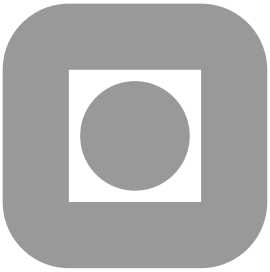
NORGES TEKNISK-NATURVITENSKAPELIGE
UNIVERSITET

Gaussian Markov Random Field Models With Applications in Spatial Statistics

by

Håvard Rue and Turid Follestad

PREPRINT
STATISTICS NO. 5/2003



NORWEGIAN UNIVERSITY OF SCIENCE AND
TECHNOLOGY
TRONDHEIM, NORWAY

This report has URL http://www.math.ntnu.no/preprint/statistics/2003/S5-2003.ps
Håvard Rue has homepage: http://www.math.ntnu.no/~hrue
E-mail: hrue@stat.ntnu.no
Address: Department of Mathematical Sciences, Norwegian University of Science and Technology, N-7034 Trondheim, Norway.

Gaussian Markov Random Field Models With Applications in Spatial Statistics

Håvard Rue and Turid Follestad
Department of Mathematical Sciences
Norwegian University of Science and Technology
N-7491 Trondheim, Norway

June 20, 2003

Abstract

Gaussian Markov Random Field (GMRF) models are frequently used in statistics, and in spatial statistics in particular. The analytical properties of the Gaussian distribution are convenient and the Markov property invaluable when constructing single site Markov chain Markov Carlo algorithms. Rue (2001) demonstrates how numerical methods for sparse matrices can be utilised to construct efficient algorithms for unconditional and various forms for conditional sampling and for the evaluation of the log normalised density. These algorithms allow for constructing block-MCMC algorithms, where all parameters involved, including hyper-parameters, can often be updated jointly in one block. The convergence properties of such algorithms are superior compared to their single-site versions.

This paper reviews the basic properties of a GMRF and how to take advantage of sparse matrix algorithms for sampling and evaluation of the log normalised density. We then discuss two topics mentioned briefly in Rue (2001): How to take advantage of more modern techniques for sparse matrices compared to more classical band-matrix methods, and how to sample a GMRF under a soft linear constraint. We apply and illustrate these techniques on two problems in spatial epidemiology. The first is a semi-parametric ecological regression problem presented by Natário and Knorr-Held (2002). The second is concerned with the modelling of a smoothly varying disease risk surface from area-level aggregated disease counts using an underlying Gaussian field model, motivated by the work of Kelsall and Wakefield (2002).

[*This paper is based on an invited talk by HR at the 19th Nordic Conference on Mathematical Statistics June 9-13, 2002 Stockholm, Sweden*]

Keywords: Gaussian Markov random fi elds, Markov chain Monte Carlo, block-sampling, conditional auto-regression, disease mapping, geostatistics, numerical methods for sparse matrices.

1 Introduction

Gaussian Markov Random Field (GMRF) models, or conditional auto-regressions, are frequently used as components in statistical models, in particular for spatial models (Besag, 1974; Besag and Kooperberg, 1995; Besag and Higdon, 1999; Cressie, 1993). All the analytical results for the Gaussian distribution, combined with a Markov property, make such models useful not only in its own right but also as components in a larger model. The Markov property is nearly a requirement for constructing single-site MCMC algorithms, as the conditional density for one component only depends on a few other components, denoted the neighbours.

Rue (2001) demonstrates that the Markov property of a GMRF makes it possible to utilise numerical methods for sparse matrices to construct fast algorithms for computations on a GMRF, when all task are formulated in terms of operations on the precision matrix. This is due to the direct connection between the non-zero pattern of the precision matrix and the Markov properties of the GMRF. Further, this provides a unified computational framework for GMRFs for which special algorithms for dynamic models based on the Kalman-filter is a special case, see Knorr-Held and Rue (2002, Appendix). When applying a GMRF in a statistical model, a non-Gaussian likelihood will often make the full conditional for the GMRF non-Gaussian, but the Markov properties are in most cases retained. The fast algorithms for GMRFs make it feasible to construct GMRF approximations to the full conditional, and based on these to construct block-MCMC algorithms for exploring the posterior (Rue, 2001; Knorr-Held and Rue, 2002). Block-sampling algorithms have also been applied for generalised additive and semi-parametric mixed models with GMRF priors (Fahrmeir and Lang, 2001; Lang and Bretzger, 2002), and dynamic models (Shephard and Pitt, 1997; Gamerman, 1998; Knorr-Held, 1999). Although block-updating the GMRF will generally improve the convergence, Knorr-Held and Rue (2002) found empirically that by constructing joint block-updates of the GMRF and its hyper-parameters, further improvements in convergence were achieved at virtually no extra cost. The reason is the strong interaction between the hyper-parameters and the GMRF, which is not resolved by block-updating the GMRF only. Rue, Steinsland and Erland (2003) discuss how to improve the GMRF approximation of the full conditional of the GMRF by constructing a class of non-Gaussian approximations that are adaptive to the non-Gaussian likelihood and have the same computational complexity as the GMRF.

In this paper, we first give a short introduction to the GMRF, describing its basic properties and how computations involving a GMRF relates to numerical methods for sparse matrices. Then we discuss two topics briefly mentioned but not pursued by Rue (2001). First, we point at the benefits by using more modern and complex techniques for sparse matrices as an alternative to the classical band-matrix approach, and then we provide the details for efficient sampling under a soft linear constraint (a linear constraint observed with Gaussian error). We illustrate the methods by two case studies both concerned with modelling the geographical variation of the risk of a disease. In our first example, we re-estimate the parameters of the semi-parametric ecological regression model of Natário and Knorr-Held (2002), who model larynx cancer mortality rates using corresponding lung cancer rates, regarded as a surrogate for smoking, as a covariate. The effect of the covariate is modelled semi-parametrically as a smooth function taking one of a set discretised values, and since these values represent a global parameter vector, the band-matrix approach becomes inefficient compared to modern techniques for sparse matrices. Using these modern techniques, we show how to construct a block-MCMC algorithm for all the parameters in the model, leading to superior speed and convergence properties. We also provide a theoretical justification of this way of constructing block-MCMC algorithms, which was

missing in the initial work by Knorr-Held and Rue (2002). In our second application, we consider the problem of estimating a smooth risk surface based on disease counts aggregated in a set of disjoint areas, motivated by the recently proposed geostatistical approach by Kelsall and Wakefield (2002). They model the spatial variation of disease risk using an underlying Gaussian field. Conditional on the disease counts they specify a posterior model for the risk on the level of aggregation of the data, that requires an approximation to the joint distribution of the area-level risks using the moments of the Gaussian random field. Based on this model, they construct a single-site MCMC algorithm for inference. By utilisation of the algorithm for sampling under a soft linear constraint and making use of a GMRF as a proxy for a Gaussian field (Rue and Tjelmeland, 2002), we will demonstrate how the approximation can be avoided, and how to construct efficient block-MCMC algorithms from the initial model.

The paper is organised as follows. In Section 2 we review basic properties and operations on the GMRF including recent advances in efficient sampling from GMRFs, and we apply these methods to two applications in Section 3.

2 Basic properties of and operations on GMRFs

In this section we review the basic properties of GMRFs. We first give a definition of a GMRF and next describe how the conditional independence properties of the GMRF allows for efficient computations using numerical algorithms for sparse matrices.

2.1 Definition of a GMRF

Let $\boldsymbol{x} = (x_1, x_2, \ldots, x_n)^T$ be a Gaussian random field (GRF) with mean $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$, that is, $\boldsymbol{x} \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$. The precision matrix of $\boldsymbol{x}$ is denoted by $\boldsymbol{Q}$ and $\boldsymbol{Q} = \boldsymbol{\Sigma}^{-1}$. The GRF $\boldsymbol{x}$ is said to be a Gaussian Markov random field (GMRF) with respect to the labelled undirected graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$, if the nodes are $\mathcal{V} = \{1, \ldots, n\}$ and the edges

$$\mathcal{E} = \{\{i, j\} \in \mathcal{V} \times \mathcal{V} \,:\, Q_{ij} \neq 0 \quad \text{and} \quad i \neq j\}.$$

If $\{i, j\} \in \mathcal{E}$, then i and j are said to be neighbours, and we write this $i \sim j$. The conditional independence structure of the GMRF is related to the non-zero pattern of the precision matrix by $x_i \perp x_j \mid \boldsymbol{x}_{-ij} \Leftrightarrow Q_{ij} = 0$, $i \neq j$. Here, $\boldsymbol{x}_{-ij}$ denote all elements of $\boldsymbol{x}$ except elements i and j. As a consequence of the correspondence between the non-zero pattern of $\boldsymbol{Q}$ and the conditional independence structure of the GMRF, the GMRF is typically specified in terms of its conditional moments. These are given by $\text{Var}(x_i \mid \boldsymbol{x}_{-i}) = 1/Q_{ii}$,

$$\text{E}(x_i \mid \boldsymbol{x}_{-i}) = \mu_i - \frac{1}{Q_{ii}} \sum_{j \sim i} Q_{ij}(x_j - \mu_j), \quad \text{and} \quad \text{Corr}(x_i, x_j \mid \boldsymbol{x}_{\{-ij\}}) = -\frac{Q_{ij}}{\sqrt{Q_{ii} Q_{jj}}}.$$

These expressions provide an interpretation of the elements of $\boldsymbol{Q}$ based on conditional moments. This can be compared to the interpretation of the elements of $\boldsymbol{\Sigma}$, which are based on marginal univariate and bivariate moments.

2.2 Efficient computations on a GMRF

If the graph $\mathcal{G}$ on which the GMRF is defined is fully connected, the precision matrix will be a full matrix. However, we will focus on situations where $\boldsymbol{Q}$ is sparse, which are the ones leading to computational savings. In most cases, only $\mathcal{O}(n)$ terms in $\boldsymbol{Q}$ are non-zero while $\mathcal{O}(n^2)$ terms are zero. We will explain why sparse matrices allow for fast computations, and how we in most cases can improve the computational efficiency by reordering the indices, i.e. by finding a permutation matrix $\boldsymbol{P}$, such that $\boldsymbol{Q}^P = \boldsymbol{P}\boldsymbol{Q}\boldsymbol{P}^T$ is faster to factorise than $\boldsymbol{Q}$. The basic operations on $\boldsymbol{Q}$ are to compute the Cholesky factorisation $\boldsymbol{Q} = \boldsymbol{L}\boldsymbol{L}^T$, where $\boldsymbol{L}$ is the (lower) Cholesky triangle, and to solve the linear systems $\boldsymbol{L}\boldsymbol{v} = \boldsymbol{b}$ and $\boldsymbol{L}^T\boldsymbol{\mu} = \boldsymbol{v}$, which is also the way to find the solution of $\boldsymbol{Q}\boldsymbol{\mu} = \boldsymbol{b}$.

After discussing some computational aspects of these matrix computations, we describe how unconditional and conditional sampling from a GMRF $\boldsymbol{x} \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{Q}^{-1})$, as well as evaluation of the corresponding log-densities, can be formulated in terms of these basic operations. The algorithms that are described are all implemented in the open source `C`-library `GMRFLib` (Rue and Follestad, 2002).

2.2.1 Some basic results

The algorithm for sampling $\boldsymbol{x} \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{Q}^{-1})$ consists of three steps. First, compute the Cholesky factorisation $\boldsymbol{Q} = \boldsymbol{L}\boldsymbol{L}^T$, then solve $\boldsymbol{L}^T\boldsymbol{v} = \boldsymbol{z}$ where $\boldsymbol{z} \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{I})$, and finally let $\boldsymbol{x} = \boldsymbol{v} + \boldsymbol{\mu}$. We will return later to the computational costs involved as this depends among other factors on the graph $\mathcal{G}$, but in the case where $\boldsymbol{Q}$ is a full matrix, computing the Cholesky factorisation costs $n^3/3$ flops. Note that computing $\boldsymbol{A} = \boldsymbol{Q}^{-1}\boldsymbol{B}$ for a $n \times p$ matrix $\boldsymbol{B}$, is done as follows. Solve $\boldsymbol{L}\boldsymbol{C} = \boldsymbol{B}$ for each of the p columns of $\boldsymbol{B}$, then solve $\boldsymbol{L}^T\boldsymbol{A} = \boldsymbol{C}$ for each of the p columns of $\boldsymbol{C}$. There is no need to compute the inverse of $\boldsymbol{Q}$.

2.2.2 Operations on sparse precision matrices

We will now explain why we can construct efficient algorithms for the factorisation of sparse matrices, and which considerations that lie behind such algorithms (Dongarra and Duff, 1998).

The main idea is to take advantage of the fact that when $\boldsymbol{Q}$ is sparse, then its Cholesky triangle $\boldsymbol{L}$ inherits the non-zero pattern from $\boldsymbol{Q}$, so if $Q_{ij} \neq 0$ then $L_{ij} \neq 0$, $i \geq j$ (shown below). In addition, some other terms of $\boldsymbol{L}$ can be non-zero, and these are called *fillins*. The positions of these additional non-zero terms can be determined from $\mathcal{G}$, such that we can *compute and store only* the non-zero terms of $\boldsymbol{L}$. If the number of fillins, n_{fillin} is small, then it is fast to compute $\boldsymbol{L}$, and if it is large, then the computation will be slower.

We now describe how to identify which elements of $\boldsymbol{L}$ that are non-zero, using a statistical approach. Let $\boldsymbol{x} \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{Q}^{-1})$. Then we know that if $\boldsymbol{L}^T\boldsymbol{x} = \boldsymbol{z}$, where $\boldsymbol{z} \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{I})$, then $\boldsymbol{x} + \boldsymbol{\mu}$ has the correct distribution. Writing this out, we obtain (for $i = n, \ldots, 1$)

$$\mathrm{E}(x_i \mid \boldsymbol{x}_{(i+1):n}) = \mu_i - \frac{1}{L_{ii}} \sum_{j=i+1}^{n} L_{ji}(x_j - \mu_j) \quad \text{and} \quad \mathrm{Var}(x_i \mid \boldsymbol{x}_{(i+1):n}) = 1/L_{ii}^2. \tag{1}$$

This provides an interpretation of the elements of $\boldsymbol{L}$ from the conditional expectation and variance of x_i, conditioned on the subsequent components $\boldsymbol{x}_{i+1,\ldots,n}$. Define the set $F(i,j) = \{i+1, \ldots, j-$

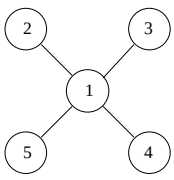
$1, j+1, \ldots, n\}$, then it follows from the global Markov property and (1) that

$$F(i,j) \text{ separates } i \text{ and } j \;\Leftrightarrow\; x_i \perp x_j \mid \boldsymbol{x}_{F(i,j)} \;\Leftrightarrow\; L_{ji} = 0. \tag{2}$$

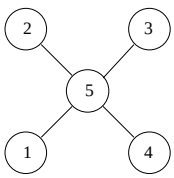
A consequence of this result is that if $i \sim j$ and $j > i$, then $L_{ji} \neq 0$. Using (2) we can determine from $\mathcal{G}$ which elements of $\boldsymbol{L}$ that are zero, and these need not to be computed.

A simple example is provided in Figure 1, for the graph of $\boldsymbol{x}' = (\mu, \boldsymbol{y}^T)^T$ (top) and $\boldsymbol{x}'' = (\boldsymbol{y}^T, \mu)^T$ (bottom), where $\boldsymbol{y} \sim \mathcal{N}(\mu\boldsymbol{1}, \boldsymbol{I})$ and $\mu \sim \mathcal{N}(0,1)$. Node 1 corresponds to μ, and the remaining nodes to $y_1, \ldots, y_n$, with $n = 4$, in increasing order. Each row shows the graph, the precision matrix and the Cholesky triangle. We see that $\boldsymbol{x}'$ makes $\boldsymbol{L}$ a full matrix (maximal number of fillins), as all elements of $\boldsymbol{y}$ depend on μ and thus none of the statements in (2) is true. The bottom row displays the case $\boldsymbol{x}''$, which don't produce any fillins. This example demonstrates that the ordering of the vertices is important for the degree of fillin. Therefore, it is common to permute $\boldsymbol{Q}$ before computing the Cholesky factorisation by choosing a permutation matrix $\boldsymbol{P}$ producing few fillins, and factorise $\boldsymbol{Q}^P = \boldsymbol{P}\boldsymbol{Q}\boldsymbol{P}^T$ instead. All equations are then solved in this permuted indices world, and then mapped back to the original indices when done. There are $n!$ possible permutations, so computing the best is not possible in general. Therefore, heuristic algorithms are used to produce, hopefully, good permutations with little fillin.

The example in Figure 1 is a special case of the *nested dissection* approach for reordering. Such schemes give quite generally few fillins, and the approach goes as follows. First, select a set of nodes dividing the graph into two disconnected subgraphs of almost equal size, next order the subsets such that the separating set has the highest indices, and then recursively repeat this division for each subgraph. More classical reordering schemes make $\boldsymbol{Q}^P$ have all the elements along the diagonal (Rue, 2001). This makes band-matrix algorithms useful. These are easy to code and runs very efficiently for long and thin graphs.



$$\boldsymbol{Q} = \begin{pmatrix} \times & \times & \times & \times & \times \\ \times & \times & & & \\ \times & & \times & & \\ \times & & & \times & \\ \times & & & & \times \end{pmatrix} \qquad \boldsymbol{L} = \begin{pmatrix} \times & & & & \\ \times & \times & & & \\ \times & \times & \times & & \\ \times & \times & \times & \times & \\ \times & \times & \times & \times & \times \end{pmatrix}$$



$$\boldsymbol{Q} = \begin{pmatrix} \times & & & & \times \\ & \times & & & \times \\ & & \times & & \times \\ & & & \times & \times \\ \times & \times & \times & \times & \times \end{pmatrix} \qquad \boldsymbol{L} = \begin{pmatrix} \times & & & & \\ & \times & & & \\ & & \times & & \\ & & & \times & \\ \times & \times & \times & \times & \times \end{pmatrix}$$

Figure 1: Effect of reordering the nodes of the graph on the fillin of the Cholesky factor $\boldsymbol{L}$ of $\boldsymbol{Q}$. The precision matrix and Cholesky factor are shown for the original graph (top) and the graph after swapping nodes 1 and 5 (bottom).

In the applications of Section 3 we use data from the 544 districts of Germany. Both the band reordering and the nested dissection reordering are illustrated in Figure 2, where the nodes of the graph are the 544 districts, defining two districts as neighbours if they share a common boundary. The left panels display the ordering of the nodes after applying the reordering schemes, and the middle and right panels give illustrations of the non-zero pattern of the precision matrix and the Cholesky triangle after

the reordering. The band-reordering scheme (Lewis, 1982) orders the districts row-wise, such that one row will make the south and north conditional independent. Hence, we obtain a band-matrix with the "row-width" as the bandwidth. The nested dissection reordering (Karypis and Kumar, 1998) splits the region into four sub-regions, but these sub-regions are so small that there is no gain in continuing the process.

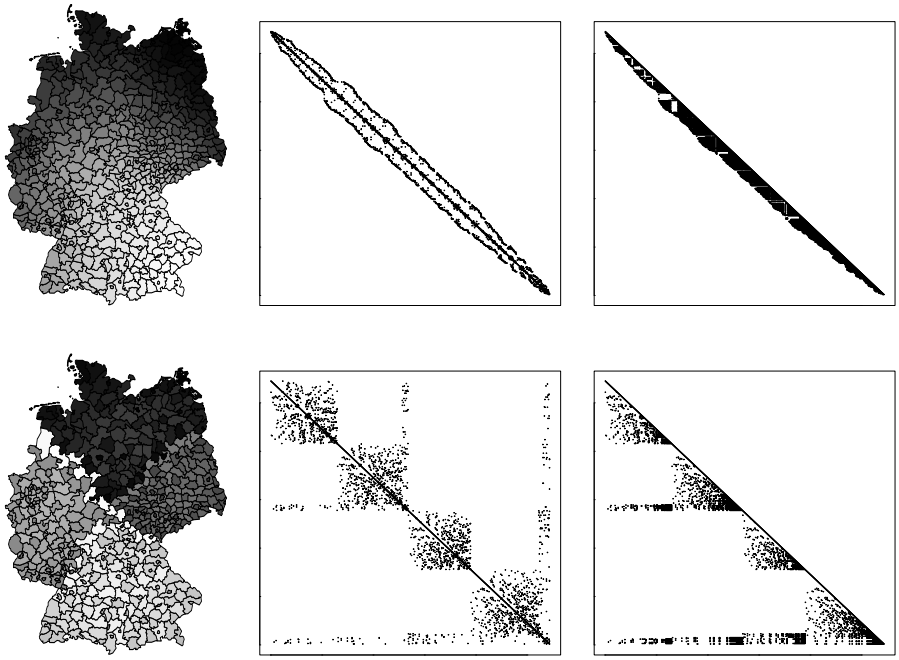


Figure 2: Illustration of the band-matrix reordering and the nested dissection reordering for the graph of Germany used in the applications of Section 3. The graph has 544 nodes each representing a disctrict, and two districts are neighbours if they share a common boundary. The left panels display the ordering of the nodes after applying the reordering schemes, and the middle and right panels give illustrations of the non-zero pattern of the precision matrix and the Cholesky triangle after reordering. The top row displays the band-matrix reordering, and the bottom row the nested dissection reordering. The ratio of non-zero terms in $\boldsymbol{L}$ and in the lower triangular part of $\boldsymbol{Q}$, is 2.5 for the nested dissection reordering and 5.2 for the band-reordering.

The nested dissection reordering is a useful tool for GMRFs where a relatively small number of nodes depends on (near) all other nodes. To exemplify this, consider a GMRF $\boldsymbol{y}$ with zero mean and a polynomial mean surface $\mu(i) = \boldsymbol{a}^T \boldsymbol{g}(i)$ which is linear in some basis functions depending on location i, with Gaussian priors on the coeffi cients $\boldsymbol{a}$. Then $\boldsymbol{x} = \boldsymbol{y} + \boldsymbol{\mu}$ is a GMRF with this property. Using band-reordering will make the bandwidth large, hence the factorisation will be slow. Using nested dissection reordering will give the global nodes a high index, as can be seen by comparing with Figure 1, such that the extra cost for including the global nodes is negligible. We will use this property in Section 3.1, where we estimate the parameters of a semi-parametric ecological regression model. In that application, the global parameter vector represents the effect of a the area-level covariate on the disease risk.

The computational complexity of factorising a precision matrix depends on the graph $\mathcal{G}$. Roughly, the GMRFs we consider are defi ned either in time or on a regular or irregular lattice in space, which may

be extended to include time as well. In time, the factorisation is $\mathcal{O}(n)$, it is $\mathcal{O}(n^{3/2})$ in space, and $\mathcal{O}(n^2)$ for space-time. This should be compared to $\mathcal{O}(n^3)$ for a full matrix. The band reordering is to be preferred for problems with long and thin graphs, otherwise we prefer to use the nested dissection reordering. For a spatial problem band reordering requires $\mathcal{O}(n^2)$ flops, compared to $\mathcal{O}(n^{3/2})$ for the nested dissection. The difference is minor for GMRFs of medium size, but is signifi cant for huge GMRF, from 30 000 nodes and upward.

It is worth mentioning that libraries for factorising sparse matrices are extremely complex and complicated software, at least compared to statistical standards. They easily require $10 - 100\,000$ lines of code, and specialist knowledge is needed to prevent loss of performance due to indirect addressing and so on. The band matrix approach is however quite simple and needs a small piece of code in comparison, and is already a part of standard libraries for numerical linear algebra. Gupta (2002) concludes in his recent comparison of such software, that

> *In this paper, we show that recent sparse solvers have significantly improved the state of the art of the direct solution of general sparse systems. ... Therefore, it would be fair to conclude that recent years have seen some remarkable advances in the general sparse direct solver algorithms and software.*

This is good news for statisticians, meaning that we can take advantage of this technology by using their software libraries for sparse matrices. However, we need to express our problems in that framework and to extend our horizon of what is possible and feasible to do, and these observations reflect the main message in our paper.

2.2.3 Unconditional and conditional sampling from a GMRF

We have previously shown that a sample $\boldsymbol{x} \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{Q}^{-1})$ can be generated by factorising $\boldsymbol{Q} = \boldsymbol{L}\boldsymbol{L}^T$, sampling $\boldsymbol{z} \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{I})$, solving $\boldsymbol{L}^T\boldsymbol{\nu} = \boldsymbol{z}$ for $\boldsymbol{\nu}$ by back-substitution and then adding the mean, such that $\boldsymbol{x} = \boldsymbol{\mu} + \boldsymbol{z}$. We will now review how to produce conditional samples effi ciently. Conditioning on a soft constraint, to be defi ned below, was only mentioned briefly as a comment in Rue (2001), but is here described in detail since it will be applied in Section 3.2.

Let $\mathcal{A}$ denote a subset of the nodes of the graph $\mathcal{G}$ and let $\mathcal{B} = -\mathcal{A}$ be the remaining nodes. Then $\pi(\boldsymbol{x}_\mathcal{A} \mid \boldsymbol{x}_\mathcal{B}) \sim \mathcal{N}(-\boldsymbol{Q}_{\mathcal{A}\mathcal{A}}^{-1}\boldsymbol{Q}_{\mathcal{A}\mathcal{B}}\boldsymbol{x}_\mathcal{B}, \boldsymbol{Q}_{\mathcal{A}\mathcal{A}}^{-1})$. This is a convenient result, as the precision matrix equals $\boldsymbol{Q}_{\mathcal{A}\mathcal{A}}$, which is a sparse sub-matrix of $\boldsymbol{Q}$, and $\boldsymbol{Q}_{\mathcal{A}\mathcal{B}}$ is non-zero only for those elements (i, j), with $i \in \mathcal{A}$ and $j \in \mathcal{B}$, for which $j \sim i$. Further, the conditional mean is found by solving a linear system involving these sparse matrices. As a consequence, $\boldsymbol{x}_\mathcal{A} \mid \boldsymbol{x}_\mathcal{B}$ is a GMRF as well, defi ned on the graph $\mathcal{G}$ restricted to $\mathcal{A}$.

We now consider the problem of sampling from a GMRF $\boldsymbol{x}$ under the linear constraint $\boldsymbol{A}\boldsymbol{x} = \boldsymbol{e}$, where $\boldsymbol{A}$ is a $p \times n$ matrix and $\boldsymbol{e}$ is a vector of length p. This conditional distribution is Gaussian as well, but does not have full rank and does not have any Markov properties in general. There is an alternative to sampling directly from this distribution, and this method is often referred to as conditional simulation using kriging (Cressie, 1993, Section 3.6.2). We fi rst generate an unconstrained sample $\boldsymbol{x}^u \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{Q}^{-1})$, and then add a correction term to produce the constrained sample $\boldsymbol{x}^c$ by

$$\boldsymbol{x}^c = \boldsymbol{x}^u - \boldsymbol{Q}^{-1}\boldsymbol{A}^T(\boldsymbol{A}\boldsymbol{Q}^{-1}\boldsymbol{A}^T)^{-1}(\boldsymbol{A}\boldsymbol{x}^u - \boldsymbol{e}). \tag{3}$$

All computations required to evaluate (3) can take advantage of the Cholesky factorisation of $\boldsymbol{Q}$. To compute $\boldsymbol{Z} = \boldsymbol{Q}^{-1}\boldsymbol{A}^T$, we first solve $\boldsymbol{L}\boldsymbol{Y} = \boldsymbol{A}^T$ by forward-substitution and then $\boldsymbol{L}^T\boldsymbol{Z} = \boldsymbol{Y}$ by back-substitution. We also note that since $\boldsymbol{A}\boldsymbol{Q}^{-1}\boldsymbol{A}^T$ is a $p \times p$ matrix, the cost of its factorisation is $\mathcal{O}(p^3)$, which is negligible for $p \ll n$.

The linear constraint $\boldsymbol{A}\boldsymbol{x} = \boldsymbol{e}$ will be denoted a *hard* constraint. A generalisation of this situation is the case when we instead of observing $\boldsymbol{A}\boldsymbol{x}$ have observed a value $\boldsymbol{e}_0$ of the Gaussian variable $\boldsymbol{e} \sim \mathcal{N}(\boldsymbol{A}\boldsymbol{x}, \boldsymbol{\Sigma}_{\boldsymbol{\epsilon}})$. We can extend (3) to cover the case of sampling from $\pi(\boldsymbol{x} \mid \boldsymbol{e} = \boldsymbol{e}_0)$. The conditional log-density is

$$\log \pi(\boldsymbol{x} \mid \boldsymbol{e} = \boldsymbol{e}_0) = -\frac{1}{2}(\boldsymbol{x} - \boldsymbol{\mu})^T\boldsymbol{Q}(\boldsymbol{x} - \boldsymbol{\mu}) - \frac{1}{2}(\boldsymbol{e}_0 - \boldsymbol{A}\boldsymbol{x})^T\boldsymbol{\Sigma}_{\boldsymbol{\epsilon}}^{-1}(\boldsymbol{e}_0 - \boldsymbol{A}\boldsymbol{x}) + \text{const.} \quad (4)$$

We write the stochastic variable $\boldsymbol{e}$ as $\boldsymbol{e} = \boldsymbol{A}\boldsymbol{x} + \boldsymbol{\epsilon}'$ with $\boldsymbol{\epsilon}' \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{\Sigma}_{\boldsymbol{\epsilon}})$, and observe that conditioning on $\boldsymbol{e} = \boldsymbol{e}_0$ in (4) is equivalent to conditioning on $\boldsymbol{A}\boldsymbol{x} = \boldsymbol{\epsilon}$, where $\boldsymbol{\epsilon} = \boldsymbol{e}_0 - \boldsymbol{\epsilon}'$ and thus $\boldsymbol{\epsilon} \sim \mathcal{N}(\boldsymbol{e}_0, \boldsymbol{\Sigma}_{\boldsymbol{\epsilon}})$. Consequently, we can reformulate the problem of sampling from $\pi(\boldsymbol{x} \mid \boldsymbol{e})$ given by (4) to the problem of sampling from $\pi(\boldsymbol{x} \mid \boldsymbol{A}\boldsymbol{x} = \boldsymbol{\epsilon})$, where $\boldsymbol{\epsilon}$ a realisation from $\boldsymbol{\epsilon} \sim \mathcal{N}(\boldsymbol{e}_0, \boldsymbol{\Sigma}_{\boldsymbol{\epsilon}})$. This is similar to the problem of sampling under a hard linear constraint described above, but replacing the fixed vector $\boldsymbol{e}$ by the stochastic variable $\boldsymbol{\epsilon}$. We denote the constraint $\boldsymbol{A}\boldsymbol{x} = \boldsymbol{\epsilon}$ a *soft* linear constraint. The expression (3) can now be modified to

$$\boldsymbol{x}^c = \boldsymbol{x}^u - \boldsymbol{Q}^{-1}\boldsymbol{A}^T(\boldsymbol{A}\boldsymbol{Q}^{-1}\boldsymbol{A}^T + \boldsymbol{\Sigma}_{\boldsymbol{\epsilon}})^{-1}(\boldsymbol{A}\boldsymbol{x}^u - \boldsymbol{\epsilon}). \quad (5)$$

Generating a sample under a soft constraint can be done by first generating an unconstrained sample $\boldsymbol{x}^u$ and a sample $\boldsymbol{\epsilon} \sim N(\boldsymbol{e}_0, \boldsymbol{\Sigma}_{\boldsymbol{\epsilon}})$, and then computing the softly constrained sample $\boldsymbol{x}^c$ from (5). As for the hard constraint, the cost of factorising the $p \times p$ matrix $\boldsymbol{A}\boldsymbol{Q}^{-1}\boldsymbol{A}^T + \boldsymbol{\Sigma}_{\boldsymbol{\epsilon}}$ is $\mathcal{O}(p^3)$. All remaining operations needed in (5) is performed by making use of the sparse matrix $\boldsymbol{Q}$. We will make use of (5) when we construct block-MCMC algorithms in Section 3.2.

2.2.4 Evaluation of the log-density

We now describe how the log-densities of the unconditional and conditional distributions described above can be evaluated at negligible cost when the factorisation $\boldsymbol{Q} = \boldsymbol{L}\boldsymbol{L}^T$ is available. To evaluate the log-density of an unconditional sample $\boldsymbol{x} \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{Q}^{-1})$, we need the terms $q = (\boldsymbol{x} - \boldsymbol{\mu})^T\boldsymbol{Q}(\boldsymbol{x} - \boldsymbol{\mu})$ and $\log|\boldsymbol{Q}|$, which are evaluated by computing $\boldsymbol{y} = \boldsymbol{Q}(\boldsymbol{x} - \boldsymbol{\mu})$ and then $q = (\boldsymbol{x} - \boldsymbol{\mu})^T\boldsymbol{y}$ and by $\log|\boldsymbol{Q}| = \log(|\boldsymbol{L}||\boldsymbol{L}^T|) = 2\sum_{i=1}^n \log L_{ii}$, respectively. The log-density of a conditional sample $\boldsymbol{x}_{\mathcal{A}} \mid \boldsymbol{x}_{\mathcal{B}}$ is found similarly. The log-density of a sample generated under a hard linear constraint $\boldsymbol{A}\boldsymbol{x} = \boldsymbol{e}$ can be evaluated using the identity

$$\pi(\boldsymbol{x} \mid \boldsymbol{A}\boldsymbol{x}) = \frac{\pi(\boldsymbol{x})\pi(\boldsymbol{A}\boldsymbol{x} \mid \boldsymbol{x})}{\pi(\boldsymbol{A}\boldsymbol{x})}. \quad (6)$$

Note that all quantities on the right hand side can be computed efficiently, see Rue (2001) for details. When we condition on a soft linear constraint, we make use of the same identity replacing $\boldsymbol{A}\boldsymbol{x}$ by the stochastic variable $\boldsymbol{\epsilon} \sim \mathcal{N}(\boldsymbol{e}_0, \boldsymbol{\Sigma}_{\boldsymbol{\epsilon}})$.

2.3 Using GMRFs as proxies for GRFs

Gaussian random fields (GRFs) are frequently used models in spatial statistics. The spatial structure of a GRF is most often specified in terms of a correlation function, where the exponential, Gaussian,

Matérn and spherical functions are among the most commonly used (Cressie, 1993). To do computations, the Gaussian field is typically discretised on a regular lattice, but there is no direct link to GMRFs in this case, as the corresponding precision matrix is full. Rue and Tjelmeland (2002) investigate the possibilities for using GMRFs as proxies for Gaussian fields, by fitting the elements of the precision matrix of a GMRF with a small neighbourhood to the corresponding elements of the precision matrix computed from the correlation functions mentioned above. Strikingly, their results show that all these correlation functions can be well approximated by a GMRF. The maximum difference in the correlation functions of the two models is less than about 0.05 when a GMRF with neighbours in a 5×5 neighbourhood around each node is used.

The result of Rue and Tjelmeland (2002) can be used to specify a model by the intuitively easier GRF formulation, and at the same time utilise the computational advantages of the GMRF. This approach is taken in our second application in Section 3.2, where we specify a lattice based GRF model for a smoothly varying risk surface based on aggregated count data.

2.4 A result for block-sampling in hidden Markov random field models

In many applications involving GMRFs, the distribution of the GMRF $\boldsymbol{x}$ is controlled by a few hyperparameters $\boldsymbol{\theta}$, and a subset $\boldsymbol{x}_{\mathcal{I}}$ of elements of $\boldsymbol{x}$ are observed by noisy observations $\boldsymbol{y}_{\mathcal{I}}$, assumed to be conditionally independent. This is the structure for the examples considered in Section 3. The joint posterior distribution of $\boldsymbol{x}$ and $\boldsymbol{\theta}$ is

$$\pi(\boldsymbol{\theta}, \boldsymbol{x} \mid \boldsymbol{y}) \propto \pi(\boldsymbol{\theta})\pi(\boldsymbol{x} \mid \boldsymbol{\theta}) \prod_{i \in \mathcal{I}} \pi(y_i \mid x_i). \tag{7}$$

A traditional MCMC algorithm for inference in this case is a single-site scheme, updating each θ_j and each x_i one at the time. This is feasible as the distribution of $x_i \mid \boldsymbol{x}_{-i}$ only depends on its neighbours. Using the fast algorithms for GMRFs, we can construct improved algorithms using a GMRF approximation to $\pi(\boldsymbol{x} \mid \boldsymbol{\theta}, \boldsymbol{y})$, computed as follows. Locate the mode $\boldsymbol{x}^*$ (which is a function of $\boldsymbol{\theta}$), expand the likelihood around $\boldsymbol{x}^*$ to second order and use this Gaussian approximation $\pi^*(\boldsymbol{x} \mid \boldsymbol{\theta}, \boldsymbol{y})$ as a proposal distribution for $\boldsymbol{x}$. Note that this approximation is a GMRF on the original graph $\mathcal{G}$, as the introduction of the likelihood only changes the diagonal terms in $\boldsymbol{Q}$ and the mean. Knorr-Held and Rue (2002) found empirically that such algorithms were feasible but seemed not to improve the convergence of $\boldsymbol{\theta}$. The problem is related to the strong interaction between $\boldsymbol{\theta}$ and the sufficient statistics for $\boldsymbol{\theta}$ based on $\pi(\boldsymbol{x} \mid \boldsymbol{\theta}, \boldsymbol{y})$. Knorr-Held and Rue (2002) propose to update $(\boldsymbol{\theta}, \boldsymbol{x})$ jointly, using the following general scheme: Sample $\boldsymbol{\theta'} \sim q(\boldsymbol{\theta'} \mid \boldsymbol{\theta})$ and $\boldsymbol{x'} \sim \pi^*(\boldsymbol{x} \mid \boldsymbol{\theta'}, \boldsymbol{y})$, and then accept/reject $(\boldsymbol{\theta'}, \boldsymbol{x'})$ jointly. The proposal for $\boldsymbol{\theta}$ is kept simple, for example a (log-)random walk. This scheme gave superior performance and the convergence is similar to what would be expected if we were able to sample from $\pi(\boldsymbol{\theta} \mid \boldsymbol{y})$ directly.

The following result represents a theoretical justification for their findings, which explain why blocking $\boldsymbol{x}$ and $\boldsymbol{\theta}$ separately makes the convergence arbitrary slow for increasing n.

Theorem 1 *Let $\pi(\mu, \boldsymbol{x}) = \pi(\mu)\pi(\boldsymbol{x} \mid \mu)$, where $\mu \sim \mathcal{N}(0, \sigma_\mu^2)$ and $\boldsymbol{x} \mid \mu \sim \mathcal{N}(\mu \boldsymbol{1}, \boldsymbol{\Sigma})$ and where $\boldsymbol{x}$ is of dimension $n > 0$. Let $\mu^{(1)}, \mu^{(2)}, \mu^{(3)}, \ldots$ be the marginal chain from the two step Gibbs sampler started in equilibrium, sampling successively $\mu \sim \pi(\mu \mid \boldsymbol{x})$ and $\boldsymbol{x} \sim \pi(\boldsymbol{x} \mid \mu)$. The marginal chain of μ is then a Gaussian $AR(1)$-process,*

$$\mu^{(t)} = \phi \mu^{(t-1)} + \epsilon_t, \tag{8}$$

where

$$\phi = \left(1 + \frac{Var(\overline{\boldsymbol{x}} \mid \mu)}{Var(\mu)}\right)^{-1} = (1 + \frac{1}{n} \times \frac{\frac{1}{n}\mathbf{1}^T\boldsymbol{\Sigma}\mathbf{1}}{\sigma_\mu^2})^{-1} = 1 - \mathcal{O}(1/n) \tag{9}$$

and $\epsilon_t \sim \mathcal{N}(0, \sigma_\mu^2(1-\phi^2))$.

The proof is simple (and therefore omitted) after noting that $\overline{\boldsymbol{x}}$ is sufficient for μ, such that we can consider the chain $\mu^{(1)} \to \overline{\boldsymbol{x}}^{(1)} \to \mu^{(2)} \to \cdots$.

Theorem 1 implies that the correlation length, defined as the minimum distance between two samples with correlation less than 0.05, of the marginal chain for μ is $\mathcal{O}(n)$. The marginal chain for μ will converge arbitrarily slow, for increasing n, even when a block sampling algorithm is used. The reason is the obvious one, block updating leads to improvements within the block and not that much between the blocks, and consequently we need to update $(\mu, \boldsymbol{x})$ jointly to break the strong interaction between μ and $\overline{\boldsymbol{x}}$. Rue et al. (2003) discuss this topic further, as well as how to construct independence samplers and approximations going beyond the Gaussian.

3 Applications

We illustrate the algorithms described in Section 2 by two applications from spatial epidemiology. Disease maps, displaying the geographical variation of disease incidence or mortality rates across a region of interest can give useful input to the formulation of etiological hypotheses on a disease. Such maps are most often generated on the basis of count data aggregated in a set of m disjoint areas. For rare and non-infectious diseases, the incidence or mortality counts $y_i,\ i = 1, \ldots, m$ are commonly assumed to be conditionally independent and to follow Poisson distributions with mean given by $E_i R_i$. The value E_i represents the expected number of cases in area i, adjusted for population size and factors like age and gender, and R_i is the area-specific relative risk, to be estimated from the data. The maximum likelihood estimate of R_i is the standardised mortality ratio SMR $= y_i/E_i$. For areas with low populations the sampling variance of SMR is high, and in addition, any evidence of extra-Poisson variation or spatial structure in the data is not taken into account. A commonly used Bayesian approach for improving on these raw estimates, first proposed by Besag, York and Mollié (1991), is to specify a log-linear model for the relative risks including spatially structured as well as unstructured Gaussian random effects, where the spatially structured effect is assigned a GMRF (intrinsic) prior. Reviews of recent work adopting this approach are given by Wakefield, Best and Waller (2000) and Mollié (1996). The model can be extended to include area-level covariates. Natário and Knorr-Held (2002) propose a semi-parametric model for the covariate effect, an approach that leads to a model formulation that is more flexible than the commonly used parametric models, that assume a log-linear relationship between the risk and the covariates. In Section 3.1 we re-estimate the parameters of the model using a full MCMC block-sampler.

In general, spatial heterogeneity of the disease risk remaining after adjusting for observed covariates will be a confounder for unmeasured spatially structured risk factors. In most cases, the risk factors are not expected to be constant within areas and disjoint across area boundaries, which is implicitly assumed by the random effects models described above. On the contrary, it seems reasonable to believe that the underlying risk surface is varying continuously over the region of study. In our second application we describe an aggregation consistent approach for estimating a smooth risk surface based on aggregated data. The approach is based on the geostatistical model of Kelsall and Wakefield (2002),

but we show that by defining their GRF model on a lattice and using a GMRF as a proxy for the GRF, we can develop an efficient sampling based approach to inference avoiding the approximation of the prior distribution of the area-level risks, required by their approach.

3.1 Semi-parametric ecological regression

Our first application illustrate how to apply the block-sampling scheme suggested by Knorr-Held and Rue (2002) to avoid the potential problem with a correlation length of order $\mathcal{O}(n)$ discussed in Section 2.4. We reconsider the model specified by Natário and Knorr-Held (2002), which is an extension of the model of Besag et al. (1991) to allow for a semi-parametric function of covariates believed to influence the risk. They use data on mortality from larynx cancer among males in the 544 districts of Germany over the period $1986 - 1990$, with estimates for lung cancer mortality as a proxy for smoking consumption as a covariate. We refer to Natário and Knorr-Held (2002) for further details and background for this application.

The model is specified as follows. The larynx cancer mortality counts $y_i,\ i = 1, \ldots, m$ in the $m = 544$ districts are assumed to be conditionally independent and Poisson distributed with mean $E_i R_i$, where $\log R_i = \eta_i,\ i = 1, \ldots, m$ are the log-relative risks of the disease, and the E_is are known constants. The prior model for the log relative risk, $\boldsymbol{\eta}$, is

$$\begin{aligned} \eta_i \mid \ldots &\sim \mathcal{N}(s_i + f(c_i), \tau^{-1}), \\ \pi(\boldsymbol{s} \mid \kappa) &\propto \kappa^{(n-1)/2} \exp(-\frac{\kappa}{2} \sum_{i \sim j} (s_i - s_j)^2), \\ \pi(\{f_j\} \mid \gamma) &\propto \gamma^{(m-2)/2} \exp(-\frac{\gamma}{2} \sum_{j} (f_j - 2f_{j-1} + f_{j-2})^2). \end{aligned} \tag{10}$$

The precision parameters τ, κ and γ are assigned vague Gamma-priors. For each area i, η_i is the sum of a spatially structured component s_i, with an intrinsic autoregressive prior defining $i \sim j$ if areas i and j share a common border, and the effect of the covariate c_i which is $f(c_i)$. In addition, a spatially unstructured random effect with precision τ is included. The covariate function $f(\cdot)$ is a random smooth function with small squared second order differences. The function $f(\cdot)$ is defined to be piecewise linear between the function values $\{f_j\}$ at 100 equally distant values of c_i, chosen to reflect the range of the covariate. We further impose the constraint $\sum s_i = 0$ to separate out the effect of the covariate.

The posterior is of the form (7) with $\boldsymbol{x} = (\{\eta_i\}, \{s_i\}, \{f_j\})$, $\boldsymbol{\theta} = (\tau, \kappa, \gamma)$ and $\mathcal{I}$ being the indices of the elements in $\boldsymbol{x}$ corresponding to $\{\eta_i\}$, and with $\pi(y_i \mid x_i)$ as the Poisson density with mean $E_i \exp(\eta_i)$ evaluated in y_i. The graph and precision matrix of $\boldsymbol{x}$ are easily found from the posterior.

Our block MCMC algorithm goes as follows. First, for each component θ_i we propose independently a new value $\theta_i' = f\theta_i$, where $\pi(f) \propto 1 + 1/f$, in the range $[1/F, F]$ where $F > 1$. Then we sample $\boldsymbol{x'}$ from the GMRF approximation $\pi^*(\boldsymbol{x} \mid \boldsymbol{\theta'}, \boldsymbol{y})$ using (3) to correct for the constraint $\sum s_i = 0$, and then accept/reject jointly. The scaling proposal for θ_i is motivated from a log-random walk, but make the proposal-terms cancel in the acceptance-ratio. The choice $F = 2$ gave an acceptance-rate of about $1/3$.

The trace-plots of $\boldsymbol{\theta}$ and the estimated posterior means of the values of the semi-parametric function $f(\cdot)$ are illustrated in Figure 3. Note the fast convergence of $\boldsymbol{\theta}$, obtained despite the fact that the

dimension of the problem is about 1 200 with three hyper-parameters. Our implementation made about 6.5 iterations in one second on a 1 200MHz laptop, using the Multifrontal Supernodal Cholesky factorisation routine in the TAUCS-library (Toledo, Chen and Rotkin, 2002).

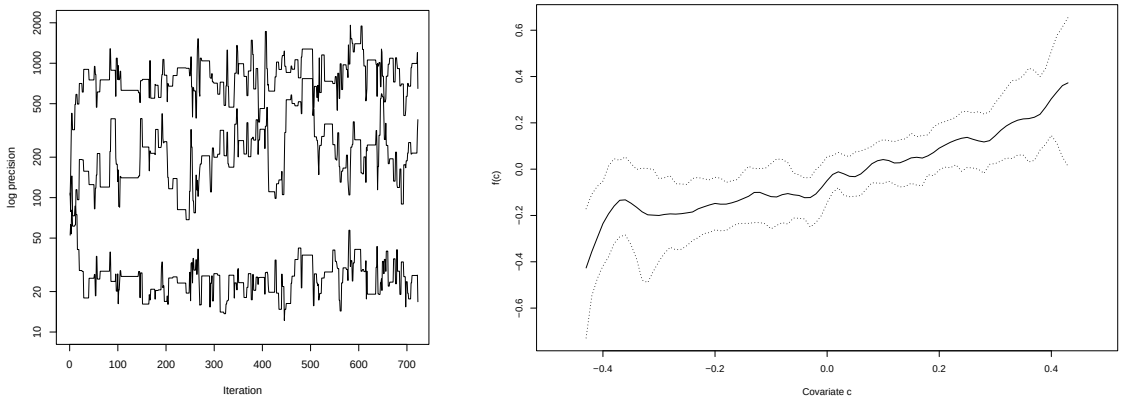


Figure 3: The trace-plots (left) of (τ, κ, γ), where τ is the lower curve, κ the middle curve and γ the upper curve. The estimated posterior means (full line) and empirical pointwise 2.5% and 97.5% quantiles (dotted lines) for the semi-parametric function $f(\cdot)$ (right).

3.2 Modelling spatial variation in the risk of a disease using GMRFs as proxies for GRFs

Our second example is the geostatistical model of Kelsall and Wakefi eld (2002). Instead of considering the risk to be constant within each area as in (10), they model the relative risk R_i in area i, denoted $\mathcal{A}_i$, as the integral of a continuous latent risk surface $R(\boldsymbol{s})$,

$$R_i = \int_{\mathcal{A}_i} R(\boldsymbol{s})\, p(\boldsymbol{s})\, d\boldsymbol{s}, \tag{11}$$

Here, $\boldsymbol{s}$ denotes spatial location and $p(\boldsymbol{s})$ the relative population density, which is assumed to be constant within each area. The motivation behind the model is that the relative risk is more likely to vary smoothly with respect to $\boldsymbol{s}$ than being constant within each region, and that such a model will be aggregation consistent. The log-risk $\log R(\cdot)$ is assumed to be a GRF with a non-zero mean and an exponential correlation function with unknown range r and precision τ. Since our data are standardised such that the overall risk for the region of study is 1, we assume the mean of the GRF to be zero. The likelihood model is as in Section 3.1, such that $y_i,\ i = 1, \ldots, 544$ are mutually independent Poisson variables with mean $E_i R_i$, for known constants $\{E_i\}$.

The approach taken by Kelsall and Wakefi eld is to approximate the joint distribution of $\{\log R_i, i =$

$1, \ldots, 544\}$ by a Gaussian distribution with second order moments

$$\begin{aligned} \mathrm{E}(\log R_i \log R_j) &= \mathrm{E}\left(\log \int_{\mathcal{A}_i} R(\boldsymbol{s})\, d\boldsymbol{s} \times \log \int_{\mathcal{A}_j} R(\boldsymbol{t})\, d\boldsymbol{t}\right) \\ &\approx \int_{A_i}\int_{A_j} \mathrm{E}\left(\log R(\boldsymbol{s}) \log R(\boldsymbol{t})\right)\, d\boldsymbol{s}\, d\boldsymbol{t}, \end{aligned} \tag{12}$$

which are computed numerically, and based on this approximated model, they construct single-site MCMC schemes for inference. The purpose of this example is to demonstrate that using a GMRF as a proxy for a GRF as discussed in Section 2.3, we can avoid approximation (12) *and* the Gaussian approximation to the joint density of $\{\log R_i\}$. Furthermore, we obtain an efficient block MCMC algorithm, making use of (5) to construct proposal distributions. We refer to the technical report Follestad and Rue (2003) for details and an extended discussion.

3.2.1 The model

We discretise the support of $R(\boldsymbol{s})$ and define the GRF model on a fine lattice covering the region of interest, consisting of the 544 districts of Germany plus a boundary region. The lattice contains $n = 31\,089$ pixels, and is shown in the left panel of Figure 4. We then apply the method for generating GMRF proxies for GRFs (Rue and Tjelmeland, 2002) using a 5×5 neighbourhood, and using discrete values for the range, defined in steps of size 0.05 in lattice coordinates. Let the fitted GMRF be denoted $\boldsymbol{x}$. For the exponential correlation function used by Kelsall and Wakefield, the maximum difference between the true and fitted correlation functions is about 0.01. Note that our methods is exact for the model defined using the GMRF proxy. The relative risk for region i is then

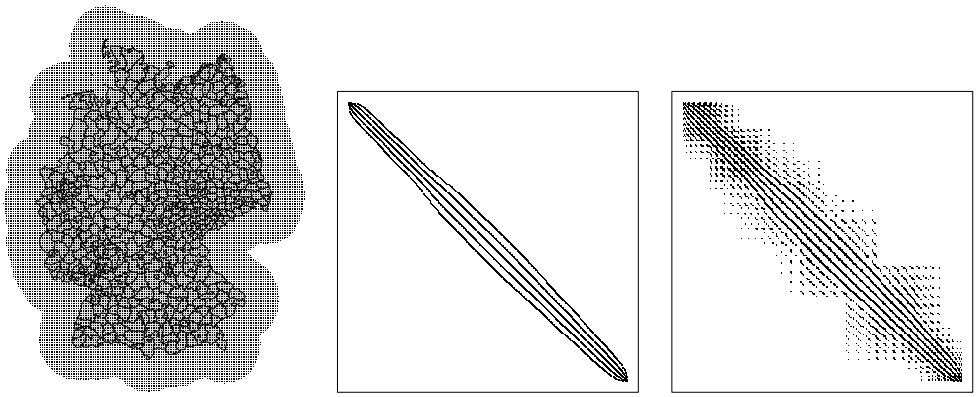
$$R_i = \frac{1}{n_i} \sum_{j:\, j \in \mathcal{A}_i} \exp(x_j), \tag{13}$$

summing over the n_i pixels j in region $\mathcal{A}_i$. We place vague priors on the discretised range and the precision of $\boldsymbol{x}$.

3.2.2 Constructing block MCMC algorithms using soft constraints

The posterior distribution of the log-risk surface $\boldsymbol{x}$ and the hyper-parameters $\boldsymbol{\theta} = (\tau, r)$ is of the form (7), but where y_i depends on $\boldsymbol{x}_{\mathcal{A}_i}$ through (13). This implies that the full conditional for x_j, $j \in \mathcal{A}_i$, depends on $\boldsymbol{x}_{\mathcal{A}_i}$ as well as the elements $x_j, j \notin \mathcal{A}_i$ for which j has a neighbouring node in $\mathcal{A}_i$. This is illustrated in Figure 4, illustrating the precision matrix for the GMRF $\boldsymbol{x}$ and a Gaussian approximation to the posterior $\boldsymbol{x}$ for a small set of regions, after using band reordering. This makes even single-site algorithms for $\boldsymbol{x}$ quite computationally expensive, as the neighbourhood in the posterior is large.

Due to Theorem 1 and the large number, $n = 31\,089$, of nodes, the construction of block MCMC algorithms is still desirable. We have not succeeded in constructing approximations that are accurate enough to allow for updating $\boldsymbol{x}$ and $\boldsymbol{\theta}$ jointly without near-zero acceptance-rates. We therefore take an intermediate approach. Let $\mathcal{A}$ be the interior region. We will update the block $\boldsymbol{x}_{\mathcal{A}}$ of interior elements keeping $\boldsymbol{\theta}$ fixed, and then $(\boldsymbol{\theta}, \boldsymbol{x}_{-\mathcal{A}})$ in one block using the scheme of Knorr-Held and Rue (2002). As $\boldsymbol{x}_{-\mathcal{A}} \mid \boldsymbol{\theta}, \boldsymbol{x}_{\mathcal{A}}$ is a GMRF, generating this joint update is straightforward. Note that this step nearly



(a) (b) (c)

Figure 4: The map of Germany with its 544 districts, overlaid by a lattice including a set of boundary nodes (a). The two right panels illustrate the conditional independence structure (after reordering) of the prior model (c) and when conditioning on the data (d), for a subset of the lattice nodes.

"integrate out" $\boldsymbol{x}_{-\mathcal{A}}$, such that the effective "n" in Theorem 1 is reduced to the number of pixels in the interior $\mathcal{A}$.

We update $\boldsymbol{x}_{\mathcal{A}}$ in sub-blocks conditioning on the remaining elements, and describe the algorithm for blocks of single regions, $\boldsymbol{x}_{\mathcal{A}_i}$. The extensions to sub-blocks of $\boldsymbol{x}_{\mathcal{A}}$ made up from several regions is similar. In practise, we update as many regions in one block as possible at the same time as obtaining reasonable acceptance-rates. We will make use of the soft constraint (4) to construct computational effi cient approximations.

The full conditional distribution of $\boldsymbol{x}_{\mathcal{A}_i}$, corresponding to the n_i nodes within area $\mathcal{A}_i$, is given by

$$\pi(\boldsymbol{x}_{\mathcal{A}_i} \mid \boldsymbol{x}_{-\mathcal{A}_i}, \boldsymbol{\theta}, \boldsymbol{y}) \propto \pi(\boldsymbol{x}_{\mathcal{A}_i} \mid \boldsymbol{x}_{-\mathcal{A}_i}, \boldsymbol{\theta}) \pi(y_i \mid \boldsymbol{x}_{\mathcal{A}_i}, \boldsymbol{\theta}). \tag{14}$$

As $\pi(\boldsymbol{x}_{\mathcal{A}_i} \mid \boldsymbol{x}_{-\mathcal{A}_i}, \boldsymbol{\theta})$ is a GMRF with precision matrix $\boldsymbol{Q}_{\mathcal{A}_i}$, given by the $n_i \times n_i$ diagonal block of $\boldsymbol{Q}$ corresponding to area $\mathcal{A}_i$, then (14) can be written as

$$\log(\pi(\boldsymbol{x}_{\mathcal{A}_i} \mid \boldsymbol{x}_{-\mathcal{A}_i}, \boldsymbol{\theta}, \boldsymbol{y})) \quad = \quad -\frac{1}{2} \boldsymbol{x}_{\mathcal{A}_i}^T \boldsymbol{Q}_{\mathcal{A}_i} \boldsymbol{x}_{\mathcal{A}_i} + \boldsymbol{d}_i^T \boldsymbol{x}_{\mathcal{A}_i} + h_i(\boldsymbol{x}) + \text{const}, \tag{15}$$

where $h(\boldsymbol{x}_i)$ is the log-likelihood of the observed count for area $\mathcal{A}_i$. The vector $\boldsymbol{d}_i$ and the matrix $\boldsymbol{Q}_{\mathcal{A}_i}$ both depend on $\boldsymbol{\theta}$, but we suppress this explicit reference for notational convenience. Due to the Poisson likelihood, the posterior distribution is non-standard. We specify a Metropolis-Hastings step, constructing a Gaussian approximation to (15) as a proposal for $\boldsymbol{x}_{\mathcal{A}_i}$. This is found by replacing the term $h_i(\boldsymbol{x})$ by a quadratic approximation around the (conditional) mode, $h_i(\boldsymbol{x}) \approx -\frac{1}{2} \boldsymbol{x}_{\mathcal{A}_i}^T \boldsymbol{B}_i \boldsymbol{x}_{\mathcal{A}_i} + \boldsymbol{b}_i^T \boldsymbol{x}_{\mathcal{A}_i}$, where $\boldsymbol{B}_i$ and $\boldsymbol{b}_i$ depend on y_i and $\boldsymbol{\theta}$. Substituting this approximation for $h_i(\boldsymbol{x})$ in (15), we obtain the Gaussian approximation $\pi_N(\boldsymbol{x}_{\mathcal{A}_i} \mid \boldsymbol{x}_{-\mathcal{A}_i}, \boldsymbol{y})$ to (15) as

$$\log(\pi_N(\boldsymbol{x}_{\mathcal{A}_i} \mid \boldsymbol{x}_{-\mathcal{A}_i}, \boldsymbol{\theta}, \boldsymbol{y})) \quad = \quad -\frac{1}{2} \boldsymbol{x}_{\mathcal{A}_i}^T (\boldsymbol{Q}_{\mathcal{A}_i} + \boldsymbol{B}_i) \boldsymbol{x}_{\mathcal{A}_i} + \boldsymbol{c}_i^T \boldsymbol{x}_{\mathcal{A}_i} + \text{const}. \tag{16}$$

Here, the precision matrix $\boldsymbol{Q}_{\mathcal{A}_i} + \boldsymbol{B}_i$ is a full matrix, such that the sparse structure is lost. However, if we use that $\boldsymbol{B}_i = \boldsymbol{D}_i + \boldsymbol{A}_i^T \boldsymbol{A}_i$ where $\boldsymbol{D}_i$ is a $n_i \times n_i$ diagonal matrix and $\boldsymbol{A}_i$ is a $1 \times n_i$ matrix,

we obtain

$$\log(\pi_N(\boldsymbol{x}_{\mathcal{A}_i} \mid \boldsymbol{x}_{-\mathcal{A}_i}, \boldsymbol{\theta}, \boldsymbol{y})) = -\frac{1}{2}\boldsymbol{x}_{\mathcal{A}_i}^T(\boldsymbol{Q}_{\mathcal{A}_i} + \boldsymbol{D}_i)\boldsymbol{x}_{\mathcal{A}_i} + \boldsymbol{c}_i^T\boldsymbol{x}_{\mathcal{A}_i} - \frac{1}{2}\boldsymbol{x}_{\mathcal{A}_i}^T\boldsymbol{A}_i^T\boldsymbol{A}_i\boldsymbol{x}_{\mathcal{A}_i} + \text{const.} \quad (17)$$

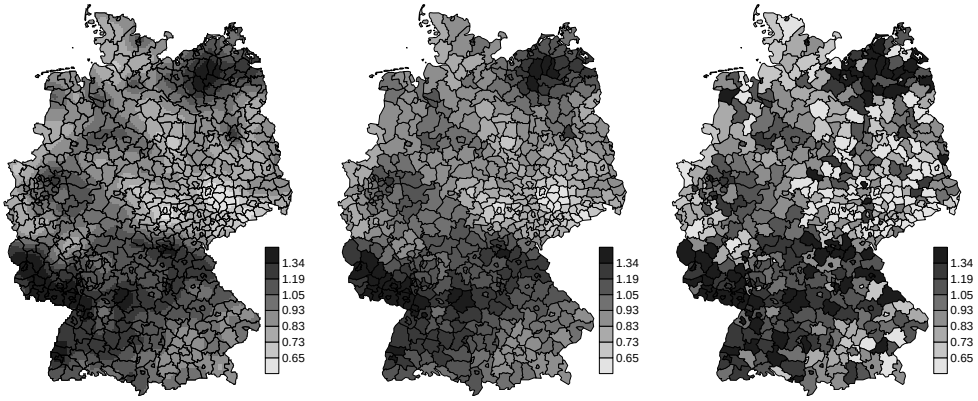
Comparing (17) to (4) in Section 2.2.3, we observe that by letting $\boldsymbol{Q} = \boldsymbol{Q}_{\mathcal{A}_i} + \boldsymbol{D}_i$, $\boldsymbol{A} = \boldsymbol{A}_i$, $\boldsymbol{\Sigma}_{\boldsymbol{\epsilon}} = \boldsymbol{I}$ and $\boldsymbol{e}_0 = \boldsymbol{0}$ in (4), we have re-formulated the problem of sampling from (16) to a constrained sampling problem, sampling from a Gaussian distribution $\boldsymbol{x}' \sim \mathcal{N}(\boldsymbol{Q}_{\mathcal{A}_i}^{-1}\boldsymbol{c}_i^T, \boldsymbol{Q}_{\mathcal{A}_i}^{-1})$ under the soft linear constraint $\boldsymbol{A}\boldsymbol{x}' = \boldsymbol{\epsilon};\ \boldsymbol{\epsilon} \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{I})$. Consequently, the efficient algorithms described in Section 2.2.3 can be applied to generate a sample from (17), where the computations make use of the non-zero pattern of the prior (and not the posterior), and this sample is used as a proposed value that is accepted or rejected in a Metropolis-Hastings step. This soft constraint approach represents large computational savings.

3.2.3 Results

We apply the method of Section 3.2.2 to a set of data on mortality from oral cavity cancer for males in Germany, observed over the period 1986-1990. The counts range from 1 to 501, with a median count of 19, and the expected number of cases $\{E_i\}$ range from 3.0 to 393.1, with a median of 19.5. The standardised mortality ratios (SMR) for the data are shown in Figure 5. Estimating the subset $\boldsymbol{x}_{\mathcal{A}}$ of $\boldsymbol{x}$ using blocks consisting of the nodes within one area and its neighbouring areas sharing a common boundary, leads to reasonable acceptance rates. To avoid boundary effects between blocks, the partition into blocks were updated randomly at every 10th step of the sampler. Studying trace plots for the MCMC updates (not shown) after running $250\,000$ iterations, we find that the convergence is fast for the elements of the log-risk surface, but that the mixing is still relatively poor for the hyper-parameters. This result is in accordance with Theorem 1. Despite the poor mixing of the individual parameters τ and r, the posterior means of the elements of $\boldsymbol{x}$ appear to be stable, and these estimates are summarised in Figure 5. We observe that the estimated spatial pattern of the risk is similar to the SMR, but the estimated risk surface is smoother. The estimated posterior means of the relative risks at the area level vary between 0.57 and 1.54. The results are similar to the ones obtained by Knorr-Held and Raßer (2000), who reported estimated posterior median relative risks in the range 0.65 and 1.42 using a Bayesian clustering approach, and between 0.56 and 1.56 using the model of Besag et al. (1991). More details on the results and the convergence properties of the block MCMC sampler are available in our technical report (Follestad and Rue, 2003).

4 Concluding remarks

Gaussian Markov Random Fields (GMRF) represent a very flexible class of models. The graph formulation allows for a unified formulation, representation and understanding of the models, and GMRFs have proved to be useful as a building block in many types of models, including temporal, spatial and geostatistical and spatio-temporal models. Computations on GMRFs have shown to be impressively fast using modern numerical methods for sparse matrices, and the same library of algorithms can be used for all GMRFs. The fast computations also make it possible to construct efficient block-MCMC algorithms for fast and more reliable inference.



Estimated risk surface Estimated $\{R_i\}$ SMR

Figure 5: Results for the German oral cavity cancer data.

5 Acknowledgements

The authors thank Leo Knorr-Held for useful comments and for providing the data.

References

Besag, J. (1974). Spatial interaction and the statistical analysis of lattice systems (with discussion), *J. Roy. Statist. Soc. Ser. B* **36**(2): 192–225.

Besag, J. and Higdon, D. (1999). Bayesian analysis of agricultural field experiments (with discussion), *J. Roy. Statist. Soc. Ser. B* **61**(4): 691–746.

Besag, J. and Kooperberg, C. (1995). On conditional and intrinsic autoregressions, *Biometrika* **82**(4): 733–746.

Besag, J., York, J. and Mollié, A. (1991). Bayesian image restoration with two applications in spatial statistics (with discussion), *Ann. Inst. Statist. Math.* **43**(1): 1–59.

Cressie, N. A. C. (1993). *Statistics for spatial data*, 2nd edn, John Wiley, New York.

Dongarra, J. and Duff, I. (1998). *Numerical linear algebra for high performance computers*, SIAM, Addison.

Fahrmeir, L. and Lang, S. (2001). Bayesian inference for generalized additive mixed models based on markov random field priors, *J. Roy. Statist. Soc. Ser. C* **50**: 201–220.

Follestad, T. and Rue, H. (2003). Modelling spatial variation in disease risk using Gaussian Markov random field proxies for Gaussian random fields, *Preprint Series in Statistics no. 3/2003*, Dept. of Mathematical Sciences, Norwegian University of Science and Technology, Trondheim, Norway.

Gamerman, D. (1998). Markov chain Monte Carlo for dynamic generalised liner models, *Biometrika* **85**(1): 215–227.

Gupta, A. (2002). Recent advances in direct methods for solving unsymmetric sparse systems of linear equations, *ACM Transactions on Mathematical Software (TOMS)* **28**(3): 301–324.

Karypis, G. and Kumar, V. (1998). METIS. A software backage for partitioning unstructured graphs, partitioning meshes, and computing fill-reducing orderings of sparse matrices. Version 4.0., *Manual*, University of Minnesota, Department of Computer Science/ Army HPC Research Center.

Kelsall, J. E. and Wakefield, J. C. (2002). Modeling spatial variation in disease risk: A geostatistical approach, *J. Amer. Statist. Assoc.* **97**(459): 692–701.

Knorr-Held, L. (1999). Conditional prior proposals in dynamic models, *Scand. J. Statist.* **26**(1): 129–144.

Knorr-Held, L. and Raßer, G. (2000). Bayesian detection of clusters and discontinuities in disease maps, *Biometrics* **56**(1): 13–21.

Knorr-Held, L. and Rue, H. (2002). On block updating in Markov random field models for disease mapping, *Scand. J. Statist.* **29**(4): 597–614.

Lang, S. and Bretzger, A. (2002). BayesX: Software for Bayesian inference based on Markov chain Monte Carlo simulation techniques. Version 0.9., *Technical report*, University of Munich.

Lewis, J. G. (1982). Algorithm 582: The Gibbs-Poole-Stockmeyer and Gibbs-King algorithms for reordering sparse matrices, *ACM Trans. Math. Softw.* **8**(2): 190–194.

Mollié, A. (1996). Bayesian mapping of disease, *in* W. R. Gilks, S. Richardson and D. J. Spiegelhalter (eds), *Markov Chain Monte Carlo in Practice*, Chapman & Hall, London, UK, pp. 359–379.

Natário, I. and Knorr-Held, L. (2002). Non-parametric ecological regression and spatial variation. Accepted for publication in *Biometrical Journal*.

Rue, H. (2001). Fast sampling of Gaussian Markov random fields, *J. Roy. Statist. Soc. Ser. B* **63**(2): 325–338.

Rue, H. and Follestad, T. (2002). GMRFLib: a C-library for fast and exact simulation of Gaussian Markov random fields, *Preprint series in statistics no. 1/2002*, Dept. of Mathematical Sciences, Norwegian University of Science and Technology, Trondheim, Norway.

Rue, H., Steinsland, I. and Erland, S. (2003). Approximating hidden Markov random fields, *Preprint Series in Statistics no. 1/2003*, Dept. of Mathematical Sciences, Norwegian University of Science and Technology, Trondheim, Norway.

Rue, H. and Tjelmeland, H. (2002). Fitting Gaussian Markov random fields to Gaussian fields, *Scand. J. Statist.* **29**(1): 31–50.

Shephard, N. and Pitt, M. K. (1997). Likelihood analysis of non-Gaussian measurement time series, *Biometrika* **84**(3): 653–667.

Toledo, S., Chen, D. and Rotkin, V. (2002). TAUCS. A library of sparse linear solvers. Version 2.0., *Manual*, School of Computer Science, Tel-Aviv University.

Wakefield, J. C., Best, N. G. and Waller, L. (2000). Bayesian approaches to disease mapping, *in* P. Elliott, J. C. Wakefield, N. G. Best and D. J. Briggs (eds), *Spatial Epidemiology. Methods and Applications*, Oxford University Press, New York, pp. 104–127.