Spatio-temporal Methods in Environmental Epidemiology

Alexandra M. Schmidt

Session 5 - Modelling point referenced data

Department of Epidemiology, Biostatistics and Occupational Health
McGill University

alexandra.schmidt@mcgill.ca

Applied Bayesian Statistics School 2025

University of Genova, Department of Architecture and Design
03-06, June 2025

Stochastic process

Basic Model

Data (**Y**) are a (partial) realization of a random process (*stochastic process* or *random field*)

$$\{Y(\mathbf{s}) : \mathbf{s} \in D\}$$

where D is a fixed subset of R^d with positive d-dimensional volume. In other words, the spatial index **s** varies *continously* throughout the region D.

Definition

A stochastic process is a collection of random variables $X(t), t \in T$ defined on a common probability space indexed by t which is in the index set T which describes the evolution of some system. For example, $X(t)$ could be the number of people in line at time t, or $Y(\mathbf{s})$ the amount of rainfall at location **s**.

Goal

- Want a method of predicting $Y(\mathbf{s}_0)$ for any $\mathbf{s}_0$ in D
- Want this method to be optimal (in some sense)

What do we need?

- Want $Y(\mathbf{s}) : \mathbf{s} \in D$ to be continuous and "smooth enough" (local stationarity)
- description of spatial covariation
- once we obtain the spatial covariation how to get predicted values

Basic Approach: given variance structure, predict.

Gaussian Processes

Definition

A function $Y(.)$ taking values $y(\mathbf{s})$ for $\mathbf{s} \in D$ has a Gaussian process distribution with mean function $m(.)$ and covariance function $c(.,.)$, denoted by

$$Y(.) \sim GP(m(.), c(.,.))$$

if for any $\mathbf{s}_1, \cdots, \mathbf{s}_n \in D$, and any $n = 1, 2, \cdots$, the joint distribution of $Y(\mathbf{s}_1), \cdots, Y(\mathbf{s}_n)$ is multivariate Normal with parameters given by

$$\begin{aligned} E\{Y(\mathbf{s}_j)\} &= m(\mathbf{s}_j) \text{ and} \\ Cov(Y(\mathbf{s}_i), Y(\mathbf{s}_j)) &= c(\mathbf{s}_i, \mathbf{s}_j) \end{aligned}$$

Multivariate Normal Distribution

The multivariate normal distribution of a n-dimensional random vector $\mathbf{X} = (X_1, \cdots, X_k)'$ can be written as $\mathbf{X} \sim N_k(\mu, \Sigma)$ with k-dimensional mean vector

$$\mu = E(\mathbf{X}) = [E(X_1), E(X_2), \cdots, E(X_k)]'$$

and $k \times k$ covariance matrix

$$\Sigma_{ij} = E[(X_i - \mu_i)(X_j - \mu_j)] = Cov(X_i, X_j)$$

and

$$\Sigma = E[(\mathbf{X} - \mu)(\mathbf{X} - \mu)] = [Cov(X_i, X_j); 1 \leq i, j \leq k]$$

The inverse of the covariance matrix is called the precision matrix, denoted by $\mathbf{Q} = \Sigma^{-1}$

Multivariate Normal Distribution - pdf and properties

$$f(\mathbf{x} \mid \mu, \Sigma) = (2\pi\,|\Sigma|)^{-1/2}\,\exp\left\{-\frac{1}{2}(\mathbf{x}-\mu)'\Sigma^{-1}(\mathbf{x}-\mu)\right\}$$

- $\mathbf{X} \sim N_k(\mu, \Sigma) \Leftrightarrow$ there exist $\mu \in \mathbb{R}^k, \mathbf{A} \in \mathbb{R}^{k \times l}$ such that $\mathbf{X} = \mathbf{AZ} + \mu$ for $Z_n \sim N(0,1)$ i.i.d. And the covariance matrix is such that $\mathbf{AA}' = \Sigma$
- If N-dimensional $\mathbf{x}$ is partitioned as follows

$$\mathbf{x} = \begin{bmatrix} \mathbf{x}_{1_{q\times 1}} \\ \mathbf{x}_{2_{(N-q)\times 1}} \end{bmatrix} \text{ and accordingly}$$

$$\mu = \begin{bmatrix} \mu_{1_{(q\times 1)}} \\ \mu_{2_{(N-q)\times 1}} \end{bmatrix} \text{ and } \Sigma = \begin{bmatrix} \Sigma_{11_{(q\times q)}} & \Sigma_{12} \\ \Sigma_{21_{(N-q)\times q}} & \Sigma_{22_{(N-q)\times(N-q)}} \end{bmatrix}$$

Then $(\mathbf{x}_1 \mid \mathbf{x}_2 = \mathbf{a}) \sim N(\mu_{1.2}, \Sigma_{1.2})$ where

$$\begin{aligned} \mu_{1.2} &= \mu_1 + \Sigma_{12}\Sigma_{22}^{-1}(\mathbf{a} - \mu_2) \\ \Sigma_{1.2} &= \Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21} \end{aligned}$$

Gaussian Processes

Visit

https://distill.pub/2019/visual-exploration-gaussian-processes/

for illustrations of GP's

Intrinsic Stationarity

- It is defined through first differences:

$$\begin{aligned} \mathrm{E}(Y(\mathbf{s}+\mathbf{h})-Y(\mathbf{s})) &= 0, \\ \mathrm{Var}(Y(\mathbf{s}+\mathbf{h})-Y(\mathbf{s})) &= 2\gamma(\mathbf{h}) \end{aligned}$$

- The quantity $2\gamma(\mathbf{h})$ is known as the variogram
- $\gamma(.)$ is known as the *semi-variogram*
- In geostatistics, $2\gamma(.)$ is treated as a *parameter* of the random process $\{Y(\mathbf{s}) : \mathbf{s} \in D\}$ (because it describes the covariance structure)

Second Order Stationarity

- Statistically speaking, some further assumptions about Y have to be made. Otherwise, the data represent an *incomplete* sampling of a single realization, making inference impossible
- A random function $Y(.)$ satisfying:

$$\begin{aligned} \mathrm{E}(Y(\mathbf{s})) &= \mu \quad \forall \mathbf{s} \in D \\ \mathrm{Cov}(Y(\mathbf{s}), Y(\mathbf{s}')) &= C(\mathbf{s}-\mathbf{s}') \quad \forall \mathbf{s}, \mathbf{s}' \in D \end{aligned}$$

  is defined to be second-order stationary
- Furthermore, if $C(\mathbf{s}-\mathbf{s}')$ is a function only of $|| \mathbf{s}-\mathbf{s}' ||$ (it is not a function of the locations), then $C(.)$ is said to be isotropic

Second Order Stationarity

- Notice that a process which is second order stationary is also intrinsic stationary, the inverse is not necessarily true
- There is a stronger type of stationarity which is called strict stationarity (joint probability distribution of the data depends only on the relative positions of the sites at which the data were taken)
- If the random process $Y(.)$ is Gaussian, we need only to specify its first-order and second-order properties, namely its *mean function* and its *covariance function*
- In practice, an assumption of second-order stationarity is often sufficient for inference purposes and it will be one of our basic assumptions

Covariogram and Correlogram

- If
$$Cov(Y(\mathbf{s}), Y(\mathbf{s}')) = C(\mathbf{s} - \mathbf{s}')$$
for all $\mathbf{s}, \mathbf{s}' \in D$, $C(.)$ is called the *covariogram*
- If $C(\mathbf{0}) > 0$, we can define
$$\rho(\mathbf{h}) = C(\mathbf{h})/C(\mathbf{0})$$
as the *correlogram*
- Properties:
  - $C(\mathbf{h}) = C(-\mathbf{h})$
  - $\rho(\mathbf{h}) = \rho(-\mathbf{h})$
  - $\rho(\mathbf{0}) = 1$
  - $C(\mathbf{0}) = Var(Y(\mathbf{s}))$ if $Y(.)$ is second order stationary

Relationships between covariogram and variogram

- Consider

$$Var(Y(\mathbf{s}) - Y(\mathbf{s}')) = Var(Y(\mathbf{s})) + Var(Y(\mathbf{s}')) - 2Cov(Y(\mathbf{s}), Y(\mathbf{s}'))$$

- If $Y(.)$ is second order stationary,

$$Var(Y(\mathbf{s}) - Y(\mathbf{s}')) = 2\{C(\mathbf{0}) - C(\mathbf{s} - \mathbf{s}')\}$$

- If $Y(.)$ is intrinsically stationary,

$$2\gamma(\mathbf{h}) = 2\{C(\mathbf{0}) - C(\mathbf{h})\}$$

- If $C(\mathbf{h}) \to 0$ as $|| \mathbf{h} || \to \infty$ then $2\gamma(\mathbf{h}) \to 2C(\mathbf{0})$
- $C(\mathbf{0})$ is the *sill* of the variogram
- The variogram estimation is to be preferred to covariogram estimation. (See Cressie, p.70 for more details)

Features of the Variogram

- $\gamma(-\mathbf{h}) = \gamma(\mathbf{h})$
- $\gamma(\mathbf{0}) = 0$
- If $\lim_{\mathbf{h}\to 0} \gamma(\mathbf{h}) = c_0 \neq 0$, then c_0 is called the **nugget effect**.
- Mathematically a nugget effect means:
  - If Y is L_2 continuous(processes $Y(.)$ for which $E(Y(\mathbf{s}+\mathbf{h}) - Y(\mathbf{s}))^2 \to 0$, as $|| \mathbf{h} || \to 0$), nugget effects cannot happen!
    - So if continuity is assumed at the microscale (very small h) in the $Y(.)$ process, the only possible reason for $c_0 > 0$ is measurement error. (Recall $2\gamma(\mathbf{0})$ is the variance of the difference between two measurements taken at *exactly* the same place).
  - In practice we only have data $\{y(\mathbf{s}_i) : i = 1, \cdots, n\}$ so we can't say much for lags $h < min\{|| \mathbf{s}_i - \mathbf{s}_j ||: 1 \leq i < j \leq n\}$

Features of the Variogram

- Often in spatial prediction we assume nugget effect is entirely due to measurement error
- In other words, it is reasonable to envisage that a measured value at location $\mathbf{s}$ could be replicated, and that the resulting multiple values would not be identical
- In this case, an alternative is to model a "white noise" zero-mean process that adds extra variation to each observation, that is

$$Y(\mathbf{s}_i) = S(\mathbf{s}_i) + Z(\mathbf{s}_i),$$

where $S(\mathbf{s})$ follows a Gaussian process with covariance function $c(u) = \sigma^2\rho(u)$ such that $\rho(0) = 1$ and the $Z(\mathbf{s}_i)$ are mutually independent, $N(0, \tau^2)$ random variables, and u is the Euclidean distance between locations

- And the nugget effect would be

$$\rho_Y(u) = \sigma^2\rho(u)/(\sigma^2 + \tau^2) \to \sigma^2/(\sigma^2 + \tau^2) < 1$$

as $u \to 0$

Properties of the variogram

(i) $2\gamma(.)$ continuous at origin implies $Y(.)$ is L_2 continuous

(ii) $2\gamma(\mathbf{h})$ does not approach 0 as $\mathbf{h} \rightarrow$ origin implies $Y(.)$ is not L_2 continuous and is highly irregular

(iii) $2\gamma(.)$ is a positive constant (except at the origin where it is zero). Then $Y(\mathbf{s})$ and $Y(\mathbf{s}')$ are uncorrelated for any $\mathbf{s} \neq \mathbf{s}'$, regardless of how close they are; $Y(.)$ is often called *white noise*

Properties of the variogram

(continuing)

(iv) $2\gamma(.)$ must be conditionally negative-definite, i.e.

$$\sum_{i=1}^{n}\sum_{j=1}^{n} a_i a_j 2\gamma(\mathbf{s}_i - \mathbf{s}_j) \leq 0$$

for any finite number of locations $\{\mathbf{s}_i : i = 1, \cdots n\}$ and real numbers $\{a_1, \cdots, a_n\}$ satisfying $\sum_{i=1}^{n} a_i = 0$.

(v) $2\gamma(\mathbf{h})/ \mid\mid \mathbf{h} \mid\mid^2 \rightarrow 0$ as $\mid\mid \mathbf{h} \mid\mid \rightarrow \infty$, i.e. $2\gamma(\mathbf{h})$ can't increase too fast with $\mid\mid \mathbf{h} \mid\mid^2$.

Some Isotropic Parametric Covariance Functions

- Most parametric variogram models used in practice will include a nugget effect, and in the stationary case will be of the form

$$2\gamma(h) = \tau^2 + \sigma^2(1 - \rho(h))$$

- $\rho(h)$ must be a positive definite function
- Also we would usually require the model for the correlation function $\rho(h)$ to incorporate the following features:
  1. $\rho(.)$ is monotone non-increasing in h;
  2. $\rho(h) \to 0$ as $h \to \infty$;
  3. at least one parameter in the model controls the rate at which $\rho(h)$ decays to zero.

Some Isotropic Parametric Covariance Functions

- In addition we may wish to include in the model some flexibility in the overall shape of the correlation function

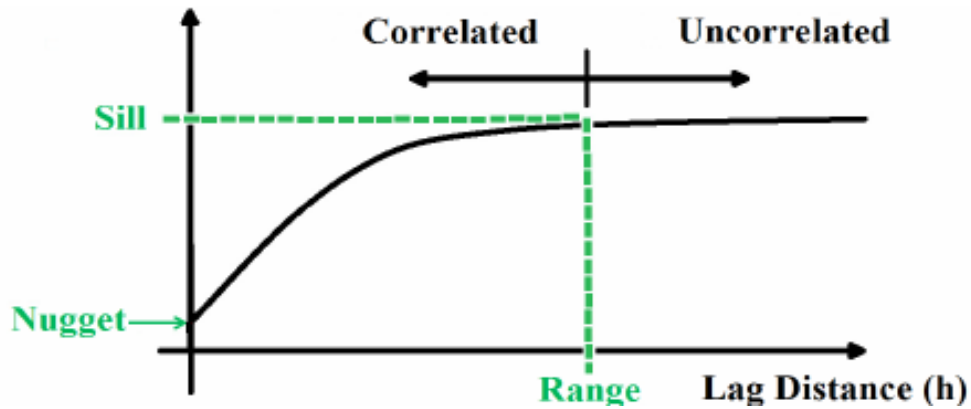
- Hence, a parametric model for the correlation function can be expected to have one or two parameters, and a model for the variogram three or four (the two correlation parameters plus the two variance components)

Variogram

How does a variogram usually look like?

- *nugget effect* - represents micro-scale variation or measurement error. It can be estimated from the empirical variogram as the value of $\gamma(h)$ for $h = 0$.

- *sill* - the $\lim_{h \to \infty} \gamma(h)$ representing the variance of the random field.
- *range* - the distance (if any) at which data are no longer autocorrelated.



Correlation functions

- The spherical family

  This one parameter family of correlation function is defined by

$$\rho(h;\phi) = \begin{cases} 1 - \frac{3}{2}(h/\phi) + \frac{1}{2}(h/\phi)^3 & ;0 \leq h \leq \phi \\ 0 & ;h > \phi \end{cases}$$

  - Because the family depends only on a scale parameter ϕ, it gives no flexibility in shape.
  - The spherical correlation function is continuous and twice-differentiable at the origin.
  - Therefore corresponds to a mean-square differentiable process $Y(\mathbf{s})$.

Correlation functions

- The powered exponential family

  This two parameter family is defined by

  $$\rho(h) = \exp\{-(h/\phi)^{\kappa}\},$$

  with $\phi > 0$ and $0 < \kappa \leq 2$.

  - The corresponding process $Y(\mathbf{s})$ is mean-square continuous (but non differentiable) if $\kappa < 2$, but becomes mean-square infinitely differentiable if $\kappa = 2$
  - The *exponential correlation function* corresponds to the case where $\kappa = 1$
  - The case $\kappa = 2$ is called the *Gaussian correlation function*

Correlation functions

- The Matérn family It is defined by

$$\rho(h;\phi;\kappa) = \{2^{\kappa-1}\Gamma(\kappa)\}^{-1}(h/\phi)^{\kappa}K_{\kappa}(h/\phi)$$

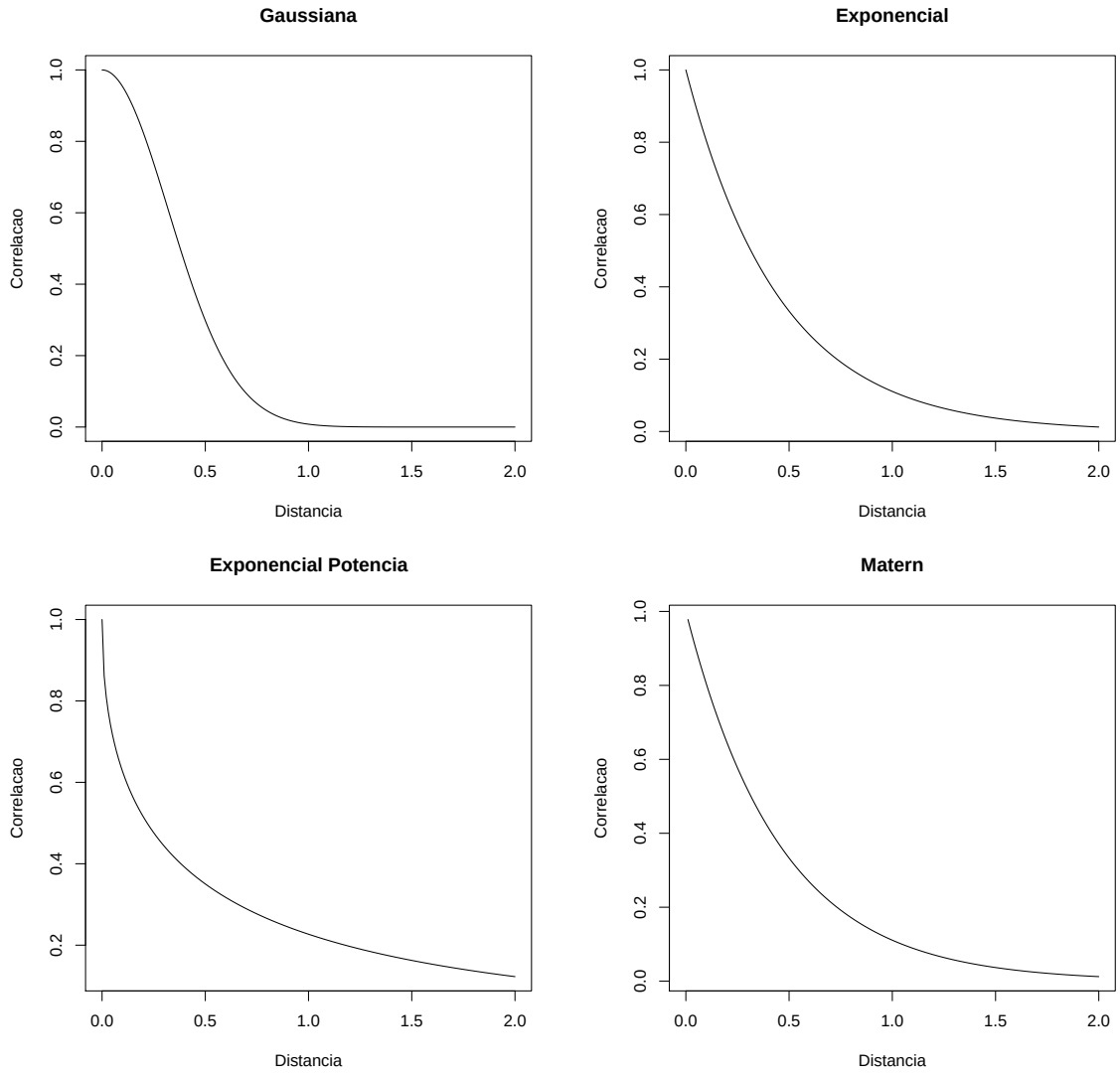
where (ϕ, κ) are parameters and $K_{\kappa}(.)$ denotes the modified Bessel function of the second kind of order κ

  - This family is valid for any $\phi > 0$ and $\kappa > 0$.
  - The case $\kappa = 0.5$ is the same as the exponential correlation function, $\rho(h) = \exp(-h/\phi)$.
  - The squared exponential correlation function is the limiting case as $\kappa \to \infty$
  - Particular cases:
    - $\kappa = 3/2$: $\rho(h) = \left(1 + \frac{\sqrt{3}h}{\phi}\right)\exp\left(-\frac{\sqrt{3}h}{\phi}\right)$
    - $\kappa = 5/2$: $\rho(h) = \left(1 + \frac{\sqrt{5}h}{\phi} + \frac{5h^2}{3\phi^2}\right)\exp\left(-\frac{\sqrt{5}h}{\phi}\right)$

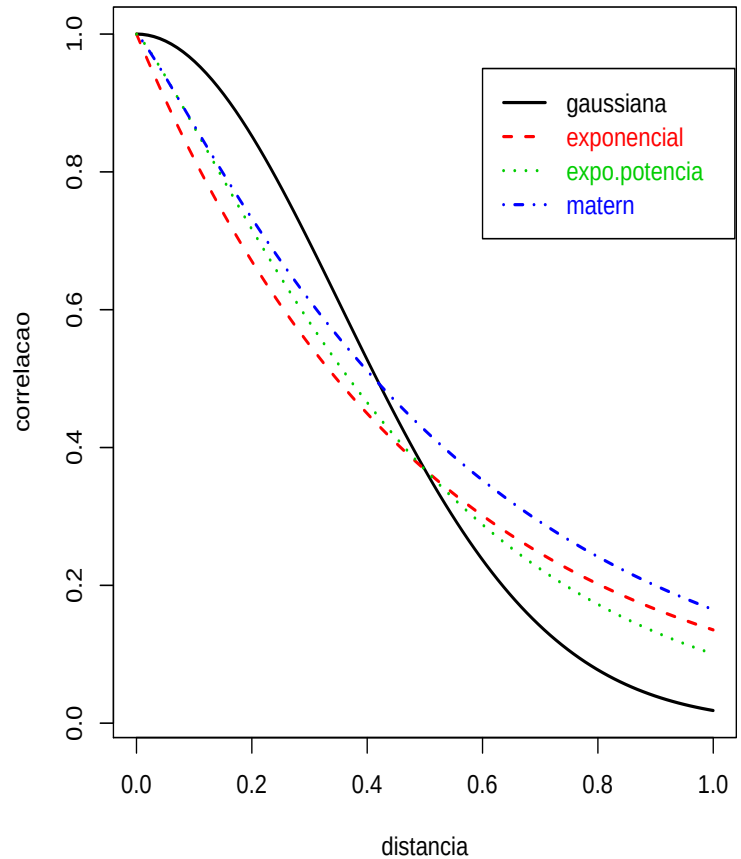
Correlation functions - Matérn family

- One attractive feature of this family is that the parameter κ controls the differentiability of the underlying process $Y(\mathbf{s})$ in a very direct way; the integer part of κ gives the number of times that $Y(\mathbf{s})$ is mean-square differentiable
- The Matérn family is probably the best choice as a flexible, yet simple (only two parameters) correlation function for general case

Correlation functions



Correlation functions



Correlation functions

For simulations of Gaussian processes with different values of the parameters of a correlation function, see document SimulateGPs.pdf

Geometrical Anisotropy

Correlation structure could be different in different directions

It can be corrected by a linear transformation,

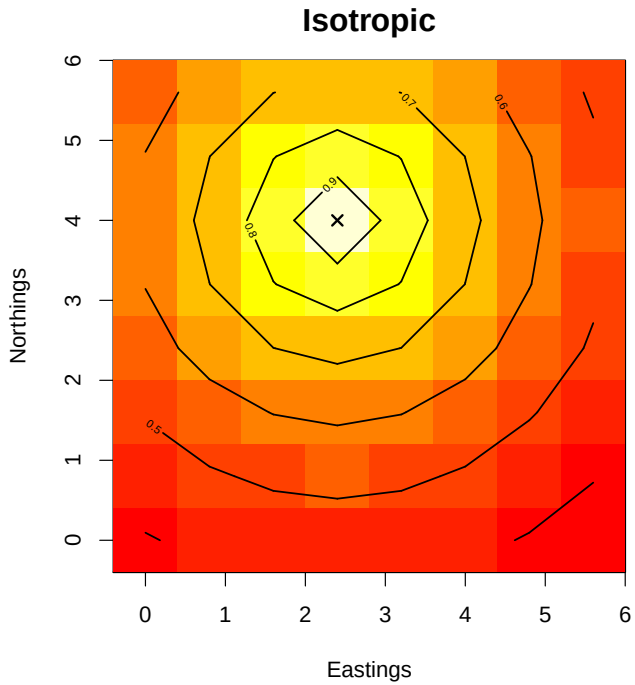
$$2\gamma(\mathbf{h}) = 2\gamma^0(|| \mathbf{Ah} ||), \quad \mathbf{h} \in R^d$$

where $\mathbf{A}$ is a $d \times d$ matrix and $2\gamma^0$ is a function of a real variable. More specifically,

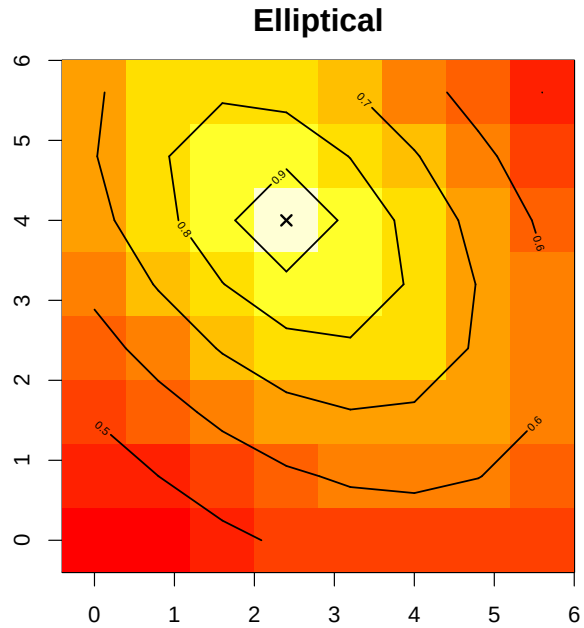
$$(s_1', s_2') = (s_1, s_2) \begin{bmatrix} \cos(\psi_A) & -\sin(\psi_A) \\ \sin(\psi_A) & \cos(\psi_A) \end{bmatrix} \begin{bmatrix} 1 & 0 \\ 0 & \psi_R^{-1} \end{bmatrix},$$

where ψ_A is called the anisotropy angle, and $\psi_R > 1$ is called the anisotropy ratio. The direction along which the correlation decays most slowly with increasing distance is called the principal axis

Comparing isotropic and anisotropic correlation functions



(e) Isotropic



(f) Geometrical anisotropy

And if the anisotropy is not geometric?

There has been a lot of research to relax the assumptions of stationarity and isotropy

- Sampson & Guttorp (JASA, 1992)
- Mardia & Goodall (Multivariate Environmental Statistics, 1993)
- Higdon, Swall & Kern (Valencia 6, 1998)
- Fuentes & Smith (Technical Report, North Carolina State University, 2001)
- Schmidt & O'Hagan (JRSS B, 2003), Zammit-Mangion et al. (JASA, 2021) (Deep GPs)
- Schmidt, Guttorp & O'Hagan (Environmetrics, 2011)
- Viana Neto, Schmidt and Guttorp (JRSS C, 2014)

See Guttorp & Schmidt (2013) for an onverview on Covariance structure of spatial and spatiotemporal processes

Estimation of the variogram

Assume that $Y(.)$ is an intrinsically stationary process

Method of Moments Estimator (or Classical Variogram Estimator)
Under the constant mean assumption, a natural estimator based on the method-of-moments, is

$$2\hat{\gamma}(\mathbf{h}) = \frac{1}{\mid N(\mathbf{h}) \mid} \sum_{N(\mathbf{h})} (Y(\mathbf{s}_i) - Y(\mathbf{s}_j))^2, \quad \mathbf{h} \in R^d$$

where

$$N(\mathbf{h}) = \{(\mathbf{s}_i, \mathbf{s}_j) : \mathbf{s}_i - \mathbf{s}_j = \mathbf{h}; i, j = 1, \cdots, n\}$$

and $\mid N(\mathbf{h}) \mid$ is the number of distinct pairs in $N(\mathbf{h})$

Estimation of the variogram

- $N(\mathbf{h}) \neq N(-\mathbf{h})$, although $2\hat{\gamma}(-\mathbf{h}) = 2\hat{\gamma}(\mathbf{h})$, preserving a property of the theoretical variogram
- $2\hat{\gamma}(.)$ gives point estimates of $2\gamma(.)$ at observed values of $\mathbf{h}$
- $2\hat{\gamma}(.)$ is not necessarily isotropic
- In irregularly spaced data, could have only one pair of locations that are $\mathbf{h}$ apart (two for $|| \mathbf{h} ||$), then lots of variability at each point estimate
- $2\hat{\gamma}(.)$ is not guaranteed to be conditionally negative definite

Estimation of the variogram

- Estimates based on empirical variogram cannot be used for spatial interpolation
- We need to estimate the function $\Rightarrow$ *theoretical variogram*
- Empirical variograms are not necessarily valid (not conditionally non-negative definite)
- Aim: estimate the theoretical variogram that is the closest to the spatial dependence observed in the data $(Y(\mathbf{s}_1), \cdots, Y(\mathbf{s}_n))'$
- Pick a parametric family of variogram and estimate the parameters that provide the best fit to the empirical variogram

Estimation of the variogram - Least Squares

Idea: Plot $2\hat{\gamma}(\mathbf{h})$ and find $\hat{\theta}$ that give a theoretical variogram "close" to the data.

- Let $2\hat{\gamma}(.)$ be the *empirical* variogram estimated at K lags, $h(1), \cdots, h(K)$.
- Let $2\gamma(h;\theta)$ be the theoretical variogram model whose form is known upto θ.
- OLS find $\hat{\theta}$ minimizing

$$\sum_{j=1}^{K} \{2\hat{\gamma}(h(j)\mathbf{e}) - 2\gamma(h(j)\mathbf{e};\theta)\}^2$$

  for some direction $\mathbf{e}$. (Could get different $\hat{\theta}$ in different directions).
- It doesn't introduce the concept of covariation into the procedure

Generalized Least Squares Fitting

- Let $2\hat{\gamma} = (2\hat{\gamma}(h(1)), \cdots, 2\hat{\gamma}(h(K)))$ with variance-covariance matrix $\mathbf{V}$ ($\mathbf{V}$ may depend on θ).
- Let $\hat{\theta}_{\mathbf{V}}$ be the vector of parameter values minimizing

$$(2\hat{\gamma} - 2\gamma(\theta))^T \mathbf{V}^{-1} (2\widehat{\gamma} - 2\gamma(\theta))$$

  where

$$2\gamma(\theta) = [2\gamma(h(1); \theta), \cdots, 2\gamma(h(K); \theta)]^T$$

- $\hat{\theta}_I$ is the OLS estimator, $\hat{\theta}_\delta$ is a weighted least squares (WLS) estimator where $\delta = diag\{Var[2\hat{\gamma}(h(1))], \cdots, Var[2\hat{\gamma}(h(K))]\}$
- Cressie (p.96) shows:

$$Var\{2\hat{\gamma}(h(j))\} \approx 2\{2\gamma(h(j); \theta)\}^2 / \mid N(h(j)) \mid$$

Generalized Least Squares Fitting

- But Cressie (1985a) notes the off-diagonal elements of $\mathbf{V}(\theta)$ can be large
- Minimizing

$$\sum_{j=1}^{K} \mid N(h(j)) \mid \left\{ \frac{\hat{\gamma}(h(j))}{\gamma(h(j);\theta)} - 1 \right\}^2 \qquad (1)$$

  is a good approximation to WLS
- But WLS is not a good approximation to GLS
- Advantages to (1):
  * More weight where more data;
  * More weight where $\gamma(\mathbf{h};\theta)$ is small.
- Equation (1) is a pragmatic compromise between OLS and GLS. Can also be used as a first step in a iterative approach to computing GLS

Variogram estimation using geoR

For variogram estimation, see document
EDA-Variograms.pdf

Kriging: Optimal Spatial Prediction

- Prediction Inference on random quantities.
- Estimation Inference on *fixed*(classical) but *unknown* parameters.
- Motivating Problem

  Observe $\mathbf{Y} = [Y(\mathbf{s}_1), \cdots, Y(\mathbf{s}_n)]'$ Want: $Y(\mathbf{s}_0)$ for any $\mathbf{s}_0 \in D$

  Need: Predictor, $p(\mathbf{Y}; \mathbf{s}_0)$, and estimate of error associated with

  its predictions

Quick Outline of our approach

1. Observe $\mathbf{Y}$;
2. Estimate variogram $2\gamma(.)$ from $\mathbf{Y}$.
   (a) Get point estimates $\{2\hat{\gamma}(h(1)), \cdots, 2\hat{\gamma}(h(K))\}$.
   (b) Pick parametric family, and estimate $\hat{\theta}$ for $2\gamma(\mathbf{h}; \theta)$ using WLS (or GLS).
3. Use $2\gamma(\mathbf{h}; \hat{\theta})$ to get $p(\mathbf{Y}, \mathbf{s}_0)$.

We also want an *optimal* predictor!

Optimal in what way?

- Focus is on *linear* predictors, i.e.

$$p(\mathbf{Y};\mathbf{s}_0) = \sum_{i=1}^{n} l_i Y(\mathbf{s}_i) + k \tag{2}$$

  where k is some constant.
- Similar to a weighted average of the data $\mathbf{Y}$. But what weights?
- Optimality criterion is to minimize *mean squared prediction error* (MSPE):

$$MSPE = E\left[(Y(\mathbf{s}_0) - p(\mathbf{Y};\mathbf{s}_0))^2\right] \tag{3}$$

- (2) and (3) imply we want to minimize

$$E\left[\left(Y(\mathbf{s}_0) - \sum_{i=1}^{n} l_i Y(\mathbf{s}_i) - k\right)^2\right] =$$

$$Var\left(Y(\mathbf{s}_0) - \sum_{i=1}^{n} l_i Y(\mathbf{s}_i)\right)$$

$$+ \quad \left(\mu(\mathbf{s}_0) - \sum_{i=1}^{n} l_i \mu(\mathbf{s}_i) - k\right)^2$$

where $\mu(\mathbf{s}) = E[Y(\mathbf{s})]$, $\mathbf{s} \in D$ is a *known* mean surface.

- Set

$$k = \mu(\mathbf{s}_0) - \sum_{i=1}^{n} l_i \mu(\mathbf{s}_i)$$

and

$$\mathbf{l} = \mathbf{c}' \Sigma^{-1}$$

where $\mathbf{c} = (C(\mathbf{s}_0, \mathbf{s}_1), \cdots, C(\mathbf{s}_0, \mathbf{s}_n))'$ and Σ is a $n \times n$ matrix with i, j^{th} element $C(\mathbf{s}_i, \mathbf{s}_j)$
NOTE: We assume that $C(.,.)$ and $\mu(.)$ are *known*

- These values of k and $\mathbf{l}$ give the optimal linear predictor

$$p^*(\mathbf{Y}; \mathbf{s}_0) = \mathbf{c}' \Sigma^{-1} (\mathbf{Y} - \mu) + \mu(\mathbf{s}_0), \tag{4}$$

where $\mu = (\mu(\mathbf{s}_1), \cdots, \mu(\mathbf{s}_n))'$.

Simple Kriging

Note that $p^*(\mathbf{Y}; \mathbf{s}_0)$ has weights depending on

- Correlations between $Y(\mathbf{s}_0)$ and each data point $Y(\mathbf{s}_i)$, $i = 1, \cdots, n$ and
- Correlations between pairs of data points $Y(\mathbf{s}_i)$ and $Y(\mathbf{s}_j)$, $i, j = 1, \cdots, n$.

MSPE is given by

$$\sigma^2_{SK}(\mathbf{s}_0) = C(\mathbf{s}_0, \mathbf{s}_0) - \mathbf{c}'\Sigma^{-1}\mathbf{c} \tag{5}$$

Matheron called (4) simple kriging

NOTE: Simple kriging requires knowing the mean and covariance structure

Ordinary Kriging

- Assume $Y(.)$ is instrinsically stationary
- Begin with $2\gamma(.)$ known, will replace with $2\gamma(\mathbf{h};\hat{\theta})$ later
- Also assume $Y(\mathbf{s}_1),\cdots,Y(\mathbf{s}_n)$ are measured without error for now. (What does this imply about $p(\mathbf{Y};\mathbf{s}_i)$)?
- Want a *linear predictor*, i.e.

$$p(\mathbf{Y};\mathbf{s}_0) = \sum_{i=1}^{n} \lambda_i Y(\mathbf{s}_i)$$

and

$$\sum_{i=1}^{n} \lambda_i = 1 \tag{6}$$

- Condition (6) guarantees uniform unbiasedness, namely

$$E\{p(\mathbf{Y};\mathbf{s}_0)\} = \mu = E[Y(\mathbf{s}_0)]$$

  and results in $p(\mathbf{Y};\mathbf{s}_0)$ being a weighted average of the data $\mathbf{Y}$.
- Condition (6) was not needed for simple kriging since μ was known.
- We want an *optimal* predictor in terms of minimizing MSPE,

$$\sigma_e^2 = E\left[(Y(\mathbf{s}_0) - p(\mathbf{Y};\mathbf{s}_0))^2\right]$$

- What do we need to find?
  The vectors of optimal weights $\lambda = (\lambda_1, \cdots, \lambda_n)'$.

Finding the optimal weights: the kriging equations

- Since we have to meet unbiasedness constraint (6), we have a constrained minimization problem to solve. One method for solving such problems is by using a *Lagrange multiplier*
- Minimize

$$E\left[\left(Y(\mathbf{s}_0) - \sum_{i=1}^{n} \lambda_i Y(\mathbf{s}_i)\right)^2\right] - 2m\left(\sum_{i=1}^{n} \lambda_i - 1\right) \tag{7}$$

with respect to $\lambda_1, \cdots, \lambda_n$ and m (a Lagrange multiplier that ensures $\sum_{i=1}^{n} \lambda_i = 1$)

- Now, $\sum_{i=1}^{n} \lambda_i = 1$ implies

$$\left(Y(\mathbf{s}_0) - \sum_{i=1}^{n} \lambda_i Y(\mathbf{s}_i) \right)^2 =$$
$$-\frac{1}{2} \sum_{i=1}^{n} \sum_{j=1}^{n} \lambda_i \lambda_j (Y(\mathbf{s}_i) - Y(\mathbf{s}_j))^2$$
$$+2 \sum_{i=1}^{n} \lambda_i (Y(\mathbf{s}_0) - Y(\mathbf{s}_i))^2 / 2$$

So,

$$E\left[\left(Y(\mathbf{s}_0) - \sum_{i=1}^{n} \lambda_i Y\mathbf{s}_i) \right)^2 \right] =$$
$$-\frac{1}{2} \sum_{i=1}^{n} \sum_{j=1}^{n} \lambda_i \lambda_j E\left[(Y(\mathbf{s}_i) - Y(\mathbf{s}_j))^2 \right]$$
$$+2 \sum_{i=1}^{n} \lambda_i E\left[Y(\mathbf{s}_0) - Y(\mathbf{s}_i))^2 \right] / 2$$

- Now (7) becomes

$$-\sum_{i=1}^{n}\sum_{j=1}^{n}\lambda_i\lambda_j\gamma(\mathbf{s}_i-\mathbf{s}_j)+2\sum_{i=1}^{n}\lambda_i\gamma(\mathbf{s}_0-\mathbf{s}_i)$$
$$-2m\left(\sum_{i=1}^{n}\lambda_i-1\right) \tag{8}$$

  So the *variogram* is involved!

- Now, to minimize (8), differentiate with respect to $\lambda_1,\cdots,\lambda_n,m$ in turn and set equal to zero. This gives the systen of equations

$$-\sum_{i=1}^{n}\lambda_j\gamma(\mathbf{s}_i-\mathbf{s}_j)+\gamma(\mathbf{s}_0-\mathbf{s}_i)-m=0,$$

  for $j=1,\cdots,n$, subject to $\sum_{i=1}^{n}=1$.

Let λ_0 denote the vector of values $(\lambda_1, \cdots, \lambda_n, m)'$ that solve the system of equations

$$\lambda_0 = \Gamma_0^{-1}\gamma_0 \tag{9}$$

where

$$\begin{aligned}
\gamma_0 &= (\gamma(\mathbf{s}_0 - \mathbf{s}_1), \cdots, \gamma(\mathbf{s}_0 - \mathbf{s}_n, 1))' \\
\Gamma_{0_{ij}} &= \begin{cases} \gamma(\mathbf{s}_i - \mathbf{s}_j) & i = 1, \cdots, n; \\ & j = 1, \cdots, n; \\ 1 & i = n+1; j = 1, \cdots, n; \\ & j = n+1; i = 1, \cdots, n; \\ 0 & i = j = n+1; \end{cases}
\end{aligned}$$

The system of equations (9) are referred to as the ordinary kriging equations.

We can write $\lambda = (\lambda_1, \cdots, \lambda_n)'$ and m in terms of

$$\begin{aligned} \Gamma &= n \times n \text{ matrix with } (i,j)^{th} \text{element } \gamma(\mathbf{s}_i - \mathbf{s}_j) \\ \gamma &= n \times 1 \text{ vector with } i^{th} \text{ element } \gamma(\mathbf{s}_0 - \mathbf{s}_i) \end{aligned}$$

giving

$$\lambda^T = \left(\gamma + \mathbf{1} \frac{(1 - \mathbf{1}'\Gamma^{-1}\gamma)}{\mathbf{1}'\Gamma^{-1}\mathbf{1}} \right)^T \Gamma^{-1}$$

and

$$m = -\left(1 - \mathbf{1}'\Gamma^{-1}\gamma\right)\left(\mathbf{1}'\Gamma^{-1}\gamma\right)$$

We denote the *ordinary kriging predictor* as $\widehat{p}(\mathbf{Y}; \mathbf{s}_0)$ or simply $\widehat{Y}(\mathbf{s}_0)$.

The minimized prediction error

The MSPE (also known as *kriging error* or *prediction error*) is

$$\begin{aligned}\sigma_k^2(\mathbf{a}_0) &= \lambda_0'\gamma_0 \\ &= \sum_{i=1}^{n}\lambda_i\gamma(\mathbf{s}_0-\mathbf{s}_i)+m \\ &= \gamma'\Gamma^{-1}\gamma-(\mathbf{1}'\Gamma^{-1}\gamma-1)^2/(\mathbf{1}'\Gamma^{-1}\mathbf{1})\end{aligned}$$

We can also write the kriging error as a function of the λ_i's alone (and leave out m), giving

$$\sigma_k^2(\mathbf{s}_0) = 2\sum_{i=1}^{n}\lambda_i\gamma(\mathbf{s}_0-\mathbf{s}_i)-\sum_{i=1}^{n}\sum_{j=1}^{n}\lambda_i\lambda_j\gamma(\mathbf{s}_i-\mathbf{s}_j).$$

For $Y(.)$ Gaussian we can construct 95% prediction intervals:

$$\hat{Y}(\mathbf{s}_0)\pm 1.96\sigma_k(\mathbf{s}_0)$$

We can also map or plot $\hat{Y}(\mathbf{s}_0)$ and $\sigma_k^2(\mathbf{s}_0)$ for many $\mathbf{s}_0$'s in D.

Comments on Ordinary Kriging

- Ordinary kriging is an *exact interpolator*, i.e. kriging surface *must* pass through all data points
- Due to assumption of no measurement error (can relax)
- Effect on maps of MSPE
  - Assumption of no measurement error means there are ”pits” in the error surface, leading to ”fried egg” contour plots
  - Pits not always obvious in `R` unless the prediction grid is very tight

Scale of Variation

Spatial prediction requires some assumptions about the process $Y(.)$ that are unverifiable:

- Intrinsic stationarity
- Correct parametric variogram family chosen
- Others about *spatial scale*

Definition The *spatial scale* of a process $Y(.)$ relates to the distance at which spatial correlation occurs.
e.g. distribution of neurons in a room vs. dist of neurons in a single brain.

- *Spatial scale* is relative to the extent of the study area D.
- *Spatial scale* is also relative to the *scale of measurement*, and the precision of the distance measures.
- The idea of spatial scale also relates to the decomposition between a (linear) trend ($\mathbf{X}\beta$) and the correlation structure ($\delta(\mathbf{s})$) mentioned earlier.

Cressie(1993) presents a decomposition of the $Y(.)$ process into components related to spatial scale:

$$Y(\mathbf{s}) = \mu(\mathbf{s}) + W(\mathbf{s}) + \eta(\mathbf{s}) + \varepsilon(\mathbf{s}), \mathbf{s} \in D \tag{10}$$

where

- $\mu(.)$ is the deterministic mean structure or large-scale variation;
- $W(.)$ is a zero-mean, L_2 continuous ($E(W(\mathbf{s}+\mathbf{h}) - W(\mathbf{s}))^2 \to 0$ as $|| \mathbf{h} || \to 0$), intrinsically stationary process, with variogram range *larger* than

$$\min\{|| \mathbf{s}_i - \mathbf{s}_j ||; 1 \leq i < j \leq n\},$$

  (i.e. at least *some* observations are correlated). This is the smooth scale variation.

- $\eta(.)$ is a zero-mean, intrinsically stationary process, independent of $W(.)$, with variogram range *smaller* than

$$\min\{|| \mathbf{s}_i - \mathbf{s}_j ||; 1 \leq i < j \leq n\},$$

  This is the microscale variation. $\varepsilon(.)$ is a zero-mean, white noise process, independent of $W(.)$ and $\eta(.)$. This is the measurement error, and let $c_{ME} = Var(\varepsilon(.))$.
  NOTE: $2\gamma_Y(.) = 2\gamma_W(.) + 2\gamma_\eta(.) + 2c_{ME}$.

Comments

- This decomposition is not unique
- Repeated measurements at the same location would give an estimate of the variance associated with $\varepsilon(.)$
- More observations at new (close) spatial locations might give estimates of microscale variation
- *Large scale* variation usually cannot be extracted uniquely
- In prediction this ambiguity does not affect prediction as much as the standard errors of prediction
- NOTE: Changing the mean structure will change the correlation structure

The nugget effect and Kriging

Using the decomposition of $Y(.)$ given in (10) as a guide:

- If the microscale variation modeled by $\eta(.)$ is L_2 continuous, then the only possible reason for a nugget effect is measurement error (i.e. $Y_1(\mathbf{s})$ and $Y_2(\mathbf{s})$ taken at the same $\mathbf{s}$ yield different values)
- However, in any single application:
  - We only have $Y(\mathbf{s}_i), i = 1, \cdots, n$
  - We can't say much about $2\gamma(.)$ for lags less than

$$\min\{|| \mathbf{s}_i - \mathbf{s}_j ||; 1 \leq i < j \leq n\},$$

  - We don't know if $\eta(.)$ is L_2 continuous or not.

Kriging with measurement error

What do we want to predict?

We want a prediction of a smooth process that accounts for measurement error. That is, we want to predict some central ("expected") value at each $\mathbf{s}_0$ we don't necessarily want to predict the measurement noise

We want to predict

$$S(\mathbf{s}_0) = \mu(\mathbf{s}_0) + W(\mathbf{s}_0) + \eta(\mathbf{s}_0).$$

$S(.)$ is called the *smoothed* or *measurement error free* process.

The kriging equations for $S(.)$ give

$$\widehat{S}(\mathbf{s}_0) = \sum_{i=1}^{n} \nu_i Y(\mathbf{s}_i)$$

with optimal weights satisfying

$$\Gamma_0 \nu_0 = \gamma_0^*$$

- The LHS is the same as the O.K. equations, and $\gamma_0^* = (\gamma^*(\mathbf{s}_0 - \mathbf{s}_1), \cdots, \gamma^*(\mathbf{s}_0 - \mathbf{s}_n), 1)'$. What is $\gamma^*(.)$?

  If $\mathbf{s}_0 \neq \mathbf{s}_i$, $\gamma^*(\mathbf{s}_0 - \mathbf{s}_i) = \gamma_Y(\mathbf{s}_0 - \mathbf{s}_i), i = 1, \cdots, n$.
  If $\mathbf{s}_0 = \mathbf{s}_i$, $\gamma^*(\mathbf{s}_0 - \mathbf{s}_i) = c_{ME} (\neq 0)$.

- The minimized MSPE is given by

$$\tau_k^2(\mathbf{s}_0) = \sum_{i=1}^{n} \nu_i \gamma^*(\mathbf{s}_0 - \mathbf{s}_i) + m - c_{ME}.$$

  Compare to

$$\sigma_k^2(\mathbf{s}_0) = \sum_{i=1}^{n} \lambda_i \gamma(\mathbf{s}_0 - \mathbf{s}_i) + m.$$

What does measurement error change?

Compare

$$\Gamma_0\lambda_0 = \gamma_0 \quad \text{no ME} \tag{11}$$
$$\Gamma_0\lambda_0 = \gamma_0^* \quad \text{ME} \tag{12}$$

- (11) is an *exact interpolator*, i.e. $p(\mathbf{Y};\mathbf{s}_i) = Y(\mathbf{s}_i)$.
- (11) "honors" the data (must go through data).
- (12) "smooths" the data, larger $c_{ME} \Rightarrow$ more smoothing.
- An extreme case gives a good comparison:

$$Y(\mathbf{s}) = \mu + \varepsilon(\mathbf{s})$$

i.e. measurement error is the *only* variation in $Y(.)$.

- Ordinary kriging (11) gives

$$\widehat{Y}(\mathbf{s}_0) = \begin{cases} \overline{Y} & \mathbf{s}_0 \notin \{\mathbf{s}_0, \cdots, \mathbf{s}_n\} \\ Y(\mathbf{s}_0) & \mathbf{s}_0 \in \{\mathbf{s}_0, \cdots, \mathbf{s}_n\} \end{cases}$$

  where $\overline{Y} = \frac{1}{n}\sum_{i=1}^{n} Y(\mathbf{s}_i)$.
- Kriging with measurement error (12) gives

$$\widehat{S}(\mathbf{s}_0) = \overline{Y}$$

  everywhere (even at data locations).
  This example shows how hard ordinary kriging works to be an exact interpolator, i.e. predict $\overline{Y}$ where there is no data, predict $Y(\mathbf{s}_i)$ at $i = 1, \cdots, n$.

A few notes

- In general, neither $\widehat{Y}(\mathbf{s}_0)$ or $\widehat{S}(\mathbf{s}_0)$ is continuous in $\mathbf{s}_0$ (but $\hat{S}(\mathbf{s}_0)$ is "more continuous" than $\widehat{Y}(\mathbf{s}_0)$)
  - No microscale variation implies $\widehat{S}(.)$ is continuous
  - No microscale variation *or* measurement error implies $\widehat{Y}(.)$ is continuous
- When $\mathbf{s}_0$ not $\in \{\mathbf{s}_1, \cdots, \mathbf{s}_n\}$, $\widehat{S}(\mathbf{s}_0) = \widehat{Y}(\mathbf{s}_0)$
- $\sigma_k^2(\mathbf{s}_0)$, $\tau_k^2(\mathbf{s}_0)$ are not equal unless $c_{ME} = 0$

Universal Kriging

Now, μ is considered an *unknown* linear combination of *known* functions $\{f_0(\mathbf{s}), \cdots, f_p(\mathbf{s})\}$, i.e.

$$Y(\mathbf{s}) = \sum_{j=1}^{p+1} f_{j-1}(\mathbf{s})\beta_{j-1} + \delta(\mathbf{s}), \mathbf{s} \in D$$

where $\beta = (\beta_0, \cdots, \beta_p)^T$ is an unknown vector of parameters; and $\delta(.)$ is a zero mean, intrinsically stationary process with variogram $2\gamma(.)$.

Linear model formulation:

$$\mathbf{Y} = \mathbf{X}\beta + \delta$$

where $\mathbf{X}$ is a $n \times (p+1)$ matrix with $(i,j)^{th}$ element equal to $f_{j-1}(\mathbf{s}_i)$.

We can think of this in terms of the decomposition of spatial variation we used before.
Specifically, for *universal kriging* we assume:

$$Y(.) = \mu(.) + W(.),$$

i.e. $Y(.)$ consists of a deterministic (non-random) trend, $\mu(.)$, and small-scale variation (with mean zero), $W(.)$.
The linear model formulation sets

$$\mu(.) = \mathbf{X}\beta$$

and

$$\delta(.) = W(.).$$

As a result, $\gamma_Y(.) = \gamma_W(.) = \gamma_\delta(.)$.

For $p(\mathbf{Y};\mathbf{s}_0)$, we want the BLUP, i.e. predictor of the form

$$p(\mathbf{Y};\mathbf{s}_0) = \sum_{i=1}^{n} \lambda_i Y(\mathbf{s}_i)$$

subject to

$$\lambda^T \mathbf{X} = \mathbf{x}^T \tag{13}$$

where $\mathbf{x}^T = (f_0(\mathbf{s}_0), \cdots, f_p(\mathbf{s}_0))^T$.

NOTE: (13) gives uniform unbiasedness:

$$E(p(\mathbf{Y};\mathbf{s}_0)) = E[\lambda^T \mathbf{Y}] = \lambda^T \mathbf{X}\beta$$

Now, to find optimal predictor, minimize the MSPE

$$\sigma_e^2 = E\{[Y(\mathbf{s}_0) - p(\mathbf{Y}; \mathbf{s}_0)]^2\}$$

subject to $\lambda^T \mathbf{X} = \mathbf{x}^T$, i.e. minimize

$$\begin{aligned} \sigma_e^2 &= E\left\{\left[Y(\mathbf{s}_0) - \sum_{i=1}^{n} \lambda_i Y(\mathbf{s}_i)\right]^2\right\} \\ &- 2\sum_{j=1}^{p+1} m_{j-1}\left\{\sum_{i=1}^{n} \lambda_i f_{j-1}(\mathbf{s}_i) - f_{j-1}(\mathbf{s}_0)\right\} \end{aligned} \tag{14}$$

with respect to $\lambda_1, \cdots, \lambda_n, m_0, \cdots, m$.
If $f_0(\mathbf{s}) = 1$, then one condition from
$\lambda^T \mathbf{X} = \mathbf{x}^T$ is $\sum_{i=1}^{n} \lambda_i = 1$

Let's write (14) in terms of $2\gamma(.)$. Note that

$$\begin{aligned}
&\left(Y(\mathbf{s}_0) - \sum_{i=1}^{n} \lambda_i Y(\mathbf{s}_i)\right)^2 = \\
&\left(\mathbf{x}^T\beta + \delta(\mathbf{s}_0) - \lambda^T\mathbf{X}\beta - \sum_{i=1}^{n} \lambda_i \delta(\mathbf{s}_i)\right)^2 \\
= \quad &\left(\delta(\mathbf{s}_0) - \sum_{i=1}^{n} \lambda_i \delta(\mathbf{s}_i)\right)^2 \\
= \quad &-\sum_{i=1}^{n}\sum_{j=1}^{n} \lambda_i\lambda_j(\delta(\mathbf{s}_i) - \delta(\mathbf{s}_j))^2/2 \\
+ \quad &2\sum_{i=1}^{n} \lambda_i(\delta(\mathbf{s}_0) - \delta(\mathbf{s}_i))^2/2
\end{aligned}$$

Take expectations and (14) becomes

$$
\begin{aligned}
- & \sum_{i=1}^{n}\sum_{j=1}^{n}\lambda_i\lambda_j\gamma(\delta(\mathbf{s}_i)-\delta(\mathbf{s}_j)) \\
+ & 2\sum_{i=1}^{n}\lambda_i\gamma(\delta(\mathbf{s}_0)-\delta(\mathbf{s}_i)) \\
- & 2\sum_{j=1}^{p+1}m_{j-1}\left\{\sum_{i=1}^{n}\lambda_i f_{j-1}(\mathbf{s}_i)-f_{j-1}(\mathbf{s}_0)\right\}
\end{aligned}
$$

Differentiating (14) and setting the derivative equal to zero gives the system of *universal kriging equations*:

$$\lambda_u = \Gamma_u^{-1}\gamma_u$$

where

$$\begin{aligned}
\lambda_u &= (\lambda_1, \cdots, \lambda_n, m_0, \cdots, m_p)^T \\
\gamma_u &= (\gamma(\mathbf{s}_0 - \mathbf{s}_1), \cdots, \gamma(\mathbf{s}_0 - \mathbf{s}_n), 1, \\
&\quad f_1(\mathbf{s}_0), \cdots, f_p(\mathbf{s}_0))^T
\end{aligned}$$

$$\Gamma_u = \begin{cases}
\gamma(\mathbf{s}_i - \mathbf{s}_j), & i = 1, \cdots, n; j = 1, \cdots, n; \\
f_{j-1-n}(\mathbf{s}_i), & i = 1, \cdots, n; \\
& j = n+1, \cdots, n+p+1; \\
0, & i = n+1, \cdots, n+p+1; \\
& j = n+1, \cdots, n+p+1.
\end{cases}$$

(with $f_0(\mathbf{s}) = 1$).
NOTE: $f_0(\mathbf{s}_0) = 1$ requires an intercept in $\mu(\mathbf{s})$ giving an unbiasedness constraint.

- If instead of intrinsic stationarity, we assume 2^{nd} order stationarity, we get kriging equations in terms of $C(.)$ rather than $\gamma(.)$ and we don't need the $f_0(\mathbf{s}) = 1$ constraint.
- The optimal universal kriging weights also give the minimized MSPE (kriging variance):

$$\begin{aligned}\sigma_k^2(\mathbf{s}_0) &= \lambda_u^T \gamma_u \\ &= 2\sum_{i=1}^{n} \lambda_i \gamma(\mathbf{s}_0 - \mathbf{s}_i) - \sum_{i=1}^{n}\sum_{j=1}^{n} \lambda_i \lambda_j \gamma(\mathbf{s}_i - \mathbf{s}_j)\end{aligned}$$

How to estimate β in Universal Kriging?

We have

$$\mathbf{Y} = \mathbf{X}\beta + \delta.$$

This is similar to a general linear model (non-independent observations) with

$$\begin{aligned} E[\mathbf{Y}] &= \mathbf{X}\beta \\ Var(\mathbf{Y}) &= \Sigma \end{aligned}$$

Now the *generalized least squares* (GLS) estimate of β is

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

To *estimate* β optimally, need to know the Σ matrix, i.e. $\{Cov(Y(\mathbf{s}_i), Y(\mathbf{s}_j)) : 1 \leq i \leq j \leq n\}$.
To *predict* $Y(\mathbf{s}_0)$ optimally, we only need to know $\{Var(Y(\mathbf{s}_i) - Y(\mathbf{s}_j)) : 1 \leq i \leq j \leq n\}$ (as long as $\mathbf{X}$ has a column of 1's).
So, we're all set (to predict) if we know $2\gamma(.)$!

Any problems?

With $\mu(\mathbf{s}) \neq \mu$, estimation of $2\gamma(.)$ more difficult.

$$\begin{aligned} & E[Y(\mathbf{s}_i) - Y(\mathbf{s}_j)]^2 = \\ = \quad & Var(Y(\mathbf{s}_i) - Y(\mathbf{s}_j)) + \{\mu(\mathbf{s}_i) - \mu(\mathbf{s}_j)\}^2 \\ = \quad & 2\gamma(\mathbf{s}_i - \mathbf{s}_j) + \left\{ \sum_{k=1}^{p+1} \beta_{k-1} (f_{k-1}(\mathbf{s}_i) - f_{k-1}(\mathbf{s}_j)) \right\}^2 \end{aligned}$$

Wishful thinking : If β known, we could base an estimate of $2\gamma(.)$ on

$$\delta \equiv \mathbf{Y}(.) - \sum_{k=1}^{p+1} \beta_{k-1} f_{k-1}(.)$$

since

$$E\left\{ [\delta(\mathbf{s}_i) - \delta(\mathbf{s}_j)]^2 \right\} = 2\gamma(\mathbf{s}_i - \mathbf{s}_j)$$

Summary

- For universal kriging, we need $\hat{\gamma}$ and $\hat{\beta}$.
- If β is known we could get $\hat{\gamma}$ based on

$$\delta(.) = Y(.) - \sum_{j=1}^{p+1} f_{j-1}(.)\beta_{j-1}$$

  since $\gamma_Y(.) = \gamma_\delta(.)$, and $\delta(.)$ has a fixed mean (zero since β is known).
- If γ known, we could get $\hat{\beta}$ using GLS, since γ known gives us Σ (the var-cov matrix).

So we're right back where we started!
This circular development leads to some dissatisfaction with universal kriging.

Some comments on estimating γ from residuals

- The *residual process* is

$$\delta(.) = Y(.) - \mathbf{X}\beta$$

- We know that, under the universal kriging assumptions, $\gamma_Y(.) = \gamma_\delta(.)$.
- Problems? β unknown.
- Result: bias in estimation, since with β unknown

$$E\left[(\delta(\mathbf{s}+\mathbf{h}) - \delta(\mathbf{s}))^2\right] = 2\gamma(\mathbf{h}) + \text{bias}$$

- The bias is negative and quadratic in $||\ \mathbf{h}\ ||$.
- So:
  - We *underestimate* γ, hence
  - We *overestimate* spatial correlation, and
  - bias is worst for large lags ($||\ \mathbf{h}\ ||$).

So what's different between universal kriging and removing a trend and ordinary kriging of residuals?
Trend removal and ordinary kriging of residuals:

- Single unbiasedness constraint ($\sum_{i=1}^{n} \lambda_i = 1$).
- One Lagrange multiplier
- $E[\hat{\delta}(\mathbf{s}_0)] = \delta(\mathbf{s}_0)$, but need additional constraints to ensure that $E[\hat{Y}(\mathbf{s}_0)] = Y(\mathbf{s}_0)$.

Universal kriging

- Expanded unbiasedness constraints ($\lambda \mathbf{X} = \mathbf{x}^T$).
- Several ($p+1$) Lagrangian multipliers.
- $E[\hat{Y}(\mathbf{s}_0)] = Y(\mathbf{s}_0)$
- Trend *must* be correct.
- $\delta(.)$ *must* be zero mean.

In practice, the prediction surfaces will be similar, but the prediction standard error surfaces will differ in magnitude.
Compare:
The minimized MSPE from ordinary kriging:

$$\begin{aligned}\delta_k^2(\mathbf{s}_0) &= \lambda_0^T \gamma_0 \\ &= \sum_{i=1}^{n} \lambda_i \gamma(\mathbf{s}_0 - \mathbf{s}_i) + m \\ &= 2\sum_{i=1}^{n} \lambda_i \gamma(\mathbf{s}_0 - \mathbf{s}_i) - \sum_{i=1}^{n}\sum_{j=1}^{n} \lambda_i \lambda_j \gamma(\mathbf{s}_i - \mathbf{s}_j).\end{aligned}$$

The minimized MSPE from universal kriging:

$$\begin{aligned}\sigma_k^2(\mathbf{s}_0) &= \lambda_u^T \gamma_u \\ &= \sum_{i=1}^{n} \lambda_i \gamma(\mathbf{s}_0 - \mathbf{s}_i) + \sum_{j=1}^{p+1} m_{j-1} f_{j-1}(\mathbf{s}_0) \\ &= 2\sum_{i=1}^{n} \lambda_i \gamma(\mathbf{s}_0 - \mathbf{s}_i) - \sum_{i=1}^{n}\sum_{j=1}^{n} \lambda_i \lambda_j \gamma(\mathbf{s}_i - \mathbf{s}_j)\end{aligned}$$

NOTE: Due to expanded unbiasedness constraints, $\lambda_0 \neq \lambda_u$!

Spatial Regression

Generalized least squares (GLS) with *known* covariance matrix

- Model:
  $\mathbf{Y} = \mathbf{X}\beta + \varepsilon, \quad E(\varepsilon) = \mathbf{0}, \quad Var(\varepsilon) = \mathbf{V},$
  where $\mathbf{V}$ is a completely specified positive definite matrix. (e.g. $V_{ij} = \sigma^2 \exp\{-\phi d_{ij}\}$, exponential family with ϕ and σ^2 known)
- GLS estimator of β:
  $\hat{\beta}_{GLS} = (\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-1}\mathbf{X}'\mathbf{V}^{-1}\mathbf{Y}.$
- But in practice ϕ is unknown and consequently $\mathbf{V}$ cannot be completely specified
- A natural solution is to replace ϕ in the evaluation of $\mathbf{V}$ by an estimator ϕ, thereby obtaining $\hat{\mathbf{V}} = \mathbf{V}(\hat{\phi})$ (Estimated GLS)
- EGLS estimator of β:
  $\hat{\beta}_{GLS} = (\mathbf{X}'\hat{\mathbf{V}}^{-1}\mathbf{X})^{-1}\mathbf{X}'\hat{\mathbf{V}}^{-1}\mathbf{Y}$

Classical Procedure

1. Estimate β using ordinary least squares
2. Estimate residuals from this β estimate
3. Calibrate the semi-variogram from the residuals
4. Use the calibration of the semi-variogram to estimate $\mathbf{V}$
5. Re-estimate β using β_{GLS}

Example in geoR

Visit

http://www.leg.ufpr.br/geoR/geoRdoc/vignette/geoRintro/geoRintrose4.html#x5-80004

for a comparison of different classical procedures to estimate parameters in the variogram

Bayesian Kriging

- Basic model:

$$Y(\mathbf{s}) = \beta' \mathbf{X}(\mathbf{s}) + Z(\mathbf{s}) + \varepsilon(\mathbf{s}), \text{ onde} \tag{15}$$

  - $\beta' \mathbf{X}(\mathbf{s})$ is a polynomial trend
  - $Z(\mathbf{s})$ follows a GP

$$\left(Z(\mathbf{s}) \mid \sigma^2, \phi^*\right) \sim GP(0, \sigma^2 \rho(|| \mathbf{s} - \mathbf{s} ||; \phi^*)), \tag{16}$$

    - σ^2 is the partial sill, and variance of $Z(.)$ and
    - $\rho(.; \phi^*)$ is a spatial correlation function which depends on parameters ϕ^* (associated with the range)
  - $\varepsilon(\mathbf{s})$ is white noise such that $\varepsilon(\mathbf{s}) \sim N(0, \tau^2)$ (τ^2 is the nugget effect)

If we marginalize the distribution of **Y** with respect to **Z**, we have that the elements of Σ are given by

$$\Sigma_{ij} = \tau^2 I(i = j) + \sigma^2 \rho(|| \mathbf{s}_i - \mathbf{s}_j ||; \phi^*),$$

where $I(i = j) = 1$, if $i = j$, and 0, otherwise.

Bayesian Kriging - Prior specification

Under the Bayesian paradigm, the parameters in $\theta = (\beta, \sigma^2, \phi^*, \tau^2)'$, are viewed as r.v. $\Rightarrow$ need to assign a prior distribution to θ.
It is common to assume prior independence, such that

- $\beta_i \sim N(0, \sigma^2_\beta)$, $i = 1, \ldots, k$, where σ^2_β is a known constant and k is the dimension of $\mathbf{X}_j$, $j = 1, \cdots, n$
- σ^2 and τ^2 inverse gamma priors with fixed mean and variances
- ϕ^* involves the parameters in the correlation function. Say that an exponential correlation function is used (i.e.$\exp(-\frac{1}{\phi^*}\, dist)$) then it is common to assign a gamma prior distribution to ϕ^* with some reasonably large variance.

  Mean prior specification usually follows the idea of practical range: the prior mean of ϕ^* is such that when the correlation is 0.05 the range is reached at half of the observed maximum distance, that is,
  $E(\phi^*) = \frac{dmax}{6}$, *a priori*

Bayesian kriging - non-informative priors

- $p(\beta)$ can be *flat*
- Without nugget, τ^2 , can't identify both σ^2 and ϕ (Zhang, 2004). With Matérn, can identify the product. So an informative prior on at least one of these parameters
- With τ^2, then ϕ and at least one of σ^2 and τ^2 require informative priors
- Suppose a Matérn covariance function. If the prior on β, σ^2, ϕ is of the form $\frac{\pi(\phi)}{(\sigma^2)^\alpha}$ with $\pi(\cdot)$ uniform, then improper posterior if $\alpha < 2$
- Problem might be visualized using priors of the form $IG(\varepsilon, \varepsilon)$ for σ^2 - "nearly" improper. Safer to use $IG(a, b)$ with $a \geq 1$

Bayesian Kriging - Posterior Distribution

Following Bayes' theorem, the posterior distribution of θ, $\pi(\theta \mid \mathbf{x})$, is proportional to the product of the likelihood function, $f_n(\mathbf{y} \mid \theta)$, by the prior distribution, $\pi(\theta)$, that is,

$$
\begin{aligned}
&\pi(\theta \mid \mathbf{y}) \propto f_n(\mathbf{y} \mid \theta)\pi(\theta) \qquad (17)\\
&\propto \underbrace{\mid \Sigma \mid^{-1/2} \exp\left\{-\frac{1}{2}(\mathbf{y}-\mathbf{X}\beta)'\Sigma^{-1}\,(\mathbf{y}-\mathbf{X}\beta)\right\}}_{f_n(\mathbf{y}\mid\theta)}\\
&\qquad\qquad \times \prod_{i=1}^{k} \exp\left\{-\tfrac{1}{2\sigma_\beta^2}\,\beta_i^2\right\}\\
&\times (\tau^2)^{-(\alpha+1)} \exp\left\{-\tfrac{\beta}{\tau^2}\right\} \times (\sigma^2)^{-(\alpha+1)} \exp\left\{-\tfrac{\beta}{\sigma^2}\right\} \Bigg\} \pi(\theta)\\
&\qquad\qquad \times \phi^{\alpha_\phi - 1} \exp\{-\beta_\phi \phi\}.
\end{aligned}
$$

There is no closed analytical solution to find the posterior distribution $\Rightarrow$ Markov Chain Monte Carlo methods (MCMC).

How to recover the spatial effects $Z(\cdot)$?

- Usually, the interest lies on the surface $\mathbf{Z} \mid \mathbf{y}$ (pattern of spatial adjustment)
- The $Z(s_i)$'s are easily recovered via composition sampling:

$$p(\mathbf{z} \mid \mathbf{y}) = \int p(\mathbf{z} \mid \theta, \mathbf{y}) p(\theta \mid \mathbf{y}) d\theta$$

- Note that

$$p(\mathbf{z} \mid \theta, \mathbf{y}) \propto f(\mathbf{y} \mid \mathbf{z}, \beta, \tau^2) p(\mathbf{z} \mid \sigma^2, \phi)$$

  is a multivariate normal distribution, resulting in easy composition sampling, in fact 1-1 with posterior samples of θ

Bayesian Kriging - spatial interpolation

Aim: prediciton of $\mathbf{Y}_u = (Y(\mathbf{s}_{u1}), \cdots, Y(\mathbf{s}_{ur}))'$ at new sites $\mathbf{s}_{u1}, \cdots, \mathbf{s}_{ur}$ with associated covariates $\mathbf{x}_0 = (\mathbf{x}(\mathbf{s}_{u1}), \cdots, \mathbf{x}(\mathbf{s}_{ur})) \Rightarrow$ Predictive distribution $(\mathbf{Y}_u \mid \mathbf{y}, \mathbf{x}, \mathbf{x}_0)$,

$$
\begin{aligned}
p(\mathbf{Y}_u \mid \mathbf{y}, \mathbf{x}, \mathbf{x}_0) &= \int_\theta p(\mathbf{Y}_u \mid \mathbf{y}, \mathbf{x}, \mathbf{x}_0, \theta) \underbrace{\pi(\theta \mid \mathbf{y}, \mathbf{x}, \mathbf{x}_0)}_{\text{posterior of } \theta} d\theta \\
&= E_{\pi(\theta \mid Y)} [p(\mathbf{Y}_u \mid \mathbf{Y}, \theta)]. \qquad (18)
\end{aligned}
$$

Bayesian Kriging - spatial interpolation

The joint distribution of $\mathbf{Y}$ and $\mathbf{Y}_u$ is given by

$$\begin{pmatrix} \mathbf{Y}_u \\ \mathbf{Y} \end{pmatrix} \mid \theta \Big) \sim N_{n+r} \left(\begin{pmatrix} \mu_u \\ \mu \end{pmatrix}; \begin{pmatrix} \Sigma_u & \Psi' \\ \Psi & \Sigma \end{pmatrix} \right), \tag{19}$$

- μ_u r-dimensional vector - mean of the unobserved sites
- Σ_u r-dimensional covariance matrix - covariances among unobserved sites
- Ψ, $n \times r$, covariance between i-th observed site and j-th unobserved site, $i = 1, \cdots, n$ and $j = 1, \cdots, r$.

Bayesian Kriging - spatial interpolation

From the properties of the partition of the multivariate normal distribution:

$$\left(\mathbf{Y}_u \mid \mathbf{Y}, \theta\right) \sim N_k(\mu_u + \Psi' \Sigma^{-1} (\mathbf{Y} - \mu); \Sigma_u - \Psi' \Sigma^{-1} \Psi).$$

Easy Monte Carlo estimate using composition with Gibbs draws $\theta^{(1)}, \ldots, \theta^{(L)}$: for each $\theta^{(l)}$ drawn from $\pi(\theta \mid Y)$, draw $\mathbf{Y}_u$ from $p(\mathbf{Y}_u \mid \mathbf{Y}, \theta^{(l)})$

$$p(\mathbf{Y}_u \mid \mathbf{Y}) \approx \frac{1}{L} \sum_{l=1}^{L} p(\mathbf{Y}_u \mid \theta^l)$$

Spatial regression using geoR

For an example of a spatial regression
see document RainfallParana.pdf

Bayesian spatial regression

For an example of an implementation of a Bayesian spatial regression using `Stan` see files fit_modelRain.r and Rainfall-1.stan

For an example of an implementation of a Bayesian spatial regression using `spBayes` see file fit_modelRain_spBayes.r

Prior sensitivity of σ^2 and ϕ

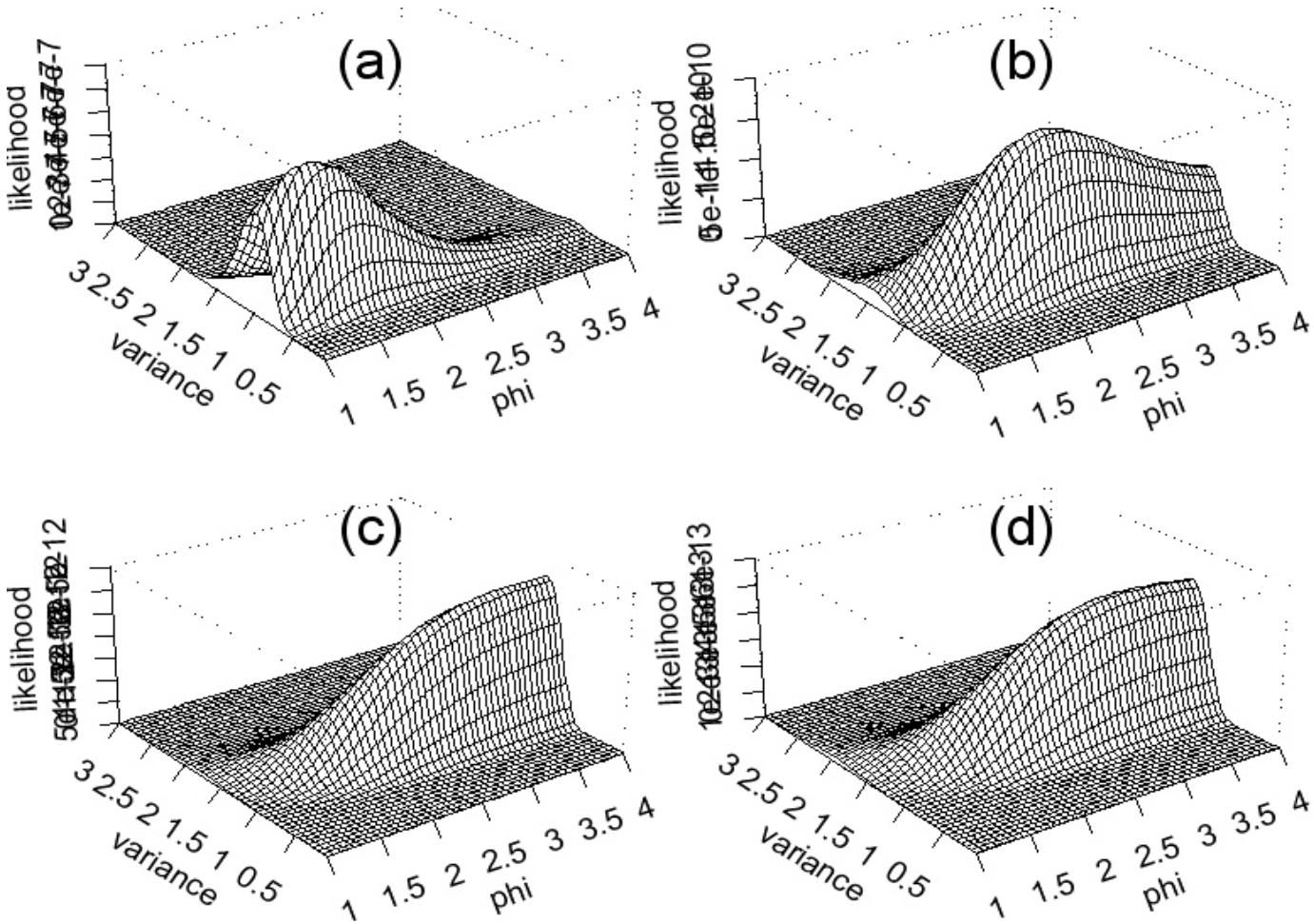


Figure: Likelihood for ϕ and σ^2 for samples (a) 1, (b) 3, (c) 10 and (d) 13. See Schmidt et al (2008) for details.

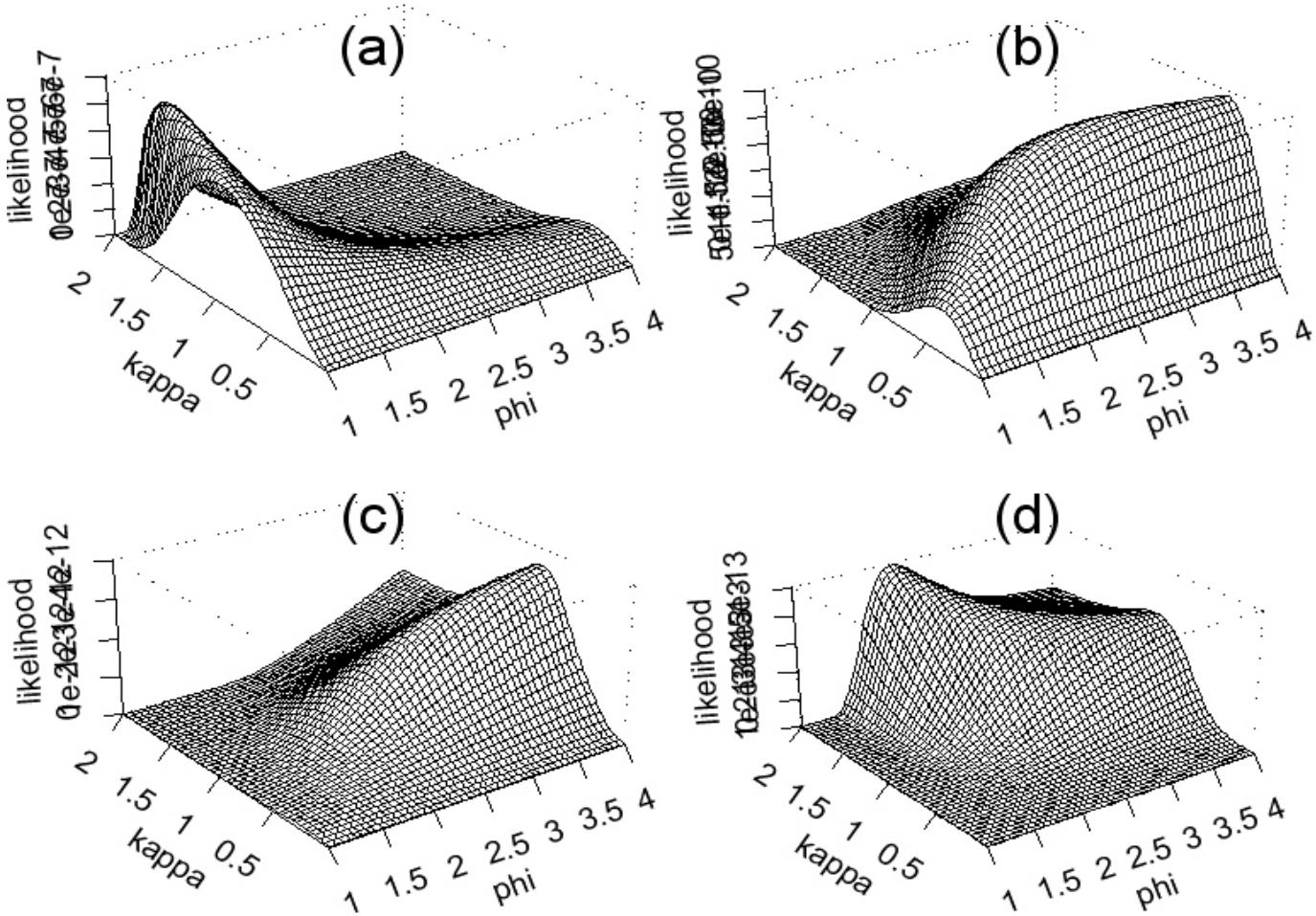


Figure: Likelihood for ϕ and κ (a) 1, (b) 3, (c) 10 and (d) 13. See Schmidt et al (2008) for details.

Reference prior for β, ϕ, σ^2:

(Berger et. al. JASA (2001))

$p^R(\beta, \sigma^2, \phi)$, is of the form

$$p^R(\beta, \sigma^2, \phi) \propto \frac{p(\phi)}{(\sigma^2)^a}, a \in I\!R, \text{ com}$$

$$a = 1 \text{ e } p(\phi) \quad \propto \quad \left\{ tr[W_\phi^2] - \frac{1}{n-p}(tr[W_\phi])^2 \right\}^{1/2},$$

where

$$\begin{aligned} W_\phi &= \left(\frac{\partial}{\partial \phi} \Sigma_\phi \right) \Sigma_\phi^{-1} P_\phi^\Sigma \text{ e} \\ P_\phi^\Sigma &= I - X(X' \Sigma_\phi^{-1} X)^{-1} X' \Sigma_\phi^{-1}; \end{aligned}$$

$(\partial/\partial\phi)\Sigma_\phi$ denotes the matrix obtained from the differentiation of Σ_ϕ element by element.

Spatial generalized linear models

- Some data sets preclude Gaussian modeling; $Y(\mathbf{s})$
- Example: $Y(\mathbf{s})$ is a binary or count variable
  - Presence/absence of a species at a location; abundance of a species at a location
  - precipitation or deposition was measurable or not
  - number of insurance claims by residents of a single family home at $\mathbf{s}$
  - Land use classification at a location (not ordinal)
- replace Gaussian likelihood by an appropriate exponential family member if possible
- See Diggle, Tawn e Moyeed (JRSS Series C, 1998)

Spatial generalized linear models

- First stage: $Y(\mathbf{s}_i)$ are conditionally independent given β and $z(\mathbf{s}_i)$, with $f(y(\mathbf{s}_i) \mid \beta, z(\mathbf{s}_i), \gamma)$ an appropriate non-Gaussian likelihood such that

$$g(E(Y(\mathbf{s}_i)) = \eta(\mathbf{s}_i) = \mathbf{x}(\mathbf{s}_i)^T \beta + z(\mathbf{s}_i),$$

  where $\eta(\cdot)$ is a canonical link function (e.g. logit) and γ is a dispersion parameter
- Second stage: Model $z(\mathbf{s})$ as a Gaussian process

$$\mathbf{Z} \sim N(\mathbf{0}, \sigma^2 R(\phi))$$

- Third stage: Priors and hyperpriors
- Lose conjugacy between first and second stage; not sensible to add a pure error term

Spatial generalized linear models

- Spatial random effects in the transformed mean with continuous covariates encourages the means of spatial variables at proximate locations to be close to each other
- Marginal spatial dependence is induced between, say, $Y(\mathbf{s})$ and $Y(\mathbf{s}')$, but the observed $Y(\mathbf{s})$ and $Y(\mathbf{s}')$ need not be close. No smoothness in $Y(\mathbf{s})$ surface
- Our second stage modeling is attractive for spatial explanation in the mean
- First stage modeling is better for encouraging proximate observations to be close
- Note that this approach offers a valid joint distribution for the $Y(\mathbf{s}_i)$, but not a spatial process model; we need not achieve a consistent stochastic process for the uncountable collection of $Y(\mathbf{s})$ values

GPs in INLA

The SPDE Approach (Rue and Lindgren, Journal of Stat. Soft., 2015)

- The stochastic partial differential equation (SPDE) approach was proposed by Lingdren *et al.* (JRSS B, 2011) and is implemented in `R-INLA`.
- The idea is to carry out the computations by approximating the full set of spatial random functions with weighted sums of simple basis functions, which allows us to hold on to the continuous interpretation of space, while the computational algorithms only see discrete structures with Markov properties

Gaussian Markov Random Fields (GMRF)

Taken from Rue and Held (2005) Gaussian Markov random fields

Definition (informal): Let $\mathbf{x} = (x_1, \cdots, x_n)'$ have a normal distribution with mean μ and covariance matrix Σ. Define the labelled graph $\mathcal{G} = (\vartheta, \varepsilon)$ where $\vartheta = \{1, 2, \cdots, n\}$ and ε be such that there is no edge between nodes i and j iff $x_i \perp x_j \mid \mathbf{x}_{-ij}$. We say that $\mathbf{x}$ is a GMRF wrt $\mathcal{G}$.

Theorem: Let $\mathbf{x}$ be normal distributed with mean μ and precision matrix $\mathbf{Q} > 0$. Then for $i \neq j$,

$$x_i \perp x_j \mid \mathbf{x}_{-ij} \leftrightarrow Q_{ij} = 0$$

Definition (formal): A random vector $\mathbf{x} = (x_1, cdots, x_n)' \in \mathbb{R}^n$ is called a GMRF wrt a labelled graph $\mathcal{G} = (\vartheta, \varepsilon)$ with mean μ and precision matrix $\mathbf{Q} > 0$, iff its density has the form

$$\pi(\mathbf{x}) = (2\pi)^{-n/2} |\mathbf{Q}|^{1/2} \exp\left(-\frac{1}{2} (\mathbf{x} - \mu)' \mathbf{Q} (\mathbf{x} - \mu \right)$$

and $Q_{ij} \neq 0 \leftrightarrow \{i, j\} \in \varepsilon \; \forall i \neq j$.

Gaussian Markov Random Fields

- If $\mathbf{Q}$ is a completely dense matrix then $\mathscr{G}$ is fully connected
- Any normal distribution with SPD covariance matrix is also a GMRF and vice versa
- Teorema: Let $\mathbf{x}$ be a GMRF wrt $\mathscr{G} = (\vartheta, \varepsilon)$ with mean μ and precision matrix $\mathbf{Q} > 0$, then

$$
\begin{aligned}
E(x_i \mid \mathbf{x}_i) &= \mu_i - \frac{1}{Q_{ii}} \sum_{j:j\sim i} Q_{ij}(x_j - \mu_j) \\
Prec(x_i \mid \mathbf{x}_{-i}) &= Q_{ii} \text{ and,} \\
Corr(x_i \mid \mathbf{x}_{-ij}) &= -\frac{Q_{ij}}{\sqrt{Q_{ii}Q_{jj}}}, \; i \neq j
\end{aligned}
$$

With a proper scaling, off-diagonal elements of $\mathbf{Q}$ provide information about the conditional correlation between x_i and x_j , given $\mathbf{x}_{ij}$

GPs in INLA

The SPDE Approach (Rue and Lindgren, Journal of Stat. Soft., 2015)

- Assume there is a spatially continuous variable that can be modelled using a Gaussian Markov Random Field (GMRF) with Matérn covariance functions,

$$Cov(S(\mathbf{x}_i), S(\mathbf{x}_j)) = \frac{\sigma^2}{2^{\nu-1}\Gamma(\nu)}(\kappa(||\mathbf{x}_i - \mathbf{x}_j||)^{\nu} K_{\nu}(\kappa||\mathbf{x}_i - \mathbf{x}_j||),$$

  σ^2 is the marginal variance; $K_{\nu}(\cdot)$ is the modified Bessel function of second kind and order $\nu > 0$; the integer part of ν determines the smoothness of the field

- Whittle (1963) shows that a GRF with Matérn covariance matrix can be represented as a solution of the following continuous domain SPDE:

$$(\kappa - \Delta)^{\alpha/2}(\tau Z(\mathbf{x})) = \mathscr{W}(\mathbf{x}),$$

  where $Z(\mathbf{x})$ represents a GRF, $\mathscr{W}(\mathbf{x})$ is a Gaussian spatial white noise process.

- Parameter α controls the smoothness of the GRF, τ controls its variance and $\kappa > 0$ is a scale parameter

GPs in INLA

The SPDE Approach (Rue and Lindgren, Journal of Stat. Soft., 2015)

- The parameters of the Matérn covariance function and the SPDE are related as follows

$$\nu = \alpha - \frac{d}{2} \text{ and } \sigma^2 = \frac{\Gamma(\nu)}{\Gamma(\alpha)(4\pi)^{d/2}\kappa^{2\nu}\tau^2}$$

- When $d = 2$ and $\nu = 1/2$, we have the exponential covariance function and $\alpha = 3/2$. In `R-INLA` the default is $\alpha = 2$, although options within $0 \leq \alpha \leq 2$ are also available

- The Finite Element method is used to find an approximate solution to the SPDE. The method involves dividing the spatial domain into a set of non-intersecting triangles, creating a triangulated mesh with n nodes and n basis functions

- They take the value 1 at vertex k, and 0 at all other vertices

GPs in INLA

The SPDE Approach (Rue and Lindgren, Journal of Stat. Soft., 2015)

- Then, the continuously indexed Gaussian field x is represented as a discretely indexed Gaussian Markov random field (GMRF) by a sum of basis functions defined on the triangulated mesh

$$S(\mathbf{x}) = \sum_{k=1}^{n} \psi_k(\mathbf{x}) z_k$$

  where n is the number of vertices of the triangulation, $\psi_k(\cdot)$ represents the piecewise linear basis functions, and $\{z_k\}$ denote zero-mean Gaussian distributed weights

- The joint distribution of the weight vector is assigned a Gaussian distribution represented as $\mathbf{Z} = (Z_1, \cdots, Z_n)' \sim N(\mathbf{0}, \mathbf{Q}^{-1}(\tau, \kappa))$

- This distribution approximates the solution $z(\mathbf{x})$ of the SPDE at the mesh nodes

- The basis functions transform the approximation $z(\mathbf{x})$ from the mesh nodes to the other spatial locations of interest

GPs in INLA

The SPDE Approach (Rue and Lindgren, Journal of Stat. Soft., 2015)

- The SPDE approach can be implemented with `R-INLA` by creating a projection matrix $\mathbf{A}$ that projects the GRF from the observations to the vertices of the triangulated mesh
- The projection matrix $\mathbf{A}$ has a number of rows equal to the number of observations, and a number of columns equal to the number of vertices of the mesh.
- Each row i of $\mathbf{A}$ corresponds to an observation at location $\mathbf{x}_i$, and may have up to three non-zero values in the columns that correspond to the vertices of the triangle containing the location
- If $\mathbf{x}_i$ lies within the triangle, these values are equal to the barycentric coordinates $\rightarrow$ they are proportional to the areas of each of the three subtriangles formed by the location $\mathbf{x}_i$ and the triangle's vertices, and they sum to 1
- If $\mathbf{x}_i$ coincides with a vertex of the triangle, row i has just one non-zero value equal to 1 in the column that corresponds to that vertex

Nearest Neighbor Gaussian Processes