Spatial Factor Models for Multivariate Spatial Data

Jeffrey Doser[1] & Andrew Finley[2]

May 15, 2023

[1]Department of Integrative Biology, Michigan State University.

[2]Department of Forestry, Michigan State University.

Multivariate spatial data

- Point-referenced spatial data often come as multivariate measurements at each location.

Multivariate spatial data

- Point-referenced spatial data often come as multivariate measurements at each location.
- Examples:
  - Environmental monitoring: stations yield measurements on ozone, NO, CO, and $PM_{2.5}$.

Multivariate spatial data

- Point-referenced spatial data often come as multivariate measurements at each location.
- Examples:
  - Environmental monitoring: stations yield measurements on ozone, NO, CO, and $PM_{2.5}$.
  - Community Ecology: assemblages/communities of species

Multivariate spatial data

- Point-referenced spatial data often come as multivariate measurements at each location.
- Examples:
  - Environmental monitoring: stations yield measurements on ozone, NO, CO, and $PM_{2.5}$.
  - Community Ecology: assemblages/communities of species
  - Forestry: measurements of stand characteristics age, total biomass, and average tree diameter.

Multivariate spatial data

- Point-referenced spatial data often come as multivariate measurements at each location.
- Examples:
  - Environmental monitoring: stations yield measurements on ozone, NO, CO, and $PM_{2.5}$.
  - Community Ecology: assemblages/communities of species
  - Forestry: measurements of stand characteristics age, total biomass, and average tree diameter.
  - Atmospheric modeling: at a given site we observe surface temperature, precipitation and wind speed

Multivariate spatial data

- Point-referenced spatial data often come as multivariate measurements at each location.
- Examples:
  - Environmental monitoring: stations yield measurements on ozone, NO, CO, and $PM_{2.5}$.
  - Community Ecology: assemblages/communities of species
  - Forestry: measurements of stand characteristics age, total biomass, and average tree diameter.
  - Atmospheric modeling: at a given site we observe surface temperature, precipitation and wind speed
- We anticipate dependence between measurements
  - at a particular location
  - across locations

Multivariate spatial generalized linear model

- Spatial generalized linear model for h-variate spatial data for $j = 1, 2, \dots, h$ and $i = 1, \dots, n$:

$$
\begin{aligned}
y_j(\mathbf{s}_i) &\sim f(\mu_j(\mathbf{s}_i), \tau_j) \\
\mu_j(\mathbf{s}_i) &= g^{-1}(\eta_j(\mathbf{s}_i)) = \mathbf{x}(\mathbf{s}_i)^\top \boldsymbol{\beta}_j + \mathrm{w}_j^*(\mathbf{s}_i)
\end{aligned}
$$

- We can imagine modeling $\mathbf{w}^*(\mathbf{s}_i) = (\mathrm{w}_1^*(\mathbf{s}_i), \mathrm{w}_2^*(\mathbf{s}_i), \dots, \mathrm{w}_h^*(\mathbf{s}_i))'$ as an h-variate Gaussian process

Multivariate spatial generalized linear model

- Spatial generalized linear model for h-variate spatial data for $j = 1, 2, \ldots, h$ and $i = 1, \ldots, n$:

$$
\begin{aligned}
y_j(\mathbf{s}_i) &\sim f(\mu_j(\mathbf{s}_i), \tau_j) \\
\mu_j(\mathbf{s}_i) &= g^{-1}(\eta_j(\mathbf{s}_i)) = \mathbf{x}(\mathbf{s}_i)^\top \boldsymbol{\beta}_j + \mathrm{w}_j^*(\mathbf{s}_i)
\end{aligned}
$$

- We can imagine modeling $\mathbf{w}^*(\mathbf{s}_i) = (\mathrm{w}_1^*(\mathbf{s}_i), \mathrm{w}_2^*(\mathbf{s}_i), \ldots, \mathrm{w}_h^*(\mathbf{s}_i))'$ as an h-variate Gaussian process
- Could model using Multivariate NNGP as disussed previously with SVCs, works well when $h < 5$.

Multivariate spatial generalized linear model

- Spatial generalized linear model for h-variate spatial data for $j = 1, 2, \ldots, h$ and $i = 1, \ldots, n$:

$$
\begin{aligned}
y_j(\mathbf{s}_i) &\sim f(\mu_j(\mathbf{s}_i), \tau_j) \\
\mu_j(\mathbf{s}_i) &= g^{-1}(\eta_j(\mathbf{s}_i)) = \mathbf{x}(\mathbf{s}_i)^\top \boldsymbol{\beta}_j + \mathrm{w}_j^*(\mathbf{s}_i)
\end{aligned}
$$

- We can imagine modeling $\mathbf{w}^*(\mathbf{s}_i) = (\mathrm{w}_1^*(\mathbf{s}_i), \mathrm{w}_2^*(\mathbf{s}_i), \ldots, \mathrm{w}_h^*(\mathbf{s}_i))'$ as an h-variate Gaussian process
- Could model using Multivariate NNGP as disussed previously with SVCs, works well when $h < 5$.
- But what about when h is large (e.g,. 10, 100)?

Spatial Factor Model

- Approximates the dependence between multivariate (spatially-dependent) outcomes through a linear combination of a (much) lower-dimensional set of spatial factors

Spatial Factor Model

- Approximates the dependence between multivariate (spatially-dependent) outcomes through a linear combination of a (much) lower-dimensional set of spatial factors
- We represent the $h \times 1$ vector $\mathbf{w}^*(\mathbf{s}_i)$ as a linear combination of latent spatial factors and factor loadings:

$$\mathbf{w}^*(\mathbf{s}_i) = \boldsymbol{\Lambda}\mathbf{w}(\mathbf{s}_i)$$

- $\boldsymbol{\Lambda}$ is an $h \times q$ loadings matrix (tall and skinny) and $\mathbf{w}(\mathbf{s}_i)$ is a $q \times 1$ vector of realizations from q *independent* spatial GPs

Spatial Factor Model

- Approximates the dependence between multivariate (spatially-dependent) outcomes through a linear combination of a (much) lower-dimensional set of spatial factors
- We represent the $h \times 1$ vector $\mathbf{w}^*(\mathbf{s}_i)$ as a linear combination of latent spatial factors and factor loadings:

$$\mathbf{w}^*(\mathbf{s}_i) = \boldsymbol{\Lambda}\mathbf{w}(\mathbf{s}_i)$$

- $\boldsymbol{\Lambda}$ is an $h \times q$ loadings matrix (tall and skinny) and $\mathbf{w}(\mathbf{s}_i)$ is a $q \times 1$ vector of realizations from q *independent* spatial GPs
- In traditional factor analysis, $\mathbf{w}(\mathbf{s}_i)$ are realizations from independent standard normal random variables.

Spatial Factor Model

- Choosing $q << m$ leads to substantial computational reductions.
- Simple to code: just sample from q independent GPs as with basic univariate models.
- Yields a non-separable multivariate cross-covariance function given by $\sum_{k=1}^{q} \mathbf{R}_k(\phi_k)\boldsymbol{\lambda}_k\boldsymbol{\lambda}_k^\top$
- Can simply replace the q full GPs with their corresponding NNGPs to yield a spatial factor NNGP model
- Identifiability constraints on $\boldsymbol{\Lambda}$: fix upper triangle to 0 and diagonal to 1. See Ren and Banerjee (2013) *Biometrics*

Priors

- Standard normal priors for the lower triangle of $\boldsymbol{\Lambda}$
- We like to model response-specific regression coefficients $\boldsymbol{\beta}_j$ hierarchically. For each $r = 1, \ldots, p$ covariate, we model $\beta_{j,r}$ following

$$\beta_{j,r} \sim N(\mu_{\beta_r}, \tau^2_{\beta_r})$$

- Gaussian hyperpriors for μ_{β_r} and IG or half-Cauchy priors for $\tau^2_{\beta_r}$
- Independent uniform priors for spatial decay parameters ϕ

Gibbs sampler

- Full conditionals are in closed form for all parameters except ϕ.
- Update ϕ with an Adaptive Metroplis-within-Gibbs algorithm (Roberts and Rosenthal 2009)
- See Taylor-Rodriguez et al. 2019 for Gaussian sampler, spOccupancy website for Pólya-Gamma sampler

Why we like spatial factor models

- Simple to code (don't need to deal with cross-covariance matrices).

Why we like spatial factor models

- Simple to code (don't need to deal with cross-covariance matrices).
- Relatively fast and efficient (well, at least for Gaussian and Binomial).

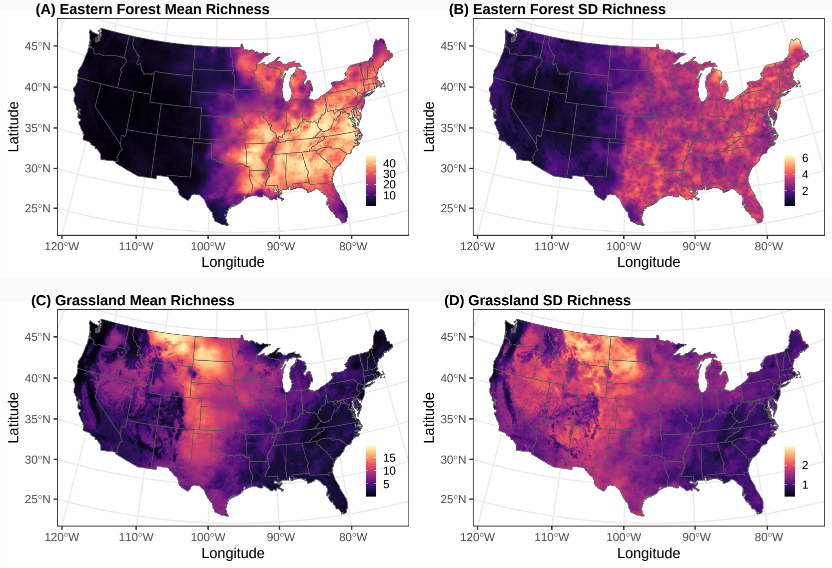
Why we like spatial factor models

- Simple to code (don't need to deal with cross-covariance matrices).
- Relatively fast and efficient (well, at least for Gaussian and Binomial).
- Factors and factor loadings can be used for model-based ordination.

Why we like spatial factor models

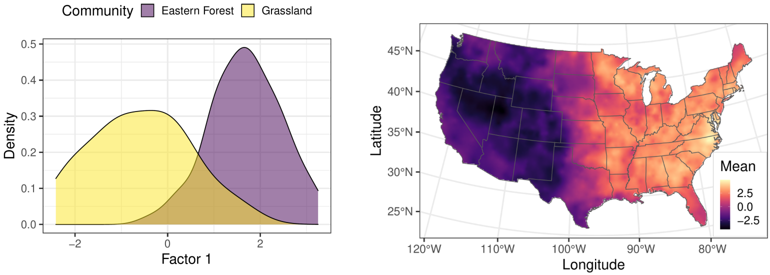
- Simple to code (don't need to deal with cross-covariance matrices).
- Relatively fast and efficient (well, at least for Gaussian and Binomial).
- Factors and factor loadings can be used for model-based ordination.
- Straightforward extensions to spatially-varying coefficient models.

Example: bird communities across the continental US



Example: bird communities across the continental US

Visualization of the first spatial factor and corresponding factor loadings



Some downsides to spatial factor models

- Convergence assessment is not always straightforward
- Sensitivity to initial values
- Order of the first q species has important implications for convergence and mixing.
- Assume a multivariate stochastic process can be represented as a linear combination of independent univariate processes

Software

- `spOccupancy`: spatial NNGP and non-spatial factor models for binary data
- `spAbundance`: Gaussian, Poisson, and NB spatial NNGP and non-spatial factor models.
- `boral`: many distributions for non-spatial and spatial factor models (Hui 2015 *MEE*; spatial use full GPs fit in JAGS)
- `Hmsc`: spatial models using NNGPs (Tikhonov et al. 2019; *MEE*)
- `spBFA`: a variety of spatial models with some nifty priors (Berchuck et al. 2022 *Bayesian Analysis*)

Exercise

Modeling the distribution of 10 tree species across Vermont