The INLA approach to estimation of SAR models

Master's thesis
Spring 2014
Claus Høstrup Vestergaard
Aalborg Universitet

Institut for Matematiske Fag
Fredrik Bajers Vej 7G
Telefon 99 40 99 40
www.math.aau.dk

Title:
The INLA approach to estimation of SAR models

Projekt type:
Master's thesis

Author:

Claus Høstrup Vestergaard

Supervisor:
Esben Høg

Circulation: 5

Pages: 58

Finished d. 11/07-2014

Synopsis:

The master's thesis principally deals with estimation of a Spatial Auto Regressive (SAR) Model by Integrated Nested Laplace Approximations (INLA). Special attention is paid to the specification of the spatial weight matrix, which is a key component of the SAR model, and to how the estimates of parameters behave depending on the particular specification.

Abstract - Overblik

Dette speciale behandler først og fremmest estimering af en Spatiel Auto Regressions (SAR) model vha. Integrerede Nestede Laplace Approximationer (INLA). Specielt fokuseres der på specifikationen af den spatielle vægt matrix, som er en nøgle ingrediens i SAR modellen, og på hvordan parameter estimater opfører sig afhængigt af specifikationen.

Der bruges også en del energi på den spatielle spill-over parameter ρ. Især de restriktioner der pålægges ρ bliver grundigt behandlet. Forfatteren bemærker en forskellighed i måden hvorpå netop dette emne bliver behandlet i literaturen, og det forsøges forklaret hvorfor visse forfattere foretrækker andre restriktioner frem for de der er strengt nødvendige.

Under databehandlingen specificeres hundredevis af vægtmatricer på en måde der så vidt vides aldrig er set før. Disse forskellige specifikationer testes og sammenlignes. I eftersøgningen af en "optimal" model, bemærker forfatteren nogle bekymrende resultater, der stiller spørgsmålstegn ved den måde resultater fra SAR modeller traditionelt rapporteres og fortolkes.

Contents

Chapter 1

Introduction

This master's thesis is written within the Statistics branch of the Department of Mathematical Sciences at Aalborg University. The thesis principally deals with estimation of a Spatial Auto Regressive (SAR) Model by Integrated Nested Laplace Approximations (INLA). Special attention is paid to the specification of the spatial weight matrix, which is a key component in the SAR model, and to how the estimates of parameters behave depending on the particular specification.

The thesis is partially an extension of a project written in the previous semester. The Chapters 3, 4, 5 and 6 deal with the theoretical framework for INLA, and are, with the exception of some minor corrections and improvements, identical to those of the 9th semester project, which can be found in AAUs project library[Vestergaard, December 2013][1]. These chapters are brought over to this thesis since they remain highly relevant to the subject. The aforementioned chapters were written in cooperation with fellow students Anne Louise Nielsen and Peter Enemark Lund under the supervision of then supervisor Poul Svante Eriksen.

Much attention is given to the spatial spill-over parameter ρ. In Chapter 8 the restrictions on ρ are rigorously dealt with. The author perceives a discrepancy in how ρ i restricted in existing literature, and attempts to explain why different mathematicians prefer different restrictions than those strictly necessary.

During the data processing hundreds of different specifications of the weight matrix W is specified, in a way that possibly breaks new ground. These specifications are tested and their results compared. In the search for an "optimal" model the author notices troubling results that asks serious question about how results from SAR models are reported and interpreted.

[1] Sections 4.2.2 and 4.2.3 do not feature in the 9th semester project. They have been added specifically for this thesis.

Chapter 2

Preliminaries

[Rue and Held, 2005, p. 19]

2.1 Undirected graphs

An undirected graph $\mathcal{G}$ is specified by a set of nodes $\mathcal{V}$ and a set of edges $\mathcal{E}$, and is denoted $\mathcal{G} = (\mathcal{V}, \mathcal{E})$. An edge $\mathcal{E}$ is denoted by an unordered pair $\{i, j\}, i \neq j$ of nodes in $\mathcal{V}$ and we say that there is an undirected edge between node i and node j. A graph is *fully connected* if $\{i, j\} \in \mathcal{E}$ for all $i, j \in \mathcal{V}$ with $i \neq j$.
Let A be a subset of $\mathcal{V}$ and let $\mathcal{G}^A$ denote the graph restricted to A, then $\mathcal{G}^A$ is called a *subgraph* of $\mathcal{G}$. If $\{i, j\} \in \mathcal{E}$ then i and j are called neighbors, and this is denoted by $i \sim j$. The *neighbors* of a node i is the set of nodes in $\mathcal{G}$ having an edge to i, so $ne(i) = \{j \in \mathcal{V} | \{i, j\} \in \mathcal{E}\}$.

2.2 Properties of the normal distribution

The normal distribution exhibits various useful properties some of which are presented in the following. Let the random variable $\boldsymbol{x} = (x_1, \ldots, x_n)^T$ be normally distributed with mean $\boldsymbol{\mu}$ ($n \times 1$ vector) and covariance matrix Σ ($n \times n$ matrix). The density of $\boldsymbol{x}$ is

$$\pi(\boldsymbol{x}) = (2\pi)^{-n/2} |\Sigma|^{-1/2} \exp\left(-\frac{1}{2}(\boldsymbol{x} - \boldsymbol{\mu})^T \Sigma^{-1} (\boldsymbol{x} - \boldsymbol{\mu})\right), \quad \boldsymbol{x} \in \mathbb{R}^n$$

This is written as $\boldsymbol{x} \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$. Also, $\mu_i = \mathrm{E}[x_i]$, $\Sigma_{ij} = \mathrm{Cov}[x_i, x_j]$, $\Sigma_{ii} = \mathrm{Var}[x_i] > 0$ and $\mathrm{Corr}[x_i, x_j] = \Sigma_{ij}/(\Sigma_{ii}\Sigma_{jj})^{1/2}$.
Sometimes it is convenient to split $\boldsymbol{x}$ into two parts $\boldsymbol{x} = (\boldsymbol{x}_A^T, \boldsymbol{x}_B^T)^T$ and split $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ according to

$$\boldsymbol{\mu} = \begin{pmatrix} \boldsymbol{\mu}_A \\ \boldsymbol{\mu}_B \end{pmatrix} \text{ and } \quad \boldsymbol{\Sigma} = \begin{pmatrix} \boldsymbol{\Sigma}_{AA} & \boldsymbol{\Sigma}_{AB} \\ \boldsymbol{\Sigma}_{BA} & \boldsymbol{\Sigma}_{BB} \end{pmatrix}$$

2.2.1 Conditional properties of the normal distribution

According to Azzalini [1996] p. 290 the conditional distribution of $\pi(\boldsymbol{x}_A|\boldsymbol{x}_B)$ is $\mathcal{N}(\boldsymbol{\mu}_{A|B}, \boldsymbol{\Sigma}_{A|B})$ where

$$\boldsymbol{\mu}_{A|B} = \boldsymbol{\mu}_A + \boldsymbol{\Sigma}_{AB}\boldsymbol{\Sigma}_{BB}^{-1}(\boldsymbol{x}_B - \boldsymbol{\mu}_B), \tag{2.1}$$

$$\boldsymbol{\Sigma}_{A|B} = \boldsymbol{\Sigma}_{AA} - \boldsymbol{\Sigma}_{AB}\boldsymbol{\Sigma}_{BB}^{-1}\boldsymbol{\Sigma}_{BA}. \tag{2.2}$$

2.2.2 Canonical Parameterization

Definition 2.1 (Canonical Parameterization)
Let $\boldsymbol{x}$ be normally distributed with mean $\boldsymbol{\mu} = \boldsymbol{Q}^{-1}\boldsymbol{b}$ and symmetric positive definite (SPD) precision matrix $\boldsymbol{Q} = \boldsymbol{\Sigma}^{-1}$. Then its canonical parameterization is $\boldsymbol{x} \sim \mathcal{N}_C(\boldsymbol{b}, \boldsymbol{Q})$ with density

$$\pi(\boldsymbol{x}) \propto \exp\left(-\frac{1}{2}\boldsymbol{x}^T\boldsymbol{Q}\boldsymbol{x} + \boldsymbol{b}^T\boldsymbol{x}\right).$$

From this it can be seen that $\mathcal{N}(\boldsymbol{\mu}, \boldsymbol{Q}^{-1}) = \mathcal{N}_C(\boldsymbol{Q}\boldsymbol{\mu}, \boldsymbol{Q})$.

Chapter 3

Gaussian Markov Random Fields

3.1 Markov Random Fields

[Rue and Held, 2005, p. 24]

Let $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ be an undirected graph and $\boldsymbol{x}$ a stochastic vector indexed by $\mathcal{V}$. Furthermore let $\boldsymbol{x}_{-i} = \{x_j\}_{j \in \mathcal{V} \setminus \{i\}}$. A Random Field $\boldsymbol{x}$ is a finite space, where each point x_i of $\boldsymbol{x}$ is a random variable. Then $\boldsymbol{x}$ is a Markov Random Field if it obeys the local Markov property, i.e.

$$\pi(x_i|\boldsymbol{x}_{-i}) = \pi(x_i|ne(i)).$$

In other words: if the distribution of x_i is independent of the rest of the graph except its neighbors.

3.2 Definition of GMRF

[Rue and Held, 2005, section 2.2.1]

Definition 3.1
Let $\pi(\boldsymbol{x})$ be a Markov Random Field and multivariate normally distributed. Then $\pi(\boldsymbol{x})$ is a Gaussian Markov Random Field (GMRF).

Let $\boldsymbol{x} = (x_1, \ldots, x_n)$ be an n-dimensional random vector, $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ an undirected graph and $\boldsymbol{x}_{-ij} = \{x_k\}_{k \in \mathcal{V} \setminus \{i,j\}}$. The set of edges $\mathcal{E}$ is constructed such that $\{i, j\} \notin \mathcal{E}$ iff x_i and x_j obey the pairwise Markov property, i.e.

$$x_i \perp x_j \mid \boldsymbol{x}_{-ij}. \tag{3.1}$$

In other words: x_i and x_j are independent given the rest of the graph.
When $\pi(\boldsymbol{x})$ is Gaussian the local and pairwise Markov properties are equivalent, [Rue and Held, 2005, p. 24]. Therefore we know that a GMRF must obey the pairwise Markov property as well as the local Markov property. This means that the pairwise Markov property now can be used to define the independence properties of the GMRF.

As it turns out in the following theorem, the precision matrix $\boldsymbol{Q}$ plays a key role with regard to (wrt) this conditional independence property, when $\boldsymbol{x}$ is normally distributed.

Theorem 3.2
Let $\boldsymbol{x}$ be normally distributed with mean $\boldsymbol{\mu}$ and SPD precision matrix $\boldsymbol{Q}$. Then for $i \neq j$

$$x_i \perp x_j \mid \boldsymbol{x}_{-ij} \Leftrightarrow Q_{ij} = 0.$$

Now the definition of a GMRF can be stated as:

Definition 3.3 (Gaussian Markov Random Field)
A random vector $\boldsymbol{x}$ is called a Gaussian Markov Random Field with regard to the graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ with mean $\boldsymbol{\mu}$ and SPD precision matrix $\boldsymbol{Q}$ if and only if its density has the form

$$\pi(\boldsymbol{x}) = (2\pi)^{-n/2} |\boldsymbol{Q}|^{1/2} \exp\left(-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^T \boldsymbol{Q}(\boldsymbol{x}-\boldsymbol{\mu})\right)$$

and

$$Q_{ij} = 0 \Leftarrow \{i,j\} \notin \mathcal{E} \text{ for all } i \neq j.$$

As it turns out, the elements in $\boldsymbol{Q}$ have the following interpretations.

Theorem 3.4
Let $\boldsymbol{x}$ be a GMRF wrt a graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ with mean $\boldsymbol{\mu}$ and SPD precision matrix. Then

$$\mathrm{E}[x_i|\boldsymbol{x}_{-i}] = \mu_i - \frac{1}{Q_{ii}} \sum_{j:j\sim i} Q_{ij}(x_j - \mu_j) \tag{3.2}$$

$$\mathrm{Prec}[x_i|\boldsymbol{x}_{-i}] = Q_{ii} \tag{3.3}$$

$$\mathrm{Corr}[x_i, x_j|\boldsymbol{x}_{-ij}] = -\frac{Q_{ij}}{\sqrt{Q_{ii}Q_{jj}}}, i \neq j \tag{3.4}$$

This theorem will not be proven. A more general proof of (3.2) and (3.3) will be presented in Section 3.3.

3.3 Conditional distribution

[Rue and Held, 2005, section 2.2.3]

When dealing with conditional distributions of *GMRF*s the canonical parameterization is useful, as updating it under successive conditioning is computationally simple, [Rue and Held, 2005] p. 26.

In the following $\boldsymbol{x}$ is split into two nonempty subsets A and B so

$$\boldsymbol{x} = \begin{pmatrix} \boldsymbol{x}_A \\ \boldsymbol{x}_B \end{pmatrix}, \quad \boldsymbol{\mu} = \begin{pmatrix} \boldsymbol{\mu}_A \\ \boldsymbol{\mu}_B \end{pmatrix}, \quad \boldsymbol{Q} = \begin{pmatrix} \boldsymbol{Q}_{AA} & \boldsymbol{Q}_{AB} \\ \boldsymbol{Q}_{BA} & \boldsymbol{Q}_{BB} \end{pmatrix}$$

The following is a generalization of Theorem 3.4 and an application of a general result for the normal distribution.

Theorem 3.5
Let $A \subset \mathcal{V}$ and $B = \mathcal{V} \setminus A$ where $A, B \neq \emptyset$. Given a GMRF $\boldsymbol{x}$ wrt $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ with mean $\boldsymbol{\mu}$ and SPD precision matrix $\boldsymbol{Q}$ then the conditional distribution $\boldsymbol{x}_A|\boldsymbol{x}_B$ is a GMRF wrt $\mathcal{G}^A$ and have mean and precision matrix as follows

$$\boldsymbol{\mu}_{A|B} = \boldsymbol{\mu}_A - \boldsymbol{Q}_{AA}^{-1}\boldsymbol{Q}_{AB}(\boldsymbol{x}_B - \boldsymbol{\mu}_B) \tag{3.5}$$

and

$$\boldsymbol{Q}_{A|B} = \boldsymbol{Q}_{AA} \tag{3.6}$$

This result shows that without computation the conditional precision matrix can be obtained since $\boldsymbol{Q}_{A|B}$ is a sub-matrix of $\boldsymbol{Q}$. It also shows that the conditional mean only depends on the mean $\boldsymbol{\mu}$ and $\boldsymbol{Q}$, through Q_{ij} where $j \in ne(i)$.

Proof. (Theorem 3.5)
Let $\tilde{\boldsymbol{x}}_A = \boldsymbol{x}_A - \boldsymbol{\mu}_A$ and $\tilde{\boldsymbol{x}}_B = \boldsymbol{x}_B - \boldsymbol{\mu}_B$, so the mean of $\tilde{\boldsymbol{x}}$ is 0. Using that $\boldsymbol{Q}_{AB} = \boldsymbol{Q}_{BA}^T$ the conditional density comes to

$$\begin{aligned}
\pi(\tilde{\boldsymbol{x}}_A|\tilde{\boldsymbol{x}}_B) &\propto \exp\left(-\frac{1}{2}(\tilde{\boldsymbol{x}}_A^T, \tilde{\boldsymbol{x}}_B^T)\begin{pmatrix} \boldsymbol{Q}_{AA} & \boldsymbol{Q}_{AB} \\ \boldsymbol{Q}_{BA} & \boldsymbol{Q}_{BB} \end{pmatrix}\begin{pmatrix} \tilde{\boldsymbol{x}}_A \\ \tilde{\boldsymbol{x}}_B \end{pmatrix}\right) \\
&= \exp\left(-\frac{1}{2}(\tilde{\boldsymbol{x}}_A^T\boldsymbol{Q}_{AA}\tilde{\boldsymbol{x}}_A + \tilde{\boldsymbol{x}}_A^T\boldsymbol{Q}_{AB}\tilde{\boldsymbol{x}}_B + \tilde{\boldsymbol{x}}_B^T\boldsymbol{Q}_{BA}\tilde{\boldsymbol{x}}_A + \tilde{\boldsymbol{x}}_B^T\boldsymbol{Q}_{BB}\tilde{\boldsymbol{x}}_B)\right) \\
&\propto \exp\left(-\frac{1}{2}(\tilde{\boldsymbol{x}}_A^T\boldsymbol{Q}_{AA}\tilde{\boldsymbol{x}}_A + (\boldsymbol{Q}_{AB}\tilde{\boldsymbol{x}}_B)^T\tilde{\boldsymbol{x}}_A + (\boldsymbol{Q}_{BA}^T\tilde{\boldsymbol{x}}_B)^T\tilde{\boldsymbol{x}}_A)\right) \\
&= \exp\left(-\frac{1}{2}\tilde{\boldsymbol{x}}_A^T\boldsymbol{Q}_{AA}\tilde{\boldsymbol{x}}_A - (\boldsymbol{Q}_{AB}\tilde{\boldsymbol{x}}_B)^T\tilde{\boldsymbol{x}}_A\right)
\end{aligned}$$

Looking at the last line and the density of a normal $\mathcal{N}(\boldsymbol{\nu}, \boldsymbol{P}^{-1})$

$$\pi(\boldsymbol{z}) \propto \exp\left(-\frac{1}{2}\boldsymbol{z}^T\boldsymbol{P}\boldsymbol{z} + (\boldsymbol{P}\boldsymbol{\nu})^T\boldsymbol{z}\right),$$

it follows that $\boldsymbol{Q}_{AA}$ is the conditional precision matrix of the conditional density and that

$$\boldsymbol{Q}_{AA}\tilde{\boldsymbol{\mu}}_{A|B} = -\boldsymbol{Q}_{AB}\tilde{\boldsymbol{x}}_B \tag{3.7}$$

Now using that $\tilde{\boldsymbol{x}}_B = \boldsymbol{x}_B - \boldsymbol{\mu}_B$ and

$$\begin{aligned}
\tilde{\boldsymbol{\mu}}_{A|B} &= \mathrm{E}[\tilde{\boldsymbol{x}}_A|\tilde{\boldsymbol{x}}_B] \\
&= \mathrm{E}[\boldsymbol{x}_A - \boldsymbol{\mu}_A|\boldsymbol{x}_B - \boldsymbol{\mu}_B] \\
&= \mathrm{E}[\boldsymbol{x}_A - \boldsymbol{\mu}_A|\boldsymbol{x}_B] \\
&= \mathrm{E}[\boldsymbol{x}_A|\boldsymbol{x}_B] - \boldsymbol{\mu}_A \\
&= \boldsymbol{\mu}_{A|B} - \boldsymbol{\mu}_A
\end{aligned}$$

and applying it to (3.7) we get

$$\boldsymbol{Q}_{AA}(\boldsymbol{\mu}_{A|B} - \boldsymbol{\mu}_A) = -\boldsymbol{Q}_{AB}(\boldsymbol{x}_B - \boldsymbol{\mu}_B).$$

From this (3.5) follows. The subgraph $\mathcal{G}^A$ comes from $\boldsymbol{Q}_{A|B} = \boldsymbol{Q}_{AA}$.

■

Notions on Canonical Parametrization

Partitioning $\{1, 2, \ldots, n\}$ into two nonempty subsets A and B yields

$$\begin{pmatrix}\boldsymbol{b}_A \\ \boldsymbol{b}_B\end{pmatrix} = \begin{pmatrix}\boldsymbol{Q}_{AA} & \boldsymbol{Q}_{AB} \\ \boldsymbol{Q}_{BA} & \boldsymbol{Q}_{BB}\end{pmatrix}\begin{pmatrix}\boldsymbol{\mu}_A \\ \boldsymbol{\mu}_B\end{pmatrix}. \tag{3.8}$$

From Theorem 3.5 a lemma regarding the canonical parametrisation, presented in Section 2.2.2, follows.

Lemma 3.6
Let $\boldsymbol{x} \sim \mathcal{N}_C(\boldsymbol{b}, \boldsymbol{Q})$ then

$$\boldsymbol{x}_A|\boldsymbol{x}_B \sim \mathcal{N}_C(\boldsymbol{b}_A - \boldsymbol{Q}_{AB}\boldsymbol{x}_B, \boldsymbol{Q}_{AA}). \tag{3.9}$$

Proof.
Let $\boldsymbol{x}_A|\boldsymbol{x}_B \sim \mathcal{N}(\boldsymbol{\mu}_{A|B}, \boldsymbol{Q}_{A|B}^{-1})$ then the canonical parameterization of this is $\mathcal{N}_C(\boldsymbol{Q}_{A|B}\boldsymbol{\mu}_{A|B}, \boldsymbol{Q}_{A|B})$. Below, the first equality is obtained from (3.5) and (3.6) and the fourth is from (3.8).

$$\begin{aligned}
\mathcal{N}_C(\boldsymbol{Q}_{A|B}\boldsymbol{\mu}_{A|B}, \boldsymbol{Q}_{A|B}) &= \mathcal{N}_C\left(\boldsymbol{Q}_{AA}\left(\boldsymbol{\mu}_A - \boldsymbol{Q}_{AA}^{-1}\boldsymbol{Q}_{AB}\left(\boldsymbol{x}_B - \boldsymbol{\mu}_B\right)\right), \boldsymbol{Q}_{AA}\right) \\
&= \mathcal{N}_C\left(\boldsymbol{Q}_{AA}\boldsymbol{\mu}_A - \boldsymbol{Q}_{AB}\left(\boldsymbol{x}_B - \boldsymbol{\mu}_B\right), \boldsymbol{Q}_{AA}\right) \\
&= \mathcal{N}_C\left(\boldsymbol{Q}_{AA}\boldsymbol{\mu}_A - \boldsymbol{Q}_{AB}\boldsymbol{x}_B + \boldsymbol{Q}_{AB}\boldsymbol{\mu}_B, \boldsymbol{Q}_{AA}\right) \\
&= \mathcal{N}_C\left(\boldsymbol{b}_A - \boldsymbol{Q}_{AB}\boldsymbol{x}_B, \boldsymbol{Q}_{AA}\right)
\end{aligned}$$

■

3.4 Specification through full conditionals

Up until this point the GMRF has been specified by its mean vector $\boldsymbol{\mu}$ and its precision matrix $\boldsymbol{Q}$. As an alternative we can specify the model by its full conditionals $\pi(x_i|\boldsymbol{x}_{-i})$ i.e. the distribution of x_i given every other point. In what follows we will construct a candidate for the specification and then prove this candidate to be unique. Suppose we specify the full conditionals as normally distributed with mean and precision

$$\mathrm{E}[x_i|\boldsymbol{x}_{-i}] = \mu_i - \sum_{j=1}^{n} \beta_{ij}(x_j - \mu_j) \tag{3.10}$$

$$\mathrm{Prec}[x_i|\boldsymbol{x}_{-i}] = \kappa_i > 0, \tag{3.11}$$

for $i = 1, \ldots, n$ and β_{ij} where $i \neq j$, and vectors $\boldsymbol{\mu}$ and $\boldsymbol{\kappa}$, i.e.

$$\pi(x_i|\boldsymbol{x}_{-i}) = \frac{1}{\sqrt{\frac{2\pi}{\kappa_i}}} \exp\left(-\frac{\kappa_i}{2}\left(x_i - (\mu_i - \sum_{j=1}^{n} \beta_{ij}(x_j - \mu_j))\right)^2\right). \tag{3.12}$$

Note that $\boldsymbol{\beta}$ takes into account the desired Markov property

$$\beta_{ij} = 0 \text{ if } i \nsim j.$$

Recalling Theorem 3.5 and comparing (3.10) and (3.11) to (3.5) and (3.6) we see that if we choose

$$\begin{aligned} Q_{ii} &= \kappa_i \\ Q_{ij} &= \kappa_i \beta_{ij} \end{aligned}$$

and require symmetry i.e.

$$Q_{ij} = \kappa_i \beta_{ij} = \kappa_j \beta_{ji} = Q_{ji}, \tag{3.13}$$

then we have a candidate for a joint density giving the specified full conditionals given that $\boldsymbol{Q}$ is SPD.

Theorem 3.7
Define

$$Q_{ij} = \begin{cases} \kappa_i \beta_{ij} & \text{for } i \neq j \\ \kappa_i & \text{for } i = j, \end{cases}$$

and assume that $\kappa_i \beta_{ij} = \kappa_j \beta_{ji}$ and that $\boldsymbol{Q}$ is SPD.

Then there exists an unique GMRF with regard to a graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ with mean $\boldsymbol{\mu}$ and precision matrix $\boldsymbol{Q}$ with the full conditionals defined in (3.12).

In order to prove the existence and uniqueness of this candidate, Brook's lemma is required.

Lemma 3.8 (Brook's lemma)
Let $\pi(\boldsymbol{x})$ be the density for $\boldsymbol{x} \in \mathbb{R}^n$ and define $\mathcal{S} = \{\boldsymbol{x} \in \mathbb{R}^n | \pi(\boldsymbol{x}) > 0\}$. Let $\boldsymbol{x}, \boldsymbol{x'} \in \mathcal{S}$, then

$$\frac{\pi(\boldsymbol{x})}{\pi(\boldsymbol{x'})} = \prod_{i=1}^{n} \frac{\pi(x_{p(i)} | x_{p(1)}, \ldots, x_{p(i-1)}, x'_{p(i+1)}, \ldots, x'_n)}{\pi(x'_{p(i)} | x_{p(1)}, \ldots, x_{p(i-1)}, x'_{p(i+1)}, \ldots, x'_p(n))}. \tag{3.14}$$

for any permutation $p = (p(1), \ldots, p(n))$ of $(1, \ldots, n)$.

Furthermore, if $\boldsymbol{x'}$ is fixed, then (3.15) represents $\pi(\boldsymbol{x})$ up to a constant.

It will be useful to note that one possible permutation of (3.14) is

$$\frac{\pi(\boldsymbol{x})}{\pi(\boldsymbol{x'})} = \prod_{i=1}^{n} \frac{\pi(x_i | x_1, \ldots, x_{i-1}, x'_{i+1}, \ldots, x'_n)}{\pi(x'_i | x_1, \ldots, x_{i-1}, x'_{i+1}, \ldots, x'_{p(n)})}. \tag{3.15}$$

Proof. (Theorem 3.7)
Assume $\boldsymbol{\mu} = 0$ and fix $\boldsymbol{x'} = 0$. Then the log of (3.15) equals

$$\begin{aligned}
\log \frac{\pi(\boldsymbol{x})}{\pi(\boldsymbol{0})} &= \log \prod_{i=1}^{n} \frac{\pi(x_i | x_1, \ldots, x_{i-1}, 0, \ldots, 0)}{\pi(0 | x_1, \ldots, x_{i-1}, 0, \ldots, 0)} \\
&= \sum_{i=1}^{n} \log \frac{\pi(x_i | x_1, \ldots, x_{i-1}, 0, \ldots, 0)}{\pi(0 | x_1, \ldots, x_{i-1}, 0, \ldots, 0)} \\
&= \sum_{i=1}^{n} \log \frac{\exp(-\frac{\kappa_i}{2}(x_i - (\mu_i - \sum_{j=1}^{i-1} \beta_{ij}(x_j - \mu_j)))^2)}{\exp(-\frac{\kappa_i}{2}(0 - (\mu_i - \sum_{j=1}^{i-1} \beta_{ij}(x_j - \mu_j)))^2)} \\
&= \sum_{i=1}^{n} \left(-\frac{\kappa_i}{2}(x_i + \sum_{j=1}^{i-1} \beta_{ij} x_j)^2 + \frac{\kappa_i}{2}(\sum_{j=1}^{i-1} \beta_{ij} x_j)^2 \right) \\
&= \sum_{i=1}^{n} \left(-\frac{\kappa_i x_i^2}{2} - \frac{\kappa_i}{2}(\sum_{j=1}^{i-1} \beta_{ij} x_j)^2 - \kappa_i x_i \sum_{j=1}^{i-1} \beta_{ij} x_j + \frac{\kappa_i}{2}(\sum_{j=1}^{i-1} \beta_{ij} x_j)^2 \right) \\
&= \sum_{i=1}^{n} \left(-\frac{\kappa_i x_i^2}{2} - \kappa_i x_i \sum_{j=1}^{i-1} \beta_{ij} x_j \right) \\
&= -\frac{1}{2} \sum_{i=1}^{n} \kappa_i x_i^2 - \sum_{i=2}^{n} \sum_{j=1}^{i-1} \kappa_i \beta_{ij} x_i x_j.
\end{aligned} \tag{3.16}$$

The second to last sum is changed from $\sum_{i=1}^{n}$ to $\sum_{i=2}^{n}$ since the last term of (3.16) does not make sense for $i = 1$. Note that the last term in (3.16) sums the rows

of the *lower* triangle of the (i, j)-matrix except for the diagonal. In the following it is shown that this is equivalent to summing the *upper* triangle except for the diagonal. Summing the upper triangle is equivalent to summing

$$\begin{aligned}\sum_{i=1}^{n-1}\sum_{j=i+1}^{n}\kappa_i\beta_{ij}x_ix_j &= \sum_{i=1}^{n-1}\sum_{j=i+1}^{n}\kappa_j\beta_{ji}x_jx_i \qquad (3.17)\\ &= \sum_{j=2}^{n}\sum_{i=1}^{j-1}\kappa_j\beta_{ji}x_jx_i\\ &= \sum_{i=2}^{n}\sum_{j=1}^{i-1}\kappa_j\beta_{ji}x_ix_j,\end{aligned}$$

where the equality in (3.17) comes from the symmetry property in (3.13). This shows that summing the upper triangle is equal to summing the lower triangle. This means that it is possible to write

$$\log\frac{\pi(\boldsymbol{x})}{\pi(\boldsymbol{0})} = \frac{1}{2}\sum_{i=1}^{n}\kappa_i x_i^2 - \frac{1}{2}\sum_{i\neq j}^{n}\kappa_i\beta_{ij}x_ix_j \qquad (3.18)$$

$$\Leftrightarrow \log\pi(\boldsymbol{x}) = c - \frac{1}{2}\sum_{i=1}^{n}\kappa_i x_i^2 - \frac{1}{2}\sum_{i\neq j}^{n}\kappa_i\beta_{ij}x_ix_j. \qquad (3.19)$$

We now see that $\boldsymbol{x}$ is a normally distributed vector with zero mean and precision $\boldsymbol{Q}$ as the distribution in (3.3), provided that $\boldsymbol{Q}$ is SPD.

■

Chapter 4

Bayesian Inference

4.1 Introduction to Bayesian Inference

[Gelman et al., 2003]

In the Bayesian framework, we are generally interested in determining the distribution of the model's parameters, $\boldsymbol{\theta}$, given the set of data at hand, $\boldsymbol{y}$. We denote this distribution $\pi(\boldsymbol{\theta}|\boldsymbol{y})$ and call it the *posterior distribution*. Consider the following equality

$$\pi(\boldsymbol{\theta}|\boldsymbol{y}) = \frac{\pi(\boldsymbol{\theta}, \boldsymbol{y})}{\pi(\boldsymbol{y})} = \frac{\pi(\boldsymbol{y}|\boldsymbol{\theta})\pi(\boldsymbol{\theta})}{\pi(\boldsymbol{y})}. \tag{4.1}$$

Since $\boldsymbol{y}$ is given $\pi(\boldsymbol{y})$ can be considered a constant. Therefore we write the unnormalized posterior distribution

$$\pi(\boldsymbol{\theta}|\boldsymbol{y}) \propto \pi(\boldsymbol{y}|\boldsymbol{\theta})\pi(\boldsymbol{\theta}), \tag{4.2}$$

which, in words, says that the posterior distribution is proportional to the *likelihood function* times the *prior distribution*.
So in order to get to the posterior, we must first assume a prior on the parameters. This prior can reflect a pre-exsisting knowledge of the parameters, perhaps results from an earlier study is available, maybe the data must satisfy a law of nature of some sort, maybe the sign of the parameter is known and so on.
If we do not have any prior knowledge about the distribution of the parameter, or if we want to "let the data speak for itself", we can choose a so-called non-informative prior. A non-informative prior reflects no pre-existing knowledge such that $\pi(\boldsymbol{\theta}) \propto 1$ i.e. constant, which means that all $\boldsymbol{\theta}$ are equally likely. If, say, $\boldsymbol{\theta} \in [0, 1]$ then unif$(0, 1)$ would be a non-informative prior, or if $\boldsymbol{\theta} \in \mathbb{R}$ then $\mathcal{N}(0, \infty)$ would work as a non-informative prior. If we choose a non-informative prior we see that the posterior is proportional to the likelihood only.

$$\pi(\boldsymbol{\theta}|\boldsymbol{y}) \propto \pi(\boldsymbol{y}|\boldsymbol{\theta}) \times 1 = \pi(\boldsymbol{y}|\boldsymbol{\theta}). \tag{4.3}$$

We also assume a certain structure of the data in order to determine a likelihood function. Say we assume the data is normally i.i.d. distributed, then $\pi(\boldsymbol{y}|\boldsymbol{\theta})$ gives us the likelihood of the data given the mean and variance. We consider the likelihood as a function of $\boldsymbol{\theta}$: $\mathbb{R}^{\dim \boldsymbol{\theta}} \to \mathbb{R}$.
Typically what we are interested in is the marginal distributions $\pi(\theta_i|\boldsymbol{y})$, in order to produce plots and calculate central posterior interval (CPI) etc. Calculating these marginals however is easier said than done. Consider that computing the marginals from the following equation

$$\pi(\theta_i|\boldsymbol{y}) = \frac{\int \pi(\boldsymbol{y}, \theta_i, \boldsymbol{\theta}_{-i}) d\boldsymbol{\theta}_{-i}}{\int \pi(\boldsymbol{y}, \boldsymbol{\theta}) d\boldsymbol{\theta}},$$

involves integration in multiple dimensions which can be a very computationally demanding task. Therefore it is more convenient to study the marginals of $\boldsymbol{\theta}$ e.g. by using a Gibbs sampler.

4.2 Model selection and model checks

In this section we will discuss methods for model selection and model checking.

4.2.1 Deviance Information Criterion

[Gelman et al., 2003, p. 182].

The Deviance Information Criterion (DIC) is a measure of the 'quality' of a model in terms of quality of fit and model complexity. A penalty for model complexity is included, since any level of fit can be obtained if enough parameters are added. DIC is a hierarchical model generalization of the Akaike information crition, and is only valid if the posterior is approximately normal.
First define the deviance as

$$D(\boldsymbol{\theta}) = -2\log(\pi(\boldsymbol{y}|\boldsymbol{\theta})),$$

and the expectation

$$\bar{D} = \mathrm{E}[D(\boldsymbol{\theta})|\boldsymbol{y}].$$

If the likelihood $\pi(\boldsymbol{y}|\boldsymbol{\theta})$ is large, i.e. the model fits the data well, the log-likelihood is large as well, which would mean that $-2\log(p(\boldsymbol{y}|\boldsymbol{\theta}))$ is small. So a model with small expectation is preferable to a model with large expectation. Define now the effective number of parameters

$$p_D = \bar{D} - D(\bar{\boldsymbol{\theta}}),$$

where $\bar{\boldsymbol{\theta}} = \mathrm{E}[\boldsymbol{\theta}|\boldsymbol{y}] = \int \boldsymbol{\theta} p(\boldsymbol{\theta}|\boldsymbol{y})$ i.e. the posterior mean. According to Spiegelhalter et al. [2002] p_D is a good measure of the effective number of parameters in the hierarchical model. We now define the DIC as

$$DIC = \bar{D} + p_D = D(\bar{\boldsymbol{\theta}}) + 2p_D.$$

In effect then $\bar{D}$ works as a penalty for poor fit and p_D acts as a penalty for model complexity. So a model leading to a low DIC is preferable to a model leading to a high DIC.

4.2.2 CPO

The conditional predictive ordinate or CPO, is a method of detecting outliers in a data set given some model. Formally we have that

$$CPO_i = \pi\left(y_i|\boldsymbol{y}_{-i}\right),$$

which is computed for all $i \in 1, \ldots, n$. If y_i is an outlier, CPO_i will be small. If a choice of model leads to many small CPOs, the model may be flawed in some way. This is similar to investigating the residuals of a model, but is a certain sense better: Whereas a residuals only measure the distance between the data and the model, the CPO measures the probability of that distance, i.e. this takes the distribution of the model into account.

4.2.3 PIT measure

The probability integral transform relies on the fact that a cumulative distribution function applied to a probability distribution function results in a uniform distribution, i.e. if y is a random variable with a continuous distribution and cdf F_y, then the random variable

$$U = F_y(y),$$

is uniformly distributed. This holds exactly if the correct distribution is used. Formally the PIT_i is defined as

$$PIT_i = p\left(Y_i \leq y_i|y_{-i}\right)$$

So if the model is 'true' the PIT_is will be uniformly distributed.

Chapter 5

Latent Gaussian Models

[Baio, 2013]

This section presents the Latent Gaussian Model (LGM) combined with Gaussian Markov Random Fields.
The GMRF $\boldsymbol{x}$ takes the form

$$\begin{aligned} &\boldsymbol{x}|\boldsymbol{\theta} \sim \mathcal{N}(0, \Sigma(\boldsymbol{\theta})) \\ &\boldsymbol{x}_i \perp \boldsymbol{x}_j \mid \boldsymbol{x}_{-ij}, \qquad\qquad i \nsim j \end{aligned}$$

where an undirected graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ denotes the conditional independence properties of $\boldsymbol{x}$, see Section 3.2 for clarification.

GMRFs are frequently used in hierarchical modeling, and these are specified in terms of the hyperprior $\pi(\boldsymbol{\theta})$, a "GMRF prior" $\pi(\boldsymbol{x}|\boldsymbol{\theta})$ and the likelihood (or data model) $\pi(\boldsymbol{y}|\boldsymbol{x}, \boldsymbol{\theta})$ accordingly

$$\begin{aligned} \boldsymbol{\theta} &\sim \pi(\boldsymbol{\theta}) \\ \boldsymbol{x}|\boldsymbol{\theta} &\sim \pi(\boldsymbol{x}|\boldsymbol{\theta}) = \mathcal{N}(0, \Sigma(\boldsymbol{\theta}_1)) \\ \boldsymbol{y}|\boldsymbol{x}, \boldsymbol{\theta} &\sim \prod_{i \in I} \pi(y_i|x_i, \boldsymbol{\theta}_2), \end{aligned} \tag{5.1}$$

where $\boldsymbol{\theta} = (\boldsymbol{\theta_1^T}, \boldsymbol{\theta_2^T})^T$ and $I \subseteq \mathcal{V}$ is a set of indices. Additional covariates $\boldsymbol{z}$ are omitted in (5.1) as they do not influence the following considerations.

Regarding the dimensions of the parameters, it is worth to notice that the hyperparameter $\boldsymbol{\theta}$ normally has low dimensionality e.g. 1-6 [Rue and Martino, 2009], while the latent field $\boldsymbol{x}$ often has the same dimension as the data vector $\boldsymbol{y}$, in the case where each observation y_i corresponds to the ith element x_i in $\boldsymbol{x}$. Concerning the hyperparameters $\boldsymbol{\theta}$, $\boldsymbol{\theta}_1$ is regarded as the hyperparameter connected to the latent field $\boldsymbol{x}$ while $\boldsymbol{\theta}_2$ is connected to the data model. In practice $\boldsymbol{\theta}_1$ often consists of an unknown precision $\boldsymbol{\tau}$ of one or more dimensions. $\boldsymbol{\theta}_2$ is comprised by

one or more hyperparameters in direct connection to the data model, for example if the data is assumed negative binomial distributed, the hyper parameter would be the dispersion parameter denoted by κ. In many applications, $\boldsymbol{\theta}_2$ is zero.

The posterior of the LGM is

$$\pi(\boldsymbol{x}, \boldsymbol{\theta}|\boldsymbol{y}) \propto \pi(\boldsymbol{\theta})\pi(\boldsymbol{x}|\boldsymbol{\theta}) \prod_{i=1}^{n} \pi(y_i|x_i, \boldsymbol{\theta_2}), \tag{5.2}$$

where $\boldsymbol{x}$ is a GMRF in this case. This leads to the marginals

$$\pi(x_i|\boldsymbol{y}) = \int \pi(x_i, \boldsymbol{\theta}|\boldsymbol{y}) d\boldsymbol{\theta} = \int \pi(\boldsymbol{\theta}|\boldsymbol{y})\pi(x_i|\boldsymbol{\theta}, \boldsymbol{y}) d\boldsymbol{\theta} \tag{5.3}$$

$$\pi(\theta_k|\boldsymbol{y}) = \int \pi(\boldsymbol{\theta}|\boldsymbol{y}) d\boldsymbol{\theta}_{-k} \tag{5.4}$$

These marginals are the required distributions for inference in the LGM. It is seen that the distributions $\pi(\boldsymbol{\theta}|\boldsymbol{y})$ and $\pi(x_i|\boldsymbol{\theta}, \boldsymbol{y})$ are needed to compute these marginals, which is why they are of certain interest.

Chapter 6

Integrated Nested Laplace Approximation

Integrated Nested Laplace Approximation (INLA) is a novel alternative to MCMC that has become increasingly used over the past years. The main attractive feature is that it, in contrast to the iterative MCMC method, is an analytic approximation.

6.1 The INLA approach

[Rue and Martino, 2009]

Recall from Chapter 5 that the posterior marginals of interest reads

$$\begin{aligned}\pi(x_i|\boldsymbol{y}) &= \int \pi(x_i|\boldsymbol{\theta},\boldsymbol{y})\pi(\boldsymbol{\theta}|\boldsymbol{y})d\boldsymbol{\theta},\\ \pi(\theta_j|\boldsymbol{y}) &= \int \pi(\boldsymbol{\theta}|\boldsymbol{y})d\boldsymbol{\theta}_{-j}.\end{aligned} \tag{6.1}$$

The INLA approach is based on the following approximation to $\pi(\boldsymbol{\theta}|\boldsymbol{y})$:

$$\begin{aligned}\tilde{\pi}(\boldsymbol{\theta}|\boldsymbol{y}) &= \frac{\pi(\boldsymbol{x},\boldsymbol{\theta}|\boldsymbol{y})}{\tilde{\pi}_G(\boldsymbol{x}|\boldsymbol{\theta},\boldsymbol{y})} \text{ evaluated in } \boldsymbol{x}=\hat{\boldsymbol{x}}(\boldsymbol{\theta})\\ &\propto \frac{\pi(\boldsymbol{x},\boldsymbol{\theta},\boldsymbol{y})}{\tilde{\pi}_G(\boldsymbol{x}|\boldsymbol{\theta},\boldsymbol{y})} \text{ evaluated in } \boldsymbol{x}=\hat{\boldsymbol{x}}(\boldsymbol{\theta}),\end{aligned} \tag{6.2}$$

where $\hat{\boldsymbol{x}}(\boldsymbol{\theta})$ is the mode and $\tilde{\pi}_G(\boldsymbol{x}|\boldsymbol{\theta},\boldsymbol{y})$ is the *Gaussian approximation* to $\pi(\boldsymbol{x}|\boldsymbol{\theta},\boldsymbol{y})$.
The posterior of the Latent Gaussian Model from (5.2) reads

$$\pi(\boldsymbol{\theta})\pi(\boldsymbol{x}|\boldsymbol{\theta})\prod_{i=1}^{n}\pi(y_i|x_i,\boldsymbol{\theta_2})$$

From this it follows that

$$\log \pi(\boldsymbol{x}, \boldsymbol{\theta}) = k(\boldsymbol{\theta}) - \frac{1}{2}\boldsymbol{x}^T \boldsymbol{Q}(\boldsymbol{\theta}_1)\boldsymbol{x} + \sum_{i \in \mathcal{I}} g_i(x_i, \boldsymbol{\theta}_2)$$

where $k(\boldsymbol{\theta})$ is constant wrt to $\boldsymbol{x}$ and $g_i(x_i, \boldsymbol{\theta}_2) = \log \pi(y_i | x_i, \boldsymbol{\theta}_2)$.

The Gaussian approximation is based on the following. Since $\boldsymbol{y}$ is fixed, the shorthand notation $\pi(\boldsymbol{x}, \boldsymbol{\theta}) = \pi(\boldsymbol{x}, \boldsymbol{\theta} | \boldsymbol{y})$ is used. The idea is to make a second order Taylor expansion of $\pi(\boldsymbol{x}, \boldsymbol{\theta})$ in the mode $\hat{\boldsymbol{x}}(\boldsymbol{\theta})$ such that

$$\begin{aligned}
\log \tilde{\pi}(\boldsymbol{x}, \boldsymbol{\theta}) &= \log \pi(\hat{\boldsymbol{x}}(\boldsymbol{\theta}), \boldsymbol{\theta}) + D(\hat{\boldsymbol{x}}(\boldsymbol{\theta}))(\boldsymbol{x} - \hat{\boldsymbol{x}}(\boldsymbol{\theta})) + \frac{1}{2}(\boldsymbol{x} - \hat{\boldsymbol{x}}(\boldsymbol{\theta}))^T D^2(\hat{\boldsymbol{x}}(\boldsymbol{\theta}))(\boldsymbol{x} - \hat{\boldsymbol{x}}(\boldsymbol{\theta})) \\
&= \log \pi(\hat{\boldsymbol{x}}(\boldsymbol{\theta}), \boldsymbol{\theta}) + \frac{1}{2}(\boldsymbol{x} - \hat{\boldsymbol{x}}(\boldsymbol{\theta}))^T D^2(\hat{\boldsymbol{x}}(\boldsymbol{\theta}))(\boldsymbol{x} - \hat{\boldsymbol{x}}(\boldsymbol{\theta})) \\
&\Updownarrow \\
\tilde{\pi}(\boldsymbol{x}, \boldsymbol{\theta}) &= \pi(\hat{\boldsymbol{x}}(\boldsymbol{\theta}), \boldsymbol{\theta}) \exp\left(\frac{1}{2}(\boldsymbol{x} - \hat{\boldsymbol{x}}(\boldsymbol{\theta}))^T D^2(\hat{\boldsymbol{x}}(\boldsymbol{\theta}))(\boldsymbol{x} - \hat{\boldsymbol{x}}(\boldsymbol{\theta}))\right) \qquad (6.3)
\end{aligned}$$

where $D = \frac{\partial}{\partial \boldsymbol{x}} \log \pi(\boldsymbol{x}, \boldsymbol{\theta})$ and $D^2 = \frac{\partial^2}{\partial \boldsymbol{x} \partial \boldsymbol{x}^T} \log \pi(\boldsymbol{x}, \boldsymbol{\theta})$. Furthermore we exploit that $D(\hat{x}(\boldsymbol{\theta})) = 0$. We now recognize an unnormalized normal distribution in (6.3) with mean $\boldsymbol{\mu} = \hat{\boldsymbol{x}}(\boldsymbol{\theta})$ and precision $-D^2$. It is noted that

$$\frac{\partial^2}{\partial x_i \partial x_j} \log \pi(\boldsymbol{x}, \boldsymbol{\theta}) = \boldsymbol{Q_{ij}}(\boldsymbol{\theta}_1) \text{ if, } \quad i \neq j$$

and

$$\frac{\partial^2}{\partial x_i^2} \log \pi(\boldsymbol{x}, \boldsymbol{\theta}) = \begin{cases} \boldsymbol{Q_{ii}}(\boldsymbol{\theta}_1) \text{ if } & i \notin \mathcal{I} \\ \boldsymbol{Q_{ii}}(\boldsymbol{\theta}_1) + \frac{\partial^2 g_i}{\partial x_i^2} \text{ if } & i \in \mathcal{I}, \end{cases}$$

Hence,

$$-D^2 = \text{diag}(\boldsymbol{c}) + \boldsymbol{Q}$$

where $\boldsymbol{Q}$ is the precision matrix and $c_i = \frac{\partial^2 g_i}{\partial x_i^2}$ if $i \in \mathcal{I}$ and zero otherwise. Since $\boldsymbol{Q}$ in our setting is sparse computing D and its higher derivatives does not demand heavy computations.
Proceeding by integrating out $\boldsymbol{x}$ in (6.3), yields

$$\begin{aligned}
\tilde{\pi}(\boldsymbol{\theta}) &= \pi(\hat{\boldsymbol{x}}(\boldsymbol{\theta}), \boldsymbol{\theta}) k \int \frac{1}{k} \exp\left(\frac{1}{2}(\boldsymbol{x} - \hat{\boldsymbol{x}}(\boldsymbol{\theta}))^T D^2(\hat{\boldsymbol{x}}(\boldsymbol{\theta}))(\boldsymbol{x} - \hat{\boldsymbol{x}}(\boldsymbol{\theta}))\right) d\boldsymbol{x} \qquad (6.4) \\
&= \pi(\hat{\boldsymbol{x}}(\boldsymbol{\theta}), \boldsymbol{\theta})(2\pi)^{(\dim \boldsymbol{\theta})/2} |D^2(\hat{\boldsymbol{x}}(\boldsymbol{\theta}))|^{-1/2}. \qquad (6.5)
\end{aligned}$$

where k in (6.4) refers to the normalizing constant in the normal distribution. This means that the Gaussian approximation is $\boldsymbol{x} | \boldsymbol{\theta} \sim \mathcal{N}(\hat{\boldsymbol{x}}(\boldsymbol{\theta}), (-D^2)^{-1})$

6.2 The INLA algorithm

We now have an approximation of the posterior, $\tilde{\pi}(\boldsymbol{\theta}|\boldsymbol{y})$. In order to determine the posterior marginal distributions $\tilde{\pi}(\theta_j|\boldsymbol{y})$ one would, intuitively, proceed by integrating over $\boldsymbol{\theta}_{-j}$ as in 6.1. This, however, is rather time consuming even for $\boldsymbol{\theta}$ of low dimensionality. In the following it will be described how the INLA algorithm circumvents this problem.

6.2.1 Exploring $\tilde{\pi}(\boldsymbol{\theta}|\boldsymbol{y})$

The INLA algorithm's first step is to compute the univariate posterior marginals of $\boldsymbol{\theta}$, that is, $\pi(\theta_j|\boldsymbol{y})$. To do this we explore the posterior for the hyperparameters $\pi(\boldsymbol{\theta}|\boldsymbol{y})$.

i) Locate the mode $\hat{\boldsymbol{\theta}}$ by optimizing $\log \tilde{\pi}(\boldsymbol{\theta}|\boldsymbol{y})$. This can be done via Newton-Raphson or other similar methods.

ii) The Hessian matrix H of $\tilde{\pi}(\hat{\boldsymbol{\theta}}|\boldsymbol{y})$ must be be negative since $\hat{\boldsymbol{\theta}}$ is a maximum. Compute therefore the negative definite Hessian such that H is SPD. H^{-1} corresponds to Σ in a normal distribution. In order to ease computations we standardize $\boldsymbol{\theta}$ as $\boldsymbol{z}$ as follows:

$$\boldsymbol{\theta}(\boldsymbol{z}) = \hat{\boldsymbol{\theta}} + V\Lambda^{1/2}\boldsymbol{z},$$

where $\Sigma = V\Lambda V^T$ is the eigendecomposition of Σ. This means that $\boldsymbol{z} \sim \mathcal{N}(0, I)$ provided that $\tilde{\pi}(\boldsymbol{\theta}|\boldsymbol{y})$ is normal. This reparametrization corrects for scale and rotation which is convenient as it accommodates the standard normal distribution.

iii) Now the exploration of $\tilde{\pi}(\boldsymbol{\theta}|\boldsymbol{y})$ takes place. Starting from $(0, 0)$ in the reparameterized coordinate system and with step size δ_z the difference

$$\log[\tilde{\pi}(\boldsymbol{\theta}(0|\boldsymbol{y}))] - \log[\tilde{\pi}(\boldsymbol{\theta}(z)|\boldsymbol{y})] \tag{6.6}$$

is calculated. Each difference below a certain value, say δ_π, is marked in Figure 6.1 with a filled dot. Thereafter the intermediate points between the black points are evaluated as well, by including them if their value exeeds δ_π. In this way, the heaviest part of the density is identified.

iv) The goal is to approximate the posterior marginal $\pi(\boldsymbol{\theta}_j|\boldsymbol{y})$. A numerically feasible approach is to use the points from step (iii) to interpolate a polynomium and use it to compute the marginals. Higher accuracy can be obtained by taking smaller steps.

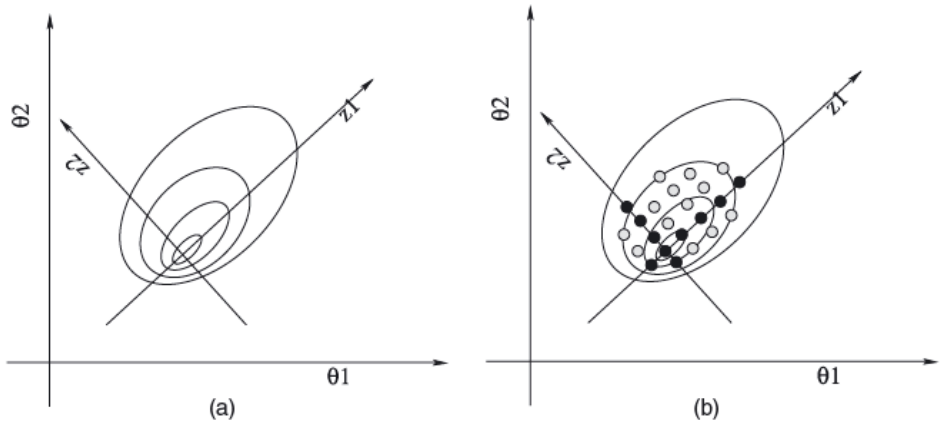

Figure 6.1: *Adapted from [Rue and Martino, 2009]. (a) The mode is located and the z-parametrization is found. (b) With step size δ_π z is explored and filled dots (•) indicates a sufficiently high log-density (see text) and grey dots indicates points filled in from exploration of the values between the black points.*

6.2.2 Approximating $\pi(x_i|\boldsymbol{\theta}, \boldsymbol{y})$

In Section 6.1 $\tilde{\pi}_G(\boldsymbol{x}|\theta, \boldsymbol{y})$ has already been estimated, so an obvious approach to estimate $\pi(x_i|\boldsymbol{\theta}, \boldsymbol{y})$ would be to marginalize x_i in $\tilde{\pi}_G$. According to Rue and Martino [2009] the Gaussian approximation is known to give reasonable results, but this approximation has the disadvantage that it does not correct for skewness due to the properties of the normal distribution.

The Laplace approximation of $\pi(x_i|\boldsymbol{\theta}, \boldsymbol{y})$

A way to improve the Gaussian approximation, is to correct for skewness. The idea is that a Laplace approximation of $\pi(x_i|\boldsymbol{\theta}, \boldsymbol{y})$, similar to the one in (6.2) is made.
The Laplace approximation of $\pi(x_i|\boldsymbol{\theta}, \boldsymbol{y})$ is

$$\tilde{\pi}_{LA}(x_i|\boldsymbol{\theta}, \boldsymbol{y}) \propto \frac{\pi(\boldsymbol{x}, \boldsymbol{\theta}, \boldsymbol{y})}{\tilde{\pi}_{GG}(\boldsymbol{x}_{-i}|x_i, \boldsymbol{\theta}, \boldsymbol{y})} \text{ evaluated in } \boldsymbol{x}_{-i} = \hat{\boldsymbol{x}}_{-i}(x_i, \boldsymbol{\theta}). \tag{6.7}$$

where $\tilde{\pi}_{GG}(\boldsymbol{x}_{-i}|x_i, \boldsymbol{\theta}, \boldsymbol{y})$ is the Gaussian approximation to $\boldsymbol{x}_{-i}|x_i, \boldsymbol{\theta}, \boldsymbol{y}$. $\tilde{\pi}_{GG}$ is infeasible to compute since it must be computed for each observation i. Thus, modifications to (6.7) are needed. The first modification is to approximate the modal configuration with

$$\hat{\boldsymbol{x}}_{-i}(x_i, \boldsymbol{\theta}) \approx \mathrm{E}_{\pi_G}[\boldsymbol{x}_{-i}|x_i] \tag{6.8}$$

where the right hand side can be derived from the distribution $\tilde{\pi}_G$ that has already been found. It is reasonable to approximate the mode of $\boldsymbol{x}_{-i}$ with the mean of

$\tilde{\pi}_G(\boldsymbol{x}_{-i}|x_i)$ since for fixed x_i:

$$\begin{aligned}\tilde{\pi}_G(x_i, \boldsymbol{x}_{-i}) &= \tilde{\pi}_G(\boldsymbol{x}_{-i}|x_i)\pi(x_i) \\ &\propto \tilde{\pi}_G(\boldsymbol{x}_{-i}|x_i) \\ &\propto \exp\left(-\frac{1}{2}(\boldsymbol{x_{-i}} - \mathrm{E}[\boldsymbol{x}_{-i}|x_i])^T\Sigma^{-1}(\boldsymbol{x_{-i}} - \mathrm{E}[\boldsymbol{x}_{-i}|x_i])\right) \qquad (6.9)\end{aligned}$$

where (6.9) is defined in (3.12). And since (6.9) is maximized by $\mathrm{E}[\boldsymbol{x}_{-i}|x_i]$ it is a reasonable choice in (6.8).

The second modification is based on the idea that only x_j's that are sufficiently "connected" to x_i in the graph should influence x_i. A simple way of doing this is to define a "region of interest" around i, $R_i(\theta)$, that defines the marginal of x_i:

$$R_i(\theta) = \{j : |\rho_{ij}(\theta)| > 0.001\}, \qquad (6.10)$$

where ρ_{ij} is the correlation coefficient $\rho_{ij} = \mathrm{Corr}[x_i, x_j]$.

The Simplified Laplace Approximation of $\pi(x_i|\boldsymbol{\theta}, \boldsymbol{y})$

The most efficient algorithm in terms of both precision and speed is the simplified Laplace approximation, which will now be presented.
The simplified Laplace approximation is based on the idea doing a Taylor series expansion of the numerator and denominator in $\tilde{\pi}_{LA}(x_i|\boldsymbol{\theta}, \boldsymbol{y})$ around $x_i = \mu_i(\boldsymbol{\theta})$. The correction for skewness is obtained by using the skew normal distribution $\pi_{SN}(z)$ introduced by Azzalini and Capitanio [1999],

$$\pi_{SN}(z) = \frac{2}{\omega}\phi\left(\frac{z-\xi}{\omega}\right)\Phi\left(a\frac{z-\xi}{\omega}\right)$$

where $\phi()$ and $\Phi()$ are the density and distribution function of the standard normal distribution, and $\xi, \omega > 0$ and a are the location, scale and skewness parameters, respectively.
This skew normal distribution is fitted to the third order expansion of the log of the Laplace approximation $\log[\tilde{\pi}_{SLA}(x_i|\boldsymbol{\theta}, \boldsymbol{y})]$. According to Rue and Martino [2009] it appears that the simplified Laplace approximation is a highly accurate method to compute the posteriors in many observational models.

Chapter 7

Method of Instrumental Variables

The Method of Instrumental Variables is used as a way to combat common problems with OLS estimation. These problems all lead to a violation of the first Gauss-Markov assumption, namely the one that states that $\mathrm{E}(\varepsilon|X) = 0$ where X are the regressors and ε is the error term[1]. If ignored these problems lead to biased and inconsistent estimates.

7.1 Three motivations for the use of Instrumental Variables

In this section the use of instrumental variables is motivated by three examples, all of which are common obstacles to the conventional OLS estimation.

- Measurement error on one or more of the regressors.
- Endogeneity of one or more of the regressors.
- Omitted variable bias.

7.1.1 Measurement error on one or more of the regressors

Consider the simple univariate regression:

$$y = \beta_0 + \beta_1 x_1^* + \epsilon, \tag{7.1}$$

which satisfies all the Gauss-Markov assumptions. Suppose now that x_1 can be expressed as follows

$$x_1 = x_1^* + e, \tag{7.2}$$

[1] This is equivalent to stating $\mathrm{Cov}(\varepsilon, X) = 0$, which will become useful later on.

where e is white noise. If (7.2) satisfies the Gauss-Markov assumptions, specifically the first one, we have that $\mathrm{Cov}(x_1^*, e) = \mathrm{E}(x_1^* e) = 0$. Therefore the covariance between the new regressor and the error term must be:

$$\begin{aligned}\mathrm{Cov}(x_1, e) &= \mathrm{E}(x_1 e) - \mathrm{E}(x_1)\,\mathrm{E}(e) = \mathrm{E}(x_1 e) = \mathrm{E}((x_1^* + e)e) \\ &= \mathrm{E}(x_1^* e) + \mathrm{E}(e^2) = \mathrm{Var}(e) = \sigma_e^2.\end{aligned}$$

This leads to a biased OLS estimate of β_1

$$\begin{aligned}\hat{\beta}_1 \xrightarrow{p} \frac{\mathrm{Cov}(y, x_1)}{\mathrm{Var}(x_1)} =& \frac{\mathrm{Cov}(\beta_0 + \beta_1(x_1 - e) + \epsilon, x_1)}{\mathrm{Var}(x_1)} \\ =& \frac{\mathrm{Cov}(\beta_0, x_1) + \mathrm{Cov}(\beta_1 x_1, x_1) - \mathrm{Cov}(\beta_1 e, x_1) + \mathrm{Cov}(\epsilon, x_1)}{\mathrm{Var}(x_1)} \\ =& \frac{\beta_1\,\mathrm{Cov}(x_1, x_1) - \beta_1\,\mathrm{Cov}(e, x_1)}{\mathrm{Var}(x_1)} \\ =& \frac{\beta_1\,\mathrm{Var}(x_1) - \beta_1\sigma_e^2}{\mathrm{Var}(x_1)} \\ =& \beta_1 - \frac{\beta_1\sigma_e^2}{\mathrm{Var}(x_1)} \\ =& \beta_1\left(1 - \frac{\sigma_e^2}{\mathrm{Var}(x_1)}\right).\end{aligned}$$

So if a measurement error on the regressor is introduced, the OLS estimator is no longer unbiased. In this case of measurement error on a single regressor it can be shown that $0 < 1 - \frac{\sigma_e^2}{\mathrm{Var}(x_1)} < 1$. However in the case of multiple regression very little can in general be said about the direction of the bias of $\hat{\boldsymbol{\beta}}$.

7.1.2 Endogeneity of one or more of the regressors

The problem of endogeneity of one or more of the covariates, sometimes referred to as simultaneity, is illustrated by the following example.

Consider the simple model known from Keynesian economics:

$$C_t = \beta Y_t + \varepsilon_t, \tag{7.3}$$

ie. consumption is proportional to income except noise. Meanwhile we also have this

$$Y_t = C_t + I_t. \tag{7.4}$$

This is a definition, there are no additional parameters, no room for error. Within the context of this simple model it is a fact that income, over time, equals consumption plus investment. Consider now a disturbance of ε_t. This would, due to (7.3), mean a disturbance in C_t, which in turn would mean a change in Y_t due to (7.4). Thus a change in ε_t leads to a change in Y_t, which means that $\text{Cov}(Y_t, \varepsilon_t) \neq 0$. This was shown in Section 7.1.1 to lead to biased estimation of β.

7.1.3 Omitted variable bias

Suppose that $\boldsymbol{y}$ is properly modelled like this:

$$\boldsymbol{y} = X\boldsymbol{\beta} + \boldsymbol{z}\delta + \boldsymbol{\varepsilon}.$$

Since an intercept is included in the model (by a column of 1's in X), it is without loss of generality assumed that $\text{E}(\boldsymbol{z}) = 0$. Suppose that $\boldsymbol{z}$ is unobservable or otherwise unavailable. If one were to omit $\boldsymbol{z}$, it would essentially be included in the error term, like so:

$$\boldsymbol{y} = X\boldsymbol{\beta} + \boldsymbol{\epsilon}. \qquad \text{where } \boldsymbol{\epsilon} = \boldsymbol{\varepsilon} + \delta\boldsymbol{z}$$

If the omitted $\boldsymbol{z}$ is uncorrelated with all the other regressors, this works fine (remember $\boldsymbol{z}$ has zero mean, so the $\boldsymbol{\epsilon}$ is also zero mean). However trouble arises if $\boldsymbol{z}$ in fact is partially correlated[2] with one or more of the regressors, which essentially means that

$$\text{Corr}(\boldsymbol{z}, \boldsymbol{x}_i^T) \neq 0, \qquad \text{for one or more } i \tag{7.5}$$

where $\boldsymbol{x}_i$ is a column of X. In this case $\text{Corr}(\boldsymbol{x}_i^T, \boldsymbol{\epsilon}) \neq 0$ for some i, which exactly is the endogeneity problem of Section 7.1.2.

7.2 Instrumental Variables

To summarize Section 7.1; if you dig a little deeper, it turns out that all three of these challenges essentially boil down to the same underlying problem. Whether you're dealing with measurement error on a variable, an endogenous or an omitted variable, the source of the bias or inconsistency of the parameter estimates is that there is correlation between a variable and the error term.

Consider, then, a model on the familiar form.

$$y = \beta_0 + \beta_1 x_1 + \ldots + \beta_p x_p + \varepsilon.$$

[2]The concept of partial correlation will elaborated in Section 7.2.

Since the three challenges of Section 7.1 are essentially the same, it's useful to think of ε as containing an error term plus an omitted variable, that is $\varepsilon = z + e$. Recall from Section 7.1.3 that this is not, in and of itself, problematic. It is however problematic if this omitted variable is correlated with one of the regressors, say, x_p.

If that's the case, the problem can be solved by finding an instrumental variable to use instead of x_p, or plainly an instrument of x_p. The first, and rather obvious, requirement for such an instrument, x_{IV}, must be that

$$\text{Corr}(x_{IV}, \varepsilon) = 0, \tag{7.6}$$

or we might say that x_{IV} must be exogenous. Another condition is that if x_p is expressed by a least squares linear predictor of all the covariates

$$x_p = \delta_0 + \delta_1 x_1 + \ldots + \delta_{p-1} x_{p-1} + \theta x_{IV} + \nu,$$

with $\text{E}(\nu) = 0$ and ν uncorrelated with $x_0, \ldots, x_{p-1}, x_{IV}$ then θ must be non-zero. In other words: If you run a least squares regression of the endogenous x_p onto all the exogenous variables, the parameter for x_{IV} must be different from zero. In the case of $p = 1$ this is equivalent to saying that x_p and x_{IV} are correlated.

It should be noted that even though an instrumental variable intuitively seems similar to a proxy variable the two are very different. Condition (7.6) requires that x_{IV} and the omitted variable z are uncorrelated, whereas a good proxy is highly correlated with the omitted variable.

Testing whether or not the second condition is true is easy; it's simply a standard t-test of the null hypothesis on the parameter θ after running a standard OLS regression. The first condition, though, is impossible to test directly. Since z is omitted for a reason, namely unavailability, one cannot actually calculate the correlation coefficient between the instrument and the omitted variable. Usually preexisting knowledge of the specific instrumental variable is used to justify the uncorrelatedness of the instrument and error term.

Example 7.1
We saw in Section 7.1.2 that the covariate Y_t was correlated with the error term e_t. An obvious choice as instrument for Y_t would be Y_{t-1}. Clearly the two are correlated (second condition), and it is very reasonable to assume that the error term at a given time is uncorrelated with the income of the previous period, Y_{t-1}, i.e. $\text{Corr}(Y_{t-1}, e_t) = 0$ (first condition) [3].

$\star$

[3] This is possible since, counter intuitively, $\text{Corr}(\cdot, \cdot)$ is not transitive, i.e. $\text{Corr}(a, b) \neq 0$ and $\text{Corr}(b, c) \neq 0 \not\Rightarrow \text{Corr}(a, c) \neq 0$.

All of this leads to the formal definition of instrumental variables.

Definition 7.2 (Instrumental Variable)
Let the regressor x_p be correlated with the omitted variable z, i.e. $Cov(x_p, z) \neq 0$. An intrumental variable x_{IV} for x_p satisfies the following two conditions

1. $\text{Cov}(x_{IV}, z) = 0$
2. $\text{Cov}(x_{IV}, x_p) \neq 0$

We say that x_{IV} is an instrument of x_p.

Chapter 8

The Spatial Autoregression model

Let's now turn to the matter of spatial correlation. It would be useful for the model to take into account any latent spatial structure. We will do so by the the Spatial Autoregression (SAR) model:

$$\boldsymbol{y}_t = X_t\boldsymbol{\beta} + \rho W\boldsymbol{y}_t + \boldsymbol{\varepsilon}, \qquad \boldsymbol{\varepsilon} \sim N_n\left(\mathbf{0}, \tau^2 I\right) \tag{8.1}$$

where X_t is a $n \times p$ matrix of covariates, $\boldsymbol{\beta}$ is a $p \times 1$ vector of regression parameters, ρ is the socalled spatial spillover effect[1] and W is a spatial weight matrix.

8.1 Designing the spatial weight matrix W

The spatial weight matrix W is $n \times n$ and defined as follows:

$$W_{(i,j)} = w_{ij} = \begin{cases} 0 & \text{if} \quad i \nsim j \\ \frac{\phi(i,j)}{\phi(i,\cdot)} & \text{if} \quad i \sim j, \\ 0 & \text{if} \quad i = j \end{cases} \tag{8.2}$$

where $\phi(i,\cdot) = \sum_j \phi(i,j)$, and $\phi(i,j)$ is a function describing the relative position of i and j in the graph. This function could be as simple as $\phi(i,j) = 1$ for all $i \sim j$ or something more complicated like a function of the distance between i and j. Note that the model implicitly assumes that every region has at least one neighbor to avoid dividing by zero. This function and indeed the concept of neighbors are somewhat user defined, and a number of different definitions of neighborhood structure and distance functions will be discussed later on in Chapter 9 and briefly in examples 8.1 and 8.2.

[1] Some restrictions on ρ are neccesary, which will be discussed in Section 8.3.

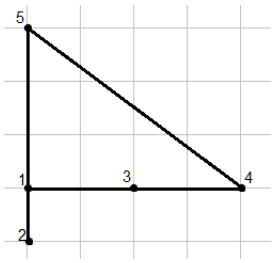

Figure 8.1

Example 8.1
Assume for these examples that $X_t\boldsymbol{\beta} = 0$. Consider the simple graph of Figure 8.1. Two vertices that are connected by an edge are said to be neighbors. If we define $\phi(i,j) = 1$ for all $i \sim j$ we get the following weight matrix

$$W = \begin{pmatrix} 0 & 1/3 & 1/3 & 0 & 1/3 \\ 1 & 0 & 0 & 0 & 0 \\ 1/2 & 0 & 0 & 1/2 & 0 \\ 0 & 0 & 1/2 & 0 & 1/2 \\ 1/2 & 0 & 0 & 1/2 & 0 \end{pmatrix}. \tag{8.3}$$

If values are assigned to each point in the graph, say $\boldsymbol{y}_t = (y_{1t} \dots y_{5t}) = (4, 11, 5, 8, 8)$, we see that in this simple case $W\boldsymbol{y}_t$ is simply the average of the neighbors.

$$\boldsymbol{y}_t = \rho W \boldsymbol{y}_t = \rho \begin{pmatrix} 0 & 1/3 & 1/3 & 0 & 1/3 \\ 1 & 0 & 0 & 0 & 0 \\ 1/2 & 0 & 0 & 1/2 & 0 \\ 0 & 0 & 1/2 & 0 & 1/2 \\ 1/2 & 0 & 0 & 1/2 & 0 \end{pmatrix} \begin{pmatrix} 4 \\ 11 \\ 5 \\ 8 \\ 8 \end{pmatrix} = \rho \begin{pmatrix} 8 \\ 4 \\ 6 \\ 6.5 \\ 6 \end{pmatrix}. \tag{8.4}$$

This is also known as the Besag model for spatial correlation.

$\star$

In the above example every neighbor is considered equally important, i.e. each neighbor of region i has equal impact on y_{it}. Often it makes sense to define different weights to each neighbor. Suppose, for example, the value of your house is related to the values of the neighboring houses. If one neighbor is located 100 yards from your house and another neighbor is located 1000 yards from your house, it seems unrealistic to assign equal weight to both neighbors: Clearly the nearer neighbor should be weighted more heavily.

Example 8.2
Continuing with the graph in Figure 8.1. If we want close neighbors to have large weights and distant neighbors to have small weights, we could define $\phi(i,j)$ by

inverse distance, so that $\phi(i,j) = \text{dist}(i,j)^{-1}$:

$$\Phi_{ij} = \text{dist}(i,j)^{-1} = \begin{pmatrix} 0 & 1 & 1/2 & 0 & 1/3 \\ 1 & 0 & 0 & 0 & 0 \\ 1/2 & 0 & 0 & 1/2 & 0 \\ 0 & 0 & 1/2 & 0 & 1/5 \\ 1/3 & 0 & 0 & 1/5 & 0 \end{pmatrix}, \tag{8.5}$$

which (after dividing by the row sums of Φ) leads to the weight matrix

$$W = \begin{pmatrix} 0 & 6/11 & 3/11 & 0 & 2/11 \\ 1 & 0 & 0 & 0 & 0 \\ 1/2 & 0 & 0 & 1/2 & 0 \\ 0 & 0 & 5/7 & 0 & 2/7 \\ 5/8 & 0 & 0 & 3/8 & 0 \end{pmatrix}. \tag{8.6}$$

Now we calculate $\boldsymbol{y}$

$$\boldsymbol{y_t} = \rho W \boldsymbol{y}_t = \rho \begin{pmatrix} 0 & 6/11 & 3/11 & 0 & 2/11 \\ 1 & 0 & 0 & 0 & 0 \\ 1/2 & 0 & 0 & 1/2 & 0 \\ 0 & 0 & 5/7 & 0 & 2/7 \\ 5/8 & 0 & 0 & 3/8 & 0 \end{pmatrix} \begin{pmatrix} 4 \\ 11 \\ 5 \\ 8 \\ 8 \end{pmatrix} = \rho \begin{pmatrix} 8.82 \\ 4 \\ 6 \\ 5.85 \\ 5.5 \end{pmatrix}. \tag{8.7}$$

$\star$

This, then, illustrates that the design of W is very flexible. Specifically the user can, depending on the available data, define a long array of different weight matrices. In Section 9.2 a number of different weight matrices based on different definitions of neighbor structure will be tested against each other.

8.2 The GMRF properties of the SAR model

To check if this model constitutes a GMRF, we must check to see if this model satisfies the conditions of Definition 3.3. First isolate $\boldsymbol{y}_t$ to arrive at the so-called data generating function of $\boldsymbol{y}_t$.

$$\begin{aligned} \boldsymbol{y}_t &= X_t\boldsymbol{\beta} + \rho W \boldsymbol{y}_t + \boldsymbol{\varepsilon} \Leftrightarrow \\ \boldsymbol{y}_t - \rho W \boldsymbol{y}_t &= X_t\boldsymbol{\beta} + \boldsymbol{\varepsilon} \Leftrightarrow \\ (I_n - \rho W)\, \boldsymbol{y}_t &= X_t\boldsymbol{\beta} + \boldsymbol{\varepsilon} \Leftrightarrow \\ \boldsymbol{y}_t &= (I_n - \rho W)^{-1} (X_t\boldsymbol{\beta} + \boldsymbol{\varepsilon}). \end{aligned}$$

Clearly this means that

$$\begin{aligned}
\mathrm{E}(\boldsymbol{y}_t) &= (I_n - \rho W)^{-1} X_t \boldsymbol{\beta} \\
\mathrm{Var}(\boldsymbol{y}_t) &= \mathrm{Var}\left((I_n - \rho W)^{-1} \boldsymbol{\varepsilon}\right) \\
&= (I_n - \rho W)^{-1} \tau^2 \left((I_n - \rho W)^T\right)^{-1} = \left((I - \rho W)^T (I - \rho W)\right)^{-1} \tau^2 \\
\mathrm{Prec}(\boldsymbol{y}_t) &= (I - \rho W)^T (I - \rho W) \frac{1}{\tau^2} = \left(I - \rho\left(W + W^T\right) + \rho^2 W^T W\right) \frac{1}{\tau^2} = Q.
\end{aligned}$$

Immediately we see that Q is symmetrical. Since Q is symmetrical and can be written as a product of a matrix and its transpose $Q = \left(\frac{1}{\tau^2} I - \frac{1}{\tau^2}\rho W\right)^T \left(\frac{1}{\tau^2} I - \frac{1}{\tau^2}\rho W\right)$ we also have that Q is positive semi definite[2].

According to Definition 3.3, Q must also be positive definite. To see when this is the case, consider the following theorem.

Theorem 8.3
The precision matrix of the SAR model, Q, is positive definite if and only if $\rho\gamma_i \neq 1$ for all i where $\gamma_1, \ldots, \gamma_n$ are the eigenvalues of W.

Proof.
Since we already know that Q is positive semi definite, i.e. $|Q| \geq 0$, we only have to prove that $|Q| \neq 0$. Since $|Q| \neq 0$ is equivalent to $|\tau^2 Q| \neq 0$, we throw out $1/\tau^2$ without loss of generality. We now have

$$\tau^2 |Q| = |\,(I - \rho W)^T (I - \rho W)\,| = |I - \rho W|^2. \tag{8.8}$$

So if $\lambda_1, \ldots, \lambda_n$ are the eigenvalues of $(I - \rho W)$ we have that $|Q| \neq 0$ iff $\lambda_i \neq 0$ for all i. The characteristic equation corresponding to the matrix $I - \rho W$ is

$$|I - \rho W - I\boldsymbol{\lambda}| = |\,-\rho W - (\boldsymbol{\lambda} - 1_n) I| = 0.$$

This means that $(\boldsymbol{\lambda} - 1_n)$ are the eigenvalues of $-\rho W$ which means that $\gamma = (\boldsymbol{\lambda} - 1_n)/-\rho$ are the eigenvalues of W. So if $\lambda_i \neq 0$ we see that $\gamma_i \neq 1/\rho$, which demonstrates that Q is positively definite if and only if $\rho\gamma_i \neq 1$ for all i.

■

This means that for any continuous prior on ρ, Q is almost surely positive definite, making the SAR-model a GMRF.[Mukherjee et al., 2014, p.4]

[2]For Q to be positive semi definite $\boldsymbol{x}^T Q \boldsymbol{x}$ must be ≥ 0.
To check, $\boldsymbol{x}^T Q \boldsymbol{x} = \boldsymbol{x}^T (I - \rho W)^T (I - \rho W) \boldsymbol{x} = ((I - \rho W)\boldsymbol{x})^T (I - \rho W)\boldsymbol{x} = (I - \rho W)\boldsymbol{x} \bullet (I - \rho W)\boldsymbol{x} \geq 0$, since the dot product is a norm.

However obvious it may be, it is important to note that Q is less sparse than W. This is, at first glance, worrisome, since this means that $Q_{ij} \neq 0$ even if regions i and j aren't neighbors. Specifically it means that $Q_{ij} \neq 0$ if i and j are neighbors *or* if they have a common neighbor, which suggests a second-order dependence between i and j. This would seem to be in conflict with the Markov assumption, that specifically says that $Q_{ij} = 0 \Leftarrow i \nsim j$.

This, however, illustrates that there are essentially two graphs at play. On one hand we have the actual graph, the W-graph. This the one with the n regions, i.e. the one $\phi(i, j)$ is defined on. On the other hand we have the more abstract graph, the Q-graph. This can be thought of as a graph induced by the choice of Q i.e. the choice of model. There is no assumption that Q_{ij} must be zero if i and j are not W-neighbors. The Markov assumption only relates to the Q-graph. This means that if i and j are not neighbors, while sharing a common neighbor, they are Q-neighbors but not W-neighbors.[Mukherjee et al., 2014, p.4]

This then is a feature, not a flaw, of the GMRF setup, since this allows for much more complex dependence structures than it would seem at first glance. It does however mean that for very connected W-graphs, Q may be relatively dense, which could slow down the calculations in INLA.

8.3 Restrictions on ρ

So far we have only made one assumption on ρ namely that $\rho \neq 1/\gamma_i$ for all i. This, as mentioned in Section 8.2, is not much of a restriction since any continuous prior on ρ guarantees this to be true with probability 1. This, however, doesn't mean that ρ can roam freely on $\mathbb{R}^n$. Recall that the eigenvectors of W are given by

$$\gamma_i = \frac{\lambda_i - 1}{-\rho},$$

where λ_i is the ith eigenvector of the matrix $(I_n - \rho W)$. We know that

$$\lambda_i = 1 - \rho\gamma_i \tag{8.9}$$

must be positive, otherwise Q couldn't be positive definite (see (8.8)). So in order to guarantee that $|Q| > 0$ we must have that

$$\lambda_i = 1 - \rho\gamma_i > 0 \qquad \forall i.$$

If $\gamma_i > 0$ this means that $-\infty < \rho < 1/\gamma_i$, and if $\gamma_i < 0$ then $-1/\gamma_i < \rho < \infty$. In other words

$$\frac{1}{\gamma_{max}^-} < \rho < \frac{1}{\gamma_{max}^+}, \tag{8.10}$$

where γ^+_{max} is the largest positive eigenvalue, and γ^-_{max} is the numerically largest negative eigenvalue. So in addition to $\rho \neq 1/\gamma_i$ we must also have that $1/\rho$ is between the largest and smallest eigenvalue of W.

The smallest possible interval is given by $\rho \in]-1, 1[$. This is equivalent to saying that the $\gamma_i \in [-1, 1]$ per (8.10). To see this consider the following definition and theorem.

Definition 8.4 (Gersghorin discs)
Let A be a complex $n \times n$ matrix with entries a_{ij}. For $i = 1, \ldots, n$ let $R_i = \sum_{j \neq i} |a_{ij}|$. The circle $C(a_{ii}, R_i)$ is then called a Gersghorin disc.[Varga, 2004, p.1]

In other words, Gersghorin discs are n circles in the complex plane centered in each diagonal element with radius equaling the sum of the absolute values of the off-diagonals.

Theorem 8.5 (Gersghorin's circle theorem)
Let A be a complex $n \times n$ matrix. Every eigenvalue of A falls within at least one of the Gersghorin discs.[Varga, 2004, p.4]

Recall that W always has zeroes along the main diagonal, and that the rows of W all sum to 1. This makes all the Gersghorin discs for W the same, namely $C(0, 1)$. All eigenvalues must lie within this circle, which means that $\gamma_i \in [-1, 1]$.

Furthermore we can prove that at least one of the eigenvalues of W must be exactly 1. Which is to say that at least one of $|\gamma^-_{max}| = 1$ and $|\gamma^+_{max}| = 1$ must be true. This can be proven by the Perron-Frobenius theorem for irreducible matrices.[Per, July 2014]

In some literature ρ is, without further ado restricted to $0 \leq \rho \leq 1$. With the above argument in mind, this seems dangerous. This restriction, however, is surely motivated by a need for proper interpretation of ρ. Consider the following simple example.

Example 8.6
Assume that $\rho < 0$ and regions i and j are neighbors. This would mean a rise in y_j would lead to a fall in y_i. Suppose j is also neighbor to k. A fall i y_k would then lead to a rise in y_j, which in turn would lead to a fall in y_i. All is OK so far. In a spatial setting however it is possible for i and k to be neighbors as well see Figure 8.2. This would suggest that a fall in y_k would lead to a rise in y_i. We now have a contradiction.

$\star$

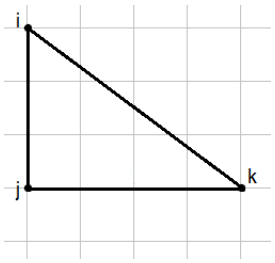

Figure 8.2

This example illustrates that if $\rho < 0$ we may end up inferring that two variables y_i and y_k are positively and negatively correlated at the same time. So for interpretations of ρ to make sense it must be positive.

8.4 The temporal case

At this point the model can only handle cross-sectional data, i.e. data that only concerns one time period. The model can quite easily be extended to handle panel data i.e. spatial-temporal data. In the case of panel data the same region is measured repeatedly. Lets say we have n regions and t measurements of each region. The data vector $\boldsymbol{y}$ then is on the form

$$\boldsymbol{y} = (y_{(1,1)}, y_{(2,1)}, \ldots, y_{(n,1)}, y_{(1,2)}, \ldots, y_{(n,2)}, \ldots, y_{(1,t)}, \ldots, y_{(n,t)})^T,$$

which is to say that the data is sorted by time and then region. This vector has length nt and is clearly too long to be multiplied with W which is only $n \times n$. Define now the new weight matrix

$$\hat{W} = I_t \otimes W, \tag{8.11}$$

where $\otimes$ is the Kronecker product. Then $\hat{W}$ is $nt \times nt$ and can be multiplied with $\boldsymbol{y}$.

This changes none of the considerations in the previous sections. It's clear that $\hat{W}$ and therefore Q is vary large in many cases. Because of the extreme sparseness of $\hat{W}$ however, this is only slightly more expensive that the non-temporal case to run in INLA.

Chapter 9

Data processing

This chapter deals with the processing of a large data set[1]. The focus of the data processing will be on the design of the spatial weight matrix W and its impact on the overall quality of the model and the estimate of ρ in the SAR model, see Chapter 8 for theoretical details. So even though the model includes a long list of covariates, the coefficients of these will consequently be of very little interest. For an analysis of the sign and magnitude of the coefficients see [Bech and Lauridsen, 2009].

9.1 Preparing the data

The data set consists of 270 time series, one for each municipality $m_1, \ldots, m_{270}$ in Denmark[2]. Each time series covers 8 years from 1997 to 2004. Actually we have partial data for more years, but the data for outpatient hospital admissions is only complete for 1997 and onwards. So in order to get a balanced data set we consider only that time period. The dependent variable is in this case each m_i's expenditure on general physicians (GPs). The independent variables is a collection of 78 economic, socio-economic, geographic and demographic variables recorded for each m_i every year. These covariates and their corresponding parameters are of no interest to this project. Results are given in[Bech and Lauridsen, 2009], along with economic arguments for the inclusion of the 78 covariates.

The data is lagged once relative to the dependent variable. This of course takes us down to 7 years worth of data, leaving us with 7 measurements of 270 municipalities, i.e. we have $270 \times 7 = 1890$ observations.

[1] The data was graciously supplied by Jørgen T. Lauridsen at the Institute of Public Health at University of Southern Denmark.

[2] The five municipalities of Bornholm are ignored because of the remoteness of this particular region

Bech and Lauridsen argue that three covariates are endogenous, namely the variables describing the number of GPs, the number of inpatient hospital admissions paid for by the municipality and the number of outpatient hospital admissions paid for by the municipality. According to Section 7.2 on page 29 we can solve the issues associated with this endogeniety by lagging those variables one more time. Since the outpatient variable is the one limiting the size of the data set, another lagging of this variable would mean losing another year of data, or at the very least losing the balancedness of the data.

The solution that has been settled for is lagging the three covariates, and replacing the missing values for outpatient admissions with a set of predicted values. The predicted values are obtained by regressing outpatient admissions linearly on two other variables: The year and a factor for municipality id. This yields a model with extremely significant coefficients and a coefficient of determination $R^2 = 0.89$. The fitted values from this model is used as substitutes for the missing values. The predicted year is year 0 (corresponding to the year 1998) in the plot in Figure 9.1.

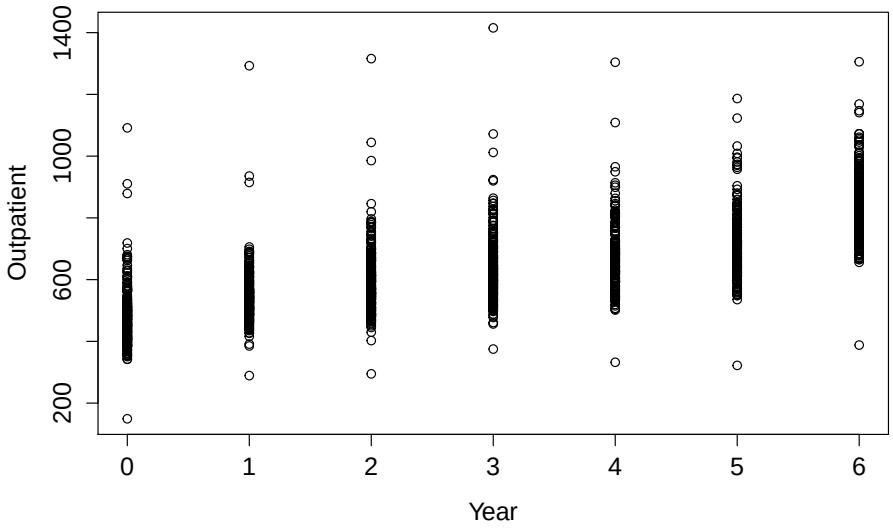

Figure 9.1: *The data for year 0 is predicted*

In addition to all of this we also have two matrices containing spatial information for the 270 municipalities. One matrix describes which municipalities share a border, and the other is a dense matrix with the distance between each pair of municipalities in kilometers. These enable us to design a huge number of different Ws, as we shall see in the following sections.

9.2 Regressions with four different W

There are, as mentioned, in Section 8.1, different ways to design the spatial weight matrix W. Recall from Section 8.1 that W is generally defined by

$$W_{(i,j)} = w_{ij} = \begin{cases} 0 & \text{if} \quad i \nsim j \\ \frac{\phi(i,j)}{\phi(i,\cdot)} & \text{if} \quad i \sim j, \\ 0 & \text{if} \quad i = j \end{cases} \tag{9.1}$$

where $\phi(i,\cdot) = \sum_j \phi(i,j)$, and $\phi(i,j)$ is a function describing the relative position of i and j in the graph. One way of doing this is to define $\phi(i,j)$ to be equal to one if and only if the municipalities i and j share a border on the map. We might call this scheme 1. Under scheme 1 we then have that

$$W^{s1}_{(i,j)} = w^{s1}_{ij} = \begin{cases} 0 & \text{if} \quad i \nsim j \\ \frac{1}{n_i} & \text{if} \quad i \sim j, \\ 0 & \text{if} \quad i = j \end{cases} \tag{9.2}$$

where n_i is the number of municipalities sharing a border with m_i.

Another scheme is to substitute the $\sim$-operator for a measure of distance. Lets define that $i \sim j$ if and only if m_i and m_j are located within d kilometers of each other. We call d the radius of interest. This scheme, scheme 2 say, obviously leads to a different weight matrix:

$$W^{s2}_{(i,j)}(d) = w^{s2}_{ij} = \begin{cases} 0 & \text{if} \quad \delta(i,j) \geq d \\ \frac{1}{n_i} & \text{if} \quad \delta(i,j) < d, \\ 0 & \text{if} \quad i = j \end{cases} \tag{9.3}$$

where $\delta(i,j)$ is the distance between m_i and m_j.

A third scheme is to define the weight matrix as seen below

$$W^{s3}_{(i,j)} = w^{s3}_{ij} = \begin{cases} 0 & \text{if} \quad i \nsim j \\ \frac{1}{\delta(i,j)} & \text{if} \quad i \sim j, \\ 0 & \text{if} \quad i = j \end{cases} \tag{9.4}$$

i.e. weight by inverse distance if m_i and m_j share a border on the map.

The fourth scheme follows intuitively from the previous two:

$$W^{s4}_{(i,j)}(d) = w^{s4}_{ij} = \begin{cases} 0 & \text{if} \quad \delta(i,j) \geq d \\ \frac{1}{\delta(i,j)} & \text{if} \quad \delta(i,j) < d. \\ 0 & \text{if} \quad i = j \end{cases} \tag{9.5}$$

9.2.1 Results

We now run the four models in INLA, and report DIC, the estimate $\hat{\rho}$ and the standard deviation on $\hat{\rho}$. In order to avoid dividing by zero during the construction of W^{s2} and W^{s4}, the radius of interest d must be larger than 40[3]. For the purposes of these runs d has arbitrarily been set to 60. The results are presented in table 9.1. CPO and PIT are also calculated and plotted in Appendix A. None of the CPO and PIT plots are critical for our choice of mode: The PITs are fairly close to uniformly distributed, and relatively few of the CPOs have low probability i.e. there are relatively few outliers in the data.

Scheme	DIC	$\hat{\rho}$	sd
s1	-5785	0.30	0.026
s2	-5304	0.41	0.038
s3	-6376	0.29	0.031
s4	-6195	0.41	0.042

Table 9.1: *Results from INLA runs for each of the four weighting schemes.*

DIC fairly clearly suggests that scheme 3 does the better job modeling the data.

Two questions comes to mind. First the obvious question: Will a change in d improve the models s2 and s4? Secondly: By weighting by inverse distance in s3 and s4, we assume that the "importance" of a neighbor declines like $1/x$ by distance. Maybe the decline is faster or maybe it is slower. To try and answer these questions we go to the next section.

9.3 Tuning of the model

In this section we will discuss how to best choose the radius of interest d and how quickly the importance of neighbors should decline as a function of distance. For this investigation a fifth and final weighting scheme is introduced.

$$W^{s5}_{(i,j)}(d,k) = w^{s5}_{ij} = \begin{cases} 0 & \text{if} \quad \delta(i,j) \geq d \\ \frac{1}{\delta(i,j)^k} & \text{if} \quad \delta(i,j) < d. \\ 0 & \text{if} \quad i = j \end{cases} \tag{9.6}$$

This gives us two tuning parameters: d and k. So really this s5 is really a family of infinitely many schemes, since there's a new scheme for each pair of d and k. Note that s4 is really a special case of s5 with $k = 1$. For example, by choosing a large value of k, we assume that the importance declines very quickly, i.e. the

[3]We divide by the row sums of W, during the construction of W^{s2} and W^{s4}, so in order to avoid dividing by zero, every region must have at least one neighbor. Skagen is the most isolated municipality, so since it's nearest neighbor is 40 kilometers away, we must have $d > 40$.

distant neighbors of m_i is given a small weight. We therefor name k the distance penalty parameter.

To the author's knowledge no such weighting scheme has previously been contemplated in any existing literature. Most literature consider only s1. Some discuss briefly the possibility of using schemes s3, i.e. weighting by inverse distance, but none of the reviewed literature mentions weighting by a different function of δ.

This model has been run for every combination of 21 values of $d = \{40.1, 45, 50, 55, \ldots, 140\}$ and 26 values of $k = \{0.5, 0.6, \ldots, 2.9, 3\}$, resulting in $21 \times 26 = 546$ runs.

If we plot DIC vs. d and k in a 3D plot, we would expect, or at least hope, to see an up-side-down cone, with a global minimum at some point (d, k). The thinking here is that if a large enough range of values d and k are tested, we'd expect DIC to converge towards some "optimal" combination of d and k. As can be seen in Figure 9.2 the results are inconclusive in the sense that there is seemingly no pattern to which values of d and k yields the better model, i.e. lower DIC. There might be a slight tilt towards higher values of k, but the difference is tiny.

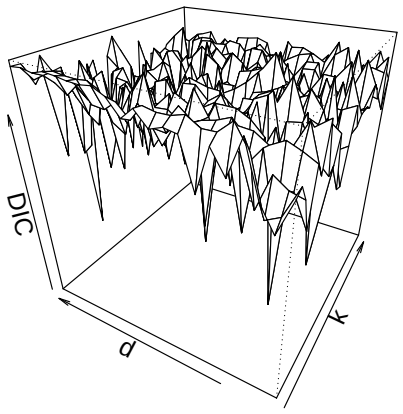

Figure 9.2: *DIC vs. d and k.*

Things do not improve much if we look at the marginal plots as in Figure 9.3. The two plots have been fitted with a second-order ploynomial, since we would expect a cone-shape. The parameters for the polynomial fitted to d are barely

significant ($p = 0.067$ and $p = 0.095$), and the polynomial, as seen on the plot, has the unexpected sign. This does not help us in choosing d at all.

The polynomial fitted to k however is slightly more interesting. First of all it has the expected sign. Second of all the estimates of the parameters are much more significant: Both have p-values $< 10^{-11}$. The polynomial has its minimum at $k = 1.86$. Even if it seems the consequence of choosing k incorrectly is small, it at least seems reasonable to choose this value of k.

If we choose $k = 1.86$ and, in lack of a better suggestion, $d = 70$, we get a DIC of -6634 and an estimate of the spatial spill-over $\hat{\rho} = 0.34$ (sd $= 0.026$). CPO and PIT are plotted for this model in Figure 9.4 and a histogram of the residuals is plotted in Figure 9.5.

9.4 Sensitivity of $\hat{\rho}$

If $\hat{\rho}$ is plotted vs. d and k as in Figure 9.6, we observe a very interesting phenomenon. First off we see that $\hat{\rho}$ fluctuates seemingly at random as a function of d, suggesting, yet again that the the output from the model is sensitive to the choice of weighting scheme for W. But plotting $\hat{\rho}$ by k is more interesting. There seems to be a clear connection between k and $\hat{\rho}$.

This suggests that, not only is $\hat{\rho}$ sensitive to the choice of scheme, it is also sensitive in a systematic way - at least for the class of models utilizing scheme 5.

This calls the use of $\hat{\rho}$ into question. We may be able to construct confidence intervals and significans tests for $\hat{\rho}$, but those are measures of confidence for the estimates for a new set of data. These measures say nothing about the robustness of $\hat{\rho}$ as a function of W.

Not only does $\hat{\rho}$ fluctuate wildly as a function of d, the pattern in the dependence on k essentially allows the user to choose $\hat{\rho}$, by making seemingly innocent changes in k.

It seems that $\hat{\rho}$ is somewhere in the interval between 0.20 and 0.45, but which one of these is the best one? Presumably the one resulting from the "optimal" $k = 1.86$. Since no one, to my knowledge, has tried weighting scheme 5, any other statistician would've picked s4, i.e. $k = 1$. This imaginary analyst would, with great confidence, report $\hat{\rho} = 0.39$ with a standard deviation of 0.024, whereas we would, with equally great confidence, report $\hat{\rho} = 0.34$ with a standard deviation of 0.026. In this example the difference may seem negligible, but the difference could be much larger depending on what value of d is chosen by the user. Very

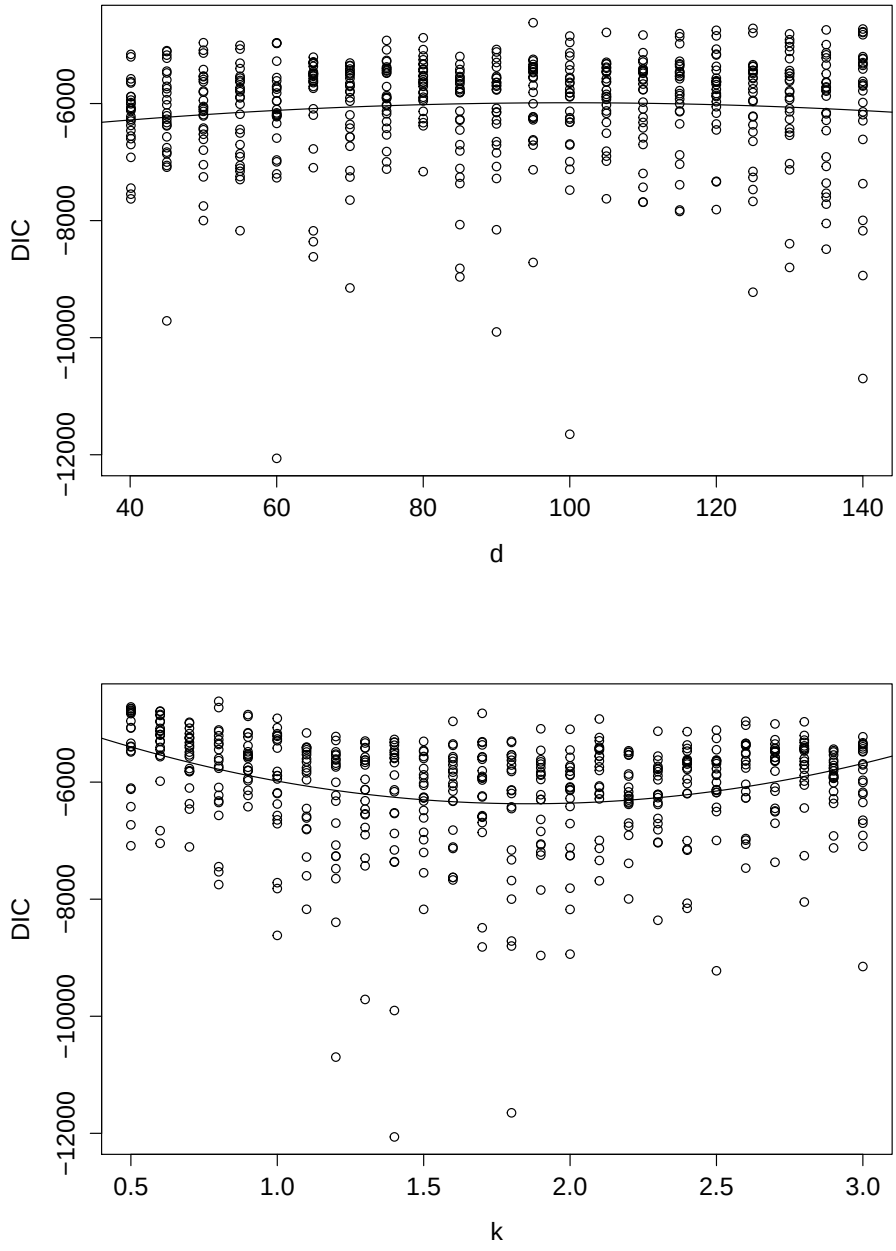

Figure 9.3: *DIC vs. d and k respectively.*

little, if any at all, of the existing literature discuss this problem.

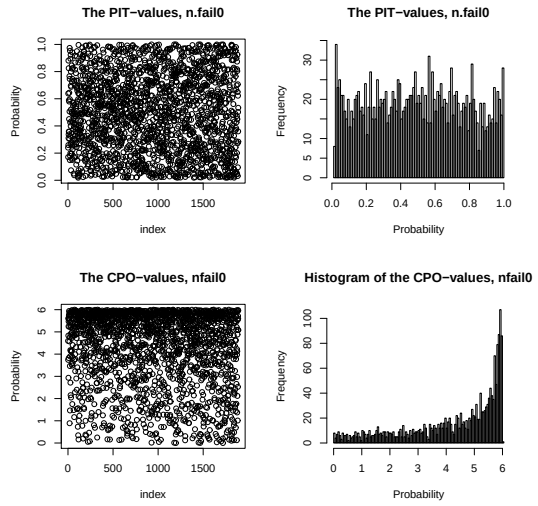

Figure 9.4: *CPO and PIT of the model utilizing scheme 5, with* $k = 1.86$ *and* $d = 70$*.*

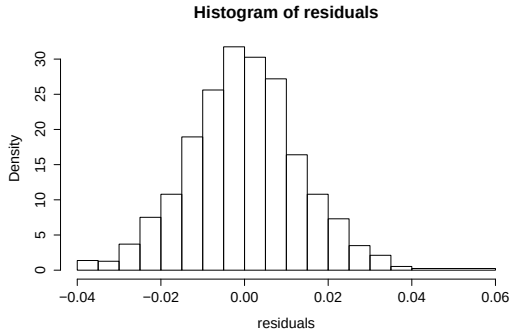

Figure 9.5: *Histogram of the residuals of the model utilizing scheme 5, with* $k = 1.86$ *and* $d = 70$*.*

9.5 Illustrations of the restrictions on ρ

As we discussed at great length in Section 8.3, we know a few things about the extreme eigenvalues of W. We know that these extremes γ_{max}^{-} and γ_{max}^{+} must be in the interval $[-1, 1]$ and that at least one of them must be on the boundary of that interval.

It turns out that all the γ_{max}^{+} of the 546 model runs equaled 1[4]. The smallest eigenvalues are plotted in Figure 9.7, and it shows a rather beautiful connection between γ_{max}^{-}, k and d. Beyond the fact that none of them are below -1, it is not

[4]Or at least so close to 1, that `.R` runs out of precision and rounds up to 1.

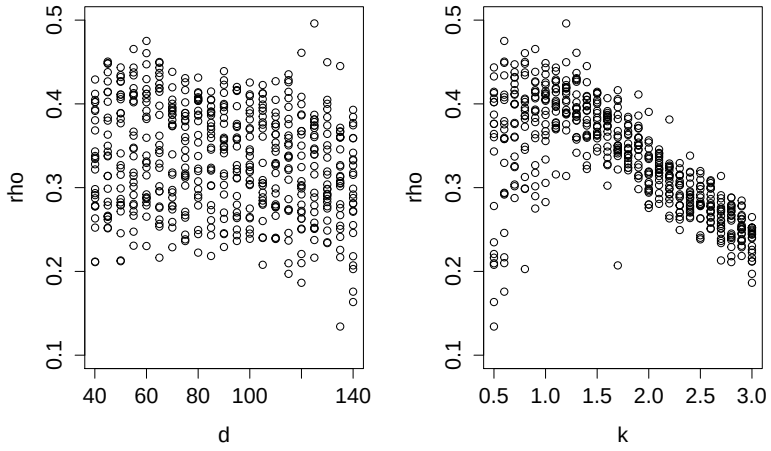

Figure 9.6: *$\hat{\rho}$ vs. k and d respectively.*

entirely clear why this aesthetically pleasing connection holds, so it only serves as an illustration that beyond what Gersghorin and Perron-Frobenius tells us about the eigenvalues of W there is still something to be learned.

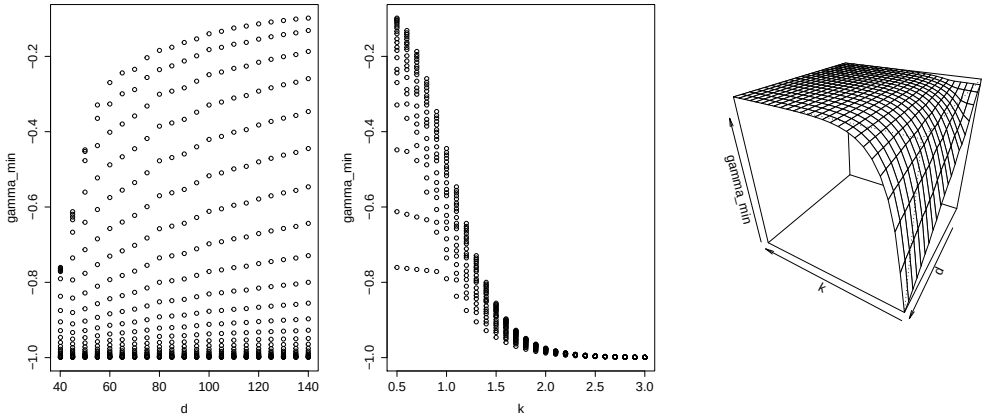

Figure 9.7: *The smallest eigenvalue of W for each of the 546 model runs.*

9.6 Summarizing the data processing

In summary, this chapter shows some discouraging results regarding the SAR model. First we see that the quality of the model, as measured by DIC, is very sensitive to the choice of weighting scheme, as it fluctuates wildly, and without

pattern, as a function of the tuning parameters k and d. The lack of pattern is worrisome since it means that DIC provides little help choosing an "optimal" pair of parameters d and k.

Secondly, and perhaps even more worrisome, we see that $\hat{\rho}$ is also sensitive to the choice of weighting scheme. This means that a slightly different W may result in a vastly different $\hat{\rho}$. There is some pattern to be found, since $\hat{\rho}$ is closely connected to k. This sensitivity of $\hat{\rho}$, is problematic since the significance tests we use for $\hat{\rho}$ fail to take variation in weighting scheme into account, i.e. even if we report a $\hat{\rho}$ with a very narrow confidence interval, a small change in d may completely change the estimate.

Lastly we note that the eigenvalues of W behave as expected, but with a very interesting pattern structure.

Chapter 10

Conclusion

During this thesis, the INLA approach has been introduced and applied to a SAR model. The SAR model has been introduced and a number of results has been shown for this type of spatial model.

Particularly a lot of attention was paid to the spatial spill-over parameter ρ and the restrictions placed on it - both those necessary from an algebraic standpoint, and the more convenient ones that arises from the need to consistently interpret the parameter. The author also notes the discrepancy in how this subject is treated in the literature, and attempts to explain why different statisticians prefer to assume different restrictions on ρ.

The data processing set off as an exercise in model choice, without much success. It turned out that model quality (as measured by DIC) fluctuates wildly and unpredictably as a function of the two parameters radius of interest d and the distance penalty parameter k.

The data processing then turned into an illustration of a particular weakness of ρ, namely its dependence on the specification of the spatial weight matrix W. This weakness raises serious questions about how we report and interpret the estimated parameter $\hat{\rho}$. To the author's knowledge no such problem has been pointed out before. This is certainly an area, that needs more research.

Appendix A

INLA output

Here are the CPO and PIT plots supplied by INLA for each of the four schemes.

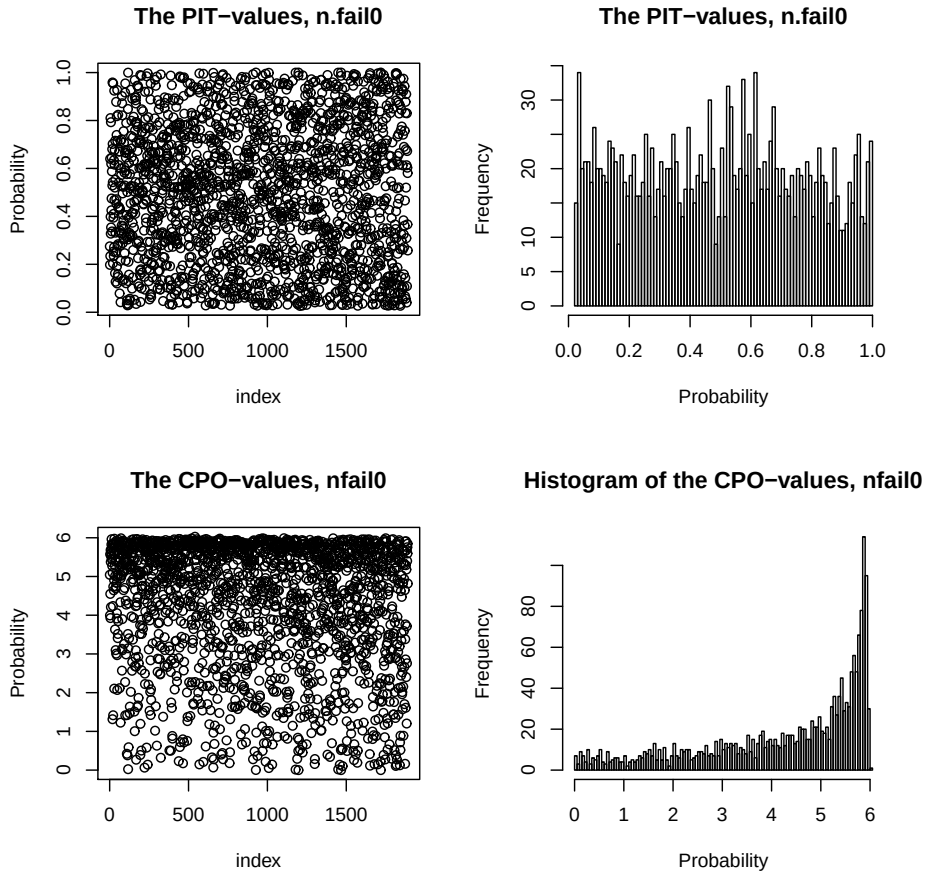

Figure A.1: *Plots of CPO and PIT for scheme 1*

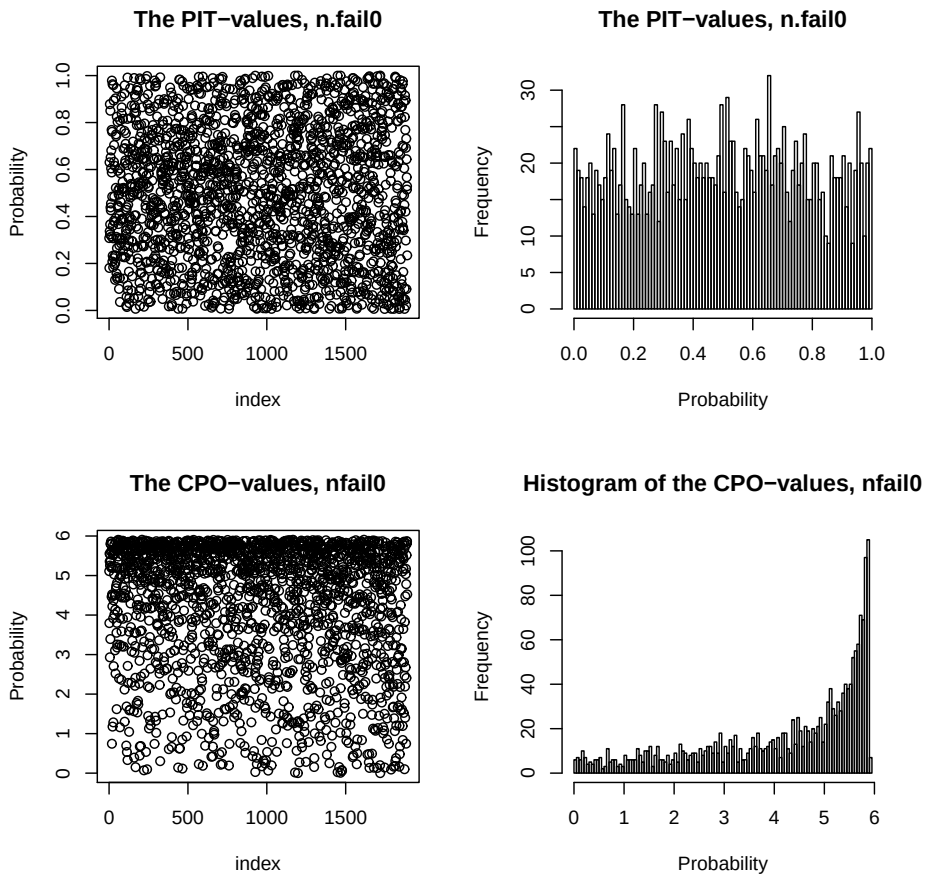

Figure A.2: *Plots of CPO and PIT for scheme 2*

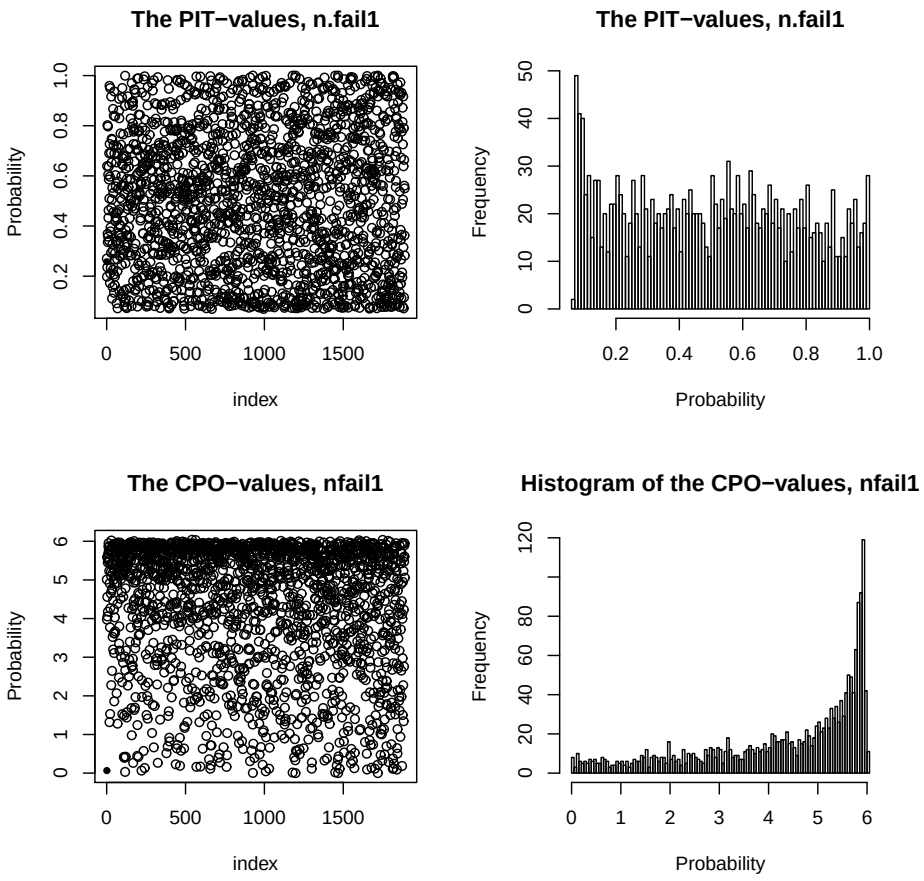

Figure A.3: *Plots of CPO and PIT for scheme 3*

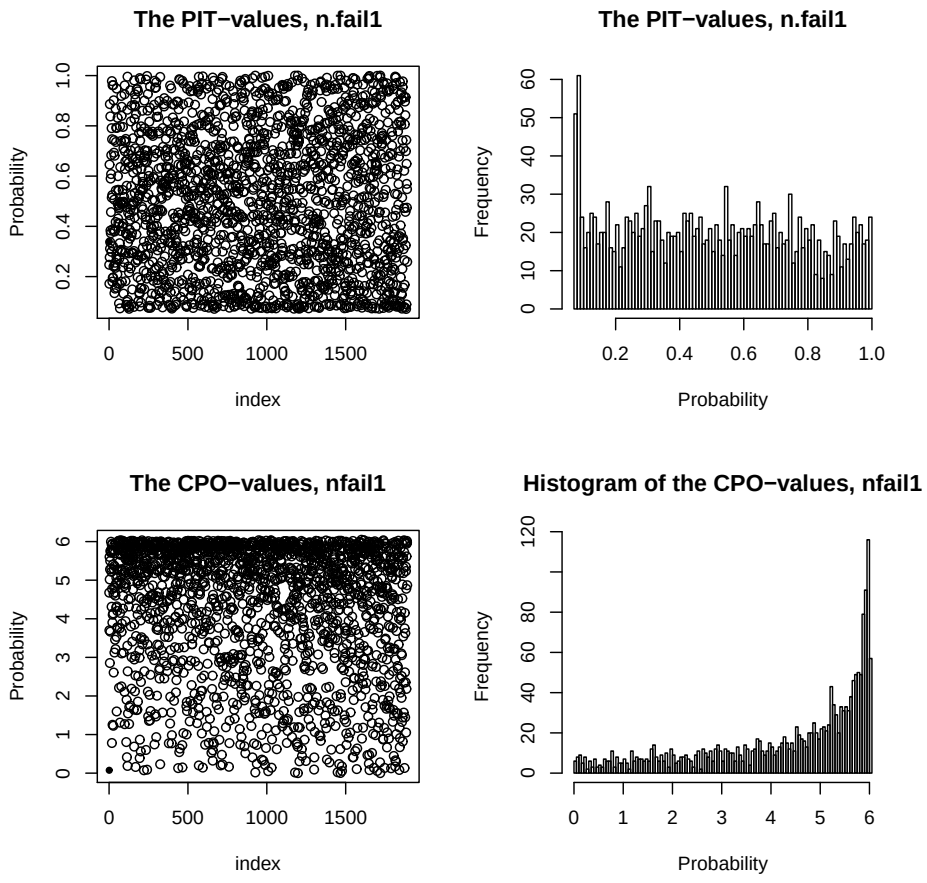

Figure A.4: *Plots of CPO and PIT for scheme 4*

Bibliography

Perron-frobenius' theorem for irreducible matrices. Wikipedia, July 2014. URL http://en.wikipedia.org/wiki/Perron%E2%80%93Frobenius_theorem#Perron.E2.80.93Frobenius_theorem_for_irreducible_matrices.

A. Azzalini and A. Capitanio. Statistical applications of the multivariate skew normal distribution. *Journal of the Royal Statistical Society. Series B (Statistical Methodology)*, 61(3):pp. 579–602, 1999. ISSN 13697412. URL http://www.jstor.org/stable/2680724.

Adelchi Azzalini. *Statistical Inference - Based on the likelihood*. Chapman & Hall/CRC, 1996. ISBN 978-0412606502.

Gianluca Baio. An introduction to inla with a comparison to jags, 5 2013. URL http://www.statistica.it/gianluca/Talks/INLA.pdf. Slides from BAYES2013.

Michael Bech and Jørgen Lauridsen. Exploring spatial patterns in gp expenditure. *European Journal of Health Economics, 10, 2009, 243-54*, 10:243–254, 2009.

Andrew Gelman, John B. Carlin, Hal S. Stern, and Donald B. Rubin. *Bayesian Data Analysis*. Chapman and Hall/CRC, 2003. ISBN 158488388X.

C. Mukherjee, P. S. Kasibhatla, and M. West. Spatially-varying sar models and bayesian inference for high-resolution lattice data. *Annals of the Institute of Statistical Mathematics*, 66:473–494, 2014. URL http://ftp.stat.duke.edu/WorkingPapers/11-02.pdf.

Håvard Rue and Leonhard Held. *Gaussian Markov Random Fields: Theory and Applications*. Chapman & Hall/CRC, 2005. ISBN 1-58488-432-0.

Håvard Rue and Sara Martino. Approximate bayesian inference for latent gaussian models using integrated nested laplace approximations. *Journal of the Royal Statistical Society*, 2009.

D. J. Spiegelhalter, N.G. Best, B. P. Carlin, and A. van der Linde. Bayesian measures of model complexity and fit (with discussion). *Journal of the Royal Statistical Society*, 2002.

R.S. Varga. *Gerghorin and His Circles*. Springer Series in Computational Mathematics. Springer, 2004. ISBN 9783540211006.

C. H. Vestergaard. The inla approach for spatially structured data. At AAU.dk (login required), December 2013. URL `http://projekter.aau.dk/projekter/files/168641932/main1.pdf`.