Project 2: Gaussian random fields and INLA

Please send your solutions to andrea.riebler@ntnu.no. **Deadline:** April 7th.

1 Multivariate normal distribution

Let $\boldsymbol{x} = (x_1, \ldots, x_n)$, $n = 100$, be multivariate normal distributed with $E(x_i) = 0$, $Var(x_i) = 1$, and $\text{Corr}(x_i, x_j) = e^{-0.1|i-j|}$.

a) Compute and image the covariance matrix $\boldsymbol{\Sigma}$ of $\boldsymbol{x}$.

b) Find the lower Cholesky factor $\boldsymbol{L}$, such that $\boldsymbol{L}\boldsymbol{L}^t = \boldsymbol{\Sigma}$, of this covariance matrix, and image that too.

c) Sample $\boldsymbol{x} = \boldsymbol{L}\boldsymbol{z}$, where $\boldsymbol{z}$ is a length n random vector of independent standard normal variables. Plot the sample.

d) Find the precision matrix $\boldsymbol{Q}$ of the covariance matrix, and compute the lower Cholesky factor $\boldsymbol{L}_Q$, such that $\boldsymbol{L}_Q\boldsymbol{L}_Q^t = \boldsymbol{Q}$, of this matrix. Image these matrices and compare them to the images obtained in a) and b).

e) Sample $\boldsymbol{x}$ by solving $\boldsymbol{L}_Q^t\boldsymbol{x} = \boldsymbol{z}$, where $\boldsymbol{z}$ is a length n random vector of independent standard normal variables. Plot the sample.

f) Permute the ordering of variables in $\boldsymbol{x}$, and redo the exercise. (It matters - there are heuristic algorithms for optimal sparseness in graphical models).

2 Gaussian random fields and Kriging

The purpose of this computer exercise is to give an introduction to parameter estimation and kriging for Gaussian random field models for spatial data.

We assume the following observation model on the unit square:

$$y(\boldsymbol{s}_j) = x(\boldsymbol{s}_j) + \epsilon_j, \quad j = 1, \dots, N,$$

where $\epsilon_j \sim N(0, \tau^2)$ are independent measurement noise terms. Further, consider a Matérn covariance function for the Gaussian random field $\boldsymbol{x}(\boldsymbol{s})$:

$$\mathrm{Cov}(x(\boldsymbol{s}_i), x(\boldsymbol{s}_j)) = \Sigma_{i,j} = \sigma^2(1 + \phi h)\exp(-\phi h),$$

where h denotes the Euclidean distance between the two sites $\boldsymbol{s}_i$ and $\boldsymbol{s}_j$.

We assume the mean increases with east and north coordinates as follows: $\mu_j = \alpha((s_{j1} - 0.5) + (s_{j2} - 0.5))$, for site $\boldsymbol{s}_j = (s_{j1}, s_{j2})$ on the unit square.

2.1 Simulation

Simulate $N = 200$ random sites in the unit square and plot them. Form the covariance matrix using $\sigma^2 = 1$, $\phi = 10$ and $\tau^2 = 0.05^2$. Take its Cholesky decomposition and simulate dependent zero-mean Gaussian data variables, then add the mean using $\alpha = 1$. Plot your observations.

(Hint: The function `rdist` of the package `fields` might be useful to compute the distance matrix).

2.2 Parameter estimation

We will now use the simulated data to estimate the model parameters $\alpha, \sigma^2, \tau^2, \phi$ using maximum likelihood estimation. Iterate between the update for the mean parameter, and updating the covariance parameters. Monitor the likelihood function at each step of the algorithm to check convergence.

2.3 Kriging

We will now use the estimated model parameters to perform kriging prediction. Predict variables $\boldsymbol{x}(s)$, where predictions sites lie on a regular grid of size 25x25 for the unit square. Visualize the Kriging surface and the prediction standard error. Compare with the true field.

3 Integrated nested Laplace approximations (INLA)

Download and install INLA by using the following command within R:

```
> install.packages("INLA",repos=c(getOption("repos"),
                INLA="https://inla.r-inla-download.org/R/stable"), dep=TRUE)
```

3.1 Simple linear regression

We consider a dataset with y_i representing a length measured in a skijumping competition and x_i the respective observation year, and assume a simple linear regression model, where

$$\mathrm{E}(y_i) = \mu + \beta x_i, \quad \mathrm{Var}(y_i) = \tau^{-1}, \quad i = 1, \ldots, n$$

1. Load package **INLA**:

```
> library(INLA)
```

2. Download the data set from the course webpage and save it in the current working directory, read "SkiJump" data into R using the supplied script

```
> skiData = read.table("SkiJump.txt", header = TRUE)
```

3. Make yourself familiar with the data set (a short summary and /or some plots)

4. Fit a simple linear regression using

```
> res = inla(Length ~ Year, data = skiData)
```

inspect the summary statistics, plot the marginal posterior for the fixed effect and hyper parameter τ, plot the marginal posterior for the hyperparameter $\sigma = \sqrt{1/\tau}$ using **inla.tmarginal()** function and get the estimates for variance σ using **inla.zmarginal()**.

3.2 GLMM with random effects

Load data "Seeds" using **data(Seeds)**. This data concerns the proportion of seeds that germinated on each of 21 plates arranged according to a 2 by 2 factorial layout by seed and type of root extract. The data are shown below, where r_i and n_i are the number of germinated and the total number of seeds on the i th plate, $i = 1, ..., N$. The model is essentially a random effects logistic, allowing for over-dispersion. If p_i is the probability of germination on the i th plate, we assume

$$\begin{aligned} r_i &\sim \mathrm{Binomial}(p_i, n_i) \\ \mathrm{logit}(p_i) &= a_0 + a_1 x_{1i} + a_2 x_{2i} + \epsilon_i \end{aligned}$$

where x_{1i} , x_{2i} are the seed type and root extract of the i th plate.

(a) Fit the above model in R-INLA and check the summary of the result.

(b) Plot the marginal posteriors.