Project 2: SPDE models and INLA

Introduction

In this project, we will look at the SPDE approach to spatial modelling and the INLA approach to statistical inference. In order to get a better understanding of the SPDE approach, we will start with doing a full implementation of a non-stationary Matérn SPDE model in one dimension, and then use it to analyse time series data. In the second part, we use the R-INLA implementation of the models to analyse more interesting spatial data, where we also use the INLA inference approach.

The first part of the project can be written in any language, whereas the second part has[1] to be done in `R` since we will use the R-INLA package. For installation instructions, see `http://www.r-inla.org/download`.

Part 1: A non-stationary random process model

In this part, you will implement a non-stationary SPDE model on $\mathbb{R}$. That is,

$$(\kappa(s)^2 - \Delta)^{\alpha/2}(\tau(s)x(s)) = W(s) \tag{1}$$

where we assume log-linear regression models for $\kappa(s)$ and $\tau(s)$,

$$\log\kappa(s) = \sum_{i=1}^{k} B_i^{\kappa}(s)\theta_i^{\kappa}, \quad \log\tau(s) = \sum_{i=1}^{k} B_i^{\tau}(s)\theta_i^{\tau}.$$

You will then use the model to analyse a time series of measurements of head accelerations in a simulated motorcycle accident, used to test crash helmets. The dataset is available in the R package `MASS`, and you can load the data through

```
> library(MASS)
> data(mcycle)
> Y <- mcycle$accel
> s <-mcycle$times
```

The response variable is in this case acceleration (in g) and s is time (in milliseconds) after impact. We will now go through the steps necessary for implementing and using the model for the data.

[1] Or actually, you are welcome to write your own implementation of the methods in a language of your choosing, though this will likely take some time.

1. Implement functions that compute the finite element matrices $\mathbf{G}$ and $\mathbf{C}$ when piecewise linear basis functions, $\{\varphi_i\}$, are used. Write the functions so that they work for irregularly spaced center locations of the basis functions.

2. Assume that $x(s)$ in (1) is sampled at some locations s_j, $j = 1, \ldots, N$, under Gaussian noise, resulting in measurements $y_j \sim \mathsf{N}(x(s_j), \sigma^2)$. First, write a function that calculates the observation matrix $\mathbf{A}$, with elements $A_{ji} = \varphi_i(s_j)$. Then, derive and implement the log-likelihood $\log \pi(\mathbf{y}|\boldsymbol{\theta})$ where $\boldsymbol{\theta} = \{\boldsymbol{\theta}^{\kappa}, \boldsymbol{\theta}^{\tau}, \sigma\}$. *Hint: Start with* $\pi(\mathbf{y}|\mathbf{x}, \boldsymbol{\theta})$ *and rewrite it so that you can integrate out* $\mathbf{x}$.

3. Assume polynomial basis functions ($B_k(s_i) = s_i^k$) for the parameters and fit the model to the data using numerical optimization of the log-likelihood. *Hint: It may be a good idea to standardise the locations to* $s \in [0, 1]$ *to avoid numerical issues.*

4. Given the estimated parameters, $\boldsymbol{\theta}^*$, calculate $\mathsf{E}(x(s)|\mathbf{y}, \boldsymbol{\theta}^*)$ as a point estimate of the latent process.

5. Examine the effect of the various choices we have in the model:

   - Do non-stationary models for κ and τ seem necessary for this data?
   - How does the posterior mean of $x(s)$ change when different orders of the polynomial basis functions are used in the parameter models?
   - Try estimating the model for a few different values of α, say $\alpha = 1, 2$, and 3. How does the smoothness parameter affect the results?
   - How do the results change if you increase or decrease the number of basis functions? Also investigate what happens if you use a basis that is not induced by the measurement locations but rather is constructed to have constant spacing between the functions.

Part 2: A spatial application using INLA

Having completed Part 1, all you would have to do to extend the model to a spatial domain is to update our functions that calculates $\mathbf{G}$, $\mathbf{C}$, and $\mathbf{A}$ to handle piecewise linear basis functions in the plane. This requires some more programming, so we will now instead use the R-INLA implementation the models.

You can now choose between two options:

1. Analyse a spatial dataset of your choice, preferably something from your own research. If you want to do this, consult with me first in order to come up with a suitable plan for the project.

2. Analyse the US temperature data as explained below.

US temperature reconstruction

In this section, we will estimate the mean summer temperature for the US. On the homepage you can find the file `USdata.RData`, which contains temperature measurements from the summer of 1997. This data is a part of the larger dataset[2] which was created from the data archives of the National Climatic Data Center. The file contains measurements at 4518 stations, and for each station we have the stations recorded temperature, longitude, latitude, and elevation. The file also contains elevation data for a fine grid over the region, obtained from the 2.5-minute Digital Elevation Model (DEM) for the Conterminous US by the PRISM Climate group[3]. Finally, since the distance to the coast may be an important covariate, I have also added the distances to both the east and the west coasts for both the measurement locations and the grid locations.

Do a full analysis of the data by finding a suitable model, validating it, and finally using it to predict the temperatures. For instance using the following steps:

1. Start with assuming a simple linear regression model for the temperatures, with elevation as explanatory variable. Explain why this model is not sufficient for the data (for example by looking at the residuals).

2. Extend the regression model by adding a spatial random effect (a SPDE model) and see if this improves the results. Also investigate how the mesh for the basis functions affects the results (this is also of interest for the final prediction later).

3. Further extend the model by including the other covariates we have access to (elevation, longitude, latitude, and distance to the two coasts) and decide on suitable models for them. Use DIC or logarithmic score to compare the different models. Once you have decided on how to include the covariates, also use DIC or the logarithmic score to investigate if the spatial effect is necessary.

4. When you have decided on a suitable model to use, calculate and plot a histogram of the model's PIT values. Does the model seem to be well-calibrated?

5. Finally, use your best model to predict the temperatures at the 2.5-minute grid over the US and also compute the standard error of the prediction and discuss the results. *Hint: The grid is quite large, so you may want to use a coarser grid. The file `thin_covariates.R` on the homepage shows one way of subsampling the grid.*

Report

Write a clear and concise report presenting your approach to the assignment, discussing the methods and results. Include figures with explanatory texts. Submit the report as a PDF and also include a zip-file with your code that can be used to run the analysis. Organize the code so that the file `proj2part1.R` runs the first part and `proj2part2.R` runs the second part. Email the files to `david.bolin@chalmers.se`. The report is due on Monday March 16, at 23:59.

[2] See `http://www.image.ucar.edu/Data/US.monthly.met/`

[3] See `http://prism.oregonstate.edu`