

State-Space Models for Biological Monitoring Data

Devin S. Johnson

University of Alaska Fairbanks

and

Jennifer A. Hoeting

Colorado State University

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.



Outline

- Biological monitoring data
- Previous methods
- Bayesian hierarchical models
 - previous models
 - Multiple trait models
 - Continuous trait models
- Analysis of fish functional traits

Biological Monitoring Data

- Organisms are sampled at several sites.
- “Individuals” are classified according to a set Φ of traits

Example:

Longevity	Trophic guild
1. ≤ 6 years	1. herbivore
2. > 6 years	2. omnivore
	3. invertivore
	4. piscivore

- Individual response vector: $\mathbf{Y} = \{Y_\phi : \phi \in \Phi\}$

Environmental Conditions

- A set, Ω , of site specific environmental measurements (covariates) are also typically recorded.

Example:

stream order, watershed area, elevation

- Let

$$\mathbf{X} = \{X_{\omega} : \omega \in \Omega\}$$

Denote the vector of environmental covariates for a single sampling site

Functional Trait vs. Species Analysis

- Distributions of functional traits are often more interesting
 - Species are geographically constrained
 - Limited ecological inference
- Analysis of functional traits is portable
- Functional traits allow inference of the biological root of species distribution and environmental adaptation.

Previous Functional Trait Analysis Methods

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

- Measure marginal association between environmental variables and traits
 - Conditional relationships give a more detailed measure of association
 - Interaction between traits can give a different view
- No predictive ability
 - Cannot predict community structure at a site using remotely sensed covariates (GIS)

State-space models for a single trait

Billhiemer and Guttorp (1997)

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

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

$$\boldsymbol{\varepsilon}_s \sim N_{D-1}(\mathbf{0}, \boldsymbol{\Sigma})$$

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

Extending the Billheimer-Guttorp model

- Generalize the BG model to explicitly allow for multiple trait inference
 - Allow for a range of trait interaction
 - Parameterize to allow parsimonious modeling
- BG model based on random effect categorical data models
 - Use graphical model structure with random effects
 - Allow inference for trait interactions

Multiple trait analysis

Notation:

i	Realization of \mathbf{Y} (cell)
\mathcal{I}	Sample space of \mathbf{Y} (not necessarily $\mathcal{I}_1 \times \dots \times \mathcal{I}_{ \Phi }$)
P_{si}	Probability density of \mathbf{Y} at site $s = 1, \dots, \mathcal{S}$ (cell probability)
C_{si}	number of individuals of type i at site s (cell count)
ϕ	Single trait ($\phi \in \Phi$)
a	Subset of traits ($a \subseteq \Phi$)

Bayesian hierarchical model

- Data model:

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

or

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

where

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

- Parameter model:

$$\ln M_{si} \sim \text{MVN}(\boldsymbol{\mu}_s, \boldsymbol{\Sigma}); \quad s = 1, \dots, \mathcal{S}$$

Interaction parameterization

$$\ln M_{si} = \sum_{a \subseteq \Phi} \mathbf{x}'_s \boldsymbol{\beta}_i^{(a)} + \sum_{a \subseteq \Phi} \varepsilon_{si}^{(a)}$$

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

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

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

Conditional independence statements

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

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

if

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

- For certain model specifications:

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

if

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

Interaction example

- Data:

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

- Saturated model:

$$\ln M_{si} = \beta_i^{(1)} + \beta_i^{(2)} + \beta_i^{(12)} + \varepsilon_{si}^{(1)} + \varepsilon_{si}^{(2)} + \varepsilon_{si}^{(12)}$$

- Conditional independence model:

$$\ln M_{si} = \beta_i^{(1)} + \beta_i^{(2)} + \varepsilon_{si}^{(1)} + \varepsilon_{si}^{(2)}$$

implies

$$Y_1 \perp Y_2 \mid \varepsilon \text{ (and in this case } Y_1 \perp Y_2 \text{)}$$

Continuous traits

- In addition, for each individual, a set, Γ , of continuous traits are measured

Example:

Shape = Body length / Body depth

(How hydrodynamic is the individual?)

- Individual response vector: $\mathbf{Y} = \{ \mathbf{Y}_{\Phi}, \mathbf{Y}_{\Gamma} \}$
 - $\mathbf{Y}_{\Gamma} \in \mathbb{R}^{|\Gamma|}$ is a vector of interval valued traits
 - (i, \mathbf{y}) represents a realization of \mathbf{Y}

Conditional Gaussian distribution

- The conditional Gaussian distribution (Lauritzen, 1996, *Graphical Models*)

$$CG(i, \mathbf{y}) \propto \exp \left\{ \sum_{a \subseteq \Phi} \lambda_i^{(a)} \right\}$$

$$\times N \left\{ \mathbf{y}; \sum_{a \subseteq \Phi} \boldsymbol{\eta}_i^{(a)}, \left(\sum_{a \subseteq \Phi} \boldsymbol{\Psi}_i^{(a)} \right)^{-1} \right\}$$

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

Random effects CG Regression

$$\lambda_{si}^{(a)} = \mathbf{x}'_s \boldsymbol{\beta}_i^{(a)} + \varepsilon_{si}^{(a)}$$

and

$$\eta_{si}^{(a)} = \mathbf{x}'_s \boldsymbol{\xi}_i^{(a)} + \delta_{si}^{(a)}$$

where,

$$\left\{ \varepsilon_{si}^{(a)} : i \in \mathcal{I} \right\} \sim N(\mathbf{0}, \mathbf{T}_a^{-1}) \quad \text{and} \quad \left\{ \delta_{si}^{(a)} : i \in \mathcal{I} \right\} \sim N(\mathbf{0}, \mathbf{K}_a^{-1})$$

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

RECG Hierarchical model

$$\begin{aligned} & \prod_{s=1}^S \prod_{j=1}^{N_s} CG(i_{sj}, \mathbf{y}_{sj} \mid \mathbf{x}_s, \boldsymbol{\beta}, \boldsymbol{\eta}, \boldsymbol{\Psi}, \boldsymbol{\varepsilon}_s, \boldsymbol{\delta}_s) \\ & \times \prod_{s=1}^S \prod_{a \subseteq \Phi} \left\{ N\left(\left\{\boldsymbol{\varepsilon}_{si}^{(a)} : i \in \mathcal{I}\right\} \mid \mathbf{0}, \mathbf{T}_f^{-1}\right) \times N\left(\left\{\boldsymbol{\delta}_{si}^{(a)} : i \in \mathcal{I}\right\} \mid \mathbf{0}, \mathbf{K}_f^{-1}\right) \right\} \\ & \times \pi(\boldsymbol{\beta}, \boldsymbol{\eta}, \boldsymbol{\Psi}, \mathbf{T}, \mathbf{K}) \end{aligned}$$

However, note the simplification

$$\prod_{j=1}^{N_s} CG(i_{sj}, \mathbf{y}_{sj} \mid \dots) \propto \text{mult}(\mathbf{C}_s \mid \dots) \times \prod_{j=1}^{N_s} N(\mathbf{y}_{sj} \mid \dots)$$

Parameter estimation

- A Gibbs sampling approach is used for parameter estimation
- Analyze $(\beta, \varepsilon, \mathbf{T})$ and $(\eta, \Psi, \delta, \mathbf{K})$ with 2 separate Gibbs chains
 - The *CG* to *Mult* $\times N$ formulation of the likelihood
 - Independent priors for $(\beta, \varepsilon, \mathbf{T})$ and $(\eta, \Psi, \delta, \mathbf{K})$
- Problem:
 - Rich random effects structure can lead to poor convergence
 - Solution: Hierarchical centering

Hierarchical centering

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

and

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

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

Fish species trait richness

- 119 stream sites visited in an EPA EMAP study
- Discrete traits:

Longevity	Trophic guild
1. ≤ 6 years	1. herbivore
2. > 6 years	2. omnivore
	3. invertivore
	4. piscivore

- Continuous trait:

Shape factor = Body length / Body depth

- $i^* = (\leq 6 \text{ years, Herbivore}) = (1, 1)$

Stream covariates

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

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

Fish trait richness model

- Interaction models:

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

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

- Random effects:

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

$$\delta_s^{(\emptyset)} \sim N\left(0, K_{\emptyset}^{-1}\right)$$

Environment effects on Longevity

Table 1. *Comparison of “null” model to model including specified covariate for Longevity. Values presented are $\approx 2\ln(BF)$.*

Covariate	> 6 years
Stream order	-0.588 (↓↓)
Elevation	4.139
Area	3.022
Use	7.319
Fish cover	5.032

Environmental effects on Trophic Guild

Table 2. Comparison of “null” model to model including specified covariate for trophic guild. Values presented are $\approx 2\ln(BF)$.

Covariate	Trophic Guild		
	Omnivore	Invertivore	Piscivore
Stream order	5.550	5.393	3.538
Elevation	-0.824 (↓)	2.166	6.368
Area	4.863	7.142	6.292
Use	5.498	7.241	0.704 (↑)
Fish cover	7.031	5.487	6.351

Environmental effects on Trophic Guild

Table 3. *Comparison of “null” model to model including specified covariate for Shape. Values presented are $\approx 2\ln(BF)$.*

Covariate	Shape
Stream order	8.116
Elevation	8.228
Area	8.225
% impacted	8.249
Fish cover	7.636

Trait interaction

Table 4. *Comparison of “null” model to model including interaction parameters.*

Interaction	$2\ln(BF)$
$\{L, T\}$	-18.492
$\{L, S\}$	-52.183
$\{T, S\}$	-104.348
$\{L, T, S\}$	-126.604

Comments / Conclusions

- Multiple traits can be analyzed with specified interaction
- Continuous traits can also be included
- Markov Random Field interpretation for trait interactions
- BG model obtainable for multi-way traits
 - Allow full interaction and correlated R.E.
 - MVN random effects imply that the cell probabilities have a “constrained” LN distribution