

Biological monitoring: A Bayesian Model for Multivariate Compositional Data

Devin S. Johnson
National Oceanic and Atmospheric Administration

Jennifer A. Hoeting and N. LeRoy Poff *
Colorado State University

October 13, 2005

Abstract

We develop a model to relate a multivariate compositional response to a number of covariates. We propose a new graphical model, called the Random Effects Discrete Regression (REDR) model, which allows for examination of the complex conditional relationships between a set of covariates and multiple discrete response variables. Our approach offers a number of advantages over previous approaches and allows for a wide range of inferences. Relationships between compositional observations can be evaluated through a set of interaction parameters and inference about the influence of covariates is possible through a set of regression coefficients. The model also allows for examination of relationships between the covariates via another set of interactions. Parameter inference via Bayesian methods and MCMC is discussed. The proposed model and MCMC methods are used to examine the relationship between compositional observations of two characteristics of fish species and a number of covariates. These relationships are of interest to the U.S. Environmental Protection Agency for stream monitoring.

KEYWORDS: *Graphical chain models, compositional data, logistic normal distribution, MCMC, random effects*

1 Introduction

In this paper we propose a class of models for compositional data based on traditional graphical models. A compositional observation $\mathbf{P} = (P_1, \dots, P_D)$ possess the two constraints: $\sum_j P_j = 1$ and $P_j \geq 0$ for $j = 1, \dots, D$. Compositional analysis is preferred to the unconstrained positive multivariate observations if relative size is a more appropriate measure than the observed counts for each

*Devin S. Johnson is Statistician, National Marine Mammal Laboratory, Alaska Fisheries Science Center, NOAA, 7600 Sand Point Way NE, Seattle WA, 98115, USA; email address: devin.johnson@noaa.gov. Jennifer A. Hoeting is Associate Professor, Department of Statistics, Colorado State University, Fort Collins, CO 80523-1877, USA; email address: jah@stat.colostate.edu. LeRoy Poff is Associate Professor, Department of Biology, Colorado State University, Fort Collins, CO 80523, USA; email address: poff@lamar.colostate.edu. Research supported by the U.S. Environmental Protection Agency as part of the STAR Research Assistance Agreement CR-829095 (Johnson and Hoeting) and STAR grant R8286301 (Poff).

vector element. A discrete multivariate compositional observation arises when numerous sampled individuals at a site are cross classified according to a number of categorical variables forming a (possibly) high dimensional contingency table. The composition of interest is the probability that a randomly selected individual falls into a particular cross classification. The goal herein is to relate a multivariate compositional response to a number of covariates. Our approach offers a number of advantages over previous approaches and allows for a wide range of inferences including inferences on the relationship between various compositional response variables as well as inferences on the relationships between compositions and the explanatory variables.

Aitchison (1986) provides an overview of models for compositional data. There are a number of challenges in modeling compositional data. 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.

The logistic-normal distribution has been proposed as a suitable distribution for compositional data in that it has a sensible covariance structure and offers a suitably rich family of distributions with which to model compositional data. If the random vector $\mathbf{X} = (X_1, \dots, X_{D-1})$, where $X_j = \log(P_j/P_D)$, follows a multi-variate normal distribution, then $\mathbf{P} = (P_1, \dots, P_D)$ follows a logistic normal distribution (Aitchison and Shen, 1980; Aitchison, 1982). This distribution is not fully suitable for the problem considered here, however, because our data are discrete compositional data with extra variability due to sampling error. Also, the logistic normal density is not defined when the proportion in a particular class, P_j , equals 0, a common occurrence when the raw data are counts. A solution to this second issue was proposed by Ghosh et al. (2002) who adopt a mixture distribution consisting of a logistic normal distribution and a distribution that is degenerate at 0.

Billheimer (1995), Billheimer and Guttorp (1997), and Billheimer et al. (2001) proposed a model for a univariate compositional response which eliminates the problems in the logistic normal model caused by zeroes and low abundance counts when modeling species compositions. We build upon and extend their model to allow for multivariate compositions.

Graphical models are distributions for analyzing the conditional relationships of a Markov random field (Lauritzen, 1996). We propose a two component graphical chain model in which a set of categorical (or discrete) random variables is modeled as a response to a set of categorical

and continuous covariates. This proposed model, called the random effects discrete regression (REDR) model, offers a number of advantages to analyzing compositional data. For the univariate composition considered in the Billheimer-Guttorm model, it is challenging to extend the results to more than three levels of response in a given category. We offer a graphical model framework which overcomes this difficulty. We use the REDR model to construct a Bayesian hierarchical model for inference. By using the REDR model and some new results on graphical models, we can examine the complex conditional relationships between the covariates and multiple response variables as a whole system through inference from a graphical chain model.

To illustrate analysis with the REDR model, we examine the relationship between two characteristics of fish species and a number of environmental covariates. These relationships are of interest to the U.S. Environmental Protection Agency for stream monitoring. Our focus here is on an application related to ecological modeling, but compositional data are prevalent in many other areas including geology, biology, economics and chemistry.

The remainder of the paper is organized as follows. In Section 2 we describe the problem of interest. Section 3 introduces the model and a Bayesian approach to parameter estimation. We discuss the results of the fish species abundance analysis in Section 4.

2 Statistical analysis of species composition

Relating the presence or absence of organisms and their biological characteristics to local environmental conditions is a challenging problem for ecologists (Legendre et al., 1997). Distributions of specific species are usually of limited interest as they are often biogeographically constrained, thereby restricting ecological inference to one geographic range. Recoding taxonomic species in terms of their membership in a number of functional trait categories allows for modeling of organism-environment relationships that can transcend biogeographic boundaries and may even apply across ecosystem types (Poff and Allan, 1995).

2.1 Previous work

The problem of relating species traits to environmental conditions is of interest to ecologists trying to understand basic natural processes. Beyond this, scientists and policy makers concerned with monitoring natural resources also find such analyses to be useful. The U.S. Environmental Protection Agency (EPA) monitors the chemical, physical, and biological quality of streams in the United

States, in part by counting the number of fish species at a number sample locations and relating this to environmental conditions at the sites. Since fish species vary over the landscape due to a large number of factors, it is useful to monitor the traits of fish instead of the specific species themselves. For example, species differ in their tolerance for high organic pollutant loads that reduce dissolved oxygen in the water; characterizing the relative proportion of species possessing such tolerance across the landscape can provide insight into water pollution. Because all species can be characterized in terms of possessing this trait, the functional approach allows for broad, interregional comparisons.

Various approaches have been used to describe the relationship between species traits and environmental conditions. One approach, canonical correspondence analysis, attempts to ordinate each species along a set of environmental axes (ter Braak, 1985). 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), called “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, i.e., multiple traits cannot be analyzed simultaneously. Secondly, both of the previous approaches measure only marginal associations between the environment and traits in question. Conditional relationships can 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 wants to predict the functional composition of a biological community at a site using remotely sensed environmental measurements, the previous methods provide no means to accomplish this task. The methodology proposed in Section 3 addresses all three of these issues. In the following section we describe the data set of interest in this paper.

2.2 Fish Traits in the MAHA Region

The EPA and other agencies are interested in assessing the ecological condition of streams and identifying any factors that might be associated with any degradation in stream conditions (U.S. Environmental Protection Agency, 2000). During 1993–1996 the EPA, along with the U.S. Fish and Wildlife Service and other contractors, surveyed 309 wadeable streams in the Mid-Atlantic Highlands as part of the Environmental Monitoring and Assessment Program (EMAP). Here, 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. Chemical samples and several physical habitat variables were measured at the sampled sites. Fish were sampled by electro-fishing and each fish species was classified according to several taxonomic and ecological categories. McCormick 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. Sites that were considered to be incapable of maintaining a fish population and sites for which environmental covariates were not recorded were eliminated from the analysis (McCormick et al., 2001). After removing these sites, 91 sites remained with between 1 and 23 fish species observed at each site.

In order to perform the functional trait analysis of species occurrence, each observed species is categorized according to two categorical variables, habit and tolerance. The habit variable describes where species live in two levels: benthic species live on or near the stream bottom and 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 caused by heavy silt loads or low dissolved oxygen. Species classified as intolerant are sensitive to human induced stream degradation, whereas tolerant species are relatively insensitive. Species between the two extreme tolerance classifications are classified as having intermediate tolerance. The proportion of benthic and intolerant species are important metrics used by the EPA to measure stream degradation (McCormick et al., 2001). Table 1 shows the cell counts for two sites. These species distributions at each site are the response in our models. Note that there is one contingency table for every sampled site. The distribution of cell counts for all sites is shown in Figure 1.

In our analysis, we are interested in the associations between the species distribution at each site and covariates measuring local environmental conditions and landscape setting (Table 2). The

Table 1: Cross tabulation of the response for two sites: the number of species observed in each combination of habit and tolerance category.

		Tolerance		
	Habit	Intolerant	Intermediate	Tolerant
Site 1	Column	0	4	3
	Benthic	0	1	5
Site 53	Column	1	3	3
	Benthic	3	3	3

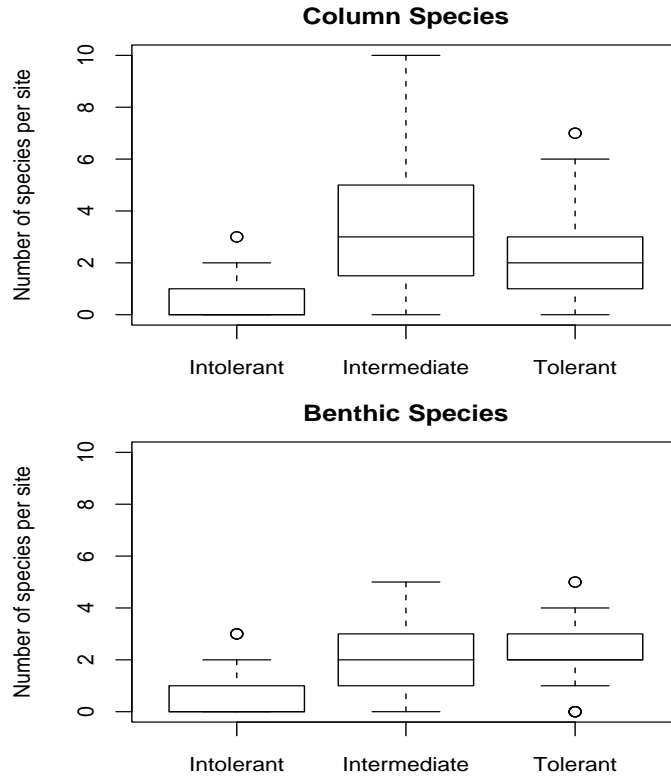


Figure 1: Distribution of fish species abundance for all 91 sites for the three different tolerance levels. Abundances are separated by habit type, column species and benthic species.

local covariates include stream site sulfate concentration ($\log \mu\text{eq/L}$), which measures atmospheric acid deposition or acid mine drainage; chloride concentration ($\log \mu\text{eq/L}$), which is associated with human activity; and water turbidity ($\log \text{NTU}$), a measure of water cloudiness caused by very fine suspended sediments. The landscape covariates include watershed area ($\log \text{km}^2$), elevation (kilometers), a measure of stream area, and mean annual watershed precipitation (meters), a measure of stream volume.

Table 2: Summary of environmental covariates for 1994 MAHA streams.

Covariate	Mean	St. Dev.	Min.	Max.
Area (km ²)	3.04	1.19	0.78	6.39
Chloride ($\mu\text{eq/L}$)	4.60	1.14	2.64	6.98
Elevation (kilometers)	0.07	0.03	0.0015	0.14
Precipitation (meters)	1.08	0.10	0.85	1.33
Sulfate ($\mu\text{eq/L}$)	5.42	0.88	3.78	8.42
Turbidity (NTU)	1.01	0.77	-0.92	3.40

3 Model Formulation

We consider models for multivariate compositional data where sampled individuals are classified according to two or more different categorical variables. The focus is on 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 (cell).

We also examine the independence structure of the discrete variables used for cross-classification. For example, for a two-way cross-classification of individuals according to the discrete variables I and J , we examine whether the classification of a random individual to category i of variable I is independent of the event that the individual is classified to category j of variable 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 $p_{ij} = \Pr\{\text{individual in cell } (i, j)\}$ and $p_i = \Pr\{\text{individual in category } i \text{ of } I\}$. 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.

We use the term “site” to index each multivariate discrete compositional observation. If only the fully saturated model (complete dependence among all classification variables) is of interest, then a simple model that combines all of the cells into a single composition can be used. This is the approach used by Dominici (2000) to combine several contingency tables, some with missing dimensions. However, if one is interested in introducing some independence structure to the cell probabilities, then an extension of the simpler model must be considered and is proposed here.

The models proposed below are motivated by the class of models known as graphical models. In standard graphical modeling every sampled individual generates a multivariate observation of categorical and continuous variables. In our case, however, there is only one observation of the

Table 3: The notation of Section 3.1 for the fish species occurrence example described in Section 2.2. The response \mathbf{y}_Φ is the response for a single species at site 53. This species has a benthic habit and is categorized as a tolerant species. Site specific indices are added in Section 3.2.

Notation	Example
Φ	Habit (column or benthic) and tolerance (intolerant, intermediate and tolerant)
D	6=2 habit levels \times 3 tolerance levels
\mathbf{y}_Φ	(2,3) for 2nd level of habit variable and 3rd level of tolerance variable
\mathbf{x}	The values of the 6 continuous environment covariates at site 53

covariate vector for all of the individuals observed at a particular site. To achieve our goal of developing a class of graphical models to address this problem, the model developed here requires a more sophisticated model structure than either the standard graphical model or linear model structures.

3.1 Developing a Single Site Model

Let Φ denote the set of categorical response variables, of which there are D possible cross-classifications on the product space of the response variable levels. For each site, we are interested in the dependence relationships *among* the covariates measured at the site and the relationship *between* the covariates and the event that a randomly selected individual is cross-classified into one of the D cells \mathbf{y}_Φ . Let $\mathbf{Y}_\Phi = \{Y_\phi : \phi \in \Phi\}$ denote the response vector for a single individual, which takes one of the D possible vectors, say \mathbf{y}_Φ , as a realization. In addition, let $\mathbf{X} = (X_1, \dots, X_p)$ 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 the analysis of fish species occurrence described in Section 2.2, we are interested in the event that a randomly selected fish species at a particular site belongs to a certain cross-classification of life-history traits. Table 3 gives such an example for a single “individual” (a single species of fish) observed at site 53. In Section 4.1 we further develop the model described below for the analysis of fish species occurrence.

Typically many individuals are sampled at any given site. We now define the likelihood model for sampling N individuals at a site and subsequently develop the specific terms of the likelihood, including the probability that a single individual will be in a single cell at a given site. We first condition on the realization of a site, which amounts to conditioning on the covariates. Sampling

N individuals at a site provides N realizations of the variable \mathbf{Y}_Φ . 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}_Φ . 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 \mathbf{Y}_Φ . Using a multinomial model for the count vector \mathbf{c} , the joint distribution for the counts and covariates, for a fixed sample size N , is

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

where $f(\mathbf{y}_\Phi|\mathbf{x})f_{CG}(\mathbf{x})$ represent the joint density of an individual classified to cell \mathbf{y}_Φ and covariates observed at the randomly selected site where the individual is observed. These terms are defined below.

The model in (1) includes terms for for the probability that a single individual will be in a particular cell at a single site, $f(\mathbf{y}_\Phi|\mathbf{x})f_{CG}(\mathbf{x})$. In the analysis of fish species occurrence described in Section 2.2, this corresponds to the probability that a randomly selected fish species at a particular site (which has a particular set of site-level covariates) belongs to a certain cross-classification of life-history traits. Using a log-linear model, we now specify a joint density for $(\mathbf{Y}_\Phi, \mathbf{X})$, which we call the Discrete Regression (DR) model,

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

and

$$f_{CG}(\mathbf{x}) = \exp \left[\sum_{d \subseteq \Delta} \left\{ \lambda_d(\mathbf{x}_\Delta) + \boldsymbol{\eta}_d(\mathbf{x}_\Delta)' \mathbf{x}_\Gamma - \frac{1}{2} \mathbf{x}'_\Gamma \boldsymbol{\Psi}_d(\mathbf{x}_\Delta) \mathbf{x}_\Gamma \right\} \right]. \quad (3)$$

In the response model (2), $\alpha_\Phi(\mathbf{x})$ is a normalizing constant with respect to $\mathbf{Y}_\Phi|\mathbf{x}$. The regression coefficients $\beta_{fcd}(\mathbf{y}_\Phi, \mathbf{x}_\Delta)$ and $\omega_{\gamma dm}(\mathbf{y}_\Phi, \mathbf{x}_\Delta)$ correspond to interaction terms which depend on \mathbf{y}_Φ and \mathbf{x}_Δ only through the variables associated with the sets $f \subseteq \Phi$ and $d \subseteq \Delta$, respectively. The summation from $m=2$ to M denotes the number of higher-order terms desired in the model for covariates x_γ . For example, if $M = 3$ the summation is over x_γ^2 and x_γ^3 . The first order terms enter into the model in the β terms with c equal to a singleton set. In the response portion of the model (2), 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}_Φ and will cancel with the

normalizing term $\alpha_\Phi(\mathbf{x})$. The conditional Gaussian (CG) density (3) is a joint distribution for continuous and discrete variables. As with the interaction terms in the response model, $\lambda_d(\mathbf{x}_\Delta)$, $\eta_d(\mathbf{x}_\Delta)$, and $\Psi_d(\mathbf{x}_\Delta)$ depend on \mathbf{x}_Δ only through the subset of variables associated with the set $d \subseteq \Delta$. The CG distribution is constructed by assuming that \mathbf{X}_Δ follows a log-linear model and $\mathbf{X}_\Gamma|\mathbf{x}_\Delta$ follows a multivariate normal distribution.

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}_Φ^* , and a reference cell for the categorical covariates, say \mathbf{x}_Δ^* . Without loss of generality, henceforth, we assume that \mathbf{y}_Φ^* and \mathbf{x}_Δ^* 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 (2) and (3) 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}_Φ^* and \mathbf{x}_Δ^* . 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 $\beta_{\phi\gamma\delta}(\mathbf{y}_\Phi, \mathbf{x}_\Delta)$ 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 .

Here we will consider only the homogeneous CG distribution, where $\Psi_d(\mathbf{x}_\Delta) = \mathbf{0}$ 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 defined by the product space of \mathbf{X}_Δ . The model can be extended to be non-homogeneous, if desired.

3.2 Random Effects Discrete Regression

In Section 3.1 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 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 usually of primary interest. Therefore, we propose a global random effects model for all sites that allows site-to-site flexibility in some of the non-zero parameter values. In order to add this flexibility as well as to model the randomness in site selection, we introduce a random error term to the response model (2).

The addition of a random effect to the response model (2) produces a full model for \mathbf{Y}_Φ , \mathbf{X} , and the random effects $\boldsymbol{\epsilon}$ of the form

$$f(\mathbf{y}_\Phi, \mathbf{x}, \boldsymbol{\epsilon}) = f_{RE}(\mathbf{y}_\Phi | \mathbf{x}, \boldsymbol{\epsilon}) f_{CG}(\mathbf{x}) f(\boldsymbol{\epsilon}). \quad (4)$$

Since there is only one observation of the explanatory variables at each site we adopt the model for the covariates, $f_{CG}(\mathbf{x})$, that is given in (3). 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,

$$f_{RE}(\mathbf{y}_\Phi | \mathbf{x}, \boldsymbol{\epsilon}) = \exp \left\{ \alpha_\Phi(\mathbf{x}) + \sum_{f \subseteq \Phi} \sum_{c \subseteq \Gamma} \sum_{d \subseteq \Delta} \beta_{fcd}(\mathbf{y}_\Phi, \mathbf{x}_\Delta) \prod_{\gamma \in c} x_\gamma \right. \\ \left. + \sum_{f \subseteq \Phi} \sum_{\gamma \in \Gamma} \sum_{d \subseteq \Delta} \sum_{m=2}^M \omega_{f\gamma dm}(\mathbf{y}_\Phi, \mathbf{x}_\Delta) x_\gamma^m + \sum_{f \subseteq \Phi} \epsilon_f(\mathbf{y}_\Phi) \right\}, \quad (5)$$

where $\epsilon_f(\mathbf{y}_\Phi) = 0$, if $y_\phi = 1$ for any $\phi \subseteq f$, to ensure identifiability. In order to allow modeling of a given independence structure for the multivariate response, we also introduce one other constraint on the random effects. For any set $f \subseteq \Phi$, if $\beta_{fcd}(\mathbf{y}_\Phi, \mathbf{x}_\Delta)$ and $\omega_{f\gamma dm}(\mathbf{y}_\Phi, \mathbf{x}_\Delta)$ are set to zero for all \mathbf{y}_Φ and \mathbf{x}_Δ then $\epsilon_f(\mathbf{y}_\Phi)$ is defined to be a zero vector. The remaining random interactions $\boldsymbol{\epsilon}_f = \{\epsilon_f(\mathbf{y}_\Phi) : y_\phi \neq 1 \text{ for any } \phi \subseteq f\}$ are given a multivariate distribution with mean $\mathbf{0}$ and covariance matrix (or scale parameter) $\boldsymbol{\Sigma}_f$.

The inclusion of random error terms in (5) has three benefits. First, the model can adjust for site-to-site variability. Secondly, the model will account for some level of over-dispersion to cell counts. Finally, every realization of the random effects provides cell probabilities that maintain the desired independence relationships among the response variable. Johnson and Hoeting (2003) provide proof of this fact using a Möbius inversion calculation similar to Lauritzen (1996, pg. 174). The site-based REDR model provides an improvement over the approach of Aitchison (1986) which only examines the average composition independence structure across sites.

In the REDR model description we did not specify the error distribution, as different situations may necessitate different error structures. If it is reasonable to assume that the error structure is symmetric with few outliers, then a Multivariate-Normal (MVN) distribution may be reasonable. In this case, the cell compositions will have a logistic-normal 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 or if there is a high level of over-dispersion in the

cell counts. For the remaining discussion of the REDR model we will assume a MVN distribution for the random effects (i.e., $f(\boldsymbol{\epsilon}_f) = f_N(\boldsymbol{\epsilon}_f; \mathbf{0}, \boldsymbol{\Sigma}_f)$).

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 (1). The conditional 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

$$f_M(\mathbf{c}|\mathbf{x}, \boldsymbol{\epsilon}) = \frac{N!}{\prod_{\mathbf{y}_\Phi} c(\mathbf{y}_\Phi)!} \prod_{\mathbf{y}_\Phi} f_{RE}(\mathbf{y}_\Phi|\mathbf{x}, \boldsymbol{\epsilon})^{c(\mathbf{y}_\Phi)}, \quad (6)$$

where $f_{RE}(\mathbf{y}_\Phi|\mathbf{x}, \boldsymbol{\epsilon})$ is given by (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

$$f(\mathbf{x}_1, \dots, \mathbf{x}_S) = \prod_{i=1}^S f_{CG}(\mathbf{x}_i|\boldsymbol{\lambda}, \boldsymbol{\eta}, \boldsymbol{\Psi}_\theta), \quad (7)$$

where \mathbf{x}_i denotes the set of observed covariates for sites $i = 1, \dots, S$ and $\boldsymbol{\lambda}$ and $\boldsymbol{\eta}$ represent the collected parameter sets $\{\lambda_d(\mathbf{x}_\Delta) : d \subseteq \Delta\}$ and $\{\eta_d(\mathbf{x}_\Delta) : d \subseteq \Delta\}$.

We now re-parameterize the homogeneous CG density in (7) 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,

$$\begin{aligned} f_{CG}(\mathbf{x}) &= f(\mathbf{x}_\Delta)f(\mathbf{x}_\Gamma|\mathbf{x}_\Delta) \\ &= \exp \left\{ \sum_{d \subseteq \Delta} \lambda_d(\mathbf{x}_\Delta) \right\} \times \frac{1}{\sqrt{2\pi}} |\boldsymbol{\Psi}_\theta|^{1/2} \\ &\quad \times \exp \left\{ \frac{1}{2} \left(\mathbf{x}_\Gamma - \sum_{d \subseteq \Delta} \boldsymbol{\tau}_d(\mathbf{x}_\Delta) \right)' \boldsymbol{\Psi}_\theta \left(\mathbf{x}_\Gamma - \sum_{d \subseteq \Delta} \boldsymbol{\tau}_d(\mathbf{x}_\Delta) \right) \right\}, \end{aligned} \quad (8)$$

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

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 (8) we can

write the joint density of the covariates over all sites (7) as

$$\begin{aligned}
f(\mathbf{x}_1, \dots, \mathbf{x}_S) &= \left\{ \prod_{i=1}^S f(\mathbf{x}_{\Delta i}) \right\} \times \left\{ \prod_{i=1}^S f_N(\mathbf{x}_{\Gamma i} | \mathbf{x}_{\Delta i}) \right\} \\
&\propto f_M(\mathbf{c}_{\Delta} | \boldsymbol{\lambda}) \prod_{i=1}^S f_N \left(\mathbf{x}_{\Gamma i}; \sum_{d \subseteq \Delta} \tau_d(\mathbf{x}_{\Delta i}), \boldsymbol{\Psi}_{\emptyset} \right),
\end{aligned} \tag{9}$$

where the explanatory observation at the i th 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

$$f_M(\mathbf{c}_{\Delta} | \boldsymbol{\lambda}) = \frac{S!}{\prod_{\mathbf{x}_{\Delta}} c(\mathbf{x}_{\Delta})!} \prod_{\mathbf{x}_{\Delta}} \exp \left\{ \sum_{d \subseteq \Delta} \lambda_d(\mathbf{x}_{\Delta}) \right\}^{c(\mathbf{x}_{\Delta})}. \tag{10}$$

The full likelihood for parameter estimation in the REDR model (4) is obtained by combining the likelihood for response variable cell counts at each site (6), the random effects density, and the explanatory variable likelihood (7).

$$\begin{aligned}
f(\{\mathbf{c}_i\}, \{\mathbf{x}_i\}, \{\boldsymbol{\epsilon}_i\}) &= \prod_{i=1}^S f_M(\mathbf{c}_i | \mathbf{x}_i, \boldsymbol{\epsilon}_i) f_{CG}(\mathbf{x}_i) f_N(\boldsymbol{\epsilon}_i) \\
&\propto \prod_{i=1}^S f_M(\mathbf{c}_i | \mathbf{x}_i) \times f_M(\mathbf{c}_{\Delta} | \boldsymbol{\lambda}) \times \prod_{i=1}^S f_N \left(\mathbf{x}_{\Gamma i}; \sum_{d \subseteq \Delta} \tau_d(\mathbf{x}_{\Delta i}), \boldsymbol{\Psi}_{\emptyset} \right) \\
&\quad \times \prod_{i=1}^S \prod_{f \subseteq \Phi} f_N(\boldsymbol{\epsilon}_{f,i}; \mathbf{0}, \boldsymbol{\Sigma}_f),
\end{aligned} \tag{11}$$

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, and $\boldsymbol{\epsilon}_i$ represents the collection of random effects vectors $\{\boldsymbol{\epsilon}_{f,i} : f \subseteq \Phi\}$ for the i th site.

3.3 Multivariate Compositions and Graphical Models

The models proposed in the previous sections were motivated by a class of models known as *graphical models* or Markov random fields. A graphical model, or more loosely a conditional independence 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. A mathematical graph involves a set of vertices, one for each element of the vector, and a set of edges connecting some of the vertices. Edges can either be *undirected* or *directed*. An undirected edge between two vertices indicates bi-directional dependence between the elements while a directed edge signifies a causal or influential effect from one element to another. If two vertices are not

connected, one can infer that the two variables which they represent are conditionally independent given their neighbors (those vertices which are connected to the pair in question). Lauritzen (1996) provides a thorough overview of graphical modeling.

The REDR model (5) and the corresponding CG model (8) for the covariates together define a chain graph model. The interaction parameters correspond to edges between response and covariate vertices. Edges between the covariates and response vertices are directed, whereas, within covariate and response vertices, the edges are undirected. A given graph represents the conditional independencies for a specific model. Edges between any two vertices are absent if and only if all interaction parameters with subscripts containing the two variable set are zero for all values of the covariates and response. Johnson and Hoeting (2003) provide a rigorous description of the graphical model properties of the REDR model.

There is one difference between the REDR model and standard graphical models. The joint models in (1) for counts and covariates are different than the standard sampling scheme for a 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.

3.4 Parameter Inference

In order to make inference about the parameters in the REDR model (4), 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 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 discussed in Johnson and Hoeting (2003), these random effects can be marginalized over in some cases without affecting the conditional independencies 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. When the goal is to predict compositions at sites for which only the explanatory variables are observed, estimates of the random effects for that site are also necessary. Modern Bayesian computational techniques can handle all of these goals with little modification.

Full development of the Bayesian approach to parameter estimation is provided in the Appendix.

4 Graphical Analysis of Fish Species Occurrence

We adopt the model described above to analyze the fish trait data described in Section 2.2. 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 proportions of benthic and intolerant species are important metrics used by the EPA to measure stream degradation (McCormick et al., 2001). We are interested in inference concerning species occurrence, or the number of species observed in each cell.

4.1 Model Specification

We consider a main-effects-only model for the analysis of the multivariate composition of habit (H) and tolerance (T) species occurrence including only the first-order linear interaction terms. We consider the following multinomial model for the cell counts,

$$f_M(\mathbf{c}_i | \mathbf{x}_i, \boldsymbol{\epsilon}_i) = \frac{N!}{\prod_{\mathbf{y}_\Phi} c(\mathbf{y}_\Phi)_i!} \prod_{\mathbf{y}_\Phi} f_{RE}(\mathbf{y}_\Phi | \mathbf{x}_i, \boldsymbol{\epsilon}_i)^{c(\mathbf{y}_\Phi)_i}, \quad (12)$$

where

$$f_{RE}(\mathbf{y}_\Phi) = \exp \left\{ \alpha_\Phi(\mathbf{x}_i) + \sum_{f \subseteq \Phi} \sum_{\gamma=0}^6 \beta_{f\gamma}(\mathbf{y}_\Phi) (x_{\gamma,i} - \bar{x}_\gamma) + \epsilon_{f,i}(\mathbf{y}_\Phi) \right\}. \quad (13)$$

Each of the environmental covariates was centered by subtracting its mean \bar{x}_γ 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}_Φ^* to be column species with intermediate tolerance level, so $\beta_{f\gamma}(\mathbf{y}_\Phi) \equiv \epsilon_{f,i}(\mathbf{y}_\Phi) \equiv 0$ for $\mathbf{y}_\Phi = (1, 2)$ and $f \subseteq \{H, T\}$ in equation (13).

We consider three different models. In the *independent* model, the habit variable is independent of the tolerance variable, so $\beta_{f\gamma}(\mathbf{y}_\Phi)$ 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}_Φ . In the *dependent* model with uncorrelated errors all interactions between cells are estimated. We assume the random effects vectors $\boldsymbol{\epsilon}_f$ are independently distributed from one another for all sites. The *dependent correlated errors* model is identical to the previous model except that the random effects vectors are correlated within each site. This is equivalent to applying a single composition model to all six cells.

Since all of the environmental covariates are continuous, the homogeneous CG model reduces to a MVN distribution and we assumed $f_{CG}(\mathbf{x}_{\gamma,i} - \bar{\mathbf{x}}_{\gamma,i}) = MVN(\mathbf{x}_{ip} - \bar{\mathbf{x}}_p; \mathbf{0}, \Psi_\theta)$. The centering here allows the elimination of the nuisance parameter $\boldsymbol{\tau}_\theta$ in (8), which is irrelevant for determining conditional independencies. Note, that in this case the posterior distribution of Ψ_θ is a Wishart distribution. However, since we are interested in the off-diagonal elements of Ψ_θ , using an MCMC sampling technique allows straightforward inference of these elements.

4.2 Model Estimation and Performance

MCMC procedures were performed with the hierarchically centered version described in the Appendix as well as with the model as originally parameterized in (13). The hierarchically centered version reached satisfactory convergence with substantially fewer MCMC iterations than (13).

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 parameters would have an acceptance rate of around 30%. The first 4000 iterations were discarded, after which the sampler was run for 20,000 iterations as a burn-in period. Finally an additional 800,000 iterations was used for model inference. Standard diagnostics suggested that the Markov chains had converged to their corresponding posterior distributions (Givens and Hoeting, 2005). Every twentieth iteration was saved for parameter inferences in order to reduce storage constraints.

The Bayesian posterior predictive p -value method of Gelman et al. (1996) was used to assess model fit for each of the three models. With this method a goodness-of-fit statistic $T(y)$, which can be a function of the observed data y and model parameters, is used to compute the Bayesian predictive p -value $P_b = Pr\{T(y^{rep}) > T(y) \mid y\}$. The data y^{rep} represent a hypothesized replicate data set that could have resulted from the model and 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 by generating a replicate data set at each iteration in the Markov chain. The goodness-of-fit statistic used for this analysis was the Freeman-Tukey statistic (Freeman and Tukey, 1950)

$$T(\mathbf{c}_1, \dots, \mathbf{c}_S) = \sum_{i=1}^S \sum_{\mathbf{y}_\Phi} \left(\sqrt{c(\mathbf{y}_\Phi)_i} - \sqrt{N_i f_{RE}(\mathbf{y}_\Phi | \mathbf{x}_i, \boldsymbol{\epsilon}_i)} \right)^2, \quad (14)$$

where $\mathbf{c}(\mathbf{y}_\Phi)_i$ is the number of species belonging to cell \mathbf{y}_Φ 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 (13) and represents the cell composition.

This naive method of approximating posterior predictive p -value can be conservative when rejecting the null hypothesis of model fit in that the MCMC approximated p -value can be too large when used without calibration (Robbins et al., 2000). Draper and Krnjajić (2005) propose a calibration method which could be included in a future analysis. The method is somewhat computationally intensive so we will omit it here and use the p -value obtained as an index of apparent model fit.

We used the Deviance Information Criterion (DIC) for model selection (Spiegelhalter et al., 2003). DIC is composed of two competing elements, a measure of goodness of fit and an measure of the number of effective parameters. For example, random effect parameters may contribute less than one parameter to the number of effective parameters. The model that minimizes the DIC criterion is optimal under this criterion.

The measure of effective parameter dimension used in DIC can be sensitive to the shape of the marginal posterior distributions of the likelihood components (David Draper, personal communication). If the marginal distributions are far from Gaussian, the calculation can be very unstable. We examined the posterior distributions of the parameters and they all appeared to be close to Gaussian in shape. While a full examination would also require inspection of the random effect distributions as well, we omitted this for convenience. After fitting the models, we examined the estimated effective number of parameters for each model. The results for all models considered were in the expected order and appeared reasonable in size, so, the calculated DIC values were judged to be sufficient for our purposes.

4.3 Results

Table 4 lists the considered models and their DIC values. The best model as measured by DIC is the independent response model followed by the uncorrelated errors model. There is a difference of 10.1 in DIC score between the independent response model and the nearest dependent response model suggesting a sizable improvement in model parsimony by selecting the independent response model. The Bayesian p -value, P_b , was greater than 0.9 for all three of the considered models. This suggests there is little evidence of lack-of-fit for any of the models. 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 5. The HPD intervals in Table 5 indicate that chloride concentration and sulfate concentration are related to the pollution tolerance response and that

Table 4: DIC and model complexity for multivariate fish species occurrence 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	1111.1	–	66.1
Dependent (Uncorrelated errors)	1117.8	6.7	106.1
Dependent (Correlated errors)	1166.8	55.7	162.5

Table 5: 95% HPD intervals for covariate interaction parameters in the independent response model for the analysis of fish species occurrence.

Covariate	Habit		Tolerance		
	Column*	Benthic	Intolerant	Intermediate*	Tolerant
Precipitation	–	(–2.17, 0.71)	(–1.64, 3.53)	–	(–2.68, 6.73)
Elevation	–	(0.17, 1.19)	(–0.73, 1.24)	–	(–0.97, 0.22)
Turbidity	–	(–0.25, 0.15)	(–0.37, 0.45)	–	(–0.19, 0.25)
Sulfate	–	(–0.21, 0.16)	(0.06, 0.71)	–	(–0.07, 0.35)
Chloride	–	(–0.17, 0.17)	(–0.77, –0.08)	–	(–0.07, 0.35)
Area	–	(–0.25, 0.02)	(–0.11, 0.42)	–	(–0.29, 0.03)

*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.

elevation is related to 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 related to either response. The HPD intervals for the off-diagonal elements of Ψ_θ are given in Table 6. Again, by examining Table 6 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 2 illustrates the chain graph for the model that is suggested by the DIC criterion. To construct this figure, we included edges between vertices only when all of the 95% HPD intervals for the parameters that correspond to the two vertices did not contain zero. For example, in Table 6, all of the intervals for precipitation contain 0, so this vertex has no edges to other vertices. The DIC criterion suggested that the independent response model is more parsimonious than a dependent response model. As noted in Section 3.2, the independent response model is a preservative model in the sense that relationships are preserved after marginalizing over the random effects (Johnson

Table 6: HPD intervals for the elements of the inverse covariance matrix Ψ_θ for the MAHA environmental variables. The intervals are presented on a correlation matrix scale.

	Elevation	Turbidity	Chloride	Sulfate	Area
Precipitation	(-0.293, 0.099)	(-0.152, 0.248)	(-0.122, 0.273)	(-0.131, 0.263)	(-0.149, 0.247)
Elevation		(-0.052, 0.336)	(0.391, 0.672)	(-0.381, -0.001)	(-0.528, -0.181)
Turbidity			(-0.392, -0.016)	(0.089, 0.456)	(-0.109, 0.284)
Chloride				(-0.623, -0.318)	(-0.485, -0.128)
Sulfate					(-0.028, 0.359)

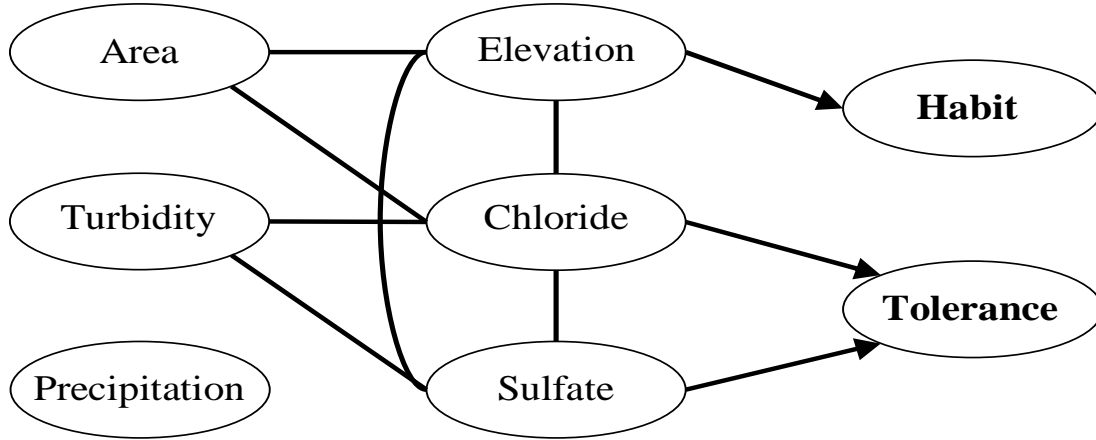


Figure 2: Data suggested chain graph for the multivariate composition of habit and tolerance. The arrows show that the covariates influence the responses, tolerance and habit. The undirected edges between covariates suggest that at least one of the HPD intervals for the covariance estimates between these covariates does not include 0. The lack of edge between any of the covariates or between the response and the covariates indicates that the corresponding intervals in Tables 5 and 6 include 0.

and Hoeting, 2003). Therefore, Figure 2 shows the subgraph for the response and covariates only.

The findings can be interpreted in light of basic understanding of stream ecology. Elevation in this dataset is a surrogate for headwater streams that are shallow and are occupied largely by benthic species. Intolerant species occurrence declines with elevated chloride levels, a surrogate for human disturbance. The positive association between high sulfate levels and intolerant species is unexpected and at odds with a previous study that used a univariate approach (McCormick et al. 2001). We note, however, that the 95% HPD interval for this association almost includes 0, and that there is a strong negative correlation of sulfate with chloride which suggests the possibility that these covariates are drawing from the same latent disturbance process to which sulfate concentration

is negatively related. And, it possible that disturbance process is influencing species occurrence. Similarly the addition of latent process to model tolerance could enhance the model. The ordering in the tolerance variable, intolerant to moderate to tolerant species was ignored in this analysis; inclusion of a latent process measuring tolerance on the continuous scale but observed on a discrete 3-point scale could but used to better address this issue.

The model developed here allows for predictions at locations where only the covariates are observed. These results were not included in this analysis due to the relatively small sample size. An example of predictions for a similar model is given in Johnson (2003).

A Appendix: Parameter Inference

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

$$\begin{aligned}
f_{\text{post}}(\boldsymbol{\beta}, \boldsymbol{\omega}, \boldsymbol{\lambda}, \boldsymbol{\tau}, \boldsymbol{\Psi}_\theta, \boldsymbol{\Sigma}, \{\boldsymbol{\epsilon}\} \mid \{\mathbf{c}\}, \{\mathbf{x}\}) &\propto \prod_{i=1}^S f_M(\mathbf{c}_i \mid \boldsymbol{\beta}, \boldsymbol{\omega}, \mathbf{x}_i, \boldsymbol{\epsilon}_i) \\
&\times f_M(\mathbf{c}_\Delta \mid S, \boldsymbol{\lambda}) \\
&\times \prod_{i=1}^S f_N(\mathbf{x}_{\Gamma_i} \mid \mathbf{x}_{\Delta_i}, \boldsymbol{\tau}, \boldsymbol{\Psi}_\theta) \\
&\times \prod_{i=1}^S \prod_{f \subseteq \Phi} f_N(\boldsymbol{\epsilon}_{f,i} \mid \boldsymbol{\Sigma}_f) \\
&\times \pi(\boldsymbol{\beta}, \boldsymbol{\omega}, \boldsymbol{\lambda}, \boldsymbol{\tau}, \boldsymbol{\Psi}_\theta, \boldsymbol{\Sigma}),
\end{aligned} \tag{15}$$

where $\{\boldsymbol{\epsilon}\} = \{\boldsymbol{\epsilon}_i : i = 1, \dots, S\}$, $\{\mathbf{c}\} = \{\mathbf{c}_i : i = 1, \dots, S\}$, and $\{\mathbf{x}\} = \{\mathbf{x}_i : i = 1, \dots, S\}$.

The posterior distribution (15) 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 a Gibbs sampling approach (e.g., Givens and Hoeting (2005)).

Assuming the CG parameters are independent of the remaining parameters simplifies the analysis with $\pi(\boldsymbol{\beta}, \boldsymbol{\omega}, \boldsymbol{\lambda}, \boldsymbol{\tau}, \boldsymbol{\Psi}_\theta, \boldsymbol{\Sigma}) = \pi(\boldsymbol{\beta}, \boldsymbol{\omega}, \boldsymbol{\Sigma}) \times \pi(\boldsymbol{\lambda}, \boldsymbol{\tau}, \boldsymbol{\Psi}_\theta)$ so the posterior distribution is given by

$$f_{\text{post}}(\boldsymbol{\beta}, \boldsymbol{\omega}, \boldsymbol{\lambda}, \boldsymbol{\tau}, \boldsymbol{\Psi}_\theta, \boldsymbol{\Sigma}, \{\boldsymbol{\epsilon}\} \mid \{\mathbf{c}\}, \{\mathbf{x}\}) \propto f_{\text{post}}(\boldsymbol{\beta}, \boldsymbol{\omega}, \boldsymbol{\Sigma}, \{\boldsymbol{\epsilon}\} \mid \{\mathbf{c}\}, \{\mathbf{x}\}) \times f_{\text{post}}(\boldsymbol{\lambda}, \boldsymbol{\tau}, \boldsymbol{\Psi}_\theta \mid \{\mathbf{x}\}).$$

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. This type of sequential estimation is often used to estimate chain model parameters (Whittaker, 1990, pg. 310).

A.1 Hierarchical Centering Parameterization

Here we present a modification to the response model parameterizations presented in (5). When using a Gibbs MCMC procedure, the Markov chains for regression coefficient parameters in random effects generalized linear models, such as β and ω , 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 the proposed 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 β and ω .

In order to describe the hierarchical centering parameterization, first recall the response portion of the REDR distribution (5). We introduce a shortened notation for the fixed effects portion of the response model for each $f \subseteq \Phi$, with

$$\mu_f(\mathbf{y}_\Phi) = \sum_{c \subseteq \Gamma} \sum_{d \subseteq \Delta} \beta_{fcd}(\mathbf{y}_\Phi, \mathbf{x}_\Delta) \prod_{\gamma \in c} x_\gamma + \sum_{\gamma \in \Gamma} \sum_{d \subseteq \Delta} \sum_{m=2}^M \omega_{f\gamma dm}(\mathbf{y}_\Phi, \mathbf{x}_\Delta) x_\gamma^m. \quad (16)$$

Then we propose the following hierarchically centered re-parameterization of (5),

$$f_{RE}^{(h)}(\mathbf{y}_\Phi | \varphi) = \exp \left\{ \sum_{f \subseteq \Phi} \varphi_f(\mathbf{y}_\Phi) \right\}, \quad (17)$$

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

$$\varphi_\emptyset = -\log \left[\sum_{\mathbf{y}_\Phi} \exp \left\{ \sum_{f \subseteq \Phi} \varphi_f(\mathbf{y}_\Phi) \right\} \right], \quad f \neq \emptyset. \quad (18)$$

If the assumption is made that $\epsilon_f \sim f_N(\mathbf{0}, \Sigma_f)$ for $f \subseteq \Phi$ that are not set to zero, as portrayed in (11), then $\varphi_f = \{\varphi_f(\mathbf{y}_\Phi) : y_\phi \neq 1 \text{ for any } \phi \subseteq f\} \sim f_N(\boldsymbol{\mu}_f, \Sigma_f)$, where $\boldsymbol{\mu}_f$ is the vector $[\mu_f(\mathbf{y}_\Phi)]$. Here we have simply changed the random effects ϵ_f from a zero mean process to a process, φ_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 mixing of the MCMC procedure. It has been our experience, however, that the re-parameterization often greatly improves mixing for these models.

The general parameterization of the full posterior distribution is given by

$$\begin{aligned} & f_{\text{post}}^{(h)}(\boldsymbol{\beta}, \boldsymbol{\omega}, \boldsymbol{\lambda}, \Psi_\emptyset, \{\varphi_i\} \mid \{\mathbf{c}_i\}, \{\mathbf{x}_i\}) \\ & \propto \left[\prod_{i=1}^S f_M^{(h)}(\mathbf{c}_i | \varphi_i) \left\{ \prod_{f \subseteq \Phi} f_N(\varphi_{fi} | \boldsymbol{\beta}_f, \boldsymbol{\omega}_f, \Sigma_f, \mathbf{x}_i) \right\} \right] \pi(\boldsymbol{\beta}, \boldsymbol{\omega}, \Sigma) \\ & \quad \times \left[\prod_{i=1}^S f_N(\mathbf{x}_{\Gamma i} | \mathbf{x}_{\Delta i}, \boldsymbol{\tau}, \Psi_\emptyset) \right] f_M(\mathbf{c}_\Delta | \boldsymbol{\lambda}) \pi(\boldsymbol{\lambda}, \boldsymbol{\tau}, \Psi). \end{aligned} \quad (19)$$

A.2 Implementing the Gibbs Sampler

In order to implement the Gibbs sampler to draw a sample from (19) 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 $\boldsymbol{\Psi}_\theta$, as can be observed from (19). Therefore we can make inferences about the response model parameters using only the first factor on the left hand side of the proportionality, while inferences about the explanatory portion of the chain graph uses only the second factor.

A.2.1 Response Model Conditional Densities

Before deriving the full conditional densities we introduce some notation. Let \mathbf{E} refer to a matrix that has S rows and r columns including a column of ones, a column corresponding to each of the explanatory variables, and a column for each interaction and powers of the continuous covariates as given in (5). 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. The vector \mathbf{E}_i , $i = 1, \dots, S$ will denote an r -vector formed from the i th 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 $\{\omega_{f\gamma dm}(\mathbf{y}_\Phi, \mathbf{x}_\Delta) : \gamma \in \Gamma, d \subseteq \Delta, m = 1, \dots, M\}$ such that the expected value (16) 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}_{f_s} . 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 i th 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 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 $\{\mathbf{T}_f : f \subseteq \Phi\}$. 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 , so $\pi(\boldsymbol{\beta}, \boldsymbol{\omega}, \boldsymbol{\Sigma}) = \prod_{f \subseteq \Phi} \pi(\boldsymbol{\beta}_f, \boldsymbol{\omega}_f) \pi(\boldsymbol{\Sigma}_f)$.

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 $\boldsymbol{\varphi}_{f_i}$ at each site and the random effect inverse covariance matrix T_f , the interaction coefficients are independent of the cell counts. If we let $\pi(\boldsymbol{\beta}_f, \boldsymbol{\omega}_f) = \pi(\mathbf{B}_{f_s}) = f_N(\mathbf{B}_{f_s}; \boldsymbol{\mu}_{B_{f_s}}, \mathbf{V}_{B_{f_s}}^{-1})$ and $\hat{\mathbf{B}}_f = (\mathbf{E}'\mathbf{E})^{-1}\mathbf{E}'\boldsymbol{\varphi}_f$ (correspondingly $\hat{\mathbf{B}}_{f_s}$ represents the stacked version) then the full conditional distribution of the interaction coefficients

is given as

$$\begin{aligned}
f(\mathbf{B}_{f_s} | \dots) &\propto \exp \left\{ -\frac{1}{2} \sum_{i=1}^S (\boldsymbol{\varphi}_{f,i} - \mathbf{B}'_f \mathbf{E}_i)' \mathbf{T}_f (\boldsymbol{\varphi}_{f,i} - \mathbf{B}'_f \mathbf{E}_i) \right\} \\
&\quad \times \exp \left\{ -\frac{1}{2} (\mathbf{B}_{f_s} - \boldsymbol{\mu}_{B_{f_s}})' \mathbf{V}_{B_{f_s}} (\mathbf{B}_{f_s} - \boldsymbol{\mu}_{B_{f_s}}) \right\} \\
&\propto \exp \left\{ -\frac{1}{2} \text{tr} \left[\mathbf{T}_f (\mathbf{B}_f - \hat{\mathbf{B}}_f)' \mathbf{E}' \mathbf{E} (\mathbf{B}_f - \hat{\mathbf{B}}_f) \right] \right\} \\
&\quad \times \exp \left\{ -\frac{1}{2} (\mathbf{B}_{f_s} - \boldsymbol{\mu}_{B_{f_s}})' \mathbf{V}_{B_{f_s}} (\mathbf{B}_{f_s} - \boldsymbol{\mu}_{B_{f_s}}) \right\} \\
&= \exp \left\{ -\frac{1}{2} (\mathbf{B}_{f_s} - \hat{\mathbf{B}}_{f_s})' (\mathbf{T}_f \otimes \mathbf{E}' \mathbf{E}) (\mathbf{B}_{f_s} - \hat{\mathbf{B}}_{f_s}) \right\} \\
&\quad \times \exp \left\{ -\frac{1}{2} (\mathbf{B}_{f_s} - \boldsymbol{\mu}_{B_{f_s}})' \mathbf{V}_{B_{f_s}} (\mathbf{B}_{f_s} - \boldsymbol{\mu}_{B_{f_s}}) \right\},
\end{aligned} \tag{20}$$

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

$$f(\mathbf{B}_{f_s} | \dots) = f_N(\mathbf{B}_{f_s}; \boldsymbol{\mu}_{f,1}, \mathbf{V}_{f,1}^{-1}), \tag{21}$$

where the mean and covariance are given by

$$\begin{aligned}
\boldsymbol{\mu}_{f,1} &= [(\mathbf{T}_f \otimes \mathbf{E}' \mathbf{E}) + \mathbf{V}_{B_{f_s}}]^{-1} [(\mathbf{T}_f \otimes \mathbf{E}' \mathbf{E}) \hat{\mathbf{B}}_{f_s} + \mathbf{V}_{B_{f_s}} \boldsymbol{\mu}_{B_{f_s}}] \\
&\quad \text{and} \\
\mathbf{V}_{f,1} &= (\mathbf{T}_f \otimes \mathbf{E}' \mathbf{E}) + \mathbf{V}_{B_{f_s}}.
\end{aligned} \tag{22}$$

Therefore in the Gibbs sampler, drawing samples of the interaction coefficients is a relatively simple draw from a multivariate normal distribution.

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

$$\begin{aligned}
\pi(\mathbf{T}_f) &= f_W(\mathbf{T}_f; a_f, \mathbf{K}_f) \\
&\propto |\mathbf{T}_f|^{(a-D_f-1)/2} \exp \left\{ -\frac{1}{2} \text{tr} [\mathbf{K}_f \mathbf{T}_f] \right\}.
\end{aligned} \tag{23}$$

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,

$$\begin{aligned}
f(\mathbf{T}_f | \dots) &\propto |\mathbf{T}_f|^{S/2} \exp \left\{ -\frac{1}{2} \sum_{i=1}^S (\boldsymbol{\varphi}_{f,i} - \mathbf{B}'_f \mathbf{E}_i)' \mathbf{T}_f (\boldsymbol{\varphi}_{f,i} - \mathbf{B}'_f \mathbf{E}_i) \right\} \\
&\quad \times |\mathbf{T}_f|^{(a-D_f-1)/2} \exp \left\{ -\frac{1}{2} \text{tr} [\mathbf{K}_f \mathbf{T}_f] \right\} \\
&= |\mathbf{T}_f|^{(a+S-D_f-1)/2} \\
&\quad \times \exp \left\{ -\frac{1}{2} \text{tr} \left[\mathbf{T}_f \left\{ \mathbf{K}_f + \sum_{i=1}^S (\boldsymbol{\varphi}_{f,i} - \mathbf{B}'_f \mathbf{E}_i)' (\boldsymbol{\varphi}_{f,i} - \mathbf{B}'_f \mathbf{E}_i) \right\} \right] \right\}
\end{aligned} \tag{24}$$

It follows, then, upon examination of (23), the full conditional distribution of \mathbf{T}_f is given by

$$f(\mathbf{T}_f | \dots) = f_W(\mathbf{T}_f; a_{f,1}, \mathbf{K}_{f,1}) \quad (25)$$

where the full conditional parameters are $a_{f,1} = a_f + S$ and

$$\mathbf{K}_{f,1} = \mathbf{K}_f + \sum_{i=1}^S (\boldsymbol{\varphi}_{f,i} - \mathbf{B}'_f \mathbf{E}_i)' (\boldsymbol{\varphi}_{f,i} - \mathbf{B}'_f \mathbf{E}_i). \quad (26)$$

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

The vector of site random effects, $\boldsymbol{\varphi}_{f,i}$ does not have a standard full conditional distribution. The full conditional density is given by

$$f(\boldsymbol{\varphi}_{f,i} | \dots) \propto f_M^{(h)}(\mathbf{c}_i | \boldsymbol{\varphi}_i) f_N(\boldsymbol{\varphi}_{f,i}; \mathbf{B}'_f \mathbf{E}_i, \mathbf{T}_f^{-1}). \quad (27)$$

Since the full conditional density is non-standard, we employ a Metropolis-within-Gibbs step to sample from this full conditional distribution.

A.2.2 Conditional Distributions for the Explanatory Variables Model

To derive the conditional distributions for the parameters in the explanatory variable CG model (8) we first note that the categorical and continuous explanatory variable parameters are functionally independent. Therefore we can perform separate posterior analyses for the discrete and continuous partitions of the explanatory model. The derivation of the required conditional distributions is similar to the derivations given above, so these are not given here. Complete derivations are provided in Johnson (2003).

Acknowledgments

The authors would like to thank David Draper for his suggested improvements to this paper and Scott Urquhart for helpful discussions.

References

- 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.
- Billheimer, D. (1995). *Statistical Analysis of Biological Monitoring Data: State-Space Models for Species Compositions*. PhD thesis, Department of Statistics, University of Washington.
- Billheimer, D. and Guttorp, P. (1997). Natural variability in benthic 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.
- Chen, M. H., Shao, Q. M., and Ibrahim, J. G. (2000). *Monte Carlo Methods in Bayesian Statistics*. Springer-Verlag, 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.
- Draper, D. and Krnjajić, M. (2005). Bayesian model specification. Technical report, Department of Applied Mathematics and Statistics, University of California, Santa Cruz.
- Freeman, M. and Tukey, J. (1950). Transformations related to the angular and square root. *Annals of Mathematical Statistics*, 21:607–611.
- Gelman, A., Meng, X.-L., and Stern, H. (1996). Posterior predictive assessment of model fitness via realized discrepancies. *Statistica Sinica*, 6:733–807.
- Ghosh, J. K., Bhanja, J., Purkayastha, S., Samanta, T., , and Sengupta, S. (2002). A statistical approach to geological mapping. *Mathematical Geology*, 34(5):505–528.
- Givens, G. H. and Hoeting, J. A. (2005). *Computational Statistics*. Wiley, New York.
- Johnson, D. (2003). *Random Effects Graphical Models for Discrete Compositional Data*. PhD thesis, Department of Statistics, Colorado State University.
- Johnson, D. S. and Hoeting, J. A. (2003). Random effects graphical models for multiple site sampling. Technical Report 2003/15, Department of Statistics, Colorado State University.
- Johnson, R. and Wichern, D. (1992). *Applied Multivariate Statistical Analysis*. Prentice-Hall, New Jersey. 642pp.
- Lauritzen, S. L. (1996). *Graphical Models*. Clarendon Press.
- Legendre, P., Galzin, R., and Harmelin-Vivien, M. (1997). Relating behavior to habitat: Solutions to the fourth-corner problem. *Ecology*, 78:547–562.
- McCormick, F., Hughes, 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 American Fisheries Society*, 130:857–877.
- 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 measurement of organs. *Proceedings of the Royal Society*, 60:489–498.
- Poff, N. and Allan, J. (1995). Functional-organization of stream fish assemblages in relation to hydrological variability. *Ecology*, 76:606–627.
- Robbins, J. M., van der Vaart, A., and Ventura, V. (2000). Asymptotic distribution of P values in composite null models. *Journal of the American Statistical Association*, 95:1143–1156.

- 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.
- ter Braak, C. (1985). Correspondence analysis of incidence and abundance data: Properties in terms of a unimodal response model. *Biometrics*, 41:859–873.
- U.S. Environmental Protection Agency (2000). Environmental Monitoring and Assessment Program, Mid-Atlantic Highlands Streams Assessment, EPA-903-R-00-015, US Environmental Protection Agency, National Health and Environmental Effects Research Laboratory, Western Ecology Division, Corvallis, OR.
- Whittaker, J. (1990). *Graphical Models in Applied Multivariate Statistics*. Wiley, Chichester.