Modelling Gaussian Fields and Geostatistical Data Using Gaussian Markov Random Fields

Outline

Chapter 1

Introduction

1.1 Spatial Data and Spatial Problems

This thesis discusses the theory behind a few ways of modelling spatial correlation and estimating the models' parameters.

To begin, we first need a clear understanding of the kinds of spatial data we might face and the kinds of spatial problems we may want to solve, since these topics dictate what kinds of models would be useful. There are a few different kinds of spatial data we may have.

First, our data might be *geostatistical* or *point-level* data, such as weather data from a set of monitoring stations. We may then be interested in the interpolated weather at any point in the space.

In contrast, with *lattice* or *areal* data our data is in the shape of a grid. In regular lattices, our observations are located at regular intervals in different dimensions across the space; we can also estimate irregular lattices. Remote sensing data, in particular, often comes in the form of tiles, which can be easily thought of as lattices. We can also think of political boundaries (*e.g.* wards) defining an irregular lattice, though this may introduce new difficulties, such as if these plots are of grossly different sizes. We can also have a lattice with missing data, such as where weather precludes some observations from being made.

Finally, the locations themselves could be the random variables of interest, such as the locations where a species of animal is observed. These data are called *point process data* or *point patterns.*

Comparing spatial data to time-series data, the closest analogue to time-series data would be the case in which we have observations along a single spatial line. However, even in this example, we have more structure in time-series data; time has an arrow.

We can also have data in which we have both spatial and temporal correlation. We can extend spatial models fairly straight-forwardly to take temporal correlation into account. For example, if we have lattice data and each datapoint has a set of spatial neighbours, we can define the set of neighbours of a node to be the union of the set of its spatial neighbours and its temporal neighbours, as illustrated below.

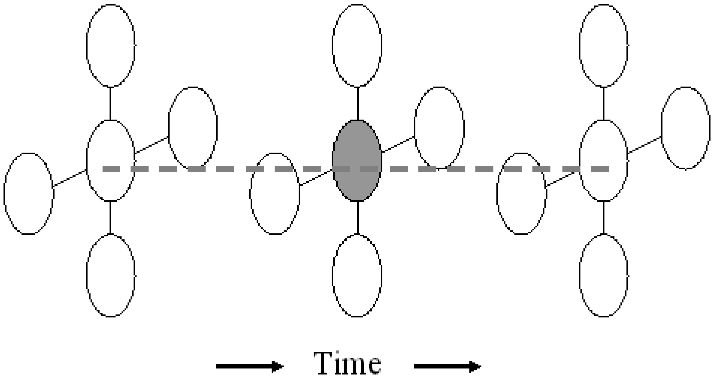

Figure 1.1: Lattice data with the grey node's neighbours in time highlighted by the dashed line.

In this thesis, I will focus largely on lattice data structures but also somewhat on geostatistical point-level data. Again, to contrast lattice and geostatistical data, with lattice data we are not concerned with the possibility of a measurement in between adjacent nodes. There are, however, obvious links between geostatistical and lattice data. In particular, we could think about geostatistical data as being on a very finely spaced lattice, with many missing observations, or we could aggregate the geostatistical data up to a grid. Models for lattice data may then be used for geostatistical data.

1.2 Methods of Modelling

There are two main approaches to spatial modelling.

Geostatistical models are applied to continuous spatial processes and typically are a function of distance and direction between geostatistical data.

Gaussian Markov Random Field models (GMRFs) model the data as being related to each other through an undirected graph. I will focus on the latter kind of model in this thesis.

The concept of GMRFs sprung from attempts to generalize a specific model put forth by the physicist Ernst Ising. Ising (1925), building on work by Lenz (1920), considered sequences of points on a line in which each point had a certain "spin". Ising models attempt to assign a probability measure to a sample space that cannot be observed, with the probability measures called Gibbs measures. *Pairwise* interactions were of initial interest in statistical physics, such as in describing gravitational and electrostatic forces, but the idea of these spatially-dependent interactions was later extended to encompass a broader area. Spitzer (1971) and Hammersley and Clifford (1971) generalized the result linking Gibbsian ensembles with Markov fields which led to GMRFs.

When might GMRFs be appropriate? GMRFs can handle data on both regular and irregular lattices. This is important since, although much attention is focused on regular lattices, these lattices do not usually arise naturally but are man-made. However, even much spatial data that appears to be in a regular lattice structure, such as tiled satellite data, may not be well estimated by GMRFs since each datapoint actually represents many (possibly unknown) values in a larger plot which have somehow been averaged or otherwise collapsed into a single value. Whittle (1954) notably discussed this as a problem in fitting a dataset from Mercer and Hall. The size of the plot that is collapsed can affect how well GMRFs can fit the data (Besag, 1974). Further, unless a dataset includes a large number of sites which are mostly neighbours, it may not be well-suited to being described by a lattice. This is because the data are presumably situated within a larger world, and if there are spatial relationships within the data there are likely relationships between those sites near the borders and sites outside the dataset, on which there are no data. Despite this, lattices have been successfully applied to carefully controlled and constructed agricultural field trials (Ripley, 1989).

GMRFs can also be useful even when we might otherwise think to use a geostatistical model, since they are less computationally-intensive than fitting the geostatistical model directly. Indeed, many have tried to use GMRFs to estimate geostatistical models and vice versa, a topic we will explore later in this thesis. Besag (1981) showed that the covariance function of a GMRF could be approximated by a Bessel function that is modified so that it decreases monotonically with distance. Griffith and Csillag (1993) tried to minimize the squared differences between the covariances of geostatistical models and GMRFs. Hrafnkelsson and Cressie (2003) showed that a class of GMRFs was approximated in an empirical setting by a geostatistical model that used a Matérn covariance model. Rue and Tjelmeland (2002) and Cressie and Verzelen (2008) represent the most recent efforts at finding good approximations through a distance minimization approach. Finally, looking for an entirely new method of approximating GGMs with GMRFs, Lindgren, Lindström and Rue (2010) find that an approximate solution to a particular stochastic partial differential equation (SPDE) can explicitly link GMRFs and GGMs for a restricted class of GGMs (a subset of the Matérn class). These papers will be discussed in more detail in chapter 6 and 7.

GMRFs can be used to model stationary or non-stationary processes. Stationary Markov Random Fields are Markov Random Fields which have a constant mean (*i.e.* one that does not depend on the location of the observation) and in which the covariance between any two nodes is a stationary covariance function, *i.e.* one in which the covariance function depends only on the vector distance between the nodes.[1] When the distribution is assumed to be Gaussian, as in a GMRF, with constant mean, we automatically have stationarity if the GMRF has full conditionals. In practice, stationarity means that the relationship between two nodes in the graph is a function solely of their position relative to one another, no matter where they are situated. If the covariance function is a function only of the Euclidean distance between the two nodes (*i.e.* no direction is involved), then the covariance function is *isotropic*. Stationarity has proved a problematic assumption in quite a few historical papers. Among them, Patankar suggested it was a problem in fitting the Mercer and Hall data (1954). It has also been noted that some plant data of Freeman (1953) exhibited different relations between the upper and lower halves of the lattice since one half was diseased (Bartlett, 1974). A stationary model would be a poor fit here. While much of this thesis will focus on stationary GMRFs, it is also possible to conceptualize non-stationary GMRFs. The chapter on intrinsic GMRFs will prove useful here, since these are often used as non-stationary priors in hierarchical Bayesian models.

With these aspects of GMRFs in mind, this thesis will be structured in the following way: first, we will discuss geostatistical models and introduce GMRFs formally, going through the necessary definitions to understand them. Second, we will solidify our understanding of GMRFs as compared to time-series models, focusing on CAR and SAR models. Third, we will discuss improper GMRFs, which are important in forming priors, which is obviously a large part of estimating GMRFs by Bayesian methods. Fourth, we will discuss the computational benefits of modelling GMRFs over other models, notably geostatistical models. Finally, over two chapters, we will consider to what extent GMRFs can approximate Gaussian geostatistical models, reviewing both traditional as well as newer methods. Overall, this thesis aims to explain the concept of GMRFs and their estimation as well as when they

[1] This is often called *second-order* or *weak stationarity*.

are practically of use, particularly in relation to geostatistical models.

Chapter 2

Geostatistical Models and Gaussian Markov Random Fields

In this chapter I introduce geostatistical models and GMRFs and discuss some of their important properties.

2.0.1 Geostatistical Models

Geostatistical models are typically based on *variograms* or *covariance functions*, functions that describe the spatial correlation between different points. Formally, given a stochastic process $Z(s)$ a variogram, $2\gamma(x, y)$, is the expected squared difference in values between two locations x and y: $2\gamma(x, y) = E(|Z(x) - Z(y)|^2)$. $\gamma(x, y)$ is the *semivariogram.* Assumptions of various forms of stationarity are often made, but can be weakened. In particular, to use the semivariogram one makes the assumption that the spatial process is *second-order stationary in first differences* (*i.e.* the process has a constant mean and the variance of the first differences, $Z(s) - Z(s + h)$, does not depend on the location of the observation). This form of stationarity is also called *intrinsic stationarity of the spatial process* and it is weaker than the assumption of *second-order stationarity of the spatial process*, which requires the mean to be constant over all locations and the covariance to depend only on the separation between points rather than their locations. Other important terms are the *nugget* (C_0), represented by the y-intercept depicted below, the *sill* ($C(0)$), the model asymptote, and the *range* (a), the distance between the y-axis and the value at which the sill level is reached (or, for asymptotic sills, the value at which the distance to the sill is negligible, conventionally defined as where the semivariance reaches 95% of the sill).

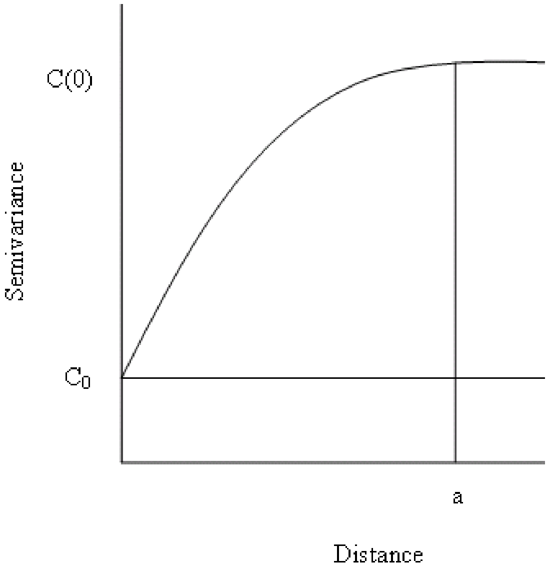

Figure 2.1: Nugget, Sill, and Range

The spatial correlation structures abled to be modelled through these kinds of models are clearly vast. The equations specifying some common semivariogram models are included in the table below.

Model	Semivariogram
Linear	$\gamma(h) = \begin{cases} 0 & h = 0 \\ C_0 + C(0)\lvert h\rvert & h \neq 0 \end{cases}$
Exponential	$\gamma(h) = \begin{cases} 0 & h = 0 \\ C_0 + C(0)\left(1 - exp\left(\frac{-\lvert h\rvert}{a}\right)\right) & h \neq 0 \end{cases}$
Spherical	$\gamma(h) = \begin{cases} 0 & h = 0 \\ C_0 + C(0)\left(1.5\frac{\lvert h\rvert}{a} - 0.5\left(\frac{\lvert h\rvert}{a}\right)^3\right) & 0 < \lvert h\rvert \leq a \\ C_0 + C(0) & \lvert h\rvert \geq a \end{cases}$
Gaussian	$\gamma(h) = \begin{cases} 0 & h = 0 \\ C_0 + C(0)\left(1 - exp\left(\frac{-\lvert h\rvert^2}{a^2}\right)\right) & h \neq 0 \end{cases}$
Power	$\gamma(h) = \begin{cases} 0 & h = 0 \\ C_0 + C(0)\lvert h\rvert^\lambda & h \neq 0 \end{cases}$
Matérn	$\gamma(h) = \begin{cases} 0 & h = 0 \\ C_0 + C(0)\left(1 - \frac{(2\sqrt{\nu}\frac{\lvert h\rvert}{a})^2}{2^{\nu-1}\Gamma(\nu)} K_\nu(2\sqrt{\nu}\frac{\lvert h\rvert}{a})\right) & h \neq 0 \end{cases}$

In this chart, K_ν is a Bessel function with order $\nu > 0$ determining the smoothness of the function and $\Gamma(\nu)$ is the Gamma function. When $\nu = \frac{1}{2}$, the Matérn covariance function reduces to an exponential covariance function and as $\nu \to \infty$ it approaches a Gaussian model.

These models are all isotropic, *i.e.* the same relationships are assumed to hold in all directions. In many cases, we might not think it would hold. For example, we might expect spatial correlation in one direction but not another, such as how we might expect temperature to vary more along a North-South axis than East-West. One alternative to isotropy is geometric anisotropy, in which the anisotropy can be transformed into isotropy by a linear transformation (Cressie, 1993).

The models that we will be discussing are also going to be based on Gaussian processes. There are cases in which we may think we are facing a non-Gaussian process. Diggle, Tawn and Moyeed give the examples of radionuclide concentrations on Rongelap Island and campylobacter infections in north Lancashire and south Cumbria as cases where we might think that if Y is our dependent variable and $S(x)$ is our spatial process, $Y_i|S(x_i)$ probably does not follow a Gaussian distribution (1998).

It was previously mentioned that the Matérn class of models specifies a "smoothness" parameter. In fact, the other classes of models have different degrees of intrinsic smoothness, as well. Banerjee and Gelfand (2003) discuss two types of continuity that may characterize a process: mean square continuity and a.s. continuity. Mean square continuity is another term for continuity in the L_2 sense, whereby a process $X(t) \in \mathbb{R}^d$ is L_2 continuous at t_0 if $lim_{t \to t_0} E[X(t) - X(t_0)]^2 = 0$. If the covariance function of a process in $\mathbb{R}^d$ is d-times continuously differentiable, then the process is mean square continuous. A process is a.s. continuous at t_0 if instead $X(t) \to X(t_0)$ a.s. as $t \to t_0$. Either type of continuity can be used to characterize smoothness. Smoothness captures the idea is that some processes are relatively continuous across different points (*e.g.* elevation levels of rolling hills) and others are relatively discontinuous (*e.g.* elevation levels of regions with cliffs).

The following figures use data simulated under some of these models to illustrate the concept. The variance of each is 1 and the range, a, is chosen for each model so that the correlation at distance 0.5 has decayed to 0.05. This is so that the effect of model smoothness may be seen. For the Matérn models, a few different ν are chosen: $\nu = 1, \nu = 2$ and $\nu = 3$; these show how the Matérn model gets closer to the Gaussian model as ν increases, and it should also be recalled that the exponential model is the same as a Matérn model with $\nu = \frac{1}{2}$.

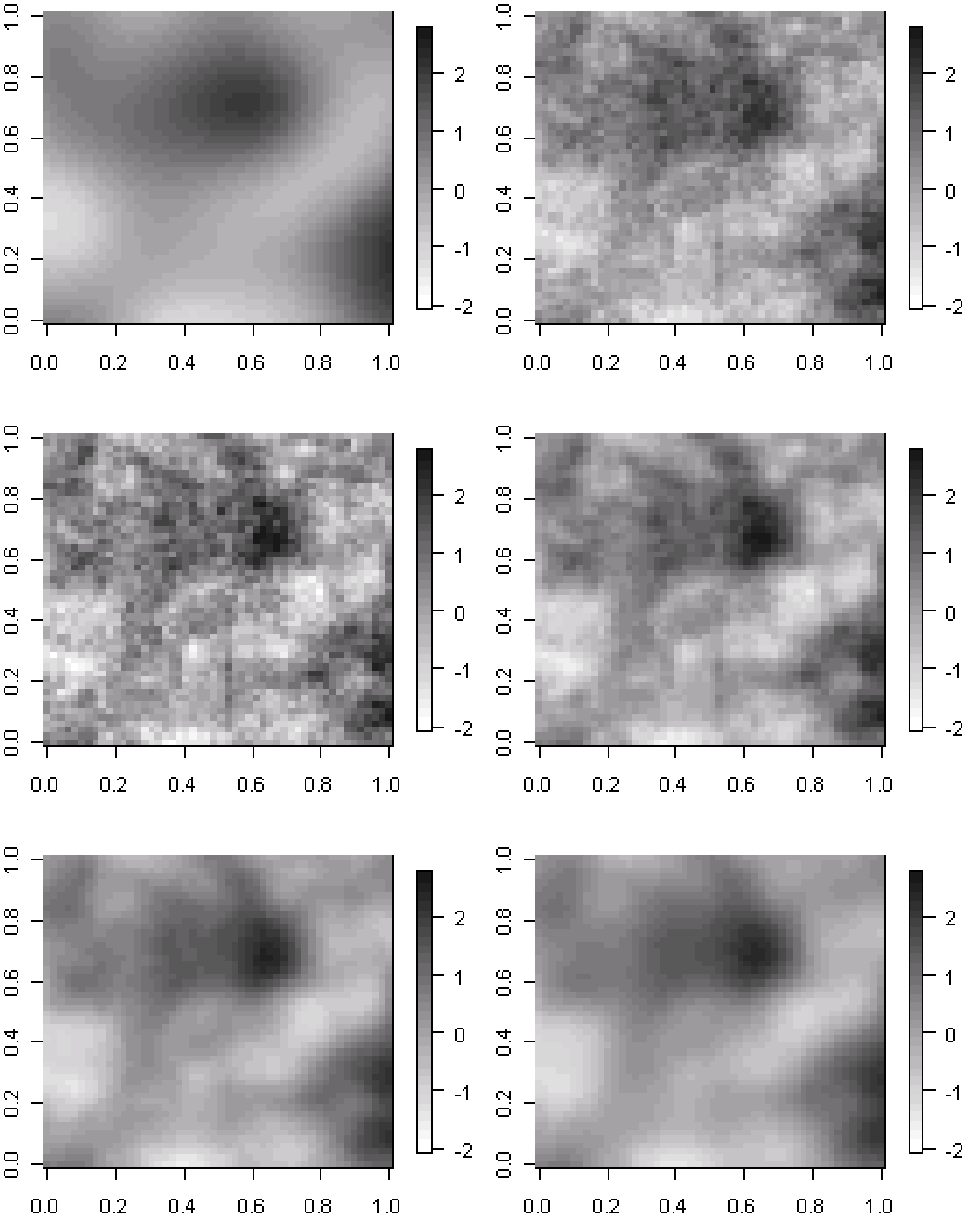

Figure 2.2: From left to right, top down: Simulations of Gaussian, Spherical, Exponential, and Matérn Models with $\nu = 1, 2, 3$

Under stationarity, the semivariogram is related to the covariance function in the following way:

$$\begin{aligned} 2\gamma(h) &= Var(Z(s+h) - Z(s)) \\ &= Var(Z(s+h)) + Var(Z(s)) - 2Cov(Z(s+h), Z(s)) \\ &= C(0) + C(0) - 2C(h) \\ &= 2(C(0) - C(h)) \\ \implies \gamma(h) &= C(0) - C(h) \end{aligned}$$

where $C(0)$ is the variance of the process, also denoted σ^2. Covariance functions are frequently used but, again, require stronger assumptions than do semivariogram models. The geostatistical literature is vast and even the requirements of intrinsic stationarity can be relaxed (Cressie, 1993). For our purposes, the key take-away is that these models are highly flexible. Still, they can be computationally intensive, a fact we will revisit in a later chapter.

2.1 Gaussian Markov Random Fields

I will now turn to discuss GMRFs. Before beginning, I will need to define some of the terms that I will use in the rest of this thesis. This introduction follows Rue and Held (2005).

2.1.1 Undirected Graphs

An *undirected graph* G is a tuple $G = (V, E)$ where V is the set of nodes in the graph and E is the set of edges $\{i, j\}$ where $i, j \in V$ and $i \neq j$. If $\{i, j\} \in E$ there is an undirected edge between node i and node j; otherwise, there is no such edge. A graph is *fully connected* if $\{i, j\} \in E \; \forall i, j \in V$ with $i \neq j$. We can label our graphs according to their nodes,

$V = \{1, 2, ..., n\}$, and these are called *labelled* graphs. The *neighbours* of node i are all the nodes in G that have an edge to node i, or

$$n(i) = \{j \in V : \{i, j\} \in E\}$$

If i and j are neighbours in a graph, this will be written as $i \sim j$. We can similarly say that the neighbours of a set $A \subset V$ are all nodes not in A but adjacent to a node in A, or

$$n(A) = \cup_{i \in A} n(i) \setminus A$$

A *path* from i_1 to i_m is a sequence of distinct nodes in V, $i_1, i_2, ..., i_m$, for which $(i_j, i_{j+1}) \in E$ for $j = 1, ..., m-1$. A subset $C \subset V$ *separates* two nodes $i \notin C, j \notin C$ if every path from i to j contains at least one node from C. Two disjoint sets $A \subset V \setminus C$ and $B \subset V \setminus C$ are separated by C if all $i \in A$ and $j \in B$ are separated by C. In other words, if we were to walk through the graph, we cannot start at a node in A and end somewhere in B without passing through C.

G^A denotes a subgraph of G defined by $\{V^A, E^A\}$, where $V^A = A$, a subset of V, and $E^A = \{\{i, j\} \in E \text{ and } \{i, j\} \in A \times A\}$. In other words, G^A is the graph that results if we start with G and remove all nodes not in A and all edges connected to at least one node which does not belong to A.

To explain some more notation used later, if we have a lattice I with sites ij where $i = 1, ..., n_1$ and $j = 1, ..., n_2$, if $C \in I$ then $x_C = \{x_i : i \in C\}$. Similarly, for $-C$, $x_{-C} = \{x_i : i \in -C\}$.

2.1.2 Other Useful Definitions

An $n \times n$ matrix A is *positive definite* iff

$$x^T Ax > 0 \; \forall x \neq 0$$

and *symmetric positive definite* (SPD) if in addition to being positive definite A is symmetric.

Suppose we have a random vector $x = (x_1, ..., x_n)^T$ with a normal distribution with mean μ and covariance Σ. As another piece of terminology, the inverse covariance matrix Σ^{-1} is also the *precision matrix* which we will denote Q.

2.1.3 Gaussian Markov Random Fields

Now we can define a GMRF. A random vector $x = (x_1, ..., x_n)^T \in R^n$ is a *GMRF* with respect to a labelled graph $G = (V, E)$ with mean μ and precision matrix $Q > 0$ iff its density has the form

$$\pi(x) = (2\pi)^{-n/2}|Q|^{1/2} exp\left(-\frac{1}{2}(x-\mu)^T Q(x-\mu)\right) \tag{2.1.1}$$

and

$$Q_{ij} \neq 0 \iff \{i, j\} \in E \; \forall i \neq j$$

This means that in the labelled graph $G = (V, E)$, where $V = \{1, ..., n\}$, E has no edge between node i and node j iff $x_i \perp x_j | x_{-\{i,j\}}$.[1] If Q is completely dense, then G is fully connected. We can also note that a GMRF is a normal distribution with a SPD covariance matrix and any normal distribution with a SPD covariance matrix is a GMRF.

In what sense is a GMRF "Markov"? There are three Markov properties in which we may be interested.

[1] Recalling the notation explained in 2.1.1, this is intuitive: conditioning on the whole graph aside from x_i and x_j, x_i and x_j should be conditionally independent if and only if there is no edge between them.

1. The *pairwise Markov property*:

$$x_i \perp x_j | x_{-\{i,j\}} \text{ if } \{i,j\} \notin E \text{ and } i \neq j$$

2. The *local Markov property*:

$$x_i \perp x_{-\{i,n(i)\}} | x_{n(i)} \; \forall i \in V$$

3. The *global Markov property*:

$$x_A \perp x_B | x_C$$

   for all disjoint sets A, B, C where C separates A and B and A and B are non-empty (if C is empty, x_A and x_B are independent).

The figure below illustrates these three Markov properties. For a GMRF, the three properties are equivalent, as proved in Speed and Kiiveri (1986).

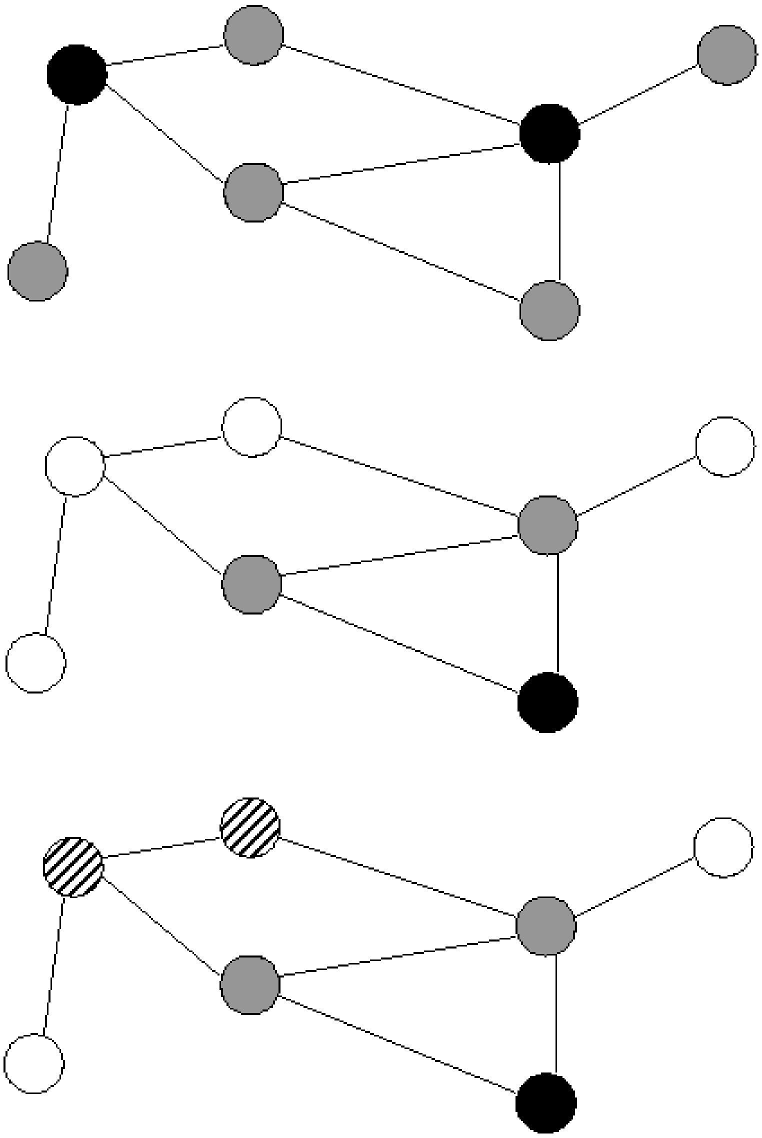

Figure 2.3: Top: The pairwise Markov property - the black nodes are conditionally independent given the grey nodes. Middle: The local Markov property - the black node is conditionally independent of the white nodes given the grey nodes. Bottom: The global Markov property - the black node is conditionally independent from the striped nodes given the grey nodes.

We will also want to consider conditional properties of GMRFs. In particular, we can split x into sets A and B (where $V = A \cup B$, $A \cap B = \emptyset$), so that

$$x = \begin{pmatrix} x_A \\ x_B \end{pmatrix}.$$

The mean and precision matrix can also be partitioned similarly:

$$\mu = \begin{pmatrix} \mu_A \\ \mu_B \end{pmatrix}$$

$$Q = \begin{pmatrix} Q_{AA} & Q_{AB} \\ Q_{BA} & Q_{BB} \end{pmatrix}.$$

With this in mind, I present the following two useful theorems from Rue and Held (2005):

Theorem 1. *Let x be a GMRF wrt $G = (V, E)$ with mean μ and precision matrix $Q > 0$. Let $A \subset V$ and $B = V \setminus A$, where $A, B \neq 0$. The conditional distribution of $x_A | x_B$ is then a GMRF wrt the subgraph G^A with mean $\mu_{A|B}$ and precision matrix $Q_{A|B} > 0$, where*

$$\mu_{A|B} = \mu_A - Q_{AA}^{-1} Q_{AB}(x_B - \mu_B) \tag{2.1.2}$$

and

$$Q_{A|B} = Q_{AA}$$

Proof. Assume $\mu = 0$ for now; we will add it back in later. Then

$$\begin{aligned} \pi(x_A|x_B) &\propto exp\left(-\frac{1}{2}(x_A, x_B)\begin{pmatrix} Q_{AA} & Q_{AB} \\ Q_{BA} & Q_{BB} \end{pmatrix}\begin{pmatrix} x_A \\ x_B \end{pmatrix}\right) \\ &\propto exp\left(-\frac{1}{2}x_A^T Q_{AA} x_A - (Q_{AB}x_B)^T x_A\right). \end{aligned}$$

The density of a normal distribution with precision P and mean γ is

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

so we see that Q_{AA} is the conditional precision matrix and the conditional mean solves $Q_{AA}\mu_{A|B} = -Q_{AB}x_B$. $Q_{AA} > 0$ since $Q > 0$. Finally, if x has mean μ then $x - \mu$ has mean 0, so replacing x by $x - \mu$ where μ is no longer assumed to be zero, we get (2.3.2). ■

Theorem 2. *Let x be normally distributed with mean μ and precision matrix $Q > 0$. Then for $i \neq j$,*

$$x_i \perp x_j | x_{-ij} \iff Q_{ij} = 0$$

where x_{-ij} stands for $x_{-\{i,j\}}$.

Proof. $x \perp y|z \iff \pi(x, y, z) = f(x, z)g(y, z)$ for some functions f and g and for all z with $\pi(z) > 0$. Assuming $\mu = 0$ and fixing $i \neq j$, WLOG, and using the definition of a GMRF, we get

$$\begin{aligned} \pi(x_i, x_j, x_{-ij}) &\propto exp\left(-\frac{1}{2}\sum_{k,l} x_k Q_{kl} x_l\right) \\ &\propto exp\left(-\frac{1}{2}x_i x_j (Q_{ij} + Q_{ji}) - \frac{1}{2}\sum_{\{k,l\}\neq\{i,j\}} x_k Q_{kl} x_l\right) \end{aligned}$$

The first term involves $x_i x_j$ iff $Q_{ij} \neq 0$ and the second term does not involve $x_i x_j$. Thus $\pi(x_i, x_j, x_{-ij}) = f(x_i, x_{-ij}) g(x_j, x_{-ij})$ for some functions f and g iff $Q_{ij} = 0$. $\blacksquare$

Finally, we may want to use the canonical parameterization for a GMRF. The *canonical parameterization* of a GMRF x wrt G is $x \sim N_C(b, Q)$ where the precision matrix is Q and the mean is $\mu = Q^{-1}b$. Note that then

$$x_A | x_B \sim N_C(b_A - Q_{AB} x_B, Q_{AA}).$$

Further, if $y|x \sim N(x, P^{-1})$ then

$$x|y \sim N_C(b + Py, Q + P).$$

These last results are useful in computing conditional densities.

Now that we have formally defined GMRFs and described their most salient properties, it might be instructive to re-visit why we might think that making conditional independence assumptions would be appropriate when dealing with spatial correlation. The intuition is that after a certain threshold distance, the spatial correlation is assumed to be negligible. Thus, conditioning on enough observations, one can get conditional independence. Apart from the previously mentioned constraints on when GMRFs are useful, GMRFs would thus require one to make more assumptions if the number of observations was small relative to the number of observations that had to be conditioned on, *i.e.* if few observations were conditionally independent.

Chapter 3

Relation to Time-Series Analogues: CAR and SAR Models

Having introduced GMRFs, it would be natural to ask how they correspond to perhaps more familiar time-series models, since time-series models are motivated by temporal correlation in much the same way as spatial models are motivated by spatial correlation. The treatment in the next section is adapted from Rue and Held (2005) and Cressie (1993).

3.1 CAR and SAR Models

In time series, we may see an AR(1) process represented thusly:

$$x_t = \phi x_{t-1} + \epsilon_t,\ \epsilon_t \sim N(0,1),\ |\phi| < 1$$

where t indexes time.

This simple process assumes conditional independence, like a GMRF. x_t is independent of x_s (where $1 \leq s < t \leq n$) conditional on $\{x_{s+1}, ..., x_{t-1}\}$, or $x_t | x_1, ..., x_{t-1} \sim N(\phi x_{t-1}, 1)$ for $t = 2, ..., n$. This is a *directed* conditional distribution and straightforwardly gives us the precision matrix

$$Q = \begin{pmatrix} 1 & -\phi & & & \\ -\phi & 1+\phi^2 & -\phi & & \\ & \ddots & \ddots & \ddots & \\ & & -\phi & 1+\phi^2 & -\phi \\ & & & -\phi & 1 \end{pmatrix}$$

We could also write the *undirected* conditional distribution, or *full conditionals* $\{\pi(x_t|x_{-t})\}$:

$$x_t|x_{-t} \sim \begin{cases} N(\phi x_{t+1}, 1) & t = 1 \\ N\left(\frac{\phi}{1+\phi^2}(x_{t-1} + x_{t+1}), \frac{1}{1+\phi^2}\right) & 1 < t < n \\ N(\phi x_{n-1}, 1) & t = n \end{cases}$$

When we move away from time series, where the observations have a natural ordering, into a domain where they do not (for example, with spatial data), this second specification is more useful. In particular, specifying each of the n full conditionals

$$x_i|x_{-i} \sim N\left(\Sigma_{j:j\neq i}\beta_{ij}x_j, \kappa_i^{-1}\right)$$

will specify the joint density of a zero mean GMRF. These models, popularized by Besag (1974, 1975), are called *conditional autoregression models* or CAR models.

The n full conditionals must satisfy some requirements to ensure that a joint normal density exists with these full conditionals. In particular, suppose we specify the full conditionals as normals with

$$E(x_i|x_{-i}) = \mu_i - \Sigma_{j:j\sim i}\beta_{ij}(x_j - \mu_j) \qquad (3.1.1)$$

$$Prec(x_i|x_{-i}) = \kappa_i > 0 \qquad (3.1.2)$$

for $i = 1, ..., n$, some $\{\beta_{ij}, i \neq j\}$ and vectors μ and κ. Since $\sim$ is symmetric, we need the requirement that if $\beta_{ij} \neq 0, \beta_{ji} \neq 0$. Recalling that a univariate normal random variable x_i with mean γ and precision κ has density proportional to

$$exp(-\frac{1}{2}\kappa x_i^2 + \kappa x_i \gamma), \tag{3.1.3}$$

and x is a GMRF wrt G only if its density has the form

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

and

$$\pi(x_A|x_{-A}) = \frac{\pi(x_A, x_{-A})}{\pi(x_{-A})} \propto \pi(x),$$

we get

$$\begin{aligned} \pi(x_i|x_{-i}) &\propto exp(-\frac{1}{2}x^T Qx) & (3.1.4) \\ &\propto exp(-\frac{1}{2}x_i^2 Q_{ii} - x_i \Sigma_{j:j\sim i} Q_{ij} x_j). & (3.1.5) \end{aligned}$$

Comparing (2.4.3) and (2.4.5), we see that

$$Prec(x_i|x_{-i}) = Q_{ii}$$

and

$$E(x_i|x_{-i}) = -\frac{1}{Q_{ii}} \Sigma_{j:j\sim i} Q_{ij} x_j$$

and since x has mean μ, if we replace x_i and x_j by $x_i - \mu_i$ and $x_j - \mu_j$ then we get

$$E(x_i|x_{-i}) = \mu_i - \frac{1}{Q_{ii}} \Sigma_{j:j\sim i} Q_{ij}(x_j - \mu_j).$$

Comparing these to (2.4.1) and (2.4.2), we see that choosing $Q_{ii} = \kappa_i$ and $Q_{ij} = \kappa_i\beta_{ij}$ and requiring Q to be symmetric and $Q > 0$ would give us a candidate for a joint density giving the specified full conditionals. Rue and Held (2005) show that this candidate is also unique.

Simultaneous autoregression models, or SAR models, are closely related to CAR models. To return to our comparisons with autoregressive time-series models, whereas CAR models are more analogous to time-series models in their Markov properties, SAR models are more analogous in their functional forms (Cressie, 1993). We can motivate SARs similar to how we motivated CARs, by assuming some spatial correlation so that there is some spatially lagged version of the variable of interest on the RHS:

$$x = \rho W x + \epsilon$$

where ρ is a spatial autoregression parameter that has to be estimated from the data and W is a spatial weighting matrix so that neighbouring values are included with some weights into the regression. W here is not necessarily symmetric, unlike in the CAR model. Re-arranging, we get

$$x = (I - \rho W)^{-1} + \epsilon$$

and then

$$Var(x) = (I - \rho W)^{-1} E(\epsilon\epsilon^T)(I - \rho W^T)^{-1}$$

or, if we assume normality, collapse ρ and W into a single parameter C, and let Λ equal the covariance matrix $E(\epsilon\epsilon^T)$,

$$x \sim N(\mu, (I - C)^{-1}\Lambda(I - C^T)^{-1}).$$

Looking back at (2.4.1), we see that in the CAR model the covariance matrix is given by

$$Var(x) = (I - B)^{-1} M$$

where M is an $n \times n$ diagonal matrix that must be estimated. This is quite similar to what we found for the SAR model. CAR and SAR models are equivalent if and only if their variance matrices are equal:

$$(I - C)^{-1}\Lambda(I - C^T)^{-1} = (I - B)^{-1}M.$$

Since M is diagonal, any SAR can be represented as a CAR but not necessarily the other way around.

3.2 Differences Between CAR and SAR Models

The main practical difference between CAR and SAR models is that the SAR model does not impose the restriction that W is symmetric and may be harder to estimate as a consequence. Indeed, if ρ is not known, the SAR model is inconsistently estimated by least squares (Whittle, 1954). In contrast, generalized least squares can be used to estimate the parameters of a simple CAR model and then the residuals can be used to estimate ρ_C, a CAR counterpart to ρ, using ordinary least squares (Haining, 1990). Since SAR models have more parameters, to estimate them in practice one typically assumes that some of the parameters are known (Waller and Gotway, 2004). SAR models are also not conveniently estimated using hierarchical Bayesian methods since it is difficult to include SAR random effects, and CAR is more convenient for computing. However, if computing time is not an issue, SAR models are very convenient for maximum likelihood estimation and their structure makes them frequently and most naturally used in econometric regressions. It has been argued that SAR models are also more appropriate for studies that involve second-order dependency captured by W (Bao, 2004).

These differences aside, in terms of results, most studies that try both types of models find no large differences between the two (*e.g.* Lichstein et al., 2002). Wall (2004) also severely critiques spatial interpretations of both CAR and SAR weighting schemes, finding unintuitive behaviour. For example, there does not seem to be much sense in a spatial model insisting, as hers did, that Missouri and Tennessee were the least spatially correlated states in the U.S., with Missouri being more correlated with Kansas than Iowa. One might think that all these models would be inappropriate over such large geographical areas with so few observations. The debate about whether CAR or SAR models are preferable or even appropriate in different circumstances continues.

Chapter 4

Intrinsic Gaussian Markov Random Fields

In this chapter I explain the details of intrinsic GMRFs and their modelling. *Intrinsic GMRFs (IGMRFs)* are *improper*, meaning that their precision matrices are not of full rank. There are a few definitions of the *order* of an IGMRF, but we will follow Rue and Held (2005) in calling the order of an IGMRF the rank deficiency of its precision matrix (see Künsch (1987) for a slightly different definition).

Intrinsic GMRFs are very important because they are the typical priors used when estimating using Bayesian methods. We will see that they have a natural relationship with different modelling choices, including the thin plate spline. The point of this chapter is to understand the correspondence between physical modelling choices and priors and the matrices used in estimation. We will work up to deriving precision matrices for a few of the more frequently-used modelling choices. The examples presented here are selected from Rue and Held (2005) and fleshed out in greater detail.

To quickly define additional terms that will be necessary, the *null space* of A is the set of all vectors x such that $Ax = 0$. The *nullity* of A is the dimension of the null space. If a singular $n \times n$ matrix A has nullity k, $|A|^*$ will denote the product of the $n - k$ non-zero eigenvalues of A.

The *first-order forward difference* of a function $f(\cdot)$ is

$$\Delta f(z) = f(z+1) - f(z)$$

and more generally *higher order forward differences* are defined recursively so that:

$$\Delta^k f(z) = \Delta \Delta^{k-1} f(z)$$

or

$$\Delta^k f(z) = (-1)^k \sum_{j=0}^{k} (-1)^j \left(\begin{array}{c} k \\ \\ j \end{array} \right) f(z+j)$$

for $k = 1, 2, ...$

For example, the second-order forward difference is:

$$\Delta^2 f(z) = f(z+2) - 2f(z+1) + f(z)$$

Proceeding along, with these preliminaries out of the way, now let Q be an $n \times n$ symmetric positive semidefinite (SPSD) matrix with rank $n-k > 0$. $x = (x_1, ..., x_n)^T$ is an *improper GMRF* of rank $n-k$ with parameters (μ, Q) if its density is

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

x is an *improper GMRF wrt to the graph G* if in addition $Q_{ij} \neq 0 \iff \{i,j\} \in \ \forall i \neq j$.

It should be noted that (μ, Q) do not represent the mean and the precision anymore since these technically no longer exist. (However, we will keep referring to them as the mean and the precision for convenience.) For intuition, we can think of a $Q(\gamma)$:

$$Q(\gamma) = Q + \gamma A^T A$$

where the columns of A^T span the null space of Q. Each element of $Q(\gamma)$ corresponds to the appropriate element in Q as $\gamma \to 0$.

An improper GMRF of rank $n-1$ where $\sum_j Q_{ij} = 0 \; \forall i$ is an *intrinsic GMRF of first order*. We will call the condition $\sum_j Q_{ij} = 0 \; \forall i$ "$Q1 = 0$" for convenience; in words, the vector 1 spans the null space of Q.

Now we will discuss IGMRFs of first order in a few different settings: on a line for regularly spaced locations; on a line for irregularly spaced locations; and on a lattice for regularly and irregularly spaced locations. We will then detail IGMRFs of higher order.

4.1 IGMRFs of First Order on the Line, Regularly Spaced

This is also known as the first-order random walk. The nodes are assumed to be located at a constant distance apart; for convenience, we will label the nodes $i = 1, ..., n$ and think of them as 1 unit of distance apart in that order. i could be thought of as time, as well as distance.

The model assumes *independent increments*

$$\Delta x_t \sim N(0, \kappa^{-1}), \; i = 1, ..., n-1$$

or

$$x_j - x_i \sim N(0, (j-i)\kappa^{-1}), \; i < j$$

If the intersection of $\{i, ..., j\}$ and $\{k, ..., l\}$ is empty for $i < j$ and $k < l$ then

$$Cov(x_j - x_i, x_l - x_k) = 0$$

The density for x is then:

$$\begin{aligned}\pi(x|\kappa) &\propto \kappa^{(n-1)/2} exp\left(-\frac{\kappa}{2}\sum_{i=1}^{n-1}(\Delta x_i)^2\right)\\ &= \kappa^{(n-1)/2} exp\left(-\frac{\kappa}{2}\sum_{i=1}^{n-1}(x_{i+1}-x_i)^2\right)\\ &= \kappa^{(n-1)/2} exp\left(-\frac{1}{2}x^TQx\right)\end{aligned}$$

where $Q = \kappa R$, with R being the *structure matrix*:

$$R = \begin{pmatrix} 1 & -1 & & & & & \\ -1 & 2 & -1 & & & & \\ & -1 & 2 & -1 & & & \\ & & \ddots & \ddots & \ddots & & \\ & & & -1 & 2 & -1 & \\ & & & & -1 & 2 & -1 \\ & & & & & -1 & 1 \end{pmatrix}$$

that follows from $\sum_{i=1}^{n-1}(\Delta x_i)^2 = (Dx)^T(Dx) = x^TD^TDx = x^TRx$ where D is an $(n-1)\times n$ matrix of the form

$$D = \begin{pmatrix} -1 & 1 & & & \\ & -1 & 1 & & \\ & & \ddots & \ddots & \\ & & & -1 & 1 \end{pmatrix}$$

In words, D takes this form because it represents a first-order forward difference.

4.2 IGMRFs of First Order on the Line, Irregularly Spaced

We will call the locations s_i and even though they are not regularly spaced anymore we can still order them $s_1 < s_2 < ... < s_n$ and call the distance $\delta_i = s_{i+1} - s_i$. The first-order random walk over continuous distances (or time) is analogous to a Wiener process in continuous time. Formally, a *Wiener process* with precision κ is a stochastic process $W(t)$ for $t \geq 0$ with $W(0) = 0$ and increments $W(t) - W(s)$ that are normally distributed with mean 0 and variance $(t - s)/\kappa$ for any $0 \leq s < t$ and independent if the time intervals do not overlap. If $\kappa = 1$ this is a *standard Wiener process.*

The model is fundamentally the same as in the regularly spaced, discrete time model. It is still true that $\sum_j Q_{ij} = 0 \; \forall i$ and the joint density of x is

$$\pi(x|\kappa) \propto \kappa^{(n-1)/2} exp\left(-\frac{\kappa}{2}\sum_{i=1}^{n-1}(x_{i+1} - x_i)^2/\delta_i\right)$$

where the only difference is that the exponential term is scaled with δ_i since $Var(x_{i+1} - x_i) = \delta_i/\kappa$.

4.3 IGMRFs of First Order on a Lattice

First we must clarify our notation. A lattice will be denoted I_n with $n = n_1 n_2$ nodes. (i, j) will denote the node in the ith row and jth column. The following discussion will be in reference to irregular lattices but the case of regular lattices is subsumed within it.

If we have two neighbouring regions i and j, we will define a normal increment to be

$$x_i - x_j \sim N(0, \kappa^{-1})$$

Assuming the increments to be independent, we get the familiar IGMRF density:

$$\pi(x) \propto \kappa^{(n-1)/2} exp\left(-\frac{\kappa}{2}\sum_{i\sim j}(x_i - x_j)^2\right)$$

where $i \sim j$ denotes the set of all unordered pairs of neighbours (unordered so as to prevent us from double-counting).

It should be noted that the increments are not truly independent, since the geometry of the lattice imposes hidden linear constraints. We can see this by noting that the number of increments $|i \sim j|$ is larger than n but the rank of the corresponding precision matrix remains $n - 1$. As a concrete example, we can consider that if we have 3 nodes, all of which are neighbours of each other, we have $x_1 - x_2 = \epsilon_1$, $x_2 - x_3 = \epsilon_2$, $x_3 - x_1 = \epsilon_3$ and $\epsilon_1 + \epsilon_2 + \epsilon_3 = 0$, a hidden constraint.

Despite the constraint, however, the density of x is still represented correctly above. Define ϵ to be $|i \sim j|$ independent increments and $A\epsilon$ to be the constraints that say that the increments sum to zero over all the graph's circuits. Then $\pi(\epsilon|A\epsilon) \propto \pi(\epsilon)$. Changing variables from ϵ to x gives the exponential term and the constant is unaffected by the constraint.

4.4 IGMRFs of Higher Orders

First, I will introduce some notation which should make the subsequent discussion easier to read. Since IGMRFs are often invariant to the addition of a polynomial of a certain degree, let us define notation for these polynomials.

For the case of an IGMRF on a line, let $s_1 < s_2 < ... < s_n$ denote the ordered locations on the line and s denote the vector $(s_1, ..., s_n)^T$. Now we will let $p_k(s_i)$ denote a polynomial of degree k evaluated at the locations s with some coefficients $\beta_k = (\beta_0, ..., \beta_k)^T$:

$$p_k(s_i) = \beta_0 + \beta_1 s_1 + \frac{1}{2}\beta_2 s_i^2 + ... + \frac{1}{k!}\beta_k s_i^k$$

or $p_k = S_k \beta_k$. S_k is the *polynomial design matrix of degree k*.

For higher dimensions, we instead let $s_i = (s_{i_1}, s_{i_2}, ..., s_{i_d})$ where s_{i_j} is the location of node i in the jth dimension, and $j = (j_1, j_2, ..., j_d)$, $s_i^j = s_{i_1}^{j_1} s_{i_2}^{j_2} ... s_{i_d}^{j_d}$, and $j! = j_1! j_2! ... j_d!$. A *polynomial trend* of degree k in d dimensions is:

$$p_{k,d}(s_i) = \sum_{0 \leq j_1 + ... + j_d \leq k} \frac{1}{j!} \beta_j s_i^j$$

or $p_{k,d} = S_{k,d} \beta_{k,d}$. It will have

$$m_{k,d} \equiv \begin{pmatrix} d+k \\ k \end{pmatrix}$$

terms.

Higher-order IGMRFs have a rank deficiency larger than one. Formally, an *IGMRF of order k on the line* is an improper GMRF of rank $n-k$ where $-\frac{1}{2}(x + p_{k-1})^T Q (x + p_{k-1}) = -\frac{1}{2} x^T Q x$ so that the density seen when first introducing IGMRFs,

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

is invariant to the addition of any polynomial of degree $k-1$,p_{k-1} to x. (We can simplify notation for this last condition and write it as $QS_{k-1} = 0$.)

Another way to state this is to say that we can decompose x into a trend and a residual component, as follows:

$$x = H_{k-1}x + (I - H_{k-1})x$$

where the projection matrix H_{k-1} projects x down to the space spanned by the column space of S_{k-1} and $I - H_{k-1}$ annihilates anything in the column space of S_{k-1}, projecting onto the space that is orthogonal to that spanned by S_{k-1}. Formally, $H_{k-1} = S_{k-1}(S_{k-1}^T S_{k-1})^{-1} S_{k-1}^T$ (*i.e.* it is just a standard projection matrix). Using projection notation, we require $-\frac{1}{2}x^T Qx = -\frac{1}{2}((I - H_{k-1})x)^T Q((I - H_{k-1})x)$ (since $QH_{k-1} = 0$ and so that term disappears). This is more readily interpretable: the density of a kth order IGMRF only depends on the residuals that remain after removing any polynomial trend of degree $k-1$.

Similarly, an *IGMRF of order k in dimension d* is an improper GMRF of rank $n - m_{k-1,d}$ where $QS_{k-1,d} = 0$.

As an example, let us consider the regularly spaced second-order random walk, first on a line and then on a lattice.

4.5 IGMRFs of Second Order on the Line, Regularly Spaced

This case is analogous to the case of IGMRFs of first order on the line, except with second-order increments $\Delta^2 x_i \sim N(0, \kappa^{-1})$. Then joint density of x is then:

$$\begin{aligned}\pi(x) &\propto \kappa^{(n-2)/2} exp\left(-\frac{\kappa}{2}\sum_{i=1}^{n-2}(x_i - 2x_{i+1} + x_{i+2})^2\right) \\ &= \kappa^{(n-2)/2} exp\left(-\frac{1}{2}x^T Qx\right)\end{aligned}$$

where the precision matrix is:

$$Q = \kappa \begin{pmatrix} 1 & -2 & 1 & & & & & & \\ -2 & 5 & -4 & 1 & & & & & \\ 1 & -4 & 6 & -4 & 1 & & & & \\ & 1 & -4 & 6 & -4 & 1 & & & \\ & & \ddots & \ddots & \ddots & \ddots & \ddots & & \\ & & & 1 & -4 & 6 & -4 & 1 & \\ & & & & 1 & -4 & 6 & -4 & 1 \\ & & & & & 1 & -4 & 5 & -2 \\ & & & & & & 1 & -2 & 1 \end{pmatrix}$$

Again, these numbers come from the fact that we can write $x_i - 2x_{i+1} + x_{i+2}$ in matrix form as

$$D \equiv \begin{pmatrix} 1 & -2 & 1 & & & \\ & 1 & -2 & 1 & & \\ & & \ddots & \ddots & \ddots & \\ & & & 1 & -2 & 1 \end{pmatrix}$$

and $(x_i - 2x_{i+1} + x_{i+2})^2 = D^T D$.

4.6 IGMRFs of Second Order on a Lattice

In this section I will discuss the example of IGMRFs of second order on a regular lattice with two dimensions. This case can be easily extended to more dimensions.

For the case of a second-order IGMRF on a regular lattice with two dimensions, let us for a simple example choose the increments

$$(x_{i+1,j} + x_{i-1,j} + x_{i,j+1} + x_{i,j-1}) - 4x_{i,j} \tag{4.6.1}$$

This is just $\left(\Delta^2_{(1,0)} + \Delta^2_{(0,1)}\right) x_{i-1,j-1}$ where Δ is generalized to account for direction so that $\Delta_{(1,0)}$ is the forward difference in direction $(1,0)$ and $\Delta_{(0,1)}$ is the forward difference in direction $(0,1)$. Analogously to the polynomials added to x in previous examples, adding the simple plane $p_{1,2}(i,j) = \beta_{00} + \beta_{10} i + \beta_{01} j$ to x will cancel in 4.7.1 for any coefficients $\beta_{00}, \beta_{10}, \beta_{01}$.

As we would expect from the previous examples, the precision matrix should correspond to:

$$-\left(\Delta^2_{(1,0)} + \Delta^2_{(0,1)}\right)^2 = -\left(\Delta^4_{(1,0)} + 2\Delta^2_{(1,0)}\Delta^2_{(0,1)} + \Delta^4_{(0,1)}\right).$$

This is actually an approximation to the *biharmonic* differential operator

$$\left(\frac{\delta^2}{\delta x^2} + \frac{\delta^2}{\delta y^2}\right)^2 = \frac{\delta^4}{\delta x^4} + 2\frac{\delta^4}{\delta x^2}\delta y^2 + \frac{\delta^4}{\delta y^4}.$$

It also therefore relates to the *thin plate spline*, a two-dimensional analogue of the cubic spline in one dimension, which is the fundamental solution of the biharmonic equation

$$\left(\frac{\delta^4}{\delta x^4} + 2\frac{\delta^4}{\delta x^2}\delta y^2 + \frac{\delta^4}{\delta y^4}\right)\phi(x,y) = 0$$

Finally, it should be noted that the choice of increments in (4.6.1) represents, visually:

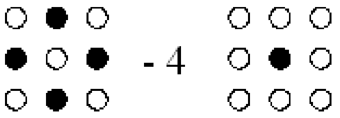

Figure 4.1: Depiction of (4.6.1).

where the black dots represent the locations in the lattice which are taken into consideration and the white dots merely make the spatial configuration clear.

This may not be the ideal choice of increments, depending on our data and our aim;

alternative choices are possible, representing different models. Our having gone through examples of IGMRFs of first and second order on the line and on a lattice, however, cover a lot of the cases one might be interested in, with many results able to be straightforwardly extended.

Chapter 5

Computational Advantages and Applications to GGMs

In this chapter we discuss the reasons why GMRFs are so computationally efficient. GMRFs' computational benefits, particularly in comparison to estimating geostatistical models, are a large source of their popularity, and this will motivate us to survey attempts to fit GMRFs to Gaussian geostatistical models, or *GGMs*, in subsequent chapters.

5.1 Taking Advantage of Sparse Matrices

The main tasks of the algorithms involved in simulating and estimating the parameters of GMRFs via likelihood-based methods are computing the Cholesky factorization of $Q = LL^T$ and solving systems of the form $Lv = b$ and $L^T = x$. But Q is sparse, and this can help us speed up the process. This section will explain why a sparse Q can help and how we can take advantage of its sparseness.

When factorizing a matrix $Q = LL^T$, what we are really doing is computing each of

$$Q_{ij} = \sum_{k=1}^{j} L_{ik} L_{jk} \; i \geq j$$

If we define $v_i = Q_{ij} - \sum_{k=1}^{j-1} L_{ik}L_{jk}$ $i \geq j$ then $L_{jj}^2 = v_j$ and $L_{ij}L_{jj} = v_i$ for $i > j$. Then if we know $\{v_i\}$ for fixed j we get the jth column in L because $L_{jj} = \sqrt{v_j}$ and $L_{ij} = \frac{v_i}{\sqrt{v_j}}$ for $i = j+1$ to n. But $\{v_i\}$ for fixed j only depends on the elements of L in the first $j-1$ columns of L, so we can go through and find all the columns in L.

Taking a step back, solving something like $Lv = b$ for v where L is lower triangular, we solve by *forward substitution*, in a forward loop:

$$v_i = \frac{1}{L_{ii}}(b_i - \sum_{j=1}^{i-1} L_{ij}v_j),\ i = 1, ..., n$$

Solving $L^T x = v$ for x where L^T is upper triangular involves *back substitution*, since the solution requires a backward loop:

$$x_i = \frac{1}{L_{ii}}(v_i - \sum_{j=1+1}^{n} L_{ji}x_j),\ i = n, ..., 1$$

Now consider the most basic case of sampling $x \sim N(\mu, Q^{-1})$ where x is a GMRF wrt to graph G with mean μ and precision matrix $Q > 0$. We would:

1. Compute the Cholesky factorization, $Q = LL^T$.

2. Sample $z \sim N(0, I)$.

3. Solve $L^T v = z$.

4. Compute $x = \mu + v$.

The theorems below from Rue and Held (2005) follow:

Theorem 3. *Let x be a GMRF wrt the graph G with mean μ and precision matrix $Q > 0$. Let L be the Cholesky triangle of Q. Then for $i \in V$:*

$$E(x_i|x_{(i+1):n}) = \mu_i - \frac{1}{L_{ii}} \sum_{j=1+1}^{n} L_{ji}(x_j - \mu_j)$$
$$Prec(x_i|x_{(i+1):n}) = L_{ii}^2$$

This also implies the following:

Theorem 4. *Let x be a GMRF wrt the graph G with mean μ and precision matrix $Q > 0$. Let L be the Cholesky triangle of Q. Define the future of i except j to be the set:*

$$F(i,j) = \{i+1, ..., j-1, j+1, ..., n\}$$

for $1 \leq i < j \leq n$. Then

$$x_i \perp x_j | x_{F(i,j)} \Longleftrightarrow L_{ji} = 0$$

Proof. Assume $\mu = 0$ WLOG and fix $1 \leq i < j \leq n$. Theorem 2 gives

$$\begin{aligned}\pi(x_{i:n}) &\propto exp\left(-\frac{1}{2}\sum_{k=i}^{n} L_{kk}^2 \left(x_k + \frac{1}{L_{kk}} \sum_{j=k+1}^{n} L_{jk} x_j\right)^2\right) \\ &= exp\left(-\frac{1}{2} x_{i:n}^T Q^{i:n} x_{i:n}\right)\end{aligned}$$

where $Q_{ij}^{i:n} = L_{ii}L_{ji}$. Theorem 2 then implies that

$$x_i \perp x_j | x_{F(i,j)} \Longleftrightarrow L_{ii}L_{ji} = 0$$

This is equivalent to

$$x_i \perp x_j | x_{F(i,j)} \Longleftrightarrow L_{ji} = 0$$

since $Q^{(i:n)} > 0$ implies $L_{ii} > 0$. ■

Thus, if we can verify that L_{ji} is zero, we do not have to compute it. Of course, we do not want to have to verify that L_{ji} is zero by computing it and comparing it to zero. Instead, we can make use of a corollary to Theorem 4:

Corollary 5. *If $F(i,j)$ separates $i < j$ in G then $L_{ji} = 0$.*

This follows from the fact that the global Markov property guarantees that if $F(i,j)$ separates $i < j$ in G then $x_i \perp x_j | x_{F(i,j)}$. Thus, by Theorem 4, $L_{ji} = 0$.

As an example, we can consider the following graph.

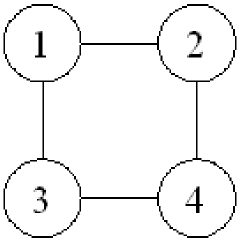

Figure 5.1: Nodes 1 and 4 are conditionally independent given nodes 2 and 3, and nodes 2 and 3 are conditionally independent given nodes 1 and 4.

The precision matrix for this graph could be represented as follows, where ×'s denote non-zero terms and blanks denote terms that may possibly be zeros given the structure of the graph:

$$Q = \begin{pmatrix} \times & \times & \times & \\ \times & \times & & \times \\ \times & & \times & \times \\ & \times & \times & \times \end{pmatrix}.$$

We cannot say that L_{32} is definitely zero, because $F(2,3) = \{4\}$, which is not a separating subset for 2 and 3 (in words, the future of 2 except 3 is just 4). We can, however, say that L_{41} is definitely zero, since $F(1,4) = \{2,3\}$, which separates 1 and 4. Thus, we can fill in the following array for L, where the possibly non-zero terms are denoted by "?":

$$L = \begin{pmatrix} \times & & & \\ \times & \times & & \\ \times & ? & \times & \\ & \times & \times & \times \end{pmatrix}$$

L will always be more or equally dense than the lower triangular part of Q. If we denote the number of possible non-zero elements in L by n_L and the number of possible non-zero elements in Q by n_Q then we are concerned with the *fill-in* $n_L - n_Q$. We obviously want to make this as small as possible to make calculations as fast as possible.

5.1.1 Minimizing the Fill-in

As it turns out, the fill-in depends on the ordering of the nodes. As a motivating example, consider the two graphs below.

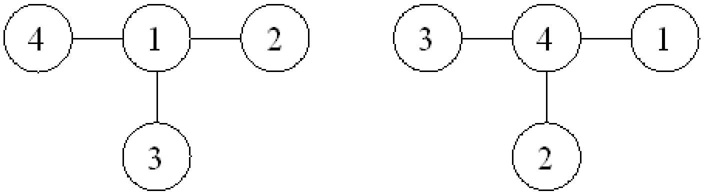

Figure 5.2: Left: graph (a). Right: graph (b).

In the case of graph (a), the precision matrix and its Cholesky triangle are:

$$Q = \begin{pmatrix} \times & \times & \times & \times \\ \times & \times & & \\ \times & & \times & \\ \times & & & \times \end{pmatrix}$$

$$L = \begin{pmatrix} \times & & & \\ \times & \times & & \\ \times & ? & \times & \\ \times & ? & ? & \times \end{pmatrix}.$$

Whereas in the case of graph (b), the precision matrix and its Cholesky triangle are:

$$Q = \begin{pmatrix} \times & & & \times \\ & \times & & \times \\ & & \times & \times \\ \times & \times & \times & \times \end{pmatrix}$$

$$L = \begin{pmatrix} \times & & & \\ & \times & & \\ & & \times & \\ \times & \times & \times & \times \end{pmatrix}.$$

In other words, in (a) the fill-in is maximal and in (b) the fill-in is minimal. The reason is that in (a) all nodes depend on node 1, hence the future of i except j is never a separating subset for $i < j$. In (b), by contrast, conditioning on node 4 makes all other nodes conditionally independent.

In general, then, a good strategy to minimize fill-in is to first select a set of nodes which, if removed, would divide the graph into two disconnected subgraphs of roughly the same size. Then order the nodes in that set after ordering all the nodes in both subgraphs. Then repeat this procedure recursively to the nodes in each subgraph. In numerical and computer science literature this is called *nested dissection*.

When Q is a band matrix we can instead try to reduce the bandwidth. A theorem states that if $Q > 0$ is a band matrix with bandwidth p and dimension n, the Cholesky triangle of Q has (lower) bandwidth p.[1] The trick with spatial data is then to try to make Q a band matrix with as small a bandwidth as possible, and again this is accomplished by re-ordering the nodes.

5.2 Timing Studies

Some studies have quantified the reduction of time required to estimate a GMRF as opposed to a geostatistical model. Using MCMC methods, an $n_r \times n_c$ lattice (with $n_r \leq n_c$) can be simulated using $\mathcal{O}(n_r^3 n_c)$ flops, and repeated i.i.d. samples only require $\mathcal{O}(n_r^2 n_c)$ flops (Rue and Tjelmeland, 2002). In contrast, general Gaussian fields typically require $\mathcal{O}(n_r^3 n_c^3)$ flops to simulate. To calculate the likelihood of a GMRF and a general Gaussian field we also need $\mathcal{O}(n_r^3 n_c)$ and $\mathcal{O}(n_r^3 n_c^3)$ flops, respectively (Rue and Tjelmeland, 2002).

More recently, Rue and Martino (2007) and Rue, Martino and Chopin (2009) have discussed an approach they call integrated nested Laplace approximations (INLAs). This approach appears to outperform MCMC methods in the time required to execute them by several orders of magnitude; simulation only takes seconds to minutes rather than hours to days. The method's primary disadvantage is that its computational cost does increase exponentially with the number of hyperparameters. Yet as Rue, Martino and Chopin conclude, this method may frequently be useful for taking a first pass at one's data, even

[1] A direct proof is in Golub and van Loan (1996).

where one expects the model one is applying to not be a great fit, because at a few seconds or a few minutes, it is practically cost-less. We will look at this method in greater detail after first discussing how GMRFs have traditionally been fit to GGMs.

Chapter 6

The Approximation of GGMs using GMRFs: Traditional Minimizations of Distance

There have been many attempts to use GMRFs to approximate GGMs, given the computational efficiency of GMRFs. The process has traditionally been as follows. First, if necessary, geostatistical data is aggregated up to lattices. Then a GMRF is used to model the aggregated data, with its parameters selected to minimize a distance measure between the GMRF and a particular GGM that has previously been fit to the data. The computations of the distance measures use a fast Fourier transformation, as we will see. To properly discuss this process, we thus need to do two things: 1) detail the distance measures used in the literature; and 2) explain how these distance measures are computed and minimized. In this chapter, we will do these two things and then discuss the benefits and limitations of this line of research.

6.1 Distance Measures

Three main pseudo-distances between Gaussian fields have been proposed: the Kullback-Leibler divergence criterion (which I will abbreviate to KL), Rue and Tjelmeland's (2002) matched-correlation (MC) criterion, and Cressie and Verzelen's (2008) conditional-mean least-squares (CMLS) criterion.

The Kullback-Leibler divergence is defined as:

$$K(f,g) \equiv \int f(x)\log\left(\frac{f(x)}{g(x)}\right) d\mu(x)$$

where f and g are densities and μ is a common measure. This pseudo-distance is historically related to Shannon's measure of entropy, but can be applied to fit the parameters of a GMRF to a Gaussian field by making the following adjustments:

$$KL(f,g;\theta) \equiv \int f(x)\log\left(\frac{f(x)}{g(x;\theta)}\right) dx$$

where f and g are densities of the true and approximated models, respectively, and θ are the parameters of the model. KL is not technically a distance since it is not symmetric and does not satisfy the triangle inequality. If f and g are the joint densities of zero-mean Gaussian fields with covariance matrices Σ_1 and Σ_2, respectively, then

$$KL(\Sigma_1,\Sigma_2) = \frac{1}{2}(log\left(\frac{|\Sigma_2|}{|\Sigma_1|}\right) + tr(\Sigma_1\Sigma_2^{-1}) - n)$$

where $|\Sigma|$ is the determinant of Σ. While this distance measure does somewhat capture the differences between the fields, it is hard to minimize. While minimizing a distance measure to find a GMRF that approximates a Gaussian field is computationally intensive and destroys the computational advantages of using GMRFs, this approach is used to show the extent to which GMRFs can approximate Gaussian fields. Dempster (1972), Besag and Kooperberg (1995), and Rue and Tjelmeland (2002) tried algorithms to minimize this pseudo-distance, which will be returned to in the next section. Dempster's can fail to converge; Rue and Tjelmeland's cannot be applied in the case of an irregular lattice.

The next possible measure that we will consider is Rue and Tjelmeland's matched-correlation criterion. They define

$$MC_\omega(Q) \equiv ||\rho - \rho'||^2_\omega \equiv \sum_{i=1}^{n_r}\sum_{j=1}^{n_c}(\rho_{ij} - \rho'_{ij})^2\omega_{ij}$$

where Q is the precision matrix, ω_{ij} are non-negative weights assigned by the user, and ρ and ρ' are correlation functions. In particular, they recommend, for isotropy:

$$\omega_{ij} \propto \begin{cases} 1 + r & \text{if } ij = 00; \\ (1 + r/d(ij,00))/d(ij,00) & \text{otherwise} \end{cases}$$

where r is the range of the neighbourhood and $d(ij, 00)$ is the Euclidean toroidal distance between (i, j) and $(0, 0)$. The computational costs of minimizing this function are also very high.

Cressie and Verzelen (2008) instead suggest the following criterion:

$$CMLS(Q) \equiv \left(\sum_{i=1}^{n}\frac{1}{Var(X_i|x_{-1})}\right)^{-1} E\left(\sum_{i=1}^{n}\frac{1}{Var(X_i|x_{-i})}(X_i - E_Q(X_i|X_{-i}))^2\right)$$

where $X = (X_1, ..., X_n)$ is a zero-mean Gaussian random vector. The weights, inversely proportional to the conditional variance, are "Gauss-Markov" type weights used in GLS. The intuition provided by the authors for using this measure is that they are interested in obtaining the best prediction of an individual component based on the other components, with weights used to scale the components so that their variability is comparable.

Unfortunately, this criterion is still computationally intensive since it uses essentially the same steps as the other two criteria, as we will see. Variants are slightly better, but this measure's main advantage remains that it can easily handle GMRFs on irregular lattices and is good at predicting unknown values given neighbouring observations (Cressie and Verzelen, 2008). In practice, the minimization of different pseudo-distances typically yields similar results (Song, Fuentes and Ghosh, 2008).

6.2 Minimizing the Distance Measures

The distance measures need to be minimized to obtain the best-fitting GMRF, but how exactly is this accomplished?

Let us start with the case of minimizing the Kullback-Leibler discrepancy since that is still the most frequently used distance measure and all the other minimizations of distances use very similar methods. In the simplest case, let f be a zero mean stationary GGM on a $n_r \times n_c$ torus with covariance function $\gamma(k,l)$ and covariance matrix Σ and let g be a zero mean stationary GMRF on the same torus with precision matrix Q parameterized with θ.[1] Then, as shown in Rue and Tjelmeland (2002),

$$KL(f,g) = -\frac{1}{2}\sum_{i=0}^{n_r-1}\sum_{j=0}^{n_c-1}(\log(\lambda_{ij}q_{ij}(\theta)) - \lambda_{ij}q_{ij}(\theta) + 1)$$

where

$$\begin{aligned}
\lambda_{ij} &= \sum_{k=0}^{n_r-1}\sum_{l=0}^{n_c-1}\gamma(k,l)exp(-2\pi\iota(ki/n_r + lj/n_c)) \\
q_{ij}(\theta) &= \sum\sum_{(k,l)\in\delta_{00}\,\cup\{00\}} Q_{00,kl}(\theta)exp(-2\pi\iota(ki/n_r + lj/n_c))
\end{aligned}$$

[1] GMRFs on a torus are much faster to estimate and thus are frequently used despite the strong assumptions required; indeed, Rue and Held describe how one might approximate a GMRF that is not on a torus with a toroidal GMRF just in order to make calculations faster (2005). But if one does not want to use a GMRF on a torus, one could instead simulate a GMRF that is not on a torus by an alternative algorithm in Rue (2001) which costs $\mathcal{O}(n_r^3 n_c)$ flops or $\mathcal{O}(n_r^2 n_c)$ flops for repeated i.i.d. samples. We already discussed a basic version of this algorithm in the chapter on simulations.

and where δ_{ij} is the neighbourhood of (i,j) and $\iota = \sqrt{-1}$.

The KL discrepancy can hence be represented in terms of the eigenvalues of the covariance and precision matrices of the GGM and GMRF. The eigenvalues themselves are the two-dimensional discrete Fourier transform of the first row of Σ and $Q(\theta)$. Thus, one can use a fast two-dimensional discrete Fourier transform algorithm to evaluate $KL(f,g)$ in $\mathcal{O}(n_r n_c \log(n_r n_c))$ flops or faster. After one has the eigenvalues one can numerically solve the minimization problem to obtain the parameter, θ_{KL}, that provides the best fit:

$$\theta_{KL} = argmin_{\theta} \sum_{i=0}^{n_r-1} \sum_{j=0}^{n_c-1} [\lambda_{ij} q_{ij}(\theta) - \log(q_{ij}(\theta))], \; q_{ij}(\theta) > 0, \forall i, j$$

There is a constraint on this minimization that all eigenvalues be positive, since that gives us a positive definite matrix. Rue and Tjelmeland (2002) suggest proceding as though it were an unconstrained minimization problem with a penalty if some of the eigenvalues are negative, a technique they find works well.

Let us look closer at the fitting procedure based on KL discrepancy. Let our GMRF be parameterized with $(2m+1)^2$ free parameters indexed $\theta_{kl}(k,l) \in \delta_{00} \cup \{00\}$. The (k',l') component of the gradient of $KL(f,g)$, evaluated at the optimal fit, is then

$$\sum_{i=0}^{n_r-1} \sum_{j=0}^{n_c-1} (\lambda_{ij} - \frac{1}{q_{ij}(\theta_{KL})}) \frac{\partial}{\partial \theta_{k'l'}} q_{ij}(\theta_{KL}) = 0 \qquad (6.2.1)$$

where

$$\frac{\partial}{\partial \theta_{k'l'}} q_{ij}(\theta_{KL}) = exp(-2\pi\iota(k'i/n_r + l'j/n_c))$$

(6.2.1) is the inverse discrete Fourier transform of $\{\lambda_{ij}\}$ and $\{1/q_{ij}\theta_{KL}\}$ at (k',l'). Then

$$\gamma(k',l') = \gamma_{KL}(k',l'), \; (k',l') \in \delta_{00} \cup \{00\}$$

since $\{1/q_{ij}\theta_{KL}\}$ are the eigenvalues of the inverse precision matrix Q^{-1}. Essentially, all covariances within $\delta_{00} \cup \{00\}$ are fitted exactly, while those outside are determined by the inversion of the fitted precision matrix. This is also true for other maximum-likelihood estimators of the precision matrix (Dempster, 1972). This is important to note because this means that GMRFs could provide very bad fits for lags outside of the neighbourhoods specified. Getting the neighbourhoods right is thus important in using GMRFs to approximate GGMs with the KL discrepancy as a distance measure.

This difficulty motivates Rue and Tjelmeland's definition of another possible distance function, their matched-correlation function, previously defined as

$$MC_{\omega}(Q) \equiv ||\rho - \rho'||^2_{\omega} \equiv \sum_{i=1}^{n_r} \sum_{j=1}^{n_c} (\rho_{ij} - \rho'_{ij})^2 \omega_{ij}$$

where Q is the precision matrix, ω_{ij} are non-negative weights assigned by the user, and ρ and ρ' are correlation functions. The idea is, in contrast to the KL discrepancy, to take lags outside of the neighbourhood into account.

The correlation function ρ' of the GMRF can then be calculated from the covariance function γ' which is the inverse discrete two-dimensional Fourier transform of $\{1/q_{ij}(\theta)\}$

$$\gamma'(k,l) = \frac{1}{n_r n_c} \sum_{i=0}^{n_r-1} \sum_{j=0}^{n_c-1} \frac{1}{q_{ij}(\theta)} exp(2\pi\iota(ki/n_r + lj/n_c))$$

This allows their distance measure to be calculated using $\mathcal{O}(n_r n_c \log(n_r n_c))$ flops. The distance measure is minimized with respect to θ for fixed $\theta_{00} = 1$ and then the solution is scaled to fit the variance, since the variance of a GMRF can always be fit exactly by scaling its precision matrix. The minimization itself is done in the same manner as for the KL discrepancy, again adding a penalty to a prospective θ if it has any non-positive eigenvalues. Rue and Tjelmeland's distance measure includes the KL discrepancy as a special case where ω is chosen to be positive for $(k,l) \in \delta_{00}$ and zero otherwise.

Finally, regarding Cressie and Verzelen's proposed distance measure

$$CMLS(Q) \equiv \left(\sum_{i=1}^{n} \frac{1}{Var(X_i|x_{-1})} \right)^{-1} E \left(\sum_{i=1}^{n} \frac{1}{Var(X_i|x_{-i})} (X_i - E_Q(X_i|x_{-i}))^2 \right),$$

given that $Var(X_i|x_{-1}) = (Q[i,i])^{-1}$ one could re-write their criterion as

$$CMLS(Q) \equiv \frac{tr(D_Q^{-1} Q \Sigma' Q)}{tr(Q)}$$

where D_Q^{-1} is the diagonal matrix whose diagonal is the same as that of Q. Σ' and Q (the covariance matrix of the GGM and the precision matrix of the GMRF, respectively) are both defined on the same torus and are Toeplitz circulant matrices that are diagonalizable using the same basis. The criterion could then just as well be written

$$CMLS(Q) = n_r n_c \sum_{i=0}^{n_r-1} \sum_{j=0}^{n_c-1} q_{ij}(Q)^2 \lambda_{ij}$$

where λ_{ij} are the eigenvalues of Σ' and are given by

$$\lambda_{ij} = \sum_{i=0}^{n_r-1} \sum_{j=0}^{n_c-1} \Sigma[00, kl] exp(-2\pi\iota(ki/n_r + lj/n_c))$$

and q_{ij} are the eigenvalues of Q and are given by

$$q_{ij} = \sum_{i=0}^{n_r-1} \sum_{j=0}^{n_c-1} Q[00, kl] exp(-2\pi\iota(ki/n_r + lj/n_c))$$

Since the criterion has again been reduced to finding these eigenvalues, one can again use the fast Fourier transform algorithm of Rue and Tjelmeland (2002) as before to evaluate $CMLS(Q)$ with $\mathcal{O}(n_r n_c \log(n_r n_c))$ flops.

The advantages of using the CMLS criterion over the KL or MC criteria come when one has an irregular lattice or wants to predict missing values such as in kriging. This can be explained by considering the evaluation function

$$\eta(Q) = E(E(X_0|x_{-0}) - E_Q(X_0|x_{-0}))^2/var(X_0|x_{-0})$$

where “0” represents the location (0,0), $E_Q(X_i|x_{-i}) = \sum_{j \neq i} -Q[i,j](Q[i,i])^{-1}x_j$ (from the CAR literature (Besag, 1974)) and $E(\cdot)$ and $var(\cdot)$ are moments under the particular Gaussian field being fit. This evaluation function provides a measure of how well a GMRF predicts the value at a node from its neighbours. Cressie and Verzelen note this is in fact what we want to do if we want to approximate a Gaussian field in a Gibbs sampler, and that kriging also uses the conditional expectation in making its predictions. It is because the CMLS criterion is so closely related to the evaluation function that it obviously performs well under it relative to the KL or MC criterion when predicting missing values.

6.3 Estimations of Different GGMs

Again, since all the distance measures have been evaluated using roughly the same algorithm, they all take a similar amount of time.

As for the closeness of the fit on different kinds of GGMs, Rue and Tjelmeland (2002) had great success in fitting the exponential and Matérn correlation functions for neighbourhood structures larger than 3 x 3, with maximal absolute differences in values of frequently less than 1% using the MC criterion (they do not report similar results for the other criteria). Gaussian and spherical functions were more difficult to fit but provided reasonably accurate fits once the neighbourhoods were larger than 5 x 5 (with maximal absolute differences below about 5%). While the accuracy of the fits increased as the neighbourhoods increased, the computational costs rose in tandem as the neighbourhoods grew. Song, Fuentes and Ghosh (2008) report this as well and also find that their fits become better as the range increases and the conditional variance decreases. While each type of distance measure has been shown to work better than the others for some kinds of problems, in practice their minimizations tend to yield similar results (Song, Fuentes and Ghosh, 2008).

It is easy to imagine that in the future more distance measures could be proposed that would also perform slightly better under special circumstances. However, for improvements in time costs, a new algorithm to compute these distances would have to be found. It is also not clear that the focus of the literature has been particularly useful, since to take advantage of GMRFs' computational efficiency one should want to apply them to the data directly and not to GGMs that were themselves fit to the data. It is true that if one had a set of known correspondences between certain GGMs and GMRFs one could find the best-fitting GMRF satisfying certain conditions and then read off the corresponding GGM, but the literature has mostly steered clear of looking for systematic relationships between GGMs and GMRFs. The main work that attempts to examine the relationships between different covariance functions and GMRFs in a systematic way is Hrafnkelsson and Cressie (2003), who restrict their attention to the Matérn class of covariance functions because of its popularity as well as the interpretability of its parameters.

Recall that the Matérn class of covariance functions is specified by two parameters - a scale parameter ($a > 0$) that controls the range of correlation and a smoothness parameter ($\nu > 0$) that controls the smoothness of the random field. For stationary and isotropic processes, which are the only kinds considered, their correlation functions are of the form:

$$\rho(h) = \frac{1}{2^{\nu-1}\Gamma(\nu)}\left(\frac{2h\sqrt{\nu}}{a}\right)^{\nu} K_\nu\left(\frac{2h\sqrt{\nu}}{a}\right); h \geq 0$$

where h is lag distance and K_ν is the modified Bessel function of order ν (Hrafnkelsson and Cressie, 2003). If S is a vector of random variables at various locations and

$$E(S(x_i)|\{s(x_j) : j \neq i\}) = \mu_i + \phi \sum_{j=1}^{n} c_{ij}\{s(x_j) - \mu_j\},$$

$$var(S(x_i)|\{s(x_j) : j \neq i\}) = \tau_i^2,$$

and $S(x_i)|\{s(x_j) : j \neq i\}$ is Gaussian, then the joint distribution of S is $N(\mu, (I - \phi C)^{-1}T)$ where C is a $n \times n$ matrix whose (i, j)th element is c_{ij} and T is an $n \times n$ matrix with $\tau_1^2 - \tau_n^2$ on the diagonal. To specify a GMRF we would need to estimate the spatial dependence parameter ϕ; we would also need to decide when the spatial dependence has decreased to 0, a point which Hrafnkelsson and Cressie (2003) call r_{max}. Then C depends on r_{max}. Hrafnkelsson and Cressie investigate a particular class of C's only and, holding everything else fixed, vary r_{max} and ϕ. For each pair (r_{max}, ϕ) they estimate the Matérn covariance parameters a and ν with non-linear least squares, minimizing:

1. $\sum_i \sum_j [\log[a_{ij}] - \log\{h_1(r_{max,i}, \phi_j)\}]^2$ where a is our first parameter (the range parameter) and $h_1(\cdot)$ is a potential parameterized function (taking the logs so that the model fits for both small and large a).

2. $\sum_i \sum_j \{\nu_{ij} - h_2(r_{max,i}, \phi_j)\}^2$ where ν is our second parameter (the smoothness parameter) and $h_2(\cdot)$ is a potential parameterized function.

In particular, they specify the following functional forms for $h_1(\cdot)$ and $h_2(\cdot)$:

$$h_1(r_{max}, \phi) = exp\left(\gamma_i e^{\lambda_1(r_{max-1})} \frac{\phi^{\beta}_{1}}{(1-\phi)^{\alpha}_{1}}\right) - 1$$

$$h_2(r_{max}, \phi) = \frac{\eta_2}{(\phi + \gamma_2(r_{max})^{\lambda_2})^{\beta_2}}$$

where $\alpha_1 > 0, \beta_1 > 0, \gamma_1 > 0, \lambda_1 > 0, \beta_2 > 0, \eta_2 > 0, \gamma_2 > 0$ and $\lambda_2 > 0$ are estimated parameters. These functional forms are not justified beyond noting that they seem to fit a surface plot between (a, ν) and r_{max} and ϕ relatively well.

While this approach is a little *ad hoc*, their estimated parameters do seem to fit previous studies. Griffith and Csillag (1993) and Griffith, Layne, and Doyle (1996), in particular, found that certain GMRFs closely approximated certain Matérn GGMs, and Hrafnkelsson and Cressie's estimates do obtain similar results (2003).

Chapter 7

The Approximation of GGMs Using GMRFs: A SPDE and INLAs

7.1 A Different Approach: Approximations Using a SPDE

The latest development in using GMRFs to approximate GGMs comes at the problem from an entirely different angle. Rather than mechanically finding parameterizations that map GMRFs to GGMs as Hrafnkelsson and Cressie did for the Matérn class of GGMs (2003), Lindgren, Lindström and Rue find that an approximate solution to a particular stochastic partial differential equation (SPDE) can explicitly link GMRFs and GGMs for a restricted class of GGMs (a subset of the Matérn class) (2010).

It has long been known that the solution $x(u)$ of the linear fractional stochastic partial differential equation

$$(\kappa^2 - \Delta)^{\alpha/2} x(u) = W(u), u \in \mathbb{R}^d, \alpha = \nu + \frac{d}{2}, \kappa > 0, \nu > 0$$

(where W is spatial Gaussian white noise with unit variance, $\Delta = \sum_{i=0}^{d} \frac{\delta^2}{\delta x_i^2}$, and κ is the scaling parameter from the Matérn covariance function between locations u and v, $\frac{\sigma^2}{\Gamma(\nu)2^{\nu-1}}(\kappa||v-u||)^{\nu}K_{\nu}(\kappa||v-u||)$ where σ^2 is the marginal variance $\frac{\Gamma(\nu)}{\Gamma(\nu+\frac{d}{2})(4\pi)^{\frac{d}{2}}\kappa^{2\nu}}$) is a Gaussian field with Matérn covariance function (Whittle, 1954, 1963).

I will state Lindgren, Lindström and Rue's main results without proof. Proofs can be found in the long appendices of their 2010 paper. To first give a quick overview, their main result is that the linear fractional SPDE given earlier can be represented as a GMRF on a regular or irregular 2D lattice. Their explicit mapping between a GGM and its GMRF representation costs only $\mathcal{O}(n)$, involving no computing.

In particular, they use the stochastic weak solution of the SPDE which requires that $\{\langle \phi_j, (\kappa^2 - \Delta)^{\alpha/2} x \rangle, j = 1, ..., n\} =^d \{\langle \phi_j, \epsilon \rangle, j = 1, ..., n\}$ for every finite set of test functions $\{\phi_j(u), j = 1, ..., n\}$. The finite element representation of the solution to the SPDE is $x(u) = \sum_{k=1}^{n} \psi_k(u) w_k$ for some basis functions ψ_k and Gaussian weights w_k (Kleoden and Platen, 1999; Kotelenez, 2007).

If C, G and K are $n \times n$ matrices such that $C_{ij} = \langle \psi_i, \psi_j \rangle$, $G_{ij} = \langle \nabla\psi_i, \nabla\psi_j \rangle$, and $K_{ij}(\kappa^2) = \kappa^2 C_{ij} + G_{ij}$ then their approximations for $Q_\alpha(\kappa^2)$ are as follows:

$$Q_1(\kappa^2) = K(\kappa^2)$$
$$Q_2(\kappa^2) = K(\kappa^2)C^{-1}K(\kappa^2)$$

and for $\alpha = 3, 4, ...$

$$Q_\alpha(\kappa^2) = K(\kappa^2)C^{-1}Q_{\alpha-2}C^{-1}K(\kappa^2)$$

These expressions make the approximations immediate; Lindgren, Lindström and Rue also provide analytic formulas for C and G in an appendix.

Their results only hold for certain values of the smoothness parameter ν in the Matérn covariance function, however: $\nu = \frac{1}{2}, \frac{3}{2}, \frac{5}{2}, ...,$ in $\mathbb{R}^1$ and $\nu = 0, 1, 2, ...,$ in $\mathbb{R}^2$. α must also be a strictly positive integer. Further, it should be emphasized that their method of representing certain GGMs using GMRFs is only an approximation. That being said, it is fast: Lindgren, Lindström and Rue tested it on Kneib and Fahrmeir's dataset (2007), which Kneib and Fahrmeir had not been able to model using a dense covariance matrix, and found it took seconds to complete a Bayesian analysis of it (2010). This dataset had a 5258 × 5258 precision matrix.

7.2 Integrated Nested Laplace Approximations

This leads to the other recent innovation; Rue, Martino and Chopin's championing of using integrated nested Laplace approximations (INLAs) rather than MCMC methods (2009). It is INLAs that Lindgren, Lindström and Rue use on Kneib and Fahrmeir's dataset. While this is a different kind of innovation, it has important ramifications for GMRF estimation and should be discussed.

For what follows, let x be the vector of all the latent Gaussian variables, θ the vector of hyperparameters, y observational variables, and $\pi(\cdot|\cdot)$ a conditional density. The idea behind INLAs is that, being interested in the posterior marginals

$$\begin{aligned} \pi(x_i|y) &= \int \pi(x_i|\theta, y)\pi(\theta|y)d\theta \\ \pi(\theta_j|y) &= \int \pi(\theta|y)d\theta_{-j} \end{aligned}$$

we can approximate them in this form:

$$\begin{aligned} \tilde{\pi}(x_i|y) &= \int \tilde{\pi}(x_i|\theta, y)\tilde{\pi}(\theta|y)d\theta \\ \tilde{\pi}(\theta_j|y) &= \int \tilde{\pi}(\theta|y)d\theta_{-j} \end{aligned}$$

$\pi(x_i|y)$ is approximated by approximating both $\pi(\theta|y)$ and $\pi(x_i|\theta, y)$ and integrating out θ through numerical integration. $\pi(\theta_j|y)$ is approximated by similarly integrating out θ_{-j} from the approximation of $\pi(\theta|y)$. I will walk through Rue, Martino and Chopin's methods for approximating $\pi(\theta|y)$ and $\pi(x_i|\theta, y)$ in turn.

7.2.1 Approximating $\tilde{\pi}(\theta|y)$

1. Optimize $\log\{\tilde{\pi}(\theta|y)\}$ with respect to θ to obtain the mode of $\{\tilde{\pi}(\theta|y)\}$ (in particular, they recommend using a quasi-Newton-Raphson method to approximate the second derivative of $\log\{\tilde{\pi}(\theta|y)\}$ by taking the difference between successive gradient vectors, the gradient in turn being approximated by using finite differences). Call this mode θ^*.

2. At θ^*, compute the negative Hessian matrix, H, using finite differences. Define $\Sigma \equiv H^{-1}$. If the density were Gaussian, this would be the covariance matrix for θ. Let $\Sigma = V\Lambda V^T$ be the eigendecomposition of Σ and define θ to be a function of a standardized variable z: $\theta(z) = \theta^* + V\Lambda^{\frac{1}{2}}z$. If $\tilde{\pi}(\theta|y)$ is Gaussian then $z \sim N(0, I)$. The point of this reparameterization is to make later numerical integration easier.

3. Now some points are selected to be used in the numerical integration. A space is searched with points being added to the collection if $\log\{\tilde{\pi}(\theta|y)\}$ is considered significant, in the following sense. Starting from the mode ($z = 0$), one travels in the positive direction of z_1 with step length δ_z as long as $\log[\tilde{\pi}\{\theta(0)|y\}]$-$\log[\tilde{\pi}\{\theta(z)|y\}] < \delta_\pi$. After travelling as far as possible in one direction along z_1 and accumulating points along this route, one switches direction and goes the opposite direction along z_1 as long as the condition holds. Then the process is repeated along z_2. These points will turn out to be on a regular grid, so that area weights Δ_k will later be able to be taken to be equal.

4. We would like to use numerical integration on $\tilde{\pi}(\theta|y)$ but this is computationally demanding since it would entail evaluating $\tilde{\pi}(\theta|y)$ at a large number of points. Instead, the authors suggest computing $\tilde{\pi}(\theta|y)$ at only the points that were selected in the previous step. They remark that for higher accuracy one should use more points and select δ_z accordingly. Rue and Martino (2007) provide some empirical guidance as to how the relative size of δ_z might affect results.

7.2.2 Approximating $\tilde{\pi}(x_i|\theta, y)$

To approximate $\tilde{\pi}(x_i|\theta, y)$, Rue, Martino and Chopin suggest a few different methods. A Gaussian approximation would be simplest; a Laplace approximation would be most accurate but would require too many computations; their preferred method is a simplified Laplace approximation. Here I will detail each in turn.

1. A Gaussian approximation $\tilde{\pi}_G(x_i|\theta, y)$ would be the simplest solution to implement and also computationally the fastest. They note that $\tilde{\pi}_G(x|\theta, y)$ was already calculated during the approximating of $\tilde{\pi}(\theta|y)$, so all that would need to be additionally calculated are the marginal variances, which could be derived by using the following recursion: $\Sigma_{ij} = \frac{\delta_{ij}^2}{L_{ii}} - \frac{1}{L_{ii}} \sum_{k=i+1}^{n} L_{ki}\Sigma_{kj}, j \geq i, i = n, ..., 1$ where L_{ij} are from the Cholesky decomposition of $Q = LL^T$ and $\delta_{ij} = 1$ if $i = j$ and 0 otherwise. In fact, we only would need to calculate Σ_{ij} for the (i, j)s for wihch we did not know L_{ij} was 0.

2. The Laplace approximation is given by:

$$\tilde{\pi}_{LA}(x_i|\theta, y) \propto \left. \frac{\pi(x, \theta, y)}{\tilde{\pi}_{GG}(x_{-i}|x_i, \theta, y)} \right|_{x_{-i}=x^*_{-i}(x_i,\theta)}$$

   where $\tilde{\pi}_{GG}$ is the Gaussian approximation to $x_{-i}|x_i, \theta, y$ and $x^*_{-i}(x_i, \theta)$ is the modal configuration. The difficulty with this approximation is that it would require $\tilde{\pi}_{GG}$ to be computed for every value of x_i and θ. One modification would be to approximate $\tilde{\pi}_{GG}$ by approximating the modal configuration as follows: $x^*_{-i}(x_i, \theta) = E_{\tilde{\pi_G}}(x_{-i}|x_i)$. Then the RHS could be evaluated using the conditional density derived from the Gaussian approximation $\tilde{\pi}_G(x_i|\theta, y)$. A second possibility would be to instead approximate $\tilde{\pi}_{LA}$ by

$$\tilde{\pi}_{LA}(x_i|\theta, y) \propto N\{x_i; \mu_i(\theta), \sigma_i^2(\theta)\}exp\{\text{cubic spline}(x_i)\}$$

The cubic spline is done on the difference of the log-density of $\tilde{\pi}_{LA}(x_i|\theta, y)$ and $\tilde{\pi}_G(x_i|\theta, y)$ at the selected points and the density is normalized using quadrature integration. The authors prefer to correct for location and skewness in the Gaussian approximation $(\tilde{\pi}_G(x_i|\theta, y))$, however, by following the method below.

3. Rue, Martino and Chopin's preferred approximation of $\tilde{\pi}(x_i|\theta, y)$ involves obtaining a simplified Laplace approximation $\tilde{\pi}_{SLA}(x_i|\theta, y)$ by expanding $\tilde{\pi}_{LA}(x_i|\theta, y)$ around $x_i = \mu_i(\theta)$. They use

$$\tilde{\pi}_{LA}(x_i|\theta, y) \propto \left.\frac{\pi(x, \theta, y)}{\tilde{\pi}_{GG}(x_{-i}|x_i, \theta, y)}\right|_{x_{-i}=x^*_{-i}(x_i,\theta)}$$

and $x^*_{-i}(x_i, \theta) = E_{\tilde{\pi_G}}(x_{-i}|x_i)$ and after some calculations obtain an estimate

$$\log\{\tilde{\pi}_{SLA}(x_i^s|\theta, y)\} = \text{constant} - \frac{1}{2}(x_i^{(s)})^2 + \gamma_i^{(1)}(\theta)x_i^{(s)} + \frac{1}{6}(x_i^{(s)})^3\gamma_i^{(3)}(\theta) + ... \quad (7.2.1)$$

where

$$\gamma_i^{(1)}(\theta) = \frac{1}{2}\sum_{j \in I_1} \sigma_j^2(\theta)\{1 - corr_{\tilde{\pi}_G}(x_i, x_j)^2\}d_j^{(3)}\{\mu_i(\theta), \theta\}\sigma_j(\theta)a_{ij}(\theta)$$

$$\gamma_i^{(3)}(\theta) = \sum_{j \in I_1} d_j^{(3)}\{\mu_i(\theta), \theta\}\{\sigma_j(\theta)a_{ij}(\theta)\}^3$$

and $d_j^{(3)}(x_i, \theta) = \frac{\partial^3}{\partial x_j^3}log\{\pi(y_j|x_j, \theta)\}\Big|_{x_j=E_{\tilde{\pi}_G}(x_j|x_i)}$, for some a_{ij} derived from $\frac{E_{\tilde{\pi}_G}-\mu_j(\theta)}{\sigma_j(\theta)} = a_{ij}(\theta)\frac{x_i-\mu_i(\theta)}{\sigma_i(\theta)}$, an equation which is in turn implied by the conditional expectation $x^*_{-i}(x_i, \theta) = E_{\tilde{\pi_G}}(x_{-i}|x_i)$.[1]

[1] (7.2.1) does not define a density since the third order term is unbounded. Rue, Martino and Chopin deal with this by fitting a skew normal density so that the third derivative at the mode is $\gamma_i^{(3)}$, the mean is

To compute this approximation one needs to calculate $a_{i\cdot}$ for each i; most of the other terms do not depend on i and hence are only computed once. The cost of computing $\log\{\tilde{\pi}_{SLA}(x_i^s|\theta, y)\}$ for a particular i is of the same order as the number of non-zero elements of the Cholesky triangle, or $\mathcal{O}(n \log(n))$. For each value of θ, repeating this process n times gives a total cost of $\mathcal{O}(n^2 \log(n))$, which they "believe" is close to the lower limit for any algorithm that approximates all n marginals. Each site must also be visited for each i since the graph of x is general, an operation which by itself costs $\mathcal{O}(n^2)$. Thus, the total cost of computing all n marginals $\tilde{\pi}(x_i|y)$ is exponential in the dimension of θ times $\mathcal{O}\{n^2 \log(n)\}$.

7.2.3 Summary

Rue, Martino and Chopin's method garnered much interest, as a new and fast method, but questions remain.

In particular, since it is such a new method, there are as yet relatively few results using it and it is not clear how accurate the approximations would be with different datasets. The asymptotics of the approximation error have not been fully explored, although Rue, Martino and Chopin claim that it produces "practically exact" results and that non-negligible bias occurs only in "pathological" cases (2009). The other main problem with their technique is simply in ease of use; many commenters seem to fear that the method would only be useful when used with a packaged "black box" program since it would take a lot of time to implement for the first time (the authors do have R code available for a limited selection of applications). Further, while the method has very low computational costs in many situations, it is not a cure-all; in particular, its computational cost is exponential in the number of hyperparameters and thus it may not be appropriate for situations in which the number of hyperparameters is large.

$\gamma_i^{(1)}$, and the variance is 1. They defend this choice in an appendix. There are also special cases for which they fit a spline-corrected Gaussian instead of a skewed normal distribution, but this is peripheral.

All that being said, whereas the traditional literature which focused more on defining new distance measures seemed to be stagnating in its reliance on the same general algorithm, these recent innovations are substantially different from what came before and have great promise to spur new work. Many of the current difficulties remaining may resolve themselves in time.

Chapter 8

Conclusion

There are clearly many ways of modelling spatial correlation. This thesis discussed some of the main methods used, focusing on Gaussian Markov Random Fields.

We saw that GMRFs can be motivated as being useful when we have observations that are far enough apart that we can call them conditionally independent given the intermediary observations. Our data should also be at an appropriate scale to capture the interactions between nodes that are modelled, and sites should neighbour each other as much as possible, with few borders with the outside world. We noted that much data modelled with GMRFs is actually not entirely appropriately modelled by GMRFs. For example, satellite data that appears as a lattice may hide some interactions between plots due to the aggregation of finer-grained data into a single value for each plot. While relationships between the plots could still be estimated, they might not be capturing the true underlying relationships.

We also discussed the benefits and drawbacks of CARs and SARs. Noting the importance of intrinsic GMRFs in specifying priors, we then discussed in detail how we might specify the precision matrices for a few important classes of models. We explored how and why GMRFs can be made to be so computationally efficient. Finally, we surveyed the literature on using GMRFs as computationally efficient proxies for GGMs. Much of this literature focused on minimizing a distance measure between a particular GMRF and GGM; we also discussed an attempt to empirically estimate a more systematic relationship between the two. Finally, we turned to the most recent literature, which finds that a small subset of GGMs can be explicitly linked to GMRFs through a SPDE, though this is only through an approximation, and which pioneers a new method for fitting models, also based on approximations which have as yet little explored errors.

In summary, we have repeatedly seen that GMRFs are only truly appropriate in a very limited number of cases. However, for those cases for which GMRFs are appropriate, they are extremely computationally efficient. Also, GMRFs show promise in being able to approximate GGMs, but there is still much room for improvement.

References

Banerjee, S. and A.E. Gelfand (2003). “On smoothness properties of spatial processes”, Journal of Multivariate Analysis, vol. 84.

Bao, S. (2004). “Literature review of spatial statistics and models”, China Data Center, University of Michigan.

Bartlett, M.S. (1974). “The statistical analysis of spatial pattern”, Advanced Applied Probability, vol. 6.

Besag, J. (1974). “Spatial interaction and the statistical analysis of lattice systems”, Journal of the Royal Statistical Society, vol. 36B.

Besag, J. (1975). “Statistical analysis of non-lattice data”, The Statistician, vol. 24.

Besag, J. (1981). “On a system of two-dimensional recurrence equations”, Journal of the Royal Statistical Society, vol. 43B.

Besag, J. and C. Kooperberg (1995). “On conditional and intrinsic autoregressions”, Biometrika, vol. 84.

Cressie, N.A.C. (1993). Statistics for Spatial Data. New York: John Wiley & Sons, Inc.

Cressie, N.A.C. and N. Verzelen (2008). “Conditional-mean least-squares of Gaussian Markov random fields to Gaussian fields”, Computational Statistics and Data Analysis, vol. 52

Dempster, A.M. (1972). “Covariance selection”, Biometrics, vol. 28.

Diggle, P.J., Tawn, J.A. and R.A. Moyeed (1998). “Model-based geostatistics”, Journal of the Royal Statistical Society, Series C: Applied Statistics, vol. 47.

Freeman, G.H. (1953). “Spread of disease in a rectangular plantation with vacancies”, Biometrika, vol. 40.

Griffith, D. and F. Csillag (1993). “Exploring relationships between semi-variogram and spatial autoregressive models”, Papers in Regional Science, vol. 72.

Griffith, D., Layne, L.J. and Doyle, P.G. (1996). “Further explorations of relationships between semi-variogram and spatial autoregressive models”, Spatial Accuracy Assessment in Natural Resources and Environmental Sciences: Second International Sy

Haining, R. (1990). Spatial data analysis in the social and environmental sciences. Cambridge: Cambridge University Press.

Hammersley, J.M. and P. Clifford (1971). “Markov field on finite graphs and lattices”, unpublished.

Hrafnkelsson, B. and N. Cressie (2003). “Hierarchical modeling of count data with application to nuclear fall-out”, Environmental and Ecological Statistics, vol. 10.

Ising, E. (1925). “Beitrag zur Theorie des Ferromagnetismus”, Zeitschrift fur Physik, vol. 31.

Kleoden, P.E. and E. Platten (1999). Numerical solution of stochastic differential equations, 3^{rd} *ed.* New York: Springer.

Kneib, T. and L. Fahrmeir (2007). "A mixed model approach for geoadditive hazard regression", Scandinavian Journal of Statistics, vol. 34.

Kotelenez, P. (2007). Stochastic ordinary and stochastic partial differential equations. New York: Springer.

Kunsch, H.R. (1987). "Statistical aspects of self-similar processes", in Y. Prohorov and V.V. Sazanov, *eds.*, Proceedings of the First World Congress of Bernoulli Society, vol. 1.

Lenz, W. (1920). "Beitrage zum Verstandnis der magnetischen Eigenschaften in festen Korpern", Physikalische Zeitschrift, vol. 21.

Lichstein, J.W. *et al.* (2002). "Spatial autocorrelation and autoregressive models in ecology", Ecological Monographs, vol. 72.

Lindgren, F., Lindström, J and H. Rue (2010). "An explicit link between Gaussian fields and Gaussian Markov random fields: The SPDE approach", Preprints in Mathematical Sciences.

Mercer, W.B. and A.D. Hall (1911). " The experimental error of field trials", Journal of Agricultural Science, vol. 4.

Patankar, V. (1954). "The goodness of fit of the frequency distribution obtained from stochastic processes", Biometrika, vol. 41.

Ripley, B.D. (1989). “Gibbsian interaction models”, in D.A. Griffiths (*ed.*), Spatial Statistics: Past, Present and Future.

Rue, H. and L. Held (2005). Gaussian Markov random fields: Theory and applications. New York: Chapman & Hall.

Rue, H., Martino, S. and Chopin, N. (2009). “Approximate Bayesian inference for latent Gaussian models by using integrated nested Laplace approximations”, Journal of the Royal Statistical Society, Series B: Statistical Methodology, vol. 71.

Rue, H. and S. Martino (2007). “Approximate Bayesian inference for hierarchical Gaussian Markov random field models”, Journal of Statistical Planning and Inference, vol. 137.

Rue, H. and H. Tjelmeland (2002). “Fitting Gaussian Markov random fields to Gaussian fields”, Scandinavian Journal of Statistics, vol. 29.

Rue, H. (2001). “Fast sampling of Gaussian Markov random fields”, Journal of the Royal Statistical Society, Series B: Statistical Methodology, vol. 63.

Song, H.-R., Fuentes, M. and S. Ghosh (2008). “A comparative study of Gaussian geostatistical models and Gaussian Markov random field models”, Journal of Multivariate Analysis, vol. 99.

Speed, T.P. and H.T. Kiiveri (1986). “Gaussian Markov distributions over finite graphs”, Annals of Statistics, vol. 14.

Spitzer, F. (1971). "Markov random fields and Gibbs ensembles", American Mathematical Monthly, vol. 78.

Waller, L. and C. Gotway (2004). Applied Spatial Statistics for Public Health Data. New York: John Wiley and Sons.

Whittle, P. (1954). "On stationary processes in the plane", Biometrika, vol. 41.

Whittle, P. (1963). "Stochastic processes in several dimensions", Bulletin of the International Statistical Institute, vol. 40.