

# Biological monitoring: Bayesian Models for a Multivariate Response

---

Jennifer A. Hoeting  
Department of Statistics  
Colorado State University  
[www.stat.colostate.edu/~jah](http://www.stat.colostate.edu/~jah)

Devin S. Johnson  
Department of Mathematical Sciences  
University of Alaska Fairbanks

---

## SOME ACKNOWLEDGMENTS AND OPPORTUNITIES

---

- I would like to thank the United States Environmental Protection Agency for funding this work.

The legalese: This work was partially funded by STAR Research Assistance Agreement CR-829095 awarded to Colorado State University by the U.S. Environmental Protection Agency (EPA). The views expressed here are solely those of authors. EPA does not endorse any products or commercial services mentioned here.

- Colorado State University Department of Statistics has **post-doctoral positions** available through its STARMAP and PRIMES programs funded by the EPA and the National Science Foundation. See my webpage for more information:

[www.stat.colostate.edu/~jah](http://www.stat.colostate.edu/~jah)

- Available this month:

**COMPUTATIONAL STATISTICS**  
by G. H. Givens and J. A. Hoeting  
Published by Wiley

See <http://www.stat.colostate.edu/computationalstatistics/>

- Compositional data are multivariate observations  $\mathbf{P} = (P_1, \dots, P_D)$  subject to the constraints that  $\sum_i P_i = 1$  and  $P_i \geq 0$
- Compositional data are usually modeled with the Logistic-Normal distribution (Aitchison 1986)
  - Scale and location parameters provide a large amount of flexibility compared to the Dirichlet model
  - LN model defined for positive compositions only
- Drawbacks:
  - With discrete counts one has a non-trivial probability of observing 0 individuals in a particular category
  - Cannot analyze multivariate compositions

- Introduce a model that allows for
  1. 0's in any category
  2. Analysis of multivariate compositions and estimation of interactions between compositions
  3. Graphical model conditional independence statements
- Adopt a Bayesian hierarchical random effects model
- Extend this model to allow for both discrete and continuous responses
- Area of application is an ecological problem: functional trait analysis

## FUNCTIONAL TRAIT VERSUS SPECIES ANALYSIS

---

- For biological monitoring, monitoring species may not be useful.

Example: Different species of fish live in different areas of the United States. A study that shows that one species of fish in Wisconsin is impacted by pollution may not be relevant for researchers in Colorado where that species is not present.

- Grouping species by functional traits may be more useful.

Example: Stream insects can be classified by their pollution tolerance. A sudden increase in the number of pollution tolerant species at a site may indicate that the site has become polluted.

## BIOLOGICAL MONITORING DATA: RESPONSE

---

- Organisms are sampled at several sites
- “Individuals” classified according to a set of traits  $\Phi$
- Individual response vector:  $\mathbf{Y} = \{Y_\phi : \phi \in \Phi\}$

### Example:

- 119 streams visited in an EPA EMAP study
- Fish were collected at each site
- Fish species were classified using 2 discrete traits
  1. Longevity: short ( $\leq 6$  years), long ( $\geq 6$  years)
  2. Trophic guild:  
herbivore, omnivore, invertivore, piscivore

A set  $\Omega$  of site specific environmental measurements (covariates) are also typically recorded

### Example

1. Sampling site elevation
2. Watershed area
3. Minimum watershed elevation

Let  $\mathbf{X} = \{X_\omega : \omega \in \Omega\}$  denote the vector of environmental covariates for a single sampling site

### 1. Ordination methods

- Canonical Correspondence Analysis (ter Braak, 1985)
- Ordinate traits along a set of environmental axes

### 2. Product moment correlations

- “Solution to the 4th corner problem” (Legendre et al. 1997)
- Estimate correlation measure between trait counts and environmental covariates

## SHORTCOMINGS OF PREVIOUS METHODS

---

1. Measure marginal association between environmental variables and traits
  - Conditional relationships give more detailed measure of association
  - Interaction between traits can give a different view
2. Models have no predictive ability
  - Cannot predict community structure at a site using remotely sense covariates (GIS covariates)

### Billheimer and Guttorp (JASA, 1997)

$$(C_{s1}, \dots, C_{sD}) \sim \text{Multinomial}(N_s; P_{s1}, \dots, P_{sD})$$

$$\log(P_{s1}/P_{sD}) = \beta_{0i} + \beta_{1i}x_s + \epsilon_{si}, \quad i = 1, \dots, (D - 1)$$

$$\epsilon_s \sim N_{D-1}(\mathbf{0}, \Sigma)$$

where

- $C_{si}$  = number of individuals belonging to category  $i$  at site  $s = 1, \dots, S$
- $x_s$  = environmental covariate at site  $s$
- Parameter estimation using a Gibbs sampler

- Generalize Billheimer–Guttorp model to explicitly allow for inference for multiple compositions
- Adopt graphical model structure with random effects
- New model allows for Markov random field interpretation for trait interactions
- Extend this model to allow both discrete and continuous responses

## SOME NOTATION

---

$i$  = Realization of  $Y$  (cell)

$\mathcal{I}$  = Sample space of  $Y$   
(not necessarily  $\mathcal{I}_1 \times \cdots \times \mathcal{I}_{|\Phi|}$ )

$P_{si}$  = Probability density of  $Y$  at site  $s = 1, \dots, S$   
(cell probability)

$C_{si}$  = number of individuals of type  $i$  at site  $s$   
(cell count)

$\phi$  = Single trait,  $\phi \in \Phi$

$a$  = Subset of traits,  $a \subseteq \Phi$

**Data model:**

$$\{C_{si} : i \in \mathcal{I}\} \sim \mathbf{Multinomial}(N_s; \{P_{si} : i \in \mathcal{I}\}) \quad s = 1, \dots, S$$

**or**

$$C_{si} \sim \mathbf{iid Poisson}(M_{si}); \quad i \in \mathcal{I}, s = 1, \dots, S$$

$$\mathbf{where } P_{si} = M_{si} / \sum_{j \in \mathcal{I}} M_{sj}$$

## INTERACTION PARAMETERIZATION

---

$$\log M_{si} = \sum_{a \subseteq \Phi} \mathbf{x}'_s \boldsymbol{\beta}_j^{(a)} + \sum_{a \subseteq \Phi} \epsilon_{si}^{(a)}$$

$$\left\{ \epsilon_{si}^{(a)} : i \in \mathcal{I} \right\} \sim MVN \left( \mathbf{0}, \mathbf{T}_a^{(-1)} \right)$$

- $\boldsymbol{\beta}_j^{(a)}$  and  $\epsilon_{si}^{(a)}$  measure the interaction between the traits in  $a$
- For model identifiability, choose reference cell  $i^*$  and set

$$\boldsymbol{\beta}_j^{(a)} = \mathbf{0} \text{ and } \epsilon_{si}^{(a)} = 0 \text{ if } i_\phi = i_\phi^* \text{ for any } \phi \in a$$

- If  $\mathcal{I} = \mathcal{I}_1 \times \cdots \times \mathcal{I}_{|\Phi|}$ , then

$$\mathbf{Y}_c \perp \mathbf{Y}_d \mid \mathbf{Y}_{\phi \setminus a}, \mathbf{X}, \boldsymbol{\epsilon} \quad \text{for } c, d \subset a \subseteq \Phi$$

$$\text{if } \boldsymbol{\beta}_i^{(a)} = \mathbf{0} \text{ and } \epsilon_i^{(a)} = 0 \quad \text{for all } i \in \mathcal{I}$$

- For certain model specifications

$$\mathbf{Y}_c \perp \mathbf{Y}_d \mid \mathbf{Y}_{\phi \setminus a}, \mathbf{X} \quad \text{for } c, d \subset a \subseteq \Phi$$

$$\text{if } \boldsymbol{\beta}_i^{(a)} = \mathbf{0} \quad \text{for all } i \in \mathcal{I}$$

- **Data:**

$$\Phi = \{1, 2\}; \mathbf{Y} = \{Y_1, Y_2\}; \text{ no covariates}$$

- **Saturated Model:**

$$\log M_{si} = \beta_i^{(1)} + \beta_i^{(2)} + \beta_i^{(12)} + \epsilon_{si}^{(1)} + \epsilon_{si}^{(2)} + \epsilon_{si}^{(12)}$$

- **Conditional independence model:**

$$\log M_{si} = \beta_i^{(1)} + \beta_i^{(2)} + \epsilon_{si}^{(1)} + \epsilon_{si}^{(2)}$$

implies

$$\mathbf{Y}_1 \perp \mathbf{Y}_2 | \epsilon \text{ (and in this case } \mathbf{Y}_1 \perp \mathbf{Y}_2)$$

Now consider an additional set of continuous traits measured for each species

**Example:**

How hydrodynamic is the fish species?

shape factor = body length/body depth

- Individual response vector:  $\mathbf{Y} = \{\mathbf{Y}_\phi, \mathbf{Y}_\Gamma\}$  where  $\mathbf{Y}_\phi$  is a vector of discrete traits and  $\mathbf{Y}_\Gamma$  is a vector of continuous traits
- $(i, \mathbf{y})$  represents a realization of  $\mathbf{Y}$

The conditional Gaussian distribution (Lauritzen, 1996)

$$\text{CG}(i, \mathbf{y}) \propto \exp \left\{ \sum_{a \subseteq \Phi} \lambda_i^{(a)} \right\} N \left\{ \mathbf{y}; \sum_{a \subseteq \Phi} \boldsymbol{\eta}_i^{(a)}, \left( \sum_{a \subseteq \Phi} \Psi_i^{(a)} \right)^{-1} \right\}$$

- $\boldsymbol{\eta}^{(a)}$  and  $\Psi^{(a)}$  measure interactions between discrete and continuous traits
- Homogeneous CG:  $\Psi_i^{(a)} = 0$  for  $a \neq \emptyset$

$$\mathbf{CG}(i, \mathbf{y}) \propto \exp \left\{ \sum_{a \subseteq \Phi} \lambda_i^{(a)} \right\} N \left\{ \mathbf{y}; \sum_{a \subseteq \Phi} \boldsymbol{\eta}_i^{(a)}, \left( \sum_{a \subseteq \Phi} \Psi_i^{(a)} \right)^{-1} \right\}$$

with

$$\lambda_{si}^{(a)} = \mathbf{x}'_{s'} \boldsymbol{\beta}_i^{(a)} + \epsilon_{si}^{(a)} \quad \text{with } \left\{ \epsilon_{si}^{(a)} : i \in \mathcal{I} \right\} \sim N(\mathbf{0}, \mathbf{T}_a^{-1})$$

$$\boldsymbol{\eta}_{si}^{(a)} = \mathbf{x}'_s \boldsymbol{\xi}_i^{(a)} + \boldsymbol{\delta}_{si}^{(a)} \quad \text{with } \left\{ \boldsymbol{\delta}_{si}^{(a)} : i \in \mathcal{I} \right\} \sim N(\mathbf{0}, \mathbf{K}_a^{-1})$$

- Reference cell identifiability constraints imposed
- Conditional independence inferred from zero-values parameters and random effects

# RANDOM EFFECTS CONDITIONAL GAUSSIAN HIERARCHICAL MODEL

---

$$\begin{aligned}
 p(\boldsymbol{\beta}, \boldsymbol{\eta}, \boldsymbol{\Psi}, \boldsymbol{\epsilon}_s, \boldsymbol{\delta}_s | \mathbf{x}_s, \mathbf{i}, \mathbf{Y}) &\propto \prod_{s=1}^S \prod_{j=1}^{N_s} \text{CG}(i_{sj}, \mathbf{y}_{sj} | \mathbf{x}_s, \boldsymbol{\beta}, \boldsymbol{\eta}, \boldsymbol{\Psi}, \boldsymbol{\epsilon}_s, \boldsymbol{\delta}_s) \\
 &\times \prod_{s=1}^S \prod_{a \subseteq \Phi} N(\{\epsilon_{si}^{(a)} : i \in \mathcal{I}\} | \mathbf{0}, \mathbf{T}_a^{-1}) \\
 &\times \prod_{s=1}^S \prod_{a \subseteq \Phi} N(\{\delta_{si}^{(a)} : i \in \mathcal{I}\} | \mathbf{0}, \mathbf{K}_a^{-1}) \\
 &\times p(\boldsymbol{\beta}, \boldsymbol{\eta}, \boldsymbol{\Psi}, \mathbf{T}, \mathbf{K})
 \end{aligned}$$

However, note the simplification

$$\prod_{j=1}^{N_s} \text{CG}(i_{sj}, \mathbf{y}_{sj} | \mathbf{x}_s, \boldsymbol{\beta}, \boldsymbol{\eta}, \boldsymbol{\Psi}, \boldsymbol{\epsilon}_s, \boldsymbol{\delta}_s) \propto \text{Mult}(\mathbf{C}_s | \dots) \prod_{j=1}^{N_s} N(\mathbf{y}_{sj} | \dots)$$

- A Gibbs sampling approach is used for parameter estimation
- Parameter sets  $(\beta, \epsilon, \mathbf{T})$  and  $(\eta, \Psi, \delta, \mathbf{K})$  can be simulated in 2 separate Gibbs chains due to
  1. The CG to Multinomial  $\times$  N formulation of the likelihood
  2. Independent priors for  $(\beta, \epsilon, \mathbf{T})$  and  $(\eta, \Psi, \delta, \mathbf{K})$
- Problem: Rich random effects structure can lead to poor convergence
- Solution: Hierarchical centering

$$\{\lambda_{si}^{(a)} : i \in \mathcal{I}\} \sim N\left(\{\mathbf{x}'_s \boldsymbol{\beta}_i^{(a)} : i \in \mathcal{I}\}, \mathbf{T}_a^{-1}\right)$$

$$\{\eta_{si}^{(a)} : i \in \mathcal{I}\} \sim N\left(\{\mathbf{x}'_s \boldsymbol{\xi}_i^{(a)} : i \in \mathcal{I}\}, \mathbf{K}_a^{-1}\right)$$

- $\boldsymbol{\beta}^{(a)}$ ,  $\boldsymbol{\xi}^{(a)}$ ,  $\mathbf{T}_a$ , and  $\mathbf{K}_a$  have closed form full conditional distributions
- $\lambda$  and  $\eta$  need to be updated with a Metropolis step in the Gibbs sampler

- 119 streams visited in an EPA EMAP study
- Fish were collected at each site
- Fish species were classified
  1. Discrete traits
    - (a) **Longevity**  
short ( $\leq 6$  years), long ( $\geq 6$  years)
    - (b) **Trophic guild**  
herbivore, omnivore, invertivore, piscivore
  2. Continuous trait: **shape factor**
- $i^* = (\leq 6 \text{ years, Herbivore})$

Environmental covariates were measured at each stream site or estimated for a watershed via GIS

1. Stream order
2. Minimum watershed elevation
3. Watershed area
4. Percent of area impacted by human use
5. Areal percent fish cover

● Interaction models

$$\lambda_{si}^{(a)} = \begin{cases} \mathbf{x}'_s \boldsymbol{\beta}_i^{(a)} + \epsilon_{si}^{(a)} & \text{for } a = \{L\}, \{T\} \\ \boldsymbol{\beta}_i^{(a)} & \text{for } a = \{L, T\} \end{cases}$$

$$\eta_{si}^{(a)} = \begin{cases} \mathbf{x}'_s \boldsymbol{\xi}_i^{(a)} + \delta_{si}^{(a)} & \text{for } a = \emptyset \\ \boldsymbol{\xi}_i^{(a)} & \text{for } a = \{L\}, \{T\}, \{L, T\} \end{cases}$$

● Random effects

$$\left\{ \epsilon_{si}^{(a)} : i \in \mathcal{I} \right\} \sim N(\mathbf{0}, \mathbf{T}_a^{-1}) \quad \text{for } a = \{L\}, \{T\}$$

$$\boldsymbol{\delta}_s^{(\emptyset)} \sim N(\mathbf{0}, \mathbf{K}_\emptyset^{-1})$$

Comparison of model with and without each covariate for longevity.

Covariate	> 6 years
Stream order	-0.59
Elevation	4.14
Watershed area	3.02
Use	7.32
Fish cover	5.03

Values presented are  $\approx 2\log(\text{BF})$ , so large positive values indicate evidence against including the covariate.

Comparison of null model to model including covariates for trophic guild

Covariate	Trophic Guild		
	Omnivore	Invertivore	Piscivore
Stream order	5.55	5.39	3.54
Elevation	-0.82	2.12	6.37
Watershed area	4.86	7.14	6.29
Use	5.50	7.24	0.70
Fish cover	7.03	5.49	6.35

Values presented are  $\approx 2\log(\text{BF})$ , so large positive values indicate evidence against including the covariate.

Comparison of null model to model including covariates for shape.

Covariate	Shape
Stream order	8.12
Elevation	8.23
Watershed area	8.22
Use	8.25
Fish cover	7.64

Values presented are  $\approx 2\log(\text{BF})$ , so large positive values indicate evidence against including the covariate.

## TRAIT INTERACTIONS

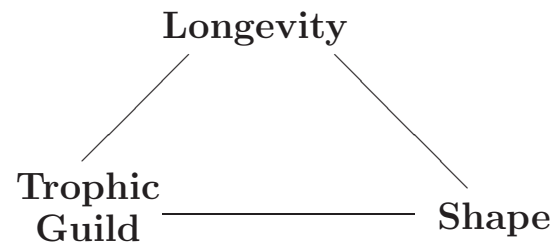
---

Comparison of null model to model including interaction parameters.

Interaction	$2\log(\text{BF})$
$\{L, T\}$	-18
$\{L, S\}$	-52
$\{T, S\}$	-104
$\{L, T, S\}$	-126

These results show strong support for evidence of interaction between the traits.

Graphical model for the response:



## CONCLUSIONS

---

- New model allows for
  1. Analysis of multiple compositions with specified interactions
  2. Analysis of discrete and continuous responses
  3. Markov random field interpretation for interactions
  4. Extension of Billheimer–Guttorp Model
    - Allows for full interaction and correlated random effects
    - MVN random effects imply that the cell probabilities have a constrained LN distribution
- Model can be applied in many other areas of application