TMA4250 Spatial Statistics
Project 1: Random Fields and Gaussian Random Fields

Spring 2023

Introduction

This project contains problems related to random fields (RFs) and Gaussian random fields (GRFs). We recommend using `R` for solving the problems, and relevant functions can be found in the `R` libraries `geoR`, `akima`, `fields`, and `ggplot2`.

Problem 1: GRFs – model characteristics

Let X be a stationary GRF on $\mathcal{D} = [1, 50] \subset \mathbb{R}$, and assume that

$$\begin{aligned} \mathrm{E}[X(s)] &= \mu = 0, \quad s \in \mathcal{D}, \\ \mathrm{Var}[X(s)] &= \sigma^2, \quad s \in \mathcal{D}, \\ \mathrm{Corr}[X(s), X(s')] &= \rho(||s - s'||), \quad s, s' \in \mathcal{D}, \end{aligned}$$

where ρ is the correlation function of X. Discretize $\mathcal{D}$ by a regular grid $\tilde{\mathcal{D}} = \{1, 2, \ldots, 50\}$ and let $\boldsymbol{X} = (X(1), \ldots, X(50))^{\mathrm{T}}$ be the discretization of X on $\tilde{\mathcal{D}}$.

Let the spatial correlation function ρ, be either powered exponential with power $\alpha \in \{1, 1.9\}$ and spatial scale $a = 10$, or Matérn with smoothness $\nu \in \{1, 3\}$ and range $a = 20$. Let the marginal variance take the values $\sigma^2 \in \{1, 5\}$. Use the definition of Matérn that we used in lectures.

a) The correlation function must be a positive semi-definite function, specify this requirement mathematically. Explain why this requirement is necessary.

Display the four correlation functions specified above for distances $h \in [0, \infty)$. Discuss the features of the spatial correlation function which are crucial for the associated GRF, and the relations between the variance and correlation function with the variogram function. Display the eight associated semi-variogram functions $\gamma : [0, \infty) \to [0, \infty)$.

NB: Think carefully about how many figures you use and how they are organized to best convey your answer.

Use the functions `cov.spatial` and/or `matern`. Compare the definitions of Matérn used by these functions to our definition (using e.g., `?cov.spatial`).

b) Determine the distribution of $\boldsymbol{X}$ and specify how to calculate the parameters of the distribution.

Simulate four realizations of X on $\tilde{\mathcal{D}}$ for each of the eight different sets of model parameters defined above and present them in eight displays with four realizations in each. Discuss the relationship between the realizations and the model parameters.

We plan to observe Y_1, Y_2 and Y_3 at locations $s_1 = 10$, $s_2 = 25$ and $s_3 = 30$, respectively. The observation model is given by

$$Y_i = X(s_i) + \epsilon_i, \quad i = 1, 2, 3,$$

where $\epsilon_1, \epsilon_2, \epsilon_3 \overset{\text{iid}}{\sim} \mathcal{N}(0, \sigma_{\text{N}}^2)$ and independent of X.

c) Determine the distribution of $\boldsymbol{Y} = (Y_1, Y_2, Y_3)^{\text{T}}$ and specify how to calculate the parameters of the distribution.

Choose one the four covariance model with $\sigma^2 = 5$, and consider the simulated realizations in **b)**. Select one realization, and use the exact values at locations s_1, s_2 and s_3 as the observation $\boldsymbol{y}$. Let the observation error variance take the values $\sigma_{\text{N}}^2 \in \{0, 0.25\}$.

d) Determine the distribution of $\boldsymbol{X}$ given the observation $\boldsymbol{Y} = \boldsymbol{y}$ and specify how to calculate the parameters of the distribution.

In this problem, use the covariance model chosen in **c)**. Use the two error variances σ_{N}^2 listed above and compute the two predictions for X on $\tilde{\mathcal{D}}$ given $\boldsymbol{Y} = \boldsymbol{y}$ with associated 90% prediction intervals. Present the results in two displays (one for each value of σ_{N}^2).

Discuss these two displays and the relation between the model parameters and the predictions with prediction intervals. Inspect carefully the appearance of the predictions at the observation locations.

e) Simulate 100 realizations of X on $\tilde{\mathcal{D}}$ given $\boldsymbol{Y} = \boldsymbol{y}$ for each value of σ_{N}^2, and estimate *empirically* the prediction and associated 90% prediction intervals.

Present the simulated realizations in two displays, one for each value of σ_{N}^2, and plot the corresponding empirically estimated predictions and prediction intervals in the same displays.

Discuss the relation between the model parameters and the realizations, and discuss the relation between the analytically and empirically obtained predictions with prediction intervals.

f) Define

$$A = \sum_{s \in \tilde{\mathcal{D}}} \mathbb{I}(X(s) > 2)(X(s) - 2),$$

where $\mathbb{I}$ is the indicator function. This is an approximation of the area under X and above level 2.

Use the 100 realizations of X on $\tilde{\mathcal{D}}$ given $\boldsymbol{Y} = \boldsymbol{y}$ and $\sigma_{\mathrm{N}}^2 = 0$ to provide a prediction $\hat{A}$ with associated prediction variance.

Continue to assume that $\sigma_{\mathrm{N}}^2 = 0$. An alternative predictor for this area is

$$\tilde{A} = \sum_{s \in \tilde{\mathcal{D}}} \mathbb{I}(\hat{X}(s) > 2)(\hat{X}(s) - 2),$$

where $\hat{X}(s)$ is the simple Kriging predictor at location s. Calculate this prediction.

Consider the two predictions and the prediction variance of the former. Compare the predictions and use Jensen's inequality to explain why one expects $\hat{A} \geq \tilde{A}$.

g) Present a short summary of the experiences you have made on evaluating the model characteristics.

Problem 2: GRF - real data

Consider the observations of terrain elevation, available in the file `topo.dat`. The 52 observations are located in the domain $\mathcal{D} = [0, 315]^2 \subset \mathbb{R}^2$. Let X be a GRF on $\mathcal{D}$, and let the vector of exact observations be $\boldsymbol{X} = (X(\boldsymbol{s}_1), \ldots, X(\boldsymbol{s}_{52}))^{\mathrm{T}}$.

a) Display the observations in various ways. Is a stationary GRF a suitable model for the terrain elevation in domain $\mathcal{D}$?

The functions `interp`, `contour` and `image.plot` in the R libraries `akima` and `fields` may be useful. `ggplot2` can be used for scatter plots with colour.

Assume the GRF X is modelled by

$$\begin{aligned} \mathrm{E}[X(\boldsymbol{s})] &= \boldsymbol{g}(\boldsymbol{s})^{\mathrm{T}}\boldsymbol{\beta}, \quad \boldsymbol{s} \in \mathcal{D}, \\ \mathrm{Var}[X(\boldsymbol{s})] &= \sigma^2, \quad \boldsymbol{s} \in \mathcal{D}, \\ \mathrm{Corr}[X(\boldsymbol{s}), X(\boldsymbol{s}')\} &= \rho(||\boldsymbol{s} - \boldsymbol{s}'||), \quad \boldsymbol{s}, \boldsymbol{s}' \in \mathcal{D}, \end{aligned}$$

where $\boldsymbol{g}(\boldsymbol{s}) = (1, g_2(\boldsymbol{s}), ..., g_{n_g}(\boldsymbol{s}))^{\mathrm{T}}$ is a n_g-vector of known explanatory spatial variables for $\boldsymbol{s} \in \mathcal{D}$, and $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_{n_g})^T$ is a n_g-vector of unknown parameters. Moreover, let the marginal variance be $\sigma^2 = 2500$ and the correlation function be $\rho(h) = \exp(-(0.01h)^{1.5})$, $h \in [0, \infty)$.

b) Develop the expressions for the minimization problem to be solved for the universal kriging predictor and the associated prediction variance at an arbitrary location $\boldsymbol{s}_0 \in \mathcal{D}$. The actual optimization need not be solved. Would you expect that the value of σ^2 needs to change for different parameterizations of the expectation function?

c) Consider the case with $\mathrm{E}[X(\boldsymbol{s})] = \beta_1$, $\boldsymbol{s} \in \mathcal{D}$; the so-called ordinary Kriging model.

Use a regular grid $\tilde{\mathcal{D}} = \{1, 2, \ldots, 315\}^2$ of $\mathcal{D}$, and calculate the Kriging predictor $\hat{X}(\boldsymbol{s})$, $\boldsymbol{s} \in \tilde{\mathcal{D}}$, with associated prediction variance $\sigma^2_{\hat{X}}(\boldsymbol{s})$, $\boldsymbol{s} \in \tilde{\mathcal{D}}$. Display the results and comment on them.

Use the function `krige.conv` and the arguments `trend.d` and `trend.l` in `krige.control` to specify the form of the expectation function - the function `expand.grid` may also be useful.

d) Let the reference variable $\boldsymbol{s} \in \mathcal{D} \subset \mathbb{R}^2$ be denoted $\boldsymbol{s} = (s_1, s_2)$, set $n_g = 6$, and define the vector-valued function of known polynomial functions $\boldsymbol{g}$ to be all polynomials $s_1^k s_2^l$ for $(k, l) \in \{(0,0), (1,0), (0,1), (1,1), (2,0), (0,2)\}$.

Specify the resulting n_g-dimensional vector $\boldsymbol{g}(\boldsymbol{s})$ and the expected value of $X(\boldsymbol{s})$ as functions of $\boldsymbol{s}$.

For the regular grid $\tilde{\mathcal{D}} = \{1, 2, \ldots, 315\}^2$ of $\mathcal{D}$, calculate the universal Kriging predictor $\hat{X}(\boldsymbol{s})$, $\boldsymbol{s} \in \tilde{\mathcal{D}}$, and its associated prediction variance, $\sigma^2_{\hat{X}}(\boldsymbol{s})$, $\boldsymbol{s} \in \tilde{\mathcal{D}}$. Display the results and comment on them.

Use the function **krige.conv** and the arguments **trend.d** and **trend.l** in **krige.control** to specify the form of the expectation function - the function **expand.grid** may also be useful.

e) Use the ordinary Kriging predictor with associated prediction variance from **c)**, and consider location $\boldsymbol{s}_0 = (100, 100)^{\mathrm{T}}$. Calculate the probability for the elevation to be higher than 850 m at this location. Further, calculate the elevation for which it is 0.90 probability that the true elevation is below it.

f) Present a short summary of the experiences you have made on evaluating the real data.

Problem 3: Parameter estimation

Consider the stationary GRF $\{X(\boldsymbol{s}); \boldsymbol{s} \in \mathcal{D} = [1, 30]^2 \subset \mathbb{R}^2\}$ with

$$\begin{aligned} \mathrm{E}[X(\boldsymbol{s})] &= \mu = 0, \quad \boldsymbol{s} \in \mathcal{D}, \\ \mathrm{Var}[X(\boldsymbol{s})] &= \sigma^2, \quad \boldsymbol{s} \in \mathcal{D}, \\ \mathrm{Corr}[X(\boldsymbol{s}), X(\boldsymbol{s}')] &= \exp(-||\boldsymbol{s} - \boldsymbol{s}'||/a), \quad \boldsymbol{s}, \boldsymbol{s}' \in \mathcal{D}. \end{aligned}$$

a) Let $\tilde{\mathcal{D}} = \{1, 2, \ldots, 30\}^2$ be a regular grid of $\mathcal{D}$. Set marginal variance to $\sigma^2 = 2$ and spatial scale to $a = 3$, and generate one realization of X on $\tilde{\mathcal{D}}$ and display it.

b) Compute the empirical semi-variogram based on **one** exact observation of all locations in $\tilde{\mathcal{D}}$, and display the estimate jointly with the true semi-variogram function. Comment on the result, particularly the precision/uncertainty of the estimates due to observing on a bounded domain $\mathcal{D}$.

You can use the function `variog` in `GeoR`.

c) Repeat point **a)** and **b)** three times. Comment on the results.

d) Select 36 locations uniformly at random in the grid $\tilde{\mathcal{D}}$. Compute the empirical semi-variogram estimate based on the corresponding 36 exact observations. Display the estimate jointly with the true semi-variogram function, and comment on the results.

Consider the model parameters, σ^2 and a, to be unknown. Estimate the parameters by a maximum likelihood criterion based on an exact observation of X at 1) all locations in $\tilde{\mathcal{D}}$ and 2) at the 36 locations. Display the corresponding estimated semi-variogram functions jointly with the true semi-variogram function, and comment on the result.

You can use the function likfit in GeoR.

Discuss and compare all the results above.

e) Repeat the procedure in **d)** with 9, 64 and 100 locations selected uniformly at random. Present the estimates jointly with the true semi-variogram function in separate displays, and comment on the results.

f) Present a short summary of the experiences you have made on parameter estimation.