Approximate Bayesian Inference for Hierarchical Gaussian Markov Random Fields Models

Håvard Rue and Sara Martino
Department of Mathematical Sciences
NTNU, Norway

September 20, 2005
Revised February 27, 2006

Abstract

Many commonly used models in statistics can be formulated as (Bayesian) hierarchical Gaussian Markov random field models. These are characterised by assuming a (often large) Gaussian Markov random field (GMRF) as the second stage in the hierarchical structure and a few hyperparameters at the third stage. Markov chain Monte Carlo is the common approach for Bayesian inference in such models. The variance of the Monte Carlo estimates is $\mathcal{O}_p(M^{-1/2})$ where M is the number of samples in the chain so, in order to obtain precise estimates of marginal densities, say, we need M to be very large.

Inspired by the fact that often one-block and independence samplers can be constructed for hierarchical GMRF models, we will in this work investigate whether MCMC is really needed to estimate marginal densities, which often is the goal of the analysis. By making use of GMRF-approximations, we show by typical examples that marginal densities can indeed be very precisely estimated by deterministic schemes. The methodological and practical consequence of these findings are indeed positive. We conjecture that for many hierarchical GMRF-models there is really no need for MCMC based inference to estimate marginal densities. Further, by making use of numerical methods for sparse matrices the computational costs of these deterministic schemes are nearly instant compared to the MCMC alternative. In particular, we discuss in detail the issue of computing marginal variances for GMRFs.

KEYWORDS: Approximate Bayesian inference, Cholesky triangle, Conditional auto-regressions, Gaussian Markov random fields, Hierarchical GMRF-models, Laplace-approximation, Marginal variances for GMRFs, Numerical methods for sparse matrices.

ADDRESS FOR CORRESPONDENCE: H. Rue, Department of Mathematical Sciences, The Norwegian University for Science and Technology, N-7491 Trondheim, Norway.

E-MAIL: hrue@math.ntnu.no and martino@math.ntnu.no

WWW-ADDRESS: http://www.math.ntnu.no/~hrue

ACKNOWLEDGEMENTS: The authors are grateful to the reviewers and the Editor for their helpful comments.

1 Introduction

A Gaussian Markov random field (GMRF) $\boldsymbol{x} = \{x_i : i \in \mathcal{V}\}$ is a $n = |\mathcal{V}|$-dimensional Gaussian random vector with additional conditional independence, or Markov properties. Assume that $\mathcal{V} = \{1, \ldots, n\}$. The conditional independence properties can be represented using an undirected graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ with vertices $\mathcal{V}$ and edges $\mathcal{E}$. Two nodes, x_i and x_j, are conditional independent given the remaining elements in $\boldsymbol{x}$, if and only if $\{i, j\} \notin \mathcal{E}$. Then, we say that $\boldsymbol{x}$ is a GMRF with respect to $\mathcal{G}$. The edges in $\mathcal{E}$ are in one-to-one correspondence with the non-zero elements of the precision matrix of $\boldsymbol{x}$, $\boldsymbol{Q}$, in the sense that $\{i, j\} \in \mathcal{E}$ if and only if $Q_{ij} \neq 0$ for $i \neq j$. When $\{i, j\} \in \mathcal{E}$ we say that i and j are neighbours, which we denote by $i \sim j$.

GMRFs are also known as conditional auto-regressions (CARs) following seminal work of Besag (1974, 1975). GMRFs (and their intrinsic versions) have a broad use in statistics, with important applications in structural time-series analysis, analysis of longitudinal and survival data, graphical models, semiparametric regression and splines, image analysis and spatial statistics. For references and examples, see Rue and Held (2005, Ch. 1).

One of the main areas of application for GMRFs is that of (Bayesian) hierarchical models. A hierarchical model is characterised by several stages of observables and parameters. The first stage, typically, consists of distributional assumptions for the observables conditionally on latent parameters. For example if we observe a time series of counts $\boldsymbol{y}$, we may assume, for y_i, $i \in \mathcal{D} \subset \mathcal{V}$ a Poisson distribution with unknown mean λ_i. Given the parameters of the observation model, we often assume the observations to be conditionally independent. The second stage consists of a prior model for the latent parameters λ_i or, more often, for a particular function of them. For example, in the Poisson case we can choose an exponential link and model the random variables $x_i = \log(\lambda_i)$. At this stage GMRFs provide a flexible tool to model the dependence between the latent parameters and thus, implicitly, the dependence between the observed data. This dependence can be of various kind, such as temporal, spatial, or even spatiotemporal. The third stage consists of prior distributions for the unknown hyperparameters $\boldsymbol{\theta}$. These are typically precision parameters in the GMRF. The posterior of interest is then

$$\pi(\boldsymbol{x}, \boldsymbol{\theta} \mid \boldsymbol{y}) \propto \pi(\boldsymbol{x} \mid \boldsymbol{\theta})\pi(\boldsymbol{\theta}) \prod_{i \in \mathcal{D}} \pi(y_i \mid x_i). \tag{1}$$

Most hierarchical GMRF-models can be written in this form. If there are unknown parameters also in the likelihood, then also the last term in (1) depends on $\boldsymbol{\theta}$. Such an extension makes only a slight notational difference in the following.

The main goal is often to compute posterior marginals, like

$$\pi(x_i \mid \boldsymbol{y}) = \int_{\boldsymbol{\theta}} \int_{\boldsymbol{x}_{-i}} \pi(\boldsymbol{x}, \boldsymbol{\theta} \mid \boldsymbol{y}) \, d\boldsymbol{x}_{-i} \, d\boldsymbol{\theta} \tag{2}$$

for each i and (sometimes also) posterior marginals for the hyperparameters θ_j. Since analytical integration is usually not possible for the posterior $\pi(\boldsymbol{x}, \boldsymbol{\theta}|\boldsymbol{y})$, it is common to use MCMC-based inference to estimate the posterior marginals. These marginals can then be used to compute marginal expectations of various statistics. Although single-site schemes, updating each element of $(\boldsymbol{x}, \boldsymbol{\theta})$ individually, are always possible, they may converge slowly due to the hierarchical structure of the problem. We refer to Rue and Held (2005, Ch. 4) for further discussion. (In some cases reparametrisation may solve the convergence problem due to the hierarchical structure (Gelfand et al., 1995; Papaspiliopoulos et al., 2003), but see also Wilkinson (2003).) In the case of disease mapping, Knorr-Held and Rue (2002) discuss various blocking strategies for updating all the unknown variables to improve the convergence, and Rue and Held (2005, Ch. 4) develop these ideas further. Even if using blocking strategies improves the convergence, MCMC techniques require a large number of samples to achieve a precise estimate. In this paper we propose a deterministic alternative to MCMC based inference

which has the advantage of being computed almost instant and which, in our examples, proves to be quite accurate. The key for fast computing time lies in the sparseness of the precision matrix $\boldsymbol{Q}$ due to the Markov properties in the GMRFs. This characteristic allows the use of efficient algorithms and, as explained in Section 2, makes it possible to compute marginal variances without the need to invert $\boldsymbol{Q}$.

One way to introduce our approximation technique is to start from the blocking strategies proposed in Knorr-Held and Rue (2002) and Rue and Held (2005, Ch. 4). The main idea behind these is to make use of the fact that the full conditional for the zero mean GMRF $\boldsymbol{x}$,

$$\pi(\boldsymbol{x} \mid \boldsymbol{\theta}, \boldsymbol{y}) \propto \exp\left(-\frac{1}{2}\boldsymbol{x}^T\boldsymbol{Q}\boldsymbol{x} + \sum_{i\in\mathcal{D}} \log \pi(y_i|x_i)\right) \tag{3}$$

can often be well approximated with a Gaussian distribution, by matching the mode and the curvature at the mode. The resulting approximation will then be

$$\widetilde{\pi}(\boldsymbol{x} \mid \boldsymbol{\theta}, \boldsymbol{y}) \propto \exp\left(-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^T(\boldsymbol{Q} + \mathrm{diag}(\boldsymbol{c}))(\boldsymbol{x}-\boldsymbol{\mu})\right) \tag{4}$$

where $\boldsymbol{\mu}$ is the mode of $\pi(\boldsymbol{x} \mid \boldsymbol{\theta}, \boldsymbol{y})$. Note that $\boldsymbol{\mu}$ and $\boldsymbol{Q}$ (and then (4)) depend on $\boldsymbol{\theta}$ but we suppress the dependence on $\boldsymbol{\theta}$ to simplify the notation. The terms of the vector $\boldsymbol{c}$ are due to the second order terms in the Taylor expansion of $\sum \log \pi(y_i|x_i)$ at the modal value $\boldsymbol{\mu}$, and these terms are zero for the nodes not directly observed through the data. We call the approximation in (4) the GMRF-approximation. The GMRF-approximation is also a GMRF with respect to the graph $\mathcal{G}$ since, by assumption, each y_i depends only on x_i, a fact that is important computationally.

Following Knorr-Held and Rue (2002) and Rue and Held (2005, Ch. 4), we can often construct a one-block sampler for $(\boldsymbol{x}, \boldsymbol{\theta})$, which proposes the new candidate $(\boldsymbol{x}', \boldsymbol{\theta}')$ by

$$\boldsymbol{\theta}' \sim q(\boldsymbol{\theta}' \mid \boldsymbol{\theta}), \quad \text{and} \quad \boldsymbol{x}' \sim \widetilde{\pi}(\boldsymbol{x} \mid \boldsymbol{\theta}', \boldsymbol{y}) \tag{5}$$

and then accept or reject $(\boldsymbol{x}', \boldsymbol{\theta}')$ jointly. This one-block algorithm, is made possible, in practise, by the outstanding computational properties of GMRFs through numerical algorithms for sparse matrices (Rue, 2001; Rue and Held, 2005). GMRFs of size up to 10^5 are indeed tractable.

In those cases where the dimension of $\boldsymbol{\theta}$ is small (less than three, say) it is possible to derive an independence sampler by reusing (4) to build an approximation of the marginal posterior for $\boldsymbol{\theta}$. The starting point is the identity

$$\pi(\boldsymbol{\theta} \mid \boldsymbol{y}) = \frac{\pi(\boldsymbol{x}, \boldsymbol{\theta} \mid \boldsymbol{y})}{\pi(\boldsymbol{x} \mid \boldsymbol{\theta}, \boldsymbol{y})}. \tag{6}$$

By approximating the denominator via expression (4) and evaluating the right-hand side at the modal value for $\boldsymbol{x}$ (for each $\boldsymbol{\theta}$), we obtain an approximation for the marginal posterior, which we denote by $\widetilde{\pi}(\boldsymbol{\theta}|\boldsymbol{y})$. This approximation is in fact the Laplace-approximation suggested by Tierney and Kadane (1986), who showed that its relative error is $\mathcal{O}(N^{-3/2})$ after renormalisation. (Here, N is the number of observations.) The approximation $\widetilde{\pi}(\boldsymbol{\theta'}|\boldsymbol{y})$ then replaces $q(\boldsymbol{\theta'}|\boldsymbol{\theta})$ in the one-block algorithm above. The independence sampler uses the approximation

$$\widetilde{\pi}(\boldsymbol{x}, \boldsymbol{\theta} \mid \boldsymbol{y}) = \widetilde{\pi}(\boldsymbol{\theta} \mid \boldsymbol{y})\, \widetilde{\pi}(\boldsymbol{x} \mid \boldsymbol{\theta}, \boldsymbol{y}). \tag{7}$$

A natural question arises here. If we can use $\widetilde{\pi}(\boldsymbol{x}, \boldsymbol{\theta}|\boldsymbol{y})$ to construct an independence sampler to explore $\pi(\boldsymbol{x}, \boldsymbol{\theta}|\boldsymbol{y})$, why not just compute approximations to the marginals from $\widetilde{\pi}(\boldsymbol{x}, \boldsymbol{\theta} \mid \boldsymbol{y})$ directly?

Since (4) is Gaussian, it is, theoretically, always possible to (approximately) compute the marginal for the x_i's as

$$\widehat{\widetilde{\pi}}(x_i \mid \boldsymbol{y}) = \sum_j \widetilde{\pi}(x_i \mid \boldsymbol{\theta}_j, \boldsymbol{y})\, \widetilde{\pi}(\boldsymbol{\theta}_j \mid \boldsymbol{y})\, \Delta_j \tag{8}$$

by simply summing out $\boldsymbol{\theta}$ by some numerical integration rule where Δ_j is the weight associated with $\boldsymbol{\theta}_j$. The approximated marginal posterior $\widetilde{\widetilde{\pi}}(x_i \mid \boldsymbol{y})$ is a mixture of Gaussians where the weights, mean and variances, are computed from (7). However, the dimension of $\boldsymbol{x}$ is usually large, thus obtaining the marginal variances for $x_i|\boldsymbol{\theta}, \boldsymbol{y}$ is computationally intensive (recall that only the precision matrix $\boldsymbol{Q}$ is explicitly known). Therefore the marginals in (8) are, in practise, possible to compute only for GMRFs since in, these cases, efficient computations are possible. A recursion algorithm to efficiently compute marginal variances for GMRFs is described in Section 2.

Although any MCMC algorithm will guarantee the correct answer in the end, the question is what happens in finite time. The Monte Carlo error is $\mathcal{O}_p(M^{-1/2})$ where M is the (effective) number of samples, hence, the strength of the MCMC approach is to provide rough (near) unbiased estimates rather quickly, on the other side, precise estimates may take unreasonable long time. Any (deterministic) approximated inference can in fact compete with a MCMC approach, as long as its squared "bias", or error, is comparable with the Monte Carlo error. The most interesting aspect of approximation (8), is that it can be computed almost instantly compared to the time any MCMC algorithm will have to run to obtain any decent accuracy.

The aim of this paper is to investigate how accurate (8) is for some typical examples of hierarchical GMRF models. In Section 3 we report some experiments using models for disease mapping on a varying scale of difficulty. We compare the marginals of interest as approximated by (8) and as estimated from very long MCMC runs. The results are very positive. Before presenting the examples, we will, in Section 2, discuss how to efficiently compute marginal variances needed in expression (8) for GMRFs. This Section also explains (implicit) why fast computations of GMRFs are possible using numerical methods for sparse matrices. Section 2 is unavoidably somewhat technical, but it is not necessary to appreciate the results in Section 3. We end with a discussion in Section 4.

2 Computing marginal variances for a GMRF

GMRFs are nearly always specified by their precision matrix $\boldsymbol{Q}$ meaning that the covariance matrix, $\boldsymbol{\Sigma} = \boldsymbol{Q}^{-1}$ is only implicitly known. Although we can formally invert $\boldsymbol{Q}$, the dimension n is typically large ($10^3 - 10^5$) so inverting $\boldsymbol{Q}$ directly is costly and inconvenient. In this section we discuss a simple and fast algorithm to compute marginal variances, applicable for GMRFs with large dimension. The starting point is the not-well-known matrix identity which appeared in a IEEE conference proceedings (Takahashi et al., 1973). In our setting, the identity is as follows. Let $\boldsymbol{L}\boldsymbol{L}^T = \boldsymbol{V}\boldsymbol{D}\boldsymbol{V}^T$ be the Cholesky-decomposition of $\boldsymbol{Q}$ where $\boldsymbol{L} = \boldsymbol{V}\boldsymbol{D}^{1/2}$ is the (lower triangular) Cholesky triangle, $\boldsymbol{D}$ is a diagonal matrix and $\boldsymbol{V}$ is a lower triangular matrix with ones on the diagonal. Then

$$\boldsymbol{\Sigma} = \boldsymbol{D}^{-1}\boldsymbol{V}^{-1} + (\boldsymbol{I} - \boldsymbol{V}^T)\boldsymbol{\Sigma}. \tag{9}$$

(The proof is simple; Since $\boldsymbol{Q}\boldsymbol{\Sigma} = \boldsymbol{I}$ then $\boldsymbol{V}\boldsymbol{D}\boldsymbol{V}^T\boldsymbol{\Sigma} = \boldsymbol{I}$. Multiplying from left with $(\boldsymbol{V}\boldsymbol{D})^{-1}$ and then adding $\boldsymbol{\Sigma}$ on both sides gives (9) after rearrangement.) A close look at (9) will reveal that the upper triangle of (9) defines recursions for Σ_{ij} (Takahashi et al., 1973), and this provide the basis for fast computations of the marginal variances of x_1 to x_n.

However, the identity (9) gives little insight in how Σ_{ij} depends on the elements of $\boldsymbol{Q}$ and on the graph $\mathcal{G}$. We will therefore, in Section 2.1, derive the recursions defined in (9) "statistically", starting from a simulation algorithm for GMRFs and using the relation between $\boldsymbol{Q}$ and its Cholesky triangle given by the global Markov property. We use the same technique to prove Theorem 1, given in Section 2.1. This theorem locates a set of indexes for which the recursions are to be solved to obtain the marginal variances. A similar result was also given in Takahashi et al. (1973), see also Erisman and Tinney (1975). We also generalise the recursions to compute marginal variances for GMRFs defined with additional soft and hard linear constraints, for example under a sum-to-zero constraint. Practical

issues appearing when implementing the algorithm using the Cholesky triangle of $\boldsymbol{Q}$ computed using sparse matrix libraries, are also discussed.

The recursions for Σ_{ij} are applicable to a GMRF with respect to any graph $\mathcal{G}$ and generalise the well known (fixed-interval) Kalman recursions for smoothing applicable for dynamic models. The computational effort needed to solve the recursions depends on both the neighbourhood structure in $\mathcal{G}$ and the size n. For typical spatial applications, the cost is $\mathcal{O}(n\log(n)^2)$ when the Cholesky triangle of $\boldsymbol{Q}$ is available.

2.1 The Recursions

The Cholesky triangle $\boldsymbol{L}$ (of $\boldsymbol{Q}$) is the starting point both for producing (unconditional and conditional) samples from a zero mean GMRF and for evaluating the log-density for any configuration. Refer to Rue and Held (2005, Ch. 2) for algorithms and further details. In short, unconditional samples are found as the solution of $\boldsymbol{L}^T\boldsymbol{x} = \boldsymbol{z}$ where $\boldsymbol{z} \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{I})$. The log-density is computed using that $\log|\boldsymbol{Q}| = 2\sum_i \log L_{ii}$.

Since the solution of $\boldsymbol{L}^T\boldsymbol{x} = \boldsymbol{z}$ is a sample from a zero mean GMRF with precision matrix $\boldsymbol{Q}$, we obtain that

$$x_i \mid x_{i+1}, \ldots, x_n \;\sim\; \mathcal{N}(-\frac{1}{L_{ii}} \sum_{k=i+1}^{n} L_{ki}x_k, 1/L_{ii}^2), \quad i = n, \ldots, 1. \tag{10}$$

Eq. (10) provides a sequential representation of the GMRF backward in "time" i, as

$$\pi(\boldsymbol{x}) = \prod_{i=n}^{1} \pi(x_i \mid x_{i+1}, \ldots, x_n).$$

Let $\boldsymbol{L}_{i:n}$ be the lower-right $(n-i-1)\times(n-i-1)$ submatrix of $\boldsymbol{L}$. It follows directly from $\boldsymbol{L}^T\boldsymbol{x} = \boldsymbol{z}$ that $\boldsymbol{L}_{i:n}\boldsymbol{L}_{i:n}^T$ is the precision matrix of $\boldsymbol{x}_{i:n} = (x_i, \ldots, x_n)^T$. The non-zero pattern in $\boldsymbol{L}$ is important for the recursions, see Rue and Held (2005, Ch. 2) for further details about the relation between $\boldsymbol{Q}$ and $\boldsymbol{L}$. Zeros in the i'th column of $\boldsymbol{L}$, $\{L_{ki}, k = 1, \ldots, n\}$, relates directly to the conditional independence properties of $\pi(\boldsymbol{x}_{i:n})$. For $i < k$, we have

$$-\frac{1}{2}\boldsymbol{x}_{i:n}^T\boldsymbol{L}_{i:n}\boldsymbol{L}_{i:n}^T\boldsymbol{x}_{i:n} = -x_i x_k L_{ii} L_{ki} + \text{remaining terms}$$

hence $L_{ki} = 0$ means that x_i and x_k are conditional independent given $x_{i+1}, \ldots, x_{k-1}, x_{k+1}, \ldots, x_n$. This is similar to the fact that $Q_{ij} = 0$ means that x_i and x_j are conditional independent given the remaining elements of $\boldsymbol{x}$. To ease the notation, define the set

$$F(i,k) = \{i+1, \ldots, k-1, k+1, \ldots, n\}, \quad 1 \le i \le k \le n$$

which is the future of i except k. Then for $i < k$

$$x_i \perp x_k \mid \boldsymbol{x}_{F(i,k)} \Longleftrightarrow L_{ki} = 0. \tag{11}$$

Unluckily it is not easy to verify that $x_i \perp x_k \mid \boldsymbol{x}_{F(i,k)}$ without computing $\boldsymbol{L}$ and checking if $L_{ki} = 0$ or not. However, the global Markov property provides a sufficient condition for L_{ki} to be zero. If i and $k > i$ are separated by $F(i,k)$ in $\mathcal{G}$, then $x_i \perp x_k \mid \boldsymbol{x}_{F(i,k)}$ and $L_{ki} = 0$. This sufficient criterion depends only on the graph $\mathcal{G}$. If we use this to conclude that $L_{ki} = 0$, then this is true for all $\boldsymbol{Q} > 0$ with fixed graph $\mathcal{G}$. In particular, if $k \sim i$ then L_{ki} is non-zero in general. This imply that the Cholesky triangle is in general more dense than the lower triangle of $\boldsymbol{Q}$.

To obtain the recursions for $\boldsymbol{\Sigma} = \boldsymbol{Q}^{-1}$, we note that (10) implies that

$$\Sigma_{ij} = \delta_{ij}/L_{ii}^2 - \frac{1}{L_{ii}} \sum_{k \in \mathcal{I}(i)} L_{ki}\Sigma_{kj}, \quad j \geq i,\ i = n, \ldots, 1, \tag{12}$$

where $\mathcal{I}(i)$ includes those k larger than i and where L_{ki} is non-zero,

$$\mathcal{I}(i) = \{k > i \ : \ L_{ki} \neq 0\} \tag{13}$$

and δ_{ij} is one if $i = j$ and zero otherwise. Note that (12) equals the upper triangle of (9). We can compute all covariances directly using (12) but the order of the indexes are important. In the outer loop i runs from n to 1 and the inner loop j runs from n to i. The first and last computed covariance is then Σ_{nn} and Σ_{11}, respectively.

It is possible to derive a similar set of equations to (12) which relates covariances to elements of $\boldsymbol{Q}$ instead of elements of $\boldsymbol{L}$, see Besag (1981). However, these equations does not define recursions.

Example 1 *Let $n = 3$, $\mathcal{I}(1) = \{2, 3\}$, $\mathcal{I}(2) = \{3\}$, then (12) gives*

$$\begin{aligned}
\Sigma_{33} &= \frac{1}{L_{33}^2} & \Sigma_{23} &= -\frac{1}{L_{22}} (L_{32}\Sigma_{33}) \\
\Sigma_{22} &= \frac{1}{L_{22}^2} - \frac{1}{L_{22}} (L_{32}\Sigma_{32}) & \Sigma_{13} &= -\frac{1}{L_{11}} (L_{21}\Sigma_{23} + L_{31}\Sigma_{33}) \\
\Sigma_{12} &= -\frac{1}{L_{11}} (L_{21}\Sigma_{22} + L_{31}\Sigma_{32}) & \Sigma_{11} &= \frac{1}{L_{11}^2} - \frac{1}{L_{11}} (L_{21}\Sigma_{21} + L_{31}\Sigma_{31})
\end{aligned}$$

where we also need to use that $\boldsymbol{\Sigma}$ is symmetric.

Our aim is to compute the marginal variances $\Sigma_{11}, \ldots, \Sigma_{nn}$. In order to do so, we need to compute Σ_{ij} (or Σ_{ji}) for all ij in some set $\mathcal{S}$, as evident from (12). Let the elements in $\mathcal{S}$ be unordered, meaning that if $ij \in \mathcal{S}$ then $ji \in \mathcal{S}$. If the recursions can be solved by only computing Σ_{ij} for all $ij \in \mathcal{S}$ we say that the recursions are solvable using $\mathcal{S}$, or simply that $\mathcal{S}$ is solvable. A sufficient condition for a set $\mathcal{S}$ to be solvable is that

$$ij \in \mathcal{S} \text{ and } k \in \mathcal{I}(i) \implies kj \in \mathcal{S} \tag{14}$$

and that $ii \in \mathcal{S}$ for $i = 1, \ldots, n$. Of course $\mathcal{S} = \mathcal{V} \times \mathcal{V}$ is such a set, but we want $|\mathcal{S}|$ to be minimal to avoid unnecessary computations. Such a minimal set depends, however, on the numerical values in $\boldsymbol{L}$ or $\boldsymbol{Q}$ implicitly. Denote by $\mathcal{S}(\boldsymbol{Q})$ a minimal set for a certain precision matrix $\boldsymbol{Q}$. The following result identifies a solvable set $\mathcal{S}^*$ containing the union of $\mathcal{S}(\boldsymbol{Q})$ for all $\boldsymbol{Q} > 0$ with a fixed graph $\mathcal{G}$.

Theorem 1 *The union of $\mathcal{S}(\boldsymbol{Q})$ for all $\boldsymbol{Q} > 0$ with fixed graph $\mathcal{G}$, is a subset of*

$$\mathcal{S}^* = \{ij \in \mathcal{V} \times \mathcal{V} \ : \ j \geq i,\ i \text{ and } j \text{ are not separated by } F(i, j)\}$$

and the recursions in (12) are solvable using $\mathcal{S}^*$.

Proof. *To prove the theorem we have to show that $\mathcal{S}^*$ is solvable and that it contains the union of $\mathcal{S}(\boldsymbol{Q})$ for all $\boldsymbol{Q} > 0$ with fixed graph $\mathcal{G}$. To verify that the recursions are solvable using $\mathcal{S}^*$, first note that $ii \in \mathcal{S}^*$, for $i = 1, \ldots, n$ since i and i are not separated by $F(i, i)$. The global Markov property ensures that if $ij \notin \mathcal{S}^*$ then $L_{ji} = 0$ for all $\boldsymbol{Q} > 0$ with fixed graph $\mathcal{G}$. Using this feature we can*

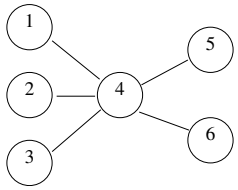
replace $\mathcal{I}(i)$ with $\mathcal{I}^*(i) = \{k > i : ik \in \mathcal{S}^*\}$ in (14). This is legal since $\mathcal{I}(i) \subseteq \mathcal{I}^*(i)$ and the difference between the two sets only identifies terms L_{ki} which are zero. Then, we have to show that

$$ij \in \mathcal{S}^* \text{ and } ik \in \mathcal{S}^* \implies kj \in \mathcal{S}^* \tag{15}$$

Eq. (15) is trivially true for $i \leq k = j$. Fix now $i < k < j$. Then $ij \in \mathcal{S}^*$ says that there exists a path $i, i_1, \ldots, i_n, j$, where $i_1, \ldots, i_n$ are all smaller than i, and $ik \in \mathcal{S}^*$ says that there exists a path $i, i'_1, \ldots, i'_{n'}, k$, where $i'_1, \ldots, i'_{n'}$ are all smaller than i. Then there is a path from k to i and from i to j where all nodes are less than or equal to i. Since $i < k$ then all the nodes in the two paths are less than k. Hence, there is a path from k and j where all nodes are less than k. This means that k and j are not separated by $F(k, j)$, so $kj \in \mathcal{S}^*$. Finally, since $\mathcal{S}^*$ only depends on $\mathcal{G}$, it must contain all $\mathcal{S}(\boldsymbol{Q})$ since each $\mathcal{S}(\boldsymbol{Q})$ is minimal, and therefore contains their union too. ■

An alternative interpretation of $\mathcal{S}^*$, is that it identifies only from the graph $\mathcal{G}$, all possible non-zero elements in $\boldsymbol{L}$. Some of these might turn out to be zero depending on the conditional independence properties of the marginal density for $\boldsymbol{x}_{i:n}$ for $i = n, \ldots, 1$, see (11). In particular, if $j \sim i$ and $j > i$ then $ij \in \mathcal{S}^*$. This provides the lower bound for the size of $\mathcal{S}^*$: $|\mathcal{S}^*| \geq n + |\mathcal{E}|$.

Example 2 *Let $\boldsymbol{x} = (x_1, \ldots, x_6)^T$ be a GMRF with respect to the graph*



Then, the set of the possible non-zero terms in $\boldsymbol{L}$ are

$$\mathcal{S}^* = \{11, 22, 33, 41, 42, 43, 44, 54, 55, 64, 65, 66\}. \tag{16}$$

The only element in $\mathcal{S}^*$ where the corresponding element in $\boldsymbol{Q}$ is zero, is 65, this because 5 and 6 are not separated by $F(5, 6) = \emptyset$ in $\mathcal{G}$ (due to 4), so $|\mathcal{S}^*| = n + |\mathcal{E}| + 1$.

The size of $\mathcal{S}^*$ depends not only on the graph $\mathcal{G}$ but also on the permutation of the vertices in the graph $\mathcal{G}$. It is possible to show that, if the graph $\mathcal{G}$ is decomposable, then there exists a permutation of the vertices, such that $|\mathcal{S}^*| = n + |\mathcal{E}|$ and $\mathcal{S}^*$ *is* the union of $\mathcal{S}(\boldsymbol{Q})$ for all $\boldsymbol{Q} > 0$ with fixed graph $\mathcal{G}$. The typical example is the following.

Example 3 *A homogeneous autoregressive model of order p satisfies*

$$x_i \mid x_1, \ldots, x_{i-1} \sim \mathcal{N}(\sum_{j=1}^{p} \phi_j x_{i-j}, 1), \quad i = 1, \ldots, n,$$

for some parameters $\{\phi_j\}$ where for simplicity we assume that $x_{-1}, \ldots, x_{-p+1}$ are fixed. Let $\{y_i\}$ be independent Gaussian observations of x_i such that $y_i \sim \mathcal{N}(x_i, 1)$. Then $\boldsymbol{x}$ conditioned on the observations is Gaussian where the precision matrix $\boldsymbol{Q}$ is a band-matrix with band-width p and $\boldsymbol{L}$ is lower triangular with the same bandwidth. When $\{\phi_j\}$ are such that $Q_{ij} \neq 0$ for all $|i - j| \leq p$, then the graph is decomposable. In this case the recursions correspond to the (fixed-interval) smoothing recursions derived from the Kalman filter for (Gaussian) linear state-space models.

Although the situation is particularly simple for decomposable graphs, most GMRFs are defined with respect to graphs that are not decomposable. This is the case for GMRFs used in spatial or

spatio-temporal applications, but also for GMRFs used in temporal models outside the state-space framework. In addition to be able to identify the set $\mathcal{S}^*$ efficiently, we also need to compute the Cholesky triangle $\boldsymbol{L}$. It is important to have efficient algorithms for these tasks as the dimension of GMRFs is typically large. Fortunately, algorithms that compute $\boldsymbol{L}$ efficiently also minimise (approximately) the size of $\mathcal{S}^*$ and then also the cost of solving the recursions. We return to this and other practical issues in Section 2.3, after discussing how to compute marginal variances for GMRFs with additional linear constraints.

2.2 Correcting for hard and soft linear constraints

We will now demonstrate how we can correct the marginal variances computed in (12) to account for additional linear constraints, for example a simple sum-to-zero constraint. Let $\boldsymbol{A}$ be a $k \times n$ matrix of rank k. The goal is now to compute the marginal variances of the GMRF under the linear constraint $\boldsymbol{A}\boldsymbol{x} = \boldsymbol{e}$. If $\boldsymbol{e}$ is fixed we denote the constraint as *hard*, and if $\boldsymbol{e}$ is a realisation of $\mathcal{N}(\boldsymbol{\mu}_{\boldsymbol{e}}, \boldsymbol{\Sigma}_{\boldsymbol{e}})$, $\boldsymbol{\Sigma}_{\boldsymbol{e}} > 0$, we denote the constraint as *soft*.

A constrained GMRF is also a GMRF, meaning that the recursions (12) are still valid using the Cholesky triangle for the constrained GMRF. Since linear constraints destroy the sparseness of the precision matrix they will not allow fast computation of the marginal variances. However, the covariance matrix under hard linear constraints, $\widetilde{\boldsymbol{\Sigma}}$, relates to the unconstrained covariance matrix $\boldsymbol{\Sigma}$ as

$$\widetilde{\boldsymbol{\Sigma}} = \boldsymbol{\Sigma} - \boldsymbol{Q}^{-1}\boldsymbol{A}^T \left(\boldsymbol{A}\boldsymbol{Q}^{-1}\boldsymbol{A}^T\right)^{-1} \boldsymbol{A}\boldsymbol{Q}^{-1}. \tag{17}$$

There is a similar relation with a soft constraint (Rue and Held, 2005, Ch. 2). In the following we assume a hard constraint. It is evident from (17) that

$$\widetilde{\Sigma}_{ii} = \Sigma_{ii} - \left(\boldsymbol{Q}^{-1}\boldsymbol{A}^T \left(\boldsymbol{A}\boldsymbol{Q}^{-1}\boldsymbol{A}^T\right)^{-1} \boldsymbol{A}\boldsymbol{Q}^{-1}\right)_{ii}, \quad i = 1, \ldots, n.$$

Hence, we can compute the diagonal of $\boldsymbol{\Sigma}$ and then correct it to account for the hard constraints. Define the $n \times k$ matrix $\boldsymbol{W}$ as $\boldsymbol{Q}^{-1}\boldsymbol{A}^T$ which is found from solving $\boldsymbol{Q}\boldsymbol{W} = \boldsymbol{A}^T$ for each of the k columns of $\boldsymbol{W}$. As the Cholesky triangle to $\boldsymbol{Q}$ is available, the j'th column of $\boldsymbol{W}$, $\boldsymbol{W}_j$, is found by solving $\boldsymbol{L}\boldsymbol{v} = \boldsymbol{A}_j^T$ and then solving $\boldsymbol{L}^T\boldsymbol{W}_j = \boldsymbol{v}$. We now see that $\widetilde{\Sigma}_{ii} = \Sigma_{ii} - C_{ii}$ where $\boldsymbol{C} = \boldsymbol{W}\left(\boldsymbol{A}\boldsymbol{W}\right)^{-1}\boldsymbol{W}^T$. We only need the diagonal of $\boldsymbol{C}$. Let $\boldsymbol{V} = \boldsymbol{W}(\boldsymbol{A}\boldsymbol{W})^{-1}$, and then $\boldsymbol{C} = \boldsymbol{V}\boldsymbol{W}^T$ and $C_{ii} = \sum_{l=1}^{k} V_{il}W_{il}$. The cost of computing $\boldsymbol{V}$ and $\boldsymbol{W}$ is for large k dominated by factorising the (dense) $k \times k$ matrix $\boldsymbol{A}\boldsymbol{W}$, which is cubic in k. As long as k is not too large it is nearly free to correct for linear soft and hard constraints.

A special case of hard constraint is to condition on a subset, B say, of the nodes in $\mathcal{G}$. This is equivalent to computing the marginal variances for $\boldsymbol{x}_A|\boldsymbol{x}_B$ where $\boldsymbol{x} = (\boldsymbol{x}_A, \boldsymbol{x}_B)$ is a zero mean GMRF. In most cases it is more efficient not to use (17), but utilise that $\boldsymbol{x}_A|\boldsymbol{x}_B$ is a GMRF with precision matrix $\boldsymbol{Q}_{AA}$ and mean $\boldsymbol{\mu}$ given by the solution of $\boldsymbol{Q}_{AA}\boldsymbol{\mu} = -\boldsymbol{Q}_{AB}\boldsymbol{x}_B$. (Note that solving for $\boldsymbol{\mu}$ require only the Cholesky triangle of $\boldsymbol{Q}_{AA}$ which is needed in any case for the recursions.) The marginal variances are then computed using (12), possibly correcting for additional linear constraints using (17).

2.3 Practical issues

Since the precision matrix $\boldsymbol{Q}$ is a sparse matrix we can take advantage of numerical algorithms for sparse symmetric positive definite matrices. Such algorithms are very efficient and make it possible to factorise precision matrices of dimension $10^3 - 10^5$ without too much effort. A major benefit is that these algorithms also minimise (approximately) the size of $\mathcal{S}^*$, and hence the cost of solving the

recursions described earlier. Rue (2001) and Rue and Held (2005) discuss numerical algorithms for sparse matrices from a statistical perspective and how to apply them for GMRFs.

An important ingredient in sparse matrix algorithms is to permute the vertices to minimise (approximately) the number of non-zero terms in $\boldsymbol{L}$. The idea is as follows, if L_{ji} is known to be zero, then L_{ji} is not computed. It turns out that the set $\mathcal{S}^*$ is exactly the set of vertices for which L_{ji} is computed, see Rue and Held (2005, Sec. 2.4.1). A permutation to efficiently compute $\boldsymbol{L}$ minimise (approximately) $|\mathcal{S}^*|$, hence is also an efficient permutation for solving the recursions. However, this implies that we have little control over which Σ_{ij}'s are computed in the recursions, apart from the diagonal and those elements where $i \sim j$.

Permutation schemes based on the idea of nested dissection are particularly useful in statistical applications. The idea is to find a small separating subset that divides the graph into two (roughly) equal parts, label the nodes in the separating set with the highest indexes, and continue recursively. For such a permutation, the computational complexity to compute $\boldsymbol{L}$ for a GMRF on a square $m \times m$ lattice with a local neighbourhood, is $\mathcal{O}(n^{3/2})$ for $n = m^2$. This also gives the optimal complexity in the order sense. The number of possible non-zero terms in $\boldsymbol{L}$ is $\mathcal{O}(n \log(n))$ which corresponds to the size of $\mathcal{S}^*$. The complexity of solving the recursions can be estimated from these numbers. We need to compute $\mathcal{O}(n \log(n))$ covariances, each involving on average $\mathcal{O}(\log(n))$ terms in $\mathcal{I}^*(i)$, which in total gives a cost of $\mathcal{O}(n \log(n)^2)$ operations. For a local GMRF on a $m \times m \times m$ cube with $n = m^3$ the size of $\mathcal{S}^*$ is $\mathcal{O}(n^{4/3})$, and the cost of solving the recursions is then $\mathcal{O}(n^{5/3})$. This cost is dominated by the cost of factorising $\boldsymbol{Q}$, which is $\mathcal{O}(n^2)$.

A practical concern arises when numerical libraries return a list with the non-zero elements in $\boldsymbol{L}$, but the set $\mathcal{S}^*$ or $\mathcal{S}(\boldsymbol{Q})$ is needed by the recursions. In fact, any easily obtainable solvable set $\mathcal{S}(\boldsymbol{Q})^+$, where $\mathcal{S}(\boldsymbol{Q}) \subseteq \mathcal{S}(\boldsymbol{Q})^+ \subseteq \mathcal{S}^*$, is acceptable. A simple approach to obtain *a* $\mathcal{S}(\boldsymbol{Q})^+$ is the following. Let $\mathcal{S}_0 = \{j \geq i \,:\, L_{ji} \neq 0\}$. Traverse the set $\mathcal{S}_0$ with i from n to 1 as the outer loop, and j from n to i such that $ij \in \mathcal{S}_0$. For each ij, check for each $k \in \mathcal{I}(i)$ if $kj \in \mathcal{S}_0$. If this is not true, then add kj to $\mathcal{S}_0$. Repeat this procedure until no changes appear in $\mathcal{S}_0$. By construction, $\mathcal{S}_0 \subseteq \mathcal{S}^*$ and $\mathcal{S}_0$ is solvable, hence we may use $\mathcal{S}(\boldsymbol{Q})^+ = \mathcal{S}_0$. Two iterations are often sufficient to obtain $\mathcal{S}(\boldsymbol{Q})^+$, where the last verify only that $\mathcal{S}_0$ is solvable. Alternatively, $\mathcal{S}^*$ can either computed directly or extracted from an intermediate result in the sparse matrix library, if this is easily accessible.

Needless to say, solving the recursions efficiently requires very careful implementation in an appropriate language, but this is the rule, not the exception when working with sparse matrices. The open-source library `GMRFLib` (Rue and Held, 2005, Appendix B) includes an efficient implementation of the recursions as well as numerous of useful routines for GMRFs. All the examples in Section 3 make extensive use of `GMRFLib`, which can be downloaded from the first author's www-page.

3 Examples

In this section, we will present some results for the approximations for the marginal posteriors computed from (7), and their comparison with estimates obtained from very long MCMC runs. We will restrict ourself to the well-known BYM-model for disease mapping (Section 3.1). The BYM-model is a hierarchical GMRF model with Poisson distributions at the first stage. We will use two different datasets, which we describe as "easy" (many counts) and "hard" (few counts). The comparison of the marginal posteriors for the hyperparameters (in this case, the precisions) are presented in Section 3.2, while the posterior marginals for the latent GMRF are presented in Section 3.3. In Section 3.4 we present some results for an extended BYM-model, where we include a semi-parametric effect of a covariate and where the latent GMRF has to obey a linear constraint.

Note that the computational speed in the following experiments is not optimal due to rather brute-

force approach taken while integrating out the hyperparameters $\boldsymbol{\theta}$. However, this step can be improved considerably, as we discuss in Section 4, while the approximation results themselves remain unaffected.

3.1 The BYM-model for disease mapping

We will now introduce the BYM-model for analysing spatial disease data (Besag et al., 1991). This model is commonly used in epidemiological applications.

The number of incidents y_i, $i = 1, \ldots, N$, of a particular disease is observed over a certain time period in a site of N districts. It is common to assume the observed counts to be conditionally independent and Poisson distributed with mean $e_i \exp(\eta_i)$, where η_i is the log-relative risk and e_i is the expected number of cases computed on some demographic parameters. Further, η_i is decomposed as $\eta_i = u_i + v_i$ where $\boldsymbol{u} = \{u_i\}$ is a spatially structured component and $\boldsymbol{v}$ is an unstructured component. An intrinsic GMRF of the following form is often assumed for the spatially structured component,

$$\pi(\boldsymbol{u} \mid \kappa_{\boldsymbol{u}}) \propto \kappa_{\boldsymbol{u}}^{(n-1)/2} \exp\left(-\frac{\kappa_{\boldsymbol{u}}}{2} \sum_{i \sim j} (u_i - u_j)^2\right) \tag{18}$$

where $\kappa_{\boldsymbol{u}}$ is the unknown precision parameter. Two districts i and j are defined to be neighbours, $i \sim j$, if they are adjacent. Further, $\boldsymbol{v}$ are independent zero mean normals with unknown precision parameter $\kappa_{\boldsymbol{v}}$. The precisions are (most commonly) assigned independent Gamma priors with fixed parameters.

The BYM-model is of course a hierarchical GMRF model, with $y_i \sim \text{Po}(e_i \exp(\eta_i))$ at the first stage. At the second stage the GMRF is $\boldsymbol{x} = (\boldsymbol{\eta}^T, \boldsymbol{u}^T)^T$. The unknown precisions $\boldsymbol{\kappa} = (\kappa_{\boldsymbol{u}}, \kappa_{\boldsymbol{v}})$ constitute the third stage. Note that we have reparametrised the GMRF using $\boldsymbol{x} = (\boldsymbol{\eta}^T, \boldsymbol{u}^T)^T$ instead of $\boldsymbol{x} = (\boldsymbol{v}^T, \boldsymbol{u}^T)^T$, in this way some of the nodes in the graph, namely the $\boldsymbol{\eta}$'s, are observed through the data $\boldsymbol{y}$. The posterior of interest if therefore

$$\pi(\boldsymbol{x}, \boldsymbol{\kappa} \mid \boldsymbol{y}) \propto \kappa_{\boldsymbol{v}}^{N/2} \kappa_{\boldsymbol{u}}^{(N-1)/2} \exp\left(-\frac{1}{2} \boldsymbol{x}^T \boldsymbol{Q} \boldsymbol{x}\right) \exp\left(\sum_{i=1}^{N} y_i x_i - e_i \exp(x_i)\right) \pi(\boldsymbol{\kappa}) \tag{19}$$

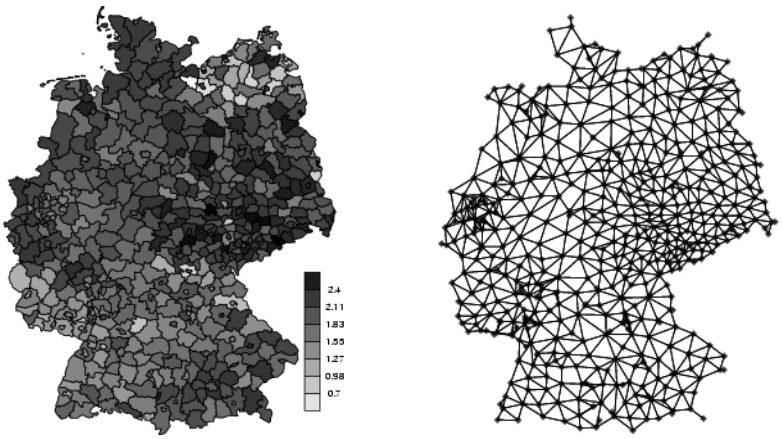
The $2N \times 2N$ precision matrix for the GMRF, $\boldsymbol{Q}$ is

$$\boldsymbol{Q} = \begin{pmatrix} \kappa_{\boldsymbol{v}} \boldsymbol{I} & -\kappa_{\boldsymbol{v}} \boldsymbol{I} \\ -\kappa_{\boldsymbol{v}} \boldsymbol{I} & \kappa_{\boldsymbol{u}} \boldsymbol{R} + \kappa_{\boldsymbol{v}} \boldsymbol{I} \end{pmatrix} \tag{20}$$

where $\boldsymbol{R}$ is the so-called structure matrix for the spatial term, R_{ii} is the number of neighbours to district i, and $R_{ij} = -1$ if $i \sim j$ (district i and j are adjacent) and zero otherwise. We set the priors of the unknown precisions to be independent and Gamma(a, b) distributed with a/b as the expected value. The values of a and b are specified later.

The two datasets we will consider in Section 3.2 and Section 3.3 are classified as the Easy-case and the Hard-case.

Easy-case The observed oral cavity cancer mortality for males in Germany (1986–1990) was previously analysed by Knorr-Held and Raßer (2000). The data have an average observed count of 28.4, median of 19, and the first and third quantile are 9 and 33. For such high counts the Poisson distribution is not too far away from a Gaussian. The observed standardised mortality ratio for the different districts of Germany are shown in Figure 1a. The corresponding graph is displayed in Figure 1b. It has $n = 544$ nodes with average 5.2, minimum 1, and maximum 11 neighbours. The parameters in the prior for the precisions are $a = 1$ and $b = 0.01$ following Rue and Held (2005, Ch. 4).



(a) (b)

Figure 1: (a) The standardised mortality ratio y_i/e_i for the oral cavity cancer counts in Germany (1986–1990). (b) The graph associated with (a) where two districts are neighbours if and only if they are adjacent.

Hard-case The observed Insulin dependent Diabetes Mellitus in Sardinia. These data were previously analysed by Bernardinelli et al. (1997) and also used by Knorr-Held and Rue (2002) as a challenging case. The graph is similar to the one in Figure 1b, and has $n = 366$ nodes with average 5.4, minimum 1 and maximum 13 neighbours. This is a sparse dataset with a total of 619 cases and median of 1. For such low counts the Poisson distribution is quite different from a Gaussian. The parameters in the prior for the precisions are $a = 1$ and $b = 0.0005$ for $\kappa_{\boldsymbol{u}}$, and $a = 1$ and $b = 0.00025$ for $\kappa_{\boldsymbol{v}}$ following Knorr-Held and Rue (2002).

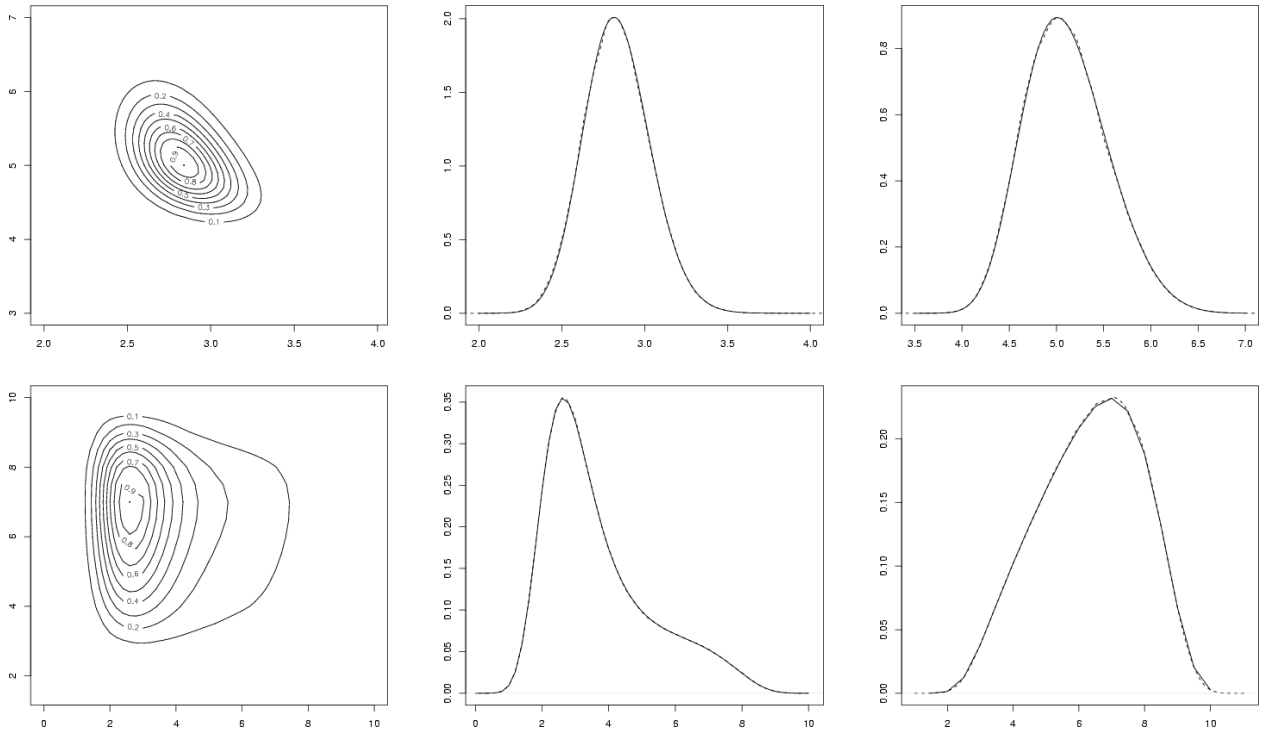
3.2 Approximating $\pi(\boldsymbol{\theta}|\boldsymbol{y})$

Our first task is to approximate the marginal posteriors for the hyperparameters $\log \kappa_{\boldsymbol{u}}$ and $\log \kappa_{\boldsymbol{v}}$, for the Easy-case and the Hard-case.

The joint marginal posterior for $\boldsymbol{\theta} = (\log \kappa_{\boldsymbol{u}}, \log \kappa_{\boldsymbol{v}})$ was estimated using the approximation to (6). This means using the GMRF-approximation (4) (depending on $\boldsymbol{\theta}$) for the full conditional $\boldsymbol{x}$ in the denominator, and then evaluate the ratio at the modal value for $\boldsymbol{x}$ for each $\boldsymbol{\theta}$. The evaluation is performed for values of $\boldsymbol{\theta}$ on a fine grid centred (approximately) at the modal value. This unnormalised density restricted to the grid is then renormalised so it integrates to one. The results are shown in column (a) in Figure 2, displaying the contour-plot of the estimated posterior marginal for $\boldsymbol{\theta}$.

The marginal posterior for the Easy-case is more symmetric than the one for the Hard-case. This is natural when we take into account the high Poisson counts which makes the likelihood more like a Gaussian. As mentioned in Section 1, this is the Laplace-approximation as derived (differently) by Tierney and Kadane (1986). The relative error in the renormalised density is $\mathcal{O}(N^{-3/2})$ where N is the number of observations, hence it is quite accurate. Note that the quality of this approximation does not change if we consider the posterior marginal for $(\kappa_{\boldsymbol{u}}, \kappa_{\boldsymbol{v}})$ instead of $(\log \kappa_{\boldsymbol{u}}, \log \kappa_{\boldsymbol{v}})$.This is, in fact, only a reparametrisation and the relative error is still $\mathcal{O}(N^{-3/2})$.

By summing out $\log \kappa_{\boldsymbol{v}}$ and $\log \kappa_{\boldsymbol{u}}$, respectively, we obtain the marginal posteriors for $\log \kappa_{\boldsymbol{u}}$ and $\log \kappa_{\boldsymbol{v}}$. These are displayed using solid lines in Figure 2 column (b) and (c). To verify these approximations, we ran MCMC algorithms based on (5) for a long time to obtain at least 10^6 near iid samples. The density estimates based on these samples are shown as dotted lines in column (b) and (c). The estimates based on the MCMC algorithms confirm the accuracy of the Laplace-approximation.



(a) (b) (c)

Figure 2: Results for the Easy-case on the top row and the for Hard-case on the bottom row. (a) Approximated marginal posterior density of $(\log \kappa_{\boldsymbol{u}}, \log \kappa_{\boldsymbol{v}})$, (b) approximated marginal posterior density of $\log \kappa_{\boldsymbol{u}}$, and (c) approximated marginal posterior density of $\log \kappa_{\boldsymbol{v}}$. In (b) and (c), the approximated marginals are shown using solid lines, while the estimated marginal posteriors from a long MCMC run are shown with dotted lines.

3.3 Approximating $\pi(x_i|\boldsymbol{y})$

Our next task is to approximate the marginal posterior for each x_i making use of (8). Note that $\widetilde{\pi}(x_i|\boldsymbol{\theta}_j, \boldsymbol{y})$ is a GMRF, hence we need to compute the marginal variances for $x_n, \ldots, x_1$. To do this, we make use of the recursions (12) and the practical advises in Section 2.3 which are implemented in `GMRFLib` (Rue and Held, 2005, Appendix B).

The results in Section 3.2 indicate that the quality of (8) depends on how well $\widetilde{\pi}(x_i|\boldsymbol{\theta}_j, \boldsymbol{y})$ approximates $\pi(x_i|\boldsymbol{\theta}_j, \boldsymbol{y})$ for those $\boldsymbol{\theta}_j$ where the probability mass is significant. For this reason, we have compared this approximation for various fixed $\boldsymbol{\theta}_j$ with the estimates for $\pi(x_i|\boldsymbol{\theta}_j, \boldsymbol{y})$ computed from long runs with a MCMC algorithm. The results are displayed in Figure 3 for the Easy-case and Figure 4 for the Hard-case.

3.3.1 Marginal posteriors for the spatially structured component for fixed $\boldsymbol{\theta}$

Easy-case Column (d) in Figure 3 shows the value of (the fixed) $\boldsymbol{\theta}_j$ relative to the marginal posterior shown in Figure 2. The first three columns show marginals of the GMRF-approximation for the spatial component $\boldsymbol{u}$ (solid lines) and the estimate obtained from very long MCMC runs (dotted lines). Only three districts are shown. They are selected such that the posterior expected value of u_i for $\boldsymbol{\theta}_j$ located at the modal value, is high (a), intermediate (b) and low (c). The results in Figure 3 indicate that the GMRF-approximation is indeed quite accurate in this case, and only small deviations from the (estimated) truth can be detected.

Hard-case Figure 4 displays the same as Figure 3 but now for the Hard-case. The results for the three first rows are quite good, although the (estimated) true marginal posteriors show some skewness not captured by the Gaussian approximation. The modal value indicated by the Gaussian approximation seems in all cases a little too high, although this is most clear for the last row. In the last row, the precisions for both the spatial structured and unstructured term are (relatively) low and outside the region with significant contribution to the probability mass for $\boldsymbol{\theta}$. With these (relatively) low precisions, we obtain a (relatively) high variance for the non-quadratic term $\exp(x_i)$ in (19), which makes the marginals more skewed. It might appear, at a first glance, that the (estimated) true marginal and the Gaussian approximation are shifted, but this is not the case. There is a skewness factor that is missing in the Gaussian approximation, which has, in this case, nearly the same effect of a shift. The results from this Hard-case are quite encouraging, as the approximations in the central part of $\pi(\boldsymbol{\theta}|\boldsymbol{y})$ are all relatively accurate.

3.3.2 Marginal posteriors for the spatially structured component

Figure 5 shows the results using (8) (solid line) to approximate the marginals for the spatial term $\boldsymbol{u}$ in the same three districts that appear in Figure 3 and Figure 4. The (estimated) truth is drawn with dotted lines. The top row shows the Easy-case while the bottom row shows the Hard-case. The columns (a) to (c) relate to the columns of Figure 3 and Figure 4 for the top and bottom row, respectively. Since the accuracy of the Gaussian approximations was verified in Figure 3 and Figure 4 to be quite satisfactory, there is no reason that integrating out $\boldsymbol{\theta}$ will result in inferior results. The approximation (8) is quite accurate for both cases but the marginals are slightly less skewed than the truth. However, the error is quite small. The bottom row demonstrates that (8), which is a mixture of Gaussians, can indeed represent also highly skewed densities.

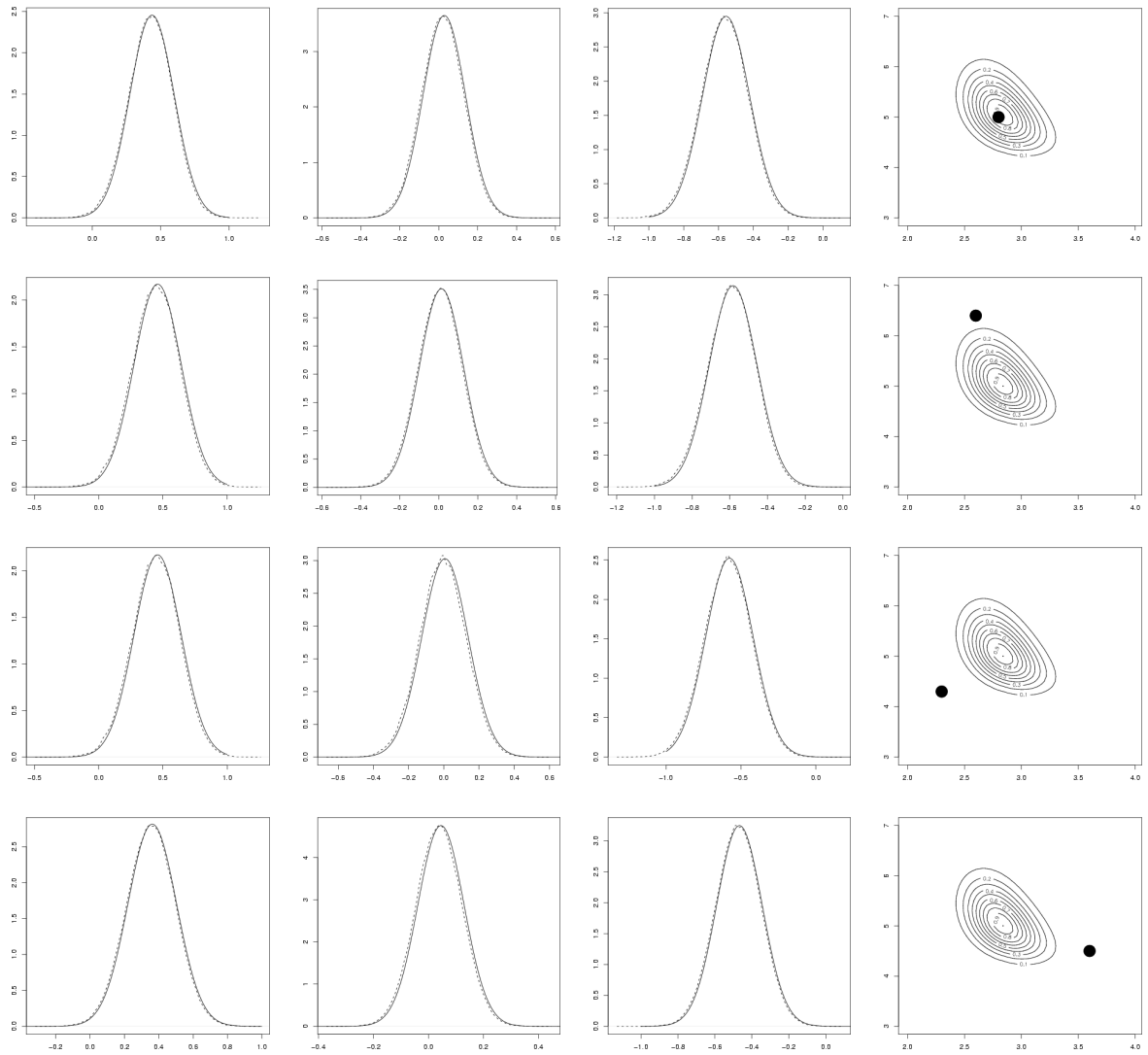
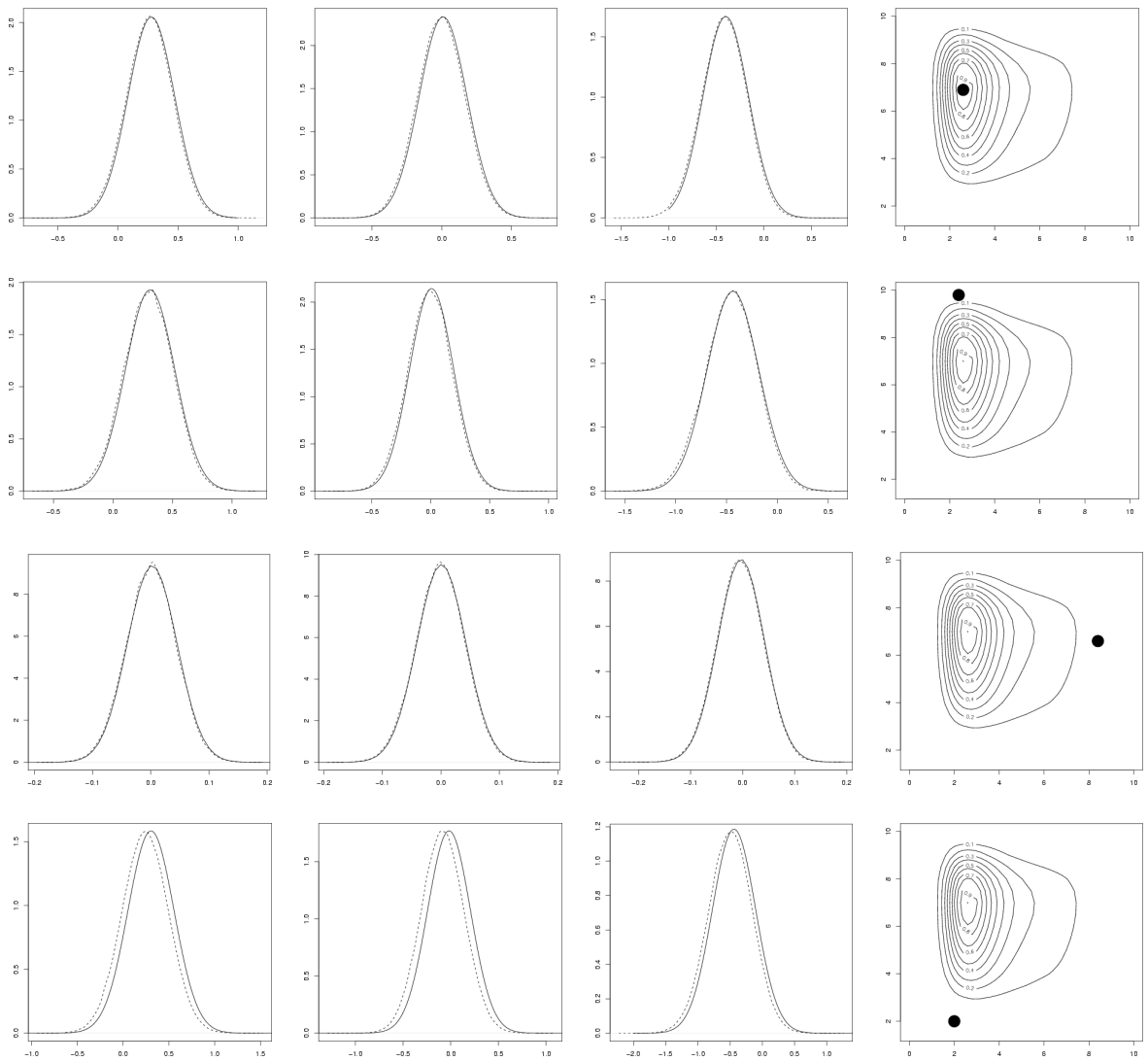
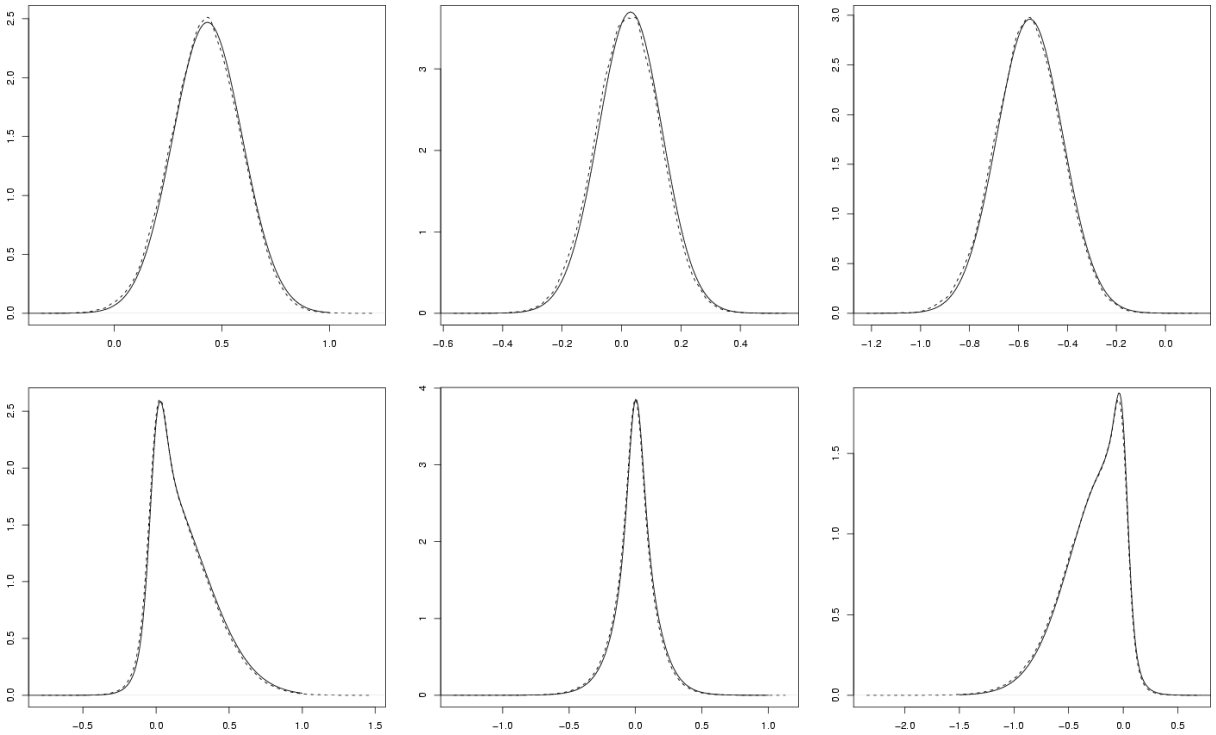


Figure 3: Results for the Easy-case. Each row shows in (d) the location of the fixed $\boldsymbol{\theta}$, and in the first three columns the (estimated) true marginal densities (dotted lines) for the spatial component at three different districts. The solid line displays the Gaussian approximation. The three districts in column (a) to (c) represent districts with (a) high, (b) intermediate, and (c) low value of the posterior expectation of u_i.



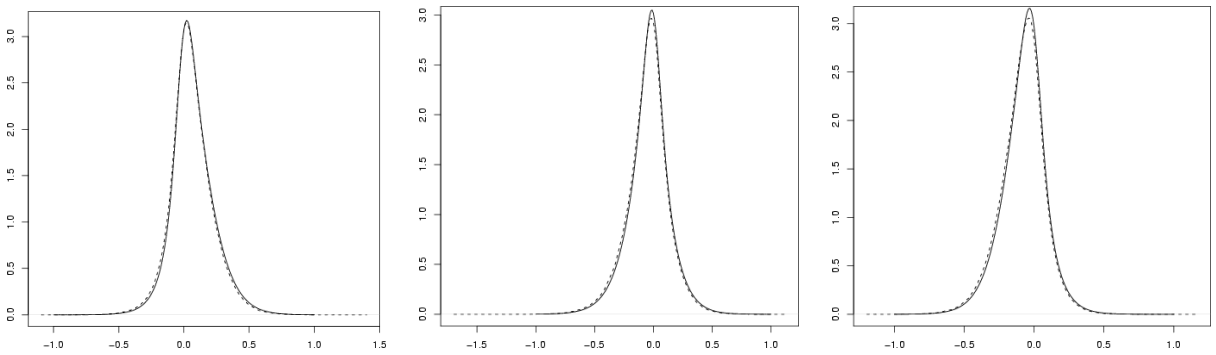
(a) (b) (c) (d)

Figure 4: Results for the Hard-case. Each row shows in (d) the location of the fixed $\boldsymbol{\theta}$, and in the first three columns the (estimated) true marginal densities (dotted lines) for the spatial component at three different districts. The solid line displays the Gaussian approximation. The three districts in column (a) to (c) represent districts with (a) high, (b) intermediate, and (c) low value of the posterior expectation of u_i.



(a) (b) (c)

Figure 5: Marginal posteriors for the spatial component in three districts. Easy-case on the top row and Hard-case on the bottom row. Columns (a) to (c) corresponds to the same columns in Figure 3 and Figure 4 for the top and bottom row, respectively. The approximations (8) are drawn with solid line and the (estimated) truth with dotted lines.



(a) (b) (c)

Figure 6: Marginal posteriors for the log-relative risk η_i in three districts for the Hard-case. Columns (a) to (c) corresponds to the same columns in Figure 4 and the bottom row in Figure 5. The approximations (8) are drawn with solid line and the (estimated) truth with dotted lines.

3.3.3 Marginal posteriors for the log-relative risk

We will now present the results for the marginal posteriors for the log-relative risk η_i for the Hard-case. It is not clear how the accuracy for these approximations should relate to those for the spatial component in Figure 5. It is η_i that is indirectly observed through y_i, but on the other hand, the difference between η_i and the spatial component u_i is only an additional unstructured component. The results are shown in Figure 6 for the same three districts shown in Figure 4 and in the last row of Figure 5. Again, the approximation (8) does not capture the right amount of skewness, for the same reason already discussed for Figure 3 and Figure 4. However, when $\boldsymbol{\theta}$ is integrated out, also the marginal posterior for $\boldsymbol{\eta}$ is quite well approximated.

3.4 Semi-parametric ecological regression

We will now consider an extension of the BYM-model (19) given by Natario and Knorr-Held (2003), which allows for adjusting the log-relative risk by a semi-parametric function of a covariate which is believed to influence the risk. The purpose of this example is to illustrate the ability of (8) to account for linear constraints, which we discuss in more details shortly. Similarly to Natario and Knorr-Held (2003), we will 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 their report for further details and background for this application.

The extension of the BYM-model is as follows. At the first stage we still assume $y_i \sim \text{Po}(e_i \exp(\eta_i))$ for each i, but now

$$\eta_i = u_i + v_i + f(c_i). \tag{21}$$

The two first terms are the spatially structured and unstructured term as in the BYM-model, whereas $f(c_i)$ is the effect of a covariate which has value c_i in district i. 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 $m = 100$ equally distant values of c_i, chosen to reflect the range of the covariate. We have scaled the covariates to the interval $[1, 100]$. The vector of $\boldsymbol{f} = (f_1, \ldots, f_m)^T$ is also a GMRF, with density

$$\pi(\boldsymbol{f} \mid \kappa_{\boldsymbol{f}}) \propto \kappa_{\boldsymbol{f}}^{(m-2)/2} \exp\left(-\frac{\kappa_{\boldsymbol{f}}}{2} \sum_{j=2}^{m} (f_j - 2f_{j-1} + f_{j-2})^2\right) \tag{22}$$

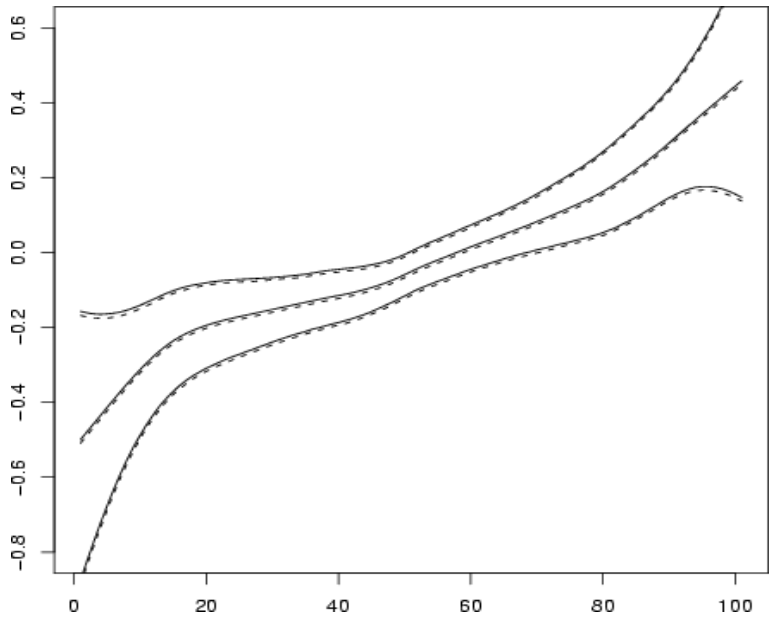


Figure 7: Marginal posteriors for the covariate effect, here represented by the mean, the 0.025 and 0.095 quantile. The approximations (8) are drawn with solid lines and the (estimated) truth with dotted lines. The middle lines are the posterior mean, the lower curves are the 0.025 quantile and the upper curves are the 0.975 quantiles.

This is a so-called second order random walk (RW2) model with (unknown) precision $\kappa_{\boldsymbol{f}}$, see for example Rue and Held (2005, Ch. 3). The density (22) can be interpreted as an approximated Galerkin solution to the stochastic differential equation, $f''(t) = dW(t)/dt$, where $W(t)$ is the Wiener process (Lindgren and Rue, 2005). We further impose the constraint $\sum_i u_i = 0$ to separate out the effect of the covariate. Note that the extended BYM-model is still a hierarchical GMRF-model but now $\boldsymbol{x} = (\boldsymbol{\eta}, \boldsymbol{u}, \boldsymbol{f})^T$. It is easy to derive the corresponding precision matrix and posterior density, but we avoid it here.

Adding a semi-parametric effect of a covariate extends directly the BYM-model presented in Section 3.1. However, the fundamental change is not the addition of the extra hyperparameter $\kappa_{\boldsymbol{f}}$, but the introduction of the linear constraint imposed to separate out the effect of the covariate. We need to make use of the correction in Section 2.2 to adjust marginal variances for the constraint, moreover, we need to do constrained optimisation to locate the mode in order to compute the GMRF-approximations. Both tasks are easily done with GMRFs and a few constraints do not slow down the computations.

We will now present the results focusing on the effect of the covariate. The other marginal posteriors are, in fact, similar to those presented in Section 3.2 and Section 3.3. The unknown precisions were all assigned Gamma-priors with parameters $a = 1$ and $b = 0.00005$ following Natario and Knorr-Held (2003). Figure 7 shows the approximated marginal posterior for $\boldsymbol{f}$, represented by the mean, the 0.025, and 0.975 quantile. The approximations (8) are drawn with solid lines and the (estimated) truth with dotted lines. The middle lines are the posterior mean, the lower curves are the 0.025 quantile and the upper curves are the 0.975 quantiles. The results show that the approximation is quite accurate. However, the approximation (8) does not capture the correct skewness, in a similar way to the last column in Figure 4. This claim is also verified by comparing the marginal posteriors for each f_j (not shown).

4 Discussion

In this report we have investigated how marginal posterior densities can be approximated using the GMRF-approximation in (8). We apply the GMRF-approximation to the full conditional for the latent GMRF component in hierarchical GMRF models. We use this to approximate both marginal posteriors for the hyperparameters and marginal posteriors for the components of the latent GMRF itself. We have also discussed how to compute marginal variances for GMRFs with and without linear constraints, and derived the recursion from a statistical point of view. The main motivation for using approximations to estimate marginal posteriors, is *only* computational efficiency. Computations with GMRFs are very efficient using numerical methods for sparse matrices, and make it possible to approximate posterior marginals nearly instant compared to the time required by MCMC algorithms. This makes the class of hierarchical GMRF-models a natural candidate for nearly instant approximated inference. The approximations were verified against very long runs of a one-block MCMC algorithm, with the following conclusions.

- The results were indeed positive in general and we obtained quite accurate approximations for all marginals investigated. Even for a quite hard dataset with low Poisson counts, the approximations were quite accurate.
- All results failed to capture the correct amount of (small) skewness, whereas the mode and the width of the density were more accurately approximated. However, the lack of skewness is a consequence of using symmetric approximations.

The range of application of these findings is, to our point of view, not only restricted to the class of BYM-models considered here but can be extended to many hierarchical GMRF-models. In particular, we want to mention hierarchical models based on log-Gaussian Cox processes (Møller et al., 1998) and model-based Geostatistics (Diggle et al., 1998). Both these popular model-classes can be considered as hierarchical GMRF-models, where Gaussian fields can be replaced by GMRFs using the results of Rue and Tjelmeland (2002), or sometimes better, using intrinsic GMRFs. The typical feature of these models is that the number of observations N is quite small. The approximation techniques we have presented, will give at least as accurate results than those presented in this paper. Another feature of these models is that, working with Gaussian fields directly, MCMC based inference is indeed challenging to implement and computationally heavy. For these reasons, the ability to use GMRFs and nearly instant approximated inference is indeed a huge step forward. All these results will be reported elsewhere.

Our approach to compute marginal posteriors is based on GMRF-approximations and the accuracy depends on the accuracy of the GMRF-approximation. Although this approximation is sufficiently accurate for many and often typical examples, is not difficult to find cases where such an approximation is not accurate enough, see for example Figure 4 last row. An important task for future work, is to construct methods that can go beyond the GMRF-approximation allowing for non-Gaussian approximations to the full conditional. One such class of approximation was introduced by Rue et al. (2004). This approximation can be applied to compute marginals as well. Preliminary results in this direction are indeed encouraging, and we are confident that improved approximation methods can be constructed without too much extra effort. These improved approximations will also serve as a validation procedure for the class of approximations considered here. They may, in fact, be used to detect if the approximations based on the GMRF-approximation are sufficiently accurate.

It is quite fast to compute our approximations even with our brute-force approach for integrating out the hyperparameters. This step can and need to be improved. This will increase the speed significantly while keeping the results nearly unchanged. There is a natural limit to the number of hyperparameters $\boldsymbol{\theta}$ our approach can deal with. Since we integrate out these numerically, we would

like $\dim(\boldsymbol{\theta}) \leq 3$. However, approximated schemes are indeed possible for higher dimensions as well, although we admit that we do not have large experience in this direction. Automatic construction of numerical quadrature rules based on the behaviour near the mode, is also a possibility which we will investigate. The benefit here, is that the numerical integration is adaptive which is also a requirement for constructing black-box algorithms for approximating marginal posteriors for hierarchical GMRF-models.

The results presented in this article imply that for many (Bayesian) hierarchical GMRF-models, namely those with a small number of hyperparameters, at least, MCMC algorithms are not needed to achieve accurate estimations of marginal posteriors. Moreover, approximated inference can be computed nearly instant compared to MCMC algorithms. This does not imply that MCMC algorithms are not needed, only that they are not needed in all cases.

References

Bernardinelli, L., Pascutto, C., Best, N. G., and Gilks, W. R. (1997). Disease mapping with errors in covariates. *Statistics in Medicine*, (16):741–752.

Besag, J. (1974). Spatial interaction and the statistical analysis of lattice systems (with discussion). *Journal of the Royal Statistical Society, Series B*, 36(2):192–225.

Besag, J. (1975). Statistical analysis of non-lattice data. *The Statistician*, 24(3):179–195.

Besag, J. (1981). On a system of two-dimentional recurrence equations. *Journal of the Royal Statistical Society, Series B*, 43(3):302–309.

Besag, J., York, J., and Mollié, A. (1991). Bayesian image restoration with two applications in spatial statistics (with discussion). *Annals of the Institute of Statistical Mathematics*, 43(1):1–59.

Diggle, P. J., Tawn, J. A., and Moyeed, R. A. (1998). Model-based geostatistics (with discussion). *Journal of the Royal Statistical Society, Series C*, 47(3):299–350.

Erisman, A. M. and Tinney, W. F. (1975). On computing certain elements of the inverse of a sparse matrix. *Communications of the ACM*, 18(3):177–179.

Gelfand, A. E., Sahu, S. K., and Carlin, B. P. (1995). Efficient parameterisations for normal linear mixed models. *Biometrika*, 82(3):479–488.

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

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

Lindgren, F. and Rue, H. (2005). A note on the second order random walk model for irregular locations. Statistics Report No. 6, Department of Mathematical Sciences, Norwegian University of Science and Technology, Trondheim, Norway.

Møller, J., Syversveen, A. R., and Waagepetersen, R. P. (1998). Log Gaussian Cox processes. *Scandinavian Journal of Statistics*, 25:451–482.

Natario, I. and Knorr-Held, L. (2003). Non-parametric ecological regression and spatial variation. *Biometrical Journal*, 45:670–688.

Papaspiliopoulos, O., Roberts, G. O., and Sköld, M. (2003). Non-centered parameterizations for hierarchical models and data augmentation (with discussion). In *Bayesian Statistics, 7*, pages 307–326. Oxford Univ. Press, New York.

Rue, H. (2001). Fast sampling of Gaussian Markov random fields. *Journal of the Royal Statistical Society, Series B*, 63(2):325–338.

Rue, H. and Held, L. (2005). *Gaussian Markov Random Fields: Theory and Applications*, volume 104 of *Monographs on Statistics and Applied Probability*. Chapman & Hall, London.

Rue, H., Steinsland, I., and Erland, S. (2004). Approximating hidden Gaussian Markov random fields. *Journal of the Royal Statistical Society, Series B*, 66(4):877–892.

Rue, H. and Tjelmeland, H. (2002). Fitting Gaussian Markov random fields to Gaussian fields. *Scandinavian Journal of Statistics*, 29(1):31–50.

Takahashi, K., Fagan, J., and Chen, M. S. (1973). Formation of a sparse bus impedance matrix and its application to short circuit study. In *8th PICA Conference proceedings*, pages 63–69. IEEE Power Engineering Society. Papers presented at the 1973 Power Industry Computer Application Conference in Minneapolis, Minnesota.

Tierney, L. and Kadane, J. B. (1986). Accurate approximations for posterior moments and marginal densities. *Journal of the American Statistical Association*, 81(393):82–86.

Wilkinson, D. J. (2003). Discussion to "Non-centered parameterizations for hierarchical models and data augmentation" by O. Papaspiliopoulos, G. O. Roberts and M. Sköld. In *Bayesian Statistics, 7*, pages 323–324. Oxford Univ. Press, New York.