Geostatistical estimation - Part II

Applied Spatial Statistics

Estimation strategies

- We now have several possible models for spatial processes
- In this lecture we discuss methods for fitting models to data
- One task is model selection:
  - Which covariates to include in $\mathbf{X}$?
  - Exponential or Matern correlation?
  - Should we include a nugget?
  - Is the covariance stationary?
- Another is parameter estimation:
  - Mean parameters $\boldsymbol{\beta} = (\beta_0, \beta_1, ..., \beta_p)$
  - Covariance parameters $\boldsymbol{\theta} = (\tau^2, \sigma^2, \phi, \nu)$

Maximum Likelihood Estimation (MLE)

- Variograms are fast and simple exploratory analysis tools
- Variograms can be used for parameter estimation
- MLE gives more precise parameter estimates
- MLE is also better for formally testing hypotheses are quantifying uncertainty
- MLE is slow for large datasets

MLE - Overview

- The likelihood function is the probability (density) of the data given the parameters

- For example, if $Y_1, ..., Y_n \sim \text{Normal}(\mu, \sigma^2)$ then the likelihood function is

$$L(\boldsymbol{\theta}) = \prod_{i=1}^{n} \frac{1}{\sigma\sqrt{2\pi}} \exp\left\{-\frac{(Y_i - \mu)^2}{2\sigma^2}\right\}$$

  for parameters $\boldsymbol{\theta} = (\mu, \sigma)$.

- The MLE is the value of $\boldsymbol{\theta}$ that maximizes this function

- This value "agrees with the data the most"

Review of the spatial model

- Recall Y_i is the observation at location $\mathbf{s}_i$
- The mean is $\mu_i(\boldsymbol{\beta}) = \mathrm{E}(Y_i) = \beta_0 + \sum_{j=1}^{p} X_{ij}\beta_j$
- The variance is $\Sigma_{ii}(\boldsymbol{\theta}) = \mathrm{V}(Y_i) = \sigma^2 + \tau^2$
- The isotropic exponential covariance is

$$\Sigma_{ij}(\boldsymbol{\theta}) = \mathrm{Cov}(Y_i, Y_j) = \sigma^2 \exp(-d_{ij}/\phi)$$

- The parameters are $\boldsymbol{\beta} = (\beta_0, ..., \beta_p)$ and $\boldsymbol{\theta} = (\sigma^2, \tau^2, \phi)$

Review of the spatial model

- As with linear regression, expressing this model in matrices cleans up notation
- The $n \times 1$ mean vector is

$$\mu(\boldsymbol{\beta}) = \begin{pmatrix} \mu_1(\boldsymbol{\beta}) \\ \vdots \\ \mu_n(\boldsymbol{\beta}) \end{pmatrix}$$

- The $n \times n$ covariance matrix is

$$\Sigma(\boldsymbol{\theta}) = \begin{pmatrix} \Sigma_{11}(\boldsymbol{\theta}) & \Sigma_{12}(\boldsymbol{\theta}) & \dots & \Sigma_{1n}(\boldsymbol{\theta}) \\ \Sigma_{21}(\boldsymbol{\theta}) & \Sigma_{22}(\boldsymbol{\theta}) & \dots & \Sigma_{2n}(\boldsymbol{\theta}) \\ \vdots & \vdots & \vdots & \vdots \\ \Sigma_{n1}(\boldsymbol{\theta}) & \Sigma_{n2}(\boldsymbol{\theta}) & \dots & \Sigma_{nn}(\boldsymbol{\theta}) \end{pmatrix}$$

Review of the spatial model

- Say $n = 3$ with $s_1 = (0, 0)$, $s_2 = (1, 0)$ and $s_3 = (2, 0)$
- Further, $p = 1$ and $X_1 = 2$, $X_2 = 4$ and $X_3 = 6$
- The 3×1 mean vector is

$$\mu(\boldsymbol{\beta}) = \begin{pmatrix} \beta_0 + 2\beta_1 \\ \beta_0 + 4\beta_1 \\ \beta_0 + 6\beta_1 \end{pmatrix}$$

- The 3×3 covariance matrix is

$$\Sigma(\boldsymbol{\theta}) = \begin{pmatrix} \sigma^2 + \tau^2 & \sigma^2 \exp(-1/\rho) & \sigma^2 \exp(-2/\rho) \\ \sigma^2 \exp(-1/\rho) & \sigma^2 + \tau^2 & \sigma^2 \exp(-1/\rho) \\ \sigma^2 \exp(-2/\rho) & \sigma^2 \exp(-1/\rho) & \sigma^2 + \tau^2 \end{pmatrix}$$

The multivariate normal distribution

- If $\mathbf{Y} = (Y_1, ..., Y_n)^T$ is jointly normal, then it follows the multivariate normal (MVN) distribution

- The MVN density function is the likelihood function

$$L(\boldsymbol{\beta}, \boldsymbol{\theta}) \propto |\Sigma(\boldsymbol{\theta})|^{-1/2} \exp\left[-\frac{1}{2}\{\mathbf{Y} - \mu(\boldsymbol{\beta})\}^T \Sigma(\boldsymbol{\theta})^{-1}\{\mathbf{Y} - \mu(\boldsymbol{\beta})\}\right]$$

- This uses the determinent (left) and inverse (right) of $\Sigma(\boldsymbol{\theta})$

- If $\sigma = 0$ and thus the observations are uncorrelated, this reduces to the product of univariate normal densities

Generalized least squares

- If θ is known, the MLE for β minimizes the generalized least squares

$$(\mathbf{Y} - \mathbf{X}\beta)^T \Sigma(\theta)^{-1} (\mathbf{Y} - \mathbf{X}\beta)$$

- The solution is

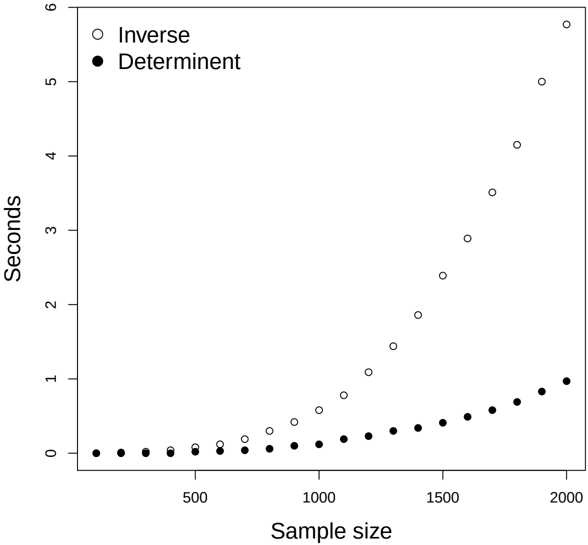
$$\hat{\beta} = \{\mathbf{X}^T \Sigma(\theta)^{-1} \mathbf{X}\}^{-1} \mathbf{X}^T \Sigma(\theta)^{-1} \mathbf{Y} \neq \{\mathbf{X}^T \mathbf{X}\}^{-1} \mathbf{X}^T \mathbf{Y}$$

- The formula is complicated, but shows that the regression estimates are not the same as least squares

Computational issues

- Evaluating the likelihood function is slow for large n
- The computational times for both the determinant and inverse of Σ increase like n^3
- For n more than a few hundreds this makes MLE hard to compute
- We will spend an entire lecture on methods application for large n

Computational times



Computational issues

- Another issue to be aware of is singularity of the covariance matrix
- A matrix is singular if its determinent is zero/inverse does not exist
- This happens if correlations are nearly one
- If there is no nugget and the dataset contains two observations at the same location, then Σ is singular
- Even if correlations are not exactly one, high correlation can pose numerical problems

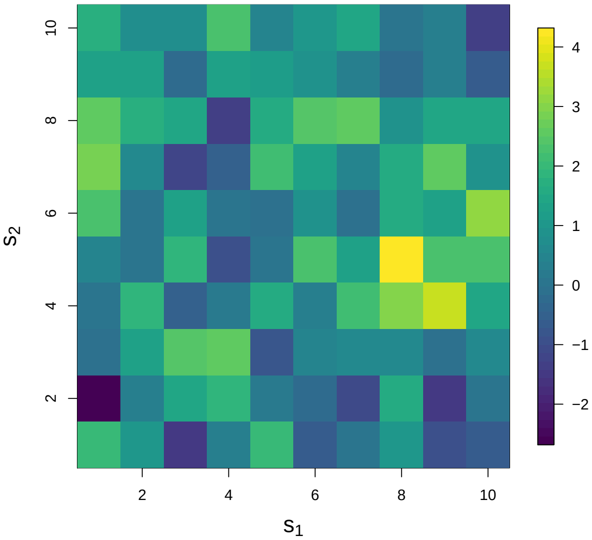
Optimizing the likelihood

- We need to find the values of β and θ that maximize $L(\beta, \theta)$
- There is no closed-form solution so we use numerical optimization
- R packages do this for us
- The idea is to start with an initial value, then follow the derivative of $L(\beta, \theta)$ to the solution
- Supplying good initial values (e.g., least squares for β, variogram for θ) can speed up this process

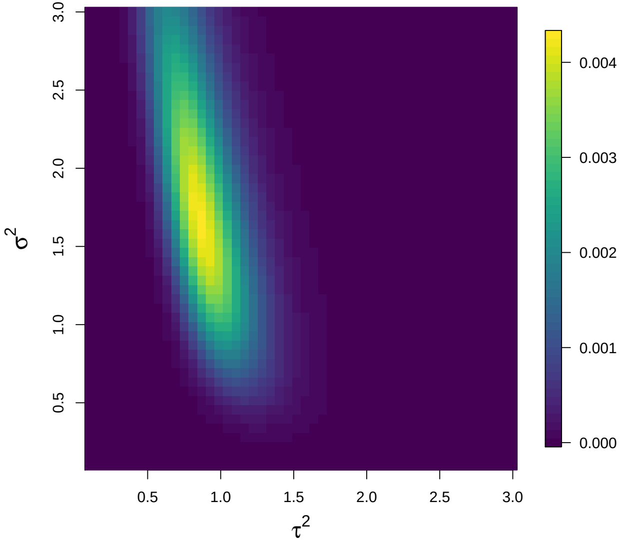
Optimizing the likelihood

- To illustrate this idea, we analyze a simulated dataset
- The data were generated with true values: $\beta_0 = 0$, $\rho = 2$, $\sigma^2 = 2$ and $\tau^2 = 1$
- Data are generated on a 10×10 grid of **s** (next slide)
- Assume only σ^2 and τ^2 are unknown
- We plot the likelihood $L(\tau^2, \sigma^2)$ for $\tau^2, \sigma^2 \in [0, 3]$
- Finally we plot the steps in a (fake) numerical optimization

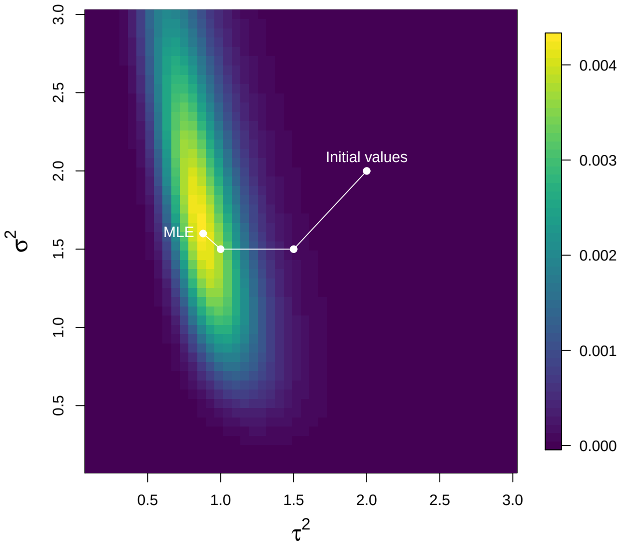
Simulated data (Y)



Likelihood function $L(\tau^2, \sigma^2)$



Numerical optimization



Standard errors

- Given $\theta = \hat{\theta}$, the estimator of β is

$$\hat{\beta} = \{\mathbf{X}^T \Sigma(\hat{\theta})^{-1}\mathbf{X}\}^{-1}\mathbf{X}^T \Sigma(\hat{\theta})^{-1}\mathbf{Y}$$

- Its covariance/standard errors are easy to compute

$$\text{Cov}(\hat{\beta}) = \{\mathbf{X}^T \Sigma(\hat{\theta})^{-1}\mathbf{X}\}^{-1}$$

- This “plug-in” approach to computing standard errors given θ ignores uncertainty in the covariance
- However, this works fine for medium/large datasets
- Confidence intervals and hypothesis tests for the regression coefficients proceed as in linear regression

Standard errors

- Standard errors for the estimator of θ can be computed under a normal approximation
- This uses the second derivatives of the likelihood function
- Unfortunately, these standard errors are unreliable unless the dataset is huge

Model comparisons

Model selection choices include:

- Which covariates to include?
- Should I use a nugget?
- Exponential or Matern correlation?

Model can be compared using cross-validation (later, since it requires prediction) or information criteria

Model comparisons

- AIC/BIC are computed as usual,

$$AIC = -2\log\{L(\hat{\beta}, \hat{\theta})\} + 2k$$

$$BIC = -2\log\{L(\hat{\beta}, \hat{\theta})\} + \log(n)k$$

  where k is the number of parameters in (β, θ)

- Models with smaller AIC/BIC are preferred

- You can use forward/backward selection for selecting covariates

- The covariates selected can depend on the covariance model