Estimation and prediction

ST733 – Spatial Statistics

Spatial statistical inference

- Parameter estimation
- Asymptotics
- Prediction
- Spatial design

Parameter estimation

- Say we have the spatial model $Y(\mathbf{s}) = \mathbf{X}(\mathbf{s})\beta + \varepsilon(\mathbf{s})$, where ε is a GP with mean zero

$$\text{Cov}(\varepsilon(\mathbf{s}_i), \varepsilon(\mathbf{s}_j)) = \sigma^2 M(d_{ij}; \phi, \nu) + \tau^2 I(i = j)$$

  where M is the Matérn correlation function with range ϕ and smoothness ν

- The parameters to be estimated are
  - The mean parameters, β
  - The covariance parameters, $\Theta = \{\sigma^2, \tau^2, \phi, \nu\}$

Parameter estimation

We will explore three estimation approaches:

- Variograms
- Maximum likelihood
- Bayes

Parameter estimation

- Denote the observed data at n locations $\mathbf{s}_1, ..., \mathbf{s}_n$ as $\mathbf{Y} = [Y(\mathbf{s}_1), ..., Y(\mathbf{s}_n)]$

- The $n \times p$ covariate matrix is $\mathbf{X}$ and the $n \times n$ covariance matrix is $\Sigma(\Theta)$ where $\Theta = \{\sigma^2, \tau^2, \phi, \nu\}$ contains the spatial covariance parameters

- The GP evaluated at the n data locations gives

$$\mathbf{Y}|\beta, \Theta \sim \text{Normal}(\mathbf{X}\beta, \Sigma(\Theta))$$

Variograms

The variogram is an exploratory plot to select a model for the correlation function

- Is there spatial correlation?
- Do we need a nugget?
- Is the Matérn correlation function a good fit?

Variograms

- Comparing the true and empirical variograms is a quick methods-of-moments estimator
- The $n \times n$ sample covariance could be used with replication
- Without replication, we need to pool information across pairs of sites separated by distance d to estimate $C(d)$

Variograms

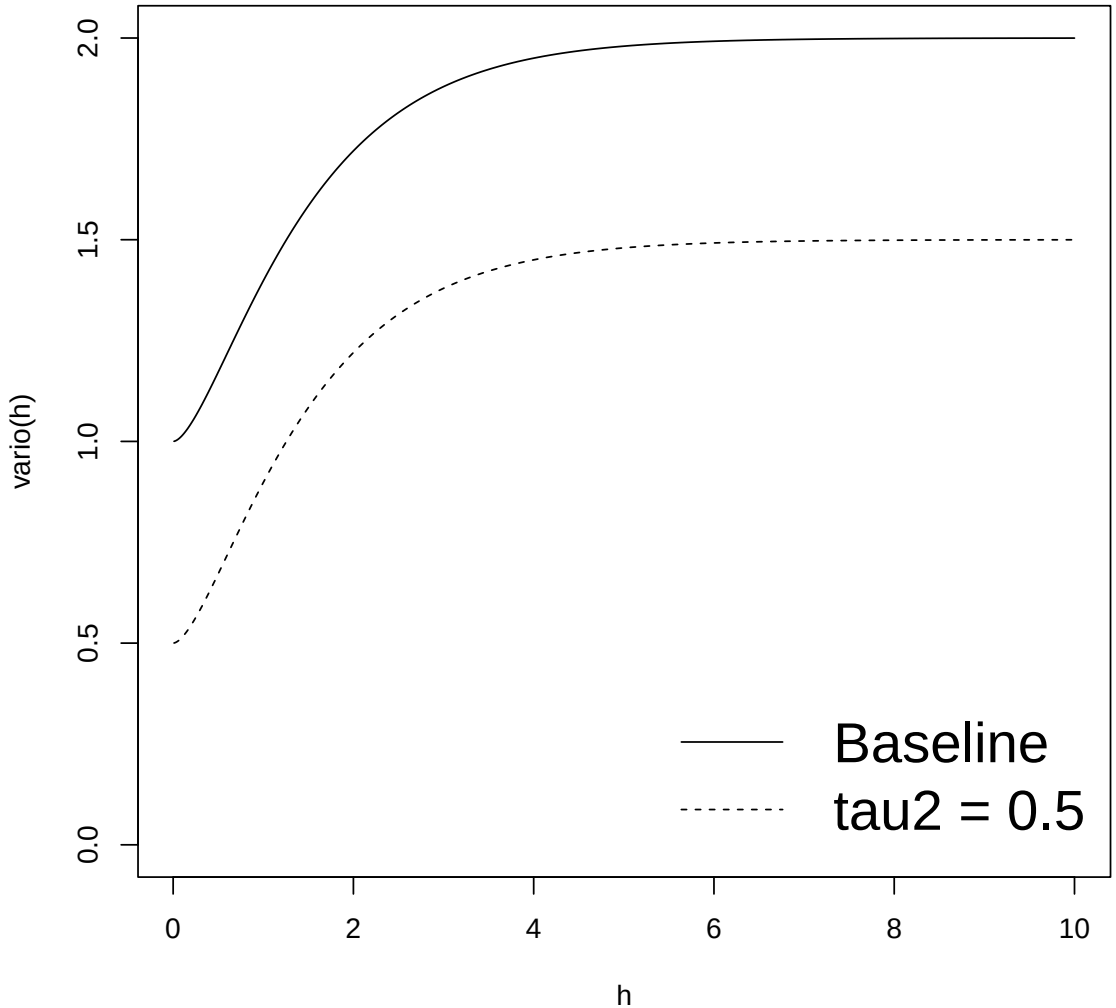
- The process (semi) variogram is

- For a stationary process, the variogram and covariance are related as

Variograms

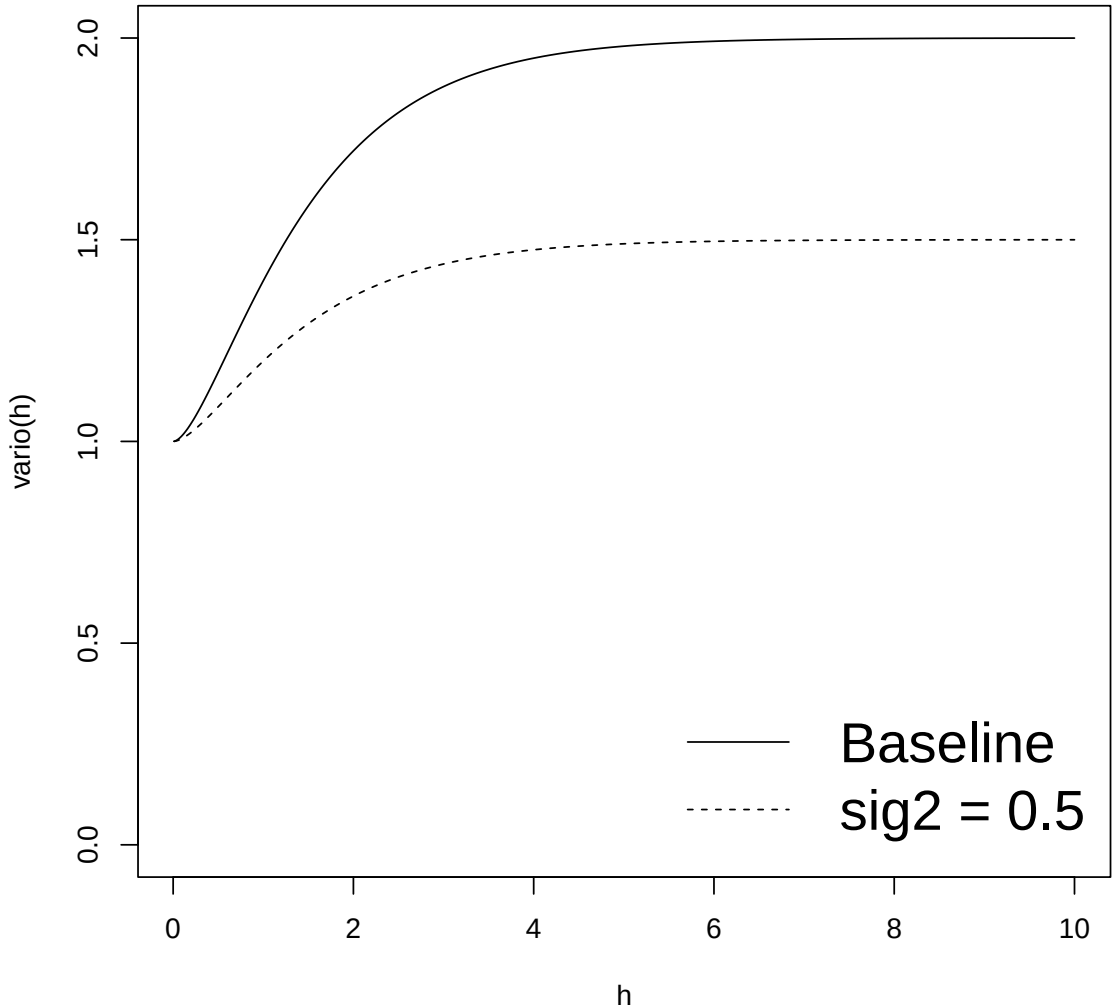
A typical variogram looks like this

The process (true) variogram for the Matérn model



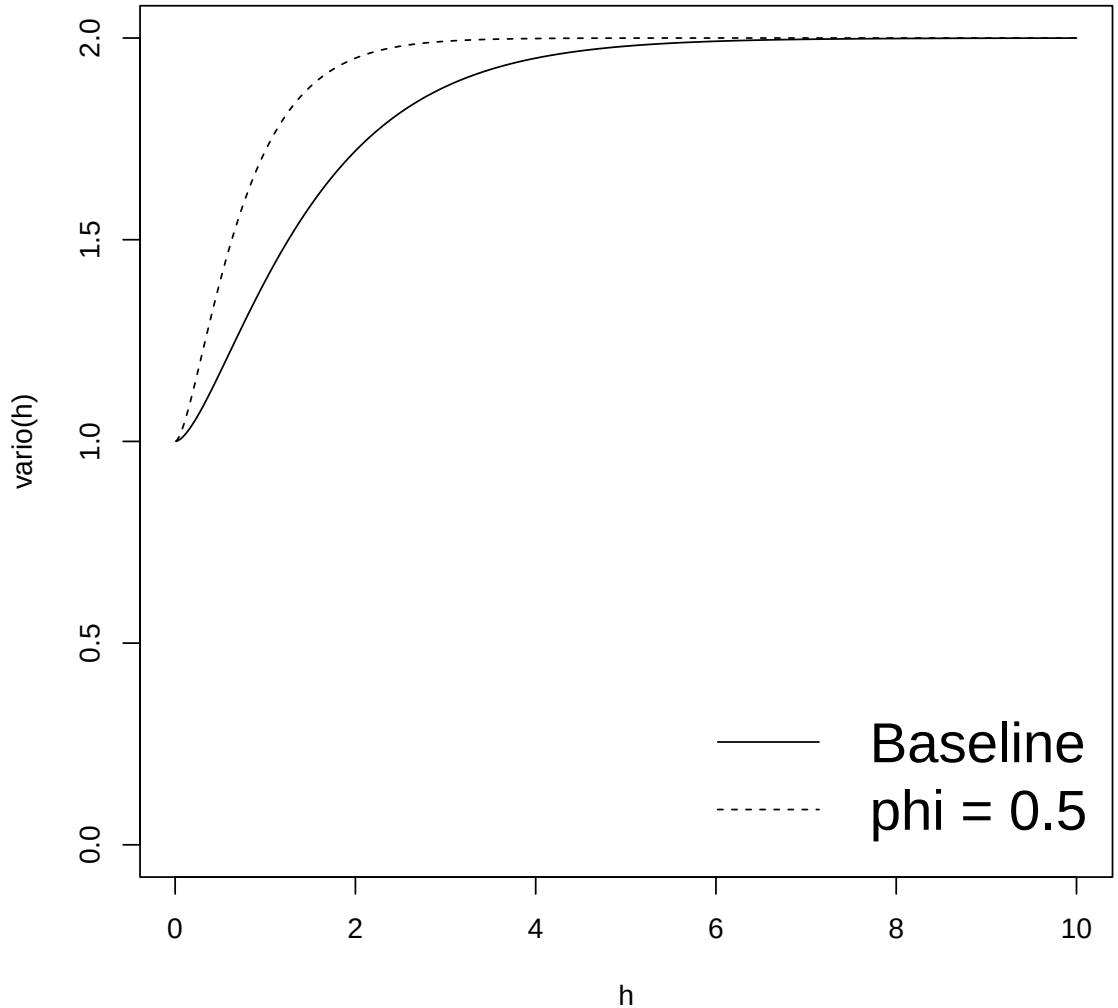
Baseline parameters: $\tau^2 = \sigma^2 = \psi = \nu = 1$

The process (true) variogram for the Matérn model



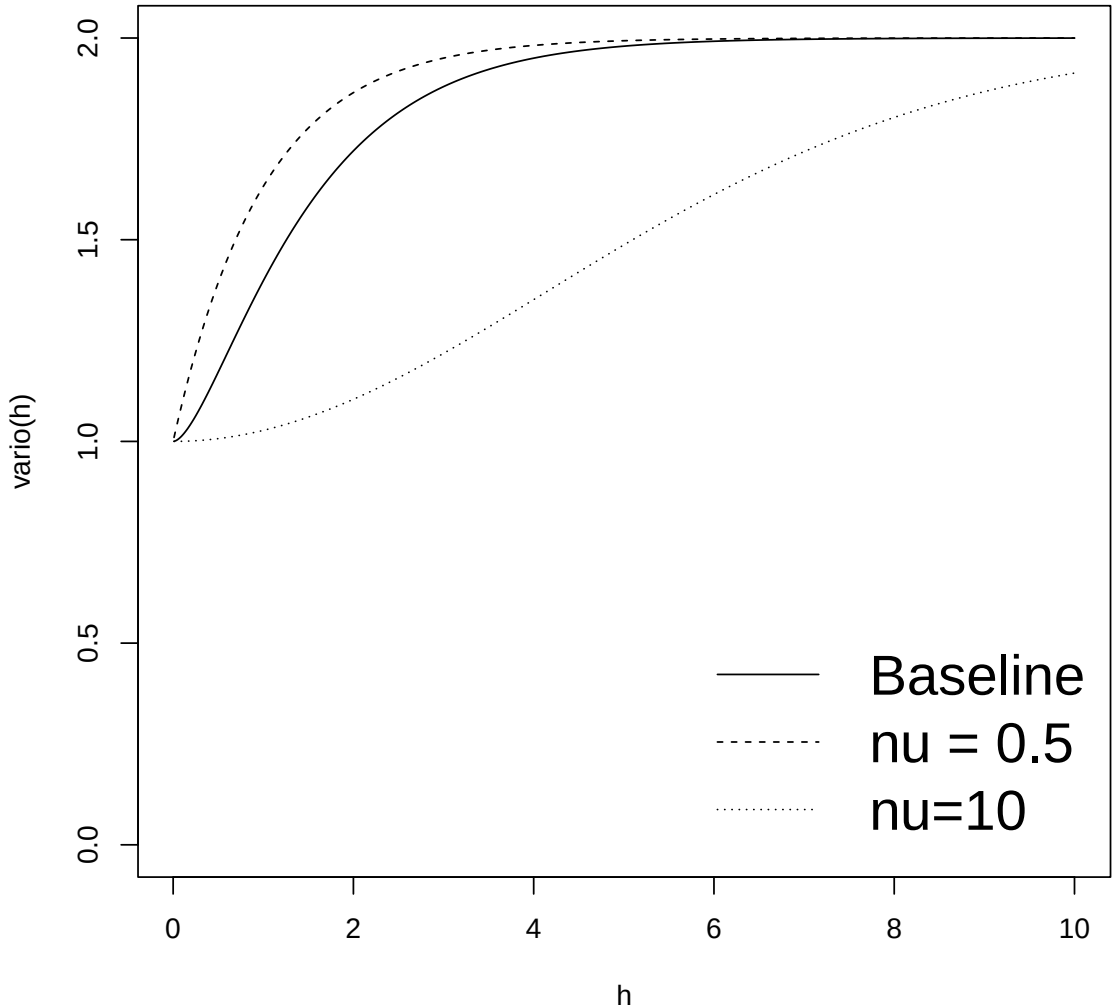
Baseline parameters: $\tau^2 = \sigma^2 = \psi = \nu = 1$

The process (true) variogram for the Matérn model



Baseline parameters: $\tau^2 = \sigma^2 = \psi = \nu = 1$

The process (true) variogram for the Matérn model



Baseline parameters: $\tau^2 = \sigma^2 = \psi = \nu = 1$

Empirical variogram

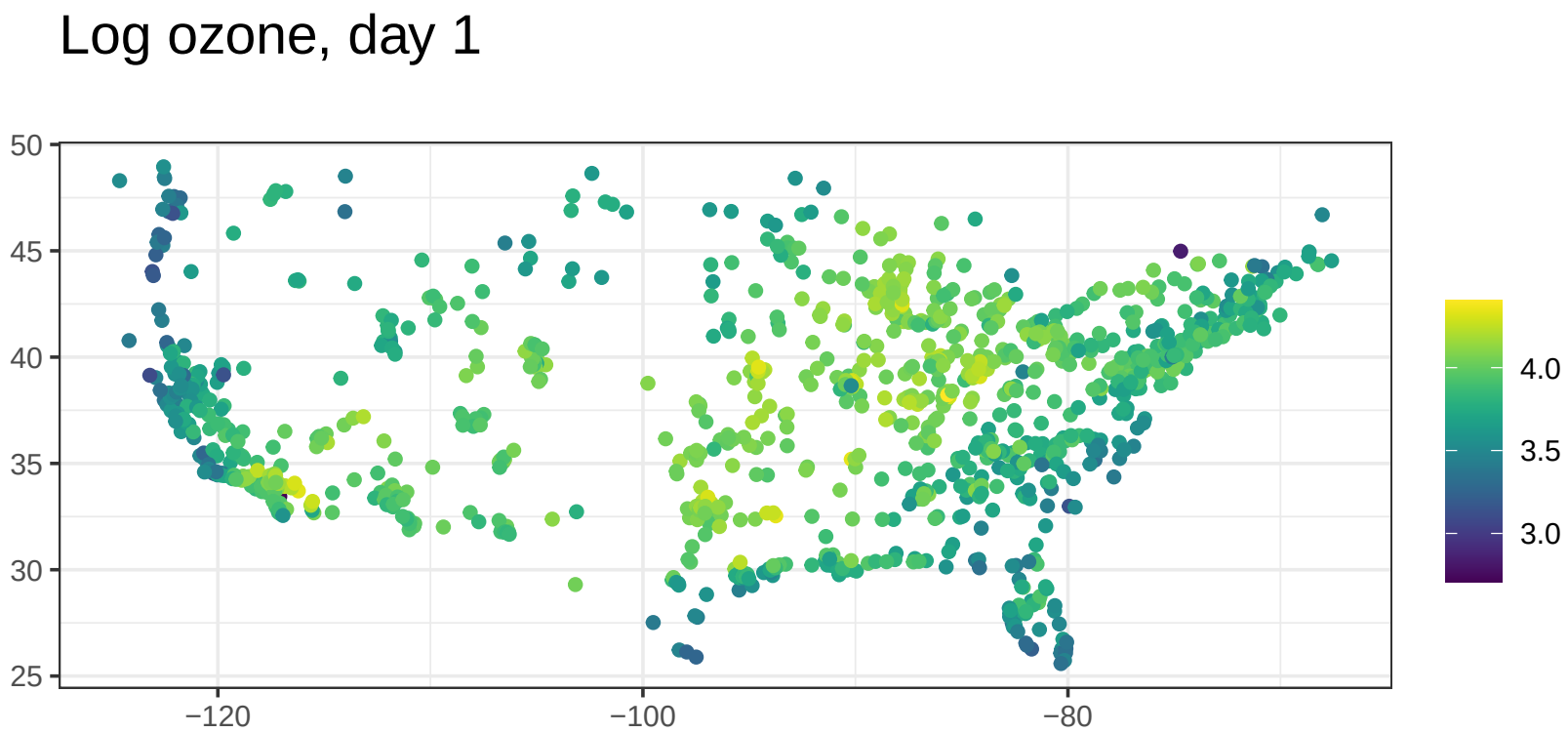
- First the mean is removed with an initial estimate of β, e.g., via least squares, giving $\hat{\varepsilon}(\mathbf{s})$
- Because the variogram is computed from local differences, it is insensitive to mean misspecifcation
- Say there are m_d pairs of points with $d_{ij} \in [d - \epsilon, d + \epsilon) = B(d)$
- The empirical variogram is at distance d is

Empirical variogram

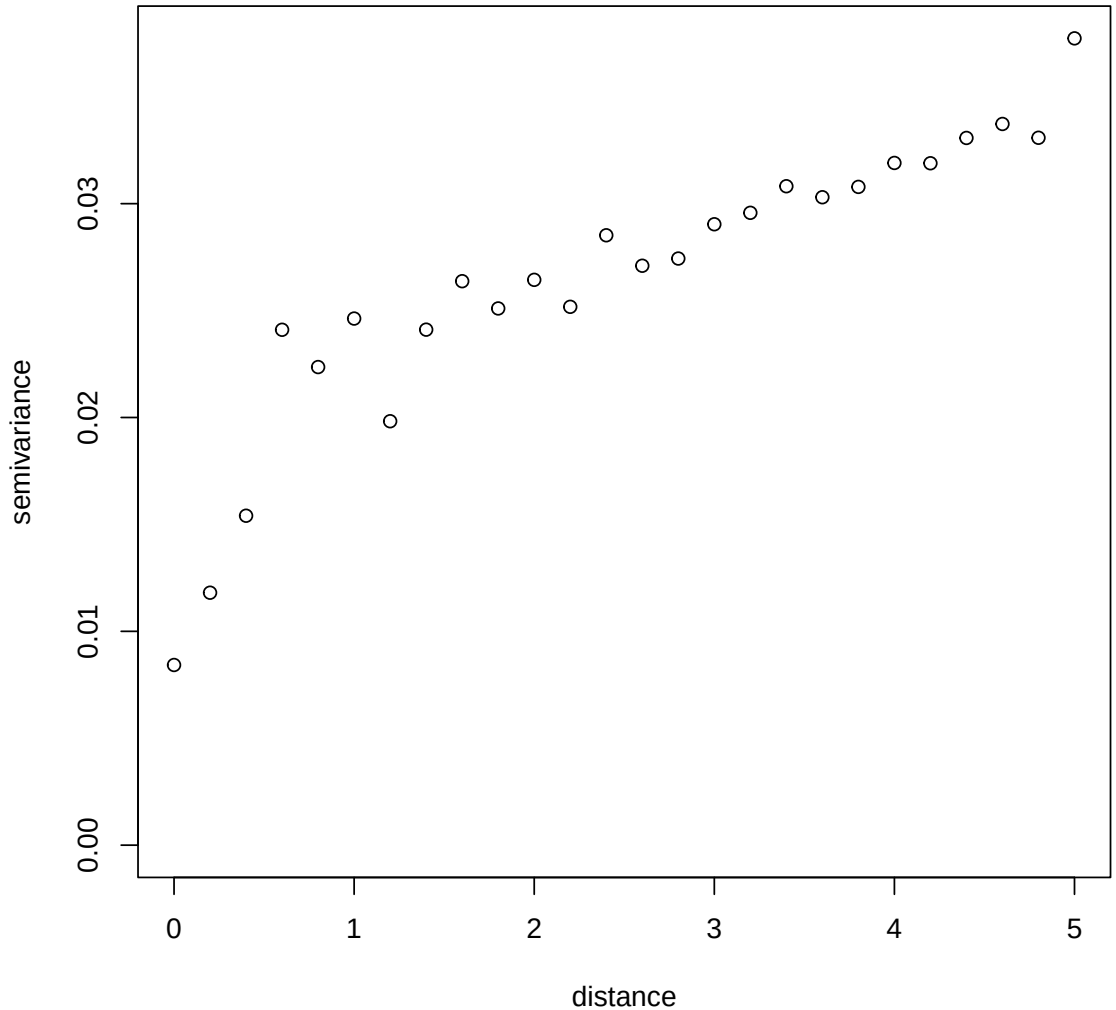
- Repeating this for several d gives the plot

- How to pick the bins?

The process (true) variogram for the Matérn model

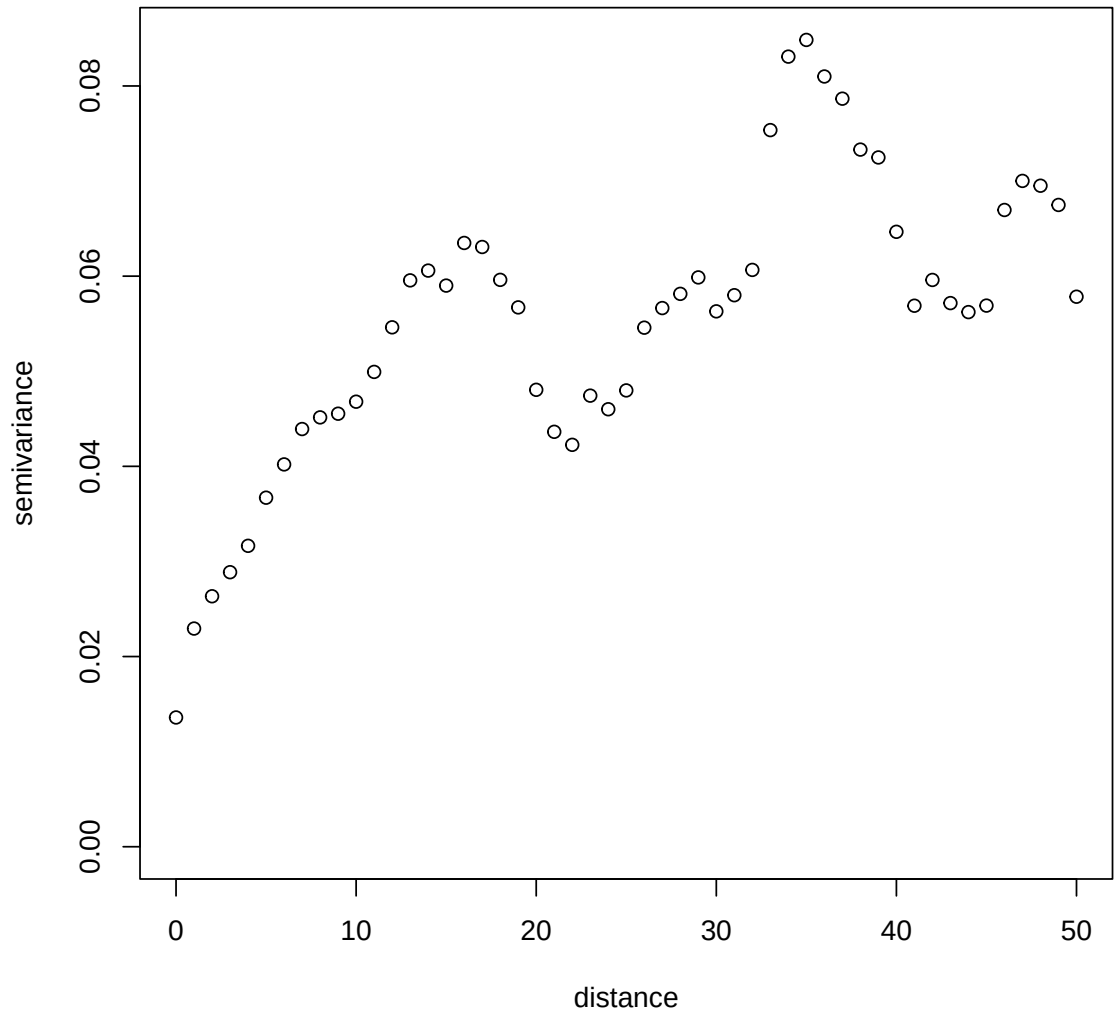


Empirical variogram

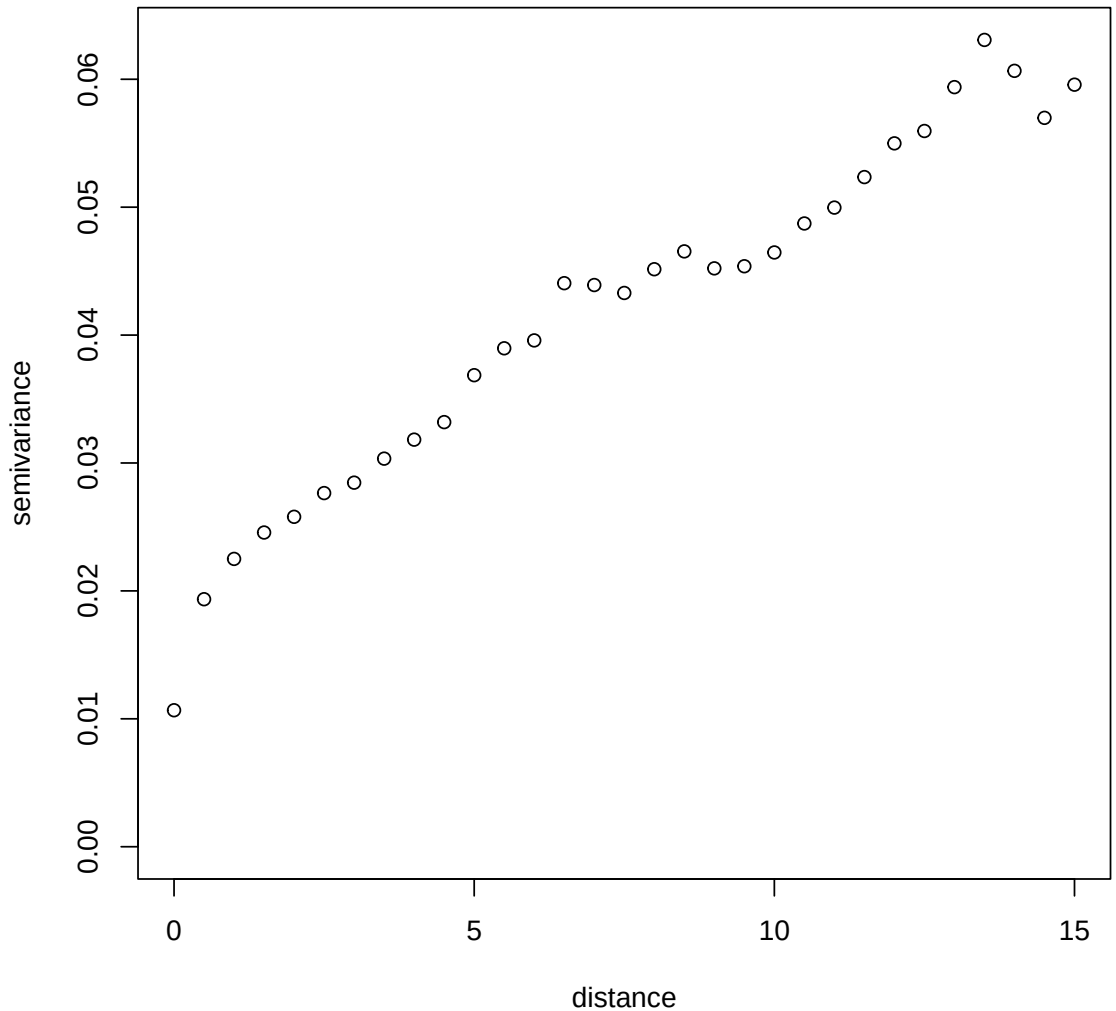



```
plot(geoR::variog(coords = s, data = Y,
     uvec=seq(0,5,.2)))
```

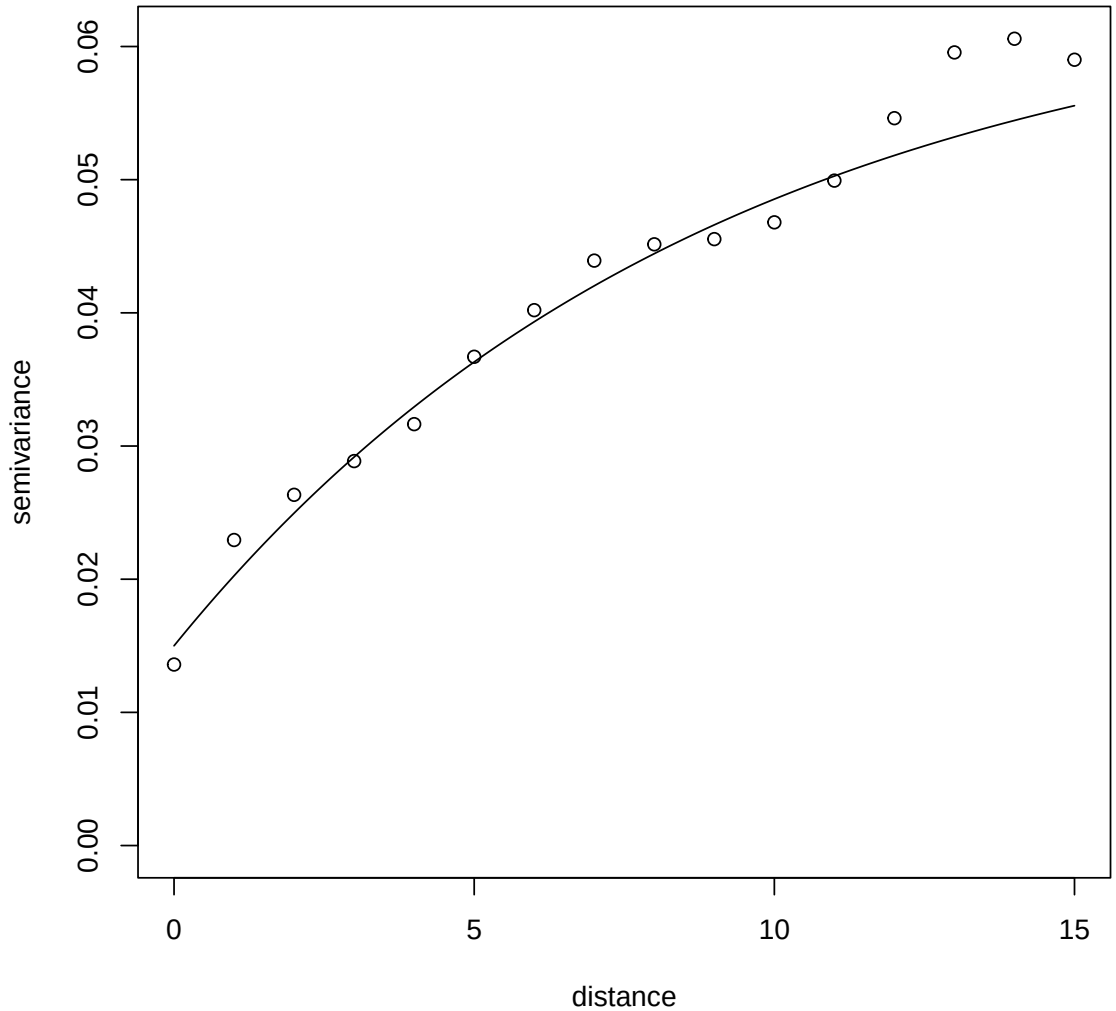
Empirical variogram



Empirical variogram



Empirical variogram



Process variogram with $\tau^2 = 0.015$, $\sigma^2 = 0.05$, $\psi = 9$ and $\nu = 0.5$

Estimating the mean

- Is the least squares estimator valid?

Estimating the mean

- Given the covariance parameters Θ, the MLE is the generalized regression estimator

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

  and the covariance is

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

- A common approach is to plug-in estimates of the covariance parameters Θ and use this for mean estimation. Is this OK?

Relative efficiency

Assume $\mathbf{X}$ is observed and we know the true covariance matrix, Σ. Derive the relative efficiency of GLS compared to OLS.

Relative efficiency of GLS v OLS

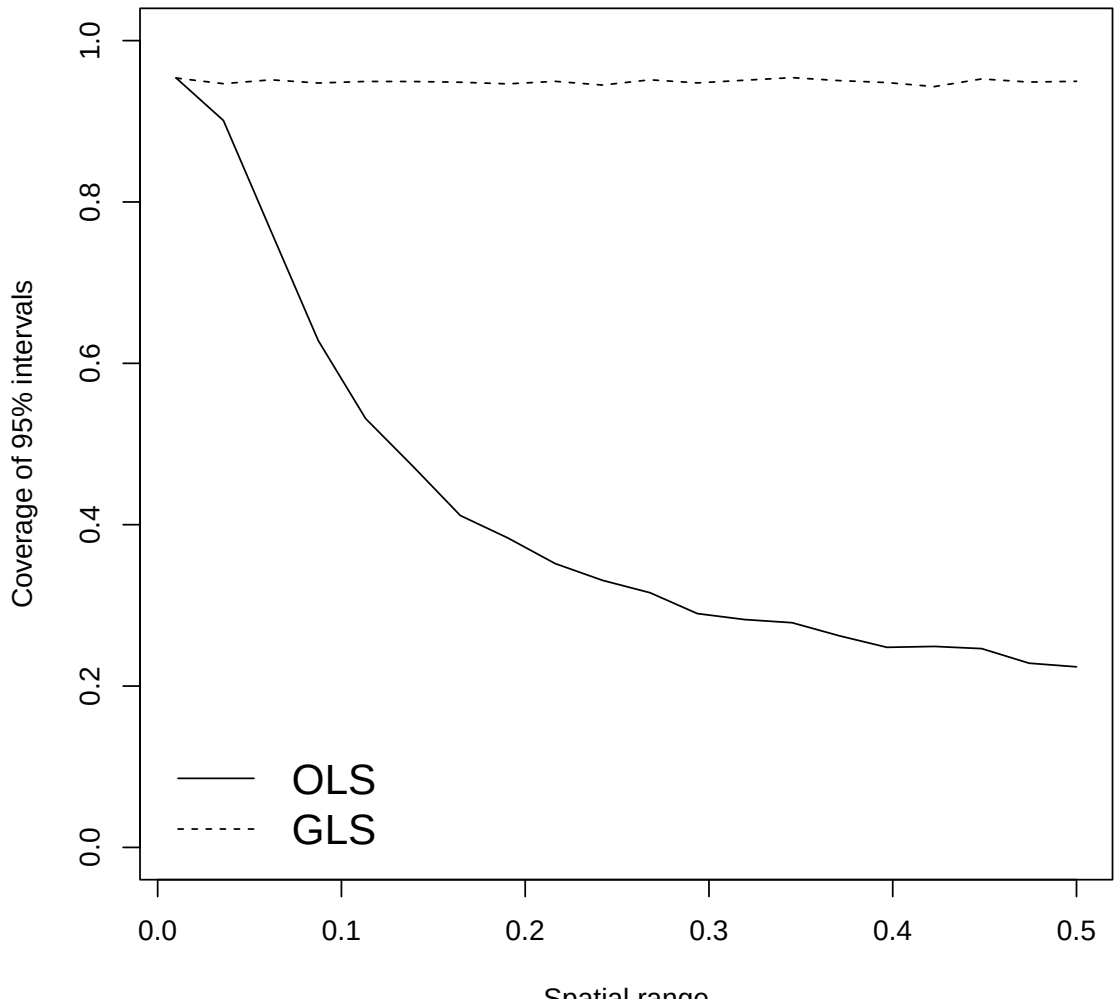
$n = 400$, $\mathbf{s}_i$ on a regular grid covering $[0, 1]^2$, and $\mathbf{X}$ and $\mathbf{Y}$ are GPs with mean zero, variance one (no nugget) and exponential correlation:

ρ_y	ρ_x	RE[1]
0.01	0.01	1.0
0.01	0.10	1.0
0.01	1.00	1.0
0.10	0.01	2.8
0.10	0.10	6.1
0.10	1.00	3.4
1.00	0.01	25.8
1.00	0.10	137.4
1.00	1.00	152.1

[1] Averaged over 100 random $\mathbf{X}$

Proper uncertainty quantification

Under these same settings with $\rho_x = \rho_y$, OLS has low coverage



Estimating the mean

- Hodges: “Adding Spatially-Correlated Errors Can Mess Up the Fixed You Love”
- The spatial estimate of β can often be very different than OLS, even a different sign
- Typically the spatial estimate is shrunk towards zero
- Typically this difference is largest when $X(\mathbf{s})$ is smooth over space

Maximum likelihood analysis

- MLE is a more formal way to estimate β and Θ
- The log likelihood is

- Since the solution for β is a known function of Θ, it can be profiled out giving the profile likelihood

Maximum likelihood analysis

- The likelihood involves the determinant and inverse of the $n \times n$ matrix $\Sigma(\Theta)$

- These are order n^3 operations, so $n > 1000$ is slow

```
> system.time(solve(diag(500)+1))[3]/60
0.000833
> system.time(solve(diag(1000)+1))[3]/60
0.005833
> system.time(solve(diag(2000)+1))[3]/60
0.046166
> system.time(solve(diag(5000)+1))[3]/60
1.153667
```

Maximum likelihood analysis

- The MLE for Θ can be found using standard optimization routines
- REML is also common to reduce bias in the variance estimator
- Good initial values (e.g., variogram estimates) help a lot
- Standard errors can be computing using the Gaussian/Fisher information matrix approximation, but be careful as asymptotics are tricky (next topic)
- Often estimates and covariance for β are obtained using the plug-in estimator previously mentioned

Asymptotic properties

- Assume mean zero and stationary, Matérn covariance
- We would like to show that the MLE for Θ is consistent and asymptotically normal to compute standard errors
- The asymptotics are challenging/interesting because we usually observe only one realization fo the process

Types of asymptotics

- **Replication**: The number of spatial location is fixed, and the number of replications increases
- **Infill**: There is only one replication and the extent of the spatial domain is fixed, but the sampling density increases
- **Increasing domain**: There is only one replication and the sampling density is fixed, but the extent of the spatial domain increases
- **Mixed**: Both infill and increasing domain

Fourier approximation

- Asymptotics in general are challenging because $|\Sigma(\Theta)|$ and $\Sigma(\Theta)^{-1}$ are intractable

- We will consider two special cases:
  - The likelihood can be written exactly for an exponential correlation and locations equally-spaced on a line
  - The likelihood for any stationary process can be approximated using spectral methods on a regular 2D grid

1D exponential case

- Say the n spatial locations are equally-spaced on a line

$$s_i = i\delta_n$$

- Increasing domain if $\delta_n \equiv \delta$ and infill if $\delta_n = 1/n$.

- Assume the data are mean zero with stationary covariance

$$\text{Cov}(h) = \exp(-h/\psi) = \rho^h$$

for $\rho = \exp(-1/\psi)$

1D exponential case

Read and summarize Section 3.1 of:

> *Hao Zhang, Dale L. Zimmerman (2005). Towards reconciling two asymptotic frameworks in spatial statistics. Biometrika,* **92***, 921–936.*

Fourier approximation for data on a 2D grid

- Say the $n = m^2$ points are configured on the square grid with grid spacing one, $\mathcal{S}_m = \{1, 2, ..., m\}^2$
- We use will the DFT at Fourier frequencies

$$\omega \in \mathcal{F}_m = \left\{ \frac{2\pi(0)}{m}, \frac{2\pi(1)}{m}, ..., \frac{2\pi(m-1)}{m} \right\}$$

- The DFT is

$$Y(\mathbf{s}) = \sum_{\omega \in \mathcal{F}_m} \exp(i\mathbf{s}_1^T \omega) Z(\omega)$$

where $Z(\omega)$ are independent complex normals with mean zero and $\mathrm{E}(||Z(\omega)||^2) = \lambda(\omega)$

Fourier approximation

- Assume no nugget ($\tau = 0$), partial sill is one ($\sigma = 1$), exponential covariance ($\nu = 1/2$) and range $1/\alpha$, then

$$\lambda(\boldsymbol{\omega}) = \left(\alpha^2 + d\right)^{-3/2}$$

where $d = ||\boldsymbol{\omega}||^2$

- The Whittle log likelihood is then

Increasing-domain asymptotics

- The expected information matrix is:

Increasing-domain asymptotics

- This increases with n

Infill asymptotics

- Now assume there is the $n = m^2$ points are configured on the square grid that is bounded in the unit square, $\mathcal{S}_m = \{1/m, 2/m, ..., m/m\}^2$
- This is equivalent the previous model but the range redefined as:
- The information matrix is:

Infill asymptotics

- The information is bounded:

Summary of asymptotic results

- Replication:
- Increasing domain:
- Infill:
- Infill + Increasing domain:

Summary of asymptotic results

- This results above deal only with the range. This paper is more general:

  Zhang, Inconsistent estimation and asymptotically equal interpolations in model-based geostatistics, JASA

- Prediction are consistent even under infill asymptotics

- How do these results affect the way you analyze the data?

Bayesian methods

- Bayesian methods are useful across statistics
- Advantages particular to spatial analysis:
- Disadvantages particular to spatial analysis:

Selecting priors

- Uninformative prior the regression coefficients:
  $\beta \sim \text{Normal}(0, 100^2 I_p)$

- I often use the reparameterization:

- Priors:

MCMC

- MCMC returns S samples from the joint posterior of (β, Θ),

$$\left(\beta^{(1)}, \Theta^{(1)}\right), ..., \left(\beta^{(S)}, \Theta^{(S)}\right)$$

- The samples for one parameter are from the marginal posterior accounting for uncertainty in all other parameters

- These samples are summarized in a table of marginal posterior means, standard deviations, and 95% intervals

- If we make a prediction for each of the S samples, prediction uncertainty includes parametric uncertainty

Spatial prediction

- A fundamental task in spatial statistics is to use the data at n location $\mathbf{s}_1, ..., \mathbf{s}_n$ to make a prediction at $\mathbf{s}_0$
- If the data follow a Gaussian processes with known covariance function, then the joint distribution of $Y(\mathbf{s}_0)$ and $\mathbf{Y} = [Y(\mathbf{s}_1), ..., Y(\mathbf{s}_n)]^T$ is
- This leads to the conditional distribution

Spatial prediction

- In an MLE analysis, the parameter Θ (and sometimes β) is plugged into this equation to give predictions
- In a Bayesian analysis, we average over posterior uncertainty in the parameters to make predictions from the posterior predictive distribution
- These approaches make assumptions (normality, linear mean, stationarity, etc.) but (IMO) are fairly robust[2]

[2] Bayesian deep learning can be shown to approximation Kriging for a specfic covariance kernel

Kriging

- Kriging approaches the problem differently by not assuming normality and solving for the best linear unbiased predictor (BLUP)

- This turns out to be equivalent to the normal-based predictions in this special case

- Types of Kriging
  - Simple Kriging: The mean is known (wlg zero)
  - Ordinary Kriging: The mean is constant but unknown
  - Universal Kriging: The mean is $\mathbf{X}\beta$ and β is unknown

Derivation of simple Kriging

- Kriging is the BLUP
  - We only consider linear predictors of the form
  - In this class, we only consider unbiased predictors
- We then seek the “best” predictor in this class as measured by

Simple Kriging

The simple Kriging solution is

Model comparisons

- Model choices include:

- AIC/BIC are applicable for MLE and DIC/WAIC are applicable for Bayes

- If the goal of the analysis is to make spatial predictions, then cross-validation is gold standard using metric such as

Model comparisons

- Choosing the mean and covariance structure simultaneously is ideal
- Spatial variants of the LASSO and Bayesian variable selection priors are available
- Selecting the mean structure using non-spatial OLS and then the covariance structure on the residuals is common
- Are you OK with this?

Spatial design

- Usually statisticians doesn't get to select the measurement locations
- But sometimes we do get to design the survey
- You get to select 100 locations for air pollution monitors in NC, where should you place them?

Preferential sampling

- Preferential sampling occurs if the location of the observation is dependent on the underlying process to be measured

- Examples:

Preferential sampling

- Preferential sampling leads to bias

Preferential sampling

- How to account for this in a statistical analysis?