TMA4250 Spatial Statistics
Assignment 1: Continuous Random Fields

IMF/NTNU

January 2017

Introduction

This assignment contains problems related to geostatistical processes and Gaussian random fields (GRF). We recommend using R for solving the problems, and relevant functions can be found in the R libraries geoR, akima and fields.

Problem 1: One-dimensional Gaussian Random Fields (GRF)

Let $\{Y(s) : s \in [1, 50] \subset \mathcal{R}^1\}$ denote the true temperature (°C) along a 50 km long road. We assume that the temperature along the road can be modeled as a stationary 1D GRF with the following properties:

$$
\begin{aligned}
E\{Y(s)\} &= \mu = 20 \\
Var\{Y(s)\} &= \sigma_1^2 \\
Corr\{Y(s), Y(s+h)\} &= \rho_Y(h) = C_Y(h)/\sigma_1^2,
\end{aligned}
$$

where $C_Y(h)$ is a covariance function and where the micro-scale variance is zero ($\sigma_0^2 = 0$). Let $\mathcal{D} = [1, 50]$ be discretised in $\mathcal{L}_{\mathcal{D}} \in \{1, 2, \ldots, 50\}$ and define the discretised GRF $\{Y(s); s \in \mathcal{L}_{\mathcal{D}}\}$.

a) Assume that the covariance function $C_Y(h)$ is either Matérn with smoothness parameter 1 ($\nu = 1$) or exponential. Display the exponential and the Matérn correlation function on $\mathcal{D}$ for different ranges θ_1 between 5 and 25. You can use the functions cov.spatial() and/or Matern() from the libraries geoR and fields.

What does the range tell us about the GRF?

Develop the relation between the correlation function and the variogram function.

b) Simulate some realisations of the GRF on $\mathcal{L}_{\mathcal{D}}$ for different covariance functions. Choose some variances σ_1^2 and ranges θ_1, and show/explain how your choice of parameters affects the resulting simulated GRF.

Assume that we measure the temperature $Y(s)$ at locations $s^* \in \{10, 25, 30\}$ in $\mathcal{L}_{\mathcal{D}}$. The observed temperatures at these locations are noisy versions of the true, underlying temperatures, and we write the observations as

$$Z(s^*) = Y(s^*) + \epsilon(s^*) \qquad\qquad s^* \in \{10, 25, 30\} \tag{1}$$

where the measurement errors $\epsilon(\cdot)$ are independent and identically distributed as $\mathcal{N}(0, \sigma_\epsilon^2)$. Further, assume that $Y(s^*)$ and $\epsilon(s^* + h)$ are independent for all h.

c) Write down the data model and the process model for temperature.

d) Consider the simulations from **b)** and choose a realisation that could be a realistic representation of the temperature differences along a 50 km long road. Assume that $\sigma_\epsilon^2 = 1$, and use the simulated values at $s^* \in \{10, 25, 30\}$ to create a set of observations (1).

Specify the pdf for the conditional discretised GRF given the observations, i.e find the distribution $[\boldsymbol{Y}|\boldsymbol{Z}]$ where $\boldsymbol{Y} = (Y(1), ..., Y(50))'$, and where $\boldsymbol{Z} = (Z(10), Z(25), Z(30))'$. Compute the expected values $E\{Y(s)|\boldsymbol{Z}\}$ and the variances values $Var\{Y(s)|\boldsymbol{Z}\}$ for each s in $\mathcal{L}_{\mathcal{D}}$, and display the results as an expectation function with associated 2σ intervals on either side.

e) Simulate 50 realisations of the conditional discretised GRF $[\boldsymbol{Y}|\boldsymbol{Z}]$. Display the realisations in one figure. For each $s \in \mathcal{L}_{\mathcal{D}}$ compute the average and the empirical variance based on the 50 realisations. Display the results as an average function with associated estimated 2σ intervals on either side. Compare the results with the results in **d)** and comment.

Problem 2: Spatial Prediction by Kriging

This problem is based on observations of terrain elevation which are available on the web site of the course in the file topo.dat. The 52 observations are in a domain $\mathcal{D} = (0, 315) \times (0, 315) \subset \mathcal{R}^2$.

a) Display the observations in various ways. The functions interp(), contour() and image.plot() in the R libraries akima and fields may be useful.

Comment the results.

Let the terrain elevation over $\mathcal{D}$ be modeled by the GRF $\{Y(\boldsymbol{s}); \boldsymbol{s} \in \mathcal{D} \subset \mathcal{R}^2\}$

with

$$\begin{aligned} E\{Y(\boldsymbol{s})\} &= \quad \boldsymbol{x}(\boldsymbol{s})'\boldsymbol{\beta} \\ Cov\{Y(\boldsymbol{s}), Y(\boldsymbol{s}+\boldsymbol{h})\} &= \quad C_Y(||\boldsymbol{h}||) \end{aligned} ,$$

where $\boldsymbol{x(s)} = (x_1(\boldsymbol{s}), ..., x_p(\boldsymbol{s}))'$ is a p-dimensional vector of known functions of $\boldsymbol{s} \in \mathcal{D}$, and $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)'$ is a vector of unknown weights.

Let the vector of observations be denoted $\mathbf{Z} = (Z(\boldsymbol{s_1}), \ldots, Z(\boldsymbol{s_{52}}))'$.

b) Show how you derive the the universal kriging predictor and prediction variance at an arbitrary location $\boldsymbol{s_0} \in \mathcal{D}$. (You don't need to solve the resulting optimisation problem.)

Let the reference variable $\boldsymbol{s} \in \mathcal{D} \subset \mathcal{R}^2$ be denoted $\boldsymbol{s} = (s_v, s_h)$, set $p = 6$ and define the set of known functions $\boldsymbol{x}(\boldsymbol{s})$ to be all polynomials $s_v^k s_h^l$ for $(k, l) \in \{(0, 0), (1, 0), (0, 1), (1, 1), (2, 0), (0, 2)\}$. Further, let the covariance function $C_Y(||\boldsymbol{h}||)$ be of exponential form with variance 2500 and range parameter 100.

c) Write down the resulting p-dimensional vector $\boldsymbol{x(s)}$ and the expected value of $Y(\boldsymbol{s})$. How can we interpret this model?

Use the function krige.conv() to compute the universal kriging surface with associated kriging variance in a (316×316) grid covering $\mathcal{D}$.

Hint : Change trend.d and trend.l in krige.conv() to specify the form of $E\{Y(\boldsymbol{s})\}$. The function expand.grid() may also be useful.

Display the results and comment.

d) Consider grid node $\boldsymbol{s_0} = (100, 100)$. What is the probability that the elevation is larger than 700 m at this location? Further, compute the elevation for which it is a 90 % probability that the true elevation is below it.

e) Add noise to the elevation data in `topo.dat`. You can assume that the noise is independent of the observations and distributed as $\mathcal{N}(0, \sigma_\epsilon^2)$. Repeat the procedure in **c)** with the noisy dataset, first with $\sigma_\epsilon^2 = 5$, then with $\sigma_\epsilon^2 = 15$. Compare the results and comment/explain.

Problem 3: Parameter estimation

Assume that the temperature (°C) in a region of size 30 km × 30 km can be modeled as a stationary GRF $\{Y(\boldsymbol{s}); \boldsymbol{s} \in \mathcal{D} \subset \mathcal{R}^2\}$ with $\mathcal{D} \in [(1, 30), (1, 30)]$, and with

$$
\begin{aligned}
E\{Y(\boldsymbol{s})\} &= \mu = 12 \\
Var\{Y(\boldsymbol{s})\} &= \sigma_1^2 = 2 \\
Cov\{Y(\boldsymbol{s}), Y(\boldsymbol{s}+\boldsymbol{h})\} &= \sigma_1^2 \exp\left\{-\frac{||h||}{\theta_1}\right\} \\
&= 2\exp\left\{-\frac{||h||}{15}\right\}.
\end{aligned}
$$

Discretise $\mathcal{D}$ into a grid $\mathcal{L}_{\mathcal{D}}$ of size 30×30.

a) Describe how the temperature in the study region is distributed based on the parameter values: What is the interpretation of the parameters μ, σ_1^2 and θ_1?

Specify the requirements for a valid spatial covariance function, and use R to compute the covariance matrix of the discretised GRF on $\mathcal{L}_{\mathcal{D}}$. The functions expand.grid() and rdist() may be useful. Use the covariance matrix to generate a simulation of the temperature on $\mathcal{L}_{\mathcal{D}}$. Display the realisation.

b) Compute the empirical variogram based on the full realisation. You can use the function variog(). Comment the results.

c) Use the realisation of the GRF from **a)** and draw 36 locations randomly from $\mathcal{L}_{\mathcal{D}}$. Compute the empirical variogram estimate based on these 36 observations. (We assume perfect observations without measurement noise.)

Assume an exponential variogram function with variance σ_1^2 and range parameter θ_1. Estimate σ_1^2 and θ_1 by maximum likelihood based on these 36 observations. Use the function likfit().

Display the variogram estimates above together with the true one.

Comment on the results.

d) Repeat the procedure in **c)** with 9, 64 and 100 observations.

Comment on the results.