

Hierarchical models for spatial data

Based on the book by Banerjee, Carlin and Gelfand *Hierarchical Modeling and Analysis for Spatial Data*, 2004. We focus on Chapters 1, 2 and 5.

- Geo-referenced data arise in agriculture, climatology, economics, epidemiology, transportation and many other areas.
- What does geo-referenced mean? In a nutshell, we know the geographic location at which an observation was collected.
- Why does it matter? Sometimes, relative location can provide information about an outcome beyond that provided by covariates.

Models for spatial data (cont'd)

- Example: infant mortality is typically higher in high poverty areas. Even after incorporating poverty as a covariate, residuals may still be spatially correlated due to other factors such as nearness to pollution sites, distance to pre and post-natal care centers, etc.
- Often data are also collected over time, so models that include spatial and temporal correlations of outcomes are needed.
- We focus on spatial, rather than spatio-temporal models.
- We also focus on models for univariate rather than multivariate outcomes.

Types of spatial data

- Point-referenced data: $Y(s)$ a random outcome (perhaps vector-valued) at location s , where s varies continuously over some region D . The location s is typically two-dimensional (latitude and longitude) but may also include altitude. Known as geostatistical data.
- Areal data: outcome Y_i is an aggregate value over an areal unit with well-defined boundaries. Here, D is divided into a finite collection of areal units. Known as lattice data even though lattices can be irregular.
- Point-pattern data: Outcome $Y(s)$ is the occurrence or not of an event and locations s are random. Example: locations of trees of a species in a forest or addresses of persons with a particular disease. Interest is often in deciding whether points occur independently in space or whether there is clustering.

Types of spatial data (cont'd)

- Marked point process data: If covariate information is available we talk about a marked point process. Covariate value at each site marks the site as belonging to a certain covariate batch or group.
- Combinations: e.g. ozone daily levels collected in monitoring stations for which we know the precise location, and number of children in a zip code reporting to the ER with respiratory problems on that day. Require data re-alignment so that outcomes and covariates obtained at different spatial resolutions can be combined in a model.

Models for point-level data

The basics

- Location index s varies continuously over region D .
- We often assume that the covariance between two observations at locations s_i and s_j depends only on the distance d_{ij} between the points.
- The spatial covariance is often modeled as exponential:

$$\text{Cov}(Y(s_i), Y(s_{i'})) = C(d_{ii'}) = \sigma^2 e^{-\phi d_{ii'}},$$

where $(\sigma^2, \phi) > 0$ are the partial sill and decay parameters, respectively.

- Covariogram: a plot of $C(d_{ii'})$ against $d_{ii'}$.
- For $i = i'$, $d_{ii'} = 0$ and $C(d_{ii'}) = \text{var}(Y(s_i))$.
- Sometimes, $\text{var}(Y(s_i)) = \tau^2 + \sigma^2$, for τ^2 the nugget effect and $\tau^2 + \sigma^2$ the sill.

Models for point-level data (cont'd)

Covariance structure

- Suppose that outcomes are normally distributed and that we choose an exponential model for the covariance matrix. Then:

$$Y|\mu, \theta \sim N(\mu, \Sigma(\theta)),$$

with

$$\begin{aligned} Y &= \{Y(s_1), Y(s_2), \dots, Y(s_n)\} \\ \Sigma(\theta)_{ii'} &= \text{cov}(Y(s_i), Y(s_{i'})) \\ \theta &= (\tau^2, \sigma^2, \phi). \end{aligned}$$

- Then

$$\Sigma(\theta)_{ii'} = \sigma^2 \exp(-\phi d_{ii'}) + \tau^2 I_{i=i'},$$

with $(\tau^2, \sigma^2, \phi) > 0$.

- This is an example of an *isotropic* covariance function: the spatial correlation is only a function of d .

Models for point-level data, details

- Basic model:

$$Y(s) = \mu(s) + w(s) + e(s),$$

where $\mu(s) = x'(s)\beta$ and the residual is divided into two components:

$w(s)$ is a realization of a zero-centered stationary Gaussian process and $e(s)$ is uncorrelated pure error.

- The $w(s)$ are functions of the partial sill σ^2 and decay ϕ parameters.
- The $e(s)$ introduces the nugget effect τ^2 .
- τ^2 interpreted as pure sampling variability or as *microscale* variability, i.e., spatial variability at distances smaller than the distance between two outcomes: the $e(s)$ are sometimes viewed as spatial processes with rapid decay.

The variogram and semivariogram

- A spatial process is said to be:
 - Strictly stationary if distributions of $Y(s)$ and $Y(s+h)$ are equal, for h the distance.
 - Weakly stationary if $\mu(s) = \mu$ and $Cov(Y(s), Y(s+h)) = C(h)$.
 - Intrinsically stationary if

$$\begin{aligned} E[Y(s+h) - Y(s)] &= 0, \text{ and} \\ E[Y(s+h) - Y(s)]^2 &= Var[Y(s+h) - Y(s)] \\ &= 2\gamma(h), \end{aligned}$$

defined for *differences* and depending only on distance.

- $2\gamma(h)$ is the variogram and $\gamma(h)$ is the *semivariogram*
- The specific form of the semivariogram will depend on the assumptions of the model.

Stationarity

- Strict stationarity implies weak stationarity but the converse is not true except in Gaussian processes.
- Weak stationarity implies intrinsic stationarity, but the converse is not true in general.
- Notice that intrinsic stationarity is defined on the differences between outcomes at two locations and thus says nothing about the joint distribution of outcomes.

Semivariogram (cont'd)

- If $\gamma(h)$ depends on h only through its length $\|h\|$, then the spatial process is *isotropic*. Else it is *anisotropic*.
- There are many choices for isotropic models. The *exponential* model is popular and has good properties. For $t = \|h\|$:

$$\begin{aligned}\gamma(t) &= \tau^2 + \sigma^2(1 - \exp(-\phi t)) \text{ if } t > 0, \\ &= 0 \text{ otherwise.}\end{aligned}$$

- See figures, page 24.
- The *powered* exponential model has an extra parameter for smoothness:

$$\gamma(t) = \tau^2 + \sigma^2(1 - \exp(-\phi t^\kappa)) \text{ if } t > 0$$

- Another popular choice is the Gaussian variogram model, equal to the exponential except for the exponent term, that is $\exp(-\phi^2 t^2)$.

Semivariogram (cont'd)

- Fitting of the variogram has been traditionally done "by eye":
 - Plot an empirical estimate of the variogram akin to the sample variance estimate or the autocorrelation function in time series
 - Choose a theoretical functional form to fit the empirical γ
 - Choose values for (τ^2, σ^2, ϕ) that fit the data well.
- If a distribution for the outcomes is assumed and a functional form for the variogram is chosen, parameter estimates can be estimated via some likelihood-based method.
- Of course, we can also be Bayesians.

Point-level data (cont'd)

- For point-referenced data, frequentists focus on spatial prediction using *kriging*.
- Problem: given observations $\{Y(s_1), \dots, Y(s_n)\}$, how do we predict $Y(s_o)$ at a new site s_o ?
- Consider the model

$$Y = X\beta + \epsilon, \text{ where } \epsilon \sim N(0, \Sigma),$$

and where

$$\Sigma = \sigma^2 H(\phi) + \tau^2 I.$$

Here, $H(\phi)_{ii'} = \rho(\phi, d_{ii'})$.

- Kriging consists in finding a function $f(y)$ of the observations that minimizes the MSE of prediction

$$Q = E[(Y(s_o) - f(y))^2 | y].$$

Classical kriging (cont'd)

- (Not a surprising!) Result: $f(y)$ that minimizes Q is the conditional mean of $Y(s_0)$ given observations y (see pages 50-52 for proof):

$$E[Y(s_o)|y] = x'_o \hat{\beta} + \hat{\gamma}' \hat{\Sigma}^{-1} (y - X \hat{\beta})$$
$$Var[Y(s_o)|y] = \hat{\sigma}^2 + \hat{\tau}^2 - \hat{\gamma}' \hat{\Sigma}^{-1} \hat{\gamma},$$

where

$$\hat{\gamma} = (\hat{\sigma}^2 \rho(\hat{\phi}, d_{o1}), \dots, \hat{\sigma}^2 \rho(\hat{\phi}, d_{on}))$$
$$\hat{\beta} = (X' \hat{\Sigma}^{-1} X)^{-1} X' \hat{\Sigma}^{-1} y$$
$$\hat{\Sigma} = \hat{\sigma}^2 H(\hat{\phi}).$$

- Solution assumes that we have observed the covariates x_o at the new site.
- If not, in the classical framework $Y(s_o), x_o$ are jointly estimated using an EM-type iterative algorithm.

Bayesian methods for estimation

- The Gaussian isotropic kriging model is just a general linear model similar to those in Chapter 15 of textbook.
- Just need to define the appropriate covariance structure.
- For an exponential covariance structure with a nugget effect, parameters to be estimated are $\theta = (\beta, \sigma^2, \tau^2, \phi)$.
- Steps:
 - Choose priors and define sampling distribution
 - Obtain posterior for all parameters $p(\theta|y)$
 - Bayesian kriging: get posterior predictive distribution for outcome at new location $p(y_o|y, X, x_o)$.

Bayesian methods (cont'd)

- Sampling distribution (marginal data model)

$$y|\theta \sim N(X\beta, \sigma^2 H(\phi) + \tau^2 I)$$

- Priors: typically chosen so that parameters are independent a priori.
- As in the linear model:
 - Non-informative prior for β is uniform or can use a normal prior too.
 - Conjugate priors for variances σ^2, τ^2 are inverse gamma priors.
- For ϕ , appropriate prior depends on covariance model. For simple exponential where

$$\rho(s_i - s_j; \phi) = \exp(-\phi \|s_i - s_j\|),$$

a Gamma prior can be a good choice.

- Be cautious with improper priors for anything but β .

Hierarchical representation of model

- Hierarchical model representation: first condition on the spatial random effects $W = \{w(s_1), \dots, w(s_n)\}$:

$$\begin{aligned} y|\theta, W &\sim N(X\beta + W, \tau^2 I) \\ W|\phi, \sigma^2 &\sim N(0, \sigma^2 H(\phi)). \end{aligned}$$

- Model specification is then completed by choosing priors for β, τ^2 and for ϕ, σ^2 (hyperparameters).
- Note that hierarchical model has n more parameters (the $w(s_i)$) than the marginal model.
- Computation with the marginal model preferable because $\sigma^2 H(\phi) + \tau^2 I$ tends to be better behaved than $\sigma^2 H(\phi)$ at small distances.

Estimation of spatial surface $W|y$

- Interest is sometimes on estimating the spatial surface using $p(W|y)$.
- If marginal model is fitted, we can still get marginal posterior for W as

$$p(W|y) = \int p(W|\sigma^2, \phi)p(\sigma^2, \phi|y)d\sigma^2d\phi.$$

- Given draws $(\sigma^{2(g)}, \phi^{(g)})$ from the Gibbs sampler on the marginal model, we can generate W from

$$p(W|\sigma^{2(g)}, \phi^{(g)}) = N(0, \sigma^{2(g)}H(\phi^{(g)})).$$

- Analytical marginalization over W is possible only if model has Gaussian form.

Bayesian kriging

- Let $Y_o = Y(s_o)$ and $x_o = x(s_o)$. Kriging is accomplished by obtaining the posterior predictive distribution

$$\begin{aligned} p(y_o|x_o, X, y) &= \int p(y_o, \theta|y, X, x_o)d\theta \\ &= \int p(y_o|\theta, y, x_o)p(\theta|y, X)d\theta. \end{aligned}$$

- Since (Y_o, Y) are jointly multivariate normal (see expressions 2.18 and 2.19 on page 51), then $p(y_o|\theta, y, x_o)$ is a conditional normal distribution.

Bayesian kriging (cont'd)

- Given MCMC draws of the parameters $(\theta^{(1)}, \dots, \theta^{(G)})$ from the posterior distribution $p(\theta|y, X)$, we draw values $y_o^{(g)}$ for each $\theta^{(g)}$ as

$$y_o^{(g)} \sim p(y_o|\theta^{(g)}, y, x_o).$$

- Draws $\{y_o^{(1)}, y_o^{(2)}, \dots, y_o^{(G)}\}$ are a sample from the posterior predictive distribution of the outcome at the new location s_o .
- To predict Y at a set of m new locations s_{o1}, \dots, s_{om} , it is best to do joint prediction to be able to estimate the posterior association among m predictions.
- Beware of joint prediction at many new locations with WinBUGS. It can take forever.

Kriging example from WinBugs

- Data were first published by Davis (1973) and consist of heights at 52 locations in a 310-foot square area.
- We have 52 $s = (x, y)$ coordinates and outcomes (heights).
- Unit of distance is 50 feet and unit of elevation is 10 feet.
- The model is

$$\text{height} = \beta + \epsilon, \text{ where } \epsilon \sim N(0, \Sigma),$$

and where

$$\Sigma = \sigma^2 H(\phi).$$

- Here, $H(\phi)_{ij} = \rho(s_i - s_j; \phi) = \exp(-\phi \|s_i - s_j\|^\kappa)$.
- Priors on (β, ϕ, κ) .
- We predict elevations at 225 new locations.