Modeling compositional data

Background

NAPAP, 1980’s

Workshop on biological monitoring, 1986

Dirichlet process: Gary Grunwald, 1987

Current framework: Dean Billheimer, 1995

Other co-workers: Adrian Raftery, Mariabeth Silkey, Eun-Sug Park

Compositional data

Vector of proportions

$$z = (z_1, \ldots, z_k)^T \quad \boxed{z_i > 0} \quad \sum_1^k z_i = 1 \quad z \in \nabla^{k-1}$$

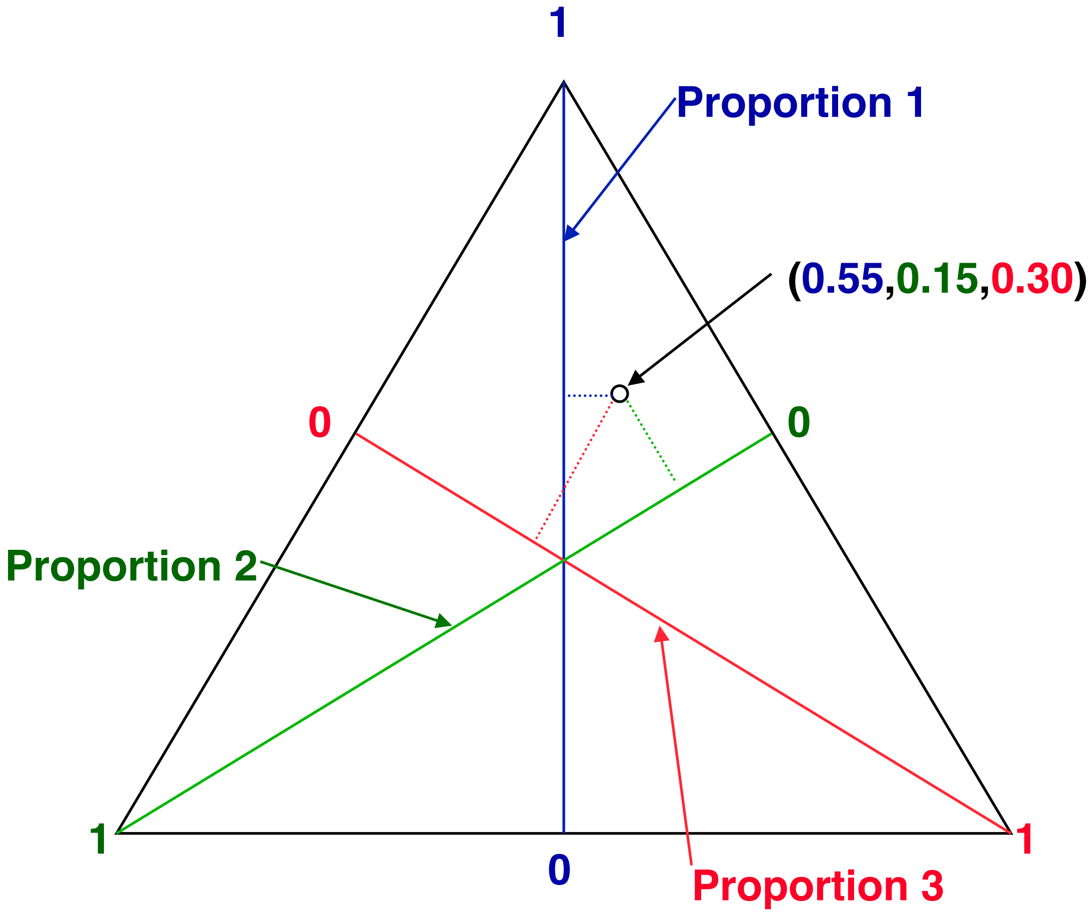
Proportion of taxes in different categories

Composition of rock samples

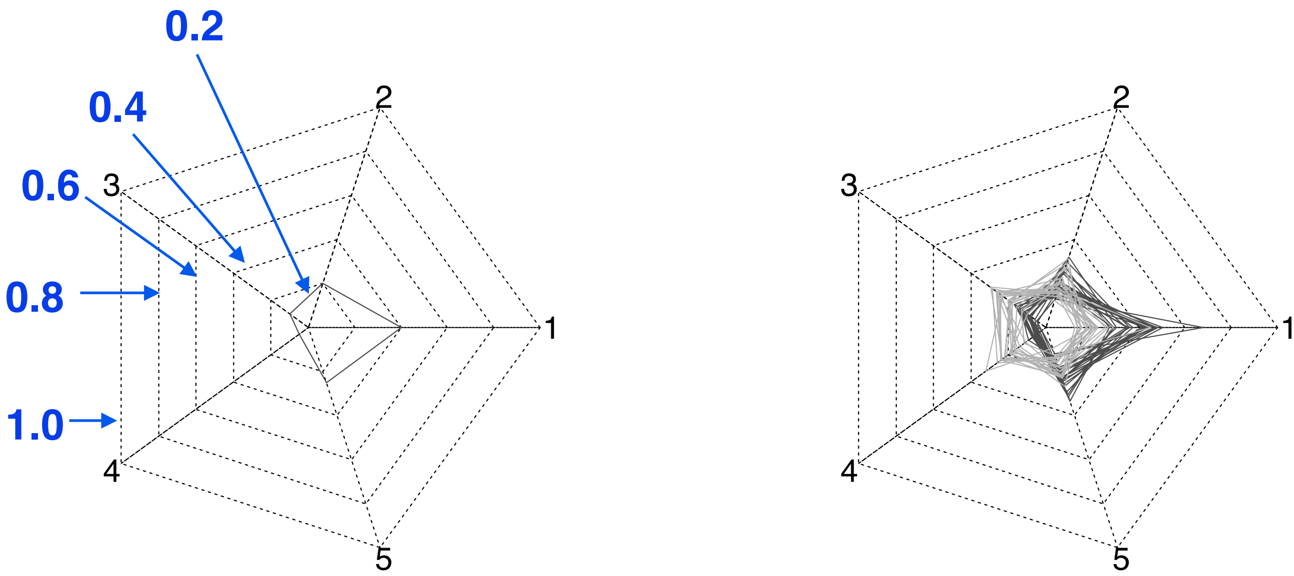
Composition of biological populations

Composition of air pollution

The triangle plot



The spider plot



(0.40,0.20,0.10,0.05,0.25)

An algebra for compositions

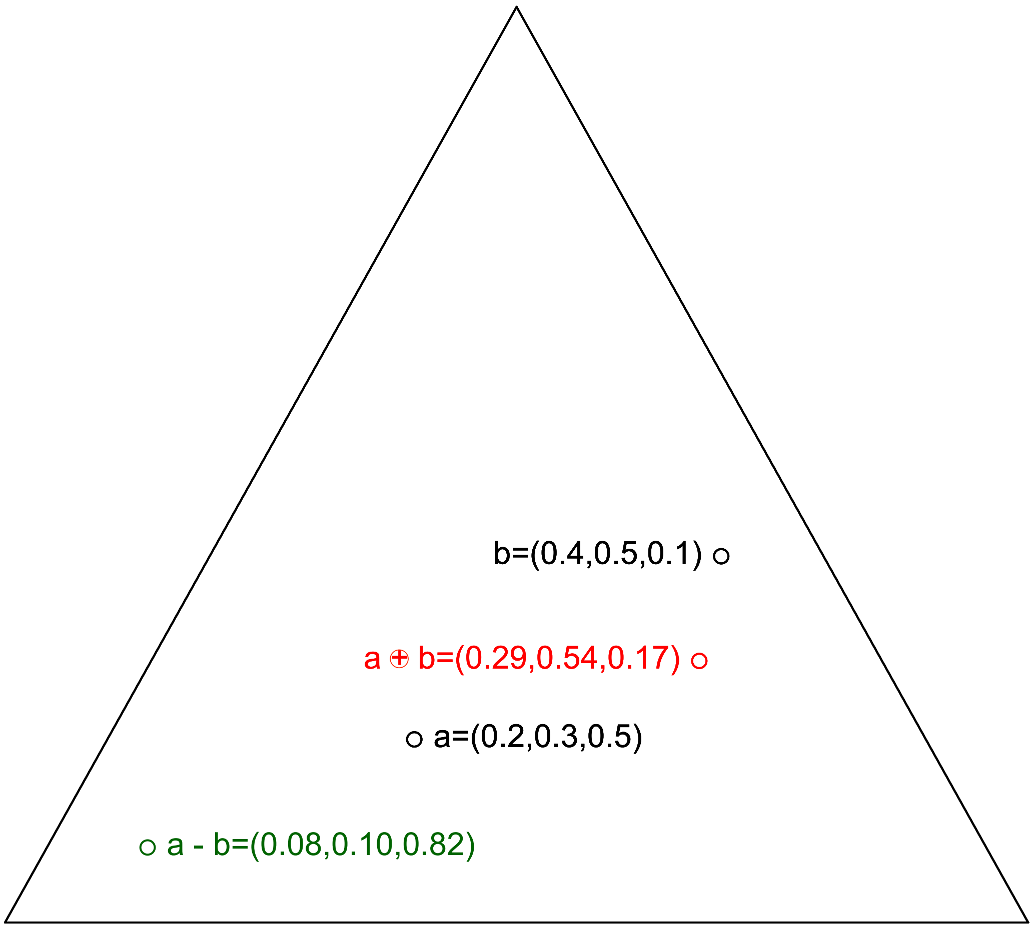
Perturbation: For $\xi,\alpha \in \nabla^{k-1}$ define

$$\xi \oplus \alpha = \left(\frac{\xi_1 \alpha_1}{\sum_1^k \xi_i \alpha_i}, \ldots, \frac{\xi_k \alpha_k}{\sum_1^k \xi_i \alpha_i} \right) \in \nabla^{k-1}$$

The composition $\iota = \left(\frac{1}{k}, \ldots, \frac{1}{k} \right)$ acts as a zero, so $\xi \oplus \iota = \xi$.

Set $\xi^{-1} = \left(\frac{1}{\xi_1}, \ldots, \frac{1}{\xi_k} \right)$ so $\xi \oplus \xi^{-1} = \iota$.

Finally define $\xi - \eta = \xi \oplus \eta^{-1}$.


b=(0.4,0.5,0.1)
a ⊕ b=(0.29,0.54,0.17)
a=(0.2,0.3,0.5)
a - b=(0.08,0.10,0.82)

The logistic normal

If $\text{alr}(z) = \left(\log\frac{z_1}{z_k}, \ldots, \log\frac{z_{k-1}}{z_k}\right)^T \sim MVN(\mu, \Sigma)$

we say that z is *logistic normal*, in short $Z \sim LN(\mu,\Sigma)$.

alr has a unique inverse.

Other distributions on the simplex:

Dirichlet — ratios of independent gammas

"Danish" — ratios of independent inverse Gaussian

Both have very limited correlation structure.

Scalar multiplication

Let a be a scalar. Define

$$\xi \otimes a = \left(\frac{\xi_1^a}{\sum \xi_i^a}, \ldots, \frac{\xi_k^a}{\sum \xi_i^a} \right)$$

$\left(\nabla^{k-1}, \oplus, \otimes\right)$ is a complete inner product space, with inner product given, e.g., by

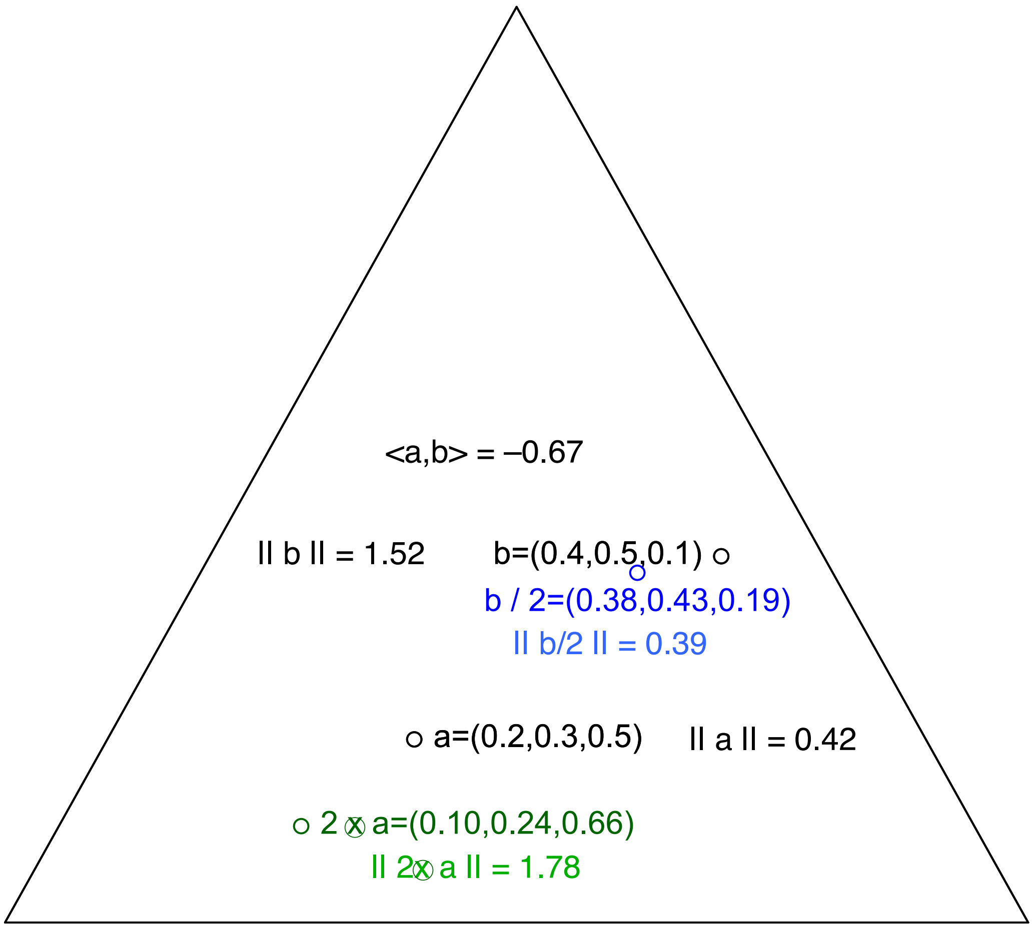
$$\langle \xi, \eta \rangle = \mathrm{alr}(\xi)^T N^{-1} \mathrm{alr}(\eta)$$

N is the precision matrix $N = I + jj^T$

j is a vector of k-1 ones.

$\|\xi\| = \langle \xi, \xi \rangle$ is a norm on the simplex.

The inner product and norm are invariant to permutations of the components of the composition.


<a,b> = −0.67
ll b ll = 1.52
b=(0.4,0.5,0.1)
b / 2=(0.38,0.43,0.19)
ll b/2 ll = 0.39
a=(0.2,0.3,0.5)
ll a ll = 0.42
2 ⊗ a=(0.10,0.24,0.66)
ll 2 ⊗ a ll = 1.78

Some models

Measurement error:

$$z_j = \xi \oplus \varepsilon_j \qquad \text{where } \varepsilon_j \sim LN(0,\Sigma)\ .$$

Regression:

$$\xi_j = \xi \oplus \gamma \otimes u_j$$

centered covariate

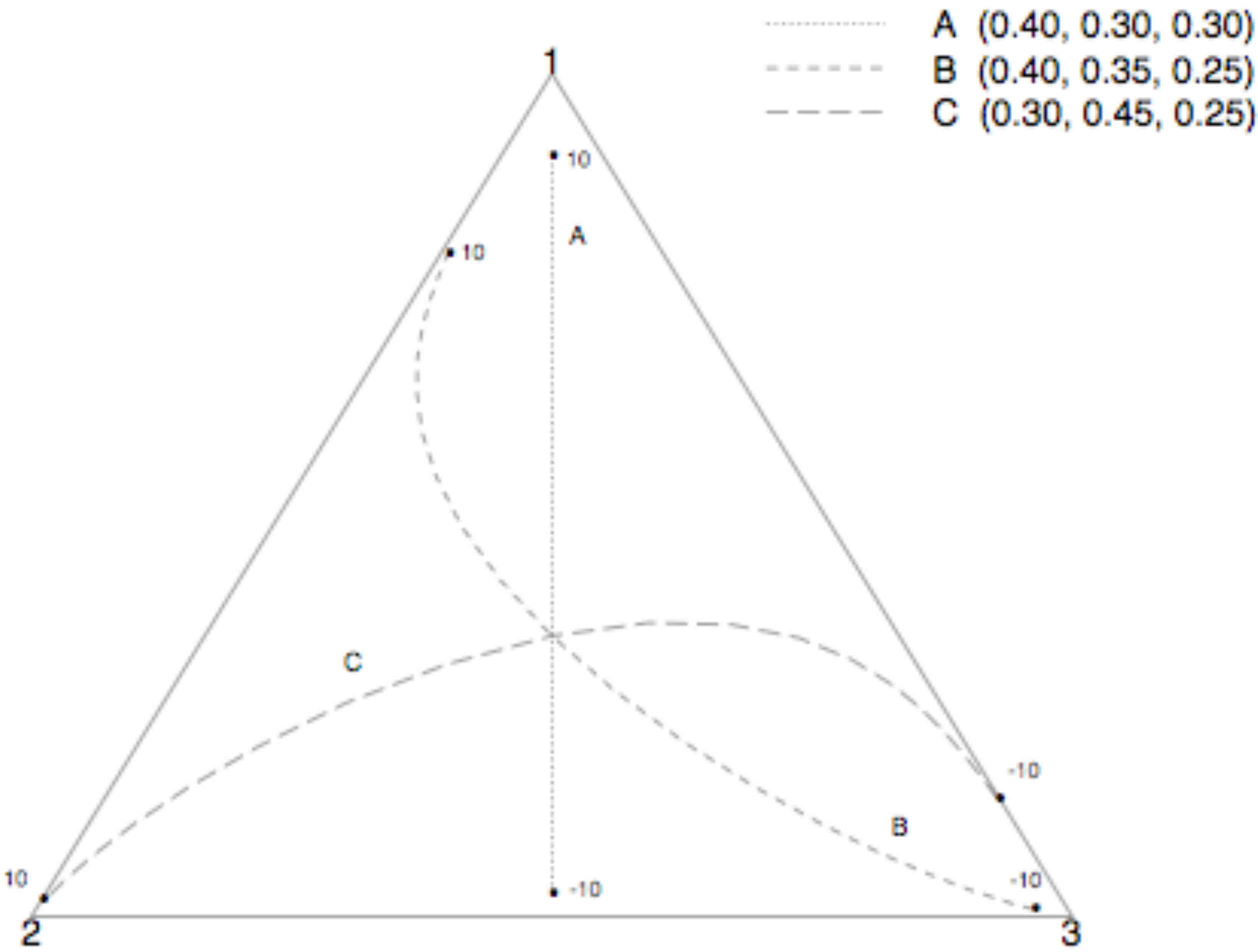
compositions

Correspondence in Euclidean space:

$$\mu_j = \beta_0 + \beta_1 (x_j - \bar{x})$$

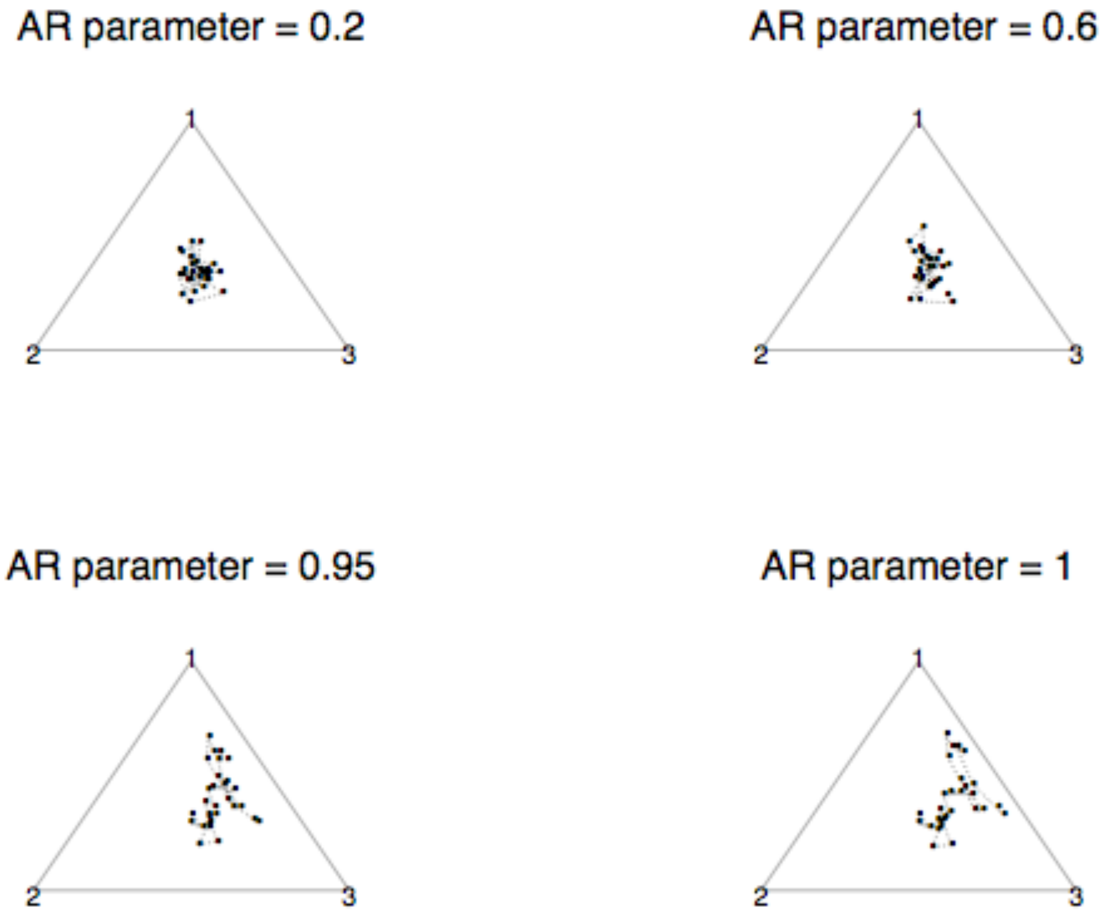
$$\underset{\xi_j}{\mathrm{alr}^{-1}(\mu_j)} = \underset{\xi}{\mathrm{alr}^{-1}(\beta_0)} \oplus \underset{\gamma}{\mathrm{alr}^{-1}(\beta_1)} \otimes \underset{u_j}{(x_j - \bar{x})}$$

Some regression lines



Time series (AR 1)

$$z_{k+1} = \phi \otimes z_k \oplus \varepsilon_k$$



A source receptor model

Observe relative concentration Y_i of k species at a location over time.

Consider p sources with chemical profiles θ_j. Let α_i be the vector of mixing proportions of the different sources at the receptor on day i.

$$EY_i = \sum_{i=1}^{p} \alpha_{ij}\theta_j = \Theta\alpha_i$$

$$Y_i = \Theta\alpha_i \oplus \varepsilon_i$$

$\Theta \sim$ **LN,** $\alpha_i \sim$ **indep LN,** $\varepsilon_i \sim$ **zero mean LN**

Juneau air quality

50 observations of relative mass of 5 chemical species. Goal: determine the contribution of wood smoke to local pollution load.

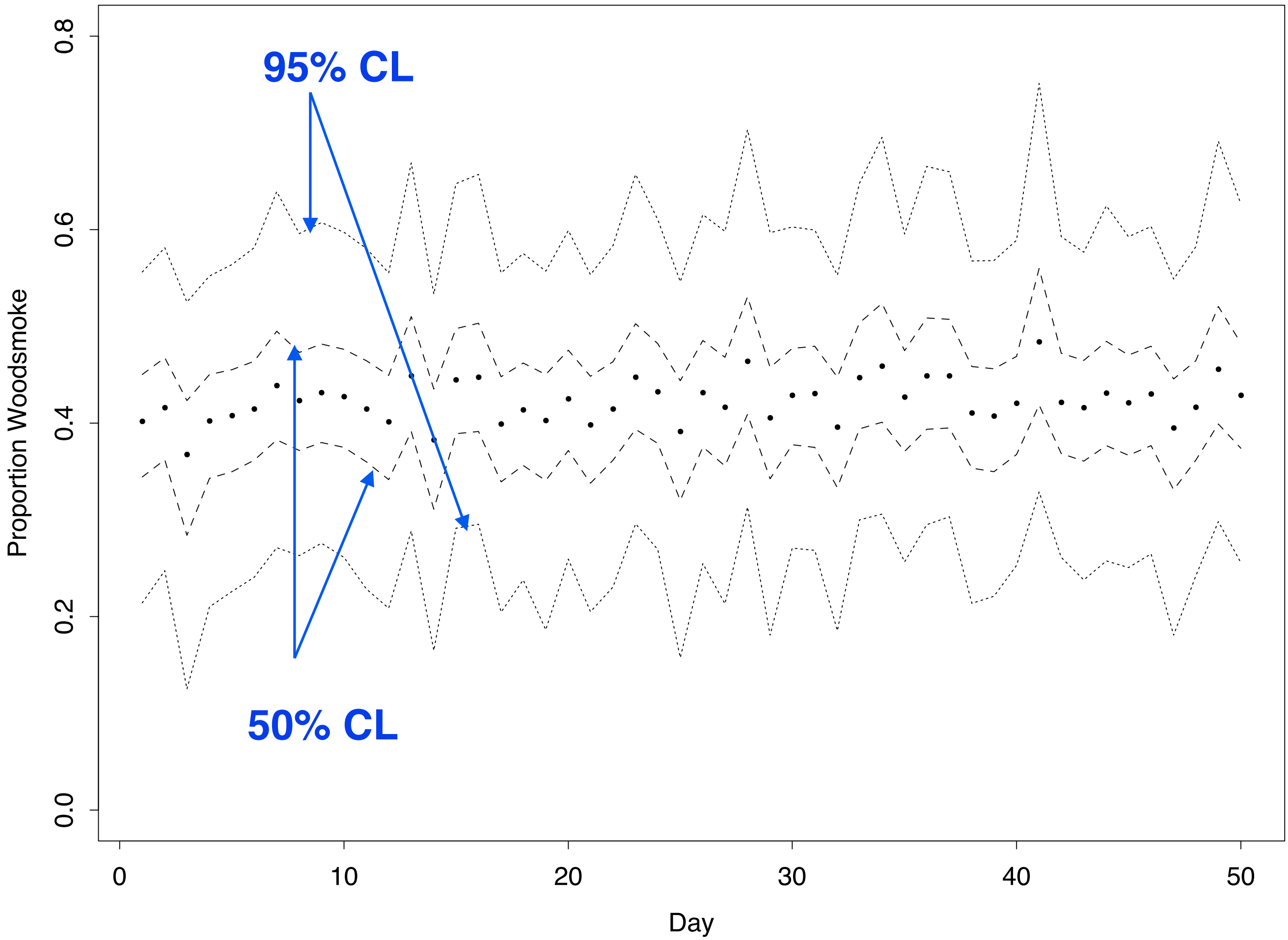
Prior specification:

$$f(\Theta, \alpha_i, \varepsilon_i, \mu_\alpha, \Gamma, \Sigma_\varepsilon) =$$

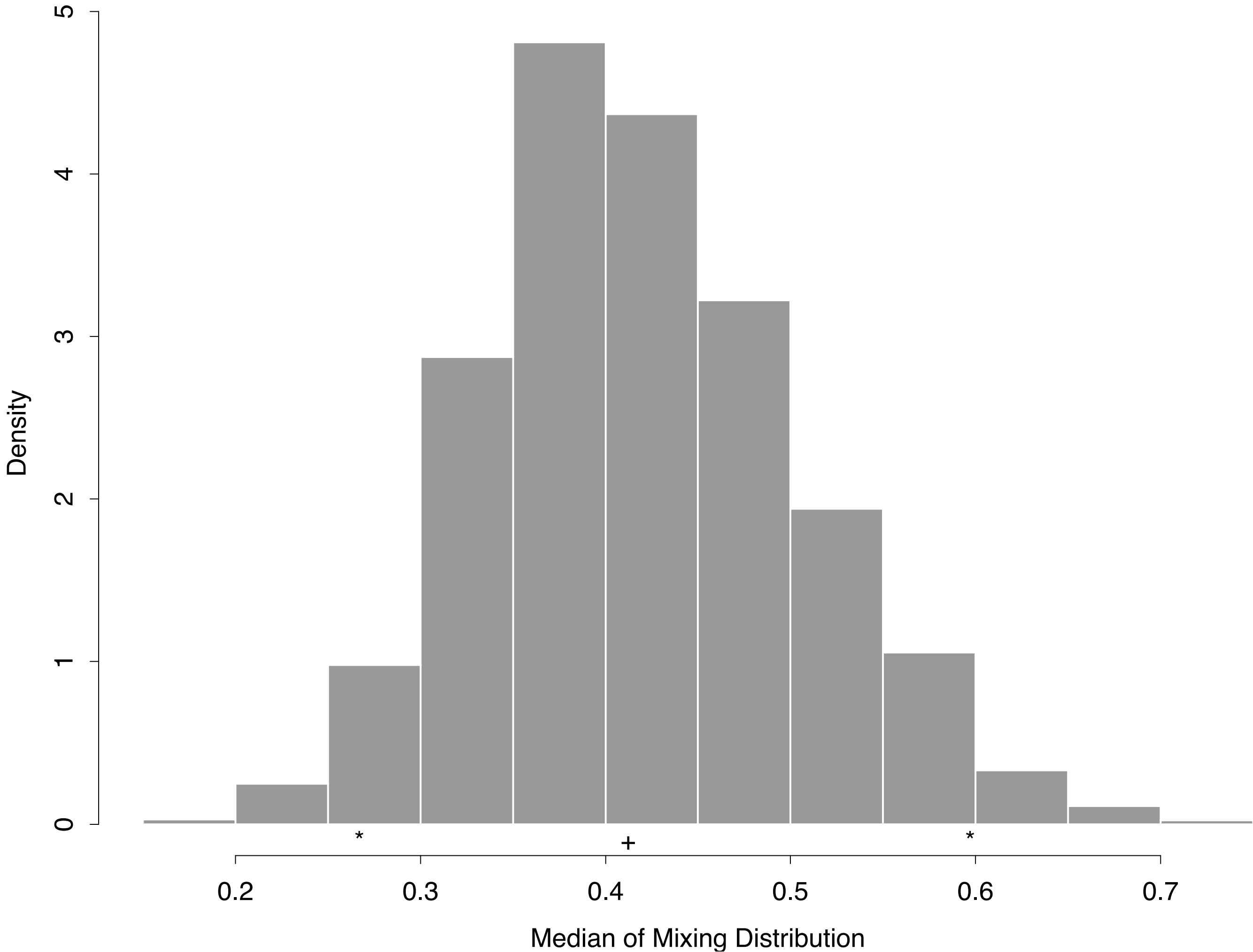
$$f(\alpha_i \mid \mu_\alpha, \Gamma) f(\varepsilon_i \mid \Sigma_\varepsilon) f(\mu_\alpha) f(\Gamma) f(\Sigma_\varepsilon)$$

Inference by MCMC.

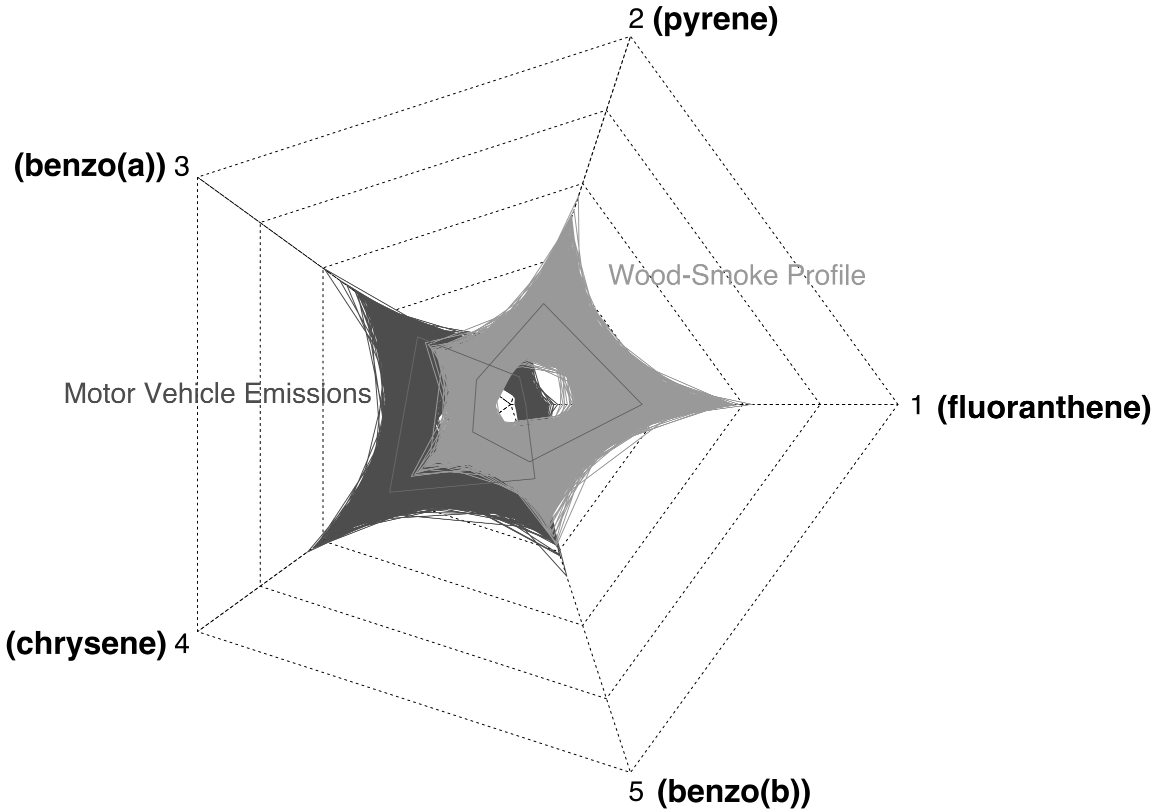
Wood smoke contribution



Wood smoke proportion



Posterior source profiles



State-space model

Space-time model of proportions

State-space model:

z_j unobservable composition ~ $LN(\mu_j, \Sigma_j)$

y_j k-vector of counts ~ $\mathrm{Mult}(\sum_{i=1}^{k} [y_j]_i, z_j)$

Inference using MCMC again

Stability of arthropod food webs

Omnivory thought to destabilize ecological communities

Stability: Capacity to recover from shock (relative abundance in trophic classes)

Mount St. Helens experiment: 6 treatments in 2-way factorial design; 5 reps.

- Predator manipulation (more omnivores, more specialists, control)
- Vegetation disturbance (50% reduction, control)

Count anthropods, 6 wks after treatment. Divide into specialized herbivores, general herbivores, predators.

Manipulated species

Omnivore:
Wolf spider

Specialist predator
Big-eyed bug

Vegetation

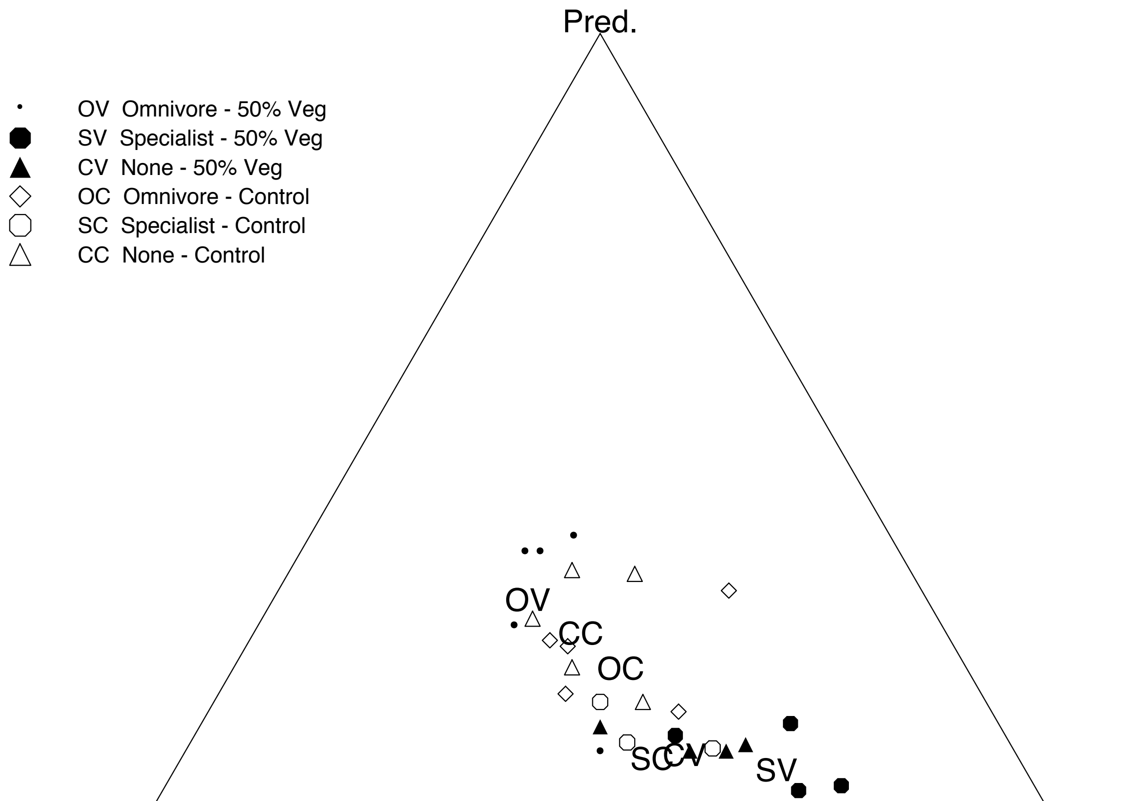
fireweed

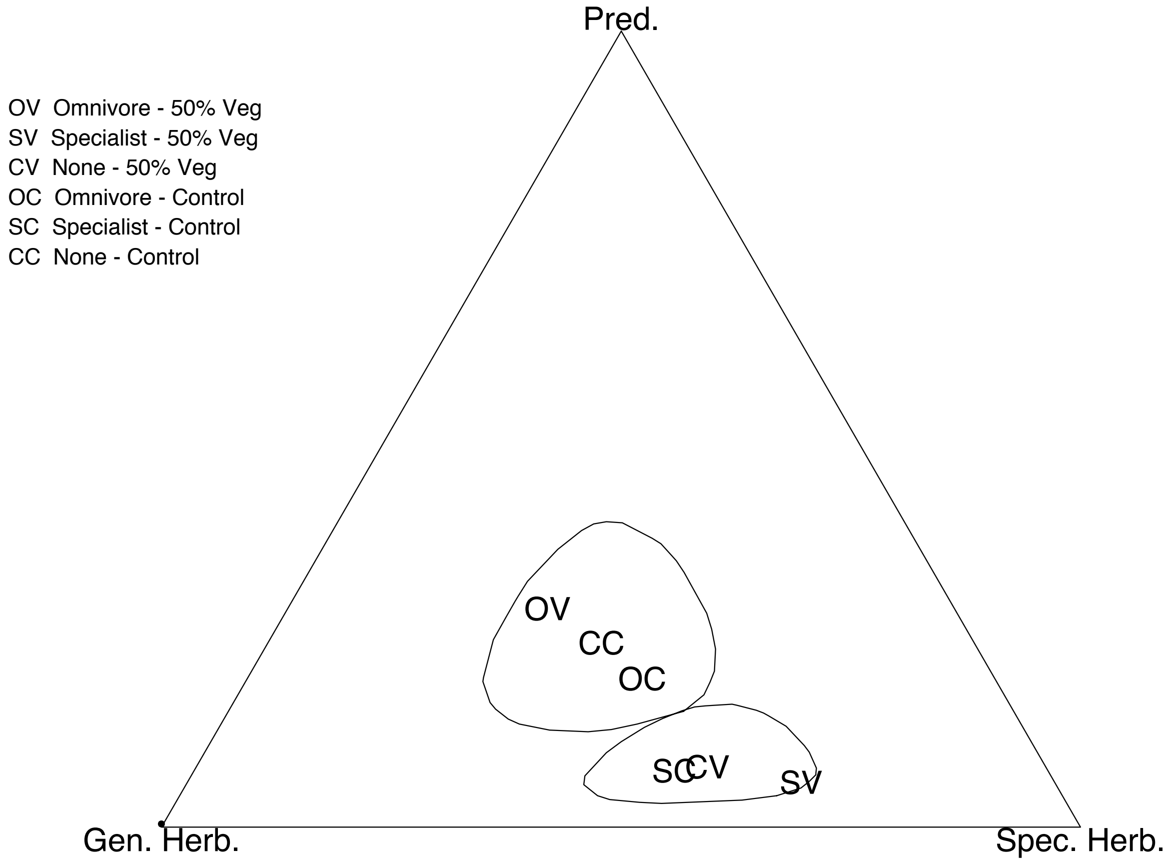
pearly-
everlasting

Specification of structure

Σ is generated from independent observations at each treatment

mean depends only on treatment


Pred.
Gen. Herb.
Spec. Herb.
OV Omnivore - 50% Veg
SV Specialist - 50% Veg
CV None - 50% Veg
OC Omnivore - Control
SC Specialist - Control
CC None - Control
OV
CC
OC
SC
CV
SV



Pred.
OV Omnivore - 50% Veg
SV Specialist - 50% Veg
CV None - 50% Veg
OC Omnivore - Control
SC Specialist - Control
CC None - Control
OV
CC
OC
SC
CV
SV
Gen. Herb.
Spec. Herb.

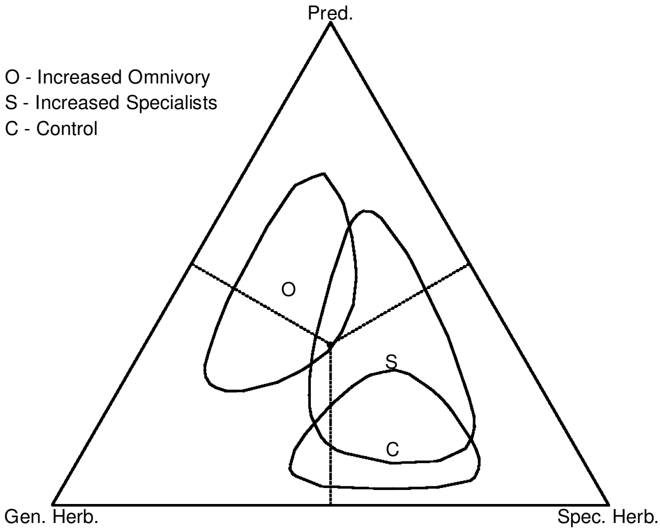
Interaction effect

ANOVA interaction effect

$$z_{ij} - \bar{z}_{i\bullet} - \bar{z}_{j\bullet} + \bar{z}_{\bullet\bullet}$$

alr inverse to get

$$\xi_{ij} - \bar{\xi}_{i\bullet} - \bar{\xi}_{j\bullet} \oplus \bar{\xi}_{\bullet\bullet}$$



Benthic invertebrates in estuary

EMAP estuaries monitoring program: Delaware Bay 1990. 25 locations, 3 grab samples of bottom sediment during summer

Invertebrates in samples classified into

- –pollution tolerant

- –pollution intolerant

- –suspension feeders (control group)

Site j, subsample t

$$z_{jt} \sim LN(\theta_j + \beta \boxed{x_j}, \Psi)$$

standardized covariate

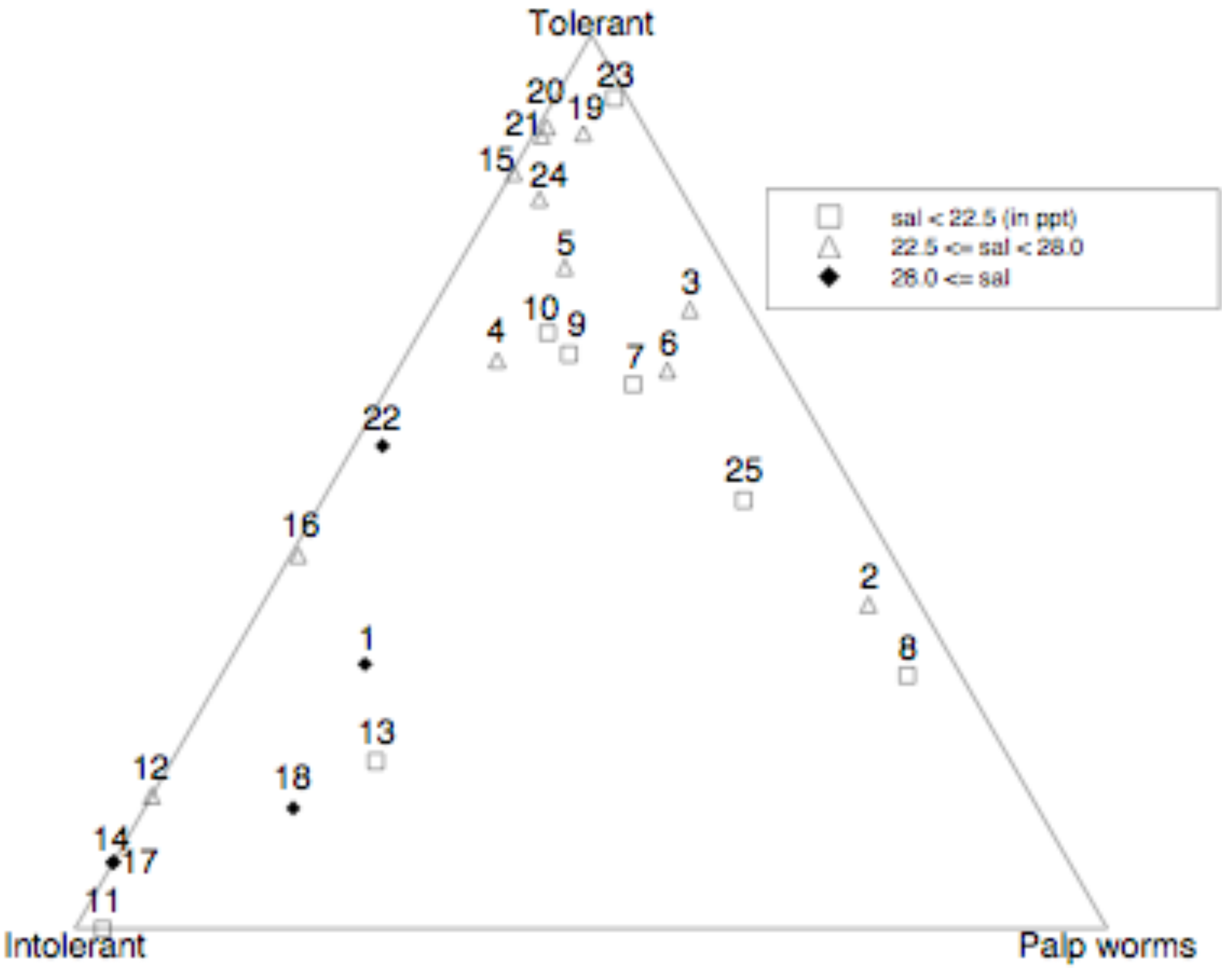
θ_j **~ CAR process**

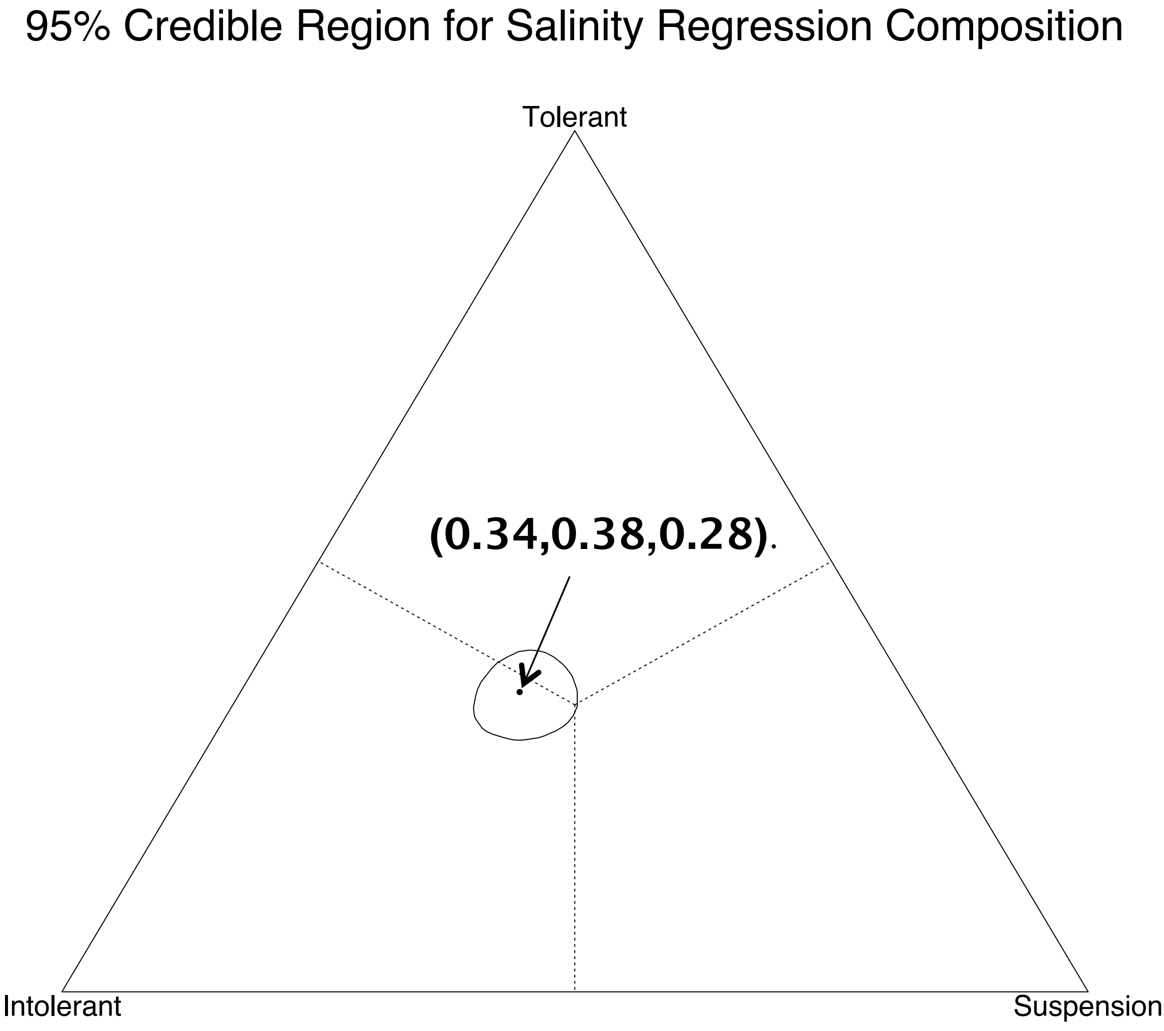
$$E(\theta_j \mid \theta_{-j}) = \mu + \sum_{k \in N(j)} \frac{\lambda}{n_j}(\theta_k - \mu)$$

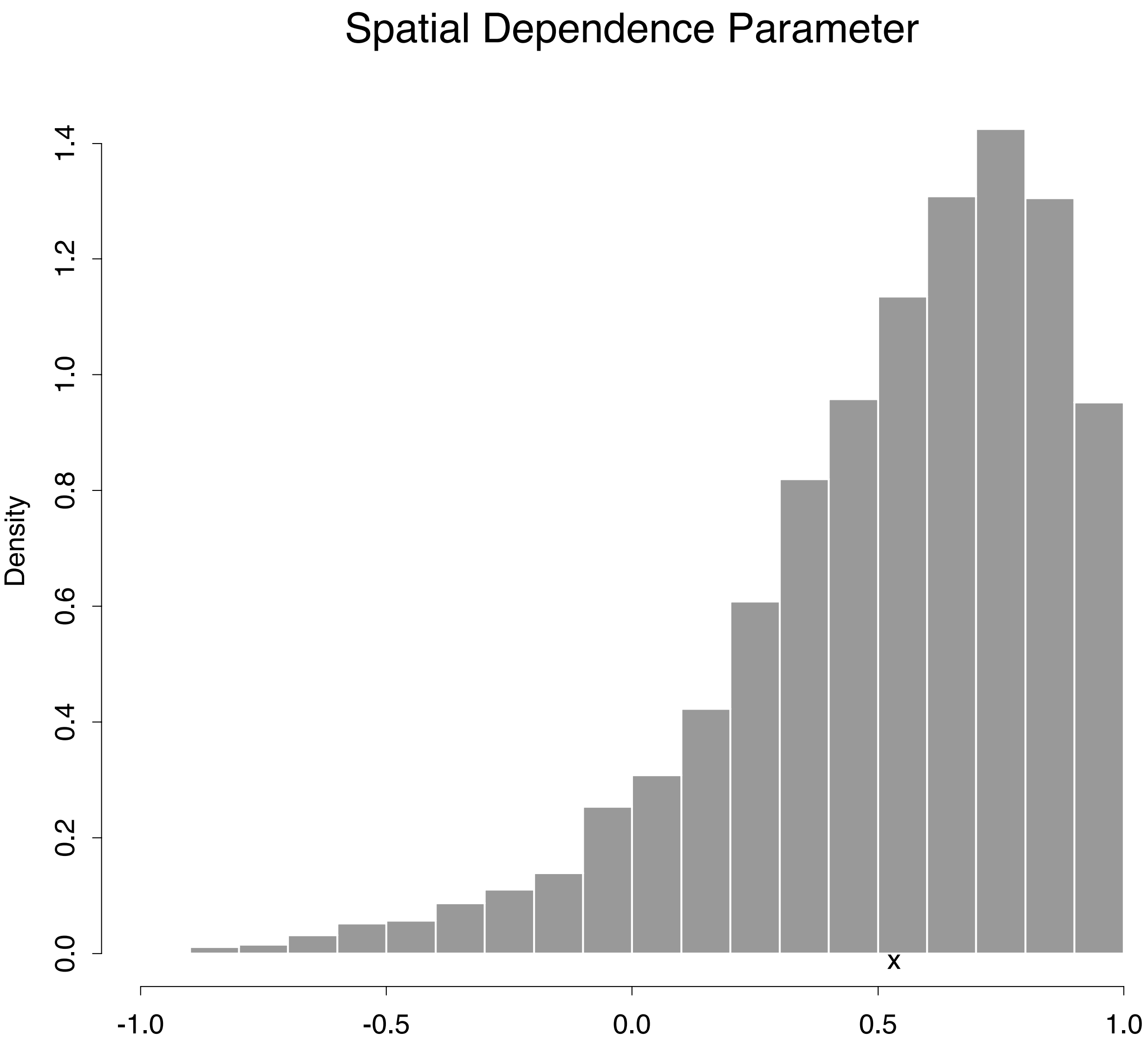
$$Var(\theta_j \mid \theta_{-j}) = \frac{\Gamma}{n_j}$$

11
13
10
25
9
8
23
7
2
6
5
4
3
24
20
21
19
17
16
18
14
15
22
12
1

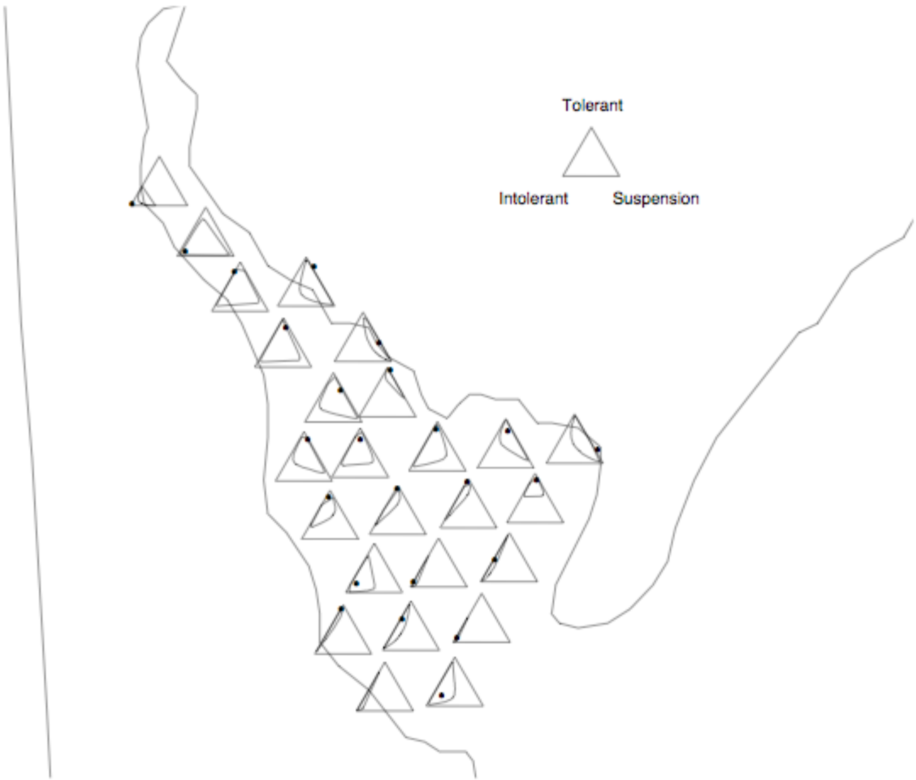
Effect of salinity



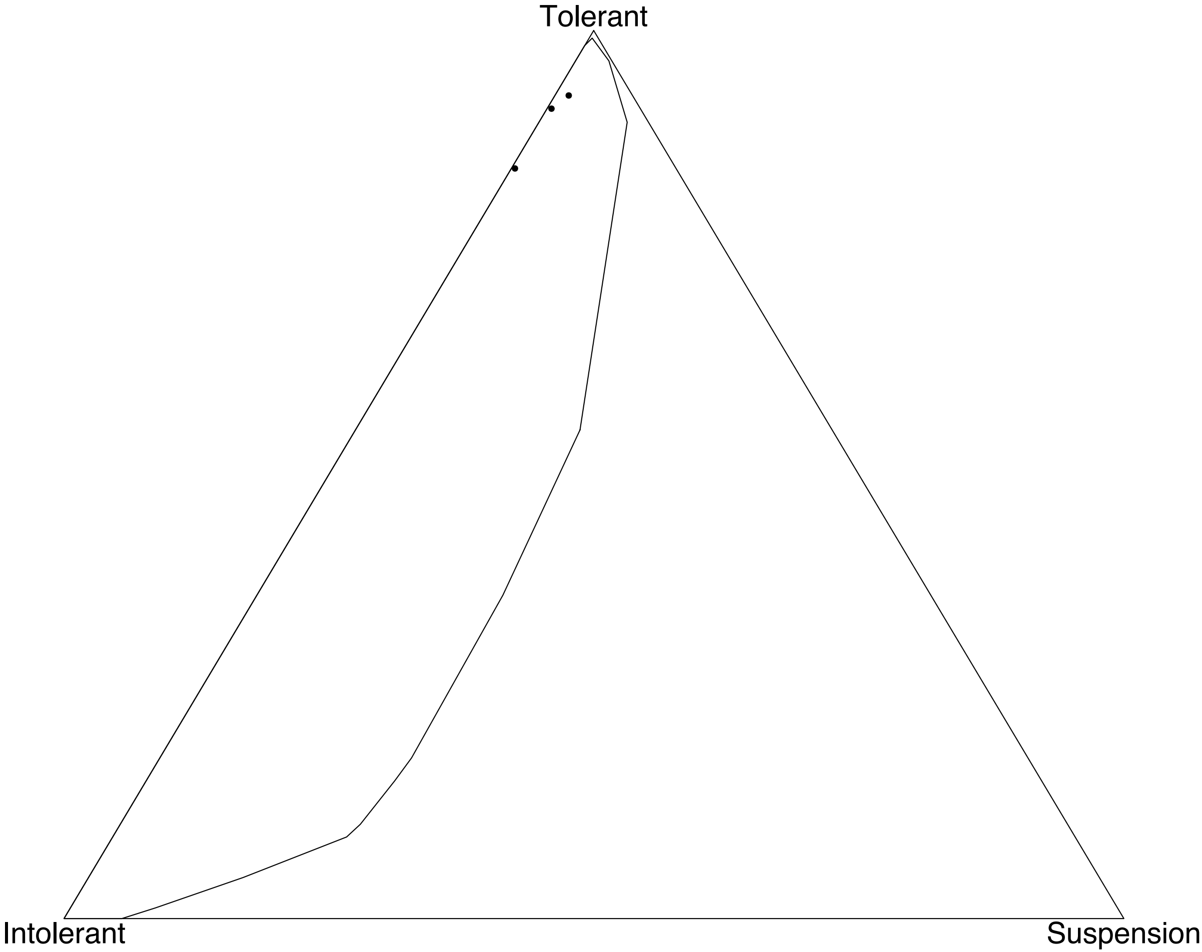

95% Credible Region for Salinity Regression Composition
Tolerant
(0.34,0.38,0.28).
Intolerant
Suspension


Spatial Dependence Parameter
Density
0.0
0.2
0.4
0.6
0.8
1.0
1.2
1.4
x
-1.0
-0.5
0.0
0.5
1.0

95% Prediction Regions for Hold-out Sub-Sample Compositions



95% Prediction Region Site 20



95% Prediction Region Site 23