Modelagem Bayesiana de Dados Composicionais Espaciais
H. TJELMELAND & K. V. LUND

Ana Beatriz Tozzo Martins - LEG/UFPR; DES/UEM
Paulo Justiniano Ribeiro Junior - LEG/UFPR

14 de dezembro de 2009

Resumo

1. **Dados composicionais** são vetores de proporções especificando frações de um todo;
2. Aitchison (1986) define **distribuições logísticas normais** p/ dados composicionais aplicando transformação logística e assumindo que os dados transformados são NM;
3. Generaliza a idéia p/ dados logísticos variando espacial/te e c/ isto define campos Gaussianos logísticos;
4. Considera o modelo numa estrutura Bayesiana e discute distribuições prioris apropriadas;
5. Considera **observações completas** e de **subcomposições** ou **porporções individuais** e discute as distribuições posterioris resultantes;
6. Em geral, a **posteriori não pode ser tratada analitica/te** $\Rightarrow$ definição de algoritmos MCMC
7. Analisa dados do Lago Ártico Aitchison (1986).

Introdução

- Em Pawlowsky-Glahn e Burger (1992) **interpolação espacial** de dado composicional é considerado via **transformação logística** em cada localização para obter dados em $\mathbb{R}^{D-1}$ seguido por **cokrigagem** $\Rightarrow$ modelo de campos Gaussianos correlacionados é adotado após transformação;
- O **artigo discute o problema de estimação** dos parâmetros incluindo o modelo numa estrutura Bayesiana hierárquica e considera observações de subcomposições e proporções individuais;
- Dados subcomposicionais se ajustam à família logística normal e permite tratamento analítico;
- Observações de proporções individuais necessitam MCMC.

Introdução (cont.)

- $x(u) = (x_1(u), ..., x_D(u))', \ u \in \mathbb{R}^k$

  $z(u) = (z_1(u), ..., z_d(u))',$

  Relação biunívoca:

$$z(u) = A_d x(u) \Longleftrightarrow x(u) = D_d z(u) + C_d$$

  $A_d \in \mathbb{R}^{d \times D}$ - matriz identidade $d \times d$ c/ coluna extra de zeros;

  $B_d \in \mathbb{R}^{D \times d}$ - matriz identidade $d \times d$ c/ uma linha extra de -1's.

  $C_d = (0, 0, ..., 0, 1)' \in \mathbb{R}^D$

Introdução (cont.)

- Transformação logística define relação biunívoca entre $\mathbb{S}^d$ e $\mathbb{R}^d$:

$$z(u) = \frac{exp\{y(u)\}}{1 + j_d' exp\{y(u)\}} \Longleftrightarrow y(u) = ln \frac{z(u)}{1 - j_d' z(u)}$$

$j_d \in \mathbb{R}^d$ - vetor coluna com eltos iguais a 1;
$y(u) = (y_1(u), ..., y_d(u))'$.

- Proporções das subcomposições e individual podem ser obtidas de $x(u)$ via matrizes seleção:
$S \in \mathbb{R}^{C \times D}$, $C \leq D$: eltos unitários onde cada 1 está localizado em cada linha e ao menos 1 em cada coluna. O restante é zero.

Existem relações para subcomposições;
Mat. permutação($C = D$): $x_p(u) = Px(u)$, e $y_p(u) \neq Py(u)$ a menos que $P = I_D$.

Campos Gaussianos Logísticos

- Como Aitchison (1986) define distribuições logísticas normais de distribuições multi normais, os autores definem **campos Gaussianos logísticos** de um processo Gaussiano multivariado.

- **Definição:** Campo Gaussiano logístico
  $x(u) = (x_1(u), ..., x_D(u))'$, $u \in \mathbb{R}^k$ ou

  $z(u) = (z_1(u), ..., z_d(u))'$, $u \in \mathbb{R}^k$

  com $\mu(u) = B_d' \mu_0(u)$ e $C(u, u') = B_d' C_0(u, u') B_d$
- **Campo Gaussiano logístico** pode ser interpretado como uma distribuição **priori** numa estrutura Bayesiana;
- Simulação não-condicional de campos Gaussiano logísticos é direta:

  Simule o processo Gaussiano multivariado $y(u)$;
  Use a transformação logística para cada localização.

Observações Completas e Subcomposicionais

- $x(u)$ Campo Gaussiano logístico com fç de parâmetros $\mu(\cdot)$ e $C(\cdot,\cdot)$;
- $u = (u_1, ..., u_n)' \Rightarrow$ para cada u_i existe matriz seleção associada $S_i \in \mathbb{R}^{C_i \times D}$ e um vetor $x_{S_i}(u_i)$

  Daí, $x(u)|x_{S_1}(u_1), ..., x_{S_n}(u_n)$ é Gaussiano logístico;

  $y(u)|y_{S_1}(u_1), ..., y_{S_n}(u_n)$ é processo Gaussiano multivariado $\Rightarrow$
  $x(u)|x_{S_1}(u_1), ..., x_{S_n}(u_n)$ campo Gaussiano logístico.
- Simulação condicional de $x(u)$ em um conj. de localizações é direta:

  Simule $y(u)|y_{S_1}(u_1), ..., y_{S_n}(u_n)$ de uma dist. Gaussiana;

  Use a transformação logística para cada localização.

Observações de Proporções individuais

- $x(u)$ Campo Gaussiano logístico com fç de parâmetros $\mu(\cdot)$ e $C(\cdot,\cdot)$;
- Observações de proporções individuais em $u_1, ..., u_n$;
- Para cada u_i existe matriz seleção associada $S_i \in \mathbb{R}^{C_i \times D}$, mas os valores observados são os vetores $S_1x(u_1), ..., S_nx(u_n)$;
- **Não existe transformação fácil** dos dados para valores correspondentes p/ $y(u)$ e a distribuição condicional resultante p/ $x(u)$ não é Gaussiana logística nem analitica/te tratável.
  $\Rightarrow$ propriedades do campo condicional deve ser obtido através de geração de realizações condicionais de $x(u)$ por**MCMC**.
- **Densidade Objetivo**

$$f(\tilde{z}_1, ..., \tilde{z}_n)(S_1x(u_1) = \tilde{x}_1, ..., S_nx(u_n) = \tilde{x}_n)$$

tem fortes correlações entre variáveis. A base Gaussiana do modelo sugere a construção de um algoritmo proposal independente.

Modelo Bayesiano Completo

Como $\mu(\cdot)$ e $C(\cdot,\cdot)$ são desconhecidos e não existe distribuição priori que permita tratamento analítico completo da posteriori **recorre-se** a MCMC para explorar a posteriori.

Distribuição Posteriori

- $\mu_\beta(u) = F(u)\beta$, $\beta \in \mathbb{R}^D$ - parâmetros desconhecidos; $F(u) \in \mathbb{R}^{d \times p}$ - fçs regressoras conhecidas.
- $\beta \sim N_p(\mu_0; \Sigma_0)$; parâmetros a serem especificados;
- **Função de covariância**: $C_{\theta,\psi}(u, u') = \alpha_\theta(u, u')\psi$ **exponencial** generalizada ou **Matérn**;
- Para obter modelo Bayesiano completa/te especificado é necessário especificar prioris p/ θ e ψ;
- $\psi^{-1} \sim W_d(q; Q) \rightarrow$ priori conjugada $\Rightarrow$ facilidade na construção do algoritmo MCMC;
  hiperparâmetros: $q > d$ - escalar; $Q_{d \times d}$ - definida positiva;
- Assumir alguma priori para $\pi(\theta)$.

Algoritmos de Simulação-Observações Completas

- Para $x(u)|\beta, \psi, \theta$ campo Gaussiano logístico, $u = (u_1, ..., u_n)'$, condicionar em x é **equivalente** a condicionar em y;
- **Distribuição posteriori**:

$$\pi(\beta, \psi, \theta|x(u_1), ..., x(u_n)) \propto \pi(\beta)\pi(\psi)\pi(\theta)f(Y|\beta, \psi, \theta),$$

onde f é densidade normal multivariada;

- **Algoritmo**:
  Atualizar β - Considerar condicional completa p/ $\pi(\beta|\psi, \theta, y)$ - NM - passo Gibbs;
  $[\psi^{-1}|\beta, \theta, y] \sim W \Rightarrow$ passo Gibbs;
  - P/ atualizar β, a proposal de ψ é também similar ao caso Gaussiano puro com condicionamento em y;
  - Diferente da situação para β e ψ, **Não Existe** proposal natural para θ e como é de baixa dimensão, propõe-se pequena mudança e MH dá convergência satisfatória.

Algoritmos de Simulação-Observações Subcomposicionais

- Para $x(u)|\beta, \psi, \theta$ campo Gaussiano logístico, $u = (u_1, ..., u_n)'$;
- Observações disponíveis são as subcomposições $x_{S_1}(u_1), ..., x_{S_n}(u_n)$;
- Condicionar em x é **equivalente** a condicionar em y;
- Introduzindo $y_S = (y_{S_1}(u_1)', ..., y_{S_n}(u_n)')'$ a posteriori é

$$\pi(\beta, \psi, \theta | x_{S_1}(u_1), ..., x_{S_n}(u_n)) \propto \pi(\beta)\pi(\psi)\pi(\theta) f(y_S|\beta, \psi, \theta);$$

- Neste caso, não é viável simular por MH $\Rightarrow \psi^{-1}$ não é Wishart ou outra distribuição tratável;
- Introduz-se y como variável latente e amostra-se de

$$\pi(\beta, \psi, \theta, y | x_{S_1}(u_1), ..., x_{S_n}(u_n)) = \pi(\beta, \psi, \theta, y | y_S);$$

e atualizar β, ψ, θ como antes e Gibbs para y.

Algoritmos de Simulação-Proporções Individuais

- $x(u)|\beta, \psi, \theta$ campo Gaussiano logístico, $u = (u_1, ..., u_n)'$, ;
- Considere proporções individuais observadas em algumas localizações;
- Sejam $S_1x(u_1), ..., S_nx(u_n)$ com $S_1, ..., S_n$ matrizes seleção;
- **Posteriori**:

  $\pi(\beta, \psi, \theta|S_1x(u_1), ..., S_nx(u_n)) \propto \pi(\beta)\pi(\psi)\pi(\theta)f(S_1x(u_1), ..., S_nx(u_n))$;
- Não existe algoritmo MH natural para esta densidade: não são tratáveis

  $\beta|\psi, \theta, S_1x(u_1), ..., S_nx(u_n)$
  $\psi|\beta, \theta, S_1x(u_1), ..., S_nx(u_n)$
- Introduz-se y como variável latente e amostra-se de

  $$\pi(\beta, \psi, \theta, y|S_1x(u_1), ..., S_nx(u_n))$$

  e atualizar β, ψ, θ como antes e Gibbs para y.

Exemplo

- Dados de um Lago no Arquipélago Ártico Canadense discutido em Aitchison (1986) sem considerar espacialização;
- **Observações** em $n = 39$ localizações;
- **Variáveis**: Profundidade da água, areia, silte e argila;
- $x(u) = (x_1(u_1), x_2(u_2), x_3(u_3))$
-

$$\mu_\beta(u) = F(u)\beta = \begin{bmatrix} 1 & 0 & ln(d(u)) & 0 \\ 0 & 1 & 0 & ln(d(u)) \end{bmatrix} = \begin{bmatrix} \beta_1 \\ \beta_2 \\ \beta_3 \\ \beta_4 \end{bmatrix}$$

d(u) - profundidade da água.
Cada elto em $y(u)$ bidimensional tem uma esperança consistindo de um termo cte e um proporcional a $ln(d(u))$.

Exemplo (cont.)

Função de correlação exponencial generalizada:

$$\alpha_\theta(u,u') = \begin{cases} 1 & \text{se} \quad u = u' \\ (1-\epsilon)\exp\left\{-\left(\frac{||u-u'||}{3R}\right)^{\nu}\right\} & \text{se} \quad u \neq u' \text{ e } \chi = 1 \\ 0 & \text{se} \quad u \neq u' \text{ e } \chi = 0 \end{cases}$$

- $||\cdot||$ - distância euclidiana;
- $R > 0$ - correlação espacial;
- $\nu \in [0,2]$ - parâmetro de forma da fç de correlação;
- $\epsilon \in [0,1]$ - efeito pepita;
- $\chi \in 0,1$- indicador de correlação espacial.

Exemplo (cont.)

- $\beta, \psi, \theta \rightarrow$ atribui prioris difusas mas próprias;
- $\beta \sim N\left(E(\beta) \rightarrow 0; \Sigma_\beta = \begin{bmatrix} 100^2 & 0 \\ 0 & 100^2 \end{bmatrix}\right)$;
- $\psi^{-1} \sim W\left(q = 4; Q = \frac{I_2}{p-3}\right)$ tq $E(\psi^{-1}) = I_2$

Comentários Finais

- O artigo define um modelo espacial p/ dado composicional e avalia o modelo dentro de um conjunto Bayesiano;
- Observações completas é a forma mais simples e define algorimtos MCMC eficientes p/ lidar com situações onde subcomposições ou proporções individuais são disponíveis;
- Algoritmos apresentam **boa**s razões de **convergência**;
- Assumir estrutura de covariância intrínsica p/o processo Gaussiano latente: **Motivos**:
  - Permite parametrização parcimoniosa da fç de covariância
  - Eficiência Computacional.
- A estrutura de covariância espacial é então especificada via matrix de covariância inversa esparsa e algoritmos especiais p/ matrizes esparsas podem ser usadas;

Comentários Finais (cont.)

- Para os eltos de θ, prioris independentes;
- $R \sim exp(15)$,
- $\nu \sim U[0,2]$, $\epsilon \sim U[0,1]$,
- $P(x=0) = P(x=1) = 1/2$
- Algoritmo MH atualizado 5 vezes p/ cada atualização de β e ψ - 1 atualização;
- β e $\psi \Rightarrow$ passo Gibbs
- **Conclusão**: MCMC pode atingir convergência rápido , boas propriedades de mistura
- Figura 4- densidades à posteriori estimadas para cada um dos 4 parâmetros.