Models for the Analysis of Discrete Compositional Data

An Application of Random Effects Graphical Models

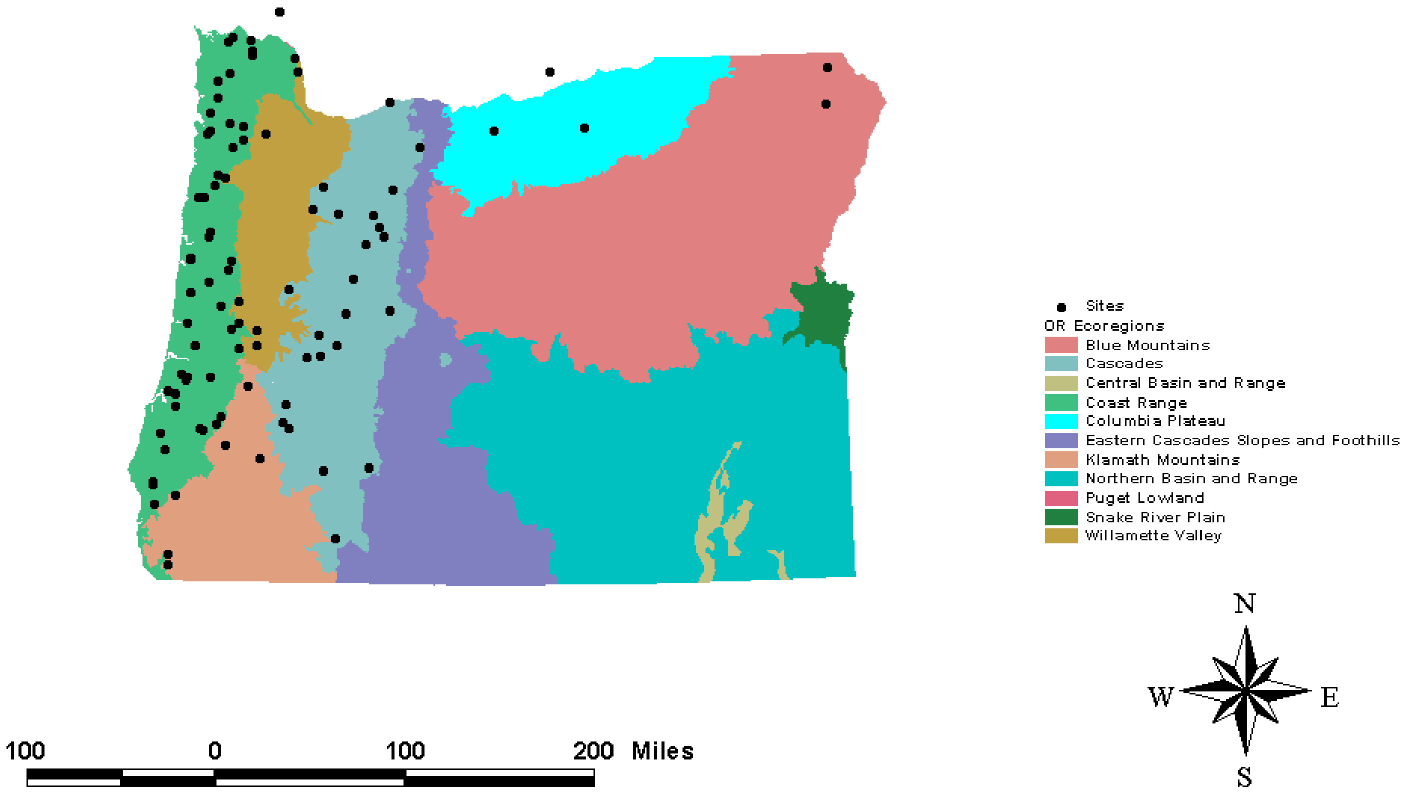
Devin S. Johnson
STARMAP
Department of Statistics
Colorado State University

Developed under the EPA STAR Research Assistance Agreement
CR-829095

Motivating Problem

- Various stream sites in Oregon were visited.
  - Benthic invertebrates collected at each site and cross categorized according to several traits (e.g. feeding type, body shape,…)
  - Environmental variables are also measured at each site (e.g. precipitation, % woody material in substrate, …)
- Total number in each category is not interesting.
- Relative proportions are more informative.
- How can we determine if collected environmental variables affect the relative proportions (which ones)?

1994-1996 REMAP and 1997 EMAP Sites in Oregon



Outline

- Motivation
  - Compositional data
  - Probability models
- Overview of graphical chain models
  - Description
  - Markov properties
- Discrete Response models
  - Modeling individual probabilities
  - Random effects DR models
- Analysis of discrete compositional data
- Conclusions and Future Research

Discrete Compositions and Probability Models

- Compositional data are multivariate observations $Z = (Z_1,\ldots,Z_D)$ subject to the constraints that $\Sigma_i Z_i = 1$ and $Z_i \geq 0$. (measures relative size of each category)

- Compositional data are usually modeled with the *Logistic-Normal* distribution (Aitchison 1986).
  - Scale and location parameters provide a large amount of flexibility
  - LN model defined for positive compositions only

- Problem: With discrete counts one has a non-trivial probability of observing 0 individuals in a particular category

Existing Compositional Data Models

- Billhiemer and Guttorp (2001) proposed using a multinomial state-space model for a single composition,

$$\left(Y_{i1},\ldots,Y_{iD}\right) \sim \text{Multinomial}\left(N_i, Z_{i1},\ldots,Z_{iD}\right)$$

$$\left(Z_{i1},\ldots,Z_{iD}\right) \sim \text{LN}\left(\boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i\right),$$

  where Y_{ij} is the number of individuals belonging to category $j = 1,\ldots,D$ at site i = 1,…,S.

  Limitations:

  - Models proportions of a single categorical variable.
  - Abstract interpretation of included covariate effects

Graphical Models

- Graph model theory (see Lauritzen 1996) has been used for many years to
  - model cell probabilities for high dimensional contingency tables
  - determine dependence relationships among categorical and continuous variables

  Limitation:
  - Graphical models are designed for a single sample (or site in the case of the Oregon stream data). Compositional data may arise at many sites

New Improvements for Compositional Data Models

- The BG state-space model can be generalized by the application of graphical model theory.
  - Generalized models can be applied to cross-classified compositions
  - Simple interpretation of covariate effects as dependence in probability

- Conversely, the class of graphical models can be expanded to include models for multiple site sampling schemes

Graphical Chain Models

- Mathematical graphs are used to illustrate complex dependence relationships in a multivariate distribution

- A random vector is represented as a set of vertices, V.

  Ex. $V = \{\alpha$ = Precipitation, β = Stream velocity,

  γ = Amount of large rock in substrate$\}$

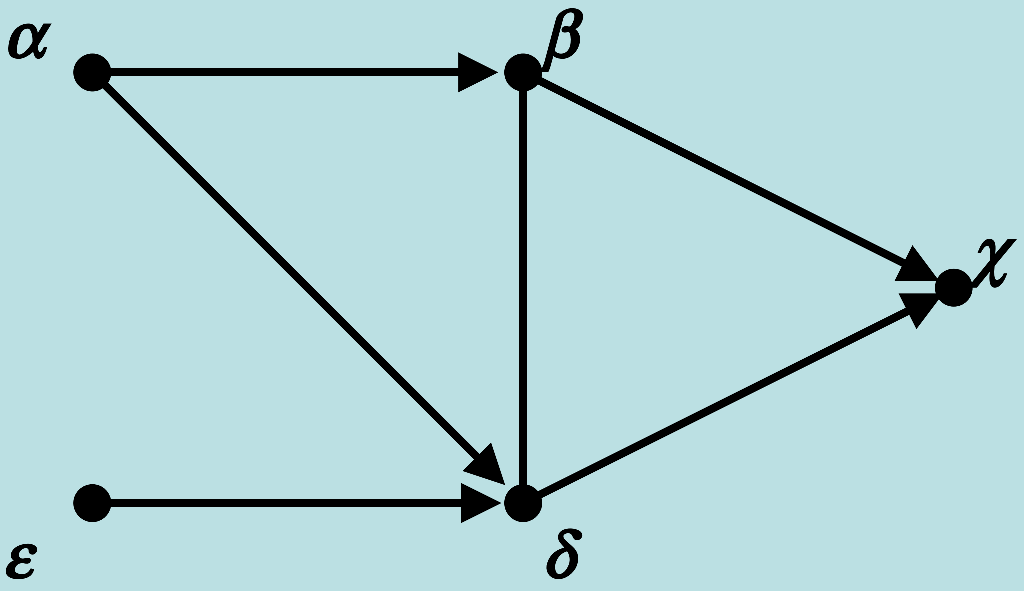
- Pairs of vertices are connected by directed or undirected edges depending on the nature of each pair's association

  Relationships are determined by a "causal" ordering

  If $\alpha < \beta$ in causal ordering, then $\alpha \rightarrow \beta$

  If $\beta = \gamma$, then $\beta - \gamma$

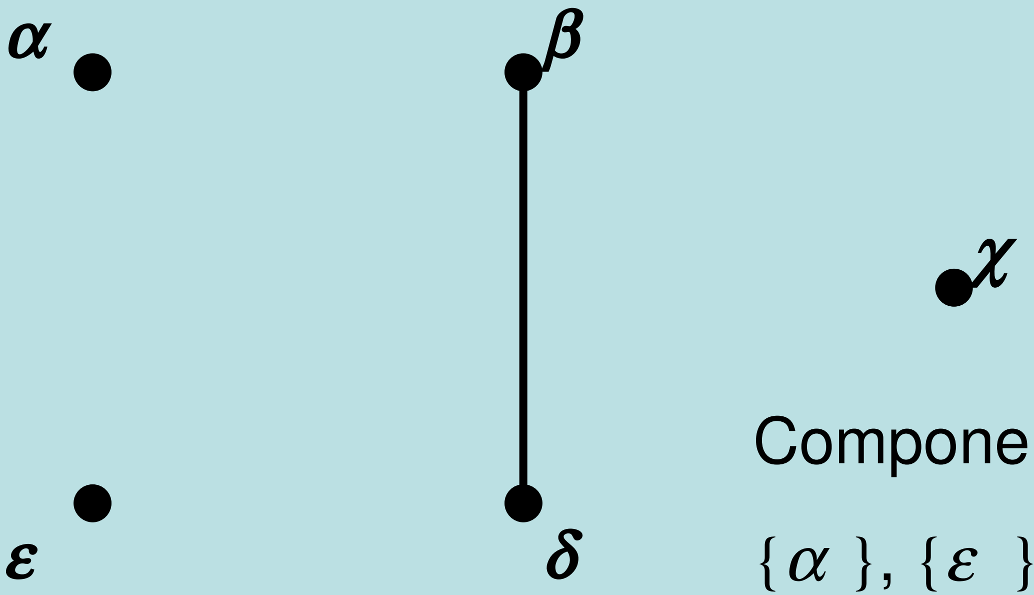
Example Chain Graph



Concepts

- **Causal ordering** $(\alpha , \varepsilon) < \beta = \delta < \chi$
- **Chain components** Sets of vertices whose elements are connected by undirected edges only

Example Chain Graph



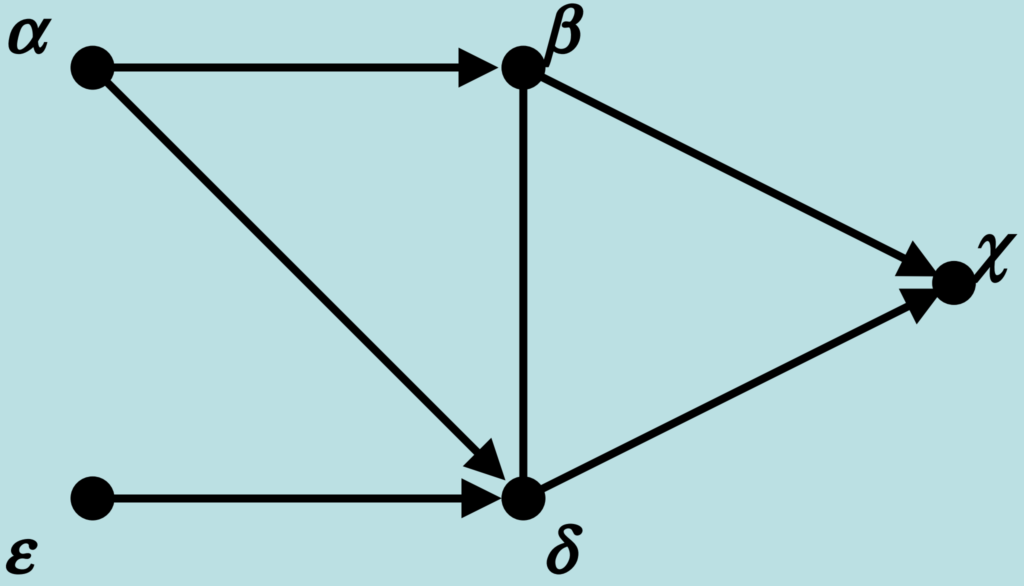
Components

$\{\alpha\}, \{\varepsilon\}, \{\beta, \delta\}, \{\chi\}$

<u>Concepts</u>

- **Causal ordering** $(\alpha, \varepsilon) < \beta = \delta < \chi$
- **Chain components** Sets of vertices whose elements are connected by undirected edges only

Example Chain Graph

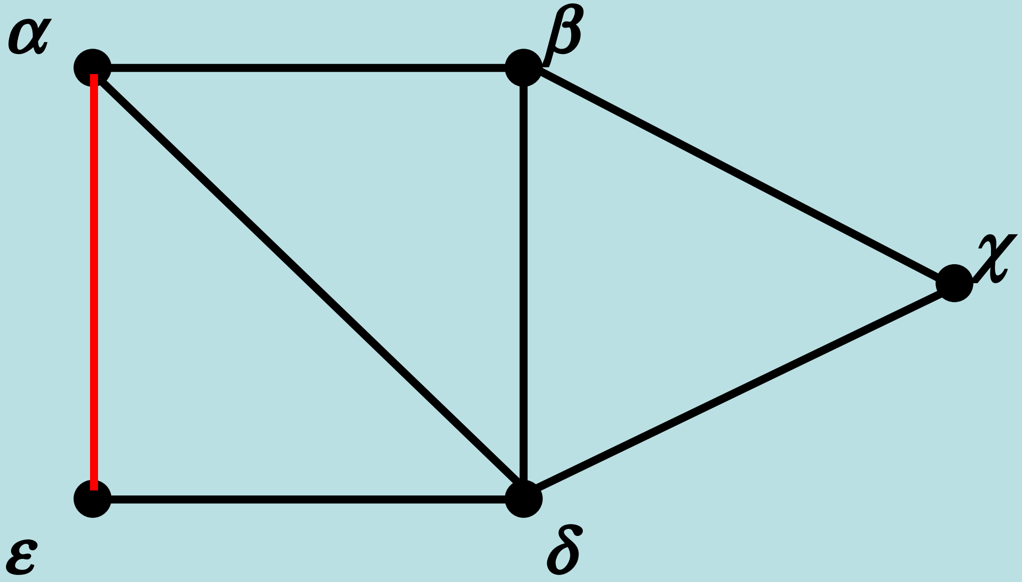


Concepts

- **Moral Graph** (G^m)**:** Graph induced by making all edges undirected and connecting parents of chain components

  Basis for determining dependence relationships between variables

Example Chain Graph

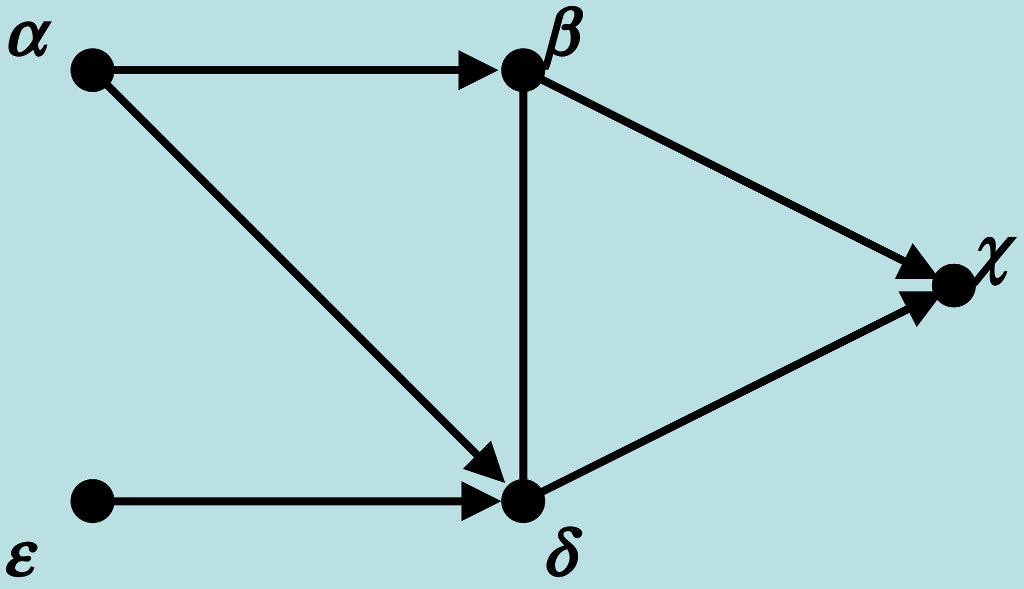


Concepts

- **Moral Graph** (G^m)**:** Graph induced by making all edges undirected and connecting parents of chain components

  Basis for determining dependence relationships between variables

Example Chain Graph



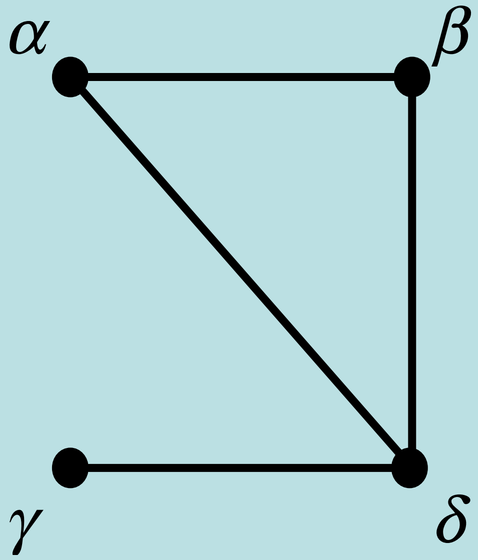
Concepts

- **Distribution models**: Joint distribution modeled as a product of conditional distributions.

  Ex. $f(\alpha , \beta , \delta , \gamma , \varepsilon) = f(\alpha) f(\varepsilon) f(\beta , \delta \mid \alpha , \varepsilon) f(\chi \mid \alpha , \varepsilon , \beta , \delta)$

Markov Properties of Undirected Graphs

- Let P denote a probability measure on the product space $X = X_\alpha \times X_\beta \times X_\gamma \times X_\delta$, and $V = \{\alpha, \beta, \gamma, \delta\}$



- Markov properties (w.r.t. P).
  - **Pairwise** Markovian $\alpha \perp \gamma \mid V \setminus \{\alpha, \gamma\}$.
  - **Local** Markovian $\beta \perp \gamma \mid (\alpha, \delta)$
  - **Global** Markovian $(\alpha, \beta) \perp \gamma \mid \delta$

Markov Properties and Factorization

- Markov relationships are related to the factorization of the joint density

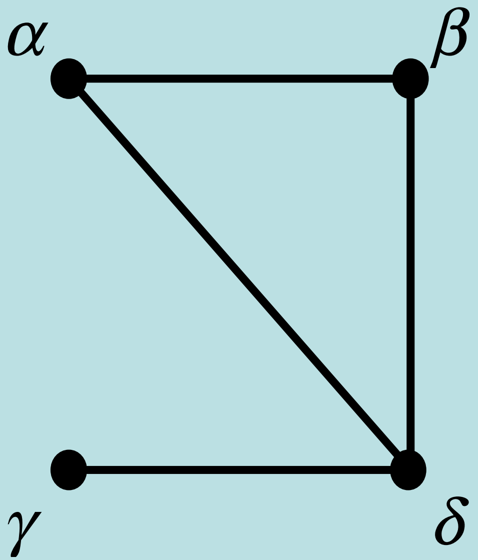
- **Theorem (Hammersley-Clifford)**.
  - G is an undirected graph
  - P has a positive and continuous density f with respect to a product measure μ .

  All three Markov properties are equivalent if and only if f factors as

$$f(\mathbf{x}) = \prod_{C \text{ complete}} h_C(\mathbf{x}_C)$$

- A **complete** set is one where all vertices in the set are connected to one another.

Factorization Example



$$
\begin{aligned}
f(\alpha,\beta,\delta,\gamma) &= f(\alpha \mid \beta,\delta,\gamma)\; f(\beta \mid \delta,\gamma)\; f(\delta \mid \gamma)\; f(\gamma) \\
&= f(\alpha \mid \beta,\delta)\; f(\beta \mid \delta)\; f(\delta \mid \gamma)\; f(\gamma) \\
&= h_{\{\alpha,\beta,\delta\}}(\alpha,\beta,\delta) \times h_{\{\delta,\gamma\}}(\delta,\gamma)
\end{aligned}
$$

Discrete Regression (DR) Chain Model

- **Response variables** (terminal chain component)
  - Set Δ of discrete categorical variables
  - Notation: $\mathbf{y}$ is a specific cell
- **Explanatory variables**
  - Set $\Gamma = \Gamma_D \cup \Gamma_C$ of categorical (Γ_D) or continuous (Γ_C) variables
  - Notation: $\mathbf{x}$ refers to a specific explanatory observation
- DR Joint distribution: $f(\mathbf{x})\, p(\mathbf{y}|\mathbf{x})$
- DR distribution is an example of a mixed variable graphical model (Lauritzen and Wermuth, 1989)

Discrete Regression Model (Response)

Model for conditional response:

$$p(\mathbf{y} \mid \mathbf{x}) = \exp \overset{\rightarrow}{} \alpha_{\Delta}(\mathbf{x}) + \underset{d\,\text{�}\,\Delta}{\text{�}} \underset{c\,\text{�}\,\Gamma}{\text{�}} \beta_{dc} \underset{\gamma \rightarrow c}{\longrightarrow} x_{\gamma}$$

$$+ \underset{d\,\text{�}\,\Delta}{\text{�}} \underset{c\,\text{�}\,\Gamma_D}{\text{�}} \underset{\gamma\,\text{�}\,\Gamma_C}{\text{�}} \overset{m}{\underset{j=2}{\text{�}}} \omega_{dc\gamma j} x_{\gamma}^{j} \overset{\rightarrow}{\text{�}}$$

- The function $\alpha_{\Delta}(\mathbf{x})$ is a normalizing constant w.r.t. $\mathbf{y}|\mathbf{x}$
- The parameters β_{dc} and $\omega_{dc\gamma\ j}$ are interaction effects that depend on $\mathbf{y}$ through the levels of the variables in d only.
- Certain interaction parameters are set to zero for identifiability of the model (analogous to interaction terms in ANOVA models)

Discrete Regression Model (Predictors)

- Model for explanatory variables (CG distribution):

$$f(\mathbf{x}) = \exp\left[\sum_{c\subseteq\Gamma_D}\lambda_c + \sum_{c\subseteq\Gamma_D}\sum_{\gamma\in\Gamma_C}\eta_{c\gamma}x_\gamma - \frac{1}{2}\sum_{c\subseteq\Gamma_D}\sum_{\mu,\gamma\in\Gamma_C}\psi_{c\mu\gamma}x_\mu x_\gamma\right]$$

- Again, interactions depend on $\mathbf{x}_{\Gamma_c}$ through the levels of the variables in the set c only, and identifiability constraints are imposed.

Markov Properties of Graphical Chain Models

- Frydenburg (1990) extended Hammersley-Clifford theorem for application to chain models
  - Markov properties are based on moral graphs constructed from “past” and “present” chain components (relative to the set of vertices in question).
  - For a distribution P with positive and continuous density f, P is Markovian if and only if f factors as

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

where C_τ represents a class of complete sets in $(G_{cl(\tau)})^m$ for all chain components.

Markov Properties of the DR Model

Proposition. A DR distribution is Markovian with respect to a chain graph G, with terminal chain component Δ and initial component Γ, if and only if

- $\boldsymbol{\beta}_{dc} \equiv \mathbf{0}$ unless d is complete and $c \subseteq pa(\delta)$ for every δ in d,
- $\boldsymbol{\omega}_{dc\gamma j} \equiv \mathbf{0}$ unless d is complete and $\{\gamma\} \cup c \subseteq pa(\delta)$ for every δ in d,
- $\boldsymbol{\lambda}_{c} \equiv \boldsymbol{\eta}_{c\gamma} \equiv \boldsymbol{\psi}_{c\mu\gamma} \equiv \mathbf{0}$ unless the sets corresponding to the subscripts are complete in G_Γ

Markov Properties of the DR Distribution

Sketch of Proof:

- LW prove conditions concerning the $\boldsymbol{\lambda}$, $\boldsymbol{\eta}$, and $\boldsymbol{\psi}$ parameters for the CG distribution, therefore, we only need look at the $\boldsymbol{\beta}$ and $\boldsymbol{\omega}$ interactions.

- If the $\boldsymbol{\beta}$ and $\boldsymbol{\omega}$ parameters are $\mathbf{0}$ for the specified sets then it is easy to see that the density factorizes on $(G_{cl(\tau\ )})^m$

- A modified version of the proof of the Hammersley-Clifford Theorem shows that if $p(\mathbf{y}|\mathbf{x})$ separates into complete factors, then, the corresponding $\boldsymbol{\beta}$ and $\boldsymbol{\omega}$ vectors for non-complete sets must be $\mathbf{0}$.

Random Effects for DR Models

- Sampling of individuals occurs at many different random sites, $i = 1,\ldots,S,$ where covariates are measured only once per site
- Hierarchical model:

$$p(\mathbf{y}_i \mid \mathbf{x}_i) = \exp \overset{\rightarrow}{\alpha}_\Delta(\mathbf{x}_i, \boldsymbol{\varepsilon}_i) + \underset{d \text{�} \Delta}{\text{�}} \underset{c \text{�} \Gamma}{\text{�}} \beta_{dc} \underset{\gamma \rightarrow c}{\longrightarrow} x_{i\gamma}$$

$$+ \underset{d \text{�} \Delta}{\text{�}} \underset{c \text{�} \Gamma_D}{\text{�}} \underset{\gamma \text{�} \Gamma_C}{\text{�}} \overset{m}{\underset{j=2}{\text{�}}} \omega_{dc\gamma j} x_{i\gamma}^{j} + \underset{d \text{�} \Delta}{\text{�}} \boldsymbol{\varepsilon}_{id} \text{�}$$

$$\boldsymbol{\varepsilon}_{id} \sim \begin{cases} \mathbf{0} & \text{if } d \text{ is not complete in } G \\ MVN\left(\mathbf{0}, \mathbf{T}_d^{-1}\right) & \text{if } d \text{ is complete in } G \end{cases}$$

- Markov properties still hold over the integrated likelihood in some cases.

Graphical Models for Discrete Compositions

- For a set Δ of categorical responses
  - Let D be the number of cross-classified cells
  - Y_{ij} = Number of observations in cell $j=1,\ldots,D$ at site $i=1,\ldots,S$

- **Likelihood**

$$(Y_{i1},\ldots,Y_{iD}) \mid X_\Gamma = x_\Gamma \sim \text{Multinomial}(N_i; p_{i1},\ldots,p_{iD}),$$

where p_{ij} is given by the DR random effects model

- **Covariate distribution**

$$X_\Gamma \sim CG(\boldsymbol{\lambda}, \boldsymbol{\eta}, \boldsymbol{\psi})$$

Parameter Estimation

- A Gibbs sampling approach is used for parameter estimation

- Hierarchical centering
  - Produces Gibbs samplers which converge to the posterior distributions faster
  - Most parameters have standard full conditionals if given conditional conjugate distributions.

- Independent priors imply that covariate and response models can be analyzed with separate MCMC procedures.

Stream Invertebrate Functional Groups

- 94 stream sites in Oregon were visited in an EPA REMAP study
- Response composition: Stream invertebrates were collected at each site and placed into 1 of 6 categories of functional feeding type

1. Collector-gatherer
2. Collector-filterer
3. Scraper
4. Engulfing predator
5. Shredder
6. Other (mostly, benthic herbivores)

Stream Covariates

- Environmental covariates: values were measured at each site for the following covariates

1. % Substrate composed of woody material
2. Alkalinity
3. Watershed area
4. Minimum basin elevation
5. Mean basin precipitation
6. % Barren land in watershed
7. Number of stream road crossings

Stream Invertebrate Model

- Composition Graphical Model:

$$\log p_{ij} = \alpha_{\Delta}\left(\mathbf{x}_i\right) + \beta_{0,j} + \sum_{\gamma=1}^{7}\beta_{\gamma,j}\left(x_{i\gamma} - \bar{x}_{\gamma}\right)s_{\gamma}^{-2} + \varepsilon_{ij}$$

$$\boldsymbol{\varepsilon}_i \sim MVN\left(\mathbf{0},\mathbf{T}_{\Delta}^{-1}\right)$$

and

$$\mathbf{x}_i\boldsymbol{\mu}\,\tilde{}\,\boldsymbol{\Psi}MVN\left(\quad,\quad{}_{\Gamma}^{-1}\right)$$

- Prior distributions

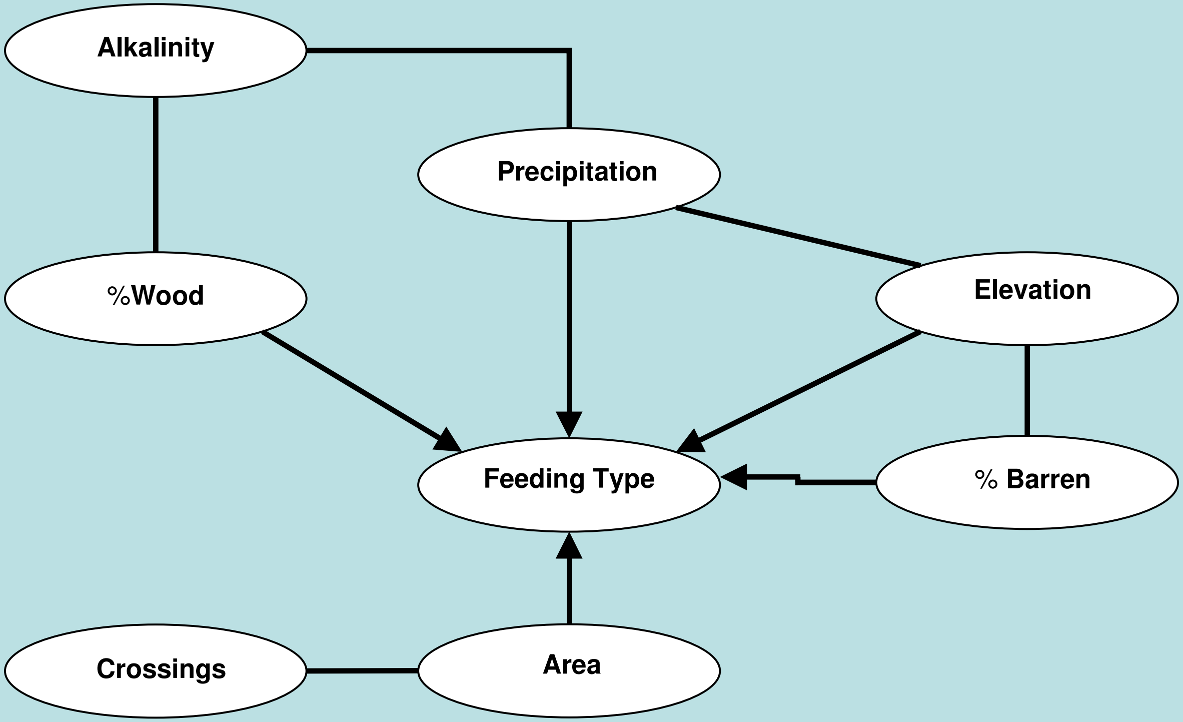
$$\beta_{\gamma,j}\left(x_{\Delta}\right) \sim \text{iid } N\left(0,\boldsymbol{\sigma}_{\gamma,j}^{2}\right),;\ldots,\qquad = 0 \qquad 7$$

$$\mathbf{T}_{\Delta} \sim \text{Wish}\left(6,\mathbf{R}\right)$$

$$\boldsymbol{\Psi}_{\Gamma} \sim \text{Wish}\left(7,\mathbf{R}\right)$$

Stream Invertebrate Functional Groups

Posterior suggested chain graph



Edge exclusion determined from 95% HPD intervals for $\boldsymbol{\beta}$ parameters and off-diagonal elements of $\boldsymbol{\psi}_{\Gamma}$.

Comments and Conclusions

- Using *Discrete Response* model with random effects, the BG model can be generalized
  - Relationships evaluated though a graphical model
  - Multiway compositions can be analyzed with specified dependence structure between cells
  - MVN random effects imply that the cell probabilities have a constrained LN distribution

- DR models also extend the capabilities of graphical models
  - Data can be analyzed from many multiple sites
  - Over dispersion in cell counts can be added

Future Work

- Model determination under a Bayesian framework
  - Models involve regression coefficients as well as many random effects

- Prediction of spatially correlated compositions over a continuous domain
  - Desirable to have a closed form predictor such as a kriging type predictor

Project Funding

The work reported here was developed under the STAR Research Assistance Agreement CR-829095 awarded by the U.S. Environmental Protection Agency (EPA) to Colorado State University. This presentation has not been formally reviewed by EPA. The views expressed here are solely those of presenter and the STARMAP, the Program he represents. EPA does not endorse any products or commercial services mentioned in this presentation.

This research is funded by

U.S. EPA - Science To Achieve Results (STAR) Program

Grant # **# CR - 829095**