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

Håvard Rue and Sara Martino†

The Norwegian University for Science and Technology, Trondheim, Norway

Nicolas Chopin

CREST-LS and ENSAE, Paris, France

Summary. Structured additive regression models are perhaps the most commonly used class of models in statistical applications. It includes, among others, (generalised) linear models, (generalised) additive models, smoothing-spline models, state-space models, semiparametric regression, spatial and spatio-temporal models, log-Gaussian Cox-processes, geostatistical and geoadditive models. In this paper we consider approximate Bayesian inference in a popular subset of structured additive regression models, *latent Gaussian models*, where the latent field is Gaussian, controlled by a few hyperparameters and with non-Gaussian response variables.

The posterior marginals are not available in closed form due to the non-Gaussian response variables. 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 not an appropriate tool for routine analysis.

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 with our approach is its generality, which makes it possible to perform Bayesian analysis in an automatic, streamlined way, and to compute model comparison criteria and various predictive measures so that models can be compared and the model under study can be challenged.

1. Introduction

1.1. Aim of the paper

This paper discusses how to perform approximate Bayesian inference in a subclass of structured additive regression models, named *latent Gaussian models*. Structured additive regression models are a flexible and extensive used class of models, see for example Fahrmeir and Tutz (2001) for a detailed account. In these models, the observation (or response) variable y_i is assumed to belong to an exponential family, where the mean μ_i is linked to

†*Address for correspondence:* Håvard Rue, Department of Mathematical Sciences, The Norwegian University for Science and Technology, N-7491 Trondheim, Norway.
E-mail: hrue@math.ntnu.no

a structured additive predictor η_i through a link-function $g(\cdot)$, so that $g(\mu_i) = \eta_i$. The structured additive predictor η_i accounts for effects of various covariates in an additive way:

$$\eta_i = \alpha + \sum_{j=1}^{n_f} f^{(j)}(u_{ji}) + \sum_{k=1}^{n_\beta} \beta_k z_{ki} + \epsilon_i. \tag{1}$$

Here, the $\{f^{(j)}(\cdot)\}$'s are unknown functions of the covariates $\boldsymbol{u}$, the $\{\beta_k\}$'s represent the linear effect of covariates $\boldsymbol{z}$ and the ϵ_i's are unstructured terms. This class of model has a wealth of applications, thanks to the very different forms that the unknown functions $\{f^{(j)}\}$ can take. Latent Gaussian models are a subset of all Bayesian additive models with a structured additive predictor (1); namely those which assign a Gaussian prior to α, $\{f^{(j)}(\cdot)\}$, $\{\beta_k\}$ and $\{\epsilon_t\}$. Let $\boldsymbol{x}$ denote the vector of all the latent Gaussian variables, and $\boldsymbol{\theta}$ the vector of hyperparameters, which are not necessarily Gaussian. In the machine learning literature, the phrase 'Gaussian process models' is often used (Rasmussen and Williams, 2006). We discuss various applications of latent Gaussian models in Section 1.2.

The main aim of this paper is twofold:

(a) To provide accurate and fast deterministic approximations to all, or some of, the n posterior marginals for x_i, the components of latent Gaussian vector $\boldsymbol{x}$, plus possibly the posterior marginals for $\boldsymbol{\theta}$ or some of its components θ_j. If needed, the marginal densities can be post-processed to compute quantities like posterior expectations, variances and quantiles.

(b) To demonstrate how to use these marginals in order *i*) to provide adequate approximations to the posterior marginal for sub-vectors $\boldsymbol{x}_S$ for any subset S, *ii*) to compute the marginal likelihood and the Deviance Information Criteria (DIC) for model comparison, and *iii*) to compute various Bayesian predictive measures.

1.2. Latent Gaussian Models: Applications

Latent Gaussian models have a numerous and wide ranging list of applications; most structured Bayesian models are in fact of this form; see for example the books by Fahrmeir and Tutz (2001), Gelman et al. (2004) and Robert and Casella (1999). We will first give some areas of applications grouped according to their physical dimension. Let $f(\cdot)$ denote one of the $f^{(j)}(\cdot)$-terms in (1) with variables $f_1, f_2, \ldots$.

Regression models Bayesian generalised linear models correspond to the linear predictor $\eta_i = \alpha + \sum_{k=1}^{n_\beta} \beta_k z_{ki}$ (Dey et al., 2000). The $f(\cdot)$-terms are used either to relax the linear relationship of the covariate as argued for by Fahrmeir and Tutz (2001), or to introduce random effects, or both. Popular models for modelling smooth effects of covariates are P-spline models (Lang and Brezger, 2004) and random walk models (Fahrmeir and Tutz, 2001; Rue and Held, 2005), or continuous indexed spline models (Wahba, 1978; Wecker and Ansley, 1983; Kohn and Ansley, 1987; Rue and Held, 2005) or Gaussian processes (O'Hagan, 1978; Chu and Ghahramani, 2005; Williams and Barber, 1998; Besag et al., 1995; Neal, 1998). Random effects make it possible to account for overdispersion caused by unobserved heterogeneity, or for correlation in longitudinal data, and can be introduced by defining $f(u_i) = f_i$ and letting $\{f_i\}$ be independent, zero-mean, and Gaussian (Fahrmeir and Lang, 2001).

Dynamic models Temporal dependency can be introduced by using i in (1) as a time index t and defining $f(\cdot)$ and covariate $\boldsymbol{u}$ so that $f(u_t) = f_t$. Then $\{f_t\}$ can model a discrete-time or continuous-time auto-regressive model, a seasonal effect or more generally the latent process of a structured time-series model (Kitagawa and Gersch, 1996; West and Harrison, 1997). Alternatively, $\{f_t\}$ can represent a smooth temporal function in the same spirit as regression models.

Spatial and spatio-temporal models Spatial dependency can be modelled similarly, using a spatial covariate $\boldsymbol{u}$ so that $f(u_s) = f_s$, where s represents the spatial location or spatial region s. The stochastic model for f_s is constructed to promote spatial smooth realisations of some kind. Popular models include the BYM model for disease-mapping with extensions for regional data (Besag et al., 1991; Held et al., 2005; Weir and Pettitt, 2000; Gschlößl and Czado, 2007; Wakefield, 2007), continuous-indexed Gaussian models (Banerjee et al., 2004; Diggle and Ribeiro, 2006), texture models (Marroquin et al., 2001; Rellier et al., 2002). Spatial and temporal dependencies can be achieved either by using a spatio-temporal covariate (s, t) or a corresponding spatio-temporal Gaussian field (Kamman and Wand, 2003; Cressie and Johannesson, 2008; Banerjee et al., 2008; Finkenstadt et al., 2006; Abellan et al., 2007; Gneiting, 2002; Banerjee et al., 2004).

In many applications, the final model may consist of a sum of various components, such as a spatial component, random effects, and both linear and smooth effects of some covariates. Furthermore, linear or sum-to-zero constraints are sometimes imposed as well in order to separate out the effects of various components in (1).

1.3. Latent Gaussian Models: Notation and Basic Properties

To simplify the following discussion, denote generically $\pi(\cdot|\cdot)$ as the conditional density of its arguments, and let $\boldsymbol{x}$ be all the n Gaussian variables $\{\eta_i\}$, α, $\{f^{(j)}\}$, $\{\beta_k\}$. The density $\pi(\boldsymbol{x}|\boldsymbol{\theta}_1)$ is Gaussian with (assumed) zero mean, precision matrix $\boldsymbol{Q}(\boldsymbol{\theta}_1)$ with hyperparameters $\boldsymbol{\theta}_1$. Denote by $\mathcal{N}(\boldsymbol{x};\ \boldsymbol{\mu},\ \boldsymbol{\Sigma})$, the $\mathcal{N}(\boldsymbol{\mu},\ \boldsymbol{\Sigma})$ Gaussian density with mean $\boldsymbol{\mu}$ and covariance (inverse precision) $\boldsymbol{\Sigma}$ at configuration $\boldsymbol{x}$. Note that we have included $\{\eta_i\}$ instead of $\{\epsilon_i\}$ into $\boldsymbol{x}$, as it simplifies the notation later on.

The distribution for the n_d observational variables $\boldsymbol{y} = \{y_i : i \in \mathcal{I}\}$ is denoted by $\pi(\boldsymbol{y}|\boldsymbol{x}, \boldsymbol{\theta}_2)$ and we assume that $\{y_i : i \in \mathcal{I}\}$ are conditionally independent given $\boldsymbol{x}$ and $\boldsymbol{\theta}_2$. For simplicity, denote by $\boldsymbol{\theta} = (\boldsymbol{\theta}_1^T, \boldsymbol{\theta}_2^T)^T$ with $\dim(\boldsymbol{\theta}) = m$. The posterior then reads (for a non-singular $\boldsymbol{Q}(\boldsymbol{\theta})$)

$$\begin{aligned}
\pi(\boldsymbol{x}, \boldsymbol{\theta} \mid \boldsymbol{y}) &\propto \pi(\boldsymbol{\theta})\ \pi(\boldsymbol{x} \mid \boldsymbol{\theta})\ \prod_{i \in \mathcal{I}} \pi(y_i \mid x_i, \boldsymbol{\theta}) \\
&\propto \pi(\boldsymbol{\theta})\ |\boldsymbol{Q}(\boldsymbol{\theta})|^{n/2} \exp\left(-\frac{1}{2}\boldsymbol{x}^T \boldsymbol{Q}(\boldsymbol{\theta})\boldsymbol{x} + \sum_{i \in \mathcal{I}} \log \pi(y_i|x_i, \boldsymbol{\theta})\right).
\end{aligned}$$

The imposed linear constraints (if any) are denoted by $\boldsymbol{A}\boldsymbol{x} = \boldsymbol{e}$ for a $k \times n$ matrix $\boldsymbol{A}$ of rank k. The main aim is to approximate the posterior marginals $\pi(x_i|\boldsymbol{y})$, $\pi(\boldsymbol{\theta}|\boldsymbol{y})$ and $\pi(\theta_j|\boldsymbol{y})$.

Many, but not all, latent Gaussian models in the literature (see Section 1.2) satisfy two basic properties which we shall assume throughout the paper. The first is that the latent field $\boldsymbol{x}$, which is often of large dimension, $n = 10^2 - 10^5$, admit conditional independence

properties. Hence, the latent field is a Gaussian Markov random field (GMRF) with a sparse precision matrix $\boldsymbol{Q}(\boldsymbol{\theta})$ (Rue and Held, 2005). This means that we can use numerical methods for sparse matrices, which are much quicker than general dense matrix calculations (Rue and Held, 2005). The second property is that the number of hyperparameters m, is small, say $m \leq 6$. Both properties are usually required to produce fast inference, but exceptions exist (Eidsvik et al., 2008).

1.4. 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 $\boldsymbol{x}$ are strongly dependent on each other. Second, $\boldsymbol{\theta}$ and $\boldsymbol{x}$ 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 $\boldsymbol{x}$ (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 $\boldsymbol{x}$. One suggestion is to use the one-block approach of Knorr-Held and Rue (2002): make a proposal for $\boldsymbol{\theta}$ to $\boldsymbol{\theta}'$, update $\boldsymbol{x}$ 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 and Wagner, 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.

1.5. 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 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 $\boldsymbol{x}$ is typically well-behaved, and looks 'almost' Gaussian. This is clearly due to the latent Gaussian prior assigned to $\boldsymbol{x}$, which has a non-negligible impact on the posterior, especially in terms of dependence between the components of $\boldsymbol{x}$.

1.6. Approximation methods in Machine Learning

A general approach towards approximate inference is the variational Bayes (VB) methodology developed in the machine learning literature (Hinton and van Camp, 1993; MacKay, 1995; Bishop, 2006). VB has provided numerous promising results in various areas, like hidden Markov models (MacKay, 1997), mixture models (Humphreys and Titterington, 2000), graphical models (Attias, 1999, 2000), state-space models (Beal, 2003), among others; see Beal (2003), Titterington (2004) and Jordan (2004) for extensive reviews.

For the sake of discussion, consider the posterior distribution $\pi(\boldsymbol{x}, \boldsymbol{\theta}|\boldsymbol{y})$ of a generic Bayesian model, with observation $\boldsymbol{y}$, latent variable $\boldsymbol{x}$, and hyperparameter $\boldsymbol{\theta}$. The principle of VB is to use as an approximation the joint density $q(\boldsymbol{x}, \boldsymbol{\theta})$ that minimises the Kullback-Leibler contrast of $\pi(\boldsymbol{x}, \boldsymbol{\theta}|\boldsymbol{y})$ with respect to $q(\boldsymbol{x}, \boldsymbol{\theta})$. The minimisation is subject to some constraint on $q(\boldsymbol{x}, \boldsymbol{\theta})$, most commonly: $q(\boldsymbol{x}, \boldsymbol{\theta}) = q_{\boldsymbol{x}}(\boldsymbol{x}) q_{\boldsymbol{\theta}}(\boldsymbol{\theta})$. Obviously, the VB approximated density $q(\boldsymbol{x}, \boldsymbol{\theta})$ does not capture the dependence between $\boldsymbol{x}$ and $\boldsymbol{\theta}$, but one hopes that its marginals (of $\boldsymbol{x}$ and $\boldsymbol{\theta}$) approximate well the true posterior marginals. The solution of this minimisation problem is approached through an iterative, EM-like algorithm.

In general, the VB approach is not without potential problems. First, even though VB seems often to approximate well the posterior mode (Wang and Titterington, 2006), the posterior variance can be (sometimes severely) under-estimated; see Bishop (2006, Chap. 10) and Wang and Titterington (2005). In the case of latent Gaussian models, this phenomenon does occur as we demonstrate in Appendix A; we show that the VB approximated variance can be up to n times smaller than the true posterior variance in a typical application. The second potential problem is that the iterative process of the basic VB algorithm is tractable for 'conjugate-exponential' models only (Beal, 2003). This implies that $\pi(\boldsymbol{\theta})$ must be conjugate with respect to the complete likelihood $\pi(\boldsymbol{x}, \boldsymbol{y}|\boldsymbol{\theta})$ and the complete likelihood must belong to an exponential family. However, few of the latent Gaussian models encountered in applications are of this type, as illustrated by our worked-through examples in Section 5. A possible remedy around this requirement is to impose restrictions on $q(\boldsymbol{x}, \boldsymbol{\theta})$, such as independence between blocks of components of $\boldsymbol{\theta}$ (Beal, 2003, Ch. 4), or a parametric form for $q(\boldsymbol{x}, \boldsymbol{\theta})$ that allow for a tractable minimisation algorithm. However, this requires case-specific solutions, and the constraints will increase the approximation error.

Another approximation scheme popular in Machine Learning is the Expectation-Propagation (EP) approach (Minka, 2001); see e.g. Zoeter et al. (2005) and Kuss and Rasmussen (2005)

for applications of EP to latent Gaussian models. EP follows principles which are somewhat similar to VB, i.e. minimises iteratively some pseudo-distance between $\pi(\boldsymbol{x}, \boldsymbol{\theta}|\boldsymbol{y})$ and the approximation $q(\boldsymbol{x}, \boldsymbol{\theta})$, subject to $q(\boldsymbol{x}, \boldsymbol{\theta})$ factorising in a 'simple' way, e.g. as a product of parametric factors, each involving a single component of $(\boldsymbol{x}, \boldsymbol{\theta})$. However, the pseudo-distance used in EP is the Kullback-Leibler contrast of $q(\boldsymbol{x}, \boldsymbol{\theta})$ relative to $\pi(\boldsymbol{x}, \boldsymbol{\theta}|\boldsymbol{y})$, rather than the other way around (as in VB). Because of this, EP usually over-estimates the posterior variance (Bishop, 2006, Chap. 10). Kuss and Rasmussen (2005) derives an EP approximation scheme for classification problems involving Gaussian processes that seems to be accurate and fast; but their focus is on approximating $\pi(\boldsymbol{x}|\boldsymbol{\theta}, \boldsymbol{y})$ for $\boldsymbol{\theta}$ set to the posterior mode, and it is not clear how to extend this approach to a fully Bayesian analysis. More importantly, deriving an efficient EP algorithm seems to require specific efforts for each class of models. With respect to computational cost, VB and EP are both designed to be faster than exact MCMC methods, but, due to their iterative nature, they are (much) slower than analytic approximations (such as those developed in this paper); see Section 5.3 for an illustration of this in one of our examples. Also, it is not clear whether EP and VB can be implemented efficiently in scenarios involving linear constraints on $\boldsymbol{x}$.

The general applicability of the VB and EP approaches does not contradict the existence of improved approximation schemes for latent Gaussian models, hopefully without the problems just discussed. How this can be done is described next.

1.7. Inference: The new approach

The posterior marginals of interests can be written as

$$\pi(x_i \mid \boldsymbol{y}) = \int \pi(x_i \mid \boldsymbol{\theta}, \boldsymbol{y})\, \pi(\boldsymbol{\theta} \mid \boldsymbol{y})\, d\boldsymbol{\theta} \quad \text{and} \quad \pi(\theta_j \mid \boldsymbol{y}) = \int \pi(\boldsymbol{\theta} \mid \boldsymbol{y})\, d\boldsymbol{\theta}_{-j},$$

and key feature of our new approach is to use this form to construct nested approximations

$$\widetilde{\pi}(x_i \mid \boldsymbol{y}) = \int \widetilde{\pi}(x_i \mid \boldsymbol{\theta}, \boldsymbol{y})\, \widetilde{\pi}(\boldsymbol{\theta} \mid \boldsymbol{y})\, d\boldsymbol{\theta} \quad \text{and} \quad \widetilde{\pi}(\theta_j \mid \boldsymbol{y}) = \int \widetilde{\pi}(\boldsymbol{\theta} \mid \boldsymbol{y})\, d\boldsymbol{\theta}_{-j}. \tag{2}$$

Here, $\widetilde{\pi}(\cdot|\cdot)$ is an approximated (conditional) density of its arguments. Approximations to $\pi(x_i|\boldsymbol{y})$ are computed by approximating $\pi(\boldsymbol{\theta}|\boldsymbol{y})$ and $\pi(x_i|\boldsymbol{\theta}, \boldsymbol{y})$, and using numerical integration (i.e. a finite sum) to integrate out $\boldsymbol{\theta}$. The integration is possible as the dimension of $\boldsymbol{\theta}$ is small, see Section 1.3. As it will become clear in the following, the nested approach makes Laplace approximations very accurate when applied to latent Gaussian models. The approximation of $\pi(\theta_j|\boldsymbol{y})$ is computed by integrating out $\boldsymbol{\theta}_{-j}$ from $\widetilde{\pi}(\boldsymbol{\theta}|\boldsymbol{y})$; we return in Section 3.1 to the practical details.

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

$$\widetilde{\pi}(\boldsymbol{\theta} \mid \boldsymbol{y}) \propto \left. \frac{\pi(\boldsymbol{x}, \boldsymbol{\theta}, \boldsymbol{y})}{\widetilde{\pi}_{\mathrm{G}}(\boldsymbol{x}|\boldsymbol{\theta}, \boldsymbol{y})} \right|_{\boldsymbol{x}=\boldsymbol{x}^\star(\boldsymbol{\theta})} \tag{3}$$

where $\widetilde{\pi}_{\mathrm{G}}(\boldsymbol{x}|\boldsymbol{\theta}, \boldsymbol{y})$ is the Gaussian approximation to the full conditional of $\boldsymbol{x}$, and $\boldsymbol{x}^\star(\boldsymbol{\theta})$ is the mode of the full conditional for $\boldsymbol{x}$, for a given $\boldsymbol{\theta}$. The proportionality sign (3) comes from the fact that the normalising constant for $\pi(\boldsymbol{x}, \boldsymbol{\theta}|\boldsymbol{y})$ 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 are not verified here; see Section 4 for a discussion of the error rate.

Note that $\widetilde{\pi}(\boldsymbol{\theta}|\boldsymbol{y})$ itself tends to depart significantly from Gaussianity. This suggests that a cruder approximation based on a Gaussian approximation to $\pi(\boldsymbol{\theta}|\boldsymbol{y})$ is not accurate enough for our purposes; this also applies to similar approximations based on 'equivalent Gaussian observations' around $\boldsymbol{x}^\star$, and evaluated at the mode of (3) (Breslow and Clayton, 1993; Ainsworth and Dean, 2006). A critical aspect of our approach is to explore and manipulate $\widetilde{\pi}(\boldsymbol{\theta}|\boldsymbol{y})$ and $\widetilde{\pi}(x_i|\boldsymbol{y})$ in a 'nonparametric' way. Rue and Martino (2007) used (3) to approximate posterior marginals for $\boldsymbol{\theta}$ for various latent Gaussian models. Their conclusion was that $\widetilde{\pi}(\boldsymbol{\theta}|\boldsymbol{y})$ 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 $\widetilde{\pi}_{\mathrm{G}}(\boldsymbol{x}|\boldsymbol{\theta},\boldsymbol{y})$ and approximate the density of $x_i|\boldsymbol{\theta},\boldsymbol{y}$ with the Gaussian marginal derived from $\widetilde{\pi}_{\mathrm{G}}(\boldsymbol{x}|\boldsymbol{\theta},\boldsymbol{y})$, i.e.

$$\widetilde{\pi}(x_i \mid \boldsymbol{\theta},\boldsymbol{y}) = \mathcal{N}\left\{x_i;\ \mu_i(\boldsymbol{\theta}), \sigma_i^2(\boldsymbol{\theta})\right\}. \tag{4}$$

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 $\boldsymbol{\theta}$, see (2), to obtain approximations of the marginals of interest for the latent field,

$$\widetilde{\pi}(x_i \mid \boldsymbol{y}) = \sum_k \widetilde{\pi}(x_i \mid \boldsymbol{\theta}_k, \boldsymbol{y}) \ \times\ \widetilde{\pi}(\boldsymbol{\theta}_k \mid \boldsymbol{y}) \ \times\ \Delta_k. \tag{5}$$

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 (4) was higher. In particular, (4) can present an error in location and/or a lack of skewness. Other issues in Rue and Martino (2007) were both the difficulty to detect the x_i's whose approximation is less accurate and the inability to improve the approximation at those locations. Moreover, they were unable to control the error of the approximations and to chose the integration points $\{\boldsymbol{\theta}_k\}$ in an adaptive and automatic way.

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 $\pi(x_i|\boldsymbol{y},\boldsymbol{\theta})$. We also present a faster alternative which corrects the Gaussian approximation (4) for error in the location and lack of skewness at moderate extra cost. The corrections are obtained by a series expansion 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 will also demonstrate how the various approximations can be used to derive tools for assessing the approximation error, approximate posterior marginals for a subset of $\boldsymbol{x}$, and to compute interesting quantities like the marginal likelihood, the Deviance Information Criteria and various Bayesian predictive measures.

1.8. Plan of paper

Section 2 contains preliminaries on GMRF's, sparse matrix computations and Gaussian approximations. Section 3 explains the INLA approach and how to approximate $\pi(\boldsymbol{\theta}|\boldsymbol{y})$, $\pi(\theta_j|\boldsymbol{y})$ and $\pi(x_i|\boldsymbol{\theta},\boldsymbol{y})$. For the latent field, 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 stochastic volatility models, a longitudinal mixed model, a spatial model for mapping of cancer incidence data and spatial log-Gaussian Cox processes. Section 6 discusses some extensions: construction of posterior marginals for subsets $\boldsymbol{x}_S$, approximations of the marginal likelihood and predictive measures, the DIC criterion for model comparison 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 GMRF's 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 $\boldsymbol{x}_{-i}$ the vector $\boldsymbol{x}$ minus its ith element and by $\Gamma(\tau;\ a, b)$ the $\Gamma(a, b)$ density (with mean a/b) at point τ.

2.1. Gaussian Markov Random Fields

A GMRF is a Gaussian random variable $\boldsymbol{x} = (x_1, \ldots, x_n)$ with Markov properties: for some $i \neq j$'s, x_i and x_j are independent conditional upon $\boldsymbol{x}_{-ij}$. These Markov properties are conveniently encoded in the precision (inverse covariance) matrix $\boldsymbol{Q}$: $Q_{ij} = 0$ if and only if x_i and x_j are independent conditional upon $\boldsymbol{x}_{-ij}$. Let the undirected graph $\mathcal{G}$ denote the conditional independence properties of $\boldsymbol{x}$, then $\boldsymbol{x}$ is said to be a GMRF with respect to $\mathcal{G}$. If the mean of $\boldsymbol{x}$ is $\boldsymbol{\mu}$, the density of $\boldsymbol{x}$ is

$$\pi(\boldsymbol{x}) = (2\pi)^{-n/2}|\boldsymbol{Q}|^{1/2} \exp\left\{-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^T\boldsymbol{Q}(\boldsymbol{x}-\boldsymbol{\mu})\right\}. \tag{6}$$

In most cases only $\mathcal{O}(n)$ of the n^2 entries of $\boldsymbol{Q}$ are non-zero, so $\boldsymbol{Q}$ is sparse. This allows for fast factorisation of $\boldsymbol{Q}$ as $\boldsymbol{L}\boldsymbol{L}^T$, where $\boldsymbol{L}$ is the (lower) Cholesky triangle. The sparseness of $\boldsymbol{Q}$ is inherited into $\boldsymbol{L}$, thanks to the global Markov property: for $i < j$, such that i and j are separated by $F(i, j) = \{i+1, \ldots, j-1, j+1, \ldots, n\}$ in $\mathcal{G}$, $L_{ji} = 0$. Thus, only non-null terms in $\boldsymbol{L}$ are computed. In addition, nodes can be re-ordered to decrease the number of non-zero terms in $\boldsymbol{L}$. The typical cost of factorising $\boldsymbol{Q}$ into $\boldsymbol{L}\boldsymbol{L}^T$ depends on the dimension of the GMRF, e.g. $\mathcal{O}(n)$ for 1D (one dimension), $\mathcal{O}(n^{3/2})$ for 2D $\mathcal{O}(n^2)$ for 3D. Solving equations which involve $\boldsymbol{Q}$ also makes use of the Cholesky triangle. For example, $\boldsymbol{Q}\boldsymbol{x} = \boldsymbol{b}$ is solved in two steps. First solve $\boldsymbol{L}\boldsymbol{v} = \boldsymbol{b}$, then solve $\boldsymbol{L}^T\boldsymbol{x} = \boldsymbol{v}$. If $\boldsymbol{z} \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{I})$ then the solution of $\boldsymbol{L}^T\boldsymbol{x} = \boldsymbol{z}$ has precision matrix $\boldsymbol{Q}$. This is the general method for producing random samples from a GMRF. The log density at any $\boldsymbol{x}$, $\log \pi(\boldsymbol{x})$, can easily be computed using (6) since $\log|\boldsymbol{Q}| = 2\sum_i \log L_{ii}$.

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

$$\Sigma_{ij} = \delta_{ij}/L_{ii}^2 - \frac{1}{L_{ii}}\sum_{k=i+1}^{n} L_{ki}\Sigma_{kj}, \quad j \geq i,\ i = n, \ldots, 1, \tag{7}$$

where $\boldsymbol{\Sigma}$ $(= \boldsymbol{Q}^{-1})$ is the covariance matrix, and $\delta_{ij} = 1$ if $i = j$ and zero otherwise. Thus Σ_{ij} can be computed from (7), 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 reduces 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 $\boldsymbol{A}\boldsymbol{x} = \boldsymbol{e}$ for a $k \times n$ matrix $\boldsymbol{A}$ of rank k, the following strategy is used: if $\boldsymbol{x}$ is a sample from the unconstrained GMRF, then

$$\boldsymbol{x}^c = \boldsymbol{x} - \boldsymbol{Q}^{-1}\boldsymbol{A}^T(\boldsymbol{A}\boldsymbol{Q}^{-1}\boldsymbol{A}^T)^{-1}(\boldsymbol{A}\boldsymbol{x} - \boldsymbol{e}) \tag{8}$$

is a sample from the constrained GMRF. The expected value of $\boldsymbol{x}^c$ can also be computed using (8). This approach is commonly called 'conditioning by Kriging', see Cressie (1993) or Rue (2001). Note that $\boldsymbol{Q}^{-1}\boldsymbol{A}^T$ is computed by solving k linear systems, one for each column of $\boldsymbol{A}^T$. 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:

$$\pi(\boldsymbol{x}) \propto \exp\left\{-\frac{1}{2}\boldsymbol{x}^T\boldsymbol{Q}\boldsymbol{x} + \sum_{i\in\mathcal{I}} g_i(x_i)\right\}. \tag{9}$$

where $g_i(x_i)$ is $\log \pi(y_i|x_i, \boldsymbol{\theta})$ in our setting. The Gaussian approximation $\widetilde{\pi}_G(\boldsymbol{x})$ is obtained by matching the modal configuration and the curvature at the mode. The mode is computed iteratively using a Newton-Raphson method, also known as the scoring algorithm and its variant, the Fisher-scoring algorithm (Fahrmeir and Tutz, 2001). Let $\boldsymbol{\mu}^{(0)}$ be the initial guess, and expand $g_i(x_i)$ around $\mu_i^{(0)}$ to the second order,

$$g_i(x_i) \approx g_i(\mu_i^{(0)}) + b_i x_i - \frac{1}{2}c_i x_i^2 \tag{10}$$

where $\{b_i\}$ and $\{c_i\}$ depend on $\boldsymbol{\mu}^{(0)}$. A Gaussian approximation is obtained, with precision matrix $\boldsymbol{Q} + \mathrm{diag}(\boldsymbol{c})$ and mode given by the solution of $(\boldsymbol{Q} + \mathrm{diag}(\boldsymbol{c}))\boldsymbol{\mu}^{(1)} = \boldsymbol{b}$. This process is repeated until it converges to a Gaussian distribution with, say, mean $\boldsymbol{x}^\star$ and precision matrix $\boldsymbol{Q}^\star = \boldsymbol{Q} + \mathrm{diag}(\boldsymbol{c}^\star)$. If there are linear constraints, the mean is corrected at each iteration using the expected value of (8).

Since the non-quadratic term in (9) 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 $\boldsymbol{Q} + \mathrm{diag}(\boldsymbol{c})$. This is computationally convenient, as the Markov properties of the GMRF are preserved.

There are some suggestions in the literature how to construct an improved Gaussian approximation to (9) with respect to the one obtained matching the mode and the curvature at the mode; see Rue (2001, Sec. 5), Rue and Held (2005, Sec. 4.4.1) and Kuss and Rasmussen (2005). We have chosen not to pursue this issue here.

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, $\pi(x_i|\boldsymbol{y})$, $i = 1, \ldots, n$. The approximation is computed in three steps. The first step (Section 3.1) approximates the posterior marginal of $\boldsymbol{\theta}$ using the Laplace approximation (3). The second step (Section 3.2) computes the Laplace approximation, or the simplified Laplace approximation, of $\pi(x_i|\boldsymbol{y}, \boldsymbol{\theta})$, for selected values of $\boldsymbol{\theta}$, in order to improve on the Gaussian approximation (4). The third step combines the previous two using numerical integration (5).

3.1. Exploring $\widetilde{\pi}(\boldsymbol{\theta}|\boldsymbol{y})$

The first step of the INLA approach is to compute our approximation to the posterior marginal of $\boldsymbol{\theta}$, see (3). The denominator in (3) is the Gaussian approximation to the full conditional for $\boldsymbol{x}$, and is computed as described in Section 2.2. The main use of $\widetilde{\pi}(\boldsymbol{\theta}|\boldsymbol{y})$ is to integrate out the uncertainty with respect to $\boldsymbol{\theta}$ when approximating the posterior marginal of x_i, see (5). For this task, we do not need to represent $\widetilde{\pi}(\boldsymbol{\theta}|\boldsymbol{y})$ parametrically, but rather to explore it sufficiently well to be able to select good evaluation points for the numerical integration. At the end of this section, we discuss how the posterior marginals $\pi(\theta_j|\boldsymbol{y})$ can be approximated. Assume for simplicity that $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_m) \in \mathbb{R}^m$, which can always be obtained by reparametrisation.

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

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

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

If $\widetilde{\pi}(\boldsymbol{\theta}|\boldsymbol{y})$ is a Gaussian density, then $\boldsymbol{z}$ is $\mathcal{N}(\boldsymbol{0}, \boldsymbol{I})$. This reparametrisation corrects for scale and rotation, and simplifies numerical integration; see for example Smith et al. (1987).

Step 3 Explore $\log \widetilde{\pi}(\boldsymbol{\theta}|\boldsymbol{y})$ using the $\boldsymbol{z}$-parametrisation. Figure 1 illustrates the procedure when $\log \widetilde{\pi}(\boldsymbol{\theta}|\boldsymbol{y})$ is unimodal. Panel (a) shows a contour plot of $\log \widetilde{\pi}(\boldsymbol{\theta}|\boldsymbol{y})$ for $m = 2$, the location of the mode and the new coordinate axis for $\boldsymbol{z}$. We want to explore $\log \widetilde{\pi}(\boldsymbol{\theta}|\boldsymbol{y})$ 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 $\log \widetilde{\pi}(\boldsymbol{\theta}|\boldsymbol{y})$ is considered as significant, and which is used in the numerical integration (5). Details are as follows. We start from the mode ($\boldsymbol{z} = \boldsymbol{0}$), and go in the positive direction of z_1 with step-length δ_z say $\delta_z = 1$, as long as

$$\log \widetilde{\pi}(\boldsymbol{\theta}(\boldsymbol{0})|\boldsymbol{y}) - \log \widetilde{\pi}(\boldsymbol{\theta}(\boldsymbol{z})|\boldsymbol{y}) < \delta_\pi \tag{11}$$

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

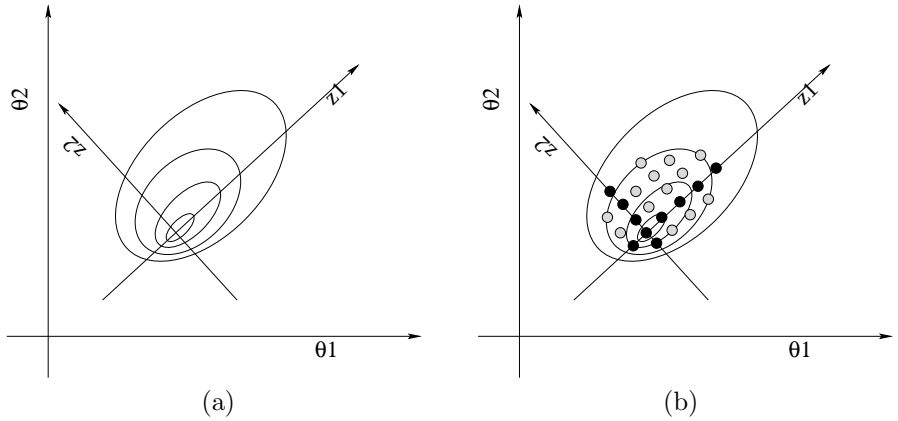


Fig. 1. Illustration of the exploration of the posterior marginal for $\boldsymbol{\theta}$. In (a) the mode is located, the Hessian and the coordinate system for $\boldsymbol{z}$ are computed. In (b) each coordinate direction is explored (black dots) until the log-density drops below a certain limit.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

in all the intermediate values by taking all different combinations of the black dots. These new points (shown as grey dots) are included if (11) holds. Since we layout the points $\boldsymbol{\theta}_k$ in a regular grid, we may take all the area-weights Δ_k in (5) to be equal.

Approximating $\pi(\theta_j|\boldsymbol{y})$. Posterior marginals for θ_j can be obtained directly from $\widetilde{\pi}(\boldsymbol{\theta}|\boldsymbol{y})$ using numerical integration. However, this is computationally demanding, as we need to evaluate $\widetilde{\pi}(\boldsymbol{\theta}|\boldsymbol{y})$ for a large number of configurations. A more feasible approach is to use the points already computed during step 1-3 to construct an interpolant to $\log \widetilde{\pi}(\boldsymbol{\theta}|\boldsymbol{y})$, and to compute marginals using numerical integration from this interpolant. If high accuracy is required, we need in practise a more dense configuration (for example $\delta_z = 1/2$ or $1/4$) than is required for the latent field $\boldsymbol{x}$; see Martino (2007) for numerical comparisons.

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

We have now a set of weighted points $\{\boldsymbol{\theta}_k\}$ to be used in the integration (5). The next step is to provide accurate approximations for the posterior marginal for the x_i's, conditioned on selected values of $\boldsymbol{\theta}$. We discuss three approximations $\widetilde{\pi}(x_i|\boldsymbol{y}, \boldsymbol{\theta}_k)$, 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 $\pi(x_i|\boldsymbol{\theta}, \boldsymbol{y})$ is the Gaussian approximation $\widetilde{\pi}_{\mathrm{G}}(x_i|\boldsymbol{\theta}, \boldsymbol{y})$, where the mean $\mu_i(\boldsymbol{\theta})$ and the marginal variance $\sigma_i^2(\boldsymbol{\theta})$ are derived using the recursions (7), and possibly correcting for linear constraints. During the exploration of $\widetilde{\pi}(\boldsymbol{\theta}|\boldsymbol{y})$, see Section 3.1, we already compute $\widetilde{\pi}_{\mathrm{G}}(\boldsymbol{x}|\boldsymbol{\theta}, \boldsymbol{y})$, 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

$$\widetilde{\pi}_{\mathrm{LA}}(x_i \mid \boldsymbol{\theta}, \boldsymbol{y}) \propto \left. \frac{\pi(\boldsymbol{x}, \boldsymbol{\theta}, \boldsymbol{y})}{\widetilde{\pi}_{\mathrm{GG}}(\boldsymbol{x}_{-i}|x_i, \boldsymbol{\theta}, \boldsymbol{y})} \right|_{\boldsymbol{x}_{-i}=\boldsymbol{x}^\star_{-i}(x_i,\boldsymbol{\theta})}. \tag{12}$$

Here, $\widetilde{\pi}_{\mathrm{GG}}$ is the Gaussian approximation to $\boldsymbol{x}_{-i}|x_i, \boldsymbol{\theta}, \boldsymbol{y}$, and $\boldsymbol{x}^\star_{-i}(x_i, \boldsymbol{\theta})$ is the modal configuration. Note that $\widetilde{\pi}_{\mathrm{GG}}$ is different from the conditional density corresponding to $\widetilde{\pi}_{\mathrm{G}}(\boldsymbol{x}|\boldsymbol{\theta}, \boldsymbol{y})$.

Unfortunately, (12) implies that $\widetilde{\pi}_{\mathrm{GG}}$ must be recomputed for each value of x_i and $\boldsymbol{\theta}$, since its precision matrix depends on x_i and $\boldsymbol{\theta}$. This is far too expensive, as it requires n factorisations of the full precision matrix. We propose two modifications to (12) which make it computationally feasible.

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

$$\boldsymbol{x}^\star_{-i}(x_i, \boldsymbol{\theta}) \approx \mathrm{E}_{\widetilde{\pi}_{\mathrm{G}}}(\boldsymbol{x}_{-i} \mid x_i). \tag{13}$$

The right-hand side is evaluated under the conditional density derived from the Gaussian approximation $\widetilde{\pi}_{\mathrm{G}}(\boldsymbol{x}|\boldsymbol{\theta}, \boldsymbol{y})$. The computational benefit is immediate. First, the conditional mean can be computed by a rank-one update from the unconditional mean, using (8). In the spatial case the cost is $\mathcal{O}(n \log n)$, for each i, which comes from solving $\boldsymbol{Q}^\star(\boldsymbol{\theta})\boldsymbol{v} = \mathbf{1}_i$, 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 (13) is that the conditional mean is continuous with respect to x_i, which is not the case when numerical optimisation is used to compute $\boldsymbol{x}^\star_{-i}(x_i, \boldsymbol{\theta})$.

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. 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 a 'region of interest' around i, $R_i(\boldsymbol{\theta})$, determine the marginal of x_i. The conditional expectation in (13) implies that

$$\frac{\mathrm{E}_{\widetilde{\pi}_{\mathrm{G}}}(x_j|x_i) - \mu_j(\boldsymbol{\theta})}{\sigma_j(\boldsymbol{\theta})} = a_{ij}(\boldsymbol{\theta}) \, \frac{x_i - \mu_i(\boldsymbol{\theta})}{\sigma_i(\boldsymbol{\theta})} \tag{14}$$

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

$$R_i(\boldsymbol{\theta}) = \{j \ : \ |a_{ij}(\boldsymbol{\theta})| > 0.001\}. \tag{15}$$

The most important computational saving using $R_i(\boldsymbol{\theta})$ comes from the calculation of the denominator of (12), where we now only need to factorise a $|R_i(\boldsymbol{\theta})| \times |R_i(\boldsymbol{\theta})|$ sparse matrix.

Expression (12), 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 (4), and choose, say, different values for the standardised variable

$$x_i^{(s)} = \frac{x_i - \mu_i(\boldsymbol{\theta})}{\sigma_i(\boldsymbol{\theta})} \tag{16}$$

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

$$\widetilde{\pi}_{\mathrm{LA}}\left(x_i \mid \boldsymbol{\theta},\boldsymbol{y}\right) \propto \mathcal{N}\left\{x_i;\ \mu_i(\boldsymbol{\theta}),\ \sigma_i^2(\boldsymbol{\theta})\right\} \times \exp\left\{\text{cubic spline}(x_i)\right\}. \tag{17}$$

The cubic spline is fitted to the difference of the log-density of $\widetilde{\pi}_{\mathrm{LA}}(x_i|\boldsymbol{\theta},\boldsymbol{y})$ and $\widetilde{\pi}_{\mathrm{G}}(x_i|\boldsymbol{\theta},\boldsymbol{y})$ 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 $\widetilde{\pi}_{\mathrm{SLA}}(x_i|\boldsymbol{\theta},\boldsymbol{y})$ by doing a series expansion of $\widetilde{\pi}_{\mathrm{LA}}(x_i|\boldsymbol{\theta},\boldsymbol{y})$ around $x_i = \mu_i(\boldsymbol{\theta})$. This allows us to correct the Gaussian approximation $\widetilde{\pi}_{\mathrm{G}}(x_i|\boldsymbol{\theta},\boldsymbol{y})$ 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 in the spatial case. Define

$$d_j^{(3)}(x_i,\boldsymbol{\theta}) = \frac{\partial^3}{\partial x_j^3} \log \pi(y_j \mid x_j,\boldsymbol{\theta})\bigg|_{x_j = \mathrm{E}_{\widetilde{\pi}_{\mathrm{G}}}(x_j|x_i)}$$

which we assume exists. The evaluation point is found from (14). The following trivial Lemma will be useful.

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

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

We expand the numerator and denominator of (12) around $x_i = \mu_i(\boldsymbol{\theta})$, using (13) and Lemma 1. Up to third order, we obtain

$$\begin{aligned} \log \pi(\boldsymbol{x},\boldsymbol{\theta},\boldsymbol{y})\bigg|_{\boldsymbol{x}_{-i} = \mathrm{E}_{\widetilde{\pi}_{\mathrm{G}}}(\boldsymbol{x}_{-i}|x_i)} &= -\frac{1}{2}(x_i^{(s)})^2 \\ &\quad + \frac{1}{6}(x_i^{(s)})^3 \sum_{j \in \mathcal{I}\setminus i} d_j^{(3)}\left(\mu_i(\boldsymbol{\theta}),\boldsymbol{\theta}\right)\left\{\sigma_j(\boldsymbol{\theta}) a_{ij}(\boldsymbol{\theta})\right\}^3 + \cdots \end{aligned} \tag{18}$$

The first and second order terms give the Gaussian approximation, whereas the third order term provides a correction for skewness. Further, the denominator of (12) reduces to

$$\log \widetilde{\pi}_{\mathrm{GG}}(\boldsymbol{x}_{-i} \mid x_i,\boldsymbol{\theta},\boldsymbol{y})\bigg|_{\boldsymbol{x}_{-i} = \mathrm{E}_{\widetilde{\pi}_{\mathrm{G}}}(\boldsymbol{x}_{-i}|x_i)} = \text{constant} + \frac{1}{2}\log\left|\boldsymbol{H} + \mathrm{diag}\left\{\boldsymbol{c}(x_i,\boldsymbol{\theta})\right\}\right| \tag{19}$$

where $\boldsymbol{H}$ is the prior precision matrix of the GMRF with ith column and row deleted, and $\boldsymbol{c}(x_i,\boldsymbol{\theta})$ is the vector of minus the second derivative of the log likelihood evaluated at $x_j = \mathrm{E}_{\widetilde{\pi}_{\mathrm{G}}}(x_j|x_i)$, see (14). Using that

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

we obtain

$$\begin{aligned}\log \widetilde{\pi}_{\mathrm{GG}}(\boldsymbol{x}_{-i} \mid x_i, \boldsymbol{\theta}, \boldsymbol{y})\Big|_{\boldsymbol{x}_{-i}=\mathrm{E}_{\widetilde{\pi}_{\mathrm{G}}}(\boldsymbol{x}_{-i}|x_i)} &= \text{constant} \\ &- \frac{1}{2} x_i^{(s)} \sum_{j \in \mathcal{I} \setminus i} \mathrm{Var}_{\widetilde{\pi}_{\mathrm{G}}}(x_j|x_i)\ d_j^{(3)}(\mu_i(\boldsymbol{\theta}), \boldsymbol{\theta})\ \sigma_j(\boldsymbol{\theta})\ a_{ij}(\boldsymbol{\theta}) + \cdots .\end{aligned} \tag{20}$$

For Gaussian data (19) is just a constant, so the first order term in (20) is the first correction for non-Gaussian observations. Note that

$$\mathrm{Var}_{\widetilde{\pi}_{\mathrm{G}}}(x_j|x_i) = \sigma_j^2(\boldsymbol{\theta}) \left\{1 - \mathrm{Corr}_{\widetilde{\pi}_{\mathrm{G}}}(x_i, x_j)^2\right\}$$

and that the covariance between x_i and x_j (under $\widetilde{\pi}_{\mathrm{G}}$) is computed while doing the rank-one update in (13), as the jth element of the solution of $\boldsymbol{Q}^\star(\boldsymbol{\theta})\boldsymbol{v} = \mathbf{1}_i$.

We now collect the expansions (18) and (20). Define

$$\begin{aligned}\gamma_i^{(1)}(\boldsymbol{\theta}) &= \frac{1}{2} \sum_{j \in \mathcal{I} \setminus i} \sigma_j^2(\boldsymbol{\theta}) \left\{1 - \mathrm{Corr}_{\widetilde{\pi}_{\mathrm{G}}}(x_i, x_j)^2\right\}\ d_j^{(3)}(\mu_i(\boldsymbol{\theta}), \boldsymbol{\theta})\ \sigma_j(\boldsymbol{\theta})\ a_{ij}(\boldsymbol{\theta}) \\ \gamma_i^{(3)}(\boldsymbol{\theta}) &= \sum_{j \in \mathcal{I} \setminus i} d_j^{(3)}(\mu_i(\boldsymbol{\theta}), \boldsymbol{\theta})\ \left\{\sigma_j(\boldsymbol{\theta}) a_{ij}(\boldsymbol{\theta})\right\}^3\end{aligned} \tag{21}$$

then

$$\log \widetilde{\pi}_{\mathrm{SLA}}(x_i^s|\boldsymbol{\theta}, \boldsymbol{y}) = \text{constant} - \frac{1}{2}(x_i^{(s)})^2 + \gamma_i^{(1)}(\boldsymbol{\theta}) x_i^{(s)} + \frac{1}{6}(x_i^{(s)})^3 \gamma_i^{(3)}(\boldsymbol{\theta}) + \cdots . \tag{22}$$

Eq. (22) 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)

$$\pi_{\mathrm{SN}}(z) = \frac{2}{\omega} \phi\left(\frac{z - \xi}{\omega}\right)\ \Phi\left(a \frac{z - \xi}{\omega}\right) \tag{23}$$

where $\phi(\cdot)$ and $\Phi(\cdot)$ are the density and distribution function of the standard normal distribution, and ξ, $\omega > 0$, and a are respectively the location, scale, and skewness parameters. We fit a Skew-Normal density to (22) 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 B for details.

We have implicitly assumed that the expansion (18) 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 (20) and fit the spline-corrected Gaussian (17) 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 $a_{i\cdot}(\boldsymbol{\theta})$, for each i; thus the 'region of interest' strategy (15) is unhelpful here. Most of the other terms in (21) do not depend on i, and thus are computed only once. The cost for computing (22), 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 $\boldsymbol{x}$ is general, we need to visit all other sites, for each i, for a potential contribution. This operation alone costs $\mathcal{O}(n^2)$. In summary, the total cost for computing all n marginals $\widetilde{\pi}(x_i|\boldsymbol{y})$, $i = 1, \ldots, n$, using (5) and the simplified Laplace approximation, is exponential in the dimension of $\boldsymbol{\theta}$ times $\mathcal{O}(n^2 \log n)$ (in the spatial case).

4. Approximation error: Asymptotics and practical issues

4.1. Approximation error of $\widetilde{\pi}(\boldsymbol{\theta}|\boldsymbol{y})$

For the sake of discussion, denote p the dimension of vector $(\boldsymbol{x}, \boldsymbol{\theta})$, i.e. $p = n + m$, and recall that n_d denotes the number of observations. Up to normalisation, $\widetilde{\pi}(\boldsymbol{\theta}|\boldsymbol{y})$ is formally equivalent to the Laplace approximation of a marginal posterior density proposed by Tierney and Kadane (1986), which, under 'standard' conditions, has error rate $\mathcal{O}(n_d^{-1})$. We want to make it clear however that these standard conditions are not relevant in many applications of latent Gaussian models. We will now discuss several asymptotic schemes and their impact on the actual error rate of $\widetilde{\pi}(\boldsymbol{\theta}|\boldsymbol{y})$.

First, assume that p is fixed while n_d goes to infinity; for instance, a GMRF model with a fixed number of nodes but a growing number of observations accumulating at each node. In this case, the usual assumptions for the asymptotic validity of a Laplace approximation, see Kass et al. (1999) or Schervish (1995, p. 453), are typically satisfied. This asymptotic scheme is obviously quite specific, but it explains the good properties of INLA in a few applications, such as for instance a GMRF model with binomial observations, $y_i|x_i \sim \text{Bin}(n_i, \text{logit}^{-1}(x_i))$, provided all the n_i take large values.

Second, if n (and therefore p) grows with n_d, then, according to Shun and McCullagh (1995), the error rate is $\mathcal{O}(n/n_d)$ as n is the dimension of the integral defining the unnormalised version of $\widetilde{\pi}(\boldsymbol{\theta}|\boldsymbol{y})$. Note that this rate is not established rigorously. This asymptotic scheme is relevant to regression models involving individual effects, in which case $n/n_d \to 0$ is not a taxing assumption. On the other hand, many GMRF models are such that n/n_d is a constant (typically 1). For such models, we have the following result. If, as $n_d \to \infty$, the true latent field $\boldsymbol{x}$ converges to a degenerate Gaussian random distribution of rank q, then the asymptotic error rate is $\mathcal{O}(q/n_d)$. Conversely, if the considered model is such that the components of $\boldsymbol{x}$ are independent, one can show that the approximation error is $\mathcal{O}(1)$ but almost never $o(1)$.

In conclusion, the accuracy of $\widetilde{\pi}(\boldsymbol{\theta}|\boldsymbol{y})$ seems to be directly related to the 'actual' dimension of $\boldsymbol{x}$. Thus, we recommend to evaluate, conditional on $\boldsymbol{\theta}$, the *effective number of parameters*, $p_{\text{D}}(\boldsymbol{\theta})$, as defined by Spiegelhalter et al. (2002). Since $\boldsymbol{x}$ given $\boldsymbol{y}$ and $\boldsymbol{\theta}$ is roughly Gaussian, $p_{\text{D}}(\boldsymbol{\theta})$ is conveniently approximated by

$$p_{\text{D}}(\boldsymbol{\theta}) \approx n - \text{Trace}\left\{\boldsymbol{Q}(\boldsymbol{\theta})\, \boldsymbol{Q}^{\star}(\boldsymbol{\theta})^{-1}\right\}, \tag{24}$$

the trace of the prior precision matrix times by the posterior covariance matrix of the Gaussian approximation (Spiegelhalter et al., 2002, Eq. (16)). (The computation of $p_{\text{D}}(\boldsymbol{\theta})$ is computationally 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 (7).) This quantity also measures to which extent the Gaussianity and the dependence structure of the prior are preserved in the posterior of $\boldsymbol{x}$, given $\boldsymbol{\theta}$. For instance, for non-informative

data, $p_{\mathrm{D}}(\boldsymbol{\theta}) = 0$, and the approximation error is zero, since the posterior equals the Gaussian prior. In all our applications, we observed that $p_{\mathrm{D}}(\boldsymbol{\theta})$ is typically small relative to n_d for values of $\boldsymbol{\theta}$ in the vicinity of the posterior mode.

Note finally that in most cases normalising the approximated density reduces further the asymptotic rate, as the dominating terms of the numerator and the denominator cancel out (Tierney and Kadane, 1986); in the standard case, normalising reduces the error rate from $\mathcal{O}(n_d^{-1})$ to $\mathcal{O}(n_d^{-3/2})$.

The discussion above of the asymptotic properties of $\widetilde{\pi}(\boldsymbol{\theta}|\boldsymbol{y})$ applies almost directly to $\widetilde{\pi}_{\mathrm{LA}}(x_i|\boldsymbol{\theta}, \boldsymbol{y})$: conditional on $\boldsymbol{\theta}$, $\widetilde{\pi}_{\mathrm{LA}}(x_i|\boldsymbol{\theta}, \boldsymbol{y})$ is a Laplace approximation of the posterior marginal density of x_i, and the dimension of the corresponding integral is the dimension of $\boldsymbol{x}_{-i}$, i.e. $n - 1$.

4.2. Assessing the approximation error

Obviously, there is only one way to assess with certainty the approximation error of our approach, which is to run a 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 $\widetilde{\pi}_{\mathrm{G}}(\boldsymbol{x}|\boldsymbol{\theta}, \boldsymbol{y})$, for each $\boldsymbol{\theta}_k$ used in the integration. We do this by computing $p_{\mathrm{D}}(\boldsymbol{\theta})$ (24) as discussed in Section 4.1, but we can also use that (3) can be rewritten as

$$\begin{aligned}\frac{\pi(\boldsymbol{\theta}|\boldsymbol{y})}{\widetilde{\pi}(\boldsymbol{\theta}|\boldsymbol{y})} &\propto |\boldsymbol{Q}^\star(\boldsymbol{\theta})|^{1/2} \int \exp\left[-\frac{1}{2}\left\{\boldsymbol{x} - \boldsymbol{x}^\star(\boldsymbol{\theta})\right\}^T \boldsymbol{Q}^\star(\boldsymbol{\theta})\left\{\boldsymbol{x} - \boldsymbol{x}^\star(\boldsymbol{\theta})\right\} + r(\boldsymbol{x}; \boldsymbol{\theta}, \boldsymbol{y})\right] d\boldsymbol{x} \\ &= \mathrm{E}_{\widetilde{\pi}_{\mathrm{G}}}\left[\exp\left\{r(\boldsymbol{x}; \boldsymbol{\theta}, \boldsymbol{y})\right\}\right],\end{aligned}$$

where the constant of proportionality is quite involved and not needed in the following discussion. Further, $\boldsymbol{x}^\star(\boldsymbol{\theta})$ and $\boldsymbol{Q}^\star(\boldsymbol{\theta})$ are the mean and precision of Gaussian distribution $\widetilde{\pi}_{\mathrm{G}}$, $r(\boldsymbol{x}; \boldsymbol{\theta}, \boldsymbol{y}) = \sum_i h_i(x_i)$, and $h_i(x_i)$ is $g_i(x_i)$ minus its Taylor expansion up to order two around $x_i^\star(\boldsymbol{\theta})$; see (9) and (10). If, for each $\boldsymbol{\theta}_k$, $p_{\mathrm{D}}(\boldsymbol{\theta})$ is small compared to n_d, and the empirical quantiles of the random variable $r(\boldsymbol{x}; \boldsymbol{\theta}, \boldsymbol{y})$ are in absolute value significantly smaller than n_d, then one has strong confidence that the Gaussian approximation is adequate. The empirical quantiles of $r(\boldsymbol{x}; \boldsymbol{\theta}, \boldsymbol{y})$ are found by sampling (for example, 1 000) independent realisations from $\widetilde{\pi}_G$.

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 (4), followed by the simplified Laplace approximation (22), then by the Laplace approximation (12). Specifically we compute the integrated marginal (5) based on both the Gaussian approximation and the simplified Laplace approximation, and compute their symmetric Kullback-Leibler divergence (SKLD). If the divergence is small then both approximations are considered as acceptable. Otherwise, compute (5) using the Laplace approximation (12) 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 (5), we can compare the SKLD between the posterior marginals obtained with a standard and those obtained with a higher

resolution. Such an 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. The computations were performed on a single-processor 2.1GHz laptop using the `inla` program (Martino and Rue, 2008) which is a user-friendly interface towards our GMRFLib-library written in C (Rue and Held, 2005, Appendix). (We will comment upon speedup strategies and parallel implementation in Section 6.5 and Section 7.) We start with some simulated examples with fixed $\boldsymbol{\theta}$ in Section 5.1, to verify the (simplified) Laplace approximation for $x_i|\boldsymbol{\theta}, \boldsymbol{y}$. We continue with a generalised linear mixed model for longitudinal data in Section 5.2, a stochastic volatility model applied to exchange rate data in Section 5.3, a spatial semi-parametric regression model for disease-mapping in Section 5.4. The dimensions get really large in Section 5.5, in which we analyse some data using a spatial log-Gaussian Cox process.

5.1. Simulated examples

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

$$f_t - \mu \mid \mu, f_1, \ldots, f_{t-1} \;\sim\; \mathcal{N}\left\{\phi(f_{t-1} - \mu),\; \sigma^2\right\}, \qquad t = 2, \ldots, 50 \tag{25}$$

and $\mu \sim \mathcal{N}(0, 1)$, $\phi = 0.85$, $\mathrm{Var}(f_t) = 1$ and $f_1 - \mu \sim \mathcal{N}(0, 1)$. In this example $\eta_t = f_t$; see (1). As our observations we take

$$\text{E1: } y_t - \eta_t \mid (\boldsymbol{\eta}, \mu) \;\sim\; \text{Student-}t_3, \quad \text{E2: } y_t \mid (\boldsymbol{\eta}, \mu) \;\sim\; \text{Bernoulli}\left\{\text{logit}^{-1}(\eta_t)\right\}$$

for $t = 1, \ldots, 50$, in experiment E1 and E2, respectively. Note that the Student-t_3 is symmetric so we need to use the full numerator in the simplified Laplace approximations as described in Section 3.2.3.

To create the observations in each experiment, we sampled first $(\boldsymbol{f}^T, \mu)^T$ from the prior, then simulated the observations. We computed $\widetilde{\pi}(f_t|\boldsymbol{\theta}, \boldsymbol{y})$ for $t = 1, \ldots, 50$ and $\widetilde{\pi}(\mu|\boldsymbol{\theta}, \boldsymbol{y})$ using the simplified Laplace approximation. We located the 'worst node', that is the node with maximum SKLD between the Gaussian and the simplified Laplace approximations. This process was repeated 100 times. Figure 2 provides the results for the 'worst of the worst nodes', that is the node that maximises our SKLD criterion among all the nodes of the 100 generated sample. The first (resp. second) column displays the results for E1 with Student-t_3 data (resp. E2 with Bernoulli data). Panel (a) and (b) display $\boldsymbol{f}$ (solid line) and the observed data (circles). In (a) the selected node is marked with a vertical line and solid dot. In (b) the node with maximum SKLD is μ, and hence is not shown. Panel (c) and (d) display the approximated marginals for the node with maximum SKLD in the standardised scale (16). 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 SKLD between the Gaussian approximation and the simplified Laplace one is 0.20 (c) and 0.05 (d). The SKLD between the simplified Laplace approximation and the Laplace one is 0.001 (c) and 0.0004 (d). Panel (e) and (f) show the simplified Laplace approximation

with a histogram based on $10,000$ (near) independent samples from $\pi(\boldsymbol{f}, \mu|\boldsymbol{\theta}, \boldsymbol{y})$. 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 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.

5.2. A generalised linear mixed model for longitudinal data

Generalised linear (mixed) models form a large class of latent Gaussian models. We consider the Epil example of the OpenBUGS (Thomas et al., 2006) manual Vol I, which is based on model III of Breslow and Clayton (1993, Sec. 6.2) and data from Thall and Vail (1990).

The data come from a clinical trial of 59 epileptics patients. Each patient i is randomised to a new drug ($\mathrm{Trt}_i = 1$) or a placebo ($\mathrm{Trt}_i = 0$), in addition to the standard chemotherapy. The observations for each patient $y_{i1}, \ldots, y_{i4}$, are the number of seizures during the two weeks before each of the four clinic visits. The covariates are age (Age), the baseline seizure counts (Base) and an indicator variable for the 4th clinic visit (V4). The linear predictor is

$$\begin{aligned}\eta_{ij} &= \beta_0 + \beta_{\mathrm{Base}} \log(\mathrm{Baseline}_j/4) + \beta_{\mathrm{Trt}}\mathrm{Trt}_j + \beta_{\mathrm{Trt}\times\mathrm{Base}}\mathrm{Trt}_j \times \log(\mathrm{Baseline}_j/4) \\ &\quad + \beta_{\mathrm{Age}}\mathrm{Age}_j + \beta_{\mathrm{V4}}\mathrm{V4}_j + \epsilon_i + \nu_{ij}, \qquad i = 1, \ldots, 59, \quad j = 1, \ldots, 4,\end{aligned}$$

using centred covariates. The observations are conditionally independent Poisson variables with mean $\exp(\eta_{ij})$. Overdispersion in the Poisson distribution is modelled using individual random effects ϵ_i and subject by visit random effects ν_{ij}. We use the same priors as in the OpenBUGS manual: $\epsilon_i \overset{\mathrm{iid}}{\sim} \mathcal{N}(0, 1/\tau_\epsilon)$, $\nu_{ij} \overset{\mathrm{iid}}{\sim} \mathcal{N}(0, 1/\tau_\nu)$, $\tau_\epsilon, \tau_\nu \sim \Gamma(0.001, 0.001)$, and all the $\beta_\cdot$'s are assigned $\mathcal{N}(0, 100^2)$ priors. In this example our latent field $\boldsymbol{x}$ is of dimension $n = 301$ and consists of $\{\eta_{ij}\}$, $\{\epsilon_i\}$, β_0, β_{Base}, β_{Trt}, $\beta_{\mathrm{Trt}\times\mathrm{Base}}$, β_{Age} and β_{V4}. The hyperparameters are $\boldsymbol{\theta} = (\tau_\epsilon, \tau_\nu)^T$.

We computed the approximate posterior marginals for the latent field using both Gaussians and simplified Laplace approximations. The node where SKLD between these two marginals is maximum, is β_0. The SKLD is 0.23. The two approximated marginals for β_0 are displayed in Figure 3(a). The simplified Laplace (solid line) approximation does correct the Gaussian approximation (dashed line) in the mean, while the correction for skewness is minor. The simplified Laplace approximation gives accurate results, as shown in Figure 3(a) where a histogram from a long MCMC using OpenBUGS is overlaid. Figure 3(b) displays the posterior marginal for τ_ϵ found by integrating out τ_ν from $\widetilde{\pi}(\boldsymbol{\theta}|\boldsymbol{y})$; again, we find no errors.

We validated the approximations at the modal value $\boldsymbol{\theta}^\star$. The effective number of parameters (24) was 121.1, which corresponds to about 2 samples for each parameter. A 95% interval for the remainder $r(\boldsymbol{x}; \boldsymbol{\theta}^\star, \boldsymbol{y})/n_d$ is $[-0.01, 0.024]$ using $1,000$ independent samples. Computing the (true) Laplace approximation for the posterior marginal of β_0 gives a negligible SKLD to the simplified Laplace one; thus indicating that the simplified Laplace approximation is adequate. The computational cost for obtaining all the latent posterior

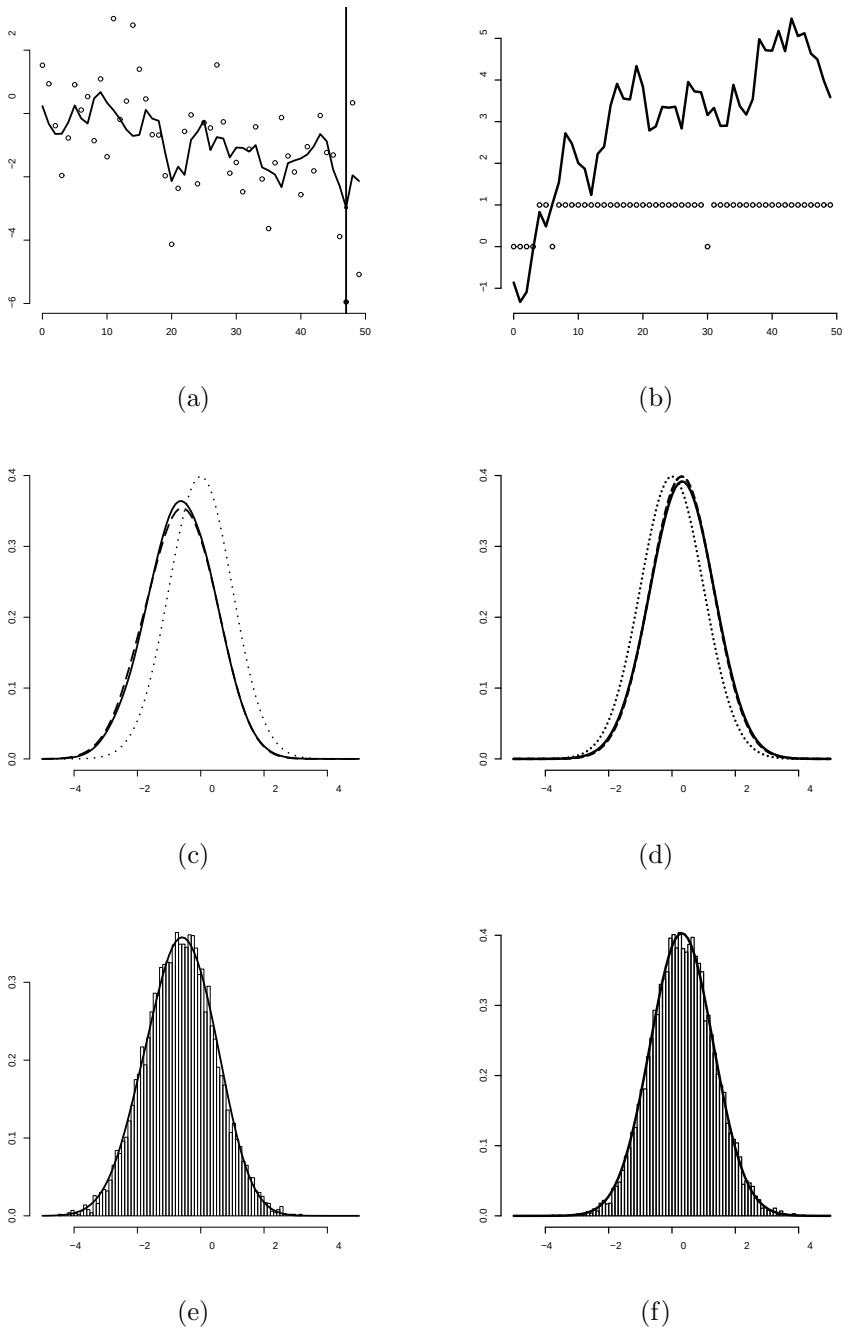


Fig. 2. First row shows the true latent Gaussian field (solid line), the observed Student-t_3 data and Bernoulli data (dots). Second row shows the approximate marginal for a selected node using various approximations; Gaussian (dotted), simplified Laplace (dashed) and Laplace (solid). Third row compares samples from a long MCMC chain with the marginal computed with the simplified Laplace approximation.

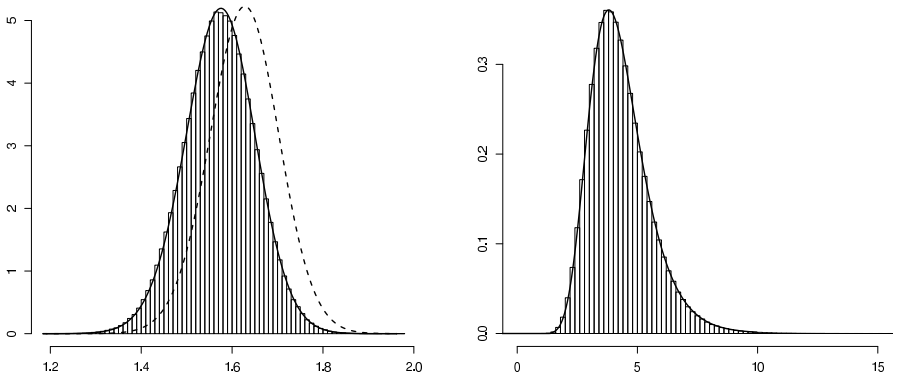


Fig. 3. The posterior marginal for β_0 (a) and τ_ϵ (b) for the example in Section 5.2. The solid line in (a) shows the marginal using simplified Laplace approximation and the dashed line the Gaussian approximation. The solid line in (b) shows the marginal for τ_ϵ after integrating out τ_ν. The histograms result from a long MCMC run using OpenBUGS.

marginals was about 1.5 seconds in total. Although OpenBUGS can provide approximate answers in minutes, we had to run it for hours to provide accurate posterior marginals.

5.3. Stochastic volatility models

Stochastic volatility models are frequently used to analyse financial time series. Figure 4(a) displays the log of the $n_d = 945$ daily difference of the pound-dollar exchange rate from October 1st, 1981, to June 28th, 1985. This data set 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).

The observations are taken to be

$$y_t \mid \eta_t \sim \mathcal{N}\{0,\ \exp(\eta_t)\}, \qquad t = 1, \ldots, n_d. \tag{26}$$

The linear predictor consists of two terms, $\eta_t = \mu + f_t$, where f_t is a first order auto-regressive Gaussian process

$$f_t \mid f_1, \ldots, f_{t-1}, \tau, \phi \sim \mathcal{N}(\phi f_{t-1}, 1/\tau), \quad |\phi| < 1,$$

and μ is a Gaussian mean value. In this example, $\boldsymbol{x} = (\mu, \eta_1, \ldots, \eta_T)^T$ and $\boldsymbol{\theta} = (\phi, \tau)^T$. The log-likelihood (with respect to η_t) 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 use the following priors: $\tau \sim \Gamma(1, 0.1)$, $\phi' \sim \mathcal{N}(3, 1)$ where $\phi = 2\exp(\phi')/(1 + \exp(\phi')) - 1$, and $\mu \sim \mathcal{N}(0, 1)$. We display the results for the Laplace approximation of the posterior marginals of the two hyperparameters and μ, but only based on only the first 50 observations in Figure 4(b)-(d), as using the full data set make the approximation problem easier. The solid line in Figure 4(d) is the marginal found using simplified Laplace approximations and the dashed line uses Gaussian approximations, but in this case there

are little difference (the SKLD is 0.05). The histograms are constructed from the output of a long MCMC run using OpenBUGS. The approximations computed are very precise and no deviance (in any node) can be detected. The results obtained using the full data set are similar but the marginals are narrower (not shown).

Following the discussion in Section 1.6, we also used this set of $n = 50$ observations to compare INLA with Zoeter et al. (2005)'s EP algorithm (with a slightly different parametrisation of the model and other priors due to constraints in their code). The latter was considerably less accurate (e.g. the posterior mean of ϕ is shifted one standard deviation to the right) and more expensive; running time was about 40 minutes for Zoeter et al. (2005)'s Matlab (`http://www.mathworks.com/`) code, to compare with 0.3 seconds for our approach.

We now extend the model to allow for Student-t_ν distributed observations, where we scale the Student-t_ν distribution to have unit variance for all $\nu > 2$. We assign a $\mathcal{N}(2.08, 1)$ prior to ν' where $\nu = 2+\exp(\nu')$. The number of hyperparameters is now three. Figure 4(e) displays the approximate posterior marginal for the degrees-of-freedom and panel (f) displays the 0.025, 0.5 and 0.975 quantiles of η_t. Also in this case, we do not find any error in the approximations which was verified on using a subset of the full data (not shown). The marginal for the degrees-of-freedom suggests that the extension to Student-t_ν is not needed in this case, but see Section 6.2 for a more formal comparison of these two models. For the latent auto-regressive process, there is little difference between the Gaussian approximation and the simplified Laplace one, for both models. The average SKLD was about 0.007 in both cases.

We validated the approximations using all the $n = 945$ observations at the modal value $\boldsymbol{\theta}^\star$. The effective number of parameters (24) was about 63, which is small compared to n_d. A 95% interval for the remainder $r(\boldsymbol{x}; \boldsymbol{\theta}^\star, \boldsymbol{y})/n_d$ is $[-0.002, 0.004]$ using $1,000$ independent samples. The computational cost for obtaining all the posterior marginals (using (26)) for the latent field, was about 11 seconds.

5.4. Disease mapping of cancer incidence data

In this example we consider spatial model for mapping cancer incidence where the stage of the disease at time of diagnosis is known. The class of "disease mapping" models are often latent Gaussians, see for example Besag et al. (1991); Wakefield et al. (2000) and Held et al. (2005) for an introduction.

The data are binary incidence cases of cervical cancer from the former East German Republic from 1979 (Knorr-Held et al., 2002). The data are stratified by district and age group. Each of the $n_d = 6990$ cases are classified into premalignant $y_i = 1$ or malignant $y_i = 0$. Denote by d_i and a_i the district and age-group for case $i = 1, \ldots, 6990$. There are 216 districts and 15 age groups ($15-19$, $20-24$, ..., > 84). We follow Rue and Held (2005, Sec. 4.3.5) and use a logistic binary regression model:

$$\text{logit}(p_i) = \eta_i = \mu + f^{(\text{a})}_{a_i} + f^{(\text{s})}_{d_i} + f^{(\text{u})}_{d_i},$$

where $\boldsymbol{f}^{(\text{a})}$ is a smooth effect of the age-group, $\boldsymbol{f}^{(\text{s})}$ is a smooth spatial field and $\boldsymbol{f}^{(\text{u})}$ are district random effects. More specifically, $\boldsymbol{f}^{(\text{a})}$ follows an intrinsic second order random

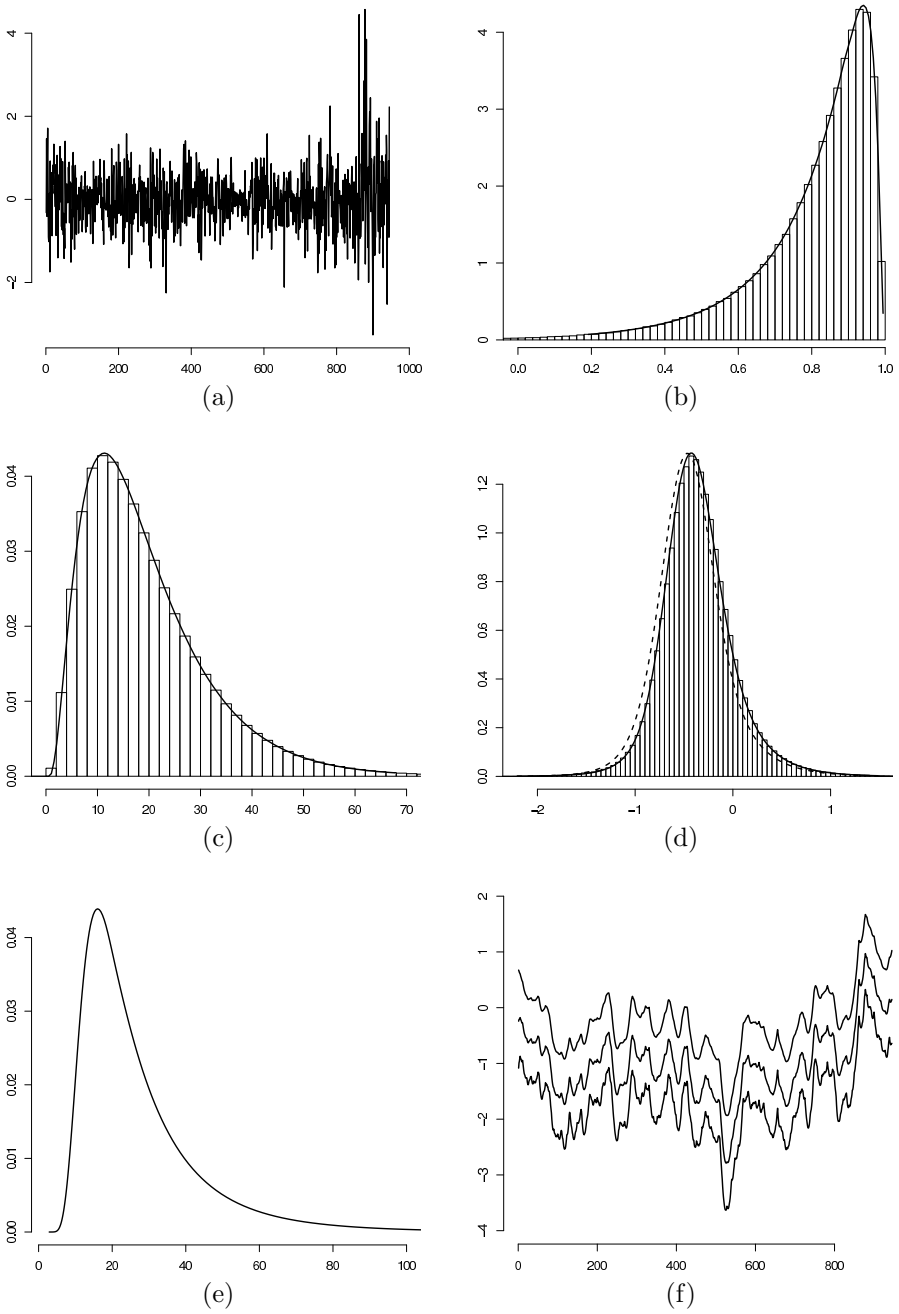


Fig. 4. Panel (a) displays the log of the daily difference of the pound-dollar exchange rate from October 1st, 1981, to June 28th, 1985. Panels (b) and (c) display the approximated posterior marginals for ϕ and τ using only the first $n = 50$ observations in (a). Overlaid are the histograms obtained from a long MCMC run using OpenBUGS. Panel (d) displays the approximated posterior marginal using simplified Laplace approximations (solid line) and Gaussian approximations (dashed line) for μ, which is the node in latent field with maximum SKLD. Panel (e) displays the posterior marginal for the degrees-of-freedom assuming Student-t_ν distributed observations, and panel (f) the 0.025, 0.5 and 0.975 posterior quantiles for η_t.

walk model (Rue and Held, 2005, Ch. 3) with precision $\kappa^{(\mathrm{a})}$,

$$\pi(\boldsymbol{f}^{(\mathrm{a})}|\kappa^{(\mathrm{a})}) \propto (\kappa^{(\mathrm{a})})^{(15-2)/2} \exp\left(-\frac{\kappa^{(\mathrm{a})}}{2}\sum_{j=3}^{15}(f_j^{(\mathrm{a})} - 2f_{j-1}^{(\mathrm{a})} + f_{j-2}^{(\mathrm{a})})^2\right). \tag{27}$$

The model for the spatial term $\boldsymbol{f}^{(\mathrm{s})}$ are defined conditionally, as

$$f_i^{(\mathrm{s})} \mid \boldsymbol{f}_{-i}^{(\mathrm{s})}, \kappa^{(\mathrm{s})} \sim \mathcal{N}\left(\frac{1}{n_i}\sum_{j\in\partial_i} f_j^{(\mathrm{s})}, \frac{1}{n_i\kappa^{(\mathrm{s})}}\right)$$

where ∂_i is the set of neighbour-districts to district i, namely those n_i districts who share a common border with district i; see Rue and Held (2005, Sec. 3.3.2) for further detail on this model. The district random effects are independent zero-mean Gaussians with precision $\kappa^{(\mathrm{u})}$. We put a zero-mean constraint on both the age and spatial effects and assign independent $\Gamma(1, 0.01)$ priors to the three hyperparameters $(\kappa^{(\mathrm{a})}, \kappa^{(\mathrm{s})}, \kappa^{(\mathrm{u})})^T$, and a $\mathcal{N}(0, 0.1)$ prior to μ. The dimension of the latent field $\boldsymbol{x}$ is $216 \times 15 + 1 = 3241$.

The results are displayed in Figure 5. Panel (a) displays the posterior marginal for the node with the largest SKLD between the approximations using simplified Laplace (solid line) and Gaussians (dashed line). The SKLD is 0.058. Overlaid is the histogram found from a long MCMC run using the block-MCMC algorithm with auxiliary variables described in Rue and Held (2005, Sec. 4.3.5); the fit is perfect. Panel (b) displays the effect of the age groups, where the solid line interpolates the posterior median and the dashed lines displays the 0.025 and 0.975 quantiles. The quantiles obtained from a long MCMC run are shown by dots; again the fit is very good. Panel (c) displays the median of the smooth spatial component, where the grey-scale goes from 0.2 (white) to 5 (black). (The shaded region is Berlin.)

We validated the approximations at the modal value $\boldsymbol{\theta}^\star$. The effective number of parameters (24) was about 101, which is small compared to n_d. A 95% interval for the remainder $r(\boldsymbol{x}; \boldsymbol{\theta}^\star, \boldsymbol{y})/n_d$ is $[-0.001, 0.001]$ using $1,000$ independent samples. The computational cost for obtaining all the posterior marginals for the latent field was about 34 seconds.

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). We illustrate in this section 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 $\boldsymbol{y} = (y_1, \ldots, y_N)^T$. 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 $|w_i|\exp(\eta_i)$.

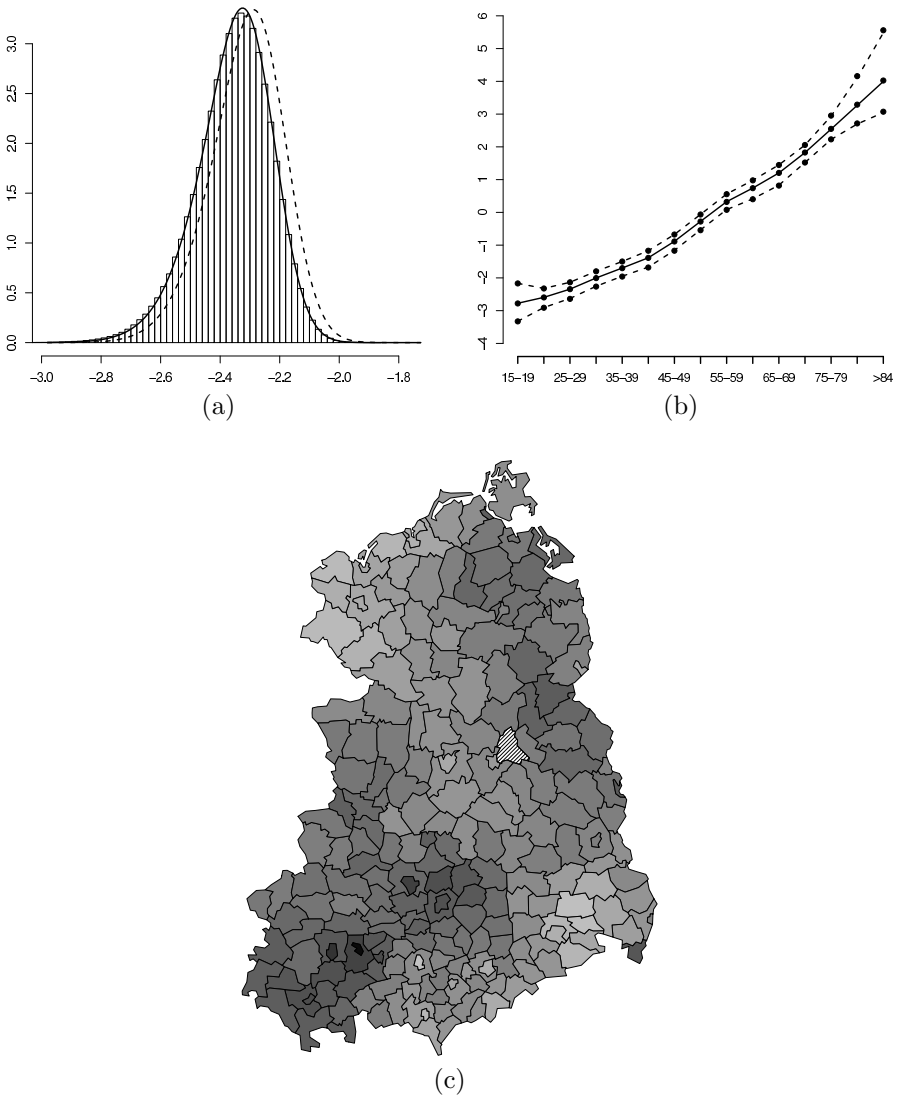


Fig. 5. The results for the cancer incidence example: (a) the posterior marginal for $f_3^{(a)}$ using simplified Laplace approximations (solid line), Gaussians approximations (dashed line) and samples from a long MCMC run (histogram). Panel (b) displays the posterior median, 0.025 and 0.975 quantiles of the age-class effect (interpolated), whereas the dots are those obtained from a long MCMC run. Panel (c) displays the posterior median of the (smooth) spatial effect.

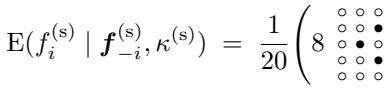
We apply model (28) to the tropical rain forest data studied by Waagepetersen (2007). These data come from a 50-hectare permanent tree plot which was established in 1980 in the tropical moist forest of Barro Colorado Island in central Panama. Censuses have been carried out every 5th year from 1980 to 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 censured 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 6(a). Sources of variation explaining the locations 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. This suggests the model

$$\eta_i = \beta_0 + \beta_{\text{Alt}}\text{Altitude}_i + \beta_{\text{Grad}}\text{Gradient}_i + f_i^{(\text{s})} + f_i^{(\text{u})}, \tag{28}$$

where $\boldsymbol{f}^{(\text{s})}$ represent the spatial component, and $\boldsymbol{f}^{(\text{u})}$ is an unstructured term. An alternative would be to use a semi-parametric model for the effect of the covariates similar to (27).

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. This makes $n_d = 20\,000$. The scaled and centred versions of the altitude and norm of the gradient, are shown in panel (b) and (c), 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)

$$\text{E}(f_i^{(\text{s})} \mid \boldsymbol{f}_{-i}^{(\text{s})}, \kappa^{(\text{s})}) = \frac{1}{20}\left(8 \;\; -2 \;\; -1 \right), \; \text{Prec}(f_i^{(\text{s})} \mid \boldsymbol{f}_{-i}^{(\text{s})}, \kappa^{(\text{s})}) = 20\kappa^{(\text{s})}. \tag{29}$$



The precision $\kappa^{(\text{s})}$ is unknown. The full conditionals are constructed to mimic the thin-plate spline. There are some corrections to (29) 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 $\boldsymbol{f}^{(\text{u})}$ are independent $\mathcal{N}(0, \kappa^{(\text{u})})$, vague $\Gamma(1.0, 0.001)$ priors are assigned to $\kappa^{(\text{s})}$ and $\kappa^{(\text{u})}$, and independent $\mathcal{N}(0, 10^3)$ priors to β_0, β_{Alt} and β_{Grad}. The latent field is $\boldsymbol{x} = (\boldsymbol{\eta}^T, (\boldsymbol{f}^{(\text{s})})^T, \beta_0, \beta_{\text{Alt}}, \beta_{\text{Grad}})^T$ with dimension $40,003$, and $\boldsymbol{\theta} = (\log \kappa^{(\text{s})}, \log \kappa^{(\text{u})})$ with dimension 2.

We computed the approximation for $20,003$ posterior marginals $\boldsymbol{f}^{(\text{s})}$, β_0, β_{Alt} and β_{Grad}, using the simplified Laplace approximation. The results are displayed in Figure 7. Panel (a) displays the estimated posterior mean of the spatial component, where we have indicated using contours, those nodes where the SKLD between the marginal computed with the Gaussian approximation and the one computed with the simplified Laplace approximation exceeds 0.25. These nodes are potential 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. As an example, we display in (b) the marginals for the "worst case" which is node $(61, 73)$ with a SKLD of 0.50: Gaussian (dashed), simplified Laplace (solid) and Laplace approximations (dash-dotted). Note that the approximations becomes more close, as we improve the approximations. Panel (c) to (e) display the posterior marginals computed with the Gaussian approximations (dashed) and the one computed with the simplified Laplace approximations (solid) for β_0, β_{Alt} and

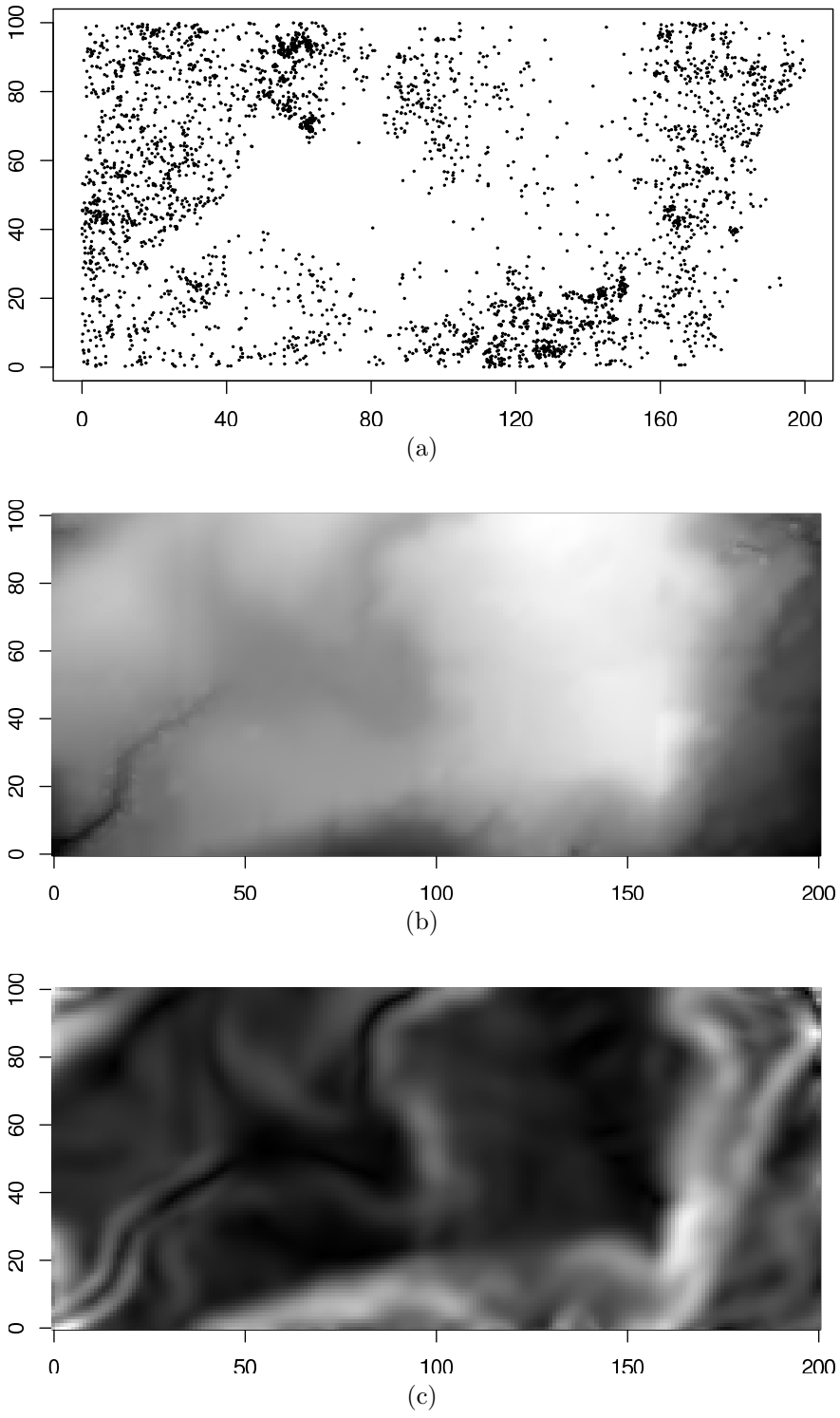


Fig. 6. Data and covariates for the log-Gaussian Cox process example: (a) locations of the $3,605$ trees, (b) altitude, and (c) norm of the gradient.

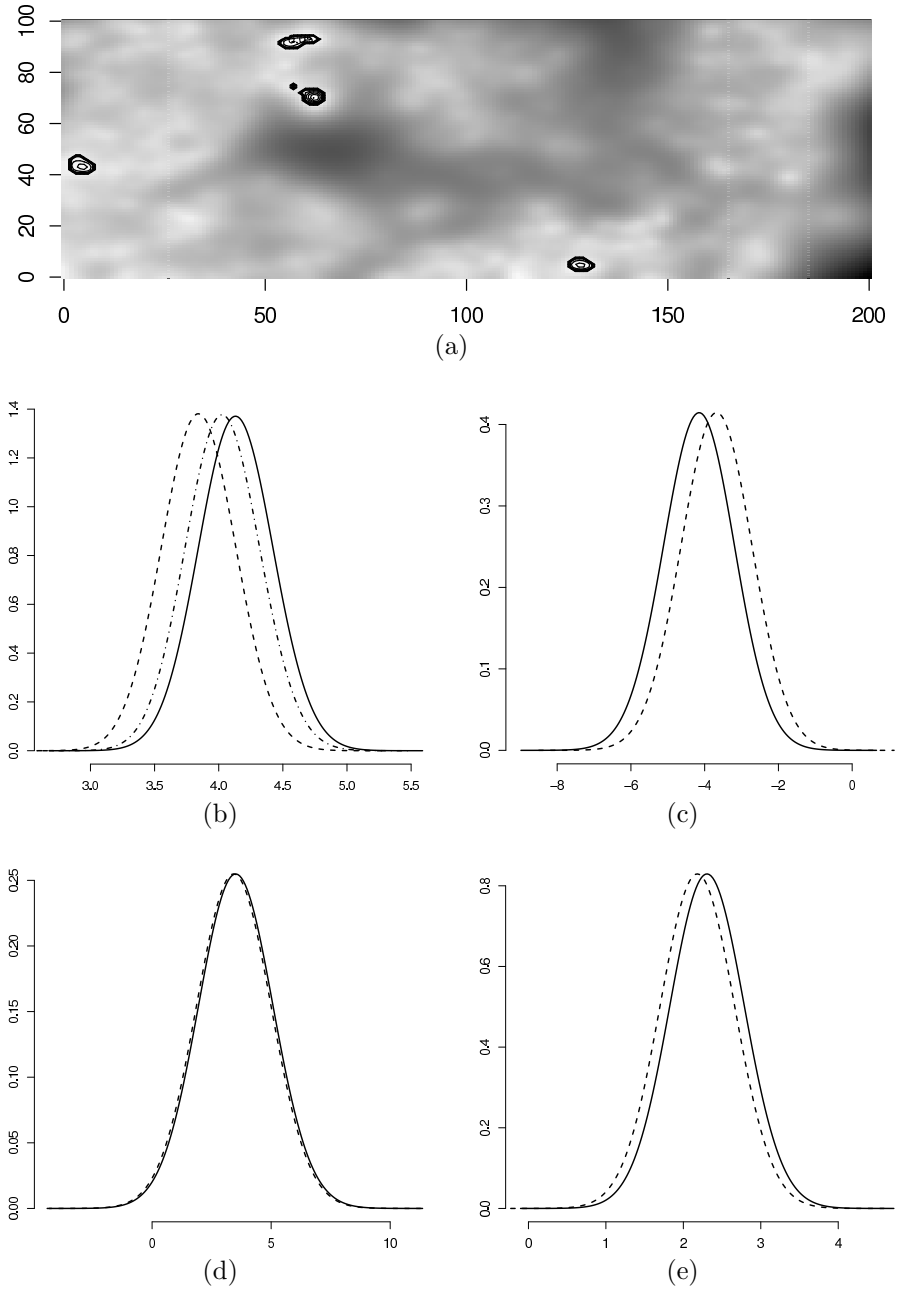


Fig. 7. LGCP example: (a) posterior mean of the spatial component with contour indicating a SKLD above 0.25, (b) the marginals for node $(61, 73)$ in the spatial component with maximum SKLD of 0.50, Gaussian (dashed), simplified Laplace (solid) and Laplace approximations (dash-dotted), (c)-(e) posterior marginals of β_0, β_{Alt} and β_{Grad} using simplified Laplace (solid) and Gaussian approximations (dashed).

β_{Grad}. The difference is mostly due to a horizontal shift, a characteristic valid for all the other nodes for this example.

This task required about 4 hours of computing due to the high dimension and the number of computed posterior marginals. The total cost can be reduced to about 10 minutes if only using the Gaussian approximation (4). To validate the approximations, we computed $p_{\text{D}}(\boldsymbol{\theta}^*) \approx 1714$ and estimated a 95% interval for the remainder $r(\boldsymbol{x}; \boldsymbol{\theta}^\star, \boldsymbol{y})/n_d$ as $[0.004, 0.01]$ 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 and the conditional marginals in the spatial field for fixed values of $\boldsymbol{\theta}$, both with excellent results. We used the one-block MCMC sampler described in Rue and Held (2005, Sec. 4.4.2).

6. Extensions

While this paper focuses on posterior marginals, INLA makes it possible to compute routinely other quantities as well. This section discusses some of these extensions.

6.1. Approximating posterior marginals for $\boldsymbol{x}_S$

A natural extension is to consider not only posterior marginals for x_i, but for a subset $\boldsymbol{x}_S = \{x_i : i \in S\}$. S can be small, say from 2 to 5, but sometimes larger sets are required. Although the Laplace approximation (12) can still be applied, replacing x_i with $\boldsymbol{x}_S$, and $\boldsymbol{x}_{-i}$ with $\boldsymbol{x}_{-S}$, practicalities get more involved. We tentatively recommend, unless extreme accuracy is required, the following approach for which the joint marginal for (near) any subset is directly available. To fix ideas, let $S = \{i, j\}$ where $i \sim j$, and keep $\boldsymbol{\theta}$ fixed. Let F_i and F_j be the (approximated) cumulative distribution functions of $x_i|\boldsymbol{\theta}, y$ and $x_j|\boldsymbol{\theta}, y$. From the Gaussian approximation $\widetilde{\pi}_G(\boldsymbol{x}|\boldsymbol{\theta}, \boldsymbol{y})$ we know the Gaussian marginal distribution for $x_i, x_j|\boldsymbol{\theta}, \boldsymbol{y}$. We have usually observed in our experiments that the correction in the mean (21) is far more important than the correction for skewness. Since correcting the mean in a Gaussian distribution does not alter the correlations, we suggest to approximate $x_i, x_j|\boldsymbol{\theta}, \boldsymbol{y}$ using the Gaussian copula and the marginals F_i and F_j. The benefit of this approach is that the marginals are kept unchanged and the construction is purely explicit. A simple choice is to use Gaussian marginals but with the mean-correction $\{\gamma_i^{(1)}\}$; see (21). Extending this approach to larger sets S is immediate, although the resulting accuracy may possibly decrease with the size of S.

6.2. Approximating the marginal likelihood

The marginal likelihood $\pi(\boldsymbol{y})$ is a useful quantity for comparing models, as Bayes factors are defined as ratios of marginal likelihoods of two competing models. It is evident from (3) that the natural approximation to the marginal likelihood is the normalising constant of $\widetilde{\pi}(\boldsymbol{\theta}|\boldsymbol{y})$,

$$\widetilde{\pi}(\boldsymbol{y}) = \int \left. \frac{\pi(\boldsymbol{\theta}, \boldsymbol{x}, \boldsymbol{y})}{\widetilde{\pi}_{\text{G}}(\boldsymbol{x}|\boldsymbol{\theta}, \boldsymbol{y})} \right|_{\boldsymbol{x}=\boldsymbol{x}^\star(\boldsymbol{\theta})} d\boldsymbol{\theta}. \tag{30}$$

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 (30) 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 (30) does not require this assumption, since we treat $\widetilde{\pi}(\boldsymbol{\theta}|\boldsymbol{y})$ in a 'nonparametric' way. This allows for taking into account the departure from Gaussianity which, for instance, appears clearly in Figure 4. Friel and Rue (2007) use a similar expression as (30) to approximate the marginal likelihood in a different context.

As an example, let us reconsider the stochastic volatility example in Section 5.3. Using (30), the log marginal likelihoods were computed to be -924.0 and -924.8 for the Gaussian and Student-t_ν observational model, respectively. The cruder approximation by Kass and Vaidyanathan (1992) gave similar results: -924.0 and -924.7. There is no evidence that a Student-t_ν observational model is required for these data.

As pointed out by a referee, this method could fail in case the posterior marginal $\pi(\boldsymbol{\theta}|\boldsymbol{y})$ is multi-modal (if not detected), but this is not specific to the evaluation of the marginal likelihood but applies to our general approach. Fortunately, latent Gaussian models generates unimodal posterior distributions in most cases.

6.3. Predictive measures

Predictive measures can be used both to validate and compare models (Gelfand, 1996; Gelman et al., 2004), and as a device to detect possible outliers or surprising observations (Pettit, 1990). One usually looks at the predictive density for the observed y_i based on all the other observations, i.e. $\pi(y_i|\boldsymbol{y}_{-i})$. We now explain how to approximate this quantity simply, without reanalysing the model. First, note that removing y_i from the dataset affects the marginals of x_i and $\boldsymbol{\theta}$ as follows:

$$\pi(x_i \mid \boldsymbol{y}_{-i}, \boldsymbol{\theta}) \propto \frac{\pi(x_i|\boldsymbol{y}, \boldsymbol{\theta})}{\pi(y_i|x_i, \boldsymbol{\theta})} \quad \text{and} \quad \pi(\boldsymbol{\theta} \mid \boldsymbol{y}_{-i}) \propto \frac{\pi(\boldsymbol{\theta}|\boldsymbol{y})}{\pi(y_i|\boldsymbol{y}_{-i}, \boldsymbol{\theta})}$$

where a one-dimensional integral is required to compute

$$\pi(y_i|\boldsymbol{y}_{-i}, \boldsymbol{\theta}) = \int \pi(y_i|x_i, \boldsymbol{\theta})\, \pi(x_i|\boldsymbol{y}_{-i}, \boldsymbol{\theta})\, dx_i.$$

The effect of $\boldsymbol{\theta}$ can then be integrated out from $\pi(y_i|\boldsymbol{y}_{-i}, \boldsymbol{\theta})$, in the same way as (5). Unusually small values of $\pi(y_i|\boldsymbol{y}_{-i})$ indicate surprising observations, but what is meant by 'small" must be calibrated with the level of x_i. Pettit (1990) suggests calibrating with the maximum value of $\pi(\cdot|\boldsymbol{y}_{-i})$, but an alternative is to compute the probability integral transform $\text{PIT}_i = \text{Prob}(y_i^{\text{new}} \leq y_i|\boldsymbol{y}_{-i})$ using the same device as above. (See also Gneiting and Raftery (2007) of a discussion of other alternatives.) An unusually small or large PIT_i (assuming continuous observations) indicates a possibly surprising observation which may require further attention. Furthermore, if the histogram of the PIT_i's is too far from a uniform, the model can be questioned (Czado et al., 2007).

As an example, let us reconsider the stochastic volatility example of Section 5.3. The Gaussian observational model indicates that three of the observations are surprising, i.e. PIT_i is close to one for $i = 331, 651$ and 862. These observations are less surprising under the Student-t_ν observation model: i.e. the same PIT_i are then about $1 - 5 \times 10^{-4}$.

6.4. Deviance Information Criteria

The Deviance Information Criterion (DIC) (Spiegelhalter et al., 2002) is a popular information criterion designed for hierarchical models, and (in most cases) is well defined for

improper priors. Its main application is Bayesian model selection, but it also provides a notion of the effective number of parameters, which we have used already; see (24). In our context, the deviance is

$$D(\boldsymbol{x}, \boldsymbol{\theta}) = -2 \sum_{i \in \mathcal{I}} \log \pi(y_i | x_i, \boldsymbol{\theta}) + \text{constant}.$$

DIC is defined as two times the mean of the deviance minus the deviance of the mean. The effective number of parameters is the mean of the deviance minus the deviance of the mean, for which (24) is a good approximation. The mean of the deviance can be computed in two steps: first, compute the conditional mean conditioned on $\boldsymbol{\theta}$ using univariate numerical integration for each $i \in \mathcal{I}$; second, integrate out $\boldsymbol{\theta}$ with respect to $\pi(\boldsymbol{\theta} \mid \boldsymbol{y})$. The deviance of the mean requires the posterior mean of each x_i, $i \in \mathcal{I}$, which is computed from the posterior marginals of x_i's. Regarding the hyperparameters, we prefer to use the posterior mode $\boldsymbol{\theta}^*$, as the posterior marginal for $\boldsymbol{\theta}$ can be severely skewed.

As an illustration, let us reconsider the example in Section 5.4. The effect of the age group was modelled as a smooth curve (7), but Figure 4(b) seems to indicate that a linear effect may be sufficient. However, this alternative model increases DIC by 33, so we reject it.

6.5. 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 discuss an alternative approach which reduces the computational cost dramatically for high m, at the expense of accuracy with respect to the numerical integration over $\widetilde{\pi}(\boldsymbol{\theta}|\boldsymbol{y})$. 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, and makes the integrand more and more Gaussian. Meaningful results can be obtained even using an extreme choice akin to empirical Bayes, that is, using only the modal configuration to integrate over $\pi(\boldsymbol{\theta}|\boldsymbol{y})$. This 'plug-in' approach will obviously underestimate variability, but it will provide reasonable results provided the uncertainty is the latent field is not dominated by the uncertainty in the hyperparameters.

An intermediate approach between full numerical integration and the 'plug-in' approach is now described. We 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

$$\begin{array}{llll}
(1,1,1,1,1,1), & (-1,1,1,1,-1), & (1,-1,1,1,-1), & (-1,-1,1,1,1), \\
(1,1,-1,1,-1), & (-1,1,-1,1,1), & (1,-1,-1,1,1), & (-1,-1,-1,1,-1), \\
(1,1,1,-1,-1), & (-1,1,1,-1,1), & (1,-1,1,-1,1), & (-1,-1,1,-1,-1), \\
(1,1,-1,-1,1), & (-1,1,-1,-1,-1), & (1,-1,-1,-1,-1) & \text{and } (-1,-1,-1,-1,1).
\end{array}$$

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 (5) and the scaling of the points, assume for simplicity that $\boldsymbol{\theta}|\boldsymbol{y}$ 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 $f_0\sqrt{m}$

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

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

The CCD integration scheme seems to provide useful results in all the cases we have considered so far. For all the examples in Section 5, as well as other models with higher dimension of $\boldsymbol{\theta}$ (Martino, 2007; Martino and Rue, 2008), the CCD scheme speeds computations up significantly while leaving the results nearly unchanged. There are cases where the integration of $\boldsymbol{\theta}$ has to be done more accurately, but these can be detected by comparing the results obtained using the empirical Bayes and the CCD approach. For these cases, the CCD integration seems to provide results half-way between the empirical and the full Bayesian approaches.

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 to analyse latent Gaussian models. A prototype of such a program, `inla`, is already available (Martino and Rue, 2008) and all the latent Gaussian models in Section 5 were specified and analysed using this program. `inla` is built upon the `GMRFLib`-library (Rue and Held, 2005, Appendix), which is open source and available from the first author's web page. (An interface to the `inla` program from `R` (R Development Core Team, 2007) is soon to come.) With respect to parallel implementation, we take advantage of the fact that we compute the approximation of $x_i|\boldsymbol{\theta}, \boldsymbol{y}$ independently for all i for fixed $\boldsymbol{\theta}$. Both the `inla` program and `GMRFLib` use the OpenMP API (see `www.openmp.org`) to speedup the computations for shared memory machines (read multi-core processors); however, we have not focused on these computational issues and speedups in this report. Parallel computing is particularly important for spatial or spatio-temporal latent Gaussian models, but also smaller models enjoy good speedup.

The main disadvantage of the INLA approach is that its computational cost is exponential with respect to 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 that it appears at first glance: the central composite design approach seems promising and provides reasonable results when m is not small, in the case where the user do not want to take an empirical Bayes approach and will not wait for a full Bayesian analysis.

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, who has no time or patience to wait for the results of a MCMC algorithm, notably if he or she has to analyse many different datasets with the same model. It also makes it possible to use latent Gaussian models as baseline models, even in cases where non-Gaussian models are more appropriate. The ability to easily validate assumptions like linear or smooth effect of a covariate is important, and our approach also gives access to Bayes factors, various predictive measures and DIC, which are useful tools to compare models and challenge the model under study.

Acknowledgement

The authors acknowledge all the good comments and questions from the Research Section Committee, the referees, and stimulating discussions with J. Eidsvik, N. Friel, A. Frigessi, J. Hasslet, L. Held, H. W. Rognebakke, J. Rousseau, H. Tjelmeland, J. Tyssedal and R. Waagepetersen. The Center for Tropical Forest Science of the Smithsonian Tropical Research Institute provided the data in Section 5.5.

A. Variational Bayes for latent Gaussian models: An example

We consider a simple latent Gaussian model defined by

$$\theta \sim \Gamma(a, b), \quad \boldsymbol{x} \mid \theta \sim \mathcal{N}\left(\mathbf{0}, \frac{1}{\theta}\boldsymbol{R}^{-1}\right) \quad \text{and} \quad \boldsymbol{y} \mid \boldsymbol{x}, \theta \sim \mathcal{N}\left(\boldsymbol{x}, \frac{1}{\kappa}\boldsymbol{I}\right)$$

where κ is a fixed hyper-parameter. Standard calculations lead to $\boldsymbol{x}|\theta, \boldsymbol{y} \sim \mathcal{N}\left(\boldsymbol{m}(\theta), \boldsymbol{Q}(\theta)^{-1}\right)$ where $\boldsymbol{m}(\theta) = \kappa \boldsymbol{Q}(\theta)^{-1}\boldsymbol{y}$, $\boldsymbol{Q}(\theta) = \theta \boldsymbol{R} + \kappa \boldsymbol{I}$ and

$$\pi(\theta \mid \boldsymbol{y}) \propto \frac{\theta^{a+n/2-1}}{|\boldsymbol{Q}(\theta)|^{1/2}} \exp\left\{-b\theta + \frac{\kappa^2}{2}\boldsymbol{y}^T \boldsymbol{Q}(\theta)^{-1}\boldsymbol{y}\right\}.$$

When $\kappa \to 0$, $\pi(\theta|\boldsymbol{y}) \to \Gamma(\theta;\ a, b)$, but in general, $\pi(\theta|\boldsymbol{y})$ is not a Gamma density. The Laplace approximation for $\theta|\boldsymbol{y}$ is exact since $\boldsymbol{y}$ is conditionally Gaussian. We now derive the VB approximation $q(\boldsymbol{x}, \theta)$ of $\pi(\theta, \boldsymbol{x}|\boldsymbol{y})$ under the assumption that $q(\boldsymbol{x}, \theta)$ minimises the Kullback-Leibler contrast of $\pi(\boldsymbol{x}, \theta|\boldsymbol{y})$ relatively to $q(\boldsymbol{x}, \theta)$, constrained to $q(\boldsymbol{x}, \theta) = q_{\boldsymbol{x}}(\boldsymbol{x}) q_\theta(\theta)$. The solution is obtained iteratively, see e.g. Beal (2003)

$$\begin{aligned} q_{\boldsymbol{x}}^{(t+1)}(\boldsymbol{x}) &\propto \exp\left\{\mathrm{E}_{q_\theta^{(t)}(\theta)} \log \pi(\boldsymbol{x}, \boldsymbol{y}|\theta)\right\}, \\ q_\theta^{(t+1)}(\theta) &\propto \pi(\theta) \exp\left\{\mathrm{E}_{q_{\boldsymbol{x}}^{(t)}(\boldsymbol{x})} \log \pi(\boldsymbol{x}, \boldsymbol{y}|\theta)\right\}. \end{aligned}$$

For our model, this gives $q_{\boldsymbol{x}}^{(t+1)}(\boldsymbol{x}) = \mathcal{N}(\boldsymbol{x};\ \boldsymbol{m}_{t+1}, \boldsymbol{Q}_{t+1}^{-1})$ where $\boldsymbol{m}_{t+1} = \kappa \boldsymbol{Q}_{t+1}^{-1}\boldsymbol{y}$, $\boldsymbol{Q}_{t+1} = \boldsymbol{R}(a + n/2)/b_t + \kappa \boldsymbol{I}$, and $q_\theta^{(t+1)}(\theta)$ is a $\Gamma(\theta;\ a + n/2, b_{t+1})$ density with $b_{t+1} = b + \boldsymbol{m}_{t+1}^T \boldsymbol{R} \boldsymbol{m}_{t+1} +$ Trace $\left(\boldsymbol{R}\boldsymbol{Q}_{t+1}^{-1}\right)$. The limit b_∞ of b_t is defined implicitly by equation:

$$\begin{aligned} b_\infty &= b + \kappa^2 \boldsymbol{y}^T \left(\frac{a + n/2}{b_\infty}\boldsymbol{R} + \kappa \boldsymbol{I}\right)^{-1} \boldsymbol{R} \left(\frac{a + n/2}{b_\infty}\boldsymbol{R} + \kappa \boldsymbol{I}\right)^{-1} \boldsymbol{y} \\ &+ \mathrm{Trace}\left(\left\{\frac{a + n/2}{b_\infty}\boldsymbol{I} + \kappa \boldsymbol{R}^{-1}\right\}^{-1}\right) \end{aligned}$$

which is not tractable. However, when $\kappa \to 0$, this transforms into $b_\infty = b + n b_\infty / \{2(a + n/2)\}$ hence $\lim_{\kappa \to 0} b_\infty = b(a + n/2)/a$. This means that, for data that are not very informative, the posterior marginal for θ is close to a $\Gamma(a, b)$ density, whereas the VB approximation is a $\Gamma(a+n/2, b(a+n/2)/a)$ density. The expectations agree, but the variance ratio is $\mathcal{O}(n)$. Numerical experiments confirmed these findings; for most reasonable values of κ, the variance estimated by VB is significantly smaller than the true posterior variance of θ. For non-Gaussian data we obtained similar empirical results.

B. Fitting the skew-Normal distribution

We explain here how to fit the skew-Normal distribution (23) to an expansion of the form

$$\log \pi(x) = \text{constant} - \frac{1}{2}x^2 + \gamma^{(1)}x + \frac{1}{6}\gamma^{(3)}x^3 + \cdots. \tag{31}$$

To second order, (31) 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 (23) 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:

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

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

$$\frac{\sqrt{2}(4-\pi)}{\pi^{3/2}}\left(\frac{a}{\omega}\right)^3 + \text{higher order terms.} \tag{32}$$

and imposes that (32) equals $\gamma^{(3)}$. This gives explicit formulae for the three parameters of the skewed-normal.

References

Abellan, J. J., S. Richardson, and N. Best (2007). Spatial versus spatiotemporal disease mapping. *Epidemiology 18*.

Ainsworth, L. M. and C. B. Dean (2006). Approximate inference for disease mapping. *Computational Statistics & Data Analysis 50*(10), 2552–2570.

Albert, J. H. and S. Chib (1993). Bayesian analysis of binary and polychotomous responce data. *Journal of the American Statistical Association 88*(422), 669–679.

Attias, H. (1999). Inferring parameters and structure of latent variable models by variational Bayes. In *Proceedings of the 15th Conf. on Uncertainty in Artificial Intelligence*, Volume 2.

Attias, H. (2000). A variational Bayesian framework for graphical models. *Advances in Neural Information Processing Systems 12*, 209–215.

Azzalini, A. and A. Capitanio (1999). Statistical applications of the multivariate skew-normal distribution. *Journal of the Royal Statistical Society, Series B 61*(4), 579–602.

Banerjee, S., B. P. Carlin, and A. E. Gelfand (2004). *Hierarchical Modeling and Analysis for Spatial Data*, Volume 101 of *Monographs on Statistics and Applied Probability*. London: Chapman & Hall.

Banerjee, S., A. E. Gelfand, A. O. Finley, and H. Sang (2008). Gaussian predictive process models for large spatial datasets. *Journal of the Royal Statistical Society, Series B xx*(xx), xx–xx. (to appear).

Beal, M. J. (2003). *Variational Algorithms for Approximate Bayesian Inference*. Ph. D. thesis, University College London.

Besag, J., P. J. Green, D. Higdon, and K. Mengersen (1995). Bayesian computation and stochastic systems (with discussion). *Statistical Science 10*(1), 3–66.

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

Bishop, C. M. (2006). *Pattern Recognition and Machine Learning*. Series: Information Science and Statistics. Springer-Verlag: New York.

Box, G. E. P. and K. B. Wilson (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 D. G. Clayton (1993). Approximate inference in generalized linear mixed models. *Journal of the American Statistical Association 88*(1), 9–25.

Brix, A. and P. J. Diggle (2001). Spatiotemporal prediction for log-Gaussian Cox processes. *Journal of the Royal Statistical Society, Series B 63*(4), 823–841.

Brix, A. and J. Møller (2001). Space-time multi type log Gaussian Cox processes with a view to modelling weeds. *Scandinavian Journal of Statistics 28*, 471–488.

Carter, C. K. and R. Kohn (1994). On Gibbs sampling for state space models. *Biometrika 81*(3), 541–543.

Chib, S., F. Nardari, and N. Shepard (2002). Markov chain Monte Carlo methods for stochastic volatility models. *Journal of Econometrics 108*, 281–316.

Chu, W. and Z. Ghahramani (2005). Gaussian processes for ordinal regression. *Journal of Machine Learning Research 6*, 1019–1041.

Cressie, N. A. C. (1993). *Statistics for spatial data*. Wiley Series in Probability and Mathematical Statistics: Applied Probability and Statistics. New York: John Wiley & Sons Inc. Revised reprint of the 1991 edition, A Wiley-Interscience Publication.

Cressie, N. A. C. and G. Johannesson (2008). Fixed rank kriging for very large spatial data sets. *Journal of the Royal Statistical Society, Series B 70*(1), 209–226.

Czado, C., T. Gneiting, and L. Held (2007). Predictive model assessment for count data. Technical Report Technical Report no 518, University of Washington, Department of Statistics.

Dey, D. K., S. K. Ghosh, and B. K. Mallick (Eds.) (2000). *Generalized Linear Models: A Bayesian Perspective*. Biostatistics series volume 5. Chapman & Hall/CRC.

Diggle, P. J. and P. J. Ribeiro (2006). *Model-based Geostatistics*. Springer Series in Statistics. Springer.

Durbin, J. and S. J. Koopman (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., S. Martino, and H. Rue (2008). Approximate Bayesian inference in spatial generalized linear mixed models. *Scandinavian Journal of Statistics xx*(xx), xx–xx. (to appear).

Fahrmeir, L. and S. Lang (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 G. Tutz (2001). *Multivariate Statistical Modelling Based on Generalized Linear Models* (2nd ed.). Berlin: Springer-Verlag.

Finkenstadt, B., L. Held, and V. Isham (Eds.) (2006). *Statistical Methods for Spatio-Temporal Systems*. Monographs on Statistics and Applied Probability. Chapman & Hall/CRC.

Friel, N. and H. Rue (2007). Recursive computing and simulation-free inference for general factorizable models. *Biometrika 94*(3), 661–672.

Frühwirth-Schnatter, S. and R. Frühwirth (2007). Auxiliary mixture sampling with applications to logistic models. *Computational Statistics & Data Analysis 51*(7), 3509–3528.

Frühwirth-Schnatter, S. and H. Wagner (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.

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 linear models. *Biometrika 85*(1), 215–227.

Gelfand, A. E. (1996). Model determination using sampling-based methods. In W. R. Gilks, S. Richardson, and D. J. Spiegelhalter (Eds.), *Markov Chain Monte Carlo in Practice*, pp. 145–161. London: Chapman & Hall.

Gelman, A., J. B. Carlin, H. S. Stern, and D. B. Rubin (2004). *Bayesian Data Analysis* (2nd ed.). Texts in Statistical Science Series. Chapman & Hall/CRC, Boca Raton, FL.

Gneiting, T. (2002). Nonseparable, stationary covariance functions for space-time data. *Journal of the American Statistical Association 97*, 590–600.

Gneiting, T. and A. E. Raftery (2007). Strictly proper scoring rules, prediction, and estimation. *Journal of the American Statistical Association 102*, 359–378.

Gschlößl, S. and C. Czado (2007). Modelling count data with overdispersion and spatial effects. *Statistical papers*. http://dx.doi.org/10.1007/s00362-006-0031-6.

Held, L., I. Natario, S. Fenton, H. Rue, and N. Becker (2005). Towards joint disease mapping. *Statistical Methods in Medical Research 14*(1), 61–82.

Hinton, G. E. and D. van Camp (1993). Keeping the neural networks simple by minimizing the description length of the weights. *Proceedings of the sixth annual conference on Computational learning theory*, 5–13.

Holmes, C. C. and L. Held (2006). Bayesian auxiliary variable models for binary and multinomial regression. *Bayesian Analysis 1*(1), 145–168.

Hsiao, C. K., S. Y. Huang, and C. W. Chang (2004). Bayesian marginal inference via candidate's formula. *Statistics and Computing 14*(1), 59–66.

Humphreys, K. and D. M. Titterington (2000). Approximate Bayesian inference for simple mixtures. In *Compstat: Proceedings in Computational Statistics, 14th Symposium Held in Utrecht, the Netherlands*. Physica Verlag.

Jordan, M. I. (2004). Graphical models. *Statistical Science 19*(1), 140–155.

Kamman, E. E. and M. P. Wand (2003). Geoadditive models. *Journal of the Royal Statistical Society, Series C 52*(1), 1–18.

Kass, R. E., L. Tierney, and J. B. Kadane (1999). The validity of posterior expansions based on Laplace's method. In S. Geisser, J. S. Hodgesa, S. J. Press, and A. Z. (eds) (Eds.), *Essays in Honor of George Bernard*, pp. 473–488. Amsterdam: North Holland.

Kass, R. E. and S. K. Vaidyanathan (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 W. Gersch (1996). *Smoothness Priors Analysis of Time Series*. Lecture Notes in Statistics no. 116. New York: Springer-Verlag.

Knorr-Held, L. (1999). Conditional prior proposals in dynamic models. *Scandinavian Journal of Statistics 26*(1), 129–144.

Knorr-Held, L., G. Raßer, and N. Becker (2002). Disease mapping of stage-specific cancer incidence data. *Biometrics 58*, 492–501.

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

Kohn, R. and C. F. Ansley (1987). A new algorithm for spline smoothing based on smoothing a stochastic process. *SIAM Journal of Scientific and Statistical Computing 8*(1), 33–48.

Kuss, M. and C. E. Rasmussen (2005). Assessing approximate inference for binary Gaussian process classification. *Journal of Machine Learning Research 6*, 1679–1704.

Lang, S. and A. Brezger (2004). Bayesian P-splines. *Journal of Computational and Graphical Statistics 13*(1).

MacKay, D. J. C. (1995). Ensemble learning and evidence maximization. In *Proceedings of the 1995 NIPS conference*.

MacKay, D. J. C. (1997). Ensemble learning for hidden Markov models. Technical report, Cavendish Laboratory, University of Cambridge.

Marroquin, J. L., F. A. Velasco, M. Rivera, and M. Nakamura (2001). Gauss-Markov measure field models for low-level vision. *IEEE Transactions on Pattern Analysis and Machine Intelligence 23*(4), 337–348.

Martino, S. (2007). *Approximate Bayesian inference for latent Gaussian models*. Ph. D. thesis, Department of Mathematical Sciences, Norwegian University of Science and Technology.

Martino, S. and H. Rue (2008). Implementing approximate Bayesian inference for latent Gaussian models using integrated nested Laplace approximations: A manual for the inla-program. Technical Report no 2, Department of mathematical sciences, Norwegian University of Science and Technology.

Minka, T. P. (2001). Expectation propagation for approximate Bayesian inference. *Uncertainty in Artificial Intelligence 17*, 362–369.

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

Møller, J. and R. Waagepetersen (2003). *Statistical inference and simulation for spatial point processes*, Volume 100 of *Monographs on Statistics and Applied Probability*. London: Chapman & Hall.

Neal, R. M. (1998). Regression and classification using Gaussian process priors. In *Bayesian Statistics, 6*, pp. 475–501. New York: Oxford University Press.

O'Hagan, A. (1978). Curve fitting and optimal design for prediction (with discussion). *Journal of the Royal Statistical Society, Series B 40*(1), 1–42.

Papaspiliopoulos, A. B. O., G. O. Roberts, and P. Fearnhead (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., G. O. Roberts, and M. Sköld (2007). A general framework for the parameterization of hierarchical models. *Statistical Science 22*(1), 59–73.

Pettit, L. I. (1990). The conditional predictive ordinate for the normal distribution. *Journal of the Royal Statistical Society, Series B 52*(1), 175–184.

R Development Core Team (2007). *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing. ISBN 3-900051-07-0.

Rasmussen, C. E. and C. K. I. Williams (2006). *Gaussian Processes for Machine Learning*. the MIT Press.

Rellier, G., X. Descombes, J. Zerubia, and F. Falzon (2002). A Gauss-Markov model for hyperspectral texture analysis of urban areas. In *Proceedings from the 16th International Conference on Pattern Recognition*, pp. I: 692–695.

Robert, C. P. and G. Casella (1999). *Monte Carlo Statistical Methods*. New York: Springer-Verlag.

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

Rue, H. and S. Martino (2007). Approximate Bayesian inference for hierarchical Gaussian Markov random fields models. *Journal of Statistical Planning and Inference 137*(10), 3177–3192. Special Issue: Bayesian Inference for Stochastic Processes.

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

Sanchez, S. M. and P. J. Sanchez (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* (2nd ed.). Springer series in statistics. New York: Springer-Verlag.

Shephard, N. (1994). Partial non-Gaussian state space. *Biometrika 81*(1), 115–131.

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

Shun, Z. and P. McCullagh (1995). Laplace approximation of high dimensional integrals. *Journal of the Royal Statistical Society, Series B 57*(4), 749–760.

Smith, A. F. M., A. M. Skene, J. E. H. Shaw, and J. C. Naylor (1987). Progress with numerical and graphical methods for practical Bayesian statistics. *The Statistician 36*(2/3), 75–82.

Spiegelhalter, D. J., N. G. Best, B. P. Carlin, and A. van der Linde (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.

Thall, P. F. and S. C. Vail (1990). Some covariance models for longitudinal count data with overdispersion. *Biometrics 46*, 657–671.

Thomas, A., B. O'Hara, U. Ligges, and S. Sturtz (2006). Making BUGS open. *R News 6*(1), 12–16.

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

Titterington, D. M. (2004). Bayesian methods for neural networks and related models. *Statistical Science 19*(1), 128–139.

Waagepetersen, R. P. (2007). An estimating function approach to inference for inhomogeneous Neyman-Scott processes. *Biometrics 63*(1), 252–258.

Wahba, G. (1978). Improper priors, spline smoothing and the problem of guarding against model errors in regression. *Journal of the Royal Statistical Society, Series B 40*(3), 364–372.

Wakefield, J. (2007). Disease mapping and spatial regression with count data. *Biostatistics 8*(2), 158–183.

Wakefield, J. C., N. G. Best, and L. A. Waller (2000). Bayesian approaches to disease mapping. In P. Elliot, J. C. Wakefield, N. G. Best, and D. J. Briggs (Eds.), *Spatial Epidemiology: Methods and Applications*, pp. 104–107. Oxford: Oxford University Press.

Wang, B. and D. M. Titterington (2005). Inadequacy of interval estimates corresponding to variational Bayesian approximations. In R. G. Cowell and Z. Ghahramani (Eds.), *Proceedings of the Tenth International Workshop on Artificial Intelligence and Statistics*, pp. 373–380.

Wang, B. and D. M. Titterington (2006). Convergence properties of a general algorithm for calculating variational Bayesian estimates for a normal mixture model. *Bayesian Analysis 1*(3), 625–650.

Wecker, W. E. and C. F. Ansley (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 A. N. Pettitt (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.

West, M. and J. Harrison (1997). *Bayesian Forecasting and Dynamic Models* (2nd ed.). Springer Series in Statistics. New York: Springer-Verlag.

Williams, C. K. I. and D. Barber (1998). Bayesian classification with Gaussian processes. *IEEE Transactions on Pattern Analysis and Machine Intelligence 20*(12), 1342–1351.

Zoeter, O., T. Heskes, and B. Kappen (2005). Gaussian quadrature based expectation propagation. In Z. Ghahramani and R. Cowell (Eds.), *Proceedings AISTATS 2005*.