Multivariate spatial analysis

Applied Spatial Statistics

Multivariate spatial data

- Say Y_{ik} is the observed value of response type k at location $\mathbf{s}_i$
- Example: Y_{i1} is the temperature in Raleigh ($\mathbf{s}_i$)
- Example: Y_{i2} is the humidity in Raleigh ($\mathbf{s}_i$)
- This is an example of a multivariate spatial process
- There can be K response types, and they can be measured at different locations

Objectives

- Test for cross-correlation between responses

- Exploit cross-correlation to improve prediction

General model

- As for univariate spatial data, we decompose the data into a mean, correlated residuals and uncorrelated residuals

$$Y_{ik} = \mu_{ik} + Z_{ik} + \varepsilon_{ik}$$

- The mean is $\mu_{ik} = \beta_{0k} + \sum_{j=1}^{p} X_{ik}\beta_{jk}$
- Z_{ik} is correlated across space and responses
- The nugget is $\varepsilon_{ik} \sim \text{Normal}(0, \tau_k^2)$, independent over i and k

Types of dependence

- The responses each have spatial covariance, $\text{Cov}(Y_{ik}, Y_{jk}) = \sigma_k^2 \rho_k(d_{ij})$
- There is also cross-covariance $\text{Cov}(Y_{ik}, Y_{ij}) = \sigma_{jk}$
- The cross-correlation can be positive or negative, but the $K \times K$ matrix of σ_{jk} must be a valid covariance matrix
- Different response types at different locations can be correlated
- The correlation is a user-defined function of σ_{jk}, ρ_j and ρ_k

Exploratory analysis

- First fit least squares regression to remove the mean trend for each response type
- Using the residuals, plot the semivariogram at each response type to select a spatial correlation model
- Use a cross-variogram to explore cross dependence

Cross-variogram

- The variogram for response type k is

$$2\gamma_k(d_{ij}) = \mathrm{E}(Y_{ik} - Y_{jk})^2$$

- The cross-variogram (assuming constant mean) for response types k and l is

$$2\gamma_{kl}(d_{ij}) = \mathrm{E}(Y_{ik} - Y_{jk})(Y_{il} - Y_{jl})$$

- If both process are strongly spatially correlated (e.g., no nugget) then $\gamma_{kl}(0) \approx 0$
- For d_{ij} larger than the range of either process,

$$\gamma_{kl}(d_{ij}) = \mathrm{Cov}(Y_{ik}, Y_{il}) = \sigma_{kl}$$

- So the height of the plateau is the cross-covariance

Separable model

- The separable model assumes that all K response types have the same spatial correlation function, $\rho(d)$

- In this case, the covariance separates as

$$\text{Cov}(Y_{ik}, Y_{jl}) = \sigma_{kl} \cdot \rho(d_{ij})$$

- Separability dramatically simplifies the analysis, but is often unrealistic

Linear model of coregionalization (LMC)

- Example: Y_{ik} is air pollution of type k
- The latent (unobserved) factors F_{i1} and F_{i2} are emissions from cars and power plants
- The loadings L_{k1} and L_{k2} determine how much each type of emission contributes to pollutant k
- Pollutants with common sources are correlated

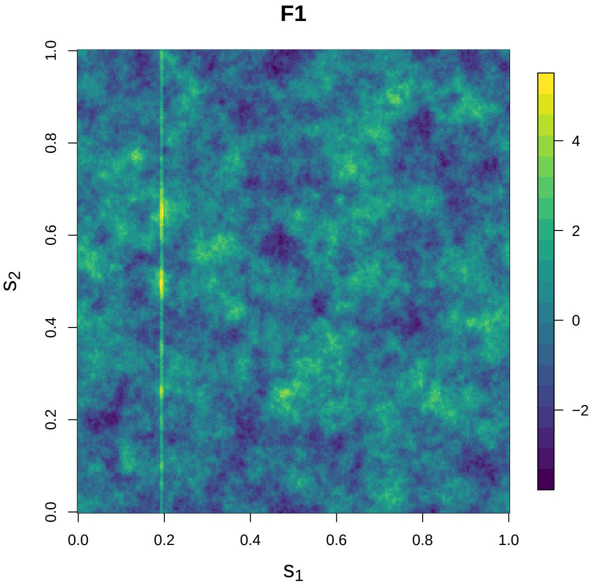
Linear model of coregionalization (LMC)

- The example to follow has two latent factors: F_1 is road emissions and F_2 is power plant emissions
- There are $K = 3$ pollutants
- The loading matrix is

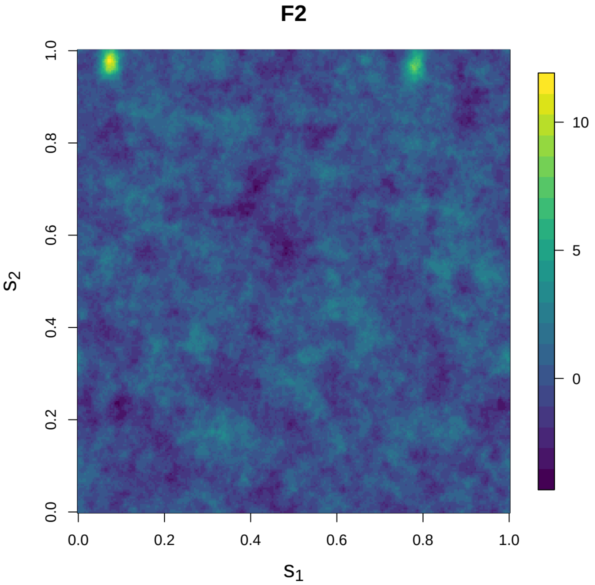
$$L = \begin{bmatrix} 5 & 5 \\ 5 & 1 \\ 1 & 5 \end{bmatrix}$$

- Pollutant 1 is an equal mix; pollutant 2 is mostly road; pollutant 3 is mostly power plants

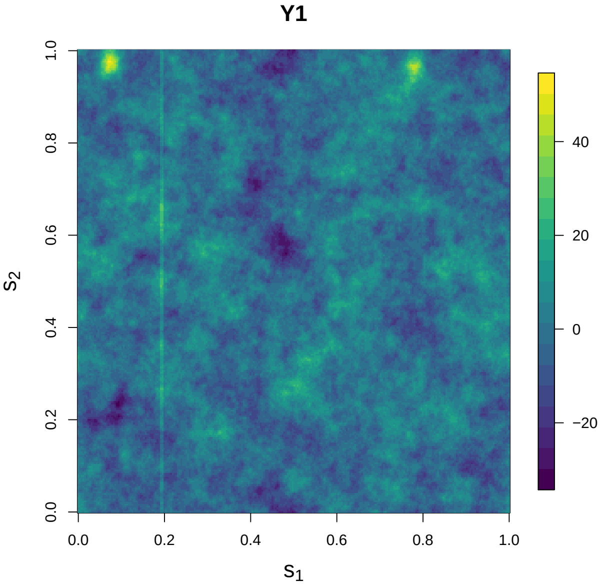
LMC - latent factor 1



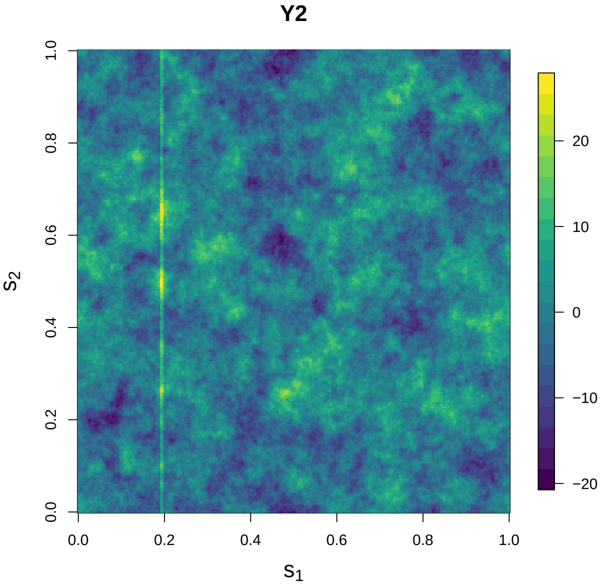
LMC - latent factor 2



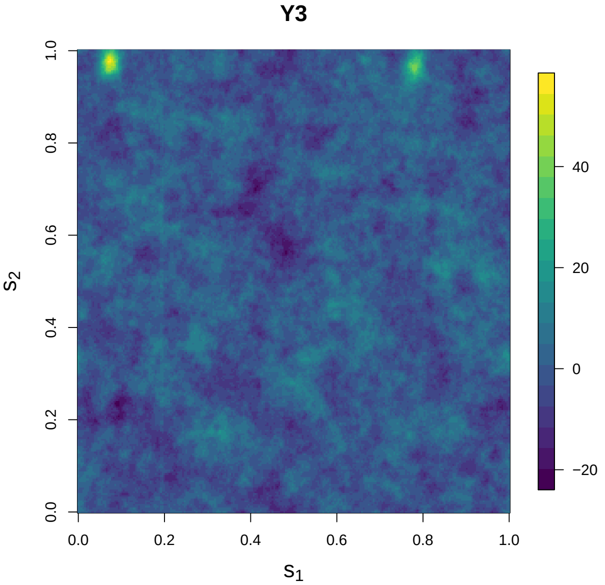
LMC - response $Y_1 = 5F_1 + 5F_2$



LMC - response $Y_2 = 5F_1 + 1F_2$



LMC - response $Y_3 = 1F_1 + 5F_2$



Linear model of coregionalization (LMC)

- It is basically factor analysis for spatial data
- Let $F_{i1}, ..., F_{iK}$ be independent spatial processes with spatial correlation functions ρ_1,...,ρ_K
- The response is modeled as

$$Y_{ik} = \sum_{u=1}^{K} L_{ku} F_{iu}$$

- The (non-separable) cross-covariance is

$$\text{Cov}(Y_{ik}, Y_{jl}) = \sum_{u=1}^{L} L_{ku} L_{lu} \rho_u(d_{ij})$$

Linear model of coregionalization (LMC)

- Say the loading matrix is

$$L = \begin{bmatrix} 5 & 5 \\ 5 & 1 \\ 1 & 5 \end{bmatrix}$$

- The cross-covariance is

$$\text{Cov}(Y_{i1}, ..., Y_{iK}) = LL^T = \begin{bmatrix} 50 & 30 & 30 \\ 30 & 26 & 10 \\ 30 & 10 & 26 \end{bmatrix}$$

- The cross-correlation is

$$\text{Cor}(Y_{i1}, ..., Y_{iK}) = \begin{bmatrix} 1.00 & 0.83 & 0.83 \\ 0.83 & 1.00 & 0.38 \\ 0.83 & 0.38 & 1.00 \end{bmatrix}$$

Co-Kriging

- As with spatiotemoral data, the Kriging equations apply for multivariate spatial data
- This requires estimating all parameters in the spatial correlation functions and the cross-correlation function
- Kriging with multiple responses is called cokriging

Software options

- The `spBayes` function `spMvLM` fits a separable model using MCMC

- The package `gstat` estimates parameters in the LMC using variograms