

TMA4250 Spatial Statistics

Assignment 1: Continuous Spatial Variables

IMF/NTNU/HO&TR

January 2018

Introduction

This assignment contains problems related to continuous spatial variables and Gaussian random fields (RF). We recommend using R for solving the problems, and relevant functions can be found in the R libraries `geoR`, `akima` and `fields`.

Problem 1: Gaussian RF - model characteristics

Consider the continuous spatial variable $\{r(x) : x \in \mathcal{D} : [1, 50] \subset \mathbb{R}^1\}$, and assume that it is modeled as a stationary 1D Gaussian RF with the following model parameters:

$$\begin{aligned}E\{r(x)\} &= \mu_r = 0 \\Var\{r(x)\} &= \sigma_r^2 \\Corr\{r(x), r(x')\} &= \rho_r(\tau)\end{aligned}$$

where $\rho_r(\tau); \tau = |x - x'|/10$ is the spatial correlation function. Let $\mathcal{D} : [1, 50]$ be discretised in $\mathbf{L} \in \{1, 2, \dots, 50\}$ and define the discretised Gaussian RF $\{r(x); x \in \mathbf{L}\}$.

Let the spatial correlation function $\rho_r(\tau)$, be either Powered exponential with parameter $\nu_r \in [1, 1.9]$ or Matern with parameter $\nu_r \in [1, 3]$. Let the variance take the values $\sigma_r^2 \in [1, 5]$.

a) The spatial correlation functions must be a positive definite function, specify this requirement mathematically. Explain why this requirement is necessary ?

Display the two spatial correlation functions for $\tau \in \mathbb{R}_\oplus$ for the model parameters specified above. Discuss the features of the spatial correlation function which is crucial for the associated Gaussian RF, and the relations between

the variance and correlation function with the variogram function. Display the associated variogram functions $\gamma_r(\tau)$.

Use the functions `cov.spatial()` and/or `Matern()`.

b) Specify the pdf for the discretized prior Gaussian model.

Simulate ten realisations of the Gaussian RF on L for each of the eight different sets of model parameters defined above and present them in eight displays with ten realizations in each.

Discuss the relation between the realizations and the model parameters.

Let the spatial variable be observed as $\{d(x); x \in [10, 25, 30] \subset L\}$ according to the acquisition model,

$$d(x) = r(x) + \epsilon(x) \quad x \in [10, 25, 30]$$

with measurement errors $\epsilon(\cdot)$ centred, independent and identically Gaussian distributed with variance σ_ϵ^2 . Further, assume that $r(x)$ and $\epsilon(x')$ are independent for all x, x' .

c) Specify the expression for corresponding likelihood model and explain why this is not a pdf, and the consequences thereof.

Consider the simulated realizations in **b)** with $\sigma_r^2 = 5$ and select one realisation, and use the values at $x \in [10, 25, 30]$ in this realization as the observed values $\{d(x); x \in [10, 25, 30] \subset L\}$. Let the observation error variance take the values $\sigma_\epsilon^2 \in [0, 0.25]$.

d) Specify the pdf for the discretised posterior Gaussian RF given the observations.

Use the prior model for the realization selected above, and the likelihood model with the two error variances listed above, and the actual observed values. Compute the corresponding two prediction for the spatial variable $\{\hat{r}(x); x \in L\}$ with associated prediction 0.9-intervals, and present the results in two displays.

Discuss these two displays and relation between the model parameters and the predictions with prediction intervals.

e) Simulate 100 realisations from each of the two discretised posterior Gaussian RF models, and estimate empirically the prediction with associated prediction 0.9-intervals based on these realizations for each model.

Present the simulated realizations in two displays, one for each model, and over-print the corresponding empirically estimated predictions and prediction intervals in each display.

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) Consider the non-linear function on $\{r(x); x \in \mathcal{D}\}$,

$$A_r = \int_{\mathcal{D}} I(r(x) > 2) dx$$

and use the 100 realizations with $\sigma_\epsilon^2 = 0$ to provide an prediction $\hat{A}_r$ with associated prediction variance. An alternative predictor is based on the predicted spatial variable with $\sigma_\epsilon^2 = 0$ being $\{\hat{r}(x); x \in \mathcal{L}\}$,

$$\tilde{A}_r = \sum_{x \in \mathcal{L}} I(\hat{r}(x) > 2).$$

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

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

Problem 2: Gaussian RF - real data

Consider observations of terrain elevation, available in the file `topo.dat`. The 52 observations are located in the domain $\mathcal{D} = (0, 315) \times (0, 315) \subset \mathbb{R}^2$. Let the 52-vector of exact observations be $\mathbf{d} = (r(\mathbf{x}_1^o), \dots, r(\mathbf{x}_{52}^o))^T$.

a) Display the observations in various ways. Is a stationary Gaussian RF 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.

Let the terrain elevation in domain $\mathcal{D}$ be modeled by the Gaussian RF $\{r(\mathbf{x}); \mathbf{x} \in \mathcal{D} \subset \mathbb{R}^2\}$ with

$$\begin{aligned} E\{r(\mathbf{x})\} &= \mathbf{g}(\mathbf{x})^T \boldsymbol{\beta}_r \\ Var\{r(\mathbf{x})\} &= \sigma_r^2 \\ Corr\{r(\mathbf{x}), r(\mathbf{x}')\} &= \rho_r(\tau/\xi) \end{aligned}$$

where $\mathbf{g}(\mathbf{x}) = (g_1(\mathbf{x}), \dots, g_{n_g}(\mathbf{x}))^T$ is n_g -vector of known explanatory spatial variables on $\mathbf{x} \in \mathcal{D}$, and $\boldsymbol{\beta}_r = (\beta_1, \dots, \beta_{n_g})^T$ is a n_g -vector of unknown parameters. Moreover, let the variance $\sigma_r^2 = 2500$ and the spatial correlation function have $\xi = 100$ with $\tau = |\mathbf{x} - \mathbf{x}'|$.

b) Develop the expression for the minimization problem to be solved for the universal kriging predictor and the associated prediction variance at an arbitrary location $\mathbf{x}_0 \in \mathcal{D}$. The actual optimization need not be solved.

Let the reference variable $\mathbf{x} \in \mathcal{D} \subset \mathbb{R}^2$ be denoted $\mathbf{x} = (x_v, x_h)$, set $n_g = 6$ and define the set of known polynomial functions $\mathbf{g}(\mathbf{x})$ to be all polynomials $x_v^k x_h^l$ for $(k, l) \in \{(0, 0), (1, 0), (0, 1), (1, 1), (2, 0), (0, 2)\}$.

c) Specify the resulting n_g -vector $\mathbf{g}(\mathbf{x})$ and the expected value of $r(\mathbf{x})$. Present an interpretation of this model, and comment on the interaction between the expectation model and the value of the variance.

Discretize the Gaussian RF to $\{r(\mathbf{x}); \mathbf{x} \in \mathcal{L}\}$ with the grid $\mathcal{L} : (315 \times 315) \in \mathcal{D}$. Calculate the universal Kriging predictor with associated prediction variance, ie $\{\hat{r}(\mathbf{x}); \mathbf{x} \in \mathcal{L}\}$ and $\{\sigma_{\hat{r}}^2(\mathbf{x}); \mathbf{x} \in \mathcal{L}\}$. Display the results and comment on them.

Use the function `krige.conv()`.

Remove the second order terms in the expectation function, and calculate the universal Kriging predictor with associated prediction variance. Is it natural to change the value of the variance if the parametrization of expectation function is changed?

Present the results and comment on them, and the relation to the results using a higher-order expectation function.

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

d) Consider grid node $\mathbf{x}_0 = (100, 100)$. Specify the probability for the elevation to be higher than 700 m at this location. Further, specify the elevation for which it is 0.90 probability that the true elevation is below it.

e) Assume in this next-to-last point that the observations in the n_g -vector $\mathbf{d}$ are associated with observation errors being centred Gaussian, independent from each other and the terrain elevation, with error variance σ_ϵ^2 . Do not change the numerical values of the observations. Calculate the Kriging prediction with associated prediction variance for the two values on the observation errors $\sigma_\epsilon^2 \in [5, 25]$, using both the expectation functions defined above.

Present the results and comment and compare.

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

Problem 3: Parameter estimation

Consider the stationary Gaussian RF $\{r(\mathbf{x}); \mathbf{x} \in \mathcal{D} \subset \mathbb{R}^2\}$ with $\mathcal{D} : [(1, 30), (1, 30)]$, with

$$\begin{aligned}E\{r(\mathbf{x})\} &= \mu_r = 0 \\Var\{r(\mathbf{x})\} &= \sigma_r^2 \\Corr\{r(\mathbf{x}), r(\mathbf{x}')\} &= \exp\{-\tau/\xi_r\}\end{aligned}$$

with $\tau = |\mathbf{x} - \mathbf{x}'|$.

a) Consider the discretized Gaussian RF $\{r(\mathbf{x}); \mathbf{x} \in \mathcal{L}\}$ on grid $\mathcal{L} : [30 \times 30] \in \mathcal{D}$. Set the model parameters $\sigma_r^2 = 2$ and $\xi_r = 15$ and generate one realization of the discretized Gaussian RF and display it.

b) Compute the empirical variogram based on exact observations of the full realisation, and display the estimate jointly with the correct variogram function. Comment on the result.

Use function `variog()`.

c) Generate 36 locations uniformly randomly in the grid $\mathcal{L}$. Compute the empirical variogram estimate based on the corresponding 36 exact observations. Display the estimate jointly with the correct variogram function, and comment on the results.

Consider the model parameters variance σ_r^2 and ξ_r to be unknown. Estimate the parameters by a maximum likelihood criterion based on exact observation of the full realization and based on the 36 exact observations. Display the corresponding estimated variogram functions jointly with the correct variogram function, and comment on the result.

Use the function `likfit()`.

Discuss and compare all the results above.

d) Repeat the procedure in **c)** with 9, 64 and 100 uniformly randomly generated exact observations from the realization. Present the estimates jointly with the correct variogram function in separate displays, and comment on the results.

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