^{S}\left(\boldsymbol{\varphi}_{f,i} - \mathbf{B}_f'\mathbf{E}_i\right)'\mathbf{T}_f\left(\boldsymbol{\varphi}_{f,i} - \mathbf{B}_f'\mathbf{E}_i\right)\right\} \\
&\qquad \times \exp\left\{-\frac{1}{2}(\mathbf{B}_{fs} - \boldsymbol{\mu}_{B_{fs}})'\mathbf{V}_{B_{fs}}(\mathbf{B}_{fs} - \boldsymbol{\mu}_{B_{fs}})\right\} \\
&\propto \exp\left\{-\frac{1}{2}tr\left[\mathbf{T}_f(\mathbf{B}_f - \hat{\mathbf{B}}_f)'\mathbf{E}'\mathbf{E}(\mathbf{B}_f - \hat{\mathbf{B}}_f)\right]\right\} \\
&\qquad \times \exp\left\{-\frac{1}{2}(\mathbf{B}_{fs} - \boldsymbol{\mu}_{B_{fs}})'\mathbf{V}_{B_{fs}}(\mathbf{B}_{fs} - \boldsymbol{\mu}_{B_{fs}})\right\} \\
&= \exp\left\{-\frac{1}{2}(\mathbf{B}_{fs} - \hat{\mathbf{B}}_{fs})'(\mathbf{T}_f \otimes \mathbf{E}'\mathbf{E})(\mathbf{B}_{fs} - \hat{\mathbf{B}}_{fs})\right\} \\
&\qquad \times \exp\left\{-\frac{1}{2}(\mathbf{B}_{fs} - \boldsymbol{\mu}_{B_{fs}})'\mathbf{V}_{B_{fs}}(\mathbf{B}_{fs} - \boldsymbol{\mu}_{B_{fs}})\right\},
\end{aligned}
\tag{4.31}
$$

where $\otimes$ represents the Kronecker product. The second proportionality statement for the random effects likelihood is due to Johnson and Wichern (1992, pg. 322).

One can now complete the square to show that

$$f(\mathbf{B}_{fs}| \ \dots \ ) = f_N(\mathbf{B}_{fs};\ \boldsymbol{\mu}_{f,1}, \mathbf{V}_{f,1}^{-1}), \tag{4.32}$$

where the mean and covariance are given by

$$\boldsymbol{\mu}_{f,1} = \left[(\mathbf{T}_f \otimes \mathbf{E}'\mathbf{E}) + \mathbf{V}_{B_{fs}}\right]^{-1} \left[(\mathbf{T}_f \otimes \mathbf{E}'\mathbf{E})\hat{\mathbf{B}}_{fs} + \mathbf{V}_{B_{fs}}\boldsymbol{\mu}_{B_{fs}}\right]$$

and (4.33)

$$\mathbf{V}_{f,1} = (\mathbf{T}_f \otimes \mathbf{E}'\mathbf{E}) + \mathbf{V}_{B_{fs}}.$$

Therefore, in the Gibbs sampler, drawing samples of the interaction coefficients is relatively simple. Updating can be done for a single parameter at a time as well, each one will have a univariate normal conditional density.

We now derive the conditional distribution for the inverse covariance matrix $\mathbf{T}_f$ of the random effects $\boldsymbol{\varphi}_f$. We assume, *a priori*, that $\mathbf{T}_f$ has a Wishart distribution, $f_W(\mathbf{T}_f;\ a_f, \mathbf{K}_f)$, with prior parameters $a > D_f - 1$, $D_f \times D_f$ positive definite matrix $\mathbf{K}_f$, and density

$$\begin{aligned} \pi(\mathbf{T}_f) &= f_W(\mathbf{T}_f;\ a_f, \mathbf{K}_f) \\ &\propto |\mathbf{T}_f|^{(a-D_f-1)/2} \exp\left\{-\frac{1}{2} tr\left[\mathbf{K}_f\mathbf{T}_f\right]\right\}. \end{aligned} \tag{4.34}$$

This is equivalent to specifying an inverse Wishart prior distribution for $\boldsymbol{\Sigma}_f$. Now, $\mathbf{T}_f$ only depends on $\boldsymbol{\varphi}_f$ and $\mathbf{B}_f$ through the random effects distribution, which is a MVN distribution. Therefore, we obtain the following full conditional distribution,

$$\begin{aligned} f(\mathbf{T}_f| \ \dots \ ) &\propto |\mathbf{T}_f|^{S/2} \exp\left\{-\frac{1}{2}\sum_{i=1}^{S}(\boldsymbol{\varphi}_{f,i} - \mathbf{B}_f'\mathbf{E}_i)'\mathbf{T}_f(\boldsymbol{\varphi}_{f,i} - \mathbf{B}_f'\mathbf{E}_i)\right\} \\ &\quad \times |\mathbf{T}_f|^{(a-D_f-1)/2} \exp\left\{-\frac{1}{2} tr\left[\mathbf{K}_f\mathbf{T}_f\right]\right\} \\ &= |\mathbf{T}_f|^{(a+S-D_f-1)/2} \\ &\quad \times \exp\left\{-\frac{1}{2} tr\left[\mathbf{T}_f\left\{\mathbf{K}_f + \sum_{i=1}^{S}(\boldsymbol{\varphi}_{f,i} - \mathbf{B}_f'\mathbf{E}_i)'(\boldsymbol{\varphi}_{f,i} - \mathbf{B}_f'\mathbf{E}_i)\right\}\right]\right\} \end{aligned} \tag{4.35}$$

It follows, then, upon examination of (4.34), the full conditional distribution of $\mathbf{T}_f$ is given by

$$f(\mathbf{T}_f| \ \dots \ ) = f_W(\mathbf{T}_f;\ a_{f,1}, \mathbf{K}_{f,1}) \tag{4.36}$$

where the full conditional parameters are

$$a_{f,1} = a_f + S$$

and (4.37)

$$\mathbf{K}_{f,1} = \mathbf{K}_f + \textstyle\sum_{i=1}^{S}(\boldsymbol{\varphi}_{f,i} - \mathbf{B}_f'\mathbf{E}_i)'(\boldsymbol{\varphi}_{f,i} - \mathbf{B}_f'\mathbf{E}_i).$$

Therefore, just like the interaction coefficients, the inverse covariance matrix $\mathbf{T}_f$ is relatively straightforward to sample from in the Gibbs algorithm.

We now turn our attention to the final parameter that needs to be updated during the Gibbs sampling algorithm, the vector of site random effects,$\boldsymbol{\varphi}_{f,i}$. Unfortunately, the random effects do not have a standard full conditional distribution. The full conditional density is given by

$$f(\boldsymbol{\varphi}_{f,i}| \ \dots \ ) \propto f^{(h)}(\mathbf{c}_i|\boldsymbol{\varphi}_i) f_N(\boldsymbol{\varphi}_{f,i};\ \mathbf{B}_f'\mathbf{E}_i, \mathbf{T}_f^{-1}), \tag{4.38}$$

where $f^{(h)}(\mathbf{c}_i|\boldsymbol{\varphi}_i)$ is either the multinomial density (4.25) or the product Poisson density (4.26). One can see from the full conditional density that for either cell count likelihood model, Poisson or multinomial, the full conditional density is non-standard. Therefore, we can employ a Metropolis-within-Gibbs step, as described in Section 1.5.2, to sample from this full conditional distribution.

Explanatory Variables Model Conditional Distributions

We now derive the conditional distributions for the parameters in the explanatory variable CG model (4.12). One can clearly see that the categorical and continuous explanatory variable parameters are functionally independent. Therefore, as with the initial separation of the response variable model and the explanatory

variable model, we can again perform separate posterior analyses for the discrete and continuous partitions of the explanatory model.

We begin with the analysis of the continuous variable model for $\mathbf{x}_\Gamma$ given $\mathbf{x}_\Delta$. We will follow the same steps as for the interaction coefficients in the response model described in the previous subsection. Let $\mathbf{E}_\Delta$ represent a matrix with S rows, and columns containing indicator variables for each variable categorical variable in Δ and all of the interactions $d \subseteq \Delta$. Again, this is identical to the classic ANOVA design matrix. The notation $\mathbf{E}_{\Delta_i}$ will represent a column vector obtained from the ith row of $\mathbf{E}_\Delta$. As with the interaction coefficients of the response model, we will place the τ parameters in a matrix $\mathbf{B}_\tau$ such that $\mathbf{B}'_\tau \mathbf{E}_{\Delta_i} = \sum_{d \subseteq \Delta} \boldsymbol{\tau}_d(\mathbf{x}_\Delta)$. Again, we also denote the "stacked" version of $\mathbf{B}_\tau$ as $\mathbf{B}_{\tau s}$. Now, using the prior distributions $\pi(\mathbf{B}_{\tau s}) = f_N(\mathbf{B}_{\tau s};\ \boldsymbol{\mu}_{\tau s}, \mathbf{V}_{\tau s}^{-1})$ and $\pi(\boldsymbol{\Psi}_\emptyset) = f_W(\boldsymbol{\Psi}_\emptyset;\ a_\psi, \mathbf{K}_\psi)$ and a set of proportionalities and equalities similar to those in (4.31) and (4.35), we obtain the full conditional distributions

$$f(\mathbf{B}_{\tau s} | \ \ldots\ ) = f_N(\mathbf{B}_{\tau s};\ \boldsymbol{\mu}_{\tau s,1}, V_{\tau s,1}^{-1}) \tag{4.39}$$

and

$$f(\boldsymbol{\Psi}_\emptyset | \ \ldots\ ) = f_W(\boldsymbol{\Psi}_\emptyset;\ a_{\psi,1}, \mathbf{K}_{\psi,1}), \tag{4.40}$$

where the parameters are given by

$$\begin{gathered}
\boldsymbol{\mu}_{\tau s,1} = \left[(\boldsymbol{\Psi}_\emptyset \otimes \mathbf{E}'_\Delta \mathbf{E}_\Delta) + \mathbf{V}_{\tau s}\right]^{-1} \left[(\boldsymbol{\Psi}_\emptyset \otimes \mathbf{E}'_\Delta \mathbf{E}_\Delta)\hat{\mathbf{B}}_{\tau s} + \mathbf{V}_{\tau s}\boldsymbol{\mu}_{\tau s}\right], \\
\mathbf{V}_{\tau s,1} = (\boldsymbol{\Psi}_\emptyset \otimes \mathbf{E}'_\Delta \mathbf{E}_\Delta) + \mathbf{V}_{\tau s}, \\
a_{\psi,1} = a_\psi + S, \\
\text{and} \\
\mathbf{K}_{\psi,1} = \mathbf{K}_\psi + \sum_{i=1}^{S} \left\{\mathbf{x}_{\Gamma i} - \sum_{d \subseteq \Delta} \boldsymbol{\tau}_d(\mathbf{x}_{\Delta i})\right\}' \left\{\mathbf{x}_{\Gamma i} - \sum_{d \subseteq \Delta} \boldsymbol{\tau}_d(\mathbf{x}_{\Delta i})\right\}.
\end{gathered} \tag{4.41}$$

The final set of parameters for the explanatory variable model is $\boldsymbol{\lambda}$. As with the random effects, λ_d, $d \subseteq \Delta$ does not have a standard full conditional distribution.

Here, we assume an independent multivariate normal distribution for each of the non-zero elements of $\boldsymbol{\lambda}_d$, so, the full conditional density is given by

$$f(\lambda_d(\mathbf{x}_\Delta)| \ \ldots\ ) \propto f_M(\mathbf{c}_\Delta|\boldsymbol{\lambda}) f_N(\boldsymbol{\lambda}_d;\ \boldsymbol{\mu}_{\boldsymbol{\lambda}_d}, \mathbf{T}^{-1}_{\boldsymbol{\lambda}_d}), \qquad d \subseteq \Delta, \tag{4.42}$$

where $f(\mathbf{c}_\Delta|\boldsymbol{\lambda})$ is the multinomial density (4.25). One can see that the full conditional density is non-standard. Therefore, we can, again, employ a Metropolis-within-Gibbs step.

4.5 Graphical Analysis of Fish Species Richness

Relating behavioral characteristics of organisms to environmental conditions at locations in which they are found is been a challenging problem for ecologists (Legendre et al., 1997). This problem was initially motivated by the n dimensional niche hypothesis. The n dimensional niche hypothesis states there is an n dimensional space upon which species unimodally distribute themselves according to environmental adaptations (Ricklefs, 1990). The unimodal distribution of a given species along an environmental axis implies the existence of an "environmental optimum" for that species. The problem of relating species traits to environmental conditions emerged from this hypothesis because distributions of specific species are usually not of interest as they are often biogeographically constrained, thereby limiting ecological inference to one geographic range. Analysis of functional traits and behaviors rather than taxonomy provides a more portable inference to community structure (Poff and Allen, 1995).

The initial attempts to describe these relationships involved the use of Canonical Correspondence Analysis (CCA) (ter Braak, 1985). The CCA approach attempts to ordinate each species along a set of environmental axes. Dolédec et al. (1996) continued the ordination approach by developing methods for marginally and jointly analyzing so called R, L, and Q tables, where R is a table with data on environmental

variables at each sampling site, L is a table of species occurrences at each site, and Q is a table of trait classifications for each species.

A more direct approach was introduced by Legendre et al. (1997). Legendre et al. (1997) termed their methodology a "solution to the fourth corner problem." For a single trait with multiple levels, the "four corners" represent four matrices: (1) a matrix of environmental variables by site, (2) an indicator matrix of species presence by site, (3) an indicator matrix of functional trait levels by species, and (4) a matrix of parameters relating environmental variables to the trait. The parameters in matrix (4) are product moment correlations between the trait counts and environmental variables and are estimated by a method of moments approach.

There are three main problems with the previous methodologies. First, these approaches only consider a single response variable at a time. Multiple traits cannot be analyzed simultaneously. Secondly, the previous approaches both measure marginal association between the environment and traits in question. The conditional relationships of a Markov random field give a more detailed measure of association between variables. For example, variables that are marginally correlated may in fact be independent upon conditioning on a third variable. This may provide evidence of possible mitigation by the third variable. Finally, the previous methods provide no predictive ability. If a researcher desires to predict community structure at a site with remotely sensed environmental measurements, the previous methods provide no means to accomplish this task.

The state-space model solution proposed by Billheimer (1995), Billheimer and Guttorp (1997), and Billheimer et al. (2001) for examining relationships between species traits and environmental covariates does possess predictive ability; however, it was only designed for analysis of a single trait. Therefore, if it is desirable to examine several traits simultaneously, one must use a fully dependent model, where all levels of interaction are present between traits, to analyze the cell counts. This

may lead to a model that is over-parameterized and provides no method for inference as to whether the response variables should be modeled as conditionally independent.

In this section we will use the full REDR distribution to model the relationship between fish species habit and pollution tolerance. Fish species habit refers to the depth in which the species inhabits, on the stream bottom or suspended in the water column. Pollution tolerance refers to a species to resist impacts due to pollution. By using the full REDR model we will be able to examine the complex conditional relationships between the environmental covariates, the habit response variable, and the tolerance response variable as a whole system through inference from a graphical chain model. The proportion of benthic and intolerant species are important metrics used by the EPA to measure stream degradation (McCormack et al., 2001). So, we are interested in inference concerning species richness, or, the number of species observed in each cell.

4.5.1 Data Description

During 1993-1996 the U. S. Environmental Protection Agency, along with the U. S. Fish and Wildlife Service and other contractors, surveyed 309 wadable streams in the Mid-Atlantic Highlands as part of the Environmental Monitoring and Assessment Program (EMAP). Herein, we will consider the data from 1994. Streams in the Mid-Atlantic Highlands Assessment (MAHA) were sampled during a 12-week period from April to July. At several sampled streams, chemical samples were taken and several physical habitat variables were measured. Fish were sampled by electro-fishing and each fish species was classified according to several taxonomic and ecological categories. McCormack et al. (2001) provides a list of all fish species in the MAHA region along with their trait classifications. A set of watershed scale environmental variables was also calculated for these sites from a GIS (Geographic Information System) model. These variables include metrics such as watershed area and average amount of precipitation in the watershed.

Some sites were eliminated from the analysis due to the fact that environmental covariates were not recorded. In addition, some sites were eliminated due to the fact that they were considered to be incapable of maintaining a fish population indefinitely. Sites for which the observed number of fish was less than 10 or the watershed area was less than 2 km^2 were eliminated. McCormack et al. (2001) judged that sites with these two criteria were incapable of maintaining a fish population indefinitely. After removal of all sites without environmental variables and those which cannot support a fish population, $S = 91$ sites remained in the 1994 data set.

In order to perform the functional trait analysis of species richness, each observed species is categorized according to two categorical variables, habit and tolerance. The habit variable has two levels, column species and benthic species. Benthic species inhabit environment at or near the bottom of a stream. Column species inhabit water depths between the surface of a stream and the bottom. Species tolerance refers to a species' ability to withstand degraded environmental conditions. Species classified as intolerant are highly sensitive to human induced stream degradation. Tolerant species can withstand large amounts of stream degradation before they are impacted. Species that are between the two extreme tolerance classifications are classified as intermediate tolerance. The distribution of cell counts is provided in Figure 4.2.

In our analysis, we are interested in the associations between habit type, tolerance level and several chemical and physical disturbance covariates as well as some climate associated covariates. The chemical covariates include stream site sulfate concentration (ln μeq/L), which measures acid deposition, chloride concentration (ln μeq/L), which is associated with human activity, and finally, a measure of water turbidity (ln NTU), or "cloudiness". The climate associated covariates include watershed area (ln km^2), elevation (m), and mean annual watershed precipitation (m). A numerical summary of the covariates in this analysis is provided in Table 4.1.

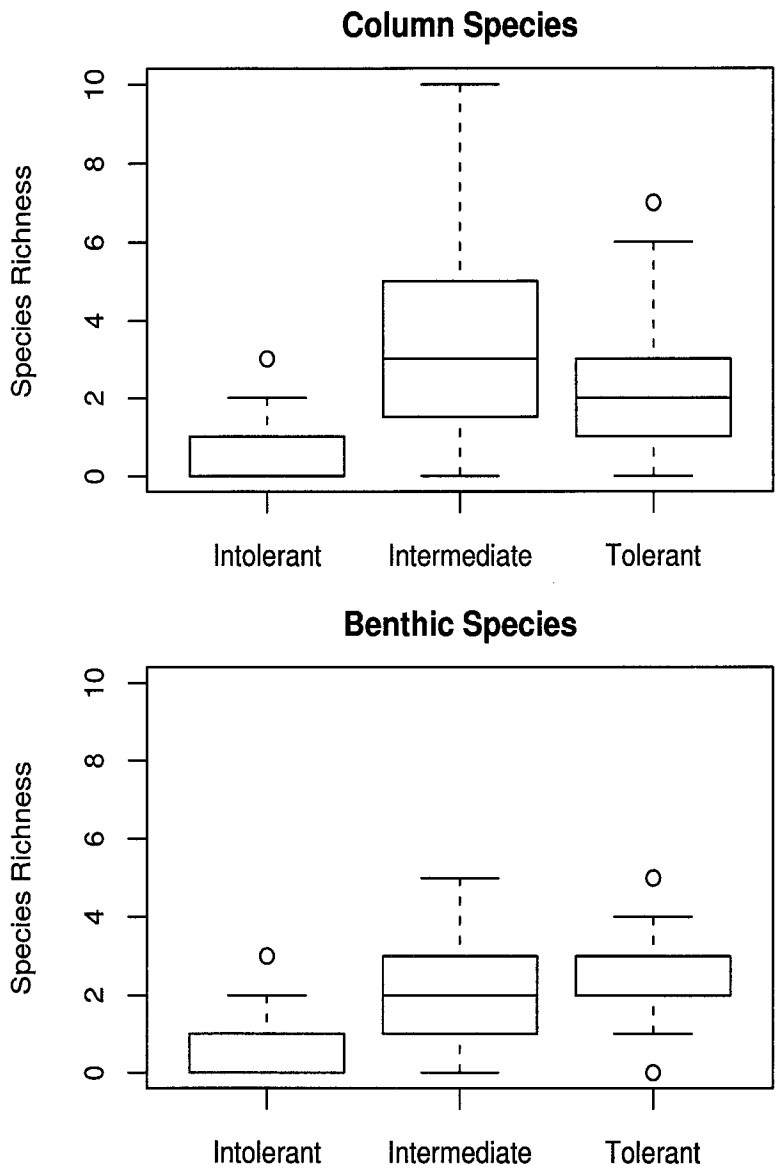

Figure 4.2: Distribution of fish species abundance for three different tolerance levels. Abundances are separated by habit type, column species and benthic species.

Table 4.1: Summary of environmental covariates for 1994 MAHA streams. Turbidity, chloride concentration, sulfate concentration, and watershed area were log transformed to give a better normal approximation.

Covariate	Mean	St. Dev.	Min.	Max.
Precipitation (m)	1.08	0.10	0.85	1.33
Elevation (m)	687.15	342.91	15.00	1389.00
Turbidity (ln NTU)	1.01	0.77	-0.92	3.40
Chloride (ln μeq/L)	4.60	1.14	2.64	6.98
Sulfate (ln μeq/L)	5.42	0.88	3.78	8.42
Area (ln km^2)	3.04	1.19	0.78	6.39

4.5.2 Model Description and Inference

Model Specification

We consider a main effects only model for the analysis of the multi-way composition of habit and tolerance species richness. We use H to denote the habit response and T to denote the pollution tolerance response. In addition, only first order linear interaction terms are included. We consider the following multinomial model for the cell counts,

$$f_M(\mathbf{c}_i|\mathbf{x}_i, \boldsymbol{\epsilon}_i) = \frac{N!}{\prod_{\mathbf{y}_\Phi} c(\mathbf{y}_\Phi)_i!} \prod_{\mathbf{y}_\Phi} f_{RE}(\mathbf{y}_\Phi|\mathbf{x}_i, \boldsymbol{\epsilon}_i)^{c(\mathbf{y}_\Phi)_i}, \tag{4.43}$$

where

$$f_{RE}(\mathbf{y}_\Phi) = \exp\left\{\alpha_\Phi(\mathbf{x}_i) + \sum_{f\subseteq\Phi}\sum_{\gamma=0}^{6} \beta_{f\gamma}(\mathbf{y}_\Phi)(x_{\gamma,i} - \bar{x}_\gamma)s_\gamma^{-1} + \epsilon_{f,i}(\mathbf{y}_\Phi)\right\}. \tag{4.44}$$

Each of the environmental covariates was centered by subtracting its mean $\bar{x}_\gamma$ and dividing by its standard deviation s_γ. This was done to improve Markov chain convergence to the posterior distribution. We chose the reference cell $\mathbf{y}_\Phi^*$ to be column species with intermediate tolerance level. Therefore, in equation (4.44) $\beta_{f\gamma}(\mathbf{y}_\Phi) \equiv \epsilon_{f,i}(\mathbf{y}_\Phi) \equiv 0$ for $\mathbf{y}_\Phi = (1,\ 2)$ and $f \subseteq \{H,\ T\}$ to ensure identifiability of the model.

In the analysis of the species richness data, we examine three different models. The first model considered is the *independent* model. In the independent model, the habit variable is independent of the tolerance variable. In other words, $\beta_{f\gamma}(\mathbf{y}_\Phi)$ and $\epsilon_{f,i}(\mathbf{y}_\Phi)$ are set to zero for $f = \{B,\ T\}$ for all covariates γ, sites i, and cells $\mathbf{y}_\Phi$. The second model is the *dependent (uncorrelated errors)* model. In the dependent model with uncorrelated errors, there are no interactions set to zero for all cells $\mathbf{y}_\Phi$. The random effects vectors $\boldsymbol{\epsilon}_f$ are independently distributed from one another for all sites. The final model is the *dependent (correlated errors).* The dependent correlated errors model is identical to the previous model except for the fact that the random effects vectors are correlated within each site. The dependent (correlated errors) model is equivalent to applying the single composition model of Chapter 3 to all six cells.

Since all of the included environmental covariates are continuous, the homogeneous CG model reduces to a MVN distribution. Therefore, we modeled the mean centered covariates as $f_{CG}(\mathbf{x}_{\gamma,i} - \bar{\mathbf{x}}_{\gamma,i}) = MVN(\mathbf{x}_{ip} - \bar{\mathbf{x}}_p;\ \mathbf{0}, \boldsymbol{\Psi}_\emptyset)$. The centering here allows the elimination of the nuisance parameter $\boldsymbol{\tau}_\emptyset$ in (4.12), which is irrelevant for determining conditional independencies. Note, that in this case one can analytically determine the posterior distribution of $\boldsymbol{\Psi}_\emptyset$ to be a Wishart distribution. However, since we are interested in the off-diagonal elements of $\boldsymbol{\Psi}_\emptyset$, using an MCMC sampling technique allows straightforward inference of these elements.

Model Estimation and Performance

For estimation purposes, the model was re-parameterized with the hierarchical centering approach of Section 4.4.1. MCMC procedures were performed with the hierarchically centered version as well as the version in (4.44). As hypothesized, the hierarchically centered version reached satisfactory convergence with substantially fewer MCMC iterations than (4.44).

In order to assess convergence of the Markov chains to the posterior distribution, the diagnostic procedure of Raftery and Lewis (1992) was employed. The diagnostic procedure of Raftery and Lewis (1992) estimates the number of iterations necessary to estimate a specified quantile of a parameter's posterior distribution within a given degree of accuracy. Since we are primarily interested in credible intervals for the β coefficients and the off-diagonal entries of $\mathbf{\Psi}_{\emptyset}$, we felt that this procedure provided the appropriate measure of convergence. For this analysis we specified the 0.025 quantile for estimation with an error no more than ± 0.005 with 95% probability.

The program WinBUGS was used to run the Gibbs sampler (Spiegelhalter et al., 2000). The Gibbs sampling algorithm was run for an initial 4000 iterations in which a MVN proposal density was tuned so that the Metropolis-within-Gibbs step for the φ_i parameters would have an acceptance rate of around 30%. The first 4000 iterations were then discarded, after which the sampler was run for an additional 800,000 iterations. Every twentieth iteration was saved for parameter inferences in order to reduce storage constraints. The Raftery and Lewis convergence diagnostic confirmed that the retained 40,000 iterations had sufficiently converged to the posterior density.

In order to assess model fitness, the Bayesian posterior predictive p-value method of Gelman et al. (1996) was used for each of the three models that were used to analyze the data. For a "goodness-of-fit" statistic $T(y)$, which can be a function of the observed data y and model parameters, the Bayesian predictive p-value is defined as $P_b = Pr\{T(y^{rep}) > T(y) \mid y\}$. The data y^{rep} represent a hypothesized replicate data set that could have resulted from the model. So, the interpretation is essentially the same as the classic p-value with the addition that the "null" distribution is the distribution of the statistic given only the observed data y. In an MCMC setting P_b is particularly easy to approximate. One simply performs the usual Gibbs sampler for parameter estimation with the addition, that at each iteration a replicate data set is generated from the sampled parameter values. At each

iteration one calculates $T(y)$ and $T(y^{rep})$ and records the proportion of iterations in which $T(y^{rep}) > T(y)$ (Gelman et al., 1996). We used this approach in our Gibbs algorithm for the feeding type REDR model. The goodness-of-fit statistic used for this analysis was the Freeman-Tukey statistic (Freeman and Tukey, 1950)

$$T(\mathbf{c}_1, \ldots, \mathbf{c}_S) = \sum_{i=1}^{S} \sum_{\mathbf{y}_\Phi} \left(\sqrt{c(\mathbf{y}_\Phi)_i} - \sqrt{N_i f_{RE}(\mathbf{y}_\Phi | \mathbf{x}_i, \boldsymbol{\epsilon}_i)} \right)^2, \tag{4.45}$$

where $\mathbf{c}(\mathbf{y}_\Phi)_i$ is the number of species belonging to cell $\mathbf{y}_\Phi$ at site i, N_i is the total number of species observed at site i, and $f_{RE}(\mathbf{y}_\Phi | \mathbf{x}_i, \boldsymbol{\epsilon}_i)$ is given by (4.44) and represents the cell composition. Other statistics could be used, such as Pearson's χ^2, however, there are many cells with small counts and the Freeman-Tukey statistic eliminates the problem of over-weighting (Brooks et al., 2000a).

To compare to appropriateness of each of the three models, we use the Deviance Information Criterion (DIC) (Spiegelhalter et al., 2003). DIC is composed of two competing elements. The first component of DIC, $\bar{D}(y)$, is a measure of model fit given by

$$\bar{D}(y) = -\frac{1}{M} \sum_{m=1}^{M} 2 \log \mathcal{L}(y | \theta_m), \tag{4.46}$$

where θ_m is a value sampled from the posterior distribution of the parameter Θ and $\mathcal{L}$ is the likelihood of the data given the parameters. The second component is a measure of model complexity, p_D given by

$$p_D = \bar{D}(y) - \hat{D}(y), \tag{4.47}$$

where $\hat{D}(y) = -2 \log \mathcal{L}(y | \bar{\theta})$ and $\bar{\theta}$ is the mean of a posterior sample of θ values. The p_D value measures the "effective number of parameters" Spiegelhalter et al. (2003). The DIC model selection criterion is then given by the sum of the two components

$$\text{DIC} = \bar{D}(y) + p_D. \tag{4.48}$$

The most appropriate model is the model that minimizes the DIC criterion.

Table 4.2: DIC and model complexity for multi-way fish species richness models. Models are listed in increasing DIC order. The column ΔDIC represents the difference in DIC from the model with the lowest DIC value. The column denoted with p_D represents the model complexity or "effective number of parameters".

Model	DIC	ΔDIC	p_D
Independent	1107.7	–	68.7
Dependent (Uncorrelated errors)	1117.8	10.1	106.1
Dependent (Correlated errors)	1166.8	59.1	162.5

4.5.3 Results and Discussion

Table 4.2 lists the considered models and their DIC values. The independent response model had the lowest DIC score followed by the uncorrelated errors model. The model with the highest DIC score is the fully dependent model with correlated errors. There is a difference of DIC scores of 10.1 between the independent response model and the nearest dependent response model. This suggests a sizable improvement in model parsimony by selecting the independent response model. Table 4.2 also shows the complexity component p_D for each model. The independent response had approximately 77.4 fewer effective parameters than the dependent response model with uncorrelated errors and approximately 100 fewer effective parameters than the dependent model with correlated errors. The Bayesian P value, P_b, was greater than 0.9 for all three of the considered models. This implies there is virtually no evidence of lack-of-fit for any of the models. So, although the dependent response models fit the data well, they seem to contain too many parameters to be parsimonious.

The 95% highest posterior density (HPD) intervals for the β interaction coefficients in the independent response model are given in Table 4.3. Proposition 4.1 states that environmental covariates must be a parent of the habit or tolerance response if at least one of the associated interaction coefficients is non-zero.

Upon examination of the HPD intervals in Table 4.3 one can see that there is ample evidence that chloride concentration and sulfate concentration are parents of the tolerance response and that elevation is a parent of the benthic response variable. The HPD intervals for the interaction terms of the remaining environmental variables, watershed area, precipitation, and turbidity, contain zero for both the habit and tolerance response variables, therefore, the analysis does not provide strong evidence that they are parents of either response. The HPD intervals for the off-diagonal elements of $\Psi_\emptyset$ are given in Table 4.4. Again, by examining Table 4.4 the intervals that do not contain zero provide strong evidence that there exists an undirected edge between the two associated variables in the marginal graph for the covariates.

Figure 4.3 illustrates the chain graph that is suggested by the DIC criterion and the 95% HPD intervals that do not contain zero. The DIC criterion suggested that the independent response model is more parsimonious than a dependent response model. The independent response model is, as mentioned in Section 4.3.1, a preservative model. Therefore, Figure 4.3 shows the subgraph for the response and covariates only, since the relationships are preserved when marginalizing over the random effects.

Table 4.3: 95% HPD intervals for covariate interaction parameters in the independent response model for the analysis of fish species richness.

	Habit		Tolerance		
Covariate	Column*	Benthic	Intolerant	Intermediate*	Tolerant
Precipitation	–	(-0.210, 0.078)	(-0.175, 0.403)	–	(-0.313, 0.057)
Elevation	–	(0.052, 0.405)	(-0.266, 0.473)	–	(-0.368, 0.089)
Turbidity	–	(-0.254, 0.117)	(-0.340, 0.366)	–	(-0.168, 0.220)
Sulfate	–	(-0.189, 0.138)	(0.007, 0.634)	–	(-0.072, 0.349)
Chloride	–	(-0.199, 0.195)	(-0.928, -0.071)	–	(-0.179, 0.319)
Area	–	(-0.301, 0.024)	(-0.133, 0.575)	–	(-0.376, 0.033)

*In this analysis, the Column habit type and the Intermediate tolerance type were used as the reference cell, therefore, the interaction coefficients are set to zero for those interaction terms referencing that cell.

Table 4.4: HPD intervals for the elements of the inverse covariance matrix $\Psi_\emptyset$ for the MAHA environmental variables.

	Elevation	Turbidity	Chloride	Sulfate	Area
Precipitation	(-0.008, 0.004)	(-1.717, 2.538)	(-1.421, 2.270)	(-1.547, 2.471)	(-1.074, 1.675)
Elevation		(-0.000, 0.000)	(0.002, 0.004)	(-0.002, 0.000)	(-0.002, -0.000)
Turbidity			(-0.754, 0.023)	(0.114, 0.945)	(-0.155, 0.400)
Chloride				(-1.194, -0.438)	(-0.625, -0.127)
Sulfate					(-0.053, 0.464)

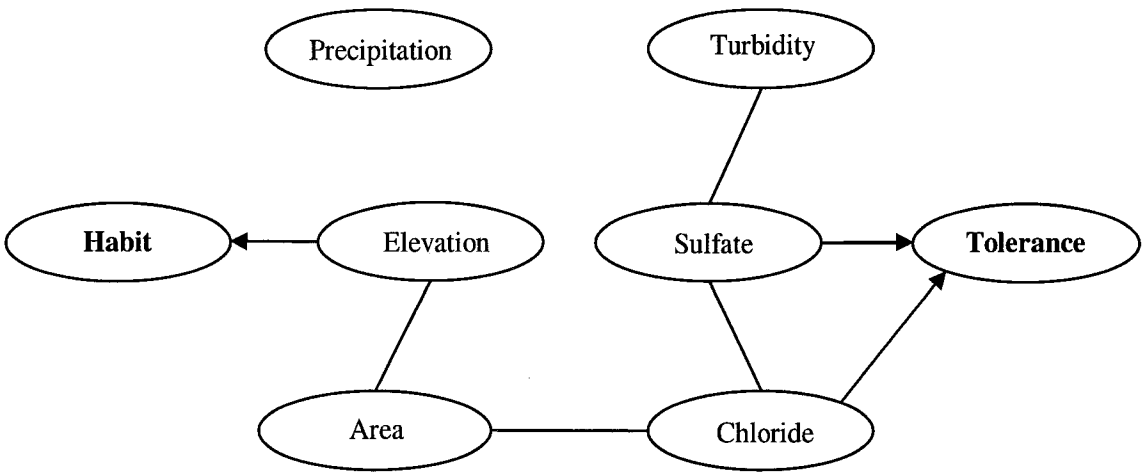

Figure 4.3: Data suggested chain graph for the multi-way composition of habit and tolerance.

Chapter 5

AUTOREGRESSIVE MODELS FOR CAPTURE-RECAPTURE DATA

5.1 Introduction

In this chapter, we present another class of models, capture-recapture models, which at first glance does not appear to be related to the random effects graphical models of Chapters 3 and 4. Here, we consider models which make inference to survival of animals living in the wild over some specified time unit. The data are collected as counts of marked and unmarked animals observed at each sampling period. The percentage of animals surviving from one time period to the next is the parameter of interest. Upon closer examination, one can observe that the structure of the data is nearly identical to the multiple site contingency table data of Chapter 4, with two exceptions: not all cells are observable, and some cells have a probability mass of zero. We give a more detailed description of the model similarity in the following section. The remaining portion of this section provides an introduction to capture-recapture modeling.

The investigation of factors that affect animal survival has become an increasingly important aspect of ecological research (Lebreton et al., 1992). It is often of interest to account for survival rates with covariates such as age, time, or weather factors (Buckland et al., 2000). Researchers have recently begun to explore the view that survival probabilities are realizations of a random process rather than fixed constants (Barry et al., 2003; Burnham, 2000; Burnham and White, 2002). Modeling survival probabilities via random effects allows one to account for overdispersion

(Barry et al., 2003) and unobserved (or random) environmental factors (Burnham, 2000). To this point, however, realizations of the survival process have been considered to be independent from one time period to the next. In some situations, this may be an unrealistic assumption. For example, survival at weekly intervals over the course of one season would likely be correlated, or high survival in one period may lead to low survival in following periods due to lack of resources. Therefore, it is reasonable to consider a time series correlation structure, such as an autoregressive structure (AR), in models where survival is considered a random process.

Vounatsou and Smith (1995), Brooks et al. (2000a), Brooks et al. (2000b), and Poole and Zeh (2002) have used Bayesian methods to estimate individual survival rates. Recently, Brooks et al. (2002) and Barry et al. (2003) have used Bayesian methods to estimate survival models with independent random effects as a way to model overdispersion. Burnham (2000) and Burnham and White (2002) have considered random effects in a non-Bayesian framework. Estimation via maximum likelihood is difficult in the context of random effects models. The likelihood is constructed by integrating over the random effects, and thus an integration must be performed over all of the random effects included in the model for each iteration of an optimization algorithm. This same difficulty is encountered in generalized linear models (Zeger and Karim, 1991).

The use of random effects allows for modeling survival in a capture-recapture model via an AR process. The Bayesian paradigm provides several advantages over maximum likelihood estimation. Using Markov Chain Monte Carlo (MCMC) procedures (Robert and Casella, 1999), point estimates can be produced by sampling from the posterior distribution of the parameters. In addition, Bayesian methods allow for estimation of the unobserved random effects as well. For example, survival probabilities can be estimated for each individual time period. This is not feasible with maximum likelihood estimation procedures when random effects are included.

We consider models for two common types of capture-recapture data: open population mark recapture (MR), where animals are recaptured and released alive, and band return (BR), where animals are recovered dead after each hunting season. For each of these data types we develop the theoretical construction and estimation procedures for a m^{th} order autoregressive, AR(m), random effects model using a Bayesian approach. The Bayesian analysis is illustrated using a long term waterfowl band recovery data set for Northern Pintails (*Anas acuta*).

5.2 Likelihood for Capture-Recapture Data

The likelihoods for open population mark-recapture (MR) data and band recovery (BR) data are structurally identical, the only major difference being a slight modification of the parameters. A complete description of the likelihood for capture-recapture data is given in Lebreton et al. (1992) and Brownie et al. (1985). We give a brief description here for the Bayesian methods presented in the next section.

Data are typically observed as an upper triangular array, $\mathbf{m}$, where the i, j^{th} element, m_{ij}, is the number of animals released at time t_i and subsequently recaptured (or reported, in the case of BR models) at time t_j (see Table 1, for example). The value I represents the number of capture occasions in which marking or banding is performed and J is the number of occasions in which recording recaptures or recoveries occurs. In MR studies, typically, $J = I$, while for BR studies, J may be greater than I due to the fact that marked animals may be harvested and reported after marking has stopped. Another component of the data is the $I \times 1$ vector $\mathbf{R} = [R_i]$, which contains the number of marked, or banded, animals released at each capture occasion. Each row of $\mathbf{m}$ is then modeled as a multinomial random variable with R_i trials and cell probabilities determined by survival probabilities and recapture or recovery probabilities. When using capture histories summarized into the sufficient statistics $\mathbf{m}$ and $\mathbf{R}$, the assumption is made that individuals have identical survival probabilities and recapture/recovery rates.

5.2.1 Open population mark-recapture likelihood

The MR model for survival estimation is also referred to as the Cormack-Jolly-Seber model (Cormack, 1964). This model is designed for studies in which a captured animal is labeled with a unique marking and released back into the wild population. At some point in the future, the marked animal may be recaptured, recorded, and released once again into the population. In MR studies, the first possible recapture of an animal marked at t_i is t_{i+1}, therefore, for these models $i = 1, \ldots, I$ and $j = i+1, \ldots, J+1$. Under the assumptions that individuals are independent and capture does not affect survival or recapture probabilities, the resulting product multinomial likelihood is

$$\mathcal{L}(\boldsymbol{\varphi}, \mathbf{p}; \mathbf{R}, \mathbf{m}) = \prod_{i=1}^{I} \binom{R_i}{m_{i,i+1}, \ldots, m_{i,J+1}, v_i} \xi_i^{v_i} \prod_{j=i+1}^{J+1} \left\{ \phi_i p_j \prod_{k=i+1}^{j-1} \phi_k (1-p_k) \right\}^{m_{ij}}, \tag{5.1}$$

where, ϕ_i is the probability that an animal survives from capture occasion t_i to t_{i+1}, $i = 1, \ldots, I$ given that it is alive at t_i, and p_j, is the probability that an animal alive at t_j, $j = 2, \ldots, J+1$, is captured at t_j. The probability that an animal is never recaptured after release at t_i is given by

$$\xi_i = 1 - \sum_{j=i+1}^{J+1} \phi_i p_j \prod_{k=i+1}^{j-1} \phi_k (1-p_k)$$

and $v_i = R_i - \sum_{j=i+1}^{J+1} m_{ij}$ is the number of animals captured at t_i and never subsequently recaptured during the study. In this section and for the remainder of this paper, a reverse order product is set equal to 1. For example, if $j = i+1$, then $\prod_{k=i+1}^{j-1} \phi_k (1-p_k) = 1$.

5.2.2 Band recovery likelihood

Band recovery models are designed for studies in which animals are captured, marked and released. Animals are then reported to the banding agency after harvesting by hunters. Therefore, at "recapture" occasions, the marked animals are

removed from the population. The structure of the data remains in the **R**, **m** format, so, the form of the likelihood is the same as (5.1), the only modifications being a change in the cell probabilities of the multinomial distribution and ranges for the i, j indices. Since an animal can be harvested and reported in the same time period in which it was banded, the index ranges are set at $i = 1, \ldots, I$ and $j = i, \ldots, J$. The resulting likelihood is

$$\mathcal{L}(\boldsymbol{\varphi}, \boldsymbol{\lambda}; \mathbf{R}, \mathbf{m}) = \prod_{i=1}^{I} \binom{R_i}{m_{ii}, \ldots, m_{iJ}, v_i} \xi_i^{v_i} \prod_{j=i}^{J} \left\{ \lambda_j \prod_{k=i}^{j-1} \phi_k \right\}^{m_{ij}}, \qquad (5.2)$$

where λ_j is the probability that a marked animal, alive at t_j, is harvested between time t_j and t_{j+1} and reported to the banding agency. In the band recovery model $\xi_i = 1 - \sum_{j=i}^{J} \lambda_j \prod_{k=i}^{j-1} \phi_k$ and $v_i = R_i - \sum_{j=i}^{J} m_{ij}$. Notice, in the BR model, that since an animal can be reported in the same time period as marking, the i, i cell probabilities involve only the λ_i parameter. Therefore, there are only $J - 1$ survival probabilities even though there are J years of data.

5.2.3 Similarity with Random Effects Discrete Regression Models

In this section, we more fully describe the relationship between the capture-recapture data and models with the random effects discrete regression models of Chapter 4. We will give examples of the similarity in terms of the band recovery model, however, the same similarities will obviously hold for the open population mark recapture models.

The capture-recapture data structure is identical to that of the random effects Discrete Regression (REDR) models of Chapter 4. Each individual animal captured and released back into the wild is placed into one of several cross-classified cells. The cell into which an animal is placed depends on the fate of the animal. The categorical variables that define these cross-classifications each have two levels and include a variable for every year from the year of release that describes whether

the animal survived or not and a variable for every year that describes whether the animal was reported to a banding agency. For example, an animal released in year one of a five year study will be cross-classified according to nine different categorical variables. It will be classified according to five reporting variables, whether or not it was reported to the banding agency in years one, two, three, four, and five. It will also be classified according to four survival variables, whether or not it survived years two, three, four, and five.

The major difference between the present band recovery models and the REDR models is that some cells are not observable or not physically plausible. For example, any cell for which an animal would have to be harvested but not reported is not observable. Another cell that is not observable is the cell for which an animal is placed if it survives all remaining years of the study once released. Cells that are not physically possible obviously include those for which an animal does not survive a given year, but survives the following year. Another physically impossible cell is a cell for which an animal is reported in a given year, but also survives that year (this is not physically impossible for the MR model though). The fact that some cells have a probability mass of zero precludes a Markov random field analysis such as in Chapter 4, but, that would not be of great interest for consumers of these models.

In terms of the model structure of Chapter 4, here, each row in the data matrix $\mathbf{m}$ represents a compositional observation ("site" in Chapter 4). The multinomial likelihood for each row of $\mathbf{m}$ is essentially the same as the multinomial count likelihood (4.4) for each "site" in Chapter 4, the difference being that each observation in the BR model has a different number of cells that are modeled. Every cell that is physically impossible is given a cell probability of zero and every cell that is not observable is combined into one cell for those that are released but not observed again in the study.

5.3 A Bayesian Approach for AR(m) Survival Models

5.3.1 Model specification

We consider a generalized linear model for the probability that an animal survives from time t_j to time t_{j+1} of the form

$$g(\phi_j) = \mathbf{X}_j'\boldsymbol{\beta} + \epsilon_j, \qquad j = 1, \ldots, J, \tag{5.3}$$

where g is an appropriate link function to constrain survival between 0 and 1, $\mathbf{X}_j$ is a $P \times 1$ matrix of covariates collected at capture occasion j, $\boldsymbol{\beta}$ is a $P \times 1$ vector of regression coefficients, and $\boldsymbol{\epsilon} = (\epsilon_1, \ldots, \epsilon_J)' \sim N(\mathbf{0}, \boldsymbol{\Sigma})$. The covariance matrix, $\boldsymbol{\Sigma}$, can be any general form. Here we consider an AR(m) model which implies that the ϵ_j error terms are realizations from the stochastic process

$$\epsilon_j = \sum_{k=1}^{m} \rho_k \epsilon_{j-k} + z_j, \qquad j = 1, \ldots, J, \tag{5.4}$$

where $z_j \sim$ i.i.d. $N(0, \sigma^2)$ and $\boldsymbol{\rho} = (\rho_1, \ldots, \rho_m)$ is a set of parameters.

We assume the process represented by (5.4) is stationary. For this model, stationarity implies that the covariance between two survival probabilities is a decreasing function of the distance between two time points and is independent of any one time point. The stationarity assumption of the error process imposes a constraint on $\boldsymbol{\rho}$ such that the roots of the characteristic equation,

$$x^m - \rho_1 x^{m-1} - \cdots - \rho_m = 0,$$

must be less than 1 in absolute value (Brockwell and Davis, 1996). In terms of the parameters, an AR(1) process is stationary if $|\rho| < 1$, while, an AR(2) process is stationary if $|\rho_1| < 2$ and $-1 < \rho_2 < 1 - |\rho_1|$.

By including random error terms in (5.3), we account for unknown environmental influences that might affect the probability of survival. Without the addition of the error terms, it is assumed that the covariates completely determine survival. Allowing for correlation between the random error terms in (5.4) provides the added

complexity that unknown environmental conditions may be similar for capture periods close together in time, so survival probabilities should also be related. Negative values for some elements of $\boldsymbol{\rho}$ might imply some density dependent effects in the population. A year in which survival is above average may lead to below average survival rates in the subsequent years due to lack of resources.

A stationary AR(m) model provides either positive or negative correlation between survival probabilities that decreases in absolute value with an increasing separation in time. So, the AR(m) model provides the type of relationship between survival probabilities that is desired. In addition, the model is relatively straightforward. Lindsey (1999, pg 106) notes that for short repeated measurement studies, elaborate time-series modeling is not necessary or possible and a simple AR process is usually adequate. The vast majority of capture-recapture datasets are no more than 50 years long (Franklin et al., 2002). Therefore, capture-recapture data certainly fit into the category of short time series data.

The model specified in (5.3) is one where the time series component appears in the error term. In other AR model formulations, the time series component appears with the mean term (Lindsey, 1999). However, we prefer model (5.3) for ease of biological interpretation. Often, the goal is to determine what covariates best model survival probability. If all of the variation in survival probability is not accounted for with the covariates sampled, only then would it be advisable to determine what associations exist between survival probabilities and different time periods.

5.3.2 Bayesian parameter estimation

We adopt a Bayesian approach for estimating the parameters for an AR(m) capture-recapture model specified in Section 5.1. The goal of this approach is to estimate the posterior distribution of the parameters to make inference about the

parameters and ecological hypotheses. This approach is relatively simple in comparison to maximum likelihood estimation (MLE). To estimate the parameters via MLE, it is necessary to evaluate the integrated likelihood of the form

$$\mathcal{L}(\boldsymbol{\varphi}, \cdot, \sigma^2, \boldsymbol{\rho}; \mathbf{R}, \mathbf{m}) = \int_{\boldsymbol{\epsilon}} \mathcal{L}(\boldsymbol{\varphi}, \cdot; \mathbf{R}, \mathbf{m}) N(\boldsymbol{\epsilon}; \mathbf{0}, \boldsymbol{\Sigma}) d\boldsymbol{\epsilon}$$

where, $\mathcal{L}(\cdot)$ is given by (5.1) or (5.2) and $N(\cdot)$ follows from (5.3) and (5.4). Therefore, for each step in an optimization algorithm a high dimensional integration must be performed. An alternative approach, quasi-likelihood (McCullagh, 1983) has been developed for random effects models in the generalized linear model setting. This approach involves the development of estimating equations that behave like likelihood functions and hence often have the same properties. In the capture-recapture setting, however, the fact that the cell probabilities are functions of the survival probabilities makes quasi-likelihood estimation difficult as well. In the Bayesian paradigm, the unobserved random effects are treated as random variables along with the parameters and the integration is performed stochastically through a Markov chain which samples from the joint conditional distribution of the parameters and the random effects given the data. From this joint conditional distribution we can obtain point estimates and confidence intervals for the parameters of the model.

In what follows, we present a general estimation procedure for both mark recapture and band recovery data. To simplify notation, we will use the notation $\mathbf{r}$ to represent either the vector of capture probabilities, $\mathbf{p}$, or band return rates, $\boldsymbol{\lambda}$, depending on the type of data being considered. We will also use the j index range $1, \ldots, J$ for both MR and BR data as this will not change the estimation procedure. The observed data, $\mathbf{m}$ and $\mathbf{R}$, as well as the covariates, $\mathbf{X}$, will collectively be denoted by D.

We assume that the parameters $\boldsymbol{\beta}$, σ^2, $\boldsymbol{\rho}$, and $\mathbf{r}$ are independent *a priori*. The posterior distribution of the parameters and random effects is then given by

$$\begin{aligned}\pi(\boldsymbol{\beta}, \sigma^2, \boldsymbol{\rho}, \beta, \mathbf{r}|D) \;\propto\; & \mathcal{L}(\boldsymbol{\beta}, \beta, \mathbf{r}; D) \times |\boldsymbol{\Sigma}|^{-1/2} \exp\left\{-\boldsymbol{\epsilon}'\boldsymbol{\Sigma}^{-1}\boldsymbol{\epsilon}/2\right\} \\ & \times \pi(\boldsymbol{\beta})\pi(\sigma^2)\pi(\boldsymbol{\rho})\pi(\mathbf{r}).\end{aligned} \tag{5.5}$$

In order to draw a sample from this distribution we will make use of the Gibbs sampler (e.g. Section 2.1 of Chen et al. (2000)), which requires the full conditional distributions for each of the parameters. A sample from the joint posterior distribution can be drawn by successively drawing from the full conditional posterior distributions for each of the parameters.

A simplification of the full conditional distributions results from the fact that the likelihood function can be broken into three parts. In addition, since the posterior is only defined up to a multiplicative constant, we can ignore the multinomial coefficients. Therefore, the likelihood can be rewritten as

$$\mathcal{L}(\boldsymbol{\beta}, \beta, \mathbf{r}; D) \propto V \times \mathcal{L}_{\boldsymbol{\phi}} \times \mathcal{L}_{\mathbf{r}},$$

where for both MR and BR data, $V = \prod_{i=1}^{I} \xi_i^{v_i}$. For MR data

$$\mathcal{L}_{\boldsymbol{\phi}} = \prod_{i=1}^{I} \prod_{j=i+1}^{J+1} \left(\phi_i \prod_{k=i+1}^{j-1} \phi_k \right)^{m_{ij}}$$

and

$$\mathcal{L}_{\mathbf{r}} = \prod_{i=1}^{I} \prod_{j=i+1}^{J+1} \left\{ p_j \prod_{k=i+1}^{j-1} (1 - p_k) \right\}^{m_{ij}},$$

and for BR data

$$\mathcal{L}_{\boldsymbol{\phi}} = \prod_{i=1}^{I} \prod_{j=i}^{J} \left\{ \prod_{k=i+1}^{j-1} \phi_k \right\}^{m_{ij}}$$

and

$$\mathcal{L}_{\mathbf{r}} = \prod_{i=1}^{I} \prod_{j=i+1}^{J} \lambda_j^{m_{ij}}.$$

Now, with the partitioned form of the likelihood we can simplify the full conditional distributions for each of the parameters. Due to the fact that all but one parameter has a nonstandard distribution, we will only give the conditional distributions up to a proportionality constant.

If the regression parameters for the covariates, $\boldsymbol{\beta}$, in (5.3) are independent *a priori*, then the full conditional of the coefficient for the l^{th} covariate, β_l, is given by

$$f(\beta_l|\boldsymbol{\beta}_{-l}, \sigma^2, \boldsymbol{\rho}, \boldsymbol{\epsilon}, \mathbf{r}, D) = f(\beta_l|\boldsymbol{\beta}_{-l}, \boldsymbol{\epsilon}, \mathbf{r}, D) \propto V \cdot \mathcal{L}_\phi \cdot \pi(\beta_l) \qquad l = 1, \ldots, P.$$

Likewise, independent priors for the components of $\mathbf{r}$, the vector of capture probabilities for MR data or band return rates for BR data, give

$$f(r_l|\mathbf{r}_{-l}, \boldsymbol{\beta}, \boldsymbol{\epsilon}, \sigma^2, \boldsymbol{\rho}, D) = f(r_l|\mathbf{r}_{-l}, \boldsymbol{\beta}, \boldsymbol{\epsilon}, D) \propto V \cdot \mathcal{L}_\mathbf{r} \cdot \pi(r_l) \qquad l = 1, \ldots, J.$$

When deriving the full conditional distribution of ϵ_l, we first note, given that we are assuming stationarity, that we can rewrite the joint distribution of the error terms in (5.3) in the form

$$\begin{aligned} f(\boldsymbol{\epsilon}|\sigma^2, \boldsymbol{\rho}) &= f(\epsilon_1) \textstyle\prod_{l=2}^{J} f(\epsilon_l|\epsilon_1, \ldots, \epsilon_{l-1}) \\ &= \textstyle\prod_{l=1}^{J} N(\nu_l,\ \sigma^2 K_l), \end{aligned} \tag{5.6}$$

where $\nu_l = E[\epsilon_l|\epsilon_1, \ldots, \epsilon_{l-1}]$, $l = 2, \ldots, m$ and K_l is a function of $\boldsymbol{\rho}$. In a stationary AR(m) process $\nu_1 = 0$ and $(\nu_l,\ K_l) = (\sum_{k=1}^{m} \rho_k \epsilon_{l-k},\ 1)$ for $l = m+1, \ldots, J$. In order to find the remaining ν's and K's one can make use of the Durbin-Levinson algorithm (Brockwell and Davis, 1996, pg 67). In the case of an AR(2) process, for example,

$$\begin{gathered} K_1 = (1-\rho_2)/[(1+\rho_2)\{(1-\rho_2)^2 - \rho_1^2\}]^{-1}, \\ K_2 = (1-\rho_2^2)^{-1}, \\ \text{and} \\ \nu_2 = \rho_1/(1-\rho_2). \end{gathered} \tag{5.7}$$

For an AR(1) process simply set $\rho_2 = 0$ in (5.7).

It is immediately apparent, due to the fact that an AR(m) process is a Markov process, that each component of β is dependent only on its m nearest neighbors.

Using this fact, the full conditional distribution of ϵ_l for the Gibbs sampler can be written as a function of the conditional normal distribution of ϵ_l given its m nearest neighbors. Therefore, the full conditional distribution for ϵ_l, $l = 1, \ldots, J$, is

$$f(\epsilon_l|\boldsymbol{\epsilon}_{-l}, \boldsymbol{\beta}, \sigma^2, \boldsymbol{\rho}, \mathbf{r}, D) \propto V \cdot \mathcal{L}_\phi \cdot \prod_{j=l}^{l+\min\{m,\ J-l\}} N(\nu_j,\ \sigma^2 K_j),$$

which, for each ϵ_l, can be condensed to the following form,

$$f(\epsilon_l|\boldsymbol{\epsilon}_{-l}, \boldsymbol{\beta}, \sigma^2, \boldsymbol{\rho}, \mathbf{r}, D) \propto V \cdot \mathcal{L}_\phi \cdot N\left(\mu_l/\eta_l,\ \sigma^2/\eta_l\right),$$

by completing the square. For an AR(2) error process

$$\mu_l = \begin{cases} \rho_1\epsilon_2 + \rho_2\epsilon_3 & l = 1 \\ \rho_1(\epsilon_1 + \epsilon_3) + \rho_2(\epsilon_4 - \rho_1\epsilon_3) & l = 2 \\ \rho_1(1-\rho_2)(\epsilon_{l-1} + \epsilon_{l+1}) + \rho_2(\epsilon_{l-2} + \epsilon_{l+2}) & l = 3, \ldots, J-2 \\ \rho_1(\epsilon_J + \epsilon_{J-2}) + \rho_2(\epsilon_{J-3} - \rho_1\epsilon_J) & l = J-1 \\ \rho_1\epsilon_{J-1} + \rho_2\epsilon_{J-2} & l = J \end{cases}$$

and

$$\eta_l = \begin{cases} 1 & l = 1 \text{ and } J \\ 1 + \rho_1^2 & l = 2 \text{ and } J-1 \\ 1 + \rho_1^2 + \rho_2^2 & l = 3, \ldots, J-2 \end{cases}.$$

Once again, in order to obtain the full conditionals in the case of an AR(1) model, simply set $\rho_2 = 0$.

Due to the stationarity constraint on the autocorrelation parameters, $\boldsymbol{\rho}$, we must consider the joint full conditional distribution of $\boldsymbol{\rho}$ instead of assuming independent priors on the components of $\boldsymbol{\rho}$. Using the decomposition of $f(\beta|\sigma^2, \boldsymbol{\rho})$ in (5.6), we can write the full conditional distribution of $\boldsymbol{\rho}$ as

$$f(\boldsymbol{\rho}|\boldsymbol{\beta}, \boldsymbol{\epsilon}, \sigma^2, \mathbf{r}, D) = f(\boldsymbol{\rho}|\boldsymbol{\epsilon}, \sigma^2) \propto \left(\prod_{j=1}^{m} K_j\right)^{-1} \exp\left\{-\frac{1}{2\sigma^2}\sum_{j=1}^{J}(\epsilon_j - \nu_j)^2/K_j\right\}\pi(\boldsymbol{\rho}).$$

The full conditional distribution of σ^2 is nearly identical to that of $\boldsymbol{\rho}$. Using the decomposition of $f(\beta|\sigma^2, \boldsymbol{\rho})$, the full conditional of σ^2 is

$$f(\sigma^2|\boldsymbol{\beta}, \boldsymbol{\epsilon}, \boldsymbol{\rho}, \mathbf{r}, D) = f(\sigma^2|\boldsymbol{\epsilon}, \rho) \propto \sigma^{-J} \exp\left\{-\frac{1}{2\sigma^2}\sum_{j=1}^{J}(\epsilon_j - \nu_j)^2/K_j\right\}\pi(\sigma^2),$$

which is the form of an inverse gamma distribution with shape and scale parameters $J/2+1$ and $C(\boldsymbol{\epsilon},\boldsymbol{\rho})/2 = \sum_{j=1}^{J}(\epsilon_j - \nu_j)^2/2K_j$, respectively. Therefore, if $\pi(\sigma^2)$ is an inverse gamma distribution with parameters a_0 and b_0, $\Gamma^{-1}(a_0,\ b_0)$, then the resulting conditional is an inverse gamma distribution with parameters $J/2+a_0+1$ and $C(\boldsymbol{\epsilon},\boldsymbol{\rho})/2+b_0$. The full conditional of σ^2 is the only standard density.

When implementing Bayesian methodology, it is necessary to choose priors for the parameters. It is a standard practice in generalized linear models with random effects to assign the vague priors $\pi(\beta_l) = N(0, 1/\tau)$ for $l = 1,\ldots,P$ and $\pi(\sigma^2) = \Gamma^{-1}(\varepsilon,\varepsilon)$ where τ and ε are small (Dey et al., 2000, pg 400). In past Bayesian capture-recapture analyses, $\pi(r_l)$ has been chosen to be a beta distribution for $l = 1,\ldots,J$ (Brooks et al., 2000a) of which the uniform distribution is a special case for vague prior information. All of these priors can be easily modified to produce informative priors as desired.

When there is little or no prior information concerning the parameter $\boldsymbol{\rho}$, a uniform distribution on the region of stationarity would be the obvious choice for a noninformative prior distribution. This uniform distribution, however, may produce marginal priors which are not as vague as the researcher would like. In the AR(2) case for example, a uniform distribution for the AR parameters ρ_1 and ρ_2 over the region of stationarity produces marginal distributions which are not uniform. In addition, a majority of the mass for the marginal distribution of ρ_2 will be located over negative values, producing a prior mean which is negative. This problem can often occur when building priors for parameter vectors over a constrained space. Barnard et al. (2000) illustrate the same dilemma when constructing priors for positive-definite covariance matrices.

In previous analyses using AR processes, the prior for the AR parameters was taken to be uniform over the stationary space of the parameters or a normal distribution if stationarity was not a concern or possibility (Huerta and West, 1999).

Informative priors can also be constructed by truncating a multivariate normal to the stationary space. There is another approach, suggested by Sun and Berger (1998), that is useful for constrained parameter spaces. If we are concerned with the parameter vector $(\theta_1,\ \theta_2)$, then a prior can be built in the form $\pi(\theta_1)\pi(\theta_2|\theta_1)$. Using this method, we can often build a sufficiently noninformative prior that has better marginal properties. For example, in the AR(2) model, if we take $\pi(\rho_2) = U(-1, 1)$ and $\pi(\rho_1|\rho_2) = U(-(1-\rho_2),\ 1-\rho_2)$, we obtain a prior that approximates a joint uniform with marginal distributions centered on 0. The partial information approach can also be used to specify informative priors for some of the AR parameters, while leaving others vague.

Another practical aspect for the Bayesian analysis of AR(m) capture-recapture models is that a modified Gibbs sampler must be used due to the non-standard conditional distributions. In the following example, a Metropolis within Gibbs sampler (Gelman and Rubin, 1993) was used. Instead of successively sampling from the full conditional distributions to obtain a sample from the joint posterior, an observation is first drawn from a proposal distribution and then either accepted or rejected with a given probability.

5.4 Example: Northern Pintails

In order to illustrate the fitting of an AR(m) model to capture-recapture data we applied the Gibbs sampler methodology to a Northern Pintail band recovery data set for females in California. These data (Table 1) were first analyzed by Franklin et al. (2002) as part of a meta-analysis on long-term trends in avian survival for many North American bird species. The previous analysis was performed using a linear trend model, an identity link function, and independent yearly random effects. The trend parameters as well as the variance component were estimated using the shrinkage estimation method of Burnham (2000). The slope estimate

from the previous analysis is 0.0023 with an estimated standard error of 0.0051. The variance component is estimated to be 0.212.

The previous analyses detected no significant trend to survival probabilities over time. We will include a slope parameter in this example, however, as an illustration of the use of covariates in our estimation procedure. Therefore, we will use the model

$$\text{logit } \phi_j = \beta_0 + \beta_1(j - 14) + \epsilon_j \quad j = 1, \ldots, 27 \tag{5.8}$$

to illustrate the application of an AR(m) model. In this example, the covariate vector $\mathbf{X}_j'$ in (5.3) is given by $(1,\ j - 14)$. The time index is centered to reduce correlation of the β_1 sample with the β_0 sample, which leads to better exploration of the posterior density for each variable. In addition, since there seems to be no significant trend based on the previous analysis, we also analyzed the data without a slope parameter.

We chose to estimate separate reporting probabilities, λ_j's, for each year. Barry et al. (2003) note that separate λ_j's in (5.2) tend to confound the effects of a random survival process and this has been our experience as well. However, we have adopted a conservative strategy for making inference about a random survival process, by allowing for fluctuating reporting rates.

For this example, we have chosen to fit an AR(2) model to the data. This implies that the error terms in (5.8) follow the stochastic process

$$\epsilon_j = \rho_1\epsilon_{j-1} + \rho_2\epsilon_{j-2} + z_j, \qquad j = 1, \ldots, 27$$

The second order AR model was chosen based on a correlogram of the maximum likelihood estimates, using (5.2), of yearly survival probabilities from the program MARK (White and Burnham, 1999). By examining the correlogram we are treating the MLE survival estimates as time series data instead of estimates of time series data. So, if there is insignificant correlation at certain lags, it is likely that the

corresponding AR coefficients will also be insignificant when they are simultaneously estimated with the covariate parameters.

The logit link function was chosen due to the fact that it is the most commonly used link in capture-recapture models. Capture-recapture data usually are not detailed enough to detect subtle differences in the shape of the link used to constrain the survival probability to (0, 1). Even in the logistic regression scenario, it is often hard to distinguish between different link functions. For probabilities in the range 0.1 to 0.9, McCullagh and Nelder (1989, pg 109) note that it is difficult to discriminate between probit and logit links based on goodness-of-fit tests and for probabilities near 0.5, all four of the common links for binary data are close to one another. Survival probabilities are usually not near the extremes of 0 or 1 for North American duck species, so, it is reasonable to use the logit link function for these data.

The Bayesian software WinBUGS (Spiegelhalter et al., 2000) was used to select the MCMC sample from the posterior distribution of $(\beta_0, \rho_1, \rho_2, \sigma, \beta)$. As was mentioned previously, there is only one parameter in which the full conditional is a standard density, therefore, a hybrid Gibbs sampler must be used. To accomplish this, winBUGS uses a Metropolis within Gibbs sampler where the proposal distribution is a normal distribution in which the variance adapts over the first 4,000 iterations to obtain an acceptance rate between 20% and 40%.

The priors chosen for the parameters were as follows:

$$(\beta_0, \beta_1)^T \sim N(\mathbf{0},\ 1/0.01\ \mathbf{I}), \qquad \sigma^{-2} \sim \Gamma(0.001,\ 0.001),$$
$$\rho_2 \sim U(-1,\ 1), \qquad \rho_1|\rho_2 \sim U(-(1-\rho_2),\ 1-\rho_2), \qquad \text{and}$$
$$\lambda_j \sim \text{i.i.d. } U(0,1) \quad j = 1, \ldots, 28.$$

These values were chosen to be sufficiently vague in order to induce little prior knowledge. The joint distribution of ρ_1 and ρ_2 was constructed to be a vague density over the region of stationarity with mean (0,0). A vague gamma distribution was

chosen for σ^{-2} in order to take advantage of the standard full conditional distribution of σ^2.

In order to select the sample, two independent chains of 15,000 iterations each were run following a burn-in period of 5,000 iterations to allow the normal proposal distribution to finish adapting. The chains appeared to have converged well before the end of the burn-in period. Figure 1 shows Gaussian kernel density estimates of the marginal posterior distributions for each of the parameters.

These results suggest that, although the posterior density of ρ_1 seems to be centered directly over 0, the majority of the posterior mass for ρ_2 seems to be located over negative values indicating that there seems to be a significant influence on the error terms at the second lag. This suggests that if survival is high in one year, it will be low in 2 years (lag 2). In addition, the posterior mass of σ seems to be located well away from 0, indicating that there is also a significant amount of random variation from year to year. The intercept parameter β_0 is also significantly greater than 0, which increases survival above approximately 0.5 on average. The posterior distribution of the slope parameter, β_1 appears to be highly concentrated near 0. While not directly comparable, the trend parameters and variance component are in qualitative agreement with the previous analysis. Figure 1 also illustrates the robustness of the marginal parameters to the presence or absence of the slope parameter. The marginal density estimates remain virtually unchanged. Posterior means, standard deviations, and 90% highest probability density (HPD) interval estimates are given in Table 2. The confidence intervals and approximate expected values support the conclusions that there exists a significant amount of variation not explained by the linear trend. There is also a high posterior probability that the slope parameter is approximately 0 and the second AR parameter is less than 0. In addition, since we have simulated values of β as well, we can estimate yearly survival as well. Figure 2 shows a plot of yearly survival with a 90% HPD confidence interval band for the intercept-only model.

The posterior densities remained virtually unchanged when the prior for the AR parameters is given by $\rho_1 \sim U(-2,\ 2)$ and $\rho_2|\rho_1 \sim U(-1,\ 1-|\rho_1|)$ as opposed to the priors used previously. For these priors, the parameter estimates and corresponding 90% HPD intervals for the intercept-only model were β_0: 0.606 (0.395, 0.811), ρ_1: 0.018 (-0.505, 0.516), ρ_2: -0.544 (-0.915, -0.185), σ: 0.617 (0.299, 0.921). The posterior distributions of the AR parameters seem to be robust to different non-informative priors.

5.5 Discussion

Software to implement the methodology described here for an AR(2) band recovery model is available at no charge at `www.stat.colostate.edu/~jah/`. It is relatively straightforward to modify and implement this software for specific problems. The software is written in winBUGS, software for the Bayesian analysis of statistical models using Markov chain Monte Carlo methods, which is available at no charge at `www.mrc-bsu.cam.ac.uk/bugs`.

Bayesian methodology allows for time-series modeling of capture-recapture data not previously available. In addition to the univariate time-series models considered here, the random effects models could also be expanded to allow for other forms of dependence. For example, it might be of interest to model recapture or recovery rates with AR random effects. In that case, the recovery or recapture parameters are treated the same as the survival procedures presented. Another example is the consideration of gender in survival models. Common practice is to include gender as a covariate. In an AR model, this would imply that unknown environmental factors have the same effect on survival of males and females in each year and the level of association of survival across time remains the same between males and females as well. This may be an unrealistic assumption, so, it might be wise to model separate AR errors for males and females. To account for correlated errors between sexes, a

multivariate AR process could be used with very little modification to the models proposed here.

Even though AR models can provide additional insight to the survival process, there are situations where estimation of the AR parameters may prove difficult. First, if there are fewer than 20 capture occasions in the data, there may not be enough data to greatly alter the prior distributions of the AR parameters. This is a problem often encountered in time series analysis. Secondly, if the recovery/recapture rates are very small there may be insufficient data to estimate AR parameters as well as covariate parameters. This second problem is common to all capture-recapture data analysis. Finally, if there is very little error variation, an AR model is unlikely to provide any additional information. This last situation is really not a problem though, since a biologist's goal is usually to model survival with covariates. If all of the error in the survival process is accounted for, one can be confident of having a good description of the survival process. The AR models are implemented to account for unobserved environmental variation.

An implication of using AR models with capture-recapture data is that the estimate of survival for any time period will have larger uncertainty than the simple covariate model. This variability is controlled by both variability of the white noise term in (5.4) and the AR parameters. For example, for an AR(1) model each random effect has a variance of $\sigma^2/(1-\rho^2)$. For σ held fixed, the variance of the random effect can get vary large as $|\rho| \to 1$. One can also observe, that for a fixed noise variance, the AR models will have larger variance than the independent random effects model.

One extension of the methodology described here is model selection. In general, model section is not an easy task for capture-recapture data in a Bayesian framework. Recently, King and Brooks (2003a) and King and Brooks (2003b) have explored using Reverse Jump MCMC procedures (Green, 1995) for capture-recapture

models with multiple strata and integrated recovery/recapture models. This provides the most promising solution, but, these procedures are not easily implemented for analysis on a regular basis. Another solution for Bayesian model selection is the Deviance Information Criterion (DIC) (Spiegelhalter et al., 2003). A nice feature of DIC is that the MCMC sample selected for parameter estimation can be used to construct a DIC score. Current formulations of the DIC, however, do not allow for distinguishing between different order AR processes for the capture-recapture models described here. If one wishes to fit an AR model to capture-recapture data, an initial step to select an appropriate order for the AR model is to fit a independent random effect model then plot a correlogram of the random effect point estimates and choose the order based on the plot. This will provide a conservative order for the model.

Overall, these models have the potential of providing wildlife biologists new insights into factors affecting survival for animals studied via capture-recapture studies.

Table 5.1: Northern Pintail recovery data for banding years 1955 - 1983. The R_i represent the number of banded ducks released each year. Birds were banded in January of each Banding Year

Banding		Year of Recovery																										
Year	R_i	55	56	57	58	59	60	61	62	63	64	65	66	67	68	69	70	71	72	73	74	75	76	77	78	79	80	81	82
55	270	7	6	3	2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
56	693	21	10	4	2	3	0	0	0	0	0	1	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0
57	1612	32	20	8	5	1	2	0	2	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
58	858	26	12	5	6	4	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
59	1471	21	18	6	5	0	0	1	0	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
60	1051	18	4	6	4	1	2	0	1	1	0	0	1	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0
61	796	24	6	4	0	3	3	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
62	277	10	9	6	6	4	1	2	4	2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
63	903	15	8	1	8	4	0	2	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
64	621	6	4	1	6	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
65	584	10	4	3	7	3	1	2	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
66	822	25	6	10	4	4	2	0	0	2	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
67	1344	28	27	8	11	3	1	4	1	2	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
68	566	10	13	6	2	2	1	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
69	481	9	7	3	2	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
70	695	11	11	5	2	2	1	1	0	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
71	632	22	10	2	4	0	1	2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
72	1114	21	11	8	3	5	3	2	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
73	639	9	10	10	2	3	0	2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
74	926	16	9	9	2	5	1	2	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
75	858	14	12	3	5	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
76	369	13	2	4	4	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
77	450	8	3	4	1	2	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
78	212	6	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
79	1680	18	28	8	4	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
80	421	14	1	2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
81	118	2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
82	60	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0

Table 5.2: Posterior means, standard deviations, and 90% highest probability density (HPD) intervals for the AR(2) model parameters.

Model	Parameter	Mean	St. Dev.	90% HPD* Interval
Intercept and slope	β_0	0.600	0.159	(0.390, 0.850)
	β_1	-0.007	0.026	(-0.046, 0.033)
	ρ_1	0.014	0.288	(-0.458, 0.485)
	ρ_2	-0.452	0.307	(-0.928, 0.004)
	σ	0.688	0.222	(0.336, 1.015)
Intercept only	β_0	0.612	0.140	(0.409, 0.857)
	ρ_1	-0.003	0.286	(-0.483, 0.456)
	ρ_2	-0.503	0.263	(-0.918, -0.109)
	σ	0.644	0.201	(0.330, 0.950)

* Estimated according to the algorithm presented by Chen et al. (2000).

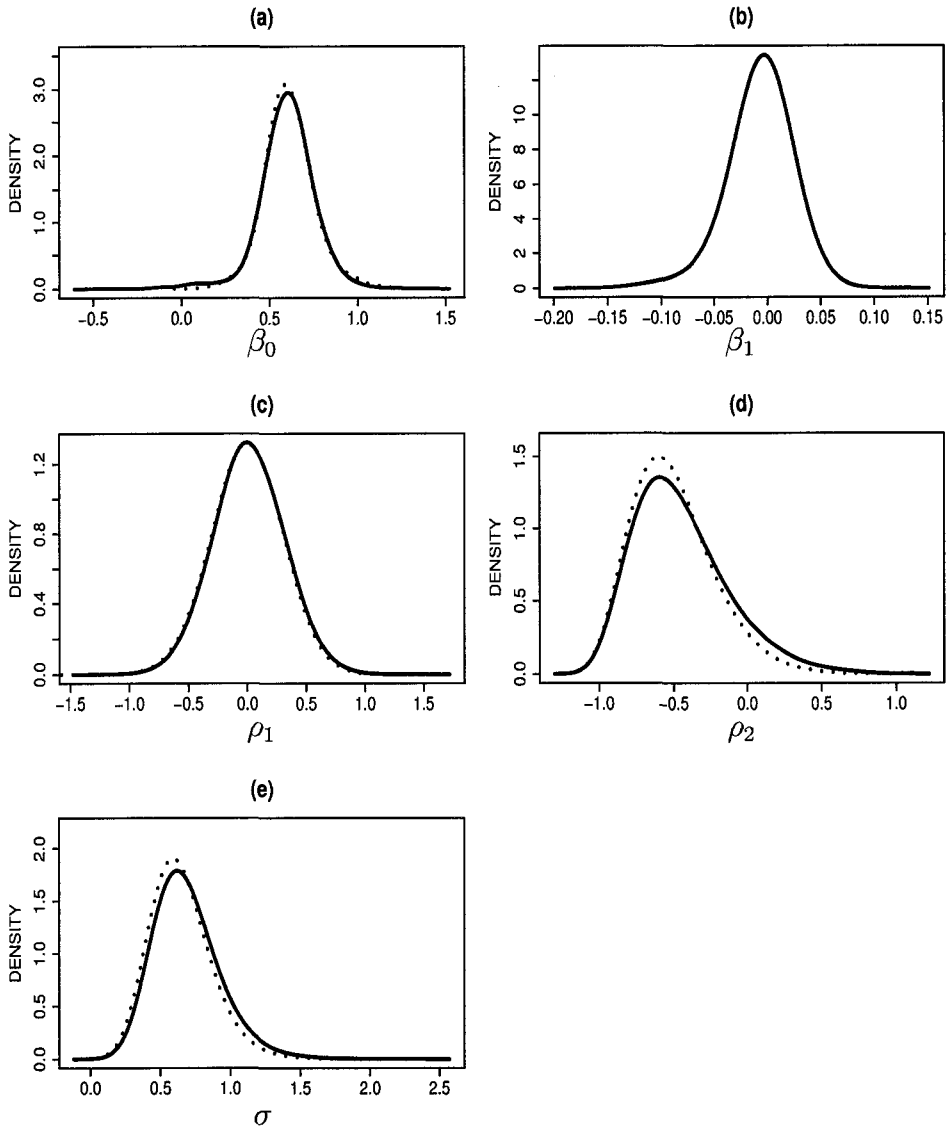

Figure 5.1: Marginal Posterior densities for (a): β_0, (b) β_1, (c): ρ_1, (d): ρ_2, and (e): σ from the Pintail data. The solid lines represent posterior densities from the time trend model, while the dotted lines represent the posterior densities when the trend parameter is absent.

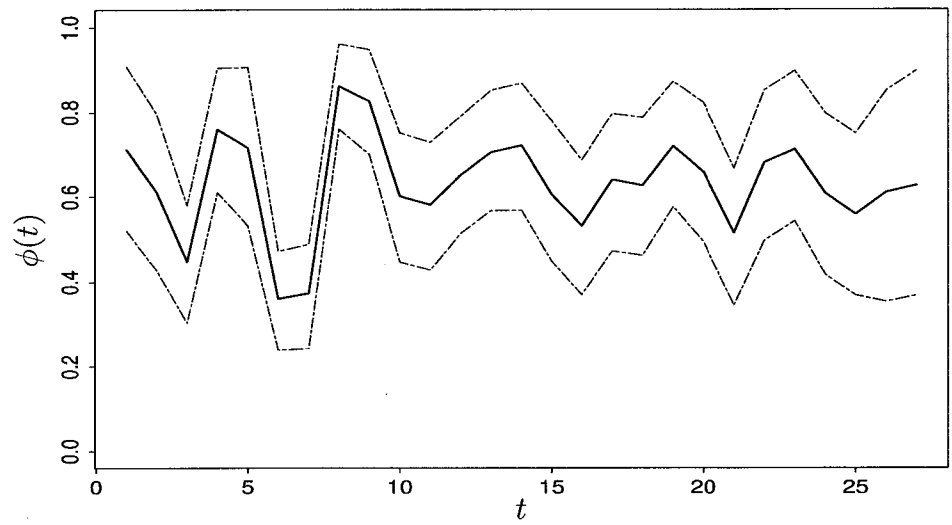

Figure 5.2: Plot of yearly survival estimates for Northern Pintail dataset with no linear time trend. The solid line is the estimated posterior mean survival and the dashed lines represent a 90% HPD interval.

Chapter 6

CONCLUSIONS AND FUTURE WORK

6.1 Discussion

In this dissertation, we have examined the analysis of discrete compositional data through the use of state-space models. Compositional data are non-negative multivariate observations that have been normalized so that the elements sum to one (or other constant value). Discrete compositional data are compositions formed from multivariate, integer observations. State-space models eliminate many of the problems encountered when analyzing discrete compositional data with the popular logistic-normal distribution. First, state-space models allow for zeros to be present in the data. In addition, the state-space models provide automatic variance inflation. High variances are often present in discrete compositional data when the total number of individuals comprising a compositional observation is small.

Two types of models were considered in this dissertation. First, models for traditional discrete compositional data were examined. Traditional discrete compositional models are designed for data for which all cells are assumed to have a positive value for the true unobserved composition. These models permit examination of various Markov properties due to their positive everywhere nature. The second type of model considered is the class of random effects capture-recapture models. As was illustrated, these models are essentially multi-way composition models with the exceptions that not all cells are observed or physically possible.

The discrete composition state-space models in this dissertation are constructed by providing a random effects formulation of a graphical chain model. The discrete

regression distribution, constructed in Section 2.4, is proposed as the base from which the compositional state-space models are constructed. The discrete regression (DR) distribution provides a general graphical chain model in which a set of discrete and/or continuous variables serve as covariates, or explanatory variables, for a set of discrete response variables. In the discrete regression distribution, the covariates as well as the response are modeled together as a multivariate observation. We have shown that the regression-like coefficients of the discrete regression distribution can be used to easily determine a graphical representation of the distribution from which complex conditional independence relationships can be derived.

A random effects formulation of the DR distribution is proposed for the analysis of discrete compositional data, by adding a random intercept term to the response portion of the model to allow for site-to-site variability. The variable response model plays the role of the unobserved "true" composition of individuals at each site. When there is only a single categorical response variable and the random effects are assumed to be multivariate normally distributed, the conditional response portion of the model is nearly identical to the independent observation model of Billheimer (1995, Ch. 4). The random effects graphical model allows us to model the joint distribution of several covariates and the categorical response of individuals simultaneously. In addition, the complex conditional relationships of this joint distribution can be represented by an extended independence chain graph based on the parameters of the DR model and the random effects. We provide necessary and sufficient conditions for constructing an independence graph for the random effects formulation of the DR model.

If one is interested primarily in determining general relationships between a categorical response and a set of covariates without regard to a specific site, such as the analysis of stream invertebrate feeding type in Section 3.5, then upon examination of the extended independence graph, one can simply ignore the random effects

vertex. This results from the fact that the random effects DR distribution can be marginalized over the random effects without disturbing the conditional independencies of the covariates and response variable. In this sense, we have also extended the single response DR model to allow sampling of individuals at a large number of sites, where individuals sampled at the same site are correlated, but, individuals sampled at different sites are independent.

The random effects graphical model formulation provides a direct extension for analysis of multi-way compositions. Multi-way discrete compositional data arise from the cross-classification of individuals according to several discrete variables. The state-space model of Billheimer (1995) can accommodate multi-way data, however, each cross-classified cell must be treated as a category in a single response model. A random effect formulation of the general DR distribution is not held to this constraint. Conditional independencies between the response variables can be added into the model. These independencies can either be used for inference through a model selection procedure, or, can be employed simply to obtain a more parsimonious model.

Unfortunately, not all multi-way random effect DR models can be marginalized over the random effects as in the single response model. We examined the class of *preservative* random effects DR models. We have shown that marginalization does not affect relationships between any response variables and covariates if the random effects model is a member of the preservative class. We conjecture that this class of models is equal to, not simply contained in, the class of models for which conditional independence relationships remain unchanged after marginalization. Showing equality of the two model classes essentially involves proving that being an element of the preservative class is not only sufficient, but, necessary for preserving conditional independencies between and within the response variables and covariates after marginalization over the random effects. There is no direct method to do this,

since there is no closed form expression for the integrated distribution. It may still be possible, however, to prove necessity though indirect means.

The final set of models considered in this dissertation are state-space models for capture-recapture data. We have extended the work of others in this area by modeling the unobserved survival rate as the realization of compositional autoregressive time series model. As was mentioned previously, the state-space capture-recapture models are similar to the multi-way state-space composition models except there are some cells which are not observable or not physically possible. This fact precludes examination of Markov properties in the same way as the traditional composition models, however, that type of inference is usually not applicable.

6.2 Future Work

There are several direct extensions to the research presented in this dissertation that we would like to pursue. The main emphasis of our future work can be grouped into two main categories, model selection methodology and spatio-temporal models. In this dissertation model selection was performed based HPD intervals or a combination of HPD intervals and the DIC criterion. While, this method can provide plausible models, a more rigorous methodology should be investigated. Secondly, models that allow spatial and temporal association for the response model as well as the covariate model in the random effects DR distribution should be considered.

6.2.1 Model Selection

Model selection methodology in a Bayesian framework is usually based on the Bayes factor for comparing two models based on evidence contained in the data. Kass and Raftery (1995) provide a comprehensive review of Bayes factors. The Bayes factor between two models M_1 and M_2 is given by the ratio of likelihoods $B_{12} = f(Data|M_1)/f(Data|M_2)$. The Bayes factor B_{12} is equal to the ratio of posterior model probabilities $f(M_1|Data)/f(M_2|Data)$ if the prior distribution on

the model space is uniform. If there are many models to compare, say K, the posterior model probabilities $f(M_k|Data)$, $k = 1, \ldots, K$ are often used for selection. Unfortunately, posterior model probabilities do not exist, in closed form, for any model proposed in this dissertation. This implies that approximation methods need to be investigated for these models.

For the response portion of the single response composition models and the AR capture-recapture models, selection is simplified somewhat due to the fact that model selection only involves parameters. The random effects are present in all of these models and different models are obtained by setting covariate or AR coefficients to zero. One method for approximation of the Bayes factor that would be applicable in this case is the generalized version of the Savage-Dickey density ratio (Verdinelli and Wasserman, 1995). The Savage-Dickey density ratio allows an approximation of the posterior model probabilities based on an MCMC simulation from the saturated model with all of the parameters present.

A Bayesian methodology for selection of the covariate portion of the discrete regression graphical model has been proposed by Madigan and Raftery (1994) and Dellapotas and Forster (1999), for purely discrete covariates, and Guidici and Green (1999) for the purely continuous case. To date, a methodology for Bayesian determination of a CG model has not be examined.

Model determination of response portion of the multi-way model is more challenging due to the fact that changing models involves selection of random effects as well as interaction parameters. The reverse jump MCMC of Green (1995) may be a viable solution. The reverse jump MCMC algorithm is an extension of the usual Metropolis-Hastings algorithm, in Section 1.5.1, that allows the chain to move over different parameter spaces. The problem with the multi-way models is that proposal of many random effects will lead to many rejections and an inefficient sampler. The proposal distribution construction methods of Brooks et al. (2003) may provide ways to overcome this problem.

6.2.2 Spatio-Temporal Models

The next area of future research we would like to explore for discrete compositional data is the explicit modeling of spatial-temporal association within the state-space model. Billheimer (1995) and Billheimer and Guttorp (1997) first considered spatial models for discrete composition by letting the error terms in the random effects DR distribution (3.10) follow a lattice Gaussian random field. Tjelmeland and Lund (2003) followed by proposing a continuous spatial domain model for continuous compositions by modeling the additive logistic ratio (ALR) (1.4) transformed data as a continuous Gaussian random field.

We would like to extend our research by examining the Markov properties for these spatial models. This will provide insight into spatial graphical models for contingency table data in general. On possible approach is to consider a *super* graphical model. The super interaction graph is constructed by constructing an extended independence graph for each site, then the random effects are connected by a model such as a lattice spatial model. One could analyze the properties of this giant graphical model. This approach is similar to the methodology proposed by Fienberg and Kim (1999) for combining log-linear graphical models. This is essentially the task we are performing by integrating a spatial process(es) as part of the state-space model. A graphical model describing spatial between site relationships of the random effects and covariates is combined with a graphical model for the random effects, covariates, and response variables within each site. Therefore, this methodology provides promise for either lattice or continuous spatial models.

Dahlhaus (2000) provides a methodology for analyzing graphical models for multivariate Gaussian data collected in a vector time series. When data are collected as a time series, the spectral matrix of the vector process plays the same role as the inverse correlation matrix. These same spectral methods could be extended to analyze spatial models within a discrete regression framework.

Appendix A

WINBUGS CODE FOR COMPOSITIONAL DATA AND AR(2) CAPTURE-RECAPTURE MODELS

A.1 WinBUGS Code for a Single Composition Model

Here we present the WinBUGS code used for the analysis of stream invertebrate functional groups in Chapter 3. The model statement includes the multinomial cell count model, prior distributions for all parameters, and goodness-of-fit statistic for the Bayesian P-value calculation. In addition, the likelihood model is specified for the missing observations that are to be predicted.

```
model
{
########################################
#   This portion defines the likelihood
########################################

for(i in 1:n){
  for(j in 1:J){

### Observed data model ###
    y[i,1:J] ~ dmulti(pr[i,1:J], N[i])

### Replicated data mode for Bayesian P-value ###
    y.rep[i,1:J] ~ dmulti(pr.cut[i,1:J], N[i])

### Define category composition model ###
    log(lambda[i,j]) <- phi[i,j]
    pr[i,j] <- lambda[i,j]/sum(lambda[i,])
    mu.phi[i,j] <- b.0[j] + inprod(x.adj[i,1:P], b[1:P,j])
    pr.cut[i,j] <- cut(pr[i,j])

### Calculate fit statistics for each cell ###
    D.obs1[i,j] <- pow(sqrt(y[i,j])-sqrt(N[i]*pr[i,j]), 2)
    D.rep1[i,j] <- pow(sqrt(y.rep[i,j])-sqrt(N[i]*pr.cut[i,j]), 2)
  }
  D.obs2[i] <- sum(lr2.obs1[i,1:J])
  D.rep2[i] <- sum(lr2.rep1[i,1:J])

### Define random effects model for hierarchical center ###
```

```
  phi[i,2:J] ~ dmnorm(mu.phi[i, 2:J], T[1:K, 1:K])
  phi[i,1] <- 0

### standardize covariates ###
  for(p in 1:P){
    x.adj[i,p] <- (x[i,p]-mean(x[,p])) / sd(x[,p])
  }
}
##########################################

##########################################
#   This portion calculates the P-value
##########################################
D.obs <- sum(D.obs2[1:n])
D.rep <- sum(D.rep2[1:n])
P.val <- step(D.rep - D.obs)
##########################################

##########################################
#   This portion defines prediction likelihood
##########################################
for(i in 1:n2){
  for(j in 1:J){
        log(lambda[(i+n),j]) <- phi[(i+n),j]
        pr2[i,j] <- lambda[(i+n),j]/sum(lambda[(i+n),])
        mu.phi[(i+n),j] <- b.0[j] + inprod(x.adj[(i+n),1:P], b[1:P,j])
  }
  y[(i+n),1:J] ~ dmulti(pr2[i,1:J], N[(i+n)])
  phi[(i+n),2:J] ~ dmnorm(mu.phi[(i+n), 2:J], T[1:K, 1:K])
  phi[(i+n),1] <- 0
  for(p in 1:P){
    x.adj[(i+n),p] <- (x[(i+n),p]-mean(x[,p])) / sd(x[,p])
  }
}
##########################################

#########################################
#   This section defines the Prior dist.
#########################################
for(p in 1:P){
    for(j in 2:J){b[p,j] ~ dnorm(0, 0.01)}
    b[p,1] <- 0
} ## same prior for all b's
for(j in 2:J){
b.0[j] ~ dnorm(0, 0.01)
}
b.0[1] <- 0
T[1:K, 1:K] ~ dwish(R[1:K, 1:K], K)
#########################################
}
```

A.2 WinBUGS Code for a Multi-Way Composition Model

The example code provided here is for the dependent response model with independent error structure that is analyzed in Section 4.5. If an independent response is desired, one can simply eliminate all terms in the code associated with the $\phi_{\{B,I\}}$ terms. If a completely dependent model with dependent errors is desired, the code in the previous section can be used.

```
model
{
########################################
#   This portion defines the likelihood
########################################
for(s in 1:S){
### Observed data model ###
    y[s,1:6] ~ dmulti(pr[s,1:6],y[s,7])
### Replicated data model for Bayesian P-value ###
    y.rep[s,1:6] ~ dmulti(pr.c[s,1:6],y[s,7])


    for(i in 1:2){
        for(j in 1:3){
### Define log-linear mode for cell probabilities ###
            log(lambda[s,(3*(i-1)+j)]) <- phi.B[s,i] + phi.I[s,j]
                                          + phi.BI[s,(3*(i-1)+j)]
            pr[s,(3*(i-1)+j)] <- lambda[s,(3*(i-1)+j)]/sum(lambda[s,])
            pr.c[s,(3*(i-1)+j)] <- cut(pr[s,(3*(i-1)+j)])

### Calculate goodness-of-fit test statistics ###
            D1[s,(3*(i-1)+j)] <-
                pow(sqrt(y[s,(3*(i-1)+j)])
                        -sqrt(y[s,7]*pr[s,(3*(i-1)+j)]), 2)
            D1.rep[s,(3*(i-1)+j)] <-
                pow(sqrt(y.rep[s,(3*(i-1)+j)])
                        -sqrt(y[s,7]*pr.c[s,(3*(i-1)+j)]), 2)
        }
    }
    D2[s] <- sum(D1[s,])
    D2.rep[s] <- sum(D1.rep[s,])

### Define random effects models for fully dependent response ###
###       with independent error structure                    ###
    phi.B[s,2] ~ dnorm(mu.B[s], T.B)
            phi.B[s,1] <- 0
    phi.I[s,2:3] ~  dmnorm(mu.I[s,1:2], T.I[1:2,1:2])
            phi.I[s,1] <- 0
    phi.BI[s,5:6] ~ dmnorm(mu.BI[s,1:2], T.BI[1:2,1:2])
        for(j in 1:4){phi.BI[s,j]<- 0}
```

```
    mu.B[s] <- a.B + wsarea.B*z[s,1] + precip.B*z[s,2] + elev.B*z[s,3]
                              + ph.B*z[s,5] + cl.B*z[s,6]
                    + ptl.B*z[s,8]
    for(j in 1:2){
        mu.I[s,j] <- a.I[j] + wsarea.I[j]*z[s,1] + precip.I[j]*z[s,2]
                            + elev.I[j]*z[s,3] + ph.I[j]*z[s,5]
                            + cl.I[j]*z[s,6] + ptl.I[j]*z[s,8]
    }
    for(j in 1:2){
        mu.BI[s,j] <- a.BI[j] + wsarea.BI[j]*z[s,1] + precip.BI[j]*z[s,2]
                              + elev.BI[j]*z[s,3] + ph.BI[j]*z[s,5]
                              + cl.BI[j]*z[s,6] + ptl.BI[j]*z[s,8]
    }

### Normalize covariates ###
    for(p in 1:8){
        z[s,p] <- (x[s,p]-mean(x[,p]))/sd(x[,p])
    }
}
##########################################

##########################################
#   The Bayesian P-value is approximated with this step
##########################################
D <- sum(D2[1:S])
D.rep <- sum(D2.rep[1:S])
P.val <- step(D.rep-D)
##########################################

##########################################
#   This section states the prior dist.
##########################################
a.B ~ dnorm(0, 0.01)
wsarea.B ~ dnorm(0, 0.01)
precip.B ~ dnorm(0, 0.01)
elev.B ~ dnorm(0, 0.01)
ph.B ~ dnorm(0, 0.01)
cl.B ~ dnorm(0, 0.01)
ptl.B ~ dnorm(0, 0.01)
for(j in 1:2){
    a.I[j] ~ dnorm(0, 0.01)
    wsarea.I[j] ~ dnorm(0, 0.01)
    precip.I[j] ~ dnorm(0, 0.01)
    elev.I[j] ~ dnorm(0, 0.01)
    ph.I[j] ~ dnorm(0, 0.01)
    cl.I[j] ~ dnorm(0, 0.01)
    ptl.I[j] ~ dnorm(0, 0.01)
    a.BI[j] ~ dnorm(0, 0.01)
    wsarea.BI[j] ~ dnorm(0, 0.01)
    precip.BI[j] ~ dnorm(0, 0.01)
    elev.BI[j] ~ dnorm(0, 0.01)
```

```
    ph.BI[j] ~ dnorm(0, 0.01)
    cl.BI[j] ~ dnorm(0, 0.01)
    ptl.BI[j] ~ dnorm(0, 0.01)
}

T.B ~ dgamma(0.01, 0.01)
T.I[1:2,1:2] ~ dwish(R[1:2,1:2], 2)
T.BI[1:2,1:2] ~ dwish(R[1:2,1:2], 2)
########################################
}
```

A.3 WinBUGS Code for an AR(2) Capture-Recapture Model

Here we present the WinBUGS code used in the example analysis of a long term capture-recapture data set of Pintails in Chapter 5. The model statement includes likelihood specification, random effects distribution for an AR(2) model, and the prior distributions used for the analysis.

```
model;
{
##########################################
#   This portion defines the likelihood
##########################################
for(i in 1:I){ D[i, 2:(I+2)] ~ dmulti(C[i,], D[i, 1]); }
# Cell probabilities ####################
for(i in 1:(I-1)){
    lphi[i] <- log(phi[i])
    logit(phi[i]) <- beta + e[i]
    for(j in (i+1):I){
            C[i, j] <- lambda[j]*exp(sum(lphi[i:(j-1)]))
    }
    for (j in 1:i){
        C[i+1, j] <- 0
    }
}
for(i in 1:I){
    C[i, i] <- lambda[i]
    C[i, I+1] <- 1 - sum(C[i, 1:I])
}
##########################################

##########################################
#   This portion defines the model for epsilon
##########################################

e[1] ~ dnorm(mu[1], tau1)
e[2] ~ dnorm(mu[2], tau2)
mu[1] <- 0
mu[2] <- (rho[1]/(1-rho[2]))*e[1]
tau1 <- ((1+rho[2])/(1-rho[2]))*((1-rho[2])*(1-rho[2]) - rho[1]*rho[1])*tau
tau2 <- tau*(1 - rho[2]*rho[2])
for(i in 3:(I-1)){
    e[i] ~ dnorm(mu[i],tau)
    mu[i] <- rho[1]*e[i-1] + rho[2]*e[i-2]
}
sigma <- 1/sqrt(tau)
##########################################

##########################################
#  This section states the prior dist.
##########################################
```

```
# for(i in 1:2){beta[i] ~ dnorm(0, 0.01)}
beta ~ dnorm(0, 0.01)
for(i in 1:I){lambda[i] ~ dunif(0, 1)}
tau ~ dgamma(0.001, 0.001)

## Prior for rho is approx. uniform on the AR(2) stationary triangle ##
rho[1] ~ dunif(l, u)
u <- abs(1 - rho[2])
l <- -u
rho[2] ~ dunif(-1, 1)
##########################################

}
```

Bibliography

Abbitt, P. and Breidt, F. J. (2001). A hierarchical model for estimating distribution profiles of soil texture. *Case Studies in Bayesian Statistics*, 5:263–278.

Aitchison, J. (1982). The statistical analysis of compositional data (with discussion). *Journal of the Royal Statistical Society - B*, 44:139–177.

Aitchison, J. (1986). *The Statistical Analysis of Compositional Data.* Chapman and Hall, New York.

Aitchison, J. and Shen, S. M. (1980). Logistic-normal distribution: Some properties and uses. *Biometrika*, 67:261–272.

Andersson, S. A., Madigan, D., and Pearlman, M. D. (1997). On the Markov equivalence of chain graphs, undirected graphs, and acyclic digraphs. *Scandinavian Journal of Statistics*, 24:81–120.

Barnard, J., McCulloch, R., and Meng, X. L. (2000). Modeling covariance matrices in terms of standard deviations and correlations, with application to shrinkage. *Statistica Sinica*, 10:1281–1311.

Barndorff-Nielsen, O. and Jørgensen, B. (1991). Some parametric models on the simplex. *Journal of Multivariate Analysis*, 39:106–116.

Barry, S. C., Brooks, S. P., Catchpole, E. A., and Morgan, B. J. T. (2003). The analysis of ring recovery data using random effects. *Biometrics*, 59:54–65.

Besag, J., Green, P., Higdon, D., and Mengersen, K. (1995). Bayesian computation and stochastic systems (with Discussion). *Statistical Science*, 10:3–66.

Billheimer, D. (1995). *Statistical Analysis of Biological Monitering Data: State-Space Models for Species Compositions.* PhD thesis, University of Washington.

Billheimer, D. and Guttorp, P. (1997). Natural variablity in benthis species composition in the Delaware Bay. *Environmental and Ecological Statistics*, 4:95–115.

Billheimer, D., Guttorp, P., and Fagen, W. F. (2001). Statistical interpretation of species composition. *Journal of the American Statistical Association*, 96:1205–1214.

Birch, M. W. (1963). Maximum-likelihood in three-way contingency tables. *Journal of the Royal Statistical Society, B*, 25:220–233.

Brockwell, P. J. and Davis, R. A. (1996). *Introduction to Time Series and Forecasting.* Springer, New York.

Brooks, S. P., Catchpole, E. A., and Morgan, B. J. T. (2000a). Bayesian animal survival estimation. *Statistical Science*, 15:357–376.

Brooks, S. P., Catchpole, E. A., and Morgan, B. J. T. (2002). "Bayesian methods for analysing ringing data" In Statistical Analysis of Data from Marked Bird Populations. *Journal of Applied Statistics*, 29(1-4):187–206. B.J.T. Morgan and D.L. Thomson, eds.

Brooks, S. P., Catchpole, E. A., Morgan, B. J. T., and Barry, S. C. (2000b). On the Bayesian analysis of ring recovery data. *Biometrics*, 56:951–956.

Brooks, S. P., Guidici, P., and Roberts, G. O. (2003). Efficient construction of reverse jump Markov chain Monte Carlo proposal distributions. *Journal of the Royal Statistical Society - B*, 65:3–55.

Brownie, C., Anderson, D. R., Burnham, K. P., and Robson, D. S. (1985). Statistical inference from band recovery data - a handbook. United States Department of the Interior Fish and Wildlife Service, Resource Publication 156.

Brunsdon, T. M. and Smith, T. (1998). The time series analysis of compositonal data. *Journal of Official Statistics*, 14:237–253.

Buckland, S. T., Goudie, I. B. J., and Borchers, D. L. (2000). Wildlife population assessment: past developments and future directions. *Biometrics*, 56:1–12.

Burnham, K. P. (2000). On random effects models for capture-recapture. Preprint available from author, Colorado State University.

Burnham, K. P. and White, G. C. (2002). "Evaluation of some random effects methodology applicable to ringing data" In Statistical Analysis of Data from Marked Bird Populations. *Journal of Applied Statistics*, 29(1-4):245-262. B.J.T. Morgan and D.L. Thomson, eds.

Chen, M. H., Shao, Q. M., and Ibrahim, J. G. (2000). *Monte Carlo Methods in Bayesian Statistics.* Springer-Verlag, New York.

Christensen, R. (1990). *Log-linear models.* Springer, New York.

Cormack, R. M. (1964). Estimates of survival from the sighting of marked animals. *Biometrika*, 51:429–438.

Cox, D. R. and Wermuth, N. (1993). Linear dependencies represented by chain graphs. *Statistical Science*, 8:204–218.

Cummins, K. W. and Klug, M. J. (1979). Feeding ecology of stream invertebrates. *Annual Review of Ecology and Systematics*, 10:147–172.

Dahlhaus, R. (2000). Graphical interaction models for multivariate time series. *Metrika*, 51:157–172.

Darroch, J. N., Lauritzen, S. L., and Speed, T. S. (1980). Markov fields and log-linear interaction models for contingency tables. *The Annals of Statistics*, 8:522–539.

Dellapotas, P. and Forster, J. J. (1999). Markov chain Monte Carlo model determination for hierarchical and graphical log-linear models. *Biometrika*, 86:615–633.

Dempster, A. P. (1972). Covariance selection. *Biometrics*, 28:157–175.

Dey, D. K., Ghosh, S. K., and Mallick, B. K. (2000). *Generalized Linear Models: A Bayesian Perspective.* Marcel Dekker, Inc., New York.

Dolédec, S., Chessel, D., ter Braak, C., and Champely, S. (1996). Matching species traits to environmental variables: A new three-table ordination method. *Environmental and Ecological Statistics*, 3:143–166.

Dominici, F. (2000). Combining contingency table data with missing observations. *Biometrics*, 56:546–553.

Durbin, J. and Koopman, S. J. (2000). Time series analysis of non-Gaussian observations based on state space models from both classical and Bayesian perspectives (with discussion). *Journal of the Royal Statistical Society - B*, 62:3–56.

Edwards, D. (1990). Hierarchical interaction models. *Journal of the Royal Statistical Society, B*, 52:3–20.

Fienberg, S. E. and Kim, S. H. (1999). Combining conditional log-linear structures. *Journal of the American Statistical Association*, 94:229–239.

Franklin, A. B., Anderson, D. R., and Burnham, K. P. (2002). Estimation of long-term trends and variation in avian survival probabilities using random effects

models" In Statistical Analysis of Data from Marked Bird Populations. *Journal of Applied Statistics*, 29(1-4):267-289. B.J.T. Morgan and D.L. Thomson, eds.

Freeman, M. and Tukey, J. (1950). Transformations related to the angular and square root. *Annals of Mathematical Statistics*, 21:607–611.

Frydenburg, M. (1990). The chain graph Markov property. *Scandinavian Journal of Statistics*, 17:333–353.

Gelman, A., Meng, X.-L., and Stern, H. (1996). Posterior predictive assessment of model fitness via realized discrepancies. *Statistica Sinica*, 6:733–807.

Gelman, A. and Rubin, D. B. (1993). *in* Discussion on the meeting on the Gibbs sampler and other MCMC methods. *Journal of the Royal Statistical Society-Series B*, 55:73.

Gilks, W. and Wild, P. (1992). Adaptive rejection sampling for Gibbs sampling. *Applied Statistics*, 2:337–348.

Green, P. J. (1995). Reversible jump Markov Chain Monte Carlo computations and Bayesian model selection. *Biometrika*, 82:711–732.

Guidici, P. and Green, P. J. (1999). Decomposable graphical Gaussian model determination. *Biometrika*, 86:785–801.

Gupta, R. and Richards, D. (2001). The history of the Dirichlet and Liouville distributions. *International Statistical Review*, 69:433–446.

Hammersley, J. M. and Clifford, P. (1971). Markov fields on finite graphs and lattices. Unpublished manuscript.

Hastings, W. K. (1970). Monte Carlo sampling methods using Markov chains and their applications. *Biometrika*, 57:97–109.

Huerta, G. and West, M. (1999). Priors and component structures in autoregressive time series models. *Journal of the Royal Statistical Society - B*, 61:881–899.

Iyenger, M. and Dey, D. K. (2002). A semiparametric model for compositional data analysis in the presence of covariates on the simplex. *Test*, 11:303–315.

Johnson, R. and Wichern, D. (1992). *Applied Multivariate Statistical Analysis.* Prentice-Hall, New Jersey. 642pp.

Kass, R. E. and Raftery, A. E. (1995). Bayes factors. *Journal of the American Statistical Association*, 90:773–795.

Kiiveri, H. T., Speed, T. S., and Carlin, J. B. (1984). Recursive causal models. *Journal of the Australian Mathematical Society*, 36:30–52.

King, R. and Brooks, S. P. (2003a). Bayesian model discrimination for multiple strata capture-recapture. *Biometrika*, 89:785–806.

King, R. and Brooks, S. P. (2003b). Models selection for integrated recovery/recapture data. *Biometrics*, 58:841–851.

Lauritzen, S. L. (1996). *Graphical Models.* Claredon Press.

Lauritzen, S. L. and Wermuth, N. (1989). Graphical models for associations between variables, some of which are qualitative and some quantitative. *The Annals of Statistics*, 17:31–57.

Lebreton, J., Burnham, K. P., Clobert, J., and Anderson, D. R. (1992). Modeling survival and testing biological hypotheses using marked animals: a unified approach with case studies. *Ecological Monographs*, 62:67–118.

Legendre, P., Galzin, R., and Harmelin-Vivien, M. (1997). Relating behavior to habitat: Solutions to the fourth-corner problem. *Ecology*, 78:547–562.

Levitz, M., Pearman, M. D., and Madigan, D. (2001). Separation and completeness properties for AMP chain graph Markov models. *The Annals of Statistics*, 29:1751–1784.

Lindsey, J. K. (1999). *Models for Repeated Measurements.* Oxford University Press Inc., New York.

Madigan, D. and Raftery, A. E. (1994). Model selection and accounting for model uncertainty in graphical models ussing Occam's window. *Journal of the American Statistical Association*, 89:1535–1546.

McCormack, F., Hughs, R., Kaufmann, P., Peck, D., Stoddard, J., and Herlihy, A. (2001). Development of an index of biotic integrity for the Mid-Atlantic Highlands Region. *Transactions of the North American Fisheries Society*, 130:857–877.

McCullagh, P. (1983). Quasi-likelihood functions. *The Annals of Statistics*, 11:59–67.

McCullagh, P. and Nelder, J. (1989). *Generalized Linear Models*, volume 37 of *Monographs on Statistics and Applied Probability.* Chapman & Hall/CRC, 2nd edition.

Metropolis, N., Rosenbluth, A., Rosenbluth, M., Teller, A., and Teller, M. (1953). Equations of state calculations by fast computiong machines. *Journal of Chemical Physics*, 21:1087–1091.

Pearl, J. and Paz, A. (1987). Graphoids: A graph based logic for reasoning about relevancy relations. In *Advances in Artificial Intelligence - II.* American Association of Artificial Intelligence, (eds. B.D. Boulay, D. Hogg, and L. Steel), pp 357-363. North-Holland, Amsterdam.

Pearson, K. (1897). Mathematical contributions to the theory of evolution. on a form of spurious correlation which may arise when indices are used in the measurment of organs. *Proceedings of the Royal Society*, 60:489–498.

Pizzi, D. (2002). Multi-scale relationships between environmental factors and macroinvertebrates in colorado and oregon watersheds. Master's thesis, Colorado State University.

Poff, N. (1997). Landscape filters and species traits: Towards mechanistic understanding and prediction in stream ecology. *Journal of the North American Benthological Society*, 16:391–409.

Poff, N. and Allen, J. (1995). Functional-organization of stream fish assemblages in relation to hydrological variability. *Ecology*, 76:606–627.

Poole, D. and Zeh, J. E. (2002). Estimation of adult bowhead whale survival rates using Markov chain Monte Carlo methods. *Journal of Agriculture, Biological, and Environmental Statistics*, 7:1–13.

Raftery, A. and Lewis, S. (1992). Comment: One long run with diagnostics: Implimentation strategies for Markov Chain Monte Carlo. *Statistical Science*, 7:493–497.

Ricklefs, R. (1990). *Ecology.* Feeman, New York.

Robert, C. P. and Casella, G. (1999). *Monte Carlo Statistical Methods.* Springer-Verlag, New York.

Rohatgi, V. (1976). *An Introduction to Probability Theory and Mathematical Statistics.* John Wiley and Sons, New York.

Smith, B. and Rayens, W. (2002). Conditional generalized Liouville distributions on the simplex. *Statistics*, 36:185–194.

Speed, T. P. and Kiiveri, H. T. (1986). Gaussian Markov distributions over finite graphs. *The Annals of Statistics*, 14:138–150.

Spiegelhalter, D., Thomas, A., and Best, N. (2000). WinBUGS version 1.3, User Manual. MRC Biostatistics Unit, Institute of Public Health, Cambridge UK.

Spiegelhalter, D. J., Best, N. G., Carlin, B. P., and van der Lind, A. (2003). Bayesian measures of complexity and fit. *Journal of the Royal Statistical Society - B*, 64:583–639.

Stephens, M. (1982). Use of the von Mises distribution to analyze continuous proportions. *Biometrika*, 69:197–203.

Studený, M. and Bouckaert, R. R. (1998). On chain graph models for description of conditional independence. *The Annals of Statistics*, 26:1434–1495.

Sun, D. and Berger, J. O. (1998). Reference priors with partial information. *Biometrika*, 85:55–71.

ter Braak, C. (1985). Correspondence analysis of incidence and abundence data: Properties in terms of a unimodal response model. *Biometrics*, 41:859–873.

Tierney, L. (1994). Markov chains for exploring posterior distributions. *The Annals of Statistics*, 22:1701–1762.

Tjelmeland, H. and Lund, K. (2003). Baysian modeling of spatial compositional data. *Journal of Applied Statistics*, 30:87–100.

Verdinelli, I. and Wasserman, L. (1995). Computing Bayes factors using a generalization of the Savage-Dickey density ratio. *Journal of the American Statistical Association*, 90:614–618.

Vounatsou, F. and Smith, A. F. M. (1995). Bayesian analysis of ring recovery data via Markov Chain Monte Carlo simulation. *Biometrics*, 51:687–708.

Wermuth, N. (1976). Analogies between multiplicative models in contingency tables and covariance selection. *Biometrics*, 32:95–108.

Wermuth, N. (1980). Linear recursive equations, covariance selection and path analysis. *Journal of the American Statistical Association*, 75:963–972.

Wermuth, N. and Lauritzen, S. L. (1990). On substantive research hypotheses, conditional independence graphs and graphical chain models. *Journal of the Royal Statistical Society, B*, 52:21–50.

White, G. C. and Burnham, K. P. (1999). Program MARK: survival estimation from populations of marked animals. *Bird Study*, 46 suppl:120–139.

Whittaker, J. (1990). *Graphical Models in Applied Multivariate Statistics.* Wiley, Chichster.

Wright, S. (1934). The method of path coefficients. *The Annals of Mathematical Statistics*, 5:161–215.

Zeger, S. L. and Karim, M. R. (1991). Generalized linear models with random effects: a Gibbs sampling approach. *Journal of the American Statistical Association*, 86:79–86.