

Approximate Bayesian Inference for Latent Gaussian Models Using Integrated Nested Laplace Approximations

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

Nicolas Chopin
CREST-LS and ENSAE, Paris
France

February 9, 2007

Abstract

We are concerned with Bayesian inference for latent Gaussian models, that is models involving a Gaussian latent field (in a broad sense), controlled by few parameters. This is perhaps the class of models most commonly encountered in applications: the latent Gaussian field can represent, for instance, a mix of smoothing splines or smooth curves, temporal and spatial processes. Hence, popular smoothing-spline models, state-space models, semiparametric regression, spatial and spatio-temporal models, log-Gaussian Cox-processes, and geostatistical models, all fall in this category.

We consider the case where the observational model is non-Gaussian, so that the posterior marginals are not available in closed form. Prominent examples are Poisson and Binomial count data. For such models, Markov chain Monte Carlo methods can be implemented, but they are not without problems, both in terms of convergence and computational time. In some practical applications, the extent of these problems is such that Markov chain Monte Carlo is simply non feasible.

We show that, by using an integrated nested Laplace approximation and its simplified version, we can directly compute very accurate approximations to the posterior marginals. The main benefit of these approximations is computational: where MCMC algorithms need hours and days to run, our approximations provide more precise estimates in seconds and minutes. Another advantage is their ease of use, which should facilitate and automate the analysis of data generated from latent Gaussian models.

KEYWORDS: Approximate Bayesian inference, Gaussian Markov random fields, Hierarchical GMRF-models, Laplace approximation, Numerical methods for sparse matrices, Parallel computing

AMS SUBJECT CLASSIFICATION:: Primary 62F15; secondary 62H99

ADDRESS FOR CORRESPONDENCE: H. Rue, Department of Mathematical Sciences, The Norwegian University for Science and Technology, N-7491 Trondheim, Norway. Email: hrue@math.ntnu.no, WWW-address: <http://www.math.ntnu.no/~hrue>, Voice: +47-7359-3533, Fax: +47-7359-3524.

1 Introduction

1.1 Latent Gaussian models

Latent Gaussian models are widely used in Bayesian analysis. Such models assume a latent Gaussian field $\mathbf{x} = (x_1, \dots, x_n)^T$, which is observed pointwise through n_d conditional independent data \mathbf{y} . In its simplest form, the covariance matrix of the latent Gaussian field and the likelihood are governed by a few parameters $\boldsymbol{\theta} = (\theta_1, \dots, \theta_m)^T$, say $m \leq 6$. Linear constraints of the form $\mathbf{A}\mathbf{x} = \mathbf{e}$, where the matrix \mathbf{A} has rank k , may also be imposed. The posterior then reads

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

In this paper, we assume that the main goal of the inference is to compute all, or some of, the n posterior marginals for x_i plus possibly the posterior marginals for $\boldsymbol{\theta}$ or some θ_j . If needed, the marginal densities can be post-processed to compute posterior expectations, variances, quantiles etc. We are concerned with the case where $\pi(y_i \mid x_i, \boldsymbol{\theta})$ is well-behaved, albeit non-Gaussian, so that the posterior marginals $\pi(x_i \mid \mathbf{y})$ and $\pi(\theta_j \mid \mathbf{y})$ are not available in closed form.

A few examples will demonstrate the wide use of latent Gaussian models. We loosely classify them with respect to their ‘physical dimension’, like 1D, 2D and 3D. In 1D, the latent process is often a mix of unstructured Gaussian effects and smooth processes in continuous or discrete ‘time’, such as integrated Wiener processes or random walk models. These can be used in a temporal context in various applications (Wecker and Ansley, 1983; Carter and Kohn, 1994; Fahrmeir and Tutz, 2001; Kitagawa and Gersch, 1996; Durbin and Koopman, 2000), or to model semiparametrically the effect of covariates in a regression setup (Lang and Brezger, 2004; Biller and Fahrmeir, 1997). In 2D, typical examples are model-based geostatistics (Diggle et al., 1998; Diggle and Ribeiro, 2006), and more generic smoothing models similar to the well-known BYM model for disease mapping (Besag et al., 1991; Weir and Pettitt, 2000), see also Banerjee et al. (2004) for many more examples. Models for spatial log-Gaussian Cox processes (Møller et al., 1998) are also in this class. Spatial models can also include 1D structures, like splines which model various covariate effects, see for example Natario and Knorr-Held (2003) and Fahrmeir and Lang (2001). 3D examples are usually an extension of a spatial model to a temporal or depth dimension, e.g. Allcroft and Glasbey (2003); Carlin and Banerjee (2003); Knorr-Held (2000); Knorr-Held and Besag (1998) and Wikle et al. (1998).

1.2 Inference: MCMC approaches

The common approach to inference for latent Gaussian models is Markov chain Monte Carlo (MCMC). It is well known however that MCMC tends to exhibit poor performance when applied to such models. Various factors explain this. First, the components of the latent field \mathbf{x} are strongly dependent on each other. Second, $\boldsymbol{\theta}$ and \mathbf{x} are also strongly dependent, especially when n is large. A common approach to (try to) overcome this first problem, is to construct a joint proposal based on a Gaussian approximation to the full conditional of \mathbf{x} (Gamerman, 1997, 1998; Carter and Kohn, 1994; Knorr-Held, 1999; Knorr-Held and Rue, 2002; Rue et al., 2004). The second problem requires, at least partially, a joint update of both $\boldsymbol{\theta}$ and \mathbf{x} . One suggestion is to (try to) use the one-block approach of Knorr-Held and Rue (2002): make a proposal for $\boldsymbol{\theta}$ to $\boldsymbol{\theta}'$, update \mathbf{x} from the Gaussian approximation conditional on $\boldsymbol{\theta}'$, then accept/reject jointly; see Rue and Held (2005, Ch. 4) for variations on this approach. Some models can alternatively be reparameterised to overcome the second problem (Papaspiliopoulos et al., 2007). Independence samplers can also sometimes be constructed (Rue et al., 2004). For some (observational) models, auxiliary variables can be introduced to simplify the construction of Gaussian approximations (Shephard, 1994; Albert and Chib, 1993; Holmes and Held, 2006; Frühwirth-Schnatter et al., 2006; Frühwirth-Schnatter and Frühwirth, 2007; Rue and Held, 2005). Despite all these developments, MCMC remains painfully slow from the end user’s point of view.

1.3 Inference: Deterministic approximations

Gaussian approximations play a central role in the development of more efficient MCMC algorithms. This remark leads to the following questions:

- Can we bypass MCMC entirely, and base our inference solely on such closed-form approximations?
- To which extent can we advocate an approach that leads to a (presumably) small approximation error over another approach giving rise to a (presumably) large MCMC error?

Obviously, MCMC errors seem preferable, as they can be made arbitrarily small, for arbitrarily large computational time. We argue however that, for a given computational cost, the deterministic approach developed in this paper outperforms MCMC algorithms to such an extent that, for latent Gaussian models, resorting to MCMC rarely makes sense in practice.

It is useful to provide some orders of magnitude. In typical spatial examples where the dimension n is a few thousands, our approximations for all the posterior marginals can be computed in (less than) a minute or a few minutes. The corresponding MCMC samplers need hours or even days to compute accurate posterior marginals. The approximation bias is in typical examples much less than the MCMC error and negligible in practice. More formally, on one hand it is well-known that MCMC is a last resort solution: Monte Carlo averages are characterised by additive $\mathcal{O}_p(N^{-1/2})$ errors, where N is the simulated sample size. Thus, it is easy to get rough estimates, but nearly impossible to get accurate ones; an additional correct digit requires 100 times more computational power. More importantly, the implicit constant in $\mathcal{O}_p(N^{-1/2})$ often hides a curse of dimensionality with respect to the dimension n of the problem, which explains the practical difficulties with MCMC mentioned above. On the other hand, Gaussian approximations are intuitively appealing for latent Gaussian models. For most real problems and datasets, the conditional posterior of \mathbf{x} is typically well-behaved, and looks ‘almost’ Gaussian. This is clearly due to the latent Gaussian prior assigned to \mathbf{x} , which has a non-negligible impact on the posterior, especially in terms of dependence between the components of \mathbf{x} .

1.4 Inference: The new approach

Our approach is based on the following approximation $\tilde{\pi}(\boldsymbol{\theta}|\mathbf{y})$ for the marginal posterior of $\boldsymbol{\theta}$:

$$\tilde{\pi}(\boldsymbol{\theta} | \mathbf{y}) \propto \frac{\pi(\mathbf{x}, \boldsymbol{\theta}, \mathbf{y})}{\tilde{\pi}_G(\mathbf{x}|\boldsymbol{\theta}, \mathbf{y})} \Bigg|_{\mathbf{x}=\mathbf{x}^*(\boldsymbol{\theta})} \quad (1)$$

where $\tilde{\pi}_G(\mathbf{x}|\boldsymbol{\theta}, \mathbf{y})$ is the Gaussian approximation to the full conditional of \mathbf{x} , and $\mathbf{x}^*(\boldsymbol{\theta})$ is the mode of the full conditional for \mathbf{x} , for a given $\boldsymbol{\theta}$. The proportionality sign (1) comes from the fact that the normalising constant for $\pi(\mathbf{x}, \boldsymbol{\theta}|\mathbf{y})$ is unknown. This expression is equivalent to Tierney and Kadane (1986)’s Laplace approximation of a marginal posterior distribution and this suggests that the approximation error is relative and of order $\mathcal{O}(n_d^{-3/2})$ after renormalisation. However, since n is not fixed but depends on n_d , standard asymptotic assumptions usually invoked for Laplace expansions, see for example Schervish (1995, p. 453), are not verified here. We will discuss the error rate for this case in more detail in Section 4.

Note that $\tilde{\pi}(\boldsymbol{\theta}|\mathbf{y})$ itself tends to depart significantly from Gaussianity. This suggests that a cruder approximation based on a Gaussian approximation to $\pi(\boldsymbol{\theta}|\mathbf{y})$ is not accurate enough for our purposes; this also applies to similar approximations based on ‘equivalent Gaussian observations’ around \mathbf{x}^* , and evaluated at the mode of (1) (Breslow and Clayton, 1993; Ainsworth and Dean, 2006). A critical aspect of our approach is to explore and manipulate $\tilde{\pi}(\boldsymbol{\theta}|\mathbf{y})$ and $\tilde{\pi}(x_i|\mathbf{y})$ in a ‘nonparametric’ way.

Rue and Martino (2007) used (1) to approximate posterior marginals for $\boldsymbol{\theta}$ for various latent Gaussian models. Their conclusion was that $\tilde{\pi}(\boldsymbol{\theta}|\mathbf{y})$ is particularly accurate: even long MCMC runs could not detect any error in it. For the posterior marginals of the latent field, they proposed to start from $\tilde{\pi}_G(\mathbf{x}|\boldsymbol{\theta}, \mathbf{y})$ and approximate the density of $x_i|\boldsymbol{\theta}, \mathbf{y}$ with the Gaussian marginal derived from $\tilde{\pi}_G(\mathbf{x}|\boldsymbol{\theta}, \mathbf{y})$, i.e.

$$\tilde{\pi}(x_i | \boldsymbol{\theta}, \mathbf{y}) = \mathcal{N} \{x_i; \mu_i(\boldsymbol{\theta}), \sigma_i^2(\boldsymbol{\theta})\}. \quad (2)$$

Here, $\boldsymbol{\mu}(\boldsymbol{\theta})$ is the mean (vector) of the Gaussian approximation, whereas $\boldsymbol{\sigma}^2(\boldsymbol{\theta})$ is a vector of corresponding marginal variances. This approximation can be integrated numerically with respect to $\boldsymbol{\theta}$ using (1), to obtain approximations of the marginals of interest for the latent field,

$$\tilde{\pi}(x_i | \mathbf{y}) = \sum_k \tilde{\pi}(x_i | \boldsymbol{\theta}_k, \mathbf{y}) \times \tilde{\pi}(\boldsymbol{\theta}_k | \mathbf{y}) \times \Delta_k. \quad (3)$$

The sum is over values of $\boldsymbol{\theta}$ with area-weights Δ_k . Rue and Martino (2007) showed that the approximate posterior marginals for $\boldsymbol{\theta}$ were accurate, while the error in the Gaussian approximation (2) was higher. In particular, (2) can present a slight error in location and/or a lack of skewness. Another issue in Rue and Martino (2007) was the difficulty to detect the x_i 's whose approximation is less accurate. Friel and Rue (2007) made use of similar ideas to perform approximate Bayesian inference for factorisable models (in particular, binary Markov random fields) that allow for recursive computing (Bartolucci and Besag, 2002; Reeves and Pettitt, 2004).

In this paper, we solve all the remaining issues in Rue and Martino (2007), and present a fully automatic approach for approximate inference in latent Gaussian models which we name *Integrated Nested Laplace Approximations* (INLA). The main tool is to apply the Laplace approximation once more, this time to $\pi(x_i | \mathbf{y}, \boldsymbol{\theta})$. We also present a faster alternative which corrects the Gaussian approximation (2) for error in the location and lack of skewness at moderate extra cost. The corrections are obtained by a series expansions of the Laplace approximation. This faster alternative is a natural first choice, because of its low computational cost and high accuracy. It is our experience that INLA outperforms without comparison any MCMC alternative, both in terms of accuracy and computational speed. We also derive tools for assessing the approximation error.

Most of the latent fields in the literature admit conditional independence properties, hence the latent field \mathbf{x} is a Gaussian Markov random field (GMRF). Thus, we base INLA on sparse matrix calculations, which are much quicker than dense matrix calculations, see Section 2. An exception are geostatistical models, but fast approximate inference is still possible in this case, using a different approach (Eidsvik et al., 2006), or combining the INLA approach with GMRF-proxies to Gaussian fields (Rue and Tjelmeland, 2002).

1.5 Plan of paper

Section 2 contains preliminaries on GMRFs, sparse matrix computations and Gaussian approximations. Section 3 explains how to approximate $\pi(\boldsymbol{\theta} | \mathbf{y})$ and $\pi(x_i | \boldsymbol{\theta}, \mathbf{y})$, using the Integrated nested Laplace approximation (INLA) approach. For the latter distributions, three approximations are discussed: Gaussian, Laplace and simplified Laplace. Section 4 discusses the error rates of the Laplace approximations used in INLA. Section 5 illustrates the performance of INLA through simulated and real examples, which include multiscale analysis of non-Gaussian time-series data, stochastic volatility models, spatial semi-parametric ecological regression and spatial log-Gaussian Cox processes. Section 6 discuss two extensions: approximations of the marginal likelihood and an alternative integration scheme for cases where the number of hyperparameters is not small but moderate. We end with a general discussion in Section 7.

2 Preliminaries

We present here basic properties of GMRFs, and explain how to perform related computations using sparse matrix algorithms. We then discuss how to compute Gaussian approximations for a latent GMRF. See Rue and Held (2005) for more details on both issues. Denote by \mathbf{x}_{-i} the vector \mathbf{x} minus its i th element, by $\mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ the Gaussian distribution, and by $\mathcal{N}(\mathbf{x}; \boldsymbol{\mu}, \boldsymbol{\Sigma})$, the Gaussian density at configuration \mathbf{x} .

2.1 Gaussian Markov Random Fields

A GMRF is a Gaussian random variable $\mathbf{x} = (x_1, \dots, x_n)$ with Markov properties: for some $i \neq j$'s, x_i and x_j are independent conditional upon \mathbf{x}_{-ij} . These Markov properties are conveniently encoded in the precision (inverse covariance) matrix \mathbf{Q} : $Q_{ij} = 0$ if and only if x_i and x_j are independent conditional upon \mathbf{x}_{-ij} . Let

the undirected graph \mathcal{G} denote the conditional independence properties of \mathbf{x} , then \mathbf{x} is said to be a GMRF with respect to \mathcal{G} . If the mean of \mathbf{x} is $\boldsymbol{\mu}$, the density of \mathbf{x} is

$$\pi(\mathbf{x}) = (2\pi)^{-n/2} |\mathbf{Q}|^{1/2} \exp \left\{ -\frac{1}{2} (\mathbf{x} - \boldsymbol{\mu})^T \mathbf{Q} (\mathbf{x} - \boldsymbol{\mu}) \right\}. \quad (4)$$

In most cases only $\mathcal{O}(n)$ of the n^2 entries of \mathbf{Q} are non-zero, so \mathbf{Q} is sparse. This allows for fast factorisation of \mathbf{Q} as $\mathbf{L}\mathbf{L}^T$, where \mathbf{L} is the (lower) Cholesky triangle. The sparseness of \mathbf{Q} is inherited into \mathbf{L} , thanks to the global Markov property: for $i < j$, such that i and j are separated by $F(i, j) = \{i + 1, \dots, j - 1, \dots, j + 1, \dots, n\}$ in \mathcal{G} , $L_{ji} = 0$. Thus, only non-null terms in \mathbf{L} are computed. In addition, nodes can be re-ordered to decrease the number of non-zero terms in \mathbf{L} . The typical cost of factorising \mathbf{Q} into $\mathbf{L}\mathbf{L}^T$ is $\mathcal{O}(n)$ for 1D, $\mathcal{O}(n^{3/2})$ for 2D and $\mathcal{O}(n^2)$ for 3D GMRFs.

Solving equations which involve \mathbf{Q} also makes use of the Cholesky triangle. For example, $\mathbf{Q}\mathbf{x} = \mathbf{b}$ is solved in two steps. First solve $\mathbf{L}\mathbf{v} = \mathbf{b}$, then solve $\mathbf{L}^T\mathbf{x} = \mathbf{v}$. If $\mathbf{z} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ then the solution of $\mathbf{L}^T\mathbf{x} = \mathbf{z}$ has precision matrix \mathbf{Q} . This is the general method for producing random samples from a GMRF. The log density at any \mathbf{x} , $\log \pi(\mathbf{x})$, can easily be computed using (4) since $\log |\mathbf{Q}| = 2 \sum_i \log L_{ii}$.

Marginal variances can also be computed efficiently. To see this, we can start with the equation $\mathbf{L}^T\mathbf{x} = \mathbf{z}$ where $\mathbf{z} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$. Recall that the solution \mathbf{x} has precision matrix \mathbf{Q} . Writing this equation out in detail, we obtain $L_{ii}x_i = z_i - \sum_{k=i+1}^n L_{ki}x_k$ for $i = n, \dots, 1$. Multiplying each side with x_j $j \geq i$, and taking expectation, we obtain

$$\Sigma_{ij} = \delta_{ij} / L_{ii}^2 - \frac{1}{L_{ii}} \sum_{k=i+1}^n L_{ki} \Sigma_{kj}, \quad j \geq i, \quad i = n, \dots, 1, \quad (5)$$

where $\Sigma (= \mathbf{Q}^{-1})$ is the covariance matrix. Thus Σ_{ij} can be computed from (5), letting the outer loop i run from n to 1 and the inner loop j from n to i . If we are only interested in the marginal variances, we only need to compute Σ_{ij} 's for which L_{ji} (or L_{ij}) is not known to be zero, see above. This reduce the computational costs to typically $\mathcal{O}(n(\log n)^2)$ in the spatial case; see Rue and Martino (2007, Sec. 2) for more details.

When the GMRF is defined with additional linear constraints, like $\mathbf{A}\mathbf{x} = \mathbf{e}$ for a $k \times n$ matrix \mathbf{A} of rank k , the following strategy is used: if \mathbf{x} is a sample from the unconstrained GMRF, then

$$\mathbf{x}^c = \mathbf{x} - \mathbf{Q}^{-1} \mathbf{A}^T (\mathbf{A} \mathbf{Q}^{-1} \mathbf{A}^T)^{-1} (\mathbf{A} \mathbf{x} - \mathbf{e}) \quad (6)$$

is a sample from the constrained GMRF. The expected value of \mathbf{x}^c can also be computed using (6). This approach is commonly called ‘conditioning by Kriging’, see Cressie (1993) or Rue (2001). Note that $\mathbf{Q}^{-1} \mathbf{A}^T$ is computed by solving k linear systems, one for each column of \mathbf{A}^T . The additional cost of the k linear constraints is $\mathcal{O}(nk^2)$. Marginal variances under linear constraints can be computed in a similar way, see Rue and Martino (2007, Sec. 2).

2.2 Gaussian Approximations

Our approach is based on Gaussian approximations to densities of the form:

$$\pi(\mathbf{x}) \propto \exp \left\{ -\frac{1}{2} \mathbf{x}^T \mathbf{Q} \mathbf{x} + \sum_{i \in \mathcal{I}} g_i(x_i) \right\}. \quad (7)$$

where $g_i(x_i)$ is $\log \pi(y_i | x_i, \boldsymbol{\theta})$ in our settings. The Gaussian approximation $\tilde{\pi}_G(\mathbf{x})$ is obtained by matching the modal configuration and the curvature at the mode. The mode is computed iteratively. Let $\boldsymbol{\mu}^{(0)}$ be the initial guess, and expand $g_i(x_i)$ around $\mu_i^{(0)}$ to the second order,

$$g_i(x_i) \approx g_i(\mu_i^{(0)}) + b_i x_i - \frac{1}{2} c_i x_i^2 \quad (8)$$

where $\{b_i\}$ and $\{c_i\}$ depend on $\boldsymbol{\mu}^{(0)}$. A Gaussian approximation is obtained, with precision matrix $\mathbf{Q} + \text{diag}(\mathbf{c})$ and mode given by the solution of $(\mathbf{Q} + \text{diag}(\mathbf{c}))\boldsymbol{\mu}^{(1)} = \mathbf{b}$. This process is repeated until it converges to a Gaussian distribution with, say, mean \mathbf{x}^* and precision matrix $\mathbf{Q}^* = \mathbf{Q} + \text{diag}(\mathbf{c})$. If there are linear constraints, the mean is corrected at each iteration using the expected value of (6).

Since the non-quadratic term in (7) is only a function of x_i and not a function of x_i and x_j , say, the precision matrix of the Gaussian approximation is of the form $\mathbf{Q} + \text{diag}(\mathbf{c})$. This is computationally convenient, as the Markov properties of the GMRF are preserved.

Density (7) may seem restrictive: a more complex density is obtained if, say, y_1 depends on the sum $x_1 + x_2$. This happens for example when the observations are a blurred version of the latent field. In such a case, we find it most convenient to alter the latent field: \mathbf{x} is augmented with x_{n+1} , where x_{n+1} is $x_1 + x_2$ plus a tiny Gaussian noise; then y_1 depends on x_{n+1} only, and (7) applies.

3 The Integrated Nested Laplace approximation (INLA)

In this section we present the INLA approach for approximating the posterior marginals of the latent Gaussian field, $\pi(x_i|\mathbf{y})$, $i = 1, \dots, n$. The approximation is computed in three steps. The first step (Section 3.1) approximates the posterior marginal of $\boldsymbol{\theta}$ using the Laplace approximation (1). The second step (Section 3.2) computes the Laplace approximation, or the simplified Laplace approximation, of $\pi(x_i|\mathbf{y}, \boldsymbol{\theta})$, for selected values of $\boldsymbol{\theta}$, in order to improve on the Gaussian approximation (2). The third step combines the previous two using numerical integration (3).

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

The first step of the INLA approach is to compute our approximation to the posterior marginal of $\boldsymbol{\theta}$, see (1). The denominator in (1) is the Gaussian approximation to the full conditional for \mathbf{x} , and is computed as described in Section 2.2. The main use of $\tilde{\pi}(\boldsymbol{\theta}|\mathbf{y})$ is to integrate out the uncertainty with respect to $\boldsymbol{\theta}$ when approximating the posterior marginal of x_i , see (3). For this task, we do not need to represent $\tilde{\pi}(\boldsymbol{\theta}|\mathbf{y})$ parametrically, but rather to explore it sufficiently well to be able to select good evaluation points for the numerical integration (3). At the end of this section, we discuss how the posterior marginals $\pi(\theta_j|\mathbf{y})$ can be approximated.

Assume for simplicity that $\boldsymbol{\theta} = (\theta_1, \dots, \theta_m) \in \mathbb{R}^m$, which can always be obtained by reparametrisation;

Step 1 Locate the mode of $\tilde{\pi}(\boldsymbol{\theta}|\mathbf{y})$, by optimising $\log \tilde{\pi}(\boldsymbol{\theta}|\mathbf{y})$ with respect to $\boldsymbol{\theta}$. This can be done using some quasi-Newton method which builds up an approximation to the second derivatives of $\log \tilde{\pi}(\boldsymbol{\theta}|\mathbf{y})$ using the difference between successive gradient vectors. The gradient is approximated using finite differences. Let $\boldsymbol{\theta}^*$ be the modal configuration.

Step 2 At the modal configuration $\boldsymbol{\theta}^*$ compute the negative Hessian matrix $\mathbf{H} > 0$, using finite differences. Let $\boldsymbol{\Sigma} = \mathbf{H}^{-1}$, which is the covariance matrix for $\boldsymbol{\theta}$ if the density were Gaussian. To aid the exploration, use standardised variables \mathbf{z} instead of $\boldsymbol{\theta}$: let $\boldsymbol{\Sigma} = \mathbf{V}\boldsymbol{\Lambda}\mathbf{V}^T$ be the eigen-decomposition of $\boldsymbol{\Sigma}$, and define $\boldsymbol{\theta}$ via \mathbf{z} , as follows

$$\boldsymbol{\theta}(\mathbf{z}) = \boldsymbol{\theta}^* + \mathbf{V}\boldsymbol{\Lambda}^{1/2}\mathbf{z}. \quad (9)$$

If $\tilde{\pi}(\boldsymbol{\theta}|\mathbf{y})$ is a Gaussian density, then \mathbf{z} is $\mathcal{N}(\mathbf{0}, \mathbf{I})$. This reparametrisation corrects for scale and rotation.

Step 3 Explore $\log \tilde{\pi}(\boldsymbol{\theta}|\mathbf{y})$ using the \mathbf{z} -parametrisation. Figure 1 illustrates the procedure when $\log \tilde{\pi}(\boldsymbol{\theta}|\mathbf{y})$ is unimodal. Panel (a) shows a contour plot of $\log \tilde{\pi}(\boldsymbol{\theta}|\mathbf{y})$ for $m = 2$. Panel (a) also displays the location of the mode and the new coordinate axis for \mathbf{z} . We want to explore $\log \tilde{\pi}(\boldsymbol{\theta}|\mathbf{y})$ in order to locate the bulk of the probability mass. The result of this procedure is displayed in panel (b). Each dot is a point where $\log \tilde{\pi}(\boldsymbol{\theta}|\mathbf{y})$ is considered as significant, and which is used in the numerical integration (3). Details are as follows. We start from the mode ($\mathbf{z} = \mathbf{0}$), and go in the positive direction of z_1 with step-length δ_z say $\delta_z = 1$, as long as

$$\log \tilde{\pi}(\boldsymbol{\theta}(\mathbf{0})|\mathbf{y}) - \log \tilde{\pi}(\boldsymbol{\theta}(\mathbf{z})|\mathbf{y}) < \delta_\pi \quad (10)$$

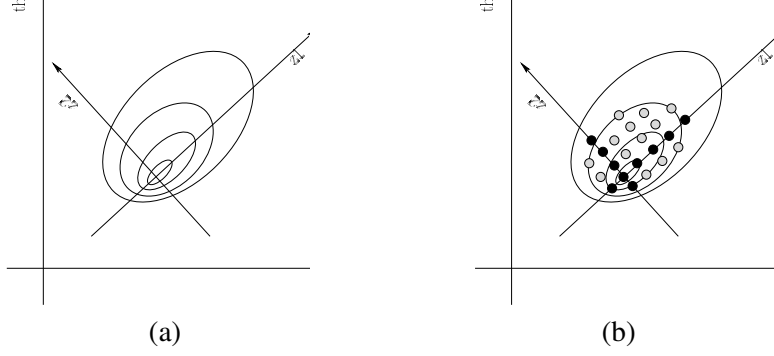


Figure 1: Illustration of the exploration of the posterior marginal for θ . In (a) the mode is located, the Hessian and the coordinate system for z are computed. In (b) each coordinate direction is explored (black dots) until the log-density drops below a certain limit. Finally the grey dots are explored.

where, for example $\delta_\pi = 2.5$. Then we switch direction and do similarly. The other coordinates are treated in the same way. This produces the black dots. We can now fill in all the intermediate values by taking all different combinations of the black dots. These new points (shown as grey dots) are included if (10) holds.

Since we layout the points θ_k in a regular grid, we take all the area-weights Δ_k in (3) to be equal.

Consider now the case where we want to compute the approximation for the posterior marginals for some or all the θ_j 's, $\tilde{\pi}(\theta_j|\mathbf{y})$. The rotation of the axis due to \mathbf{V} in (9) is inconvenient when summing out the remaining variables θ_{-j} . We can then replace the negative Hessian \mathbf{H} by its diagonal, in order to suppress the rotation while retaining the scaling.

3.2 Approximating $\pi(x_i|\theta, \mathbf{y})$

We have now a set of weighted points $\{\theta_k\}$ to be used in the integration (3). The next step is to provide accurate approximations for the posterior marginal for the x_i 's, conditioned on selected values of θ . We discuss three approximations $\tilde{\pi}(x_i|\mathbf{y}, \theta_k)$, that is the Gaussian, the Laplace, and a simplified Laplace approximation. Although the Laplace approximation is preferred in general, the much smaller cost of the simplified Laplace generally compensates for the slight loss in accuracy.

3.2.1 Using Gaussian Approximations

The simplest (and cheapest) approximation to $\pi(x_i|\theta, \mathbf{y})$ is the Gaussian approximation $\tilde{\pi}_G(x_i|\theta, \mathbf{y})$, where the mean $\mu_i(\theta)$ and the marginal variance $\sigma_i^2(\theta)$ are derived using the recursions (5), and possibly correcting for linear constraints. During the exploration of $\tilde{\pi}(\theta|\mathbf{y})$, see Section 3.1, we already compute $\tilde{\pi}_G(\mathbf{x}|\theta, \mathbf{y})$, so only marginal variances need to be additionally computed. The Gaussian approximation gives often reasonable results, but there can be errors in the location and/or errors due to the lack of skewness (Rue and Martino, 2007).

3.2.2 Using Laplace Approximations

The natural way to improve the Gaussian approximation is to compute the Laplace approximation

$$\tilde{\pi}_{\text{LA}}(x_i | \theta, \mathbf{y}) \propto \frac{\pi(\mathbf{x}, \theta, \mathbf{y})}{\tilde{\pi}_{\text{GG}}(\mathbf{x}_{-i}|x_i, \theta, \mathbf{y})} \Bigg|_{\mathbf{x}_{-i}=\mathbf{x}_{-i}^*(x_i, \theta)}. \quad (11)$$

Here, $\tilde{\pi}_{\text{GG}}$ is the Gaussian approximation to $\mathbf{x}_{-i}|x_i, \theta, \mathbf{y}$, and $\mathbf{x}_{-i}^*(x_i, \theta)$ is the modal configuration. Note that $\tilde{\pi}_{\text{GG}}$ is different from the conditional density corresponding to $\tilde{\pi}_G(\mathbf{x}|\theta, \mathbf{y})$.

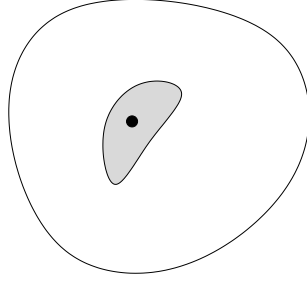


Figure 2: Illustration of the region of interest $R_i(\boldsymbol{\theta})$. The outer circle illustrates the graph of the GMRF, whereas the black dot indicates the node of interest. The conditional expectation (13) locates the nodes that are affected by a change in x_i , that is all the nodes in the grey region.

Unfortunately, (11) implies that $\tilde{\pi}_{\text{GG}}$ must be recomputed for each value of x_i and $\boldsymbol{\theta}$, since its precision matrix depends on i and $\boldsymbol{\theta}$. This is far too expensive, as it requires n factorisations of the full precision matrix. We propose two modifications to (11) which makes it computationally feasible.

Our first modification consists in avoiding the optimisation step in computing $\tilde{\pi}_{\text{GG}}(\mathbf{x}_{-i}|x_i, \boldsymbol{\theta}, \mathbf{y})$ by approximating the modal configuration,

$$\mathbf{x}_{-i}^*(x_i, \boldsymbol{\theta}) \approx \mathbf{E}_{\tilde{\pi}_{\text{G}}}(\mathbf{x}_{-i} | x_i). \quad (12)$$

The right-hand side is evaluated under the conditional density derived from the Gaussian approximation $\tilde{\pi}_{\text{G}}(\mathbf{x}|\boldsymbol{\theta}, \mathbf{y})$. The computational benefit is immediate. First, the conditional mean can be computed by a rank one update from the unconditional mean, using (6). In the spatial case the cost is $\mathcal{O}(n \log n)$, for each i , which comes from solving $\mathbf{Q}\mathbf{v} = \mathbf{1}_i$, where $\mathbf{1}_i$ equals one at position i , and zero otherwise. This rank one update is computed only once for each i , as it is linear in x_i . Although their settings are slightly different, Hsiao et al. (2004) show that deviating from the conditional mode does not necessarily degrade the approximation error. Another positive feature of using (12) is that the conditional mode is continuous with respect to x_i , a feature which does not hold in practice when numerical optimisation is used to compute $\mathbf{x}_{-i}^*(x_i, \boldsymbol{\theta})$.

Our next modification materialises the following intuition: only those x_j that are ‘close’ to x_i should have an impact on the marginal of x_i . Figure 2 illustrates this idea. The graph of \mathbf{x} is represented by the larger circle. The node i is marked with a black dot. If the dependency between x_j and x_i decays as the distance between nodes i and j increases, only those x_j ’s in the grey region are of interest regarding the marginal of x_i . Denote by $R_i(\boldsymbol{\theta})$ the ‘region of interest’ regarding the marginal of x_i . The conditional expectation in (12) implies that

$$\frac{\mathbf{E}_{\tilde{\pi}_{\text{G}}}(x_j|x_i) - \mu_j(\boldsymbol{\theta})}{\sigma_j(\boldsymbol{\theta})} = a_{ij}(\boldsymbol{\theta}) \frac{x_i - \mu_i(\boldsymbol{\theta})}{\sigma_i(\boldsymbol{\theta})} \quad (13)$$

for some $a_{ij}(\boldsymbol{\theta})$ when $j \neq i$. Hence, a simple rule for constructing the set $R_i(\boldsymbol{\theta})$ is

$$R_i(\boldsymbol{\theta}) = \{j : |a_{ij}(\boldsymbol{\theta})| > 0.001\}. \quad (14)$$

The most important computational saving using $R_i(\boldsymbol{\theta})$ comes from the calculation of the denominator of (11), where we now only need to factorise a $|R_i(\boldsymbol{\theta})| \times |R_i(\boldsymbol{\theta})|$ sparse matrix.

Expression (11), simplified as explained above, must be computed for different values of x_i in order to find the density. To select these points, we use the mean and variance of the Gaussian approximation (2), and choose, say, different values for the standardised variable

$$x_i^{(s)} = \frac{x_i - \mu_i(\boldsymbol{\theta})}{\sigma_i(\boldsymbol{\theta})} \quad (15)$$

according to the corresponding choice of abscissas given by the Gauss-Hermite quadrature rule. To represent the density $\tilde{\pi}_{\text{LA}}(x_i|\boldsymbol{\theta}, \mathbf{y})$, we use

$$\tilde{\pi}_{\text{LA}}(x_i | \boldsymbol{\theta}, \mathbf{y}) \propto \mathcal{N}\{x_i; \mu_i(\boldsymbol{\theta}), \sigma_i^2(\boldsymbol{\theta})\} \times \exp\{\text{cubic spline}(x_i)\}. \quad (16)$$

The cubic spline is fitted to the difference of the log-density of $\tilde{\pi}_{\text{LA}}(x_i|\boldsymbol{\theta}, \mathbf{y})$ and $\tilde{\pi}_{\text{G}}(x_i|\boldsymbol{\theta}, \mathbf{y})$ at the selected abscissa points, and then the density is normalised using quadrature integration.

3.2.3 Using a Simplified Laplace Approximation

In this section we derive a simplified Laplace approximation $\tilde{\pi}_{\text{SLA}}(x_i|\boldsymbol{\theta}, \mathbf{y})$ by doing a series expansion of $\tilde{\pi}_{\text{LA}}(x_i|\boldsymbol{\theta}, \mathbf{y})$ around $x_i = \mu_i(\boldsymbol{\theta})$. This allows us to correct the Gaussian approximation $\tilde{\pi}_{\text{G}}(x_i|\boldsymbol{\theta}, \mathbf{y})$ for location and skewness. For many observational models including the Poisson and the Binomial, these corrections are sufficient to obtain essentially correct posterior marginals. The benefit is purely computational: as most of the terms are common for all i , we can compute all the n marginals in only $\mathcal{O}(n^2 \log n)$ time.

Define

$$d_j^{(3)}(x_i, \boldsymbol{\theta}) = \left. \frac{\partial^3}{\partial x_j^3} \log \pi(y_j | x_j, \boldsymbol{\theta}) \right|_{x_j = \mathbb{E}_{\tilde{\pi}_{\text{G}}}(x_j|x_i)}$$

which we assume exists. The evaluation point is found from (13). The following trivial Lemma will be useful.

Lemma 1 *Let $\mathbf{x} = (x_1, \dots, x_n)^T \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma})$, then for all x_1*

$$-\frac{1}{2}(x_1, \mathbb{E}(\mathbf{x}_{-1}|x_1)^T) \boldsymbol{\Sigma}^{-1} \begin{pmatrix} x_1 \\ \mathbb{E}(\mathbf{x}_{-1}|x_1) \end{pmatrix} = -\frac{1}{2}x_1^2/\Sigma_{11}.$$

We expand the numerator and denominator of (11) around $x_i = \mu_i(\boldsymbol{\theta})$, using (12) and Lemma 1. Up to third order, we obtain

$$\begin{aligned} \log \pi(\mathbf{x}, \boldsymbol{\theta}, \mathbf{y}) \Big|_{\mathbf{x}_{-i} = \mathbb{E}_{\tilde{\pi}_{\text{G}}}(\mathbf{x}_{-i}|x_i)} &= -\frac{1}{2}(x_i^{(s)})^2 \\ &+ \frac{1}{6}(x_i^{(s)})^3 \sum_{j \in \mathcal{I} \setminus i} d_j^{(3)}(\mu_i(\boldsymbol{\theta}), \boldsymbol{\theta}) \{\sigma_j(\boldsymbol{\theta}) a_{ij}(\boldsymbol{\theta})\}^3 + \dots \end{aligned} \quad (17)$$

The first and second order terms give the Gaussian approximation, whereas the third order term provides a correction for skewness. Further, the denominator of (11) reduces to

$$\log \tilde{\pi}_{\text{GG}}(\mathbf{x}_{-i} | x_i, \boldsymbol{\theta}, \mathbf{y}) \Big|_{\mathbf{x}_{-i} = \mathbb{E}_{\tilde{\pi}_{\text{G}}}(\mathbf{x}_{-i}|x_i)} = \text{constant} + \frac{1}{2} \log |\mathbf{H} + \text{diag}\{c(x_i, \boldsymbol{\theta})\}| \quad (18)$$

where \mathbf{H} is the prior precision matrix of the GMRF with i th column and row deleted, and $c(x_i, \boldsymbol{\theta})$ is the vector of minus the second derivative of the log likelihood evaluated at $x_j = \mathbb{E}_{\tilde{\pi}_{\text{G}}}(x_j|x_i)$, see Section 2.2. Using that

$$d \log |\mathbf{H} + \text{diag}(\mathbf{c})| = \sum_j \left[\{\mathbf{H} + \text{diag}(\mathbf{c})\}^{-1} \right]_{jj} dc_j$$

we obtain

$$\begin{aligned} \log \tilde{\pi}_{\text{GG}}(\mathbf{x}_{-i} | x_i, \boldsymbol{\theta}, \mathbf{y}) \Big|_{\mathbf{x}_{-i} = \mathbb{E}_{\tilde{\pi}_{\text{G}}}(\mathbf{x}_{-i}|x_i)} &= \text{constant} \\ &- \frac{1}{2}x_i^{(s)} \sum_{j \in \mathcal{I} \setminus i} \text{Var}_{\tilde{\pi}_{\text{G}}}(x_j|x_i) d_j^{(3)}(\mu_i(\boldsymbol{\theta}), \boldsymbol{\theta}) \sigma_j(\boldsymbol{\theta}) a_{ij}(\boldsymbol{\theta}) + \dots \end{aligned} \quad (19)$$

For Gaussian data (18) is just a constant, so the first order term in (19) is the first correction for non-Gaussian observations. Note that

$$\text{Var}_{\tilde{\pi}_{\text{G}}}(x_j|x_i) = \sigma_j^2(\boldsymbol{\theta}) \{1 - \text{Corr}_{\tilde{\pi}_{\text{G}}}(x_i, x_j)^2\}$$

but the correlation between x_i and x_j is only available for some of the i 's and j 's. This is because the marginal variances are computed using (5). We approach this problem by simply replacing all correlations not computed by a default value, say 0.05.

We now collect the expansions (17) and (19). Define

$$\begin{aligned}\gamma_i^{(1)}(\boldsymbol{\theta}) &= \frac{1}{2} \sum_{j \in \mathcal{I} \setminus i} \sigma_j^2(\boldsymbol{\theta}) \{1 - \text{Corr}_{\tilde{\pi}_G}(x_i, x_j)^2\} d_j^{(3)}(\mu_i(\boldsymbol{\theta}), \boldsymbol{\theta}) \sigma_j(\boldsymbol{\theta}) a_{ij}(\boldsymbol{\theta}) \\ \gamma_i^{(3)}(\boldsymbol{\theta}) &= \sum_{j \in \mathcal{I} \setminus i} d_j^{(3)}(\mu_i(\boldsymbol{\theta}), \boldsymbol{\theta}) \{\sigma_j(\boldsymbol{\theta}) a_{ij}(\boldsymbol{\theta})\}^3\end{aligned}\tag{20}$$

then

$$\log \tilde{\pi}_{\text{SLA}}(x_i^s | \boldsymbol{\theta}, \mathbf{y}) = \text{constant} - \frac{1}{2}(x_i^{(s)})^2 + \gamma_i^{(1)}(\boldsymbol{\theta}) x_i^{(s)} + \frac{1}{6}(x_i^{(s)})^3 \gamma_i^{(3)}(\boldsymbol{\theta}) + \dots\tag{21}$$

Eq. (21) does not define a density as the third order term is unbounded. A common way to introduce skewness into the Gaussian distribution is to use the Skew-Normal distribution (Azzalini and Capitanio, 1999)

$$\pi_{\text{SN}}(z) = \frac{2}{\omega} \phi\left(\frac{z - \xi}{\omega}\right) \Phi\left(a \frac{z - \xi}{\omega}\right)\tag{22}$$

where $\phi(\cdot)$ and $\Phi(\cdot)$ are the density and distribution function of the standard normal distribution, and $\xi, \omega > 0$, and a are respectively the location, scale, and skewness parameters. We fit a Skew-Normal density to (21) so that the third derivative at the mode is $\gamma_i^{(3)}$, the mean is $\gamma_i^{(1)}$ and the variance is 1. In this way, $\gamma_i^{(3)}$ only contributes to the skewness whereas the adjustment in the mean comes from $\gamma_i^{(1)}$; see Appendix for details.

We have implicitly assumed that the expansion (17) is dominated by the third order term. This is adequate when the log-likelihood is skewed, but not for symmetric distributions with thick tails like a Student- t_ν with a low degree of freedom. For such cases, we expand only the denominator (19) and fit the spline-corrected Gaussian (16), instead of a skewed Normal. This is slightly more expensive, but is needed.

The simplified Laplace approximation appears to be highly accurate for many observational models. The computational cost is dominated by the calculation of vector $a_i(\boldsymbol{\theta})$, for each i ; thus the ‘region of interest’ strategy (14) is unhelpful here. Most of the other terms in (20) do not depend on i , and thus are computed only once. The cost for computing (21), for a given i , is of the same order as the number of non-zero elements of the Cholesky triangle, e.g. $\mathcal{O}(n \log n)$ in the spatial case. Repeating the procedure n times gives a total cost of $\mathcal{O}(n^2 \log n)$ for each value of $\boldsymbol{\theta}$. We believe this is close to the lower limit for any general algorithm that approximates all of the n marginals. Since the graph of \mathbf{x} is general, we need to visit all other sites, for each i , for a potential contribution. This operation alone costs $\mathcal{O}(n^2)$.

4 Approximation error: Asymptotics and practical issues

4.1 Approximation error of $\tilde{\pi}(\boldsymbol{\theta} | \mathbf{y})$

To simplify the discussion, we assume that the dimension of the observations \mathbf{y} , n_d , equals the dimension of the latent field \mathbf{x} , n , so that each node x_i is observed as y_i . Equation (1) can be rewritten as

$$\begin{aligned}\left\{ \frac{\tilde{\pi}_u(\boldsymbol{\theta} | \mathbf{y})}{\pi(\boldsymbol{\theta} | \mathbf{y})} \right\}^{-1} &\propto |\mathbf{Q}^*(\boldsymbol{\theta})|^{1/2} \int \exp \left[-\frac{1}{2} \{\mathbf{x} - \mathbf{x}^*(\boldsymbol{\theta})\}^T \mathbf{Q}^* \{\mathbf{x} - \mathbf{x}^*(\boldsymbol{\theta})\} + r(\mathbf{x}; \boldsymbol{\theta}, \mathbf{y}) \right] d\mathbf{x} \\ &\propto \mathbb{E}_{\tilde{\pi}_G} [\exp \{r(\mathbf{x}; \boldsymbol{\theta}, \mathbf{y})\}]\end{aligned}\tag{23}$$

where $\tilde{\pi}_u(\boldsymbol{\theta} | \mathbf{y})$ is the unnormalised version of $\tilde{\pi}(\boldsymbol{\theta} | \mathbf{y})$, $\mathbf{x}^*(\boldsymbol{\theta})$ and $\mathbf{Q}^*(\boldsymbol{\theta})$ are the mean and variance of Gaussian distribution $\tilde{\pi}_G$, $r(\mathbf{x}; \boldsymbol{\theta}, \mathbf{y}) = \sum_i h_i(x_i)$, and $h_i(x_i)$ is $g_i(x_i)$ minus its Taylor expansion up to order two around $x_i^*(\boldsymbol{\theta})$, see (7) and (8). The approximation $\tilde{\pi}(\boldsymbol{\theta} | \mathbf{y})$ is based on a Taylor expansion of order 2, but higher

orders can also be computed. Denote by $\mathbf{S}(\boldsymbol{\theta}) = (s_{ij}(\boldsymbol{\theta}))$ the inverse of $\mathbf{Q}^*(\boldsymbol{\theta})$. Straightforward calculations show that

$$\begin{aligned} \tilde{\pi}_u(\boldsymbol{\theta} | \mathbf{y}) &= \tilde{\pi}_u(\boldsymbol{\theta} | \mathbf{y}) \left[1 + \frac{1}{8} \sum_{i=1}^n s_{ii}(\boldsymbol{\theta})^2 \frac{\partial^4 g_i(x_i^*(\boldsymbol{\theta}))}{\partial x_i^4} + \frac{5}{24} \sum_{i=1}^n \left\{ \frac{\partial^3 g_i(x_i^*(\boldsymbol{\theta}))}{\partial x_i^3} \right\}^2 s_{ii}(\boldsymbol{\theta})^3 \right. \\ &\quad \left. + \frac{1}{24} \sum_{i \neq j} \frac{\partial^3 g_i(x_i^*(\boldsymbol{\theta}))}{\partial x_i^3} \frac{\partial^3 g_j(x_j^*(\boldsymbol{\theta}))}{\partial x_j^3} s_{ij}(\boldsymbol{\theta}) \{2s_{ij}(\boldsymbol{\theta})^2 + 3s_i(\boldsymbol{\theta})s_j(\boldsymbol{\theta})\} \right]^{-1} \end{aligned}$$

corresponds to an expansion up to order 7: odd orders produce null coefficients, and order 4 and 6 give the second term, and the two last terms, respectively. The density above is not necessarily positive, but if both approximations are close, this seems an indication that both are accurate. We discuss only $\tilde{\pi}(\boldsymbol{\theta} | \mathbf{y})$ from now on.

For sake of exposition, denote p the dimension of integral (23), although $p = n$ in our case. Under standard assumptions, in particular when p is fixed, this integral is $1 + \mathcal{O}(n^{-1})$; see e.g. Tierney and Kadane (1986). Shun and McCullagh (1995) consider the case where p grows with n , but do not establish rigorously the error rate. It does not seem possible in our settings to prove that the multiplicative error is always $o(1)$ with respect to n . For instance, if the x_i 's are independent, it is possible to exhibit cases where the error, for any expectation evaluated with respect to the marginal posterior, is $\mathcal{O}(1)$ but not $o(1)$. (More details are available on request.) This discussion is complicated by the difficulty of defining asymptotics in spatial models: observations may be generated in a larger and larger domain (increasing domain asymptotics), in a fixed volume (infill asymptotics), and other asymptotic schemes could be devised. Instead, we propose heuristic arguments for explaining the good accuracy observed in practical applications.

Remark 1 The ‘actual’ dimensionality of (23) is typically much smaller than n . Because of the dependency within \mathbf{x} , \mathbf{x} is well approximated by its q principal components, with $q \ll n$. A convenient measure of the dimensionality of (23) is Spiegelhalter et al. (2002)’s measure for the *effective number of parameters*, p_D . In the case of approximately Gaussian models, then

$$p_D(\boldsymbol{\theta}) \approx \text{Trace} \left\{ \mathbf{Q}(\boldsymbol{\theta}) \mathbf{Q}^*(\boldsymbol{\theta})^{-1} \right\}, \quad (24)$$

the trace of the prior precision matrix times the by posterior covariance matrix of the Gaussian approximation. The quantity $p_D(\boldsymbol{\theta})$ indicates how informative the data is, and to which extent the Gaussianity and the dependence structure of the prior are preserved in the posterior of \mathbf{x} , given $\boldsymbol{\theta}$. The calculation of $p_D(\boldsymbol{\theta})$ is cheap, since the covariances of neighbours are obtained as a by-product of the computation of the marginal variances in the Gaussian approximation based on (5).

Remark 2 The approximation error is reduced through normalisation, provided (23) is roughly constant with respect to $\boldsymbol{\theta}$ within the support of the true marginal. For the Laplace approximation with standard assumptions, renormalisation improves the relative error from $\mathcal{O}(n^{-1})$ to $\mathcal{O}(n^{-3/2})$ (Tierney and Kadane, 1986).

Remark 3 The high accuracy of our approximation which we obtain in the experiments in Section 5, seems to be due both to the Gaussian latent field and the well-behaved observational models usually considered in applications, e.g. an exponential family distribution for $\pi(y_i | x_i, \boldsymbol{\theta})$.

Following these remarks, a more direct way to assess the approximation error is simply to evaluate the order of magnitude of $r(\mathbf{x}; \boldsymbol{\theta}, \mathbf{y})$: simulate independent samples $\{\mathbf{x}^j\}$ from $\tilde{\pi}_G(\mathbf{x} | \boldsymbol{\theta}, \mathbf{y})$, and compute the 0.025 lower and upper quantiles of the empirical distribution of the $r(\mathbf{x}; \boldsymbol{\theta}, \mathbf{y})$'s. This is a rather quick procedure. The first term in the exponential defining the integrand in (23) is distributed according to $\chi_n^2/2$, so we consider the remainder $r(\mathbf{x}; \boldsymbol{\theta}, \mathbf{y})$ to be small if quantiles are in absolute value much less than n . In the same way, empirical averages of the $r(\mathbf{x}; \boldsymbol{\theta}, \mathbf{y})$, for different values of $\boldsymbol{\theta}$, can be used to determine the variability of (23) with respect to $\boldsymbol{\theta}$.

4.2 Approximation error of $\tilde{\pi}_{\text{LA}}(x_i|\boldsymbol{\theta}, \mathbf{y})$

The approximation error of $\tilde{\pi}_{\text{LA}}(x_i|\boldsymbol{\theta}, \mathbf{y})$ admits a similar expression to that of $\tilde{\pi}(\boldsymbol{\theta}|\mathbf{y})$. To see this, consider an alternative structure for the model, where the node x_i becomes an additional component of the parameter $\boldsymbol{\theta}$, and the latent field is therefore \mathbf{x}_{-i} ; then, the same manipulations as for (23) leads eventually to

$$\left\{ \frac{\tilde{\pi}_{\text{LA},u}(x_i|\boldsymbol{\theta}, \mathbf{y})}{\pi(x_i|\boldsymbol{\theta}, \mathbf{y})} \right\}^{-1} \propto \mathbb{E}_{\tilde{\pi}_{\text{GG}}} [\exp \{r(\mathbf{x}; \boldsymbol{\theta}, \mathbf{y})\}]$$

where $\tilde{\pi}_{\text{LA},u}(x_i|\boldsymbol{\theta}, \mathbf{y})$ is the unnormalised version of $\tilde{\pi}_{\text{LA}}(x_i|\boldsymbol{\theta}, \mathbf{y})$. Thus, we essentially obtain the same result as in Section 4.1; before normalisation, the approximation error of $\tilde{\pi}_{\text{LA}}(x_i|\boldsymbol{\theta}, \mathbf{y})$ is comparable to that of $\tilde{\pi}(\boldsymbol{\theta}|\mathbf{y})$, and the assessment criteria proposed in the previous section are also good indicators of the accuracy of $\tilde{\pi}_{\text{LA}}(x_i|\boldsymbol{\theta}, \mathbf{y})$. Note however that normalisation has a different effect on $\tilde{\pi}_{\text{LA}}(x_i|\boldsymbol{\theta}, \mathbf{y})$. Under increasing domain asymptotics, new components $x_{i'}$ are generated further and further from x_i , so at some point the additional terms $h_{i'}(x_{i'})$ in the expression of the remainder $r(\mathbf{x}; \boldsymbol{\theta}, \mathbf{y})$ should be a constant with respect to x_i , and therefore should be cancelled by normalisation. Thus, we conjecture that the error is at worst $\mathcal{O}(1)$ under increasing domain asymptotics.

4.3 Assessing the approximation error

Obviously, there is only one way to assess with certainty the approximation error of our approach, which is to run an MCMC sampler for an infinite time. However, we propose to use the following two strategies to assess the approximation error, which should be reasonable in most situations.

Our first strategy is to verify the overall approximation $\tilde{\pi}_{\text{G}}(\mathbf{x}|\boldsymbol{\theta}, \mathbf{y})$, for each $\boldsymbol{\theta}_k$ used in the integration. We do this by computing $p_{\text{D}}(\boldsymbol{\theta})$ (24), and the lower and upper $\alpha/2$ quantiles in the empirical distribution for the remainder $r(\mathbf{x}; \boldsymbol{\theta}, \mathbf{y})$. If p_{D} is small compared to n , and the quantiles of $r(\mathbf{x}; \boldsymbol{\theta}, \mathbf{y})$ are in absolute value much less than n , this provides strong confidence that the Gaussian approximation is an adequate approximation.

Our second strategy is based on the simple idea of comparing elements of a sequence of more and more accurate approximations. In our case, this sequence consists of the Gaussian approximation (2), followed by the simplified Laplace approximation (21), then by the Laplace approximation (11). Specifically we compute the integrated marginal (3) based on both the Gaussian approximation and the simplified Laplace approximation, and compute their (symmetric) Kullback-Leibler divergence (KLD). If the divergence is small then both approximations are considered as acceptable. Otherwise, compute (3) using the Laplace approximation (11) and compute the divergence with the one based on the simplified Laplace approximation. Again, if the divergence is small, simplified Laplace and Laplace approximations appear to be acceptable; otherwise, the Laplace approximation is our best estimate but the label ‘problematic’ should be attached to the approximation to warn the user. (This last option has not yet happened to us.)

To assess the error due to the numerical integration (3), we can compare the KLD between the posterior marginals obtained with a standard and those obtained with a higher resolution. As a such approach is standard in numerical integration, we do not pursue this issue here.

5 Examples

This section provides examples of applications of the INLA approach, with comparisons to results obtained from intensive MCMC runs. Comparisons are expressed in terms of computational time; computations were performed on a 2.1GHz laptop, and programmed in C. We start with simple examples with fixed $\boldsymbol{\theta}$ in Section 5.1 and Section 5.2, to verify the (simplified) Laplace approximation for $x_i|\boldsymbol{\theta}, \mathbf{y}$. We continue with a stochastic volatility model applied to exchange rate data in Section 5.3 and a spatial semi-parametric ecological regression problem in Section 5.4. The dimensions gets really large in Section 5.5, in which we analyse some data using a spatial log-Gaussian Cox process.

5.1 Simple simulated examples

We start by illustrating the various approximations of $\pi(x_i|\boldsymbol{\theta}, \mathbf{y})$ in two quite challenging examples. The first model is based on a first order auto-regressive latent field with unknown mean,

$$\eta_t - \mu \mid \eta_1, \dots, \eta_{t-1}, \mu \sim \mathcal{N} \{ \phi(\eta_{t-1} - \mu), \sigma^2 \}, \quad t = 1, \dots, 50 \quad (25)$$

where $\mu \sim \mathcal{N}(0, 10)$, $\phi = 0.85$ and $\text{Var}(\eta_t) = 1$. As our observations we take

$$y_t - \eta_t \mid (\boldsymbol{\eta}, \mu) \sim \text{Student-}t_3 \quad \text{and} \quad y_t \mid (\boldsymbol{\eta}, \mu) \sim \text{Bernoulli} \{ \text{logit}^{-1}(\eta_t) \}$$

for $t = 1, \dots, 50$, in both experiments. Note that the Student- t_3 is symmetric so we use the full numerator in the simplified Laplace approximations as described in Section 3.2.3.

To create the observations, we sampled first $\mathbf{x} = (\boldsymbol{\eta}^T, \mu^T)^T$ from the prior, then simulated the observations. We computed $\tilde{\pi}(\eta_t|\boldsymbol{\theta}, \mathbf{y})$ for $t = 1, \dots, 50$ and $\tilde{\pi}(\mu|\boldsymbol{\theta}, \mathbf{y})$ using the simplified Laplace approximation and located the node with maximum Kullback-Leibler divergence (KLD) between the Gaussian and the simplified Laplace approximations. This process was repeated 100 times, and the realisation with the largest maximum KLD was selected. Figure 3 displays the results for the Student- t_3 data (first column) and the Bernoulli data (second column). Panel (a) and (b) display $\boldsymbol{\eta}$ (solid line) and the observed data (circles). In (a) the node with the maximum KLD is marked with a vertical line and solid dot. In (b) the node with the maximum KLD is μ hence not shown. Panel (c) and (d) display the approximated marginals for the node with maximum KLD in the standardised scale (15). The dotted line is the Gaussian approximation, the dashed line is the simplified Laplace and the solid line is the Laplace approximation. In both cases, the simplified Laplace and the Laplace approximation are very close to each other. The KLD between the Gaussian approximation and the simplified Laplace one is 0.20 and 0.05, respectively. The KLD between the simplified Laplace approximation and the Laplace one is 0.001 and 0.0004. Panel (e) and (f) show the simplified Laplace approximation with a histogram based on 10,000 (near) independent samples from $\pi(\boldsymbol{\eta}, \mu|\boldsymbol{\theta}, \mathbf{y})$. The fit is excellent.

The great advantage of the Laplace approximations is the high accuracy and low computational cost. In both examples, we computed all the approximations (for each experiment) in less than 0.08 seconds, whereas the MCMC samples required about 25 seconds.

The results shown in this example are rather typical and are not limited to simple time-series models like (25). The Laplace approximation only ‘sees’ the log-likelihood model and then uses some of the other nodes (see Figure 2) to compute the correction to the Gaussian approximation. Hence, the form of the log-likelihood is more important than the form of the covariance for the latent field. We expect similar results for spatial or spatio-temporal latent Gaussian models, with Student- t_ν or Binomial observations.

5.2 Bayesian multiscale analysis for time series data

This example shows a situation where it is useful to have estimates of the marginals with a relative error, so that even the tails can be evaluated accurately. We extend the Bayesian multiscale tool for exploratory analysis of time series data by Øigård et al. (2006) to allow for non-Gaussian observations. The fundamental problem is to detect significant features and important structures of a signal observed with noise. Although a noisy signal can be smoothed, often some of the features are visible only on certain scales, and may disappear if the smoothing is too severe. The multiscale idea consists in considering several levels of smoothing simultaneously. Chaudhuri and Marron (1999) introduced such ideas in nonparametric function estimation in the form of the SIZer methodology (SIGNificant ZERO crossings of derivatives), see also Erästö (2005).

Let $\eta(t)$ be the unknown continuous underlying signal with derivatives $\eta'(t)$, and level of smoothing κ . Let $\mathbf{y} = (y_1, \dots, y_n)^T$ be observations of $\eta(\cdot)$ at time-points $\mathbf{t} = (t_1, \dots, t_n)^T$. The derivative is said to be ‘significant’ positive at time point t , if

$$\text{Prob}(\eta'(t) > 0 \mid \mathbf{y}, \kappa) > 1 - \alpha/2$$

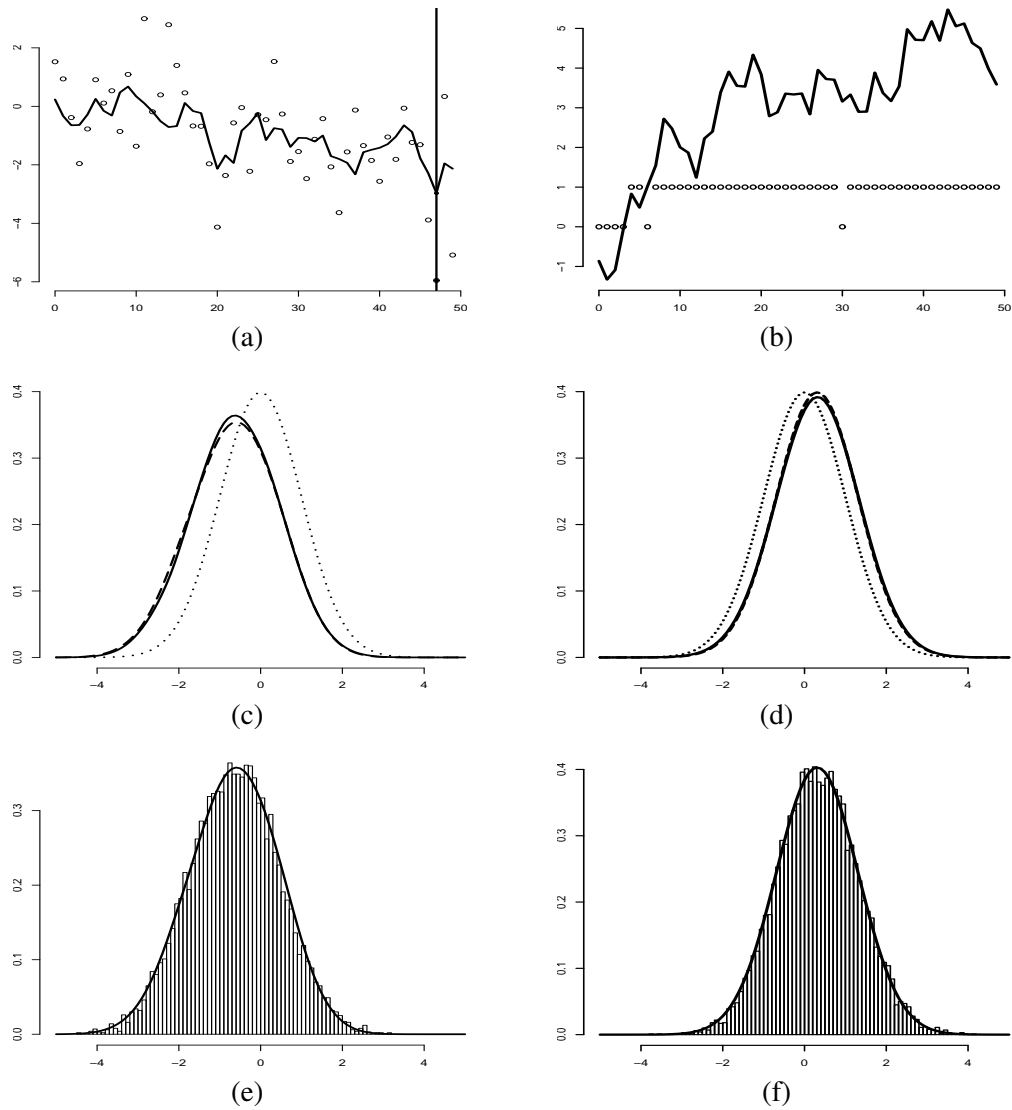


Figure 3: First row shows the true latent Gaussian field (solid line), the observed Student- t_3 data and Bernoulli data (dots). Second row shows the approximate marginal for a selected node using various approximations; Gaussian (dotted), simplified Laplace (dashed) and Laplace (solid). Last row compares samples from a long MCMC chain with the marginal computed with the simplified Laplace approximation.

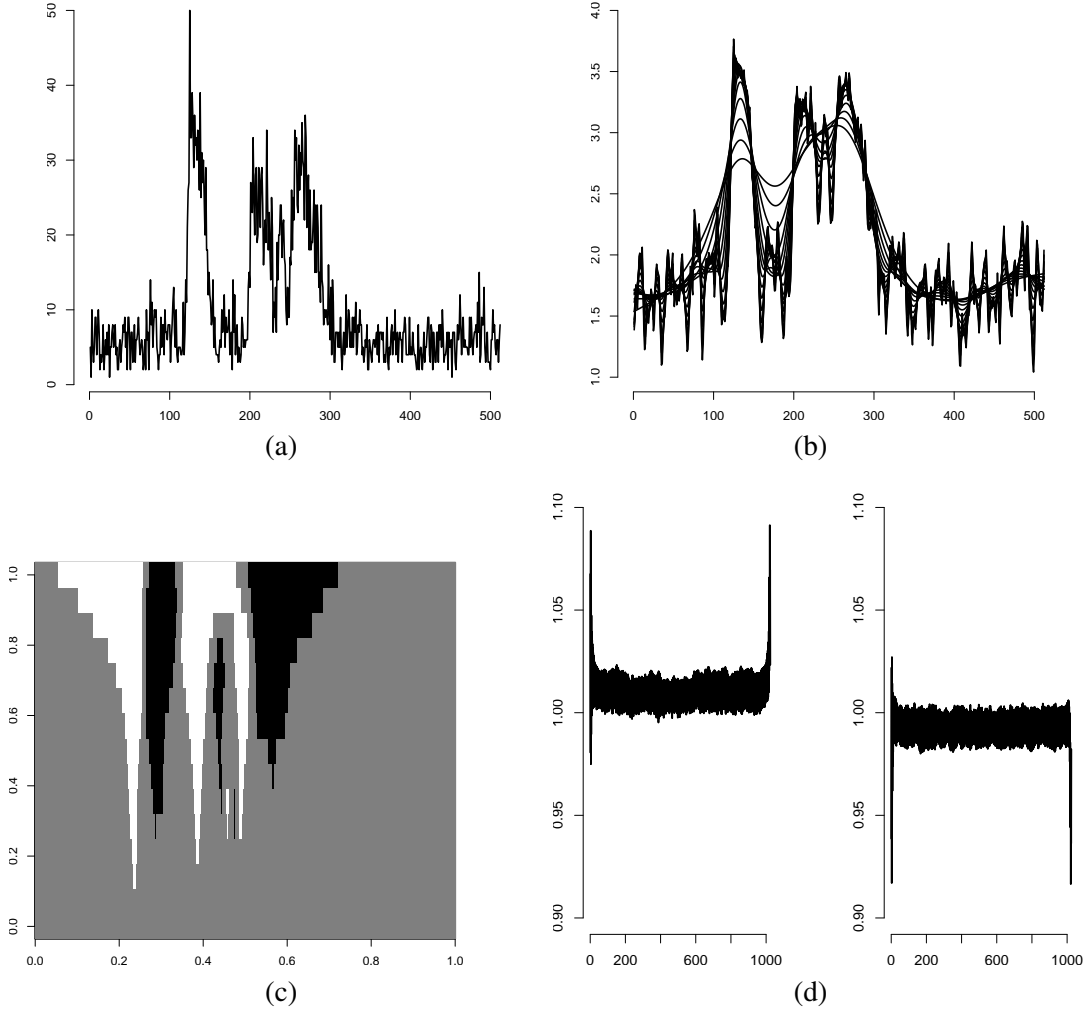


Figure 4: Results for the multiscale analysis example. Panel (a) displays the raw burst data. Panel (b) displays the posterior means for varying degree of smoothing. Panel (c) shows the SIZer map for $\alpha = 0.05$. Panel (d) displays the ratios between the estimated and the real probability of being below the approximate quantiles, lower quantiles on the left and upper quantiles on the right.

and similarly significant negative, where α is the level of significance. The SIZer map displays regions of significant positive and negative gradients for various levels of smoothing κ .

We now illustrate how to use the simplified Laplace approximation to compute the SIZer map. We use gamma ray burst intensity data previously analysed by Besbeas et al. (2004); the observations are Poisson:

$$y(t_i) \sim \text{Poisson} \{ \exp(\eta(t_i)) \}, \quad t_i = i \quad \text{for } i = 1, \dots, n = 512$$

where $\eta(t)$ is the latent Gaussian process. The data are displayed in Figure 4(a). We follow Øigård et al. (2006), and model the continuous process $\eta(t)$ as an integrated Wiener process with precision κ . Wecker and Ansley (1983) show that the integrated Wiener process is Markov if augmented with the derivatives $\eta'(t)$,

$$\left(\begin{array}{c} \eta(t_{i+1}) \\ \eta'(t_{i+1}) \end{array} \right) \middle| \left\{ \left(\begin{array}{c} \eta(s) \\ \eta'(s) \end{array} \right), s \leq t_i \right\}, \kappa \sim \mathcal{N} \left\{ \left(\begin{array}{cc} 1 & \delta_i \\ 0 & 1 \end{array} \right) \left(\begin{array}{c} \eta(t_i) \\ \eta'(t_i) \end{array} \right), \frac{1}{\kappa} \left(\begin{array}{cc} \delta_i^3/3 & \delta_i^2/2 \\ \delta_i^2/2 & \delta_i \end{array} \right) \right\}$$

where $\delta_i = t_{i+1} - t_i$. Hence, the discretely observed integrated Wiener process $\mathbf{x} = (\{\eta(t_i)\}, \{\eta'(t_i)\})^T$ is a GMRF of dimension $2n$, see Rue and Held (2005, Sec. 3.5). Note that the derivatives at t_i are a part of the GMRF, hence we can approximate their marginal densities (for a fixed κ) and check whether they are significant negative or positive.

We use the simplified Laplace approximation and compute all the $2n$ marginals for $\log \kappa = 1, \dots, 15$. This takes about 0.35 seconds for each value of $\log \kappa$. The posterior means of $\{\eta(t_i)\}$ and the SIZer map for $\alpha = 0.05$ are displayed in Figure 4(b) and (c), respectively. In the SIZer map, white indicates significant positive derivative, black indicates significant negative derivative whereas grey indicates none. The vertical scale in the SIZer map goes from $\log \kappa = 1$ to $\log \kappa = 15$.

To verify the results we ran a MCMC sampler for nine hours and estimated the probability for the chain to be below the $\alpha/2$ (where $\alpha = 0.05$) quantiles as computed from our approximation $\tilde{\pi}(\eta'(t_i)|\kappa, \mathbf{y})$. Panel (d) displays the ratios of these estimated probabilities and the true value $\alpha/2$ for all the 1024 nodes in the Markov field. Lower (resp. upper) quantiles are displayed on the left (resp. right). In both cases the error is largest at the two ends. The average ratio is 1.01 for the lower quantiles and 0.99 for the upper quantiles, so the average absolute approximation bias is 0.00025. This is indeed impressive; recall that the simplified Laplace approximation fits the skew-Normal parametric density.

Hannig and Marron (2006) develop some theory for computing more accurately (asymptotically) the SIZer map for nonparametric regression. The Bayesian approach taken here does not resort to asymptotic theory, can deal with non-Gaussian observations, can take into account covariates and unstructured effects and so on. The calculations can be done exactly for Gaussian observation models, and, as illustrated here, practically exactly for common non-Gaussian observation models.

5.3 Stochastic volatility models

Stochastic volatility models are frequently used to analyse financial time series. Figure 5(a) displays the log of the daily difference of the pound-dollar exchange rate from October 1st, 1981, to June 28th, 1985. This dataset has been analysed by Durbin and Koopman (2000), among others. There has been much interest in developing efficient MCMC methods for such models, e.g. Shephard and Pitt (1997) and Chib et al. (2002).

Following Durbin and Koopman (2000), we consider a first order auto-regressive latent Gaussian process

$$x_t | x_1, \dots, x_{t-1}, \tau, \phi \sim \mathcal{N}(\phi x_{t-1}, 1/\tau),$$

where $|\phi| < 1$ to ensure stationarity. The observations are taken to be

$$y_t | x_1, \dots, x_t, \kappa \sim \mathcal{N}\{0, \exp(x_t)/\kappa\} \quad (26)$$

where κ is an unknown precision. The log-likelihood (with respect to κ) is quite far from being Gaussian and is non-symmetric. There is some evidence that financial data have heavier tails than the Gaussian, so a Student- t_ν distribution with unknown degrees of freedom can be substituted to the Gaussian in (26); see Chib et al. (2002). We consider this modified model at the end of this example.

We display the results for the simplified Laplace approximation of the posterior marginals of the three unknown hyperparameters (properly transformed so that $\boldsymbol{\theta} \in \mathbb{R}^3$):

$$\theta_1 = \text{logit}\left(\frac{\phi + 1}{2}\right), \quad \theta_2 = \log \tau, \quad \text{and} \quad \theta_3 = \log \kappa.$$

We use vague priors for $\boldsymbol{\theta}$, as strong priors make the approximation problem easier. For the same reason, we display the results based on only the first $n = 50$ observations in Figure 5(a). The results for the full dataset are similar, but the posterior marginals for the θ_j 's are closer to Gaussians.

Figure 5(b)-(d) displays the approximate posterior marginals for θ_1 , θ_2 and θ_3 . The histograms are constructed from the output of a MCMC algorithm running for one day. The approximations computed are quite precise and no serious deviance can be detected. Figure 5(e) displays the approximate posterior marginal for $x_t|\mathbf{y}$ based on the simplified Laplace approximation, for the component of \mathbf{x} which maximises the KLD between the posterior marginal based on the Gaussian approximation and based on the simplified Laplace approximation. The KLD for all the x_t 's are quite small and roughly equal to 3×10^{-4} , so there is no particular gain in using the simplified Laplace approximation compared to the Gaussian one. The fit is quite good, although we slightly underestimate the right hand side tail. The approximation error diminishes as the

number of observations increases, but is still visible for the full dataset. A closer inspection reveals that the underestimation is due to the (default) quite rough numerical integration (3). Improving the accuracy of the numerical integration removes the underestimation.

We validated the approximations using all the $n = 945$ observations at the modal value θ^* . The effective number of parameters (24) was about 53, which is small compared to n . A 95% interval for the remainder $r(\mathbf{x}; \theta^*, \mathbf{y})/n$ is $[-0.002, 0.004]$ using 1,000 independent samples. The computational cost for obtaining all the posterior marginals was about 0.32 seconds for each value of θ , and 32 seconds in total.

We also applied the stochastic volatility model to the full dataset, see Figure 5(a), using a Student- t_ν instead of a Gaussian for the observational model in (26), and a uniform prior for $\log \nu$. The number of hyperparameters is then 4. Figure 5(f) shows predictions for future x_t 's using the full dataset.

5.4 Semi-parametric ecological regression

In this example we consider an ecological regression problem and analyse the spatial variation of disease risk in relation to a proxy exposure variable available on the same units. This is taken from Natario and Knorr-Held (2003), which is referred to for a more throughout background.

The data are male larynx cancer mortality counts in the $n = 544$ districts of Germany from 1986 to 1990:

$$y_i | \eta_i \sim \text{Poisson} \{E_i \exp(\eta_i)\}, \quad i = 1, \dots, n. \quad (27)$$

The (fixed) 'district effect' E_i accounts for the number of people in district i , its age distribution, etc., and η is the log-relative risk. The maximum likelihood estimator for η_i using (27) is y_i/E_i and is displayed in Figure 6(a). The model for η_i takes the following form,

$$\eta_i = u_i + v_i + f(c_i) \quad (28)$$

where \mathbf{u} is a spatially structured term, \mathbf{v} is a unstructured term ('random' effects) and $f(c_i)$ is an unknown effect of the exposure covariate with value c_i at district i . For the exposure covariate we use the lung cancer rate as a proxy for smoking consumption, see Figure 6(b). The spatially structured term is modelled as an intrinsic GMRF (Rue and Held, 2005, Ch. 3)

$$u_i | \mathbf{u}_{-i}, \kappa_{\mathbf{u}} \sim \mathcal{N} \left(\frac{1}{n_i} \sum_{j \sim i} u_j, \frac{1}{n_i \kappa_{\mathbf{u}}} \right)$$

where n_i are the number of neighbour districts of i and $\kappa_{\mathbf{u}}$ is the unknown precision. The unstructured term \mathbf{v} is taken as a vector of independent $\mathcal{N}(0, \kappa_{\mathbf{v}})$. The effect of the covariate \mathbf{c} is modelled as a smooth function $f(\cdot)$, parametrised as unknown values $\mathbf{f} = (f_1, \dots, f_{100})^T$ for arguments $1, \dots, n_c = 100$. The values of the covariate are scaled to the interval $[1, n_c]$. The vector \mathbf{f} is assumed to follow a second-order random walk,

$$\pi(\mathbf{f} | \kappa_{\mathbf{f}}) \propto \kappa_{\mathbf{f}}^{(n_c-1)/2} \exp \left\{ -\frac{1}{2} \kappa_{\mathbf{f}} \sum_{i=3}^{n_c} (f_i - 2f_{i-1} + f_{i-2})^2 \right\} \quad (29)$$

with unknown precision $\kappa_{\mathbf{f}}$. To separate the spatial effect and the effect of covariate, we impose $\sum_i u_i = 0$.

Following Natario and Knorr-Held (2003), we assign independent vague Gamma priors to $\theta = (\kappa_{\mathbf{u}}, \kappa_{\mathbf{v}}, \kappa_{\mathbf{f}})$. We use the simplified Laplace approximation for the marginals of $\mathbf{x} = (\boldsymbol{\eta}^T, \mathbf{u}^T, \mathbf{f}^T)^T$ with length 1,188. The computation took about 52 seconds using 53 evaluation points for the numerical integration. The posterior mean of the spatial term \mathbf{u} is displayed in Figure 6(c) whereas the posterior mean of the unstructured effect \mathbf{v} is displayed in (d). The posterior mean of the covariate effect \mathbf{f} is displayed in (e) with lower and upper 0.025 percentile, computed with the simplified Laplace approximation (solid) and the Gaussian approximation (2) (dotted). The two approximations nearly agree, which is also confirmed by relatively small values of the KLD between the two approximations shown in (f). The maximum KLD for all variables appears for f_{50} and equals 0.032. This indicates that the Gaussian approximation had been sufficient for this example, so the

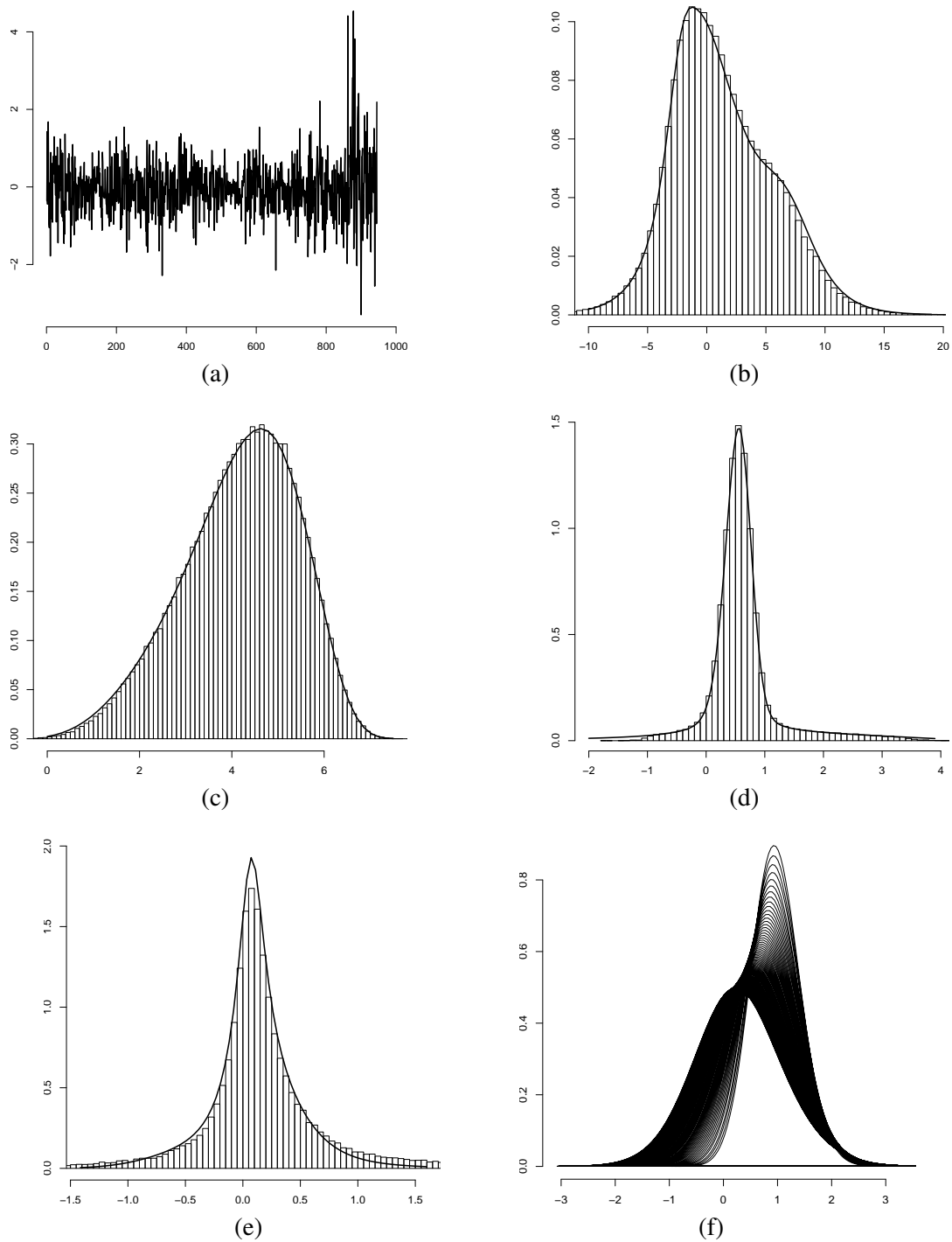


Figure 5: Panel (a) displays the log of the daily difference of the pound-dollar exchange rate from October 1st, 1981, to June 28th, 1985. Panels (b)-(d) display the approximated posterior marginals for θ_1 to θ_3 using only the first $n = 50$ observations in (a). Overlaid are the histograms obtained from a very long MCMC run. The fit is perfect. Panel (e) displays the approximated posterior marginal for the location of the latent field with maximum KLD, compared with the histograms from a very long MCMC run. Our approximation underestimate slightly the behaviour on the right hand side, but this turn out to be an effect of the default (quite rough) integration method. Panel (f) displays the predicted posterior marginals for future x_t 's conditioned on the full dataset, assuming in this case that observations are Student- t_ν distributed.

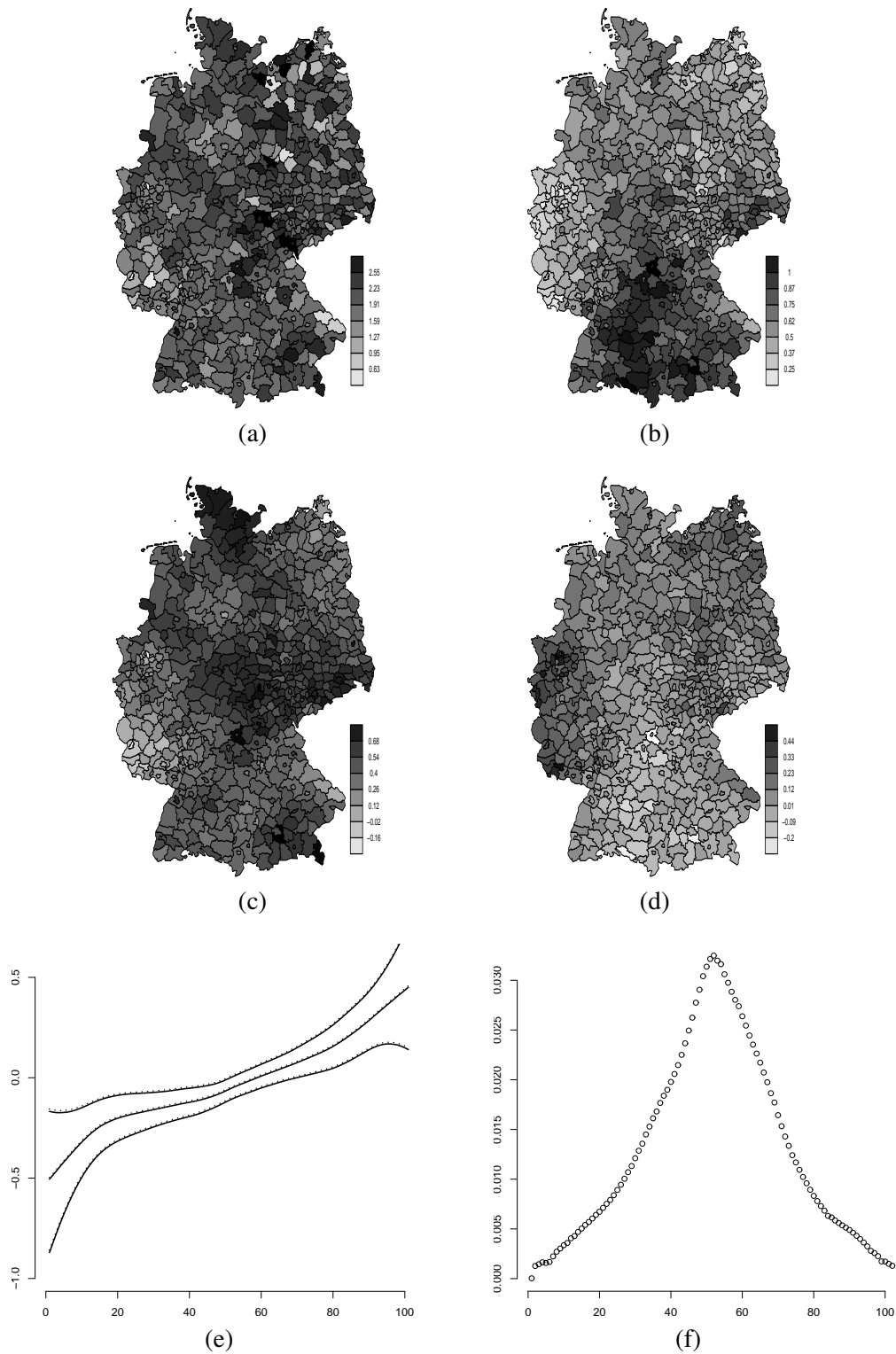


Figure 6: Semi-parametric ecological regression example: panel (a) displays the maximum likelihood estimator for the log relative risk. Panel (b) shows covariate values. Panels (c) and (d) give the posterior mean of the structured (u) and unstructured (v) effects, respectively. Panel (e) displays the posterior mean of the covariate effect with lower and upper 0.0025 percentiles. The solid line is the simplified Laplace approximation and the dotted line is the Gaussian one. Panel (f) shows the KLD for the covariate effect.

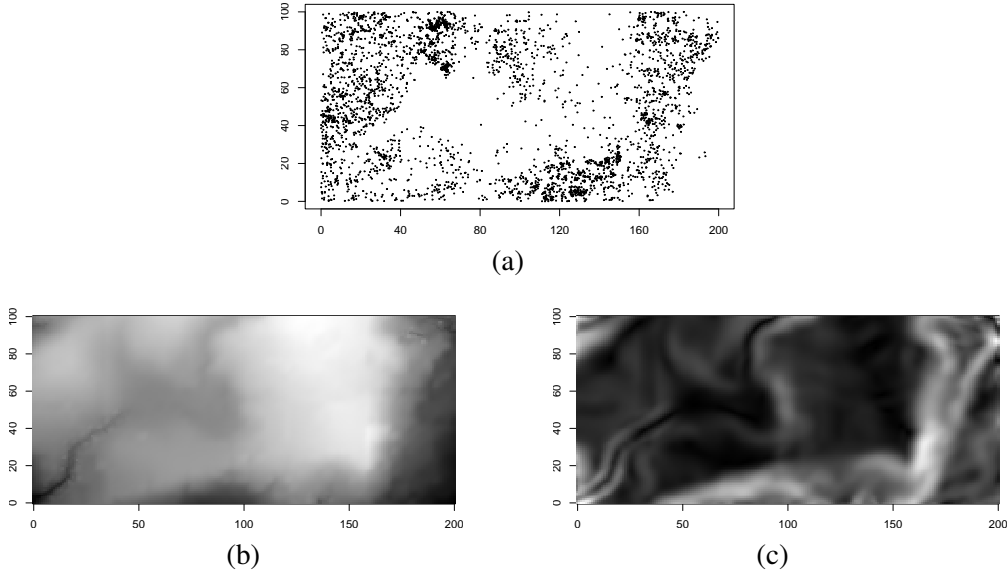


Figure 7: Data and covariates for the log-Gaussian Cox process example: (a) locations of the 3,605 trees, (b) altitude, and (c) norm of the gradient.

computational cost could have been reduced to 18 seconds (approximating all the marginals). Long MCMC runs confirm that the marginals computed using the simplified Laplace approximation are essentially correct.

We validated the approximations by computing $p_D(\boldsymbol{\theta}^*) \approx 91$ and estimated a 95% interval for the remainder $r(\boldsymbol{x}; \boldsymbol{\theta}^*, \boldsymbol{y})/n$ as $[-0.001, 0.001]$ using 1,000 independent samples.

5.5 Log-Gaussian Cox process

Log-Gaussian Cox processes (LGCP) are a flexible class of models that have been successfully used for modelling spatial or spatio-temporal point processes, see for example Møller et al. (1998), Brix and Møller (2001), Brix and Diggle (2001) and Møller and Waagepetersen (2003). In this section we will illustrate how LGCP models can be analysed using our approach for approximate inference.

A LGCP is a hierarchical Poisson process: \boldsymbol{Y} in $W \subset \mathbb{R}^d$ is a Poisson point process with a random intensity function $\lambda(\boldsymbol{\xi}) = \exp(Z(\boldsymbol{\xi}))$, where $Z(\boldsymbol{\xi})$ is a Gaussian field at $\boldsymbol{\xi} \in \mathbb{R}^d$. In this way, the dependency in the point-pattern is modelled through a common latent Gaussian variable $Z(\cdot)$. In the analysis of LGCP, it is common to discretise the observation window W . Divide W into N disjoint cells $\{w_i\}$ located at $\boldsymbol{\xi}_i$ each with area $|w_i|$. Let y_i be the number of occurrences of the realised point pattern within w_i and let $\boldsymbol{y} = (y_1, \dots, y_N)^T$. Let η_i be the random variable $Z(\boldsymbol{\xi}_i)$. Clearly $\pi(\boldsymbol{y}|\boldsymbol{\eta}) = \prod_i \pi(y_i|\eta_i)$ and $y_i|\eta_i$ is Poisson distributed with mean $|w_i| \exp(\eta_i)$; the same likelihood as for the semi-parametric ecological regression example (27). A straightforward generalisation is to allow for covariates: η_i can be decomposed in the same way as (28), say

$$\eta_i = \beta_0 + \beta_1 c_{1i} + \beta_2 c_{2i} + u_i + v_i, \quad (30)$$

where \boldsymbol{u} represent the spatial component, and \boldsymbol{v} is an unstructured term. An alternative would be to use a semi-parametric model for the effect of the covariates similar to (29).

We apply model (30) to the tropical rain forest data studied by Waagepetersen (2007). These data come from a 50-hectare permanent tree plot which was established in 1980 in the tropical moist forest of Barro Colorado Island in central Panama. Censuses have been carried out every 5th year from 1980 to 2005, where all free-standing woody stems at least 10 mm diameter at breast height were identified, tagged, and mapped. In total, over 350,000 individual trees species have been censused over 25 years. We will be looking at the tree species *Beilschmiedia pendula* Lauraceae using data collected from the first four census periods. The positions of the 3605 trees are displayed in Figure 7(a). Sources of variation explaining the locations

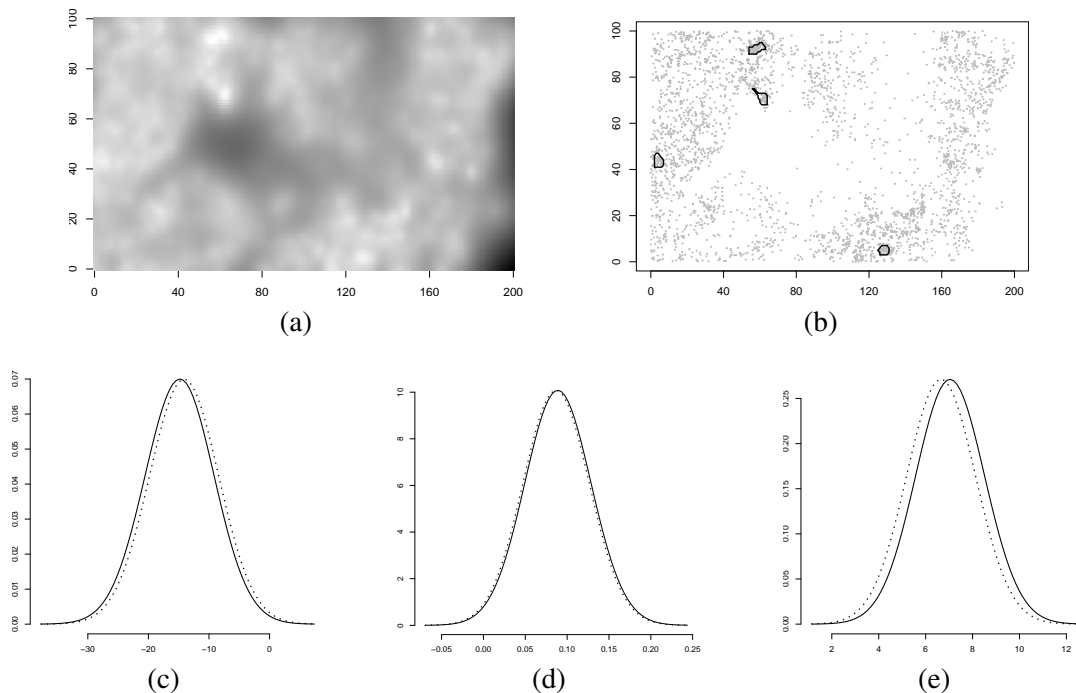


Figure 8: LGCP example: (a) posterior mean of the spatial component \mathbf{u} , (b) Nodes where the KLD between simplified Laplace and Gaussian approximations exceeds 0.2, (c)-(e) posterior marginals of β_0 , β_1 and β_2 using simplified Laplace (solid) and Gaussian approximations (dotted).

include the elevation and the norm of the gradient. There may be clustering or aggregation due to unobserved covariates or seed dispersal. The unobserved covariates can be either spatially structured or unstructured.

We start by dividing the area of interest into a 200×100 regular lattice, where each square pixel of the lattice represent 25 square metres. Denote elevation and norm of the gradient by c_1 and c_2 , respectively. The scaled versions of these covariates are shown in panel (b) and (c), for c_1 and c_2 , respectively. For the spatial structured term, we use a second order polynomial intrinsic GMRF (see Rue and Held (2005, Sec. 3.4.2)), with following full conditionals in the interior (with obvious notation)

$$E(x_i | \mathbf{x}_{-i}, \kappa_{\mathbf{u}}) = \frac{1}{20} \begin{pmatrix} \begin{matrix} \circ & \circ & \circ & \circ & \circ \\ \circ & \bullet & \bullet & \bullet & \circ \\ \circ & \bullet & \bullet & \bullet & \circ \\ \circ & \bullet & \bullet & \bullet & \circ \\ \circ & \circ & \circ & \circ & \circ \end{matrix} & -2 & \begin{matrix} \circ & \circ & \circ & \circ & \circ \\ \circ & \bullet & \bullet & \bullet & \circ \\ \circ & \bullet & \bullet & \bullet & \circ \\ \circ & \bullet & \bullet & \bullet & \circ \\ \circ & \circ & \circ & \circ & \circ \end{matrix} & -1 & \begin{matrix} \circ & \circ & \bullet & \circ & \circ \\ \circ & \bullet & \bullet & \bullet & \circ \\ \circ & \bullet & \bullet & \bullet & \circ \\ \circ & \bullet & \bullet & \bullet & \circ \\ \circ & \circ & \bullet & \circ & \circ \end{matrix} \end{pmatrix}, \quad \text{Prec}(x_i | \mathbf{x}_{-i}, \kappa_{\mathbf{u}}) = 20\kappa_{\mathbf{u}}. \quad (31)$$

The precision $\kappa_{\mathbf{u}}$ is unknown. The full conditionals are constructed to mimic the thin-plate spline. There are some corrections to (31) near the boundary, which can be found using the stencils in Terzopoulos (1988). We impose a sum-to-zero constraint on the spatial term due to β_0 . The unstructured terms \mathbf{v} are independent $\mathcal{N}(0, \kappa_{\mathbf{v}})$, vague Gamma (resp. Gaussian) priors are assigned to $\kappa_{\mathbf{u}}$ and $\kappa_{\mathbf{v}}$ (resp. to β_0, β_1 and β_2). The GMRF is $\mathbf{x} = (\boldsymbol{\eta}^T, \mathbf{u}^T, \beta_0, \beta_1, \beta_2)^T$ with dimension 40, 003, and $\boldsymbol{\theta} = (\log \kappa_{\mathbf{u}}, \log \kappa_{\mathbf{v}})$.

We computed the approximation for 20, 003 posterior marginals using the simplified Laplace approximation, thus ignoring the unstructured components. This task required about 4 hours of computing or about 24 minutes for each value of $\boldsymbol{\theta}$. The high computational cost is due to the large number of computed posterior marginals. The total cost can be reduced to only 10 minutes if using the Gaussian approximation (2). The results are displayed in Figure 8. Panel (a) displays the estimated posterior mean of the spatial component. In (b) we have marked areas where the KLD between the marginal computed with the Gaussian approximation and the one computed with the simplified Laplace approximation exceeds 0.2. These nodes are candidates for further investigation, so we computed their posteriors using also the Laplace approximation; the results agreed well with those obtained from the simplified Laplace approximation. Panel (c) to (e) display the posterior marginals computed with the Gaussian approximation (dotted) and the one computed with the simplified Laplace approximation (solid) for β_0 , β_1 and β_2 . The difference is mostly due to a horizontal shift, a characteristic valid for all the other nodes as well.

To validate the approximations, we computed $p_D(\boldsymbol{\theta}^*) \approx 1714$ and estimated a 95% interval for the remainder $r(\boldsymbol{x}; \boldsymbol{\theta}^*, \boldsymbol{y})/n$ as $[0.002, 0.005]$ using 1,000 independent samples. Varying $\boldsymbol{\theta}$ gave similar results. There are no indications that the approximations does not works well in this case. Due to the size of the GMRF, the comparison with results from long MCMC runs were performed on a cruder grid, with excellent results. We also compared the conditional marginals in the spatial field for fixed values of $\boldsymbol{\theta}$, and again obtained excellent results.

6 Extensions

6.1 Approximating the marginal likelihood

The marginal likelihood $\pi(\boldsymbol{y})$ is a useful quantity for comparing models, as the Bayes factor is its ratio for two competing models. It is evident from (1) that the natural approximation to the marginal likelihood is the normalising constant for $\tilde{\pi}(\boldsymbol{\theta}|\boldsymbol{y})$,

$$\tilde{\pi}(\boldsymbol{y}) = \int \frac{\pi(\boldsymbol{\theta}, \boldsymbol{x}, \boldsymbol{y})}{\tilde{\pi}_G(\boldsymbol{x}|\boldsymbol{\theta}, \boldsymbol{y})} \Big|_{\boldsymbol{x}=\boldsymbol{x}^*(\boldsymbol{\theta})} d\boldsymbol{\theta}. \quad (32)$$

where $\pi(\boldsymbol{\theta}, \boldsymbol{x}, \boldsymbol{y}) = \pi(\boldsymbol{\theta})\pi(\boldsymbol{x}|\boldsymbol{\theta})\pi(\boldsymbol{y}|\boldsymbol{x}, \boldsymbol{\theta})$. An alternative, cruder estimate of the marginal likelihood is obtained by assuming that $\boldsymbol{\theta}|\boldsymbol{y}$ is Gaussian; then (32) turns into some known constant times $|\boldsymbol{H}|^{-1/2}$, where \boldsymbol{H} is the Hessian matrix in Section 3.1, see Kass and Vaidyanathan (1992). Our approximation (32) does not require this assumption, since we treat $\tilde{\pi}(\boldsymbol{\theta}|\boldsymbol{y})$ in a ‘nonparametric’ way. This allows for taking into account the departure from Gaussianity which, for instance, appears clearly in Figure 5. We have limited experience with using (32) for computing Bayes factors, and for this reason, we have not stressed this issue in the examples. Friel and Rue (2007) use a similar expression as (32) to approximate the marginal likelihood in a different context.

6.2 Moderate number of hyperparameters

Integrating out the hyperparameters as described in Section 3.1 can be quite expensive if the number of hyperparameters, m , is not small but moderate, say, in the range of 6 to 12. Using, for example, $\delta_z = 1$ and $\delta_\pi = 2.5$, the integration scheme proposed in Section 3.1 will require, if $\boldsymbol{\theta}|\boldsymbol{y}$ is Gaussian, $\mathcal{O}(5^m)$ evaluation points. Even if we restrict ourselves to three evaluation points in each dimension, the cost $\mathcal{O}(3^m)$ is still exponential in m . In this section we will discuss an alternative approach which will reduce the computational cost dramatically for high m , but, at the same time, it will also reduce the accuracy of the numerical integration over $\tilde{\pi}(\boldsymbol{\theta}|\boldsymbol{y})$. The aim is to be able to provide useful results even when the number of hyperparameters is so large that the more direct approach in Section 3.1 is unfeasible.

Although many hyperparameters make the integration harder, it is often the case that increasing the number of hyperparameters increases also variability and the regularity, so that the integrand simplifies. Meaningful results can sometimes be obtained even using an extreme choice, that is using only the modal configuration to integrate over $\pi(\boldsymbol{\theta}|\boldsymbol{y})$. This ‘plug-in’ approach will obviously underestimate variability.

We can consider the integration problem as a design problem where we layout some ‘points’ in a m -dimensional space. Based on the measured response, we estimate the response surface at each point. As a first approximation, we can consider only response surfaces of second order, and use a classical quadratic design like the central-composite design (CCD) (Box and Wilson, 1951). A CCD contains an embedded factorial or fractional factorial design with centre points augmented with a group of $2m + 1$ ‘star points’ which allow for estimating the curvature. For $m = 5$, the design points are chosen (up to an arbitrary scaling) as

$$\begin{aligned} &(1, 1, 1, 1, 1), \quad (-1, 1, 1, 1, -1), \quad (1, -1, 1, 1, -1), \quad (-1, -1, 1, 1, 1), \\ &(1, 1, -1, 1, -1), \quad (-1, 1, -1, 1, 1), \quad (1, -1, -1, 1, 1), \quad (-1, -1, -1, 1, -1), \\ &(1, 1, 1, -1, -1), \quad (-1, 1, 1, -1, 1), \quad (1, -1, 1, -1, 1), \quad (-1, -1, 1, -1, -1), \\ &(1, 1, -1, -1, 1), \quad (-1, 1, -1, -1, -1), \quad (1, -1, -1, -1, -1) \quad \text{and} \quad (-1, -1, -1, -1, 1). \end{aligned}$$

They are all on the surface of the m dimensional sphere with radius \sqrt{m} . The star points consist of $2m$ points located along each axis at distance $\pm\sqrt{m}$ and the central point in the origin. For $m = 5$ this makes $n_p = 27$ points in total, which is small compared to $5^5 = 3,125$ or $3^5 = 243$. The number of design-points is 8 for $m = 3$, 16 for $m = 4$ and 5, 32 for $m = 6$, 64 for $m = 7$ and 8, 128 for $m = 9$, 10 and 11, and 256 from $m = 12$ to 17; see Sanchez and Sanchez (2005) for how to compute such designs. For all designs, there are additional $2m + 1$ star-points.

To determine the integration weights Δ_k in (3) and the scaling of the points, assume for simplicity that $\boldsymbol{\theta}|\mathbf{y}$ is standard Gaussian. We require that the integral of 1 equals 1, and that the integral of $\boldsymbol{\theta}^T \boldsymbol{\theta}$ equals m . This gives the integration weight for the points on the sphere with radius $f_0\sqrt{m}$

$$\Delta = \left[(n_p - 1) (f_0^2 - 1) \left\{ 1.0 + \exp\left(-\frac{mf_0^2}{2}\right) \right\} \right]^{-1}$$

where $f_0 > 1$ is any constant. The integration weight for the central point is $1 - (n_p - 1)\Delta$.

To validate the CCD integration, we recomputed the posterior marginal for the stochastic volatility model in Section 5.3 using Student- t_ν distributed observations, and for the semi-parametric ecological regression example in Section 5.4 using this integration method instead of the grid search in Section 3.1. The results were indeed positive. The predictions in Figure 5(f) for future x_t 's using the full dataset were nearly indistinguishable from those obtained using the CCD integration using only 1/15 of the computational cost. Same remarks apply for the results shown in Figure 6. The number of hyperparameters in these two cases is 4 and 3 respectively. Although this is not a large number, we should still be able to detect if the CCD integration is too rough, and this does not seem to be the case. Although the CDD integration is not as thoroughly verified as the INLA itself, the results obtained can be viewed as a ‘proof of concept’ at this stage. We hope to provide further empirical evidence that the CCD integration is adequate when the number of hyperparameters is moderate.

7 Discussion

We have presented a new approach to approximate posterior marginals in latent Gaussian models, based on integrated nested Laplace approximations (INLA). The results obtained are very encouraging: we obtain practically exact results over a wide range of commonly used latent Gaussian models. We also provide tools for assessing the approximation error, which are able to detect cases where the approximation bias is non-negligible; we note however that this seems to happen only in pathological cases.

We are aware that our work goes against a general trend of favouring ‘exact’ Monte Carlo methods over non-random approximations, as advocated for instance by Papaspiliopoulos et al. (2006) in the context of diffusions. Our point however is that, in the specific case of latent Gaussian models, the orders of magnitude involved in the computational cost of both approaches are such that this idealistic point of view is simply untenable for these models. As we said already, our approach provides precise estimates in seconds and minutes, even for models involving thousands of variables, in situations where any MCMC computation typically takes hours or even days.

The advantages of our approach are not only computational. It also allows for greater automation and parallel implementation. The core of the computational machinery is based on sparse matrix algorithms, which automatically adapt to any kind of latent field, e.g. 1D, 2D, 3D and so on. All the examples considered in this paper were computed using the same general code, with essentially no tuning. In practice, INLA can be used almost as a black box. The code is now part of `GMRFLib`-library (Rue and Held, 2005, Appendix) and available from the first author’s web page. With respect to parallel implementation, the INLA approach computes the approximation of $x_i|\boldsymbol{\theta}, \mathbf{y}$ independently for all i for fixed $\boldsymbol{\theta}$. Hence, parallel computing is trivial to implement. This is particularly important for spatial or spatio-temporal latent Gaussian models.

The main disadvantage of the INLA approach is that the computational cost is exponential in the number of hyperparameters m . In most applications m is small, but applications where m goes up to 10 do exist. This problem may be less severe than it appears at first glance: the central composite design approach seems

promising, and provides reasonable results when m is not small, but this track needs more research. In fact, we doubt that any MCMC algorithm which would explore the m -dimensional space of θ in a random fashion would provide more accurate results for the same cost.

It is our view that the prospects of this work are more important than this work itself. Near instant inference will make latent Gaussian models more applicable, useful and appealing for the end user, which has no time or patience to wait for the results of an MCMC algorithm, or has to analyse many different dataset with the same model, or both. Further, near instant inference makes it much easier to challenge the model itself: Bayes factors can be computed through the normalising constant for $\tilde{\pi}(\theta|\mathbf{y})$, the model can be assessed through cross-validation, residual analysis, etc., in a reasonable time.

Acknowledgement

The authors acknowledge Jo Eidsvik, Nial Friel, Arnoldo Frigessi, John Hasslet, Leonhard Held, Hanne W. Rognebakke, Judith Rousseau, Håkon Tjelmeland, John Tyssedal and Rasmus Waagepetersen for stimulating discussions related to this work, and the Center for Tropical Forest Science of the Smithsonian Tropical Research Institute for providing the data in Section 5.5.

A Fitting the skew-Normal distribution

We explain here how to fit the skew-Normal distribution (22) to an expansion of the form

$$\log \pi(x) = \text{constant} - \frac{1}{2}x^2 + \gamma^{(1)}x + \frac{1}{6}\gamma^{(3)}x^3 + \dots \quad (33)$$

To second order, (33) is Gaussian with mean $\gamma^{(1)}$ and variance 1. The mean and the variance of the skew-Normal distribution are $\xi + \omega\delta\sqrt{2/\pi}$ and $\omega^2(1 - 2\delta^2/\pi)$, respectively, where $\delta = a/\sqrt{1+a^2}$. We keep these fixed to $\gamma^{(1)}$ and 1, respectively, but adjust a so the third derivative at the mode in (22) equals $\gamma^{(3)}$. This gives three equations to determine (ξ, ω, a) . The modal configuration is not available analytically, but a series expansion of the log skew-Normal density around $x = \xi$ gives:

$$x^* = \left(\frac{a}{\omega}\right) \frac{\sqrt{2\pi} + 2\xi\left(\frac{a}{\omega}\right)}{\pi + 2\left(\frac{a}{\omega}\right)^2} + \text{higher order terms.}$$

We now compute the third derivative of the log-density of the skew-Normal at x^* . In order to obtain an analytical (and computationally fast) fit, we expand this third order derivative with respect to a/ω :

$$\frac{\sqrt{2}(4-\pi)}{\pi^{3/2}} \left(\frac{a}{\omega}\right)^3 + \text{higher order terms.} \quad (34)$$

and imposes that (34) equals $\gamma^{(3)}$. This gives explicit formulae for the three parameters of the skewed-normal.

References

- Ainsworth, L. M. and Dean, C. B. (2006). Approximate inference for disease mapping. *Computational Statistics & Data Analysis*, 50(10):2552–2570.
- Albert, J. H. and Chib, S. (1993). Bayesian analysis of binary and polychotomous response data. *Journal of the American Statistical Association*, 88(422):669–679.
- Allcroft, D. J. and Glasbey, C. A. (2003). A latent Gaussian Markov random field model for spatio-temporal rainfall disaggregation. *Journal of the Royal Statistical Society, Series C*, 52:487–498.
- Azzalini, A. and Capitanio, A. (1999). Statistical applications of the multivariate skew-normal distribution. *Journal of the Royal Statistical Society, Series B*, 61(4):579–602.

- Banerjee, S., Carlin, B. P., and Gelfand, A. E. (2004). *Hierarchical Modeling and Analysis for Spatial Data*, volume 101 of *Monographs on Statistics and Applied Probability*. Chapman & Hall, London.
- Bartolucci, F. and Besag, J. (2002). A recursive algorithm for Markov random fields. *Biometrika*, 89:724–730.
- 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.
- Besbeas, P., Feis, I. D., and Sapatinas, T. (2004). A comparative simulation study of wavelet shrinkage estimators for poisson counts. *International Statistical Review*, 72(2):209–237.
- Biller, C. and Fahrmeir, L. (1997). Bayesian spline-type smoothing in generalized regression models. *Computational Statistics*, 12:135–151.
- Box, G. E. P. and Wilson, K. B. (1951). On the experimental attainment of optimum conditions (with discussion). *Journal of the Royal Statistical Society, Series B*, 13(1):1–45.
- Breslow, N. E. and Clayton, D. G. (1993). Approximate inference in generalized linear mixed models. *Journal of the American Statistical Association*, 88(1):9–25.
- Brix, A. and Diggle, P. J. (2001). Spatiotemporal prediction for log-Gaussian Cox processes. *Journal of the Royal Statistical Society, Series B*, 63(4):823–841.
- Brix, A. and Møller, J. (2001). Space-time multi type log Gaussian Cox processes with a view to modelling weeds. *Scandinavian Journal of Statistics*, 28:471–488.
- Carlin, B. P. and Banerjee, S. (2003). Hierarchical multivariate CAR models for spatio-temporally correlated survival data (with discussion). In *Bayesian Statistics*, 7, pages 45–63. Oxford Univ. Press, New York.
- Carter, C. K. and Kohn, R. (1994). On Gibbs sampling for state space models. *Biometrika*, 81(3):541–543.
- Chaudhuri, P. and Marron, J. S. (1999). SiZer for exploration of structures in curves. *Journal of the American Statistical Association*, 94:807–823.
- Chib, S., Nardari, F., and Shepard, N. (2002). Markov chain Monte Carlo methods for stochastic volatility models. *Journal of Econometrics*, 108:281–316.
- Cressie, N. A. C. (1993). *Statistics for spatial data*. Wiley Series in Probability and Mathematical Statistics: Applied Probability and Statistics. John Wiley & Sons Inc., New York. Revised reprint of the 1991 edition, A Wiley-Interscience Publication.
- Diggle, P. J. and Ribeiro, P. J. (2006). *Model-based Geostatistics*. Springer Series in Statistics. Springer.
- 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.
- Durbin, J. and Koopman, S. J. (2000). Time series analysis of non-Gaussian observations based on state space models from both classical and Bayesian perspectives (with discussion). *Journal of the Royal Statistical Society, Series B*, 62(1):3–56.
- Eidsvik, J., Martino, S., and Rue, H. (2006). Approximate bayesian inference in spatial generalized linear mixed models. Technical Report no 2, Department of mathematical sciences, Norwegian University of Science and Technology.
- Erästö, P. (2005). *Studies in trend detection of scatter plots with visualization*. PhD thesis, Department of Mathematics and Statistics, University of Helsinki, Finland.
- Fahrmeir, L. and Lang, S. (2001). Bayesian inference for generalized additive mixed models based on Markov random field priors. *Journal of the Royal Statistical Society, Series C*, 50(2):201–220.
- Fahrmeir, L. and Tutz, G. (2001). *Multivariate Statistical Modelling Based on Generalized Linear Models*. Springer-Verlag, Berlin, 2nd edition.
- Friel, N. and Rue, H. (2007). Recursive computing and simulation-free inference for general factorizable models. *Biometrika*, xx(xx):xx–xx. (in revision).

- Frühwirth-Schnatter, S., , and Wagner, H. (2006). Auxiliary mixture sampling for parameter-driven models of time series of small counts with applications to state space modelling. *Biometrika*, 93(4):827–841.
- Frühwirth-Schnatter, S. and Frühwirth, R. (2007). Auxiliary mixture sampling with applications to logistic models. *Computational Statistics & Data Analysis*, xx(xx):xx–xx. (to appear).
- Gamerman, D. (1997). Sampling from the posterior distribution in generalized linear mixed models. *Statistics and Computing*, 7(1):57–68.
- Gamerman, D. (1998). Markov chain Monte Carlo for dynamic generalised liner models. *Biometrika*, 85(1):215–227.
- Hannig, J. and Marron, J. S. (2006). Advanced distribution theory for SiZer. *Journal of the American Statistical Association*, 101:254–269.
- Holmes, C. C. and Held, L. (2006). Bayesian auxiliary variable models for binary and multinomial regression. *Bayesian Analysis*, 1(1):145–168.
- Hsiao, C. K., Huang, S. Y., and Chang, C. W. (2004). Bayesian marginal inference via cadidate’s formula. *Statistics and Computing*, 14(1):59–66.
- Kass, R. E. and Vaidyanathan, S. K. (1992). Approximate Bayes factors and orthogonal parameters, with application to testing equality of two binomial proportions. *Journal of the Royal Statistical Society, Series B*, 54(1):129–144.
- Kitagawa, G. and Gersch, W. (1996). *Smoothness Priors Analysis of Time Series*. Lecture Notes in Statistics no. 116. Springer-Verlag, New York.
- Knorr-Held, L. (1999). Conditional prior proposals in dynamic models. *Scandinavian Journal of Statistics*, 26(1):129–144.
- Knorr-Held, L. (2000). Bayesian modelling of inseparable space-time variation in disease risk. *Statistics in Medicine*, 19(17-18):2555–2567.
- Knorr-Held, L. and Besag, J. (1998). Modelling risk from a disease in time and space. *Statistics in Medicine*, 17(18):2045–2060.
- 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.
- Lang, S. and Brezger, A. (2004). Bayesian P-splines. *Journal of Computational and Graphical Statistics*, 13(1).
- Møller, J., Syversveen, A. R., and Waagepetersen, R. P. (1998). Log Gaussian Cox processes. *Scandinavian Journal of Statistics*, 25:451–482.
- Møller, J. and Waagepetersen, R. (2003). *Statistical inference and simulation for spatial point processes*, volume 100 of *Monographs on Statistics and Applied Probability*. Chapman & Hall, London.
- Nataro, I. and Knorr-Held, L. (2003). Non-parametric ecological regression and spatial variation. *Biometrical Journal*, 45:670–688.
- Øigård, T. A., Rue, H., and Godtlielsen, F. (2006). Bayesian multiscale analysis for time series data. *Computational Statistics & Data Analysis*, 51(3):1719–1730.
- Papaspiliopoulos, A. B. O., Roberts, G. O., and Fearnhead, P. (2006). Exact and computationally efficient likelihood-based estimation for discretely observed diffusion processes (with discussion). *Journal of the Royal Statistical Society, Series B*, 68(3):333–382.
- Papaspiliopoulos, O., Roberts, G. O., and Sköld, M. (2007). A general framework for the parameterisation of hierarchical models. *Statistical Science*, xx(xx):xx–xx. (to appear).
- Reeves, R. and Pettitt, A. N. (2004). Efficient recursions for general factorisable models. *Biometrika*, 91(3):751–757.
- 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. and Martino, S. (2007). Approximate Bayesian inference for hierarchical Gaussian Markov random fields models. *Journal of Statistical Planning and Inference*, xx(xx):xx–xx. (to appear).
- 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.
- Sanchez, S. M. and Sanchez, P. J. (2005). Very large fractional factorials and central composite designs. *ACM Transactions on Modeling and Computer Simulation*, 15(4):362–377.
- Schervish, M. J. (1995). *Theory of statistics*. Springer series in statistics. Springer-Verlag, New York, 2nd edition.
- Shephard, N. (1994). Partial non-Gaussian state space. *Biometrika*, 81(1):115–131.
- Shephard, N. and Pitt, M. K. (1997). Likelihood analysis of non-Gaussian measurement time series. *Biometrika*, 84(3):653–667.
- Shun, Z. and McCullagh, P. (1995). Laplace approximation of high dimensional integrals. *Journal of the Royal Statistical Society, Series B*, 57(4):749–760.
- Spiegelhalter, D. J., Best, N. G., Carlin, B. P., and van der Linde, A. (2002). Bayesian measures of model complexity and fit (with discussion). *Journal of the Royal Statistical Society, Series B*, 64(2):583–639.
- Terzopoulos, D. (1988). The computation of visible-surface representations. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 10(4):417–438.
- 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.
- Waagepetersen, R. P. (2007). An estimating function approach to inference for inhomogeneous Neyman-Scott processes. *Biometrics*, xx(xx):xx–xx. (to appear).
- Wecker, W. E. and Ansley, C. F. (1983). The signal extraction approach to nonlinear regression and spline smoothing. *Journal of the American Statistical Association*, 78(381):81–89.
- Weir, I. S. and Pettitt, A. N. (2000). Binary probability maps using a hidden conditional autoregressive Gaussian process with an application to Finnish common toad data. *Journal of the Royal Statistical Society, Series C*, 49(4):473–484.
- Wikle, C. K., Berliner, L. M., and Cressie, N. A. C. (1998). Hierarchical Bayesian space-time models. *Environmental and Ecological Statistics*, 5(2):117–154.