TMA4250 Spatial Statistics
Project 3: Gaussian Markov Random Fields

Spring 2024

Introduction

This project contains problems related to Gaussian Markov random fields (GMRFs). We recommend using `R` for solving the problems, and relevant functions can be found in the `R` libraries `ggplot2`, `rgdal`, and `spdep`. You can use dense matrices when you solve these problems, but you have to do computations in terms of **precision matrices** and not covariance matrices. **NB:** the `R` function `chol` in `base` returns an upper-triangular matrix, i.e., $\mathbf{L}^{\mathrm{T}}$, and not $\mathbf{L}$. *Keep in mind throughout that it is always good to use the same color scale on figures that you compare.*

You will work with the geography of Nigeria, which has two nested subdivisions. The first administrative level is called admin1 for short and consists of 37 admin1 areas. The second administrative level is called admin2 for short and consists of 775 admin2 areas. *Actually, one of the admin2 areas in the map is a lake and there are only 774 admin2 areas, but we will ignore this issue throughout.*

The geography of Nigeria (shown in Figure 1) is contained in the files:

- `Admin1Geography.RData`: Contains an object `nigeriaAdm1` that contains the borders of the 37 admin1 areas
- `Admin2Geography.RData`: Contains an object `nigeriaAdm2` that contains the borders of the 775 admin2 areas
- `Admin1Graph.txt`: a 37×37 matrix giving the admin1 neighbourhood structure
- `Admin2Graph.txt`: a 775×775 matrix giving the admin2 neighbourhood structure
- `functions.R`: contains a function `plotAreaCol`, which plots values on the 37 admin1 areas or the 775 admin2 areas in Nigeria. The figure is saved directly to disk and not displayed in `R`. This is because plotting fine-scale maps directly in `R` with this function is **very** slow.

Data used in Problem 2 is contained in file:

- `DirectEstimates.txt`: See more information in problem description

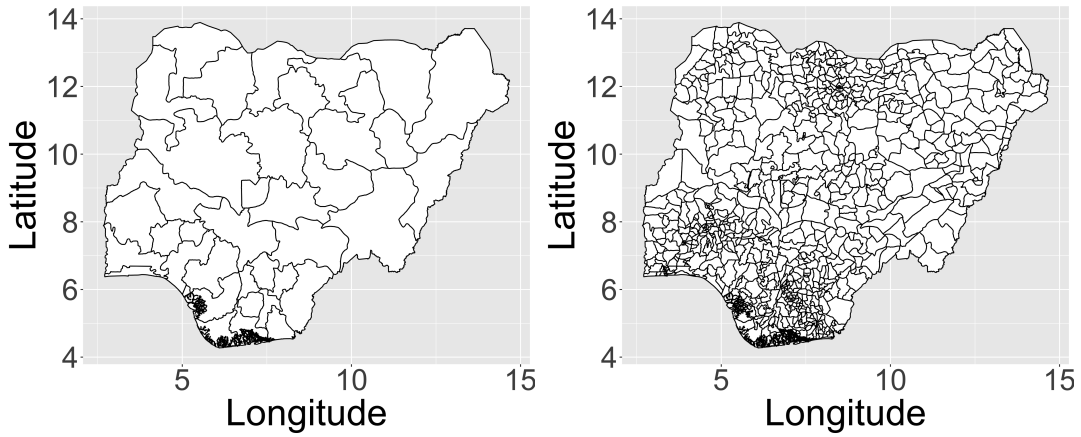

Figure 1: Left: map of 37 admin1 areas, and right: map of 775 admin2 areas.

Problem 1: Simulation and Visualization

In this problem we will use the term *admin1 graph* to refer to the graph structure arising from connecting admin1 areas that share a border, and the term *admin2 graph* to refer to the graph structure arising from connecting admin2 areas that share a border. These graphs are found in `Admin1Graph.txt` and `Admin2Graph.txt`.

a) Describe shortly how to use the neighbourhood matrices to construct the precision matrices of the Besag model on the admin1 graph, $\mathbf{Q}_1 = \tau_1 \mathbf{R}_1$, and on the admin2 graph, $\mathbf{Q}_2 = \tau_2 \mathbf{R}_2$. Here $\tau_1 > 0$ and $\tau_2 > 0$ are the precision parameters, and $\mathbf{R}_1$ and $\mathbf{R}_2$ are the structure matrices. What is the dimension of each of the precision matrices, and what is the rank of each of the precision matrices?

Compute the proportion of non-zero elements and display the sparsity pattern for each of the precision matrices. Discuss the benefits of treating the Besag models as GMRFs (and not standard multivariate Gaussian distributions).
Hint: The function `image` *is useful for plotting the non-zero patterns.*

b) Describe shortly how you simulate from the following GMRFs, and generate two realizations from each distribution:

- the Besag model on the admin1 graph with $\tau_1 = 1$, and sum-to-zero constraint

- $\mathcal{N}_{37}(\mathbf{0}, \mathbf{I}_{37})$

Display the realizations on the Nigeria map and comment them. What is the purpose of the sum-to-zero constraint? Discuss similarities and differences between simulations from the two distributions.

c) Generate two realizations each from

- the Besag model on the admin2 graph with $\tau_2 = 1$, and sum-to-zero constraint
- $\mathcal{N}_{775}(\mathbf{0}, \mathbf{I}_{775})$

Display the realizations on the Nigeria map and comment them. Discuss similarities and differences between the simulations from the two distributions. Discuss whether the differences between the two distributions are more clear for the admin2 graph than for the admin1 graph in **b)**.

d) Generate 100 realizations from the Besag model on the admin2 graph with $\tau_2 = 1$ and the sum-to-zero constraint. Compute empirically the marginal variance in each admin2 area and display the computed empirical variances on the map. Does the Besag model appear to be stationary?

Consider Gubio (admin2 area number 150) and based on the 100 realizations, compute the empirical correlations between this admin2 area and all others. Explain why areas far apart can be dependent even though the Besag model satisfies the pairwise Markov property, and explain why there are negative correlations.

Problem 2: Small Area Estimation

In this problem we will consider the estimation of *vaccination coverages* (proportion vaccinated) for children in the 37 admin1 areas in Nigeria. This is challenging because Nigeria does not have a complete registration system for vaccines, and a huge amount of effort is used to coordinate the collection of a sample of children in each admin1 area in such a way that this sample can be used to produce estimates of vaccination coverages for the admin1 areas. In this project, we will not consider the details for how this is done (but you can Google **"direct estimation"** and **"Demographic and Health Surveys"** if you are interested). A key feature of the procedure is that there is no sharing of information between admin1 areas. I.e., no spatial modelling.

Let p_a denote the true proportion of children who are vaccinated in area a for $a = 1, \ldots, 37$. Let $\hat{P}_a$ be the estimator for p_a that results from the above procedure. A common assumption is that

$$\text{logit}(\hat{P}_a) \sim \mathcal{N}(\text{logit}(p_a), V_a), \quad a = 1, \ldots, 37,$$

where $\text{logit}(x) = \log(x/(1-x))$, $V_1, \ldots, V_{37}$ are **known** variances, and $\hat{P}_1, \ldots, \hat{P}_{37}$ are independent. The **goal** of our analysis is to learn about $\boldsymbol{p} = (p_1, \ldots, p_{37})^{\mathrm{T}}$ based on observing $\hat{\boldsymbol{p}} = (\hat{p}_1, \ldots, \hat{p}_{37})^{\mathrm{T}}$. The dataset `DirectEstimation.txt` contains three columns: `Admin1` is the names of the admin1 areas, `Observation` is the values of $\text{logit}(\hat{p}_a)$, and `StdDev` is the values of $\sqrt{V_a}$.

We will consider hierarchical spatial models where we imagine the true proportions to be stochastic variables. We use the notation $\boldsymbol{X} = (\text{logit}(P_1), \ldots, \text{logit}(P_{37}))^{\mathrm{T}}$ and $\boldsymbol{Y} = (\text{logit}(\hat{P}_1), \ldots, \text{logit}(\hat{P}_{37}))^{\mathrm{T}}$ for the the random vector of true proportions and the random vector of observed proportions, respectively.

a) Display the observed proportions (i.e., $\hat{p}_a$, and **not** y_a) on the Nigeria map. Discuss whether borrowing strength in space to reduce uncertainty seems reasonable.

b) Determine the distribution of $\boldsymbol{Y}|\boldsymbol{X}$ and its parameters.

Given the vague prior $\boldsymbol{X} \sim \mathcal{N}_{37}(\boldsymbol{0}, \sigma^2 \mathbf{I}_{37})$, where $\sigma^2 = 100^2$, determine the distribution of $\boldsymbol{X}|\boldsymbol{Y} = \boldsymbol{y}$ and its parameters. Discuss what happens in the limit $\sigma^2 \to \infty$. What is the distribution of $P_a|\boldsymbol{Y} = \boldsymbol{y}$ and its parameters for $a = 1, \ldots, 37$ in this limiting case?

Compute the median and the coefficient of variation for $P_a|\boldsymbol{Y} = \boldsymbol{y}$ empirically based on 100 samples, and display them in two maps of Nigeria. Comment what you see, and compare with the estimated values in **a)**.

c) Assume that $\boldsymbol{X}$ *a priori* follows a Besag model, defined on the admin1 area graph, with precision parameter τ. Determine the expected value and precision matrix of $\boldsymbol{X}|\boldsymbol{Y} = \boldsymbol{y}$ as a function of τ. Is this an intrinsic GMRF or a proper GMRF?

Assume that $\tau = 1$. Compute the median and the coefficient of variation for $P_a|\boldsymbol{Y} = \boldsymbol{y}$ empirically based on 100 samples, and display them next in two maps of Nigeria. Comment what you see, and discuss differences and similarities with **b)**.

d) Imagine that in addition to the national survey that gave rise to $\boldsymbol{Y}$, an independent survey in (the admin1 area) Kaduna gave rise to a much more precise estimate of the proportion for that state. Assume that

$$Y_{38}|P_{\text{Kaduna}} \sim \mathcal{N}(\text{logit}(P_{\text{Kaduna}}), 0.1^2),$$

where $Y_{38}|\boldsymbol{P}$ is independent of $\boldsymbol{Y}|\boldsymbol{P}$. We want to produce better estimates for the vaccination coverages using all 38 observations. Let $\tilde{\boldsymbol{Y}} = (Y_1, \ldots, Y_{37}, Y_{38})^{\mathrm{T}}$, and assume that $\boldsymbol{X}$ has the same Besag model as in **c)**. Determine the expected value and precision matrix of $\boldsymbol{X}|\tilde{\boldsymbol{Y}} = \tilde{\boldsymbol{y}}$ as a function of τ. *Hint: you can view this as updating* $\boldsymbol{X}|\boldsymbol{Y} = \boldsymbol{y}$ *with the additional information* $Y_{38} = y_{38}$*. I.e.,* $\boldsymbol{X}|\boldsymbol{Y} = \boldsymbol{y}, Y_{38} = y_{38}$.

Assume that the value $y_{38} = 0.5$ is observed and that $\tau = 1$. Compute the median and the coefficient of variation for $P_a|\tilde{\boldsymbol{Y}} = \tilde{\boldsymbol{y}}$ empirically based on 100 samples, and

display them in two maps of Nigeria. Comment what you see, and discuss differences and similarities with **b)** and **c)**.

e) In this problem, you will investigate the sensitivity to τ. Repeat problem **c)** for $\tau = 0.1$ and $\tau = 10$, and compare with the case $\tau = 1$ computed in **c)**. Discuss the similarities and differences in medians and coefficients of variation across the three different values of τ. Based on what you see, is it important to estimate τ correctly?

f) Show that the log-likelihood for the model in **c)** can be computed by

$$l(\tau; \boldsymbol{y}) = \log(f(\boldsymbol{y}; \tau)) = \log(f(\boldsymbol{x}; \tau)) + \log(f(\boldsymbol{y}|\boldsymbol{x})) - \log(f(\boldsymbol{x}|\boldsymbol{y}; \tau)) \tag{1}$$

for any $\boldsymbol{x} \in \mathbb{R}^{37}$.

Show that

$$\begin{aligned} l(\tau; \boldsymbol{y}) = \text{Const} + \frac{37-1}{2}\log(\tau) - \frac{\tau}{2}\boldsymbol{x}^{\mathrm{T}}\mathbf{R}\boldsymbol{x} - \frac{1}{2}(\boldsymbol{y}-\boldsymbol{x})^{\mathrm{T}}\mathbf{D}^{-1}(\boldsymbol{y}-\boldsymbol{x}) \\ - \frac{1}{2}\log|\mathbf{Q}_{\mathrm{C}}| + \frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu}_{\mathrm{C}})^{\mathrm{T}}\mathbf{Q}_{\mathrm{C}}(\boldsymbol{x}-\boldsymbol{\mu}_{\mathrm{C}}), \end{aligned}$$

where $\boldsymbol{X}|\boldsymbol{Y} = \boldsymbol{y} \sim \mathcal{N}_{37}(\boldsymbol{\mu}_{\mathrm{C}}, \mathbf{Q}_{\mathrm{C}}^{-1})$, $\mathbf{R}$ is the structure matrix of the Besag distribution, and $\mathbf{D}$ is a 37×37 diagonal matrix with $d_{a,a} = V_a$, $a = 1, \ldots, 37$. Find (numerically) the maximum likelihood estimate $\hat{\tau}$ of τ. Which value do you estimate? *Hint: you can use the function* `optimize` *for minimizining/maximizing a univariate function. The computational efficiency of your approach is not important.*

Assume that $\tau = \hat{\tau}$. Compute the median and the coefficient of variation for $P_a|\boldsymbol{Y} = \boldsymbol{y}$ empirically based on 100 samples, and display them in two maps of Nigeria. Comment what you see, and discuss differences and similarities with **b)** and **c)**.