

Proofs subject to correction. Not to be reproduced without permission. Contributions to the discussion must not exceed 400 words. Contributions longer than 400 words will be cut by the editor.

J. R. Statist. Soc. B (2009)
71, Part 2, pp. 1–35

Approximate Bayesian inference for latent Gaussian models by using integrated nested Laplace approximations

Håvard Rue and Sara Martino

Norwegian University for Science and Technology, Trondheim, Norway

and Nicolas Chopin

Centre de Recherche en Economie et Statistique and Ecole Nationale de la Statistique et de l'Administration Economique, Paris, France

[Read before The Royal Statistical Society at a meeting organized by the Research Section on Wednesday, October 15th, 2008, Professor I. L. Dryden in the Chair]

Summary. Structured additive regression models are perhaps the most commonly used class of models in statistical applications. It includes, among others, (generalized) linear models, (generalized) additive models, smoothing spline models, state space models, semiparametric regression, spatial and spatiotemporal models, log-Gaussian Cox processes and geostatistical and geospatial models. 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 owing to the non-Gaussian response variables. For such models, Markov chain Monte Carlo methods can be implemented, but they are not without problems, in terms of both convergence and computational time. In some practical applications, the extent of these problems is such that Markov chain Monte Carlo sampling 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 Markov chain Monte Carlo algorithms need hours or days to run, our approximations provide more precise estimates in seconds or 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.

Keywords: Approximate Bayesian inference; Gaussian Markov random fields; Generalized additive mixed models; Laplace approximation; Parallel computing; Sparse matrices; Structured additive regression models

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 extensively 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

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

1 belong to an exponential family, where the mean μ_i is linked to a structured additive predictor
 2 η_i through a link function $g(\cdot)$, so that $g(\mu_i) = \eta_i$. The structured additive predictor η_i accounts
 3 for effects of various covariates in an additive way:

$$4 \quad \eta_i = \alpha + \sum_{j=1}^{n_f} f^{(j)}(u_{ji}) + \sum_{k=1}^{n_\beta} \beta_k z_{ki} + \varepsilon_i. \quad (1)$$

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

16 The main aim of this paper is twofold:

- 17 (a) to provide accurate and fast deterministic approximations to all, or some of, the n pos-
 18 terior marginals for x_i , the components of latent Gaussian vector \mathbf{x} , plus possibly the
 19 posterior marginals for $\boldsymbol{\theta}$ or some of its components θ_j (if needed, the marginal densities
 20 can be post-processed to compute quantities like posterior expectations, variances and
 21 quantiles);
- 22 (b) to demonstrate how to use these marginals
 - 23 (i) to provide adequate approximations to the posterior marginal for subvectors \mathbf{x}_S for
 24 any subset S ,
 - 25 (ii) to compute the marginal likelihood and the deviance information criterion (DIC)
 26 for model comparison and
 - 27 (iii) to compute various Bayesian predictive measures.

30 1.2. Latent Gaussian models: applications

31 Latent Gaussian models have a numerous and wide ranging list of applications; most structured
 32 Bayesian models are in fact of this form; see for example Fahrmeir and Tutz (2001), Gelman *et al.*
 33 (2004) and Robert and Casella (1999). We shall first give some areas of applications grouped
 34 according to their physical dimension. Let $f(\cdot)$ denote one of the $f^{(j)}(\cdot)$ terms in equation (1)
 35 with variables f_1, f_2, \dots

- 36 (a) *Regression models*: Bayesian generalized linear models correspond to the linear predictor
 37 $\eta_i = \alpha + \sum_{k=1}^{n_\beta} \beta_k z_{ki}$ (Dey *et al.*, 2000). The $f(\cdot)$ terms are used either to relax the linear
 38 relationship of the covariate as argued for by Fahrmeir and Tutz (2001), or to introduce
 39 random effects or both. Popular models for modelling smooth effects of covariates are
 40 penalized spline models (Lang and Brezger, 2004) and random-walk models (Fahrmeir
 41 and Tutz, 2001; Rue and Held, 2005), or continuous indexed spline models (Wahba,
 42 1978; Wecker and Ansley, 1983; Kohn and Ansley, 1987; Rue and Held, 2005) or Gauss-
 43 ian processes (O’Hagan, 1978; Chu and Ghahramani, 2005; Williams and Barber, 1998;
 44 Besag *et al.*, 1995; Neal, 1998). Random effects make it possible to account for overdi-
 45 spersion caused by unobserved heterogeneity, or for correlation in longitudinal data, and
 46 can be introduced by defining $f(u_i) = f_i$ and letting $\{f_i\}$ be independent, zero mean and
 47 Gaussian (Fahrmeir and Lang, 2001).

- (b) *Dynamic models*: temporal dependence can be introduced by using i in equation (1) as a time index t and defining $f(\cdot)$ and covariate \mathbf{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.
- (c) *Spatial and spatiotemporal models*: spatial dependence can be modelled similarly, using a spatial covariate \mathbf{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 realizations of some kind. Popular models include the Besag–York–Mollié 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 dependences can be achieved either by using a spatiotemporal covariate (s, t) or a corresponding spatiotemporal 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 to separate the effects of various components in equation (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 \mathbf{x} be all the n Gaussian variables $\{\eta_i\}$, α , $\{f^{(j)}\}$ and $\{\beta_k\}$. The density $\pi(\mathbf{x}|\boldsymbol{\theta}_1)$ is Gaussian with (assumed) zero mean and precision matrix $\mathbf{Q}(\boldsymbol{\theta}_1)$ with hyperparameters $\boldsymbol{\theta}_1$. Denote by $\mathcal{N}(\mathbf{x}; \boldsymbol{\mu}, \boldsymbol{\Sigma})$, the $\mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ Gaussian density with mean $\boldsymbol{\mu}$ and covariance (inverse precision) $\boldsymbol{\Sigma}$ at configuration \mathbf{x} . Note that we have included $\{\eta_i\}$ instead of $\{\varepsilon_i\}$ in \mathbf{x} , as it simplifies the notation later.

The distribution for the n_d observational variables $\mathbf{y} = \{y_i : i \in \mathcal{I}\}$ is denoted by $\pi(\mathbf{y}|\mathbf{x}, \boldsymbol{\theta}_2)$ and we assume that $\{y_i : i \in \mathcal{I}\}$ are conditionally independent given \mathbf{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 $\mathbf{Q}(\boldsymbol{\theta})$)

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

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

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 \mathbf{x} , which is often of large dimension, $n = 10^2$ – 10^5 , admits conditional independence properties. Hence, the latent field is a Gaussian Markov random field (GMRF) with a sparse precision matrix $\mathbf{Q}(\boldsymbol{\theta})$ (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: Markov chain Monte Carlo approaches

The common approach to inference for latent Gaussian models is Markov chain Monte Carlo (MCMC) sampling. It is well known, however, that MCMC methods tend to exhibit poor performance when applied to such models. Various factors explain this. First, the components of the latent field \mathbf{x} are strongly dependent on each other. Second, $\boldsymbol{\theta}$ and \mathbf{x} are also strongly dependent, especially when n is large. A common approach to (try to) overcome this first problem is to construct a joint proposal based on a Gaussian approximation to the full conditional of \mathbf{x} (Ganerman, 1997, 1998; Carter and Kohn, 1994; Knorr-Held, 1999; Knorr-Held and Rue, 2002; Rue *et al.*, 2004). The second problem requires, at least partially, a joint update of both $\boldsymbol{\theta}$ and \mathbf{x} . One suggestion is to use the one-block approach of Knorr-Held and Rue (2002): make a proposal for $\boldsymbol{\theta}$ to $\boldsymbol{\theta}'$, update \mathbf{x} from the Gaussian approximation conditional on $\boldsymbol{\theta}'$, then accept or reject jointly; see Rue and Held (2005), chapter 4, for variations on this approach. Some models can alternatively be reparameterized 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 sampling 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.

- (a) Can we bypass MCMC methods entirely and base our inference on such closed form approximations?
- (b) 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 that is developed in this paper outperforms MCMC algorithms to such an extent that, for latent Gaussian models, resorting to MCMC sampling 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 sampling is a last resort solution: Monte Carlo averages are characterized by additive $\mathcal{O}_p(N^{-1/2})$ errors, where N is the simulated sample size. Thus, it is easy to obtain rough estimates, but nearly impossible to obtain 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 sampling that were mentioned above. On the other hand, Gaussian approximations are intuitively appealing for latent Gaussian models. For most real problems and data sets, the conditional posterior of \mathbf{x} is typically well behaved, and looks 'almost' Gaussian. This is clearly due to the latent Gaussian prior that is assigned to

1 \mathbf{x} , which has a non-negligible effect on the posterior, especially in terms of dependence between
2 the components of \mathbf{x} .

3 4 1.6. Approximation methods in machine learning

5 A general approach towards approximate inference is the variational Bayes (VB) methodology
6 that was developed in the machine learning literature (Hinton and van Camp, 1993; MacKay,
7 1995; Bishop, 2006). VB methodology has provided numerous promising results in various
8 areas, like hidden Markov models (MacKay, 1997), mixture models ((Humphreys and Titter-
9 ington, 2000), graphical models (Attias, 1999, 2000) and state space models (Beal, 2003), among
10 others; see Beal (2003), Titterington (2004) and Jordan (2004) for extensive reviews.

11 For the sake of discussion, consider the posterior distribution $\pi(\mathbf{x}, \boldsymbol{\theta}|\mathbf{y})$ of a generic Bayesian
12 model, with observation \mathbf{y} , latent variable \mathbf{x} and hyperparameter $\boldsymbol{\theta}$. The principle of VB meth-
13 ods is to use as an approximation the joint density $q(\mathbf{x}, \boldsymbol{\theta})$ that minimizes the Kullback–Leibler
14 contrast of $\pi(\mathbf{x}, \boldsymbol{\theta}|\mathbf{y})$ with respect to $q(\mathbf{x}, \boldsymbol{\theta})$. The minimization is subject to some constraint on
15 $q(\mathbf{x}, \boldsymbol{\theta})$, most commonly $q(\mathbf{x}, \boldsymbol{\theta}) = q_{\mathbf{x}}(\mathbf{x})q_{\boldsymbol{\theta}}(\boldsymbol{\theta})$. Obviously, the VB approximated density $q(\mathbf{x}, \boldsymbol{\theta})$
16 does not capture the dependence between \mathbf{x} and $\boldsymbol{\theta}$, but one hopes that its marginals (of \mathbf{x} and
17 $\boldsymbol{\theta}$) approximate well the true posterior marginals. The solution of this minimization problem is
18 approached through an iterative, EM-like algorithm.

19 In general, the VB approach is not without potential problems. First, even though VB meth-
20 ods seem often to approximate well the posterior mode (Wang and Titterington, 2006), the
21 posterior variance can be (sometimes severely) underestimated; see Bishop (2006), chapter 10,
22 and Wang and Titterington (2005). In the case of latent Gaussian models, this phenomenon
23 does occur as we demonstrate in Appendix A; we show that the VB-approximated variance can
24 be up to n times smaller than the true posterior variance in a typical application. The second
25 potential problem is that the iterative process of the basic VB algorithm is tractable for ‘conju-
26 gate exponential’ models only (Beal, 2003). This implies that $\pi(\boldsymbol{\theta})$ must be conjugate with respect
27 to the complete likelihood $\pi(\mathbf{x}, \mathbf{y}|\boldsymbol{\theta})$ and the complete likelihood must belong to an exponential
28 family. However, few of the latent Gaussian models that are encountered in applications are
29 of this type, as illustrated by our worked-through examples in Section 5. A possible remedy
30 around this requirement is to impose restrictions on $q(\mathbf{x}, \boldsymbol{\theta})$, such as independence between
31 blocks of components of $\boldsymbol{\theta}$ (Beal (2003), chapter 4), or a parametric form for $q(\mathbf{x}, \boldsymbol{\theta})$ that allows
32 for a tractable minimization algorithm. However, this requires case-specific solutions, and the
33 constraints will increase the approximation error.

34 Another approximation scheme that is popular in machine learning is the expectation–prop-
35 agation (EP) approach (Minka, 2001); see for example Zoeter *et al.* (2005) and Kuss and Ras-
36 mussen (2005) for applications of EP to latent Gaussian models. EP follows principles which
37 are quite similar to VB, i.e. it minimizes iteratively some pseudodistance between $\pi(\mathbf{x}, \boldsymbol{\theta}|\mathbf{y})$ and
38 the approximation $q(\mathbf{x}, \boldsymbol{\theta})$, subject to $q(\mathbf{x}, \boldsymbol{\theta})$ factorizing in a ‘simple’ way, e.g. as a product of
39 parametric factors, each involving a single component of $(\mathbf{x}, \boldsymbol{\theta})$. However, the pseudodistance
40 that is used in EP is the Kullback–Leibler contrast of $q(\mathbf{x}, \boldsymbol{\theta})$ relative to $\pi(\mathbf{x}, \boldsymbol{\theta}|\mathbf{y})$, rather than the
41 other way around (as in VB methods). Because of this, EP usually overestimates the posterior
42 variance (Bishop (2006), chapter 10). Kuss and Rasmussen (2005) derived an EP approximation
43 scheme for classification problems involving Gaussian processes that seems to be accurate and
44 fast; but their focus is on approximating $\pi(\mathbf{x}|\boldsymbol{\theta}, \mathbf{y})$ for $\boldsymbol{\theta}$ set to the posterior mode, and it is not
45 clear how to extend this approach to a fully Bayesian analysis. More importantly, deriving an
46 efficient EP algorithm seems to require specific efforts for each class of models. With respect
47 to computational cost, VB and EP methods are both designed to be faster than exact MCMC
48 methods, but, owing to their iterative nature, they are (much) slower than analytic approxima-

tions (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 methods can be implemented efficiently in scenarios involving linear constraints on \mathbf{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 interest can be written as

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

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

$$\begin{aligned}\tilde{\pi}(x_i|\mathbf{y}) &= \int \tilde{\pi}(x_i|\boldsymbol{\theta}, \mathbf{y}) \tilde{\pi}(\boldsymbol{\theta}|\mathbf{y}) \, d\boldsymbol{\theta}, \\ \tilde{\pi}(\theta_j|\mathbf{y}) &= \int \tilde{\pi}(\boldsymbol{\theta}|\mathbf{y}) \, d\boldsymbol{\theta}_{-j}.\end{aligned}\tag{2}$$

Here, $\tilde{\pi}(\cdot|\cdot)$ is an approximated (conditional) density of its arguments. Approximations to $\pi(x_i|\mathbf{y})$ are computed by approximating $\pi(\boldsymbol{\theta}|\mathbf{y})$ and $\pi(x_i|\boldsymbol{\theta}, \mathbf{y})$, 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 will become clear in what follows, the nested approach makes Laplace approximations very accurate when applied to latent Gaussian models. The approximation of $\pi(\theta_j|\mathbf{y})$ is computed by integrating out $\boldsymbol{\theta}_{-j}$ from $\tilde{\pi}(\boldsymbol{\theta}|\mathbf{y})$; we return in Section 3.1 to the practical details.

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

$$\tilde{\pi}(\boldsymbol{\theta}|\mathbf{y}) \propto \frac{\pi(\mathbf{x}, \boldsymbol{\theta}, \mathbf{y})}{\tilde{\pi}_G(\mathbf{x}|\boldsymbol{\theta}, \mathbf{y})} \Big|_{\mathbf{x}=\mathbf{x}^*(\boldsymbol{\theta})}\tag{3}$$

where $\tilde{\pi}_G(\mathbf{x}|\boldsymbol{\theta}, \mathbf{y})$ is the Gaussian approximation to the full conditional of \mathbf{x} , and $\mathbf{x}^*(\boldsymbol{\theta})$ is the mode of the full conditional for \mathbf{x} , for a given $\boldsymbol{\theta}$. The proportionality sign in expression (3) comes from the fact that the normalizing constant for $\pi(\mathbf{x}, \boldsymbol{\theta}|\mathbf{y})$ is unknown. This expression is equivalent to Tierney and Kadane's (1986) 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 renormalization. However, since n is not fixed but depends on n_d , standard asymptotic assumptions that are usually invoked for Laplace expansions are not verified here; see Section 4 for a discussion of the error rate.

$\tilde{\pi}(\boldsymbol{\theta}|\mathbf{y})$ itself tends to depart significantly from Gaussianity. This suggests that a cruder approximation based on a Gaussian approximation to $\pi(\boldsymbol{\theta}|\mathbf{y})$ is not sufficiently accurate for our purposes; this also applies to similar approximations that are based on 'equivalent Gaussian observations' around \mathbf{x}^* , and evaluated at the mode of expression (3) (Breslow and Clayton, 1993; Ainsworth and Dean, 2006). A critical aspect of our approach is to explore and manipulate $\tilde{\pi}(\boldsymbol{\theta}|\mathbf{y})$ and $\tilde{\pi}(x_i|\mathbf{y})$ in a 'non-parametric' way. Rue and Martino (2007) used expression (3) to approximate posterior marginals for $\boldsymbol{\theta}$ for various latent Gaussian models. Their conclusion was that $\tilde{\pi}(\boldsymbol{\theta}|\mathbf{y})$ is particularly accurate: even long MCMC runs could not detect any error in it. For the posterior marginals of the latent field, they proposed to start from $\tilde{\pi}_G(\mathbf{x}|\boldsymbol{\theta}, \mathbf{y})$ and to

approximate the density of $x_i|\boldsymbol{\theta}, \mathbf{y}$ with the Gaussian marginal derived from $\tilde{\pi}_G(\mathbf{x}|\boldsymbol{\theta}, \mathbf{y})$, i.e.

$$\tilde{\pi}(x_i|\boldsymbol{\theta}, \mathbf{y}) = \mathcal{N}\{x_i; \mu_i(\boldsymbol{\theta}), \sigma_i^2(\boldsymbol{\theta})\}. \quad (4)$$

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 expression (2)), to obtain approximations of the marginals of interest for the latent field,

$$\tilde{\pi}(x_i|\mathbf{y}) = \sum_k \tilde{\pi}(x_i|\boldsymbol{\theta}_k, \mathbf{y}) \times \tilde{\pi}(\boldsymbol{\theta}_k|\mathbf{y}) \times \Delta_k. \quad (5)$$

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, whereas the error in the Gaussian approximation (4) was higher. In particular, equation (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 could not control the error of the approximations and choose 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* (INLAs). The main tool is to apply the Laplace approximation once more, this time to $\pi(x_i|\mathbf{y}, \boldsymbol{\theta})$. We also present a faster alternative which corrects the Gaussian approximation (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, in terms of both accuracy and computational speed. We shall also demonstrate how the various approximations can be used to derive tools for assessing the approximation error, to approximate posterior marginals for a subset of \mathbf{x} , and to compute interesting quantities like the marginal likelihood, the DIC and various Bayesian predictive measures.

1.8. Plan of paper

Section 2 contains preliminaries on GMRFs, sparse matrix computations and Gaussian approximations. Section 3 explains the INLA approach and how to approximate $\pi(\boldsymbol{\theta}|\mathbf{y})$, $\pi(\theta_j|\mathbf{y})$ and $\pi(x_i|\boldsymbol{\theta}, \mathbf{y})$. For the latent field, three approximations are discussed: Gaussian, Laplace and simplified Laplace. Section 4 discusses the error rates of the Laplace approximations that are 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 \mathbf{x}_S , approximations of the marginal likelihood and predictive measures, the DIC 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 GMRFs and explain how to perform related computations using sparse matrix algorithms. We then discuss how to compute Gaussian approximations for

a latent GMRF. See Rue and Held (2005) for more details on both issues. Denote by \mathbf{x}_{-i} the vector \mathbf{x} minus its i th element 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 $\mathbf{x} = (x_1, \dots, x_n)$ with Markov properties: for some $i \neq j$, x_i and x_j are independent conditional on \mathbf{x}_{-ij} . These Markov properties are conveniently encoded in the precision (inverse covariance) matrix \mathbf{Q} : $Q_{ij} = 0$ if and only if x_i and x_j are independent conditional on \mathbf{x}_{-ij} . Let the undirected graph \mathcal{G} denote the conditional independence properties of \mathbf{x} ; then \mathbf{x} is said to be a GMRF with respect to \mathcal{G} . If the mean of \mathbf{x} is $\boldsymbol{\mu}$, the density of \mathbf{x} is

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

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

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

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

where $\boldsymbol{\Sigma} (= \mathbf{Q}^{-1})$ is the covariance matrix, and $\delta_{ij} = 1$ if $i = j$ and $\delta_{ij} = 0$ otherwise. Thus Σ_{ij} can be computed from expression (7), letting the outer loop i run from n to 1 and the inner loop j from n to i . If we are interested only in the marginal variances, we need to compute only Σ_{ij} s for which L_{ji} (or L_{ij}) is not known to be 0; see above. This reduces the computational costs to typically $\mathcal{O}\{n \log(n)^2\}$ in the spatial case; see Rue and Martino (2007) section 2, for more details.

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

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

is a sample from the constrained GMRF. The expected value of \mathbf{x}^c can also be computed by using equation (8). This approach is commonly called ‘conditioning by kriging’; see Cressie (1993) or Rue (2001). Note that $\mathbf{Q}^{-1}\mathbf{A}^T$ is computed by solving k linear systems, one for each column of \mathbf{A}^T . The additional cost of the k linear constraints is $\mathcal{O}(nk^2)$. Marginal variances under linear constraints can be computed in a similar way; see Rue and Martino (2007), section 2.

2.2. Gaussian approximations

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

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

where $g_i(x_i)$ is $\log\{\pi(y_i|x_i, \boldsymbol{\theta})\}$ in our setting. The Gaussian approximation $\tilde{\pi}_G(\mathbf{x})$ is obtained by matching the modal configuration and the curvature at the mode. The mode is computed iteratively by using a Newton–Raphson method, which is 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,

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

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

Since the non-quadratic term in expression (9) is only a function of x_i and not a function of x_j and x_k , say, the precision matrix of the Gaussian approximation is of the form $\mathbf{Q} + \text{diag}(\mathbf{c})$. This is computationally convenient, as the Markov properties of the GMRF are preserved.

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

3. The integrated nested Laplace approximation

In this section we present the INLA approach for approximating the posterior marginals of the latent Gaussian field, $\pi(x_i|\mathbf{y})$, $i = 1, \dots, n$. The approximation is computed in three steps. The first step (Section 3.1) approximates the posterior marginal of $\boldsymbol{\theta}$ by using the Laplace approximation (3). The second step (Section 3.2) computes the Laplace approximation, or the simplified Laplace approximation, of $\pi(x_i|\mathbf{y}, \boldsymbol{\theta})$, for selected values of $\boldsymbol{\theta}$, to improve on the Gaussian approximation (4). The third step combines the previous two by using numerical integration (5).

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

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

- (a) *Step 1:* locate the mode of $\tilde{\pi}(\boldsymbol{\theta}|\mathbf{y})$, by optimizing $\log\{\tilde{\pi}(\boldsymbol{\theta}|\mathbf{y})\}$ with respect to $\boldsymbol{\theta}$. This can be done by using some quasi-Newton method which builds up an approximation to the

second derivatives of $\log\{\tilde{\pi}(\boldsymbol{\theta}|\mathbf{y})\}$ by using the difference between successive gradient vectors. The gradient is approximated by using finite differences. Let $\boldsymbol{\theta}^*$ be the modal configuration.

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

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

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

- (c) *Step 3*: explore $\log\{\tilde{\pi}(\boldsymbol{\theta}|\mathbf{y})\}$ by using the \mathbf{z} -parameterization. Fig. 1 illustrates the procedure when $\log\{\tilde{\pi}(\boldsymbol{\theta}|\mathbf{y})\}$ is unimodal. Fig. 1(a) shows a contour plot of $\log\{\tilde{\pi}(\boldsymbol{\theta}|\mathbf{y})\}$ for $m=2$, the location of the mode and the new co-ordinate axis for \mathbf{z} . We want to explore $\log\{\tilde{\pi}(\boldsymbol{\theta}|\mathbf{y})\}$ to locate the bulk of the probability mass. The result of this procedure is displayed in Fig. 1(b). Each dot is a point where $\log\{\tilde{\pi}(\boldsymbol{\theta}|\mathbf{y})\}$ is considered as significant, and which is used in the numerical integration (5). Details are as follows. We start from the mode ($\mathbf{z}=\mathbf{0}$) and go in the positive direction of z_1 with step length δ_z say $\delta_z=1$, as long as

$$\log[\tilde{\pi}\{\boldsymbol{\theta}(\mathbf{0}|\mathbf{y})\}] - \log[\tilde{\pi}\{\boldsymbol{\theta}(\mathbf{z})|\mathbf{y}\}] < \delta_\pi \quad (11)$$

where, for example $\delta_\pi = 2.5$. Then we switch direction and do similarly. The other co-ordinates are treated in the same way. This produces the black dots. We can now fill in all the intermediate values by taking all different combinations of the black dots. These new points (which are shown as grey dots) are included if condition (11) holds. Since we lay out the points $\boldsymbol{\theta}_k$ in a regular grid, we may take all the area weights Δ_k in equation (5) to be equal.

- (d) *Approximating $\pi(\theta_j|\mathbf{y})$* . Posterior marginals for θ_j can be obtained directly from $\tilde{\pi}(\boldsymbol{\theta}|\mathbf{y})$ by using numerical integration. However, this is computationally demanding, as we need to evaluate $\tilde{\pi}(\boldsymbol{\theta}|\mathbf{y})$ for a large number of configurations. A more feasible approach is to use the points that were already computed during steps 1–3 to construct an interpolant

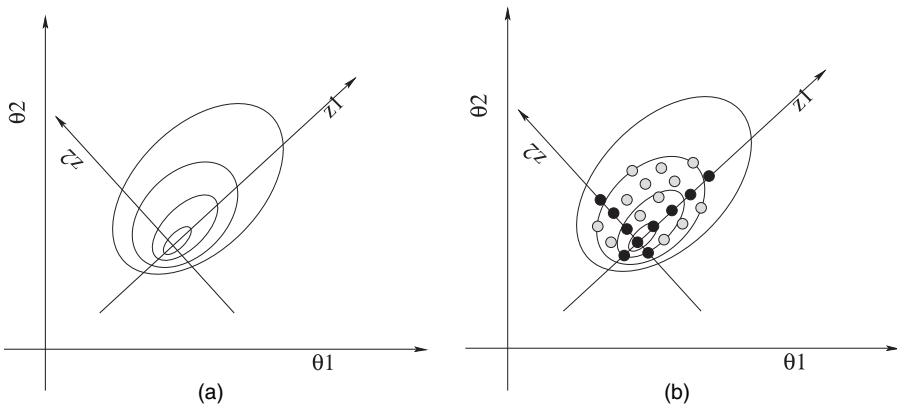


Fig. 1. Illustration of the exploration of the posterior marginal for $\boldsymbol{\theta}$: in (a) the mode is located and the Hessian and the co-ordinate system for \mathbf{z} are computed; in (b) each co-ordinate direction is explored (\bullet) until the log-density drops below a certain limit; finally the new points (\circ) are explored

to $\log\{\tilde{\pi}(\boldsymbol{\theta}|\mathbf{y})\}$, and to compute marginals by using numerical integration from this interpolant. If high accuracy is required, we need in practice a more dense configuration (for example $\delta_{\mathbf{z}} = \frac{1}{2}$ or $\delta_{\mathbf{z}} = \frac{1}{4}$) than is required for the latent field \mathbf{x} ; see Martino (2007) for numerical comparisons.

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

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 $\tilde{\pi}(x_i|\mathbf{y}, \boldsymbol{\theta}_k)$, i.e. 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 approximation generally compensates for the slight loss in accuracy.

3.2.1. Using Gaussian approximations

The simplest (and cheapest) approximation to $\pi(x_i|\boldsymbol{\theta}, \mathbf{y})$ is the Gaussian approximation $\tilde{\pi}_{\text{G}}(x_i|\boldsymbol{\theta}, \mathbf{y})$, where the mean $\mu_i(\boldsymbol{\theta})$ and the marginal variance $\sigma_i^2(\boldsymbol{\theta})$ are derived by using the recursions (7), and possibly correcting for linear constraints. During the exploration of $\tilde{\pi}(\boldsymbol{\theta}|\mathbf{y})$ (see Section 3.1), we already compute $\tilde{\pi}_{\text{G}}(\mathbf{x}|\boldsymbol{\theta}, \mathbf{y})$, so only marginal variances need to be additionally computed. The Gaussian approximation often gives reasonable results, but there can be errors in the location and/or errors due to the lack of skewness (Rue and Martino, 2007).

3.2.2. Using Laplace approximations

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

$$\tilde{\pi}_{\text{LA}}(x_i|\boldsymbol{\theta}, \mathbf{y}) \propto \frac{\pi(\mathbf{x}, \boldsymbol{\theta}, \mathbf{y})}{\tilde{\pi}_{\text{GG}}(\mathbf{x}_{-i}|x_i, \boldsymbol{\theta}, \mathbf{y})} \Big|_{\mathbf{x}_{-i}=\mathbf{x}_{-i}^*(x_i, \boldsymbol{\theta})}. \quad (12)$$

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

Unfortunately, expression (12) implies that $\tilde{\pi}_{\text{GG}}$ 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 factorizations of the full precision matrix. We propose two modifications to expression (12) which make it computationally feasible.

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

$$\mathbf{x}_{-i}^*(x_i, \boldsymbol{\theta}) \approx E_{\tilde{\pi}_{\text{G}}}(\mathbf{x}_{-i}|x_i). \quad (13)$$

The right-hand side is evaluated under the conditional density that is derived from the Gaussian approximation $\tilde{\pi}_{\text{G}}(\mathbf{x}|\boldsymbol{\theta}, \mathbf{y})$. The computational benefit is immediate. First, the conditional mean can be computed by a rank 1 update from the unconditional mean, by using equation (8). In the spatial case the cost is $\mathcal{O}\{n \log(n)\}$, for each i , which comes from solving $\mathbf{Q}^*(\boldsymbol{\theta})\mathbf{v} = \mathbf{1}_i$, where $\mathbf{1}_i$ equals 1 at position i and 0 otherwise. This rank 1 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) showed that deviating from the conditional mode does not necessarily degrade the approximation error. Another positive feature of approximation (13) is that the conditional mean is continuous with respect to x_i , which is not so when numerical optimization is used to compute $\mathbf{x}_{-i}^*(x_i, \boldsymbol{\theta})$.

Our next modification materializes the following intuition: only those x_j that are ‘close’ to x_i should have an effect on the marginal of x_i . If the dependence 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 approximation (13) implies that

$$\frac{E_{\tilde{\pi}_G}(x_j|x_i) - \mu_j(\boldsymbol{\theta})}{\sigma_j(\boldsymbol{\theta})} = a_{ij}(\boldsymbol{\theta}) \frac{x_i - \mu_i(\boldsymbol{\theta})}{\sigma_i(\boldsymbol{\theta})} \quad (14)$$

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

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

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

Expression (12), simplified as explained above, must be computed for different values of x_i 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 standardized variable

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

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

$$\tilde{\pi}_{LA}(x_i|\boldsymbol{\theta}, \mathbf{y}) \propto \mathcal{N}\{x_i; \mu_i(\boldsymbol{\theta}), \sigma_i^2(\boldsymbol{\theta})\} \times \exp\{\text{cubic spline}(x_i)\}. \quad (17)$$

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

3.2.3. Using a simplified Laplace approximation

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

$$d_j^{(3)}(x_i, \boldsymbol{\theta}) = \frac{\partial^3}{\partial x_j^3} \log\{\pi(y_j|x_j, \boldsymbol{\theta})\} \Big|_{x_j = E_{\tilde{\pi}_G}(x_j|x_i)},$$

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

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

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

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

$$\log\{\pi(\mathbf{x}, \boldsymbol{\theta}, \mathbf{y})\} \Big|_{\mathbf{x}_i = E_{\tilde{\pi}_G}(\mathbf{x}_{-i}|x_i)} = -\frac{1}{2}(x_i^{(s)})^2 + \frac{1}{6}(x_i^{(s)})^3 \sum_{j \in \mathcal{I} \setminus i} d_j^{(3)}\{\mu_i(\boldsymbol{\theta}), \boldsymbol{\theta}\} \{\sigma_j(\boldsymbol{\theta}) a_{ij}(\boldsymbol{\theta})\}^3 + \dots \quad (18)$$

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

$$\log\{\tilde{\pi}_{\text{GG}}(\mathbf{x}_{-i}|x_i, \boldsymbol{\theta}, \mathbf{y})\} \Big|_{\mathbf{x}_{-i}=E_{\tilde{\pi}_{\text{G}}}(\mathbf{x}_{-i}|x_i)} = \text{constant} + \frac{1}{2} \log |\mathbf{H} + \text{diag}\{\mathbf{c}(x_i, \boldsymbol{\theta})\}| \quad (19)$$

where \mathbf{H} is the prior precision matrix of the GMRF with i th column and row deleted, and $\mathbf{c}(x_i, \boldsymbol{\theta})$ is the vector of minus the second derivative of the log-likelihood evaluated at $x_j = E_{\tilde{\pi}_{\text{G}}}(x_j|x_i)$; see equation (14). Using that

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

we obtain

$$\begin{aligned} & \log\{\tilde{\pi}_{\text{GG}}(\mathbf{x}_{-i}|x_i, \boldsymbol{\theta}, \mathbf{y})\} \Big|_{\mathbf{x}_{-i}=E_{\tilde{\pi}_{\text{G}}}(\mathbf{x}_{-i}|x_i)} \\ &= \text{constant} - \frac{1}{2} x_i^{(s)} \sum_{j \in \mathcal{I} \setminus i} \text{var}_{\tilde{\pi}_{\text{G}}}(x_j|x_i) d_j^{(3)} \{\mu_i(\boldsymbol{\theta}), \boldsymbol{\theta}\} \sigma_j(\boldsymbol{\theta}) a_{ij}(\boldsymbol{\theta}) + \dots \end{aligned} \quad (20)$$

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

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

and that the covariance between x_i and x_j (under $\tilde{\pi}_{\text{G}}$) is computed while doing the rank 1 update in approximation (13), as the j th element of the solution of $\mathbf{Q}^*(\boldsymbol{\theta})\mathbf{v} = \mathbf{1}_i$.

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

$$\begin{aligned} \gamma_i^{(1)}(\boldsymbol{\theta}) &= \frac{1}{2} \sum_{j \in \mathcal{I} \setminus i} \sigma_j^2(\boldsymbol{\theta}) \{1 - \text{corr}_{\tilde{\pi}_{\text{G}}}(x_i, x_j)^2\} d_j^{(3)} \{\mu_i(\boldsymbol{\theta}), \boldsymbol{\theta}\} \sigma_j(\boldsymbol{\theta}) a_{ij}(\boldsymbol{\theta}) \\ \gamma_i^{(3)}(\boldsymbol{\theta}) &= \sum_{j \in \mathcal{I} \setminus i} d_j^{(3)} \{\mu_i(\boldsymbol{\theta}), \boldsymbol{\theta}\} \{\sigma_j(\boldsymbol{\theta}) a_{ij}(\boldsymbol{\theta})\}^3; \end{aligned} \quad (21)$$

then

$$\log\{\tilde{\pi}_{\text{SLA}}(x_i^s|\boldsymbol{\theta}, \mathbf{y})\} = \text{constant} - \frac{1}{2} (x_i^{(s)})^2 + \gamma_i^{(1)}(\boldsymbol{\theta}) x_i^{(s)} + \frac{1}{6} (x_i^{(s)})^3 \gamma_i^{(3)}(\boldsymbol{\theta}) + \dots \quad (22)$$

Equation (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)

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

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 equation (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, $\gamma_i^{(3)}$ contributes only 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 -distribution 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 distribution. This is slightly more expensive, but it is needed.

The simplified Laplace approximation appears to be highly accurate for many observational models. The computational cost is dominated by the calculation of vector $a_i(\boldsymbol{\theta})$, for each i ; thus the ‘region of interest’ strategy (15) is unhelpful here. Most of the other terms in expression (21) do not depend on i and thus are computed only once. The cost for computing equation (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 that this is close to the lower limit for any general algorithm that approximates all the n marginals. Since the graph of \mathbf{x} is general, we need to visit all other sites, for each i , for a potential contribution. This operation alone costs $\mathcal{O}(n^2)$. In summary, the total cost for computing all n marginals $\tilde{\pi}(x_i|\mathbf{y})$, $i = 1, \dots, n$, using equation (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 $\tilde{\pi}(\boldsymbol{\theta}|\mathbf{y})$

For the sake of discussion, denote p the dimension of vector $(\mathbf{x}, \boldsymbol{\theta})$, i.e. $p = n + m$, and recall that n_d denotes the number of observations. Up to normalization, $\tilde{\pi}(\boldsymbol{\theta}|\mathbf{y})$ is formally equivalent to the Laplace approximation of a marginal posterior density that was 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 shall now discuss several asymptotic schemes and their influence on the actual error rate of $\tilde{\pi}(\boldsymbol{\theta}|\mathbf{y})$.

First, assume that p is fixed while $n_d \rightarrow \infty$; 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), page 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 a GMRF model with binomial observations, $y_i|x_i \sim \text{Bin}\{n_i, \text{logit}^{-1}(x_i)\}$, provided that all the n_i take large values.

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

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

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

the trace of the prior precision matrix times by the posterior covariance matrix of the Gaussian approximation (Spiegelhalter *et al.* (2002), equation (16)). (The computation of $p_D(\boldsymbol{\theta})$ is

1 computationally cheap, since the covariances of neighbours are obtained as a by-product of the
 2 computation of the marginal variances in the Gaussian approximation based on equation (7).
 3 This quantity also measures to what extent the Gaussianity and the dependence structure of
 4 the prior are preserved in the posterior of \mathbf{x} , given $\boldsymbol{\theta}$. For instance, for non-informative data,
 5 $p_D(\boldsymbol{\theta})=0$, and the approximation error is zero, since the posterior equals the Gaussian prior.
 6 In all our applications, we observed that $p_D(\boldsymbol{\theta})$ is typically small relative to n_d for values of $\boldsymbol{\theta}$
 7 neat the posterior mode.

8 Note finally that in most cases normalizing the approximated density reduces further the
 9 asymptotic rate, as the dominating terms of the numerator and the denominator cancel out
 10 (Tierney and Kadane, 1986); in the standard case, normalizing reduces the error rate from
 11 $\mathcal{O}(n_d^{-1})$ to $\mathcal{O}(n_d^{-3/2})$.

12 The discussion above of the asymptotic properties of $\tilde{\pi}(\boldsymbol{\theta}|\mathbf{y})$ applies almost directly to
 13 $\tilde{\pi}_{\text{LA}}(x_i|\boldsymbol{\theta}, \mathbf{y})$: conditional on $\boldsymbol{\theta}$, $\tilde{\pi}_{\text{LA}}(x_i|\boldsymbol{\theta}, \mathbf{y})$ is a Laplace approximation of the posterior mar-
 14 ginal density of x_i , and the dimension of the corresponding integral is the dimension of \mathbf{x}_{-i} , i.e.
 15 $n-1$.

17 4.2. Assessing the approximation error

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

22 Our first strategy is to verify the overall approximation $\tilde{\pi}_G(\mathbf{x}|\boldsymbol{\theta}, \mathbf{y})$, for each $\boldsymbol{\theta}_k$ that is used
 23 in the integration. We do this by computing $p_D(\boldsymbol{\theta})$ (24) as discussed in Section 4.1, but we can
 24 also use that expression (3) can be rewritten as

$$\begin{aligned} \frac{\pi(\boldsymbol{\theta}|\mathbf{y})}{\tilde{\pi}(\boldsymbol{\theta}|\mathbf{y})} &\propto |\mathbf{Q}^*(\boldsymbol{\theta})|^{1/2} \int \exp[-\frac{1}{2}(\mathbf{x} - \mathbf{x}^*(\boldsymbol{\theta}))^T \mathbf{Q}^*(\boldsymbol{\theta})(\mathbf{x} - \mathbf{x}^*(\boldsymbol{\theta})) + r(\mathbf{x}; \boldsymbol{\theta}, \mathbf{y})] d\mathbf{x} \\ &= E_{\tilde{\pi}_G}[\exp\{r(\mathbf{x}; \boldsymbol{\theta}, \mathbf{y})\}], \end{aligned}$$

26 where the constant of proportionality is quite involved and not needed in the following dis-
 27 cussion. Further, $\mathbf{x}^*(\boldsymbol{\theta})$ and $\mathbf{Q}^*(\boldsymbol{\theta})$ are the mean and precision of Gaussian distribution $\tilde{\pi}_G$,
 28 $r(\mathbf{x}; \boldsymbol{\theta}, \mathbf{y}) = \sum_i h_i(x_i)$, and $h_i(x_i)$ is $g_i(x_i)$ minus its Taylor expansion up to order 2 around $x_i^*(\boldsymbol{\theta})$;
 29 see expressions (9) and (10). If, for each $\boldsymbol{\theta}_k$, $p_D(\boldsymbol{\theta})$ is small compared with n_d , and the empirical
 30 quantiles of the random variable $r(\mathbf{x}; \boldsymbol{\theta}, \mathbf{y})$ are in absolute value significantly smaller than n_d ,
 31 then we have strong confidence that the Gaussian approximation is adequate. The empirical
 32 quantiles of $r(\mathbf{x}; \boldsymbol{\theta}, \mathbf{y})$ are found by sampling (e.g. 1000) independent realizations from $\tilde{\pi}_G$.

33 Our second strategy is based on the simple idea of comparing elements of a sequence of
 34 increasingly accurate approximations. In our case, this sequence consists of the Gaussian approx-
 35 imation (4), followed by the simplified Laplace approximation (22), then by the Laplace approx-
 36 imation (12). Specifically we compute the integrated marginal (5) on the basis of both the
 37 Gaussian approximation and the simplified Laplace approximation, and compute their sym-
 38 metric Kullback–Leibler divergence (SKLD). If the divergence is small then both approxi-
 39 mations are considered as acceptable. Otherwise, compute equation (5) by using the Laplace
 40 approximation (12) and compute the divergence with that based on the simplified Laplace
 41 approximation. Again, if the divergence is small, simplified Laplace and Laplace approxima-
 42 tions appear to be acceptable; otherwise, the Laplace approximation is our best estimate but the
 43 label ‘problematic’ should be attached to the approximation to warn the user. (This last option
 44 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 that are 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 with results that were obtained from intensive MCMC runs. The computations were performed on a single-processor 2.1-GHz laptop using the `inla` program (Martino and Rue, 2008) which is an easy-to-use interface to our `GMRFLib` library written in C (Rue and Held (2005), appendix). (We shall comment on speed-up strategies and parallel implementation in Section 6.5 and Section 7.) We start with some simulated examples with fixed θ in Section 5.1, to verify the (simplified) Laplace approximation for $x_i|\theta, \mathbf{y}$. We continue with a generalized linear mixed model for longitudinal data in Section 5.2, a stochastic volatility model applied to exchange rate data in Section 5.3 and a spatial semiparametric regression model for disease mapping in Section 5.4. The dimensions become really large in Section 5.5, in which we analyse some data by using a spatial log-Gaussian Cox process.

5.1. Simulated examples

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

$$f_t - \mu | \mu, f_1, \dots, f_{t-1} \sim \mathcal{N}\{\phi(f_{t-1} - \mu), \sigma^2\}, \quad t = 2, \dots, 50, \quad (25)$$

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

$$y_t - \eta_t | (\boldsymbol{\eta}, \mu) \sim \text{Student } t_3, \\ y_i | (\boldsymbol{\eta}, \mu) \sim \text{Bernoulli}\{\text{logit}^{-1}(\eta_i)\},$$

for $t = 1, \dots, 50$, in experiment 1 and 2 respectively. Note that the Student t_3 distribution 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 $(\mathbf{f}^\top, \mu)^\top$ from the prior, and then simulated the observations. We computed $\tilde{\pi}(f_t|\theta, \mathbf{y})$ for $t = 1, \dots, 50$ and $\tilde{\pi}(\mu|\theta, \mathbf{y})$ by using the simplified Laplace approximation. We located the ‘worst node’, i.e. the node with maximum SKLD between the Gaussian and the simplified Laplace approximations. This process was repeated 100 times. Fig. 2 provides the results for the ‘worst of the worst nodes’, i.e. the node that maximizes our SKLD criterion among all the nodes of the 100 generated sample. Figs 2(a), 2(c) and 2(e) display the results for experiment 1 with Student t_3 -data, and Figs 2(b), 2(d) and 2(f) display the results for experiment with Bernoulli data. Figs 2(a) and 2(b) display \mathbf{f} (full curves) and the observed data (circles). In Fig. 2(a) the selected node is marked with a vertical line and full dot. In Fig. 2(b) the node with maximum SKLD is μ and hence is not shown. Figs 2(c) and 2(d) display the approximated marginals for the node with maximum SKLD in the standardized scale (16). The dotted curve is the Gaussian approximation, the broken curve is the simplified Laplace and the full curve 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 approximation is 0.20

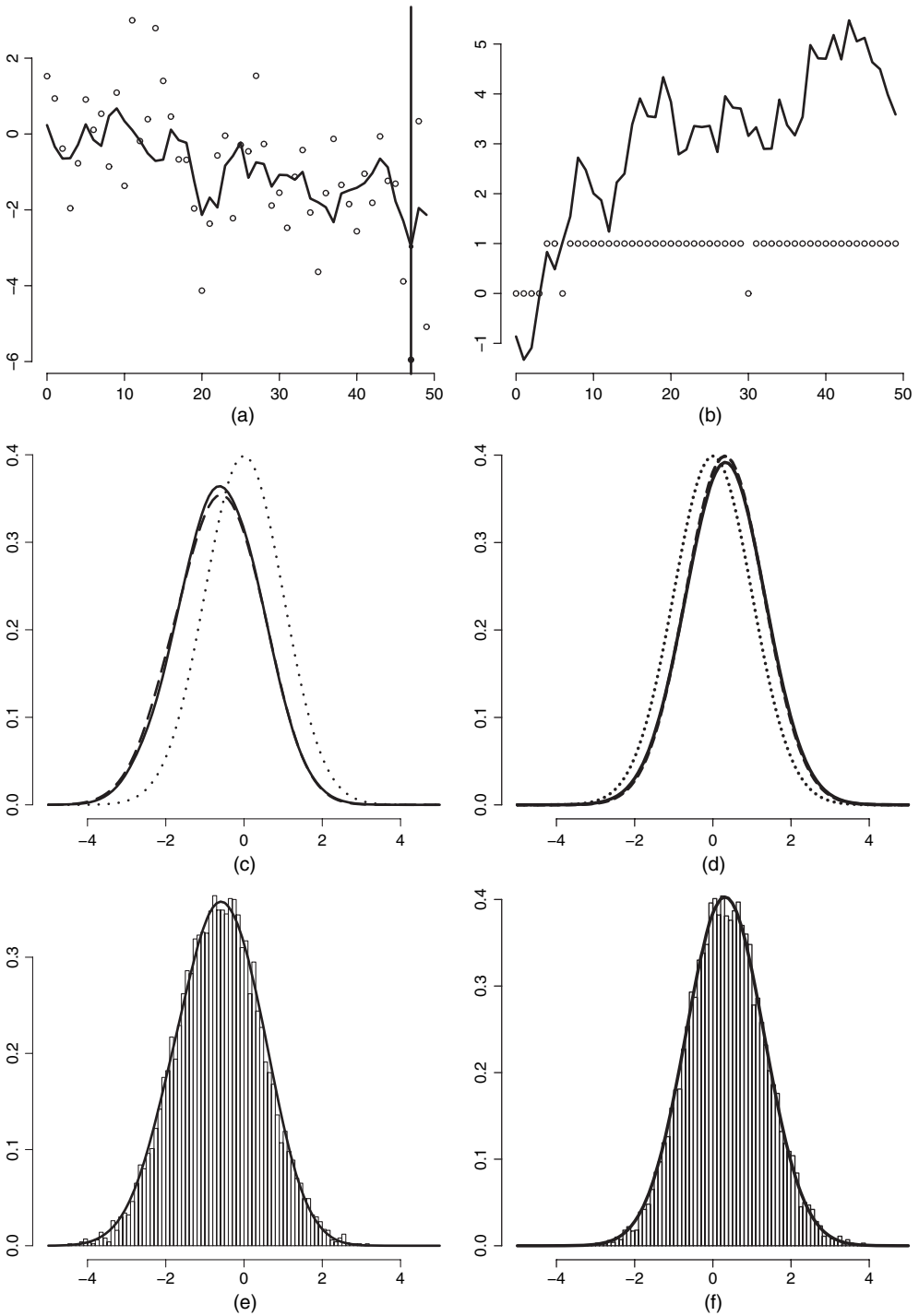


Fig. 2. (a), (b) True latent Gaussian field (—), observed Student t_3 -data and Bernoulli data (o), (c), (d) approximate marginal for a selected node by using various approximations (\cdots , Gaussian; $---$, simplified Laplace; $---$, Laplace) and (e), (f) comparison of samples from a long MCMC chain with the marginal computed with the simplified Laplace approximation

1 Fig. 2(c) and 0.05 Fig. 2(d). The SKLD between the simplified Laplace approximation and
 2 the Laplace approximation is 0.001 Fig. 2(c) and 0.0004 Fig. 2(d). Figs 2(e) and 2(f) show the
 3 simplified Laplace approximation with a histogram based on 10000 (near) independent samples
 4 from $\pi(\mathbf{f}, \mu | \boldsymbol{\theta}, \mathbf{y})$. The fit is excellent.

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

8 The results that are shown in this example are quite typical and are not limited to simple
 9 time series models like expression (25). The Laplace approximation only ‘sees’ the log-likeli-
 10 hood model and then uses some of the other nodes to compute the correction to the Gaussian
 11 approximation. Hence, the form of the log-likelihood is more important than the form of the
 12 covariance for the latent field.

13 5.2. *A generalized linear mixed model for longitudinal data*

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

17 The data come from a clinical trial of 59 epileptic patients. Each patient i is randomized to
 18 a new drug ($\text{Trt}_i = 1$) or a placebo ($\text{Trt}_i = 0$), in addition to the standard chemotherapy. The
 19 observations for each patient y_{i1}, \dots, y_{i4} , are the number of seizures during the 2 weeks before
 20 each of the four clinic visits. The covariates are age (Age), the baseline seizure counts (Base) and
 21 an indicator variable for the fourth clinic visit (V4). The linear predictor is
 22

$$23 \eta_{ij} = \beta_0 + \beta_{\text{Base}} \log(\text{Baseline}_j/4) + \beta_{\text{Trt}} \text{Trt}_j + \beta_{\text{Trt} \times \text{Base}} \text{Trt}_j \times \log(\text{Baseline}_j/4) + \beta_{\text{Age}} \text{Age}_j \\ 24 + \beta_{\text{V4}} \text{V4}_j + \varepsilon_i + \nu_{ij}, \quad i = 1, \dots, 59, \quad j = 1, \dots, 4,$$

25 using centred covariates. The observations are conditionally independent Poisson variables with
 26 mean $\exp(\eta_{ij})$. Overdispersion in the Poisson distribution is modelled by using individual ran-
 27 dom effects ε_i and subject by visit random effects ν_{ij} . We use the same priors as in the OpenBUGS
 28 manual: $\varepsilon_i \sim \text{IIDN}(0, 1/\tau_\varepsilon)$, $\nu_{ij} \sim \text{IIDN}(0, 1/\tau_\nu)$, $\tau_\varepsilon, \tau_\nu \sim \Gamma(0.001, 0.001)$, and all the β .s are as-
 29 signed $\mathcal{N}(0, 100^2)$ priors. In this example our latent field \mathbf{x} is of dimension $n = 301$ and consists
 30 of $\{\eta_{ij}\}$, $\{\varepsilon_i\}$, β_0 , β_{Base} , β_{Trt} , $\beta_{\text{Trt} \times \text{Base}}$, β_{Age} and β_{V4} . The hyperparameters are $\boldsymbol{\theta} = (\tau_\varepsilon, \tau_\nu)^T$.

31 We computed the approximate posterior marginals for the latent field by using both Gaussians
 32 and simplified Laplace approximations. The node where SKLD between these two marginals is
 33 maximum, is β_0 . The SKLD is 0.23. The two approximated marginals for β_0 are displayed in
 34 Fig. 3(a). The simplified Laplace (full curve) approximation does correct the Gaussian approx-
 35 imation (broken curve) in the mean, and the correction for skewness is minor. The simplified
 36 Laplace approximation gives accurate results, as shown in Fig. 3(a) where a histogram from a
 37 long MCMC run using OpenBUGS is overlaid. Fig. 3(b) displays the posterior marginal for τ_ε
 38 found by integrating out τ_ν from $\tilde{\pi}(\boldsymbol{\theta} | \mathbf{y})$; again, we find no errors.

39 We validated the approximations at the modal value $\boldsymbol{\theta}^*$. The effective number of parameters
 40 (24) was 121.1, which corresponds to about two samples for each parameter. A 95% interval for
 41 the remainder $r(\mathbf{x}; \boldsymbol{\theta}^*, \mathbf{y})/n_d$ is $[-0.01, 0.024]$, using 1000 independent samples. Computing the
 42 (true) Laplace approximation for the posterior marginal of β_0 gives a negligible SKLD com-
 43 pared with the simplified Laplace approximation, thus indicating that the simplified Laplace
 44 approximation is adequate. The computational cost for obtaining all the latent posterior margi-
 45 nals was about 1.5 s in total. Although OpenBUGS can provide approximate answers in minutes,
 46 we had to run it for hours to provide accurate posterior marginals.

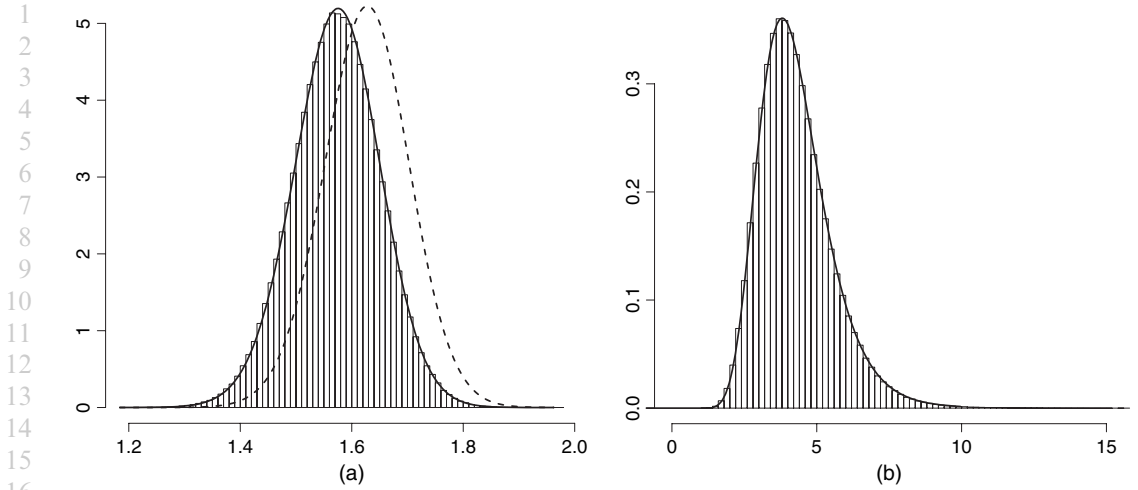


Fig. 3. Posterior marginal for (a) β_0 (—, simplified Laplace approximation; ----, Gaussian approximation) and (b) τ_ε (—, after integrating out τ_i) for the example in Section 5.2: the histograms result from a long MCMC run using OpenBUGS

5.3. Stochastic volatility models

Stochastic volatility models are frequently used to analyse financial time series. Fig. 4(a) displays the logarithm 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

$$y_t | \eta_t \sim \mathcal{N}\{0, \exp(\eta_t)\}, \quad t = 1, \dots, n_d. \quad (26)$$

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

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

and μ is a Gaussian mean value. In this example, $\mathbf{x} = (\mu, \eta_1, \dots, \eta_T)^\top$ and $\boldsymbol{\theta} = (\phi, \tau)^\top$. 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 distribution, so a Student t_ν -distribution with unknown degrees of freedom can be substituted for the Gaussian distribution in expression (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 based on only the first 50 observations in Figs 4(b)–4(d), as using the full data set makes the approximation problem easier. The full curve in Fig. 4(d) is the marginal that was found by using simplified Laplace approximations and the broken curve uses Gaussian approximations, but in this case there are little differences (the SKLD is 0.05). The histograms are constructed from the output of a long MCMC run using OpenBUGS. The approximations that were computed are very precise and no deviance (in any node) can be detected. The results that were obtained by using the full data set are similar but the marginals are narrower (not shown).

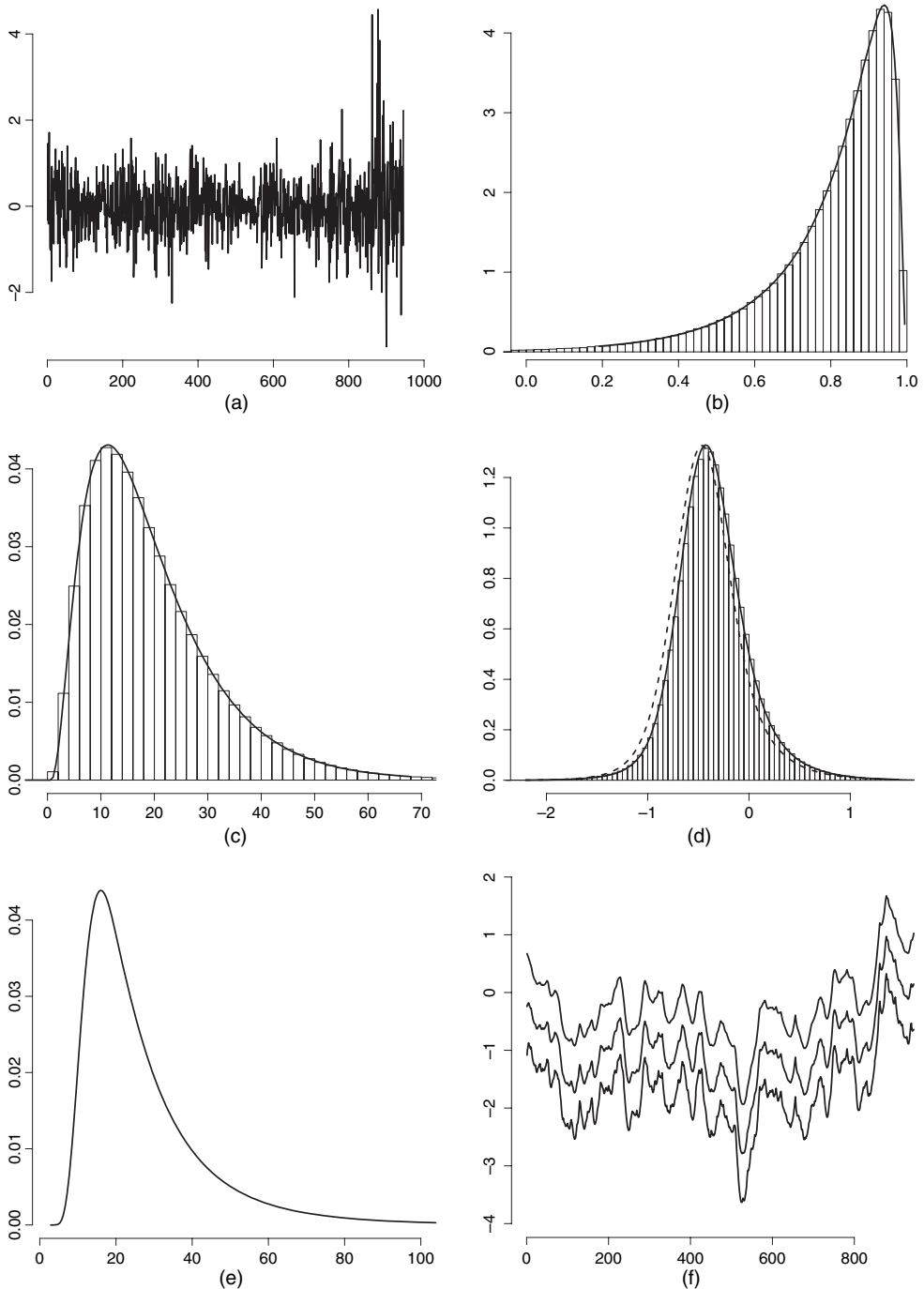


Fig. 4. (a) Log-daily-difference of the pound-dollar exchange rate from October 1st, 1981, to June 28th, 1985, (b), (c) approximated posterior marginals for ϕ and τ by using only the first $n = 50$ observations in (a) (overlaid are the histograms that were obtained from a long MCMC run using OpenBUGS),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

Following the discussion in Section 1.6, we also used this set of $n = 50$ observations to compare INLA with the EP algorithm of Zoeter *et al.* (2005) (with a slightly different parameterization of the model and other priors owing to constraints in their code). The latter was considerably less accurate (e.g. the posterior mean of ϕ is shifted 1 standard deviation to the right) and more expensive; the running time was about 40 min for the MATLAB (<http://www.mathworks.com/>) code of Zoeter *et al.* (2005) to compare with 0.3 s 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 an $\mathcal{N}(2.08, 1)$ prior to ν' where $\nu = 2 + \exp(\nu')$. The number of hyperparameters is now 3. Fig. 4(e) displays the approximate posterior marginal for the degrees of freedom and Fig. 4(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 a Student t_ν -distribution 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 approximation for both models. The average SKLD was about 0.007 in both cases.

We validated the approximations by using all the $n = 945$ observations at the modal value θ^* . The effective number of parameters (27) was about 63, which is small compared with n_d . A 95% interval for the remainder $r(\mathbf{x}; \theta^*, \mathbf{y})/n_d$ is $[-0.002, 0.004]$, using 1000 independent samples. The computational cost for obtaining all the posterior marginals (using expression (26)) for the latent field, was about 11 s.

5.4. Disease mapping of cancer incidence data

In this example we consider a spatial model for mapping cancer incidence where the stage of the disease at the time of diagnosis is known. The class of ‘disease mapping’ models is 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 $i = 1, \dots, 6990$. There are 216 districts and 15 age groups (15–19, 20–24, \dots , > 84). We follow Rue and Held (2005), Section 4.3.5, and use a logistic binary regression model:

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

where $\mathbf{f}^{(a)}$ is a smooth effect of the age group, $\mathbf{f}^{(s)}$ is a smooth spatial field and $\mathbf{f}^{(u)}$ are district random effects. More specifically, $\mathbf{f}^{(a)}$ follows an intrinsic second-order random-walk model Rue and Held (2005), chapter 3) with precision $\kappa^{(a)}$,

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

The model for the spatial term $\mathbf{f}^{(s)}$ is defined conditionally, as

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

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

The results are displayed in Fig. 5. Fig. 5(a) displays the posterior marginal for the node with the largest SKLD between the approximations by using the simplified Laplace (full curve) and Gaussian (broken curve) approximations. The SKLD is 0.058. Overlaid is the histogram that was found from a long MCMC run using the block MCMC algorithm with auxiliary variables that was described in Rue and Held (2005), Section 4.3.5; the fit is perfect. Fig. 5(b) displays the effect of the age groups, where the full curve interpolates the posterior median and the broken curves display the 0.025- and 0.975-quantiles. The quantiles that were obtained from a long MCMC run are shown by dots; again the fit is very good. Fig. 5(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 θ^* . The effective number of parameters (24) was about 101, which is small compared with n_d . A 95% interval for the remainder $r(\mathbf{x}; \theta^*, \mathbf{y})/n_d$ is $[-0.001, 0.001]$, using 1000 independent samples. The computational cost for obtaining all the posterior marginals for the latent field was about 34 s.

5.5. Log-Gaussian Cox process

Log-Gaussian Cox processes are a flexible class of models that have been successfully used for modelling spatial or spatiotemporal 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 log-Gaussian Cox process models can be analysed by using our approach for approximate inference.

A log-Gaussian Cox process is a hierarchical Poisson process: \mathbf{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 dependence in the point pattern is modelled through a common latent Gaussian variable $Z(\cdot)$. In the analysis of a log-Gaussian Cox process, it is common to discretize the observation window W . Divide W into N disjoint cells $\{w_i\}$ at $\boldsymbol{\xi}_i$ each with area $|w_i|$. Let y_i be the number of occurrences of the realized point pattern within w_i and let $\mathbf{y} = (y_1, \dots, y_N)^T$. Let η_i be the random variable $Z(\boldsymbol{\xi}_i)$. Clearly $\pi(\mathbf{y}|\boldsymbol{\eta}) = \prod_i \pi(y_i|\eta_i)$ and $y_i|\eta_i$ is Poisson distributed with mean $|w_i| \exp(\eta_i)$.

We apply model (28) to the tropical rainforest data that were studied by Waagepetersen (2007). These data come from a 50-ha 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 fifth 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 350000 individual trees species have been censused over 25 years. We shall be looking at the tree species *Beilschmiedia pendula* Lauraceae by using data that were collected from the first four census periods. The positions of the 3605 trees are displayed in Fig. 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

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

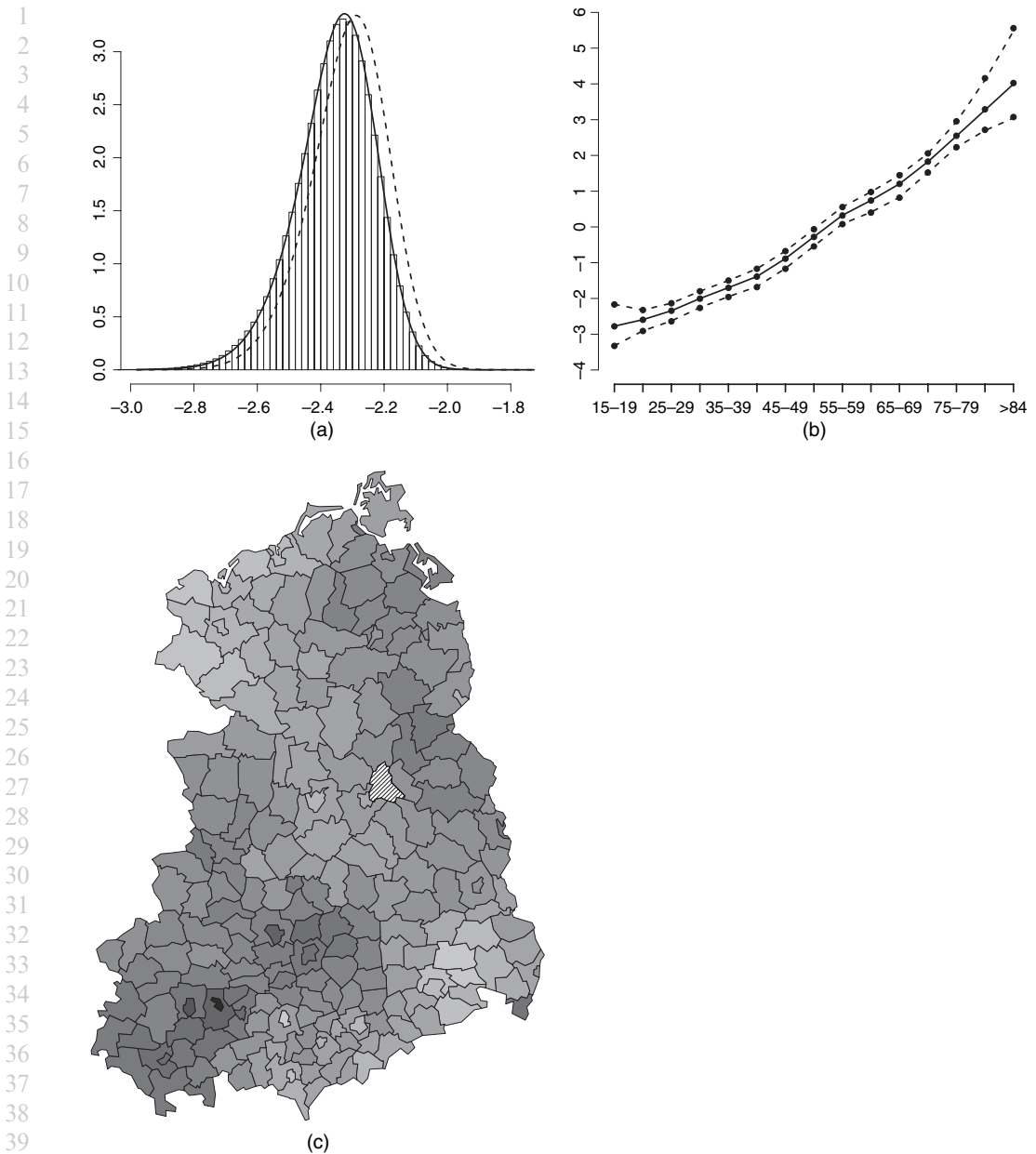


Fig. 5. Results for the cancer incidence example: (a) posterior marginal for $f_3^{(a)}$ by using simplified Laplace approximations (—), Gaussian approximations (-----) and samples from a long MCMC run (\square); (b) posterior median (—) and 0.025- and 0.975-quantiles (-----) of the age-class effect and results obtained from a long MCMC run (\bullet); (c) posterior median of the (smooth) spatial effect

where $\mathbf{f}^{(s)}$ represent the spatial component and $\mathbf{f}^{(u)}$ is an unstructured term. An alternative would be to use a semiparametric model for the effect of the covariates similar to expression (27).

We start by dividing the area of interest into a 200×100 regular lattice, where each square pixel of the lattice represent 25 m^2 . This makes $n_d = 20000$. The scaled and centred versions of the altitude and norm of the gradient, are shown in Figs 6(b) and 6(c) respectively. For the

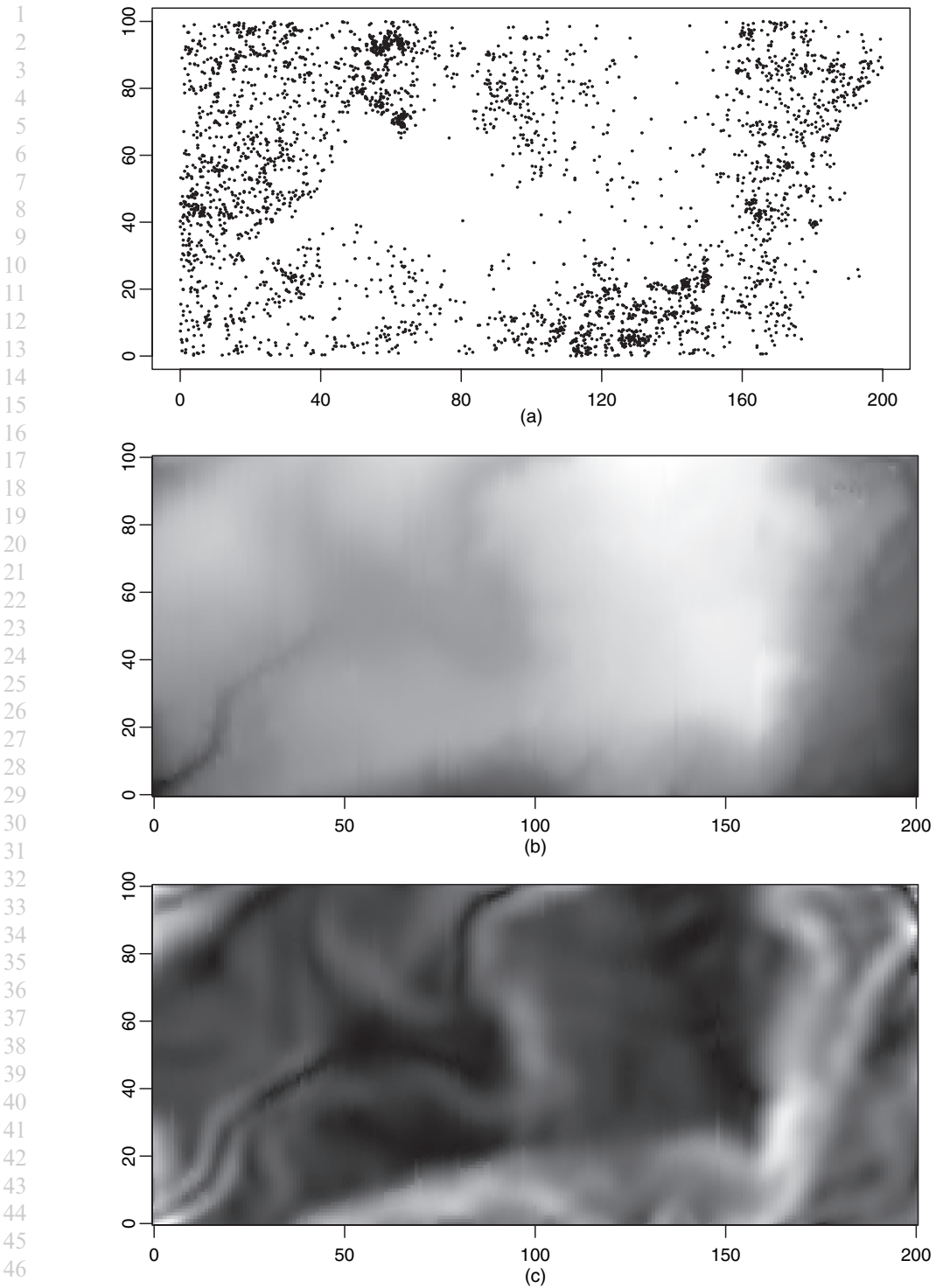


Fig. 6. Data and covariates from the log-Gaussian Cox process example: (a) locations of the 3605 trees, (b) altitude and (c) norm of the gradient

spatial structured term, we use a second-order polynomial intrinsic GMRF (see Rue and Held (2005), section 3.4.2), with following full conditionals in the interior (with obvious notation)

$$E(f_i^{(s)} | \mathbf{f}_{-i}^{(s)}, \kappa^{(s)}) = \frac{1}{20} \begin{pmatrix} \circ & \circ & \circ & \circ & \circ & & \circ & \circ & \circ & \circ & \circ & & \circ & \circ & \bullet & \circ & \circ \\ 8 & \circ & \circ & \bullet & \circ & \circ & \circ & \bullet & \circ & \bullet & \circ & & \circ & \circ & \circ & \circ & \circ \\ \circ & \bullet & \circ & \bullet & \circ & -2 & \circ & \circ & \circ & \circ & \circ & & -1 & \bullet & \circ & \circ & \circ & \bullet \\ \circ & \circ & \bullet & \circ & \circ & & \circ & \bullet & \circ & \bullet & \circ & & \circ & \circ & \circ & \circ & \circ & \circ \\ \circ & \circ & \circ & \circ & \circ & & \circ & \circ & \circ & \circ & \circ & & \circ & \circ & \bullet & \circ & \circ \end{pmatrix},$$

$$\text{Prec}(f_i^{(s)} | \mathbf{f}_{-i}^{(s)}, \kappa^{(s)}) = 20\kappa^{(s)}. \quad (29)$$

The precision $\kappa^{(s)}$ is unknown. The full conditionals are constructed to mimic the thin plate spline. There are some corrections to expression (29) near the boundary, which can be found by using the stencils in Terzopoulos (1988). We impose a sum-to-zero constraint on the spatial term due to β_0 . The unstructured terms $\mathbf{f}^{(u)}$ are independent $\mathcal{N}(0, \kappa^{(u)})$, vague $\Gamma(1.0, 0.001)$ priors are assigned to $\kappa^{(s)}$ and $\kappa^{(u)}$, and independent $\mathcal{N}(0, 10^3)$ priors to $\beta_0, \beta_{\text{Alt}}$ and β_{Grad} . The latent field is $\mathbf{x} = (\boldsymbol{\eta}^T, (\mathbf{f}^{(s)})^T, \beta_0, \beta_{\text{Alt}}, \beta_{\text{Grad}})^T$ with dimension 40003, and $\boldsymbol{\theta} = (\log(\kappa^{(s)}), \log(\kappa^{(u)}))$ with dimension 2.

We computed the approximation for 20003 posterior marginals $\mathbf{f}^{(s)}, \beta_0, \beta_{\text{Alt}}$ and β_{Grad} , using the simplified Laplace approximation. The results are displayed in Fig. 7. Fig. 7(a) displays the estimated posterior mean of the spatial component, where we have indicated by using contours those nodes where the SKLD between the marginal computed with the Gaussian approximation and that computed with the simplified Laplace approximation exceeds 0.25. These nodes are potential candidates for further investigation, so we computed their posteriors by using the Laplace approximation also; the results agreed well with those obtained from the simplified Laplace approximation. As an example, we display in Fig. 7(b) the marginals for the ‘worst case’ which is node (61, 73) with an SKLD of 0.50: Gaussian (broken curve) simplified Laplace (full curve) and Laplace approximations (chain curve). Note that the approximations becomes more close as we improve the approximations. Figs 7(c)–7(e) display the posterior marginals computed with the Gaussian approximations (broken curve) and that computed with the simplified Laplace approximations (full curve) for $\beta_0, \beta_{\text{Alt}}$ and β_{Grad} . The difference is mostly due to a horizontal shift, a characteristic that is valid for all the other nodes for this example.

This task required about 4 h of computing owing to the high dimension and the number of computed posterior marginals. The total cost can be reduced to about 10 min if only using the Gaussian approximation (4). To validate the approximations, we computed $p_D(\boldsymbol{\theta}^*) \approx 1714$ and estimated a 95% interval for the remainder $r(\mathbf{x}; \boldsymbol{\theta}^*, \mathbf{y})/n_d$ as [0.004, 0.01], using 1000 independent samples. Varying $\boldsymbol{\theta}$ gave similar results. There are no indications that the approximations does not work well in this case. Owing 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 that was described in Rue and Held (2005), section 4.4.2.

6. Extensions

Although 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 \mathbf{x}_S

A natural extension is to consider not only posterior marginals for x_i but also for a subset $\mathbf{x}_S = \{x_i : i \in S\}$. S can be small, say from 2 to 5, but sometimes larger sets are required. Although

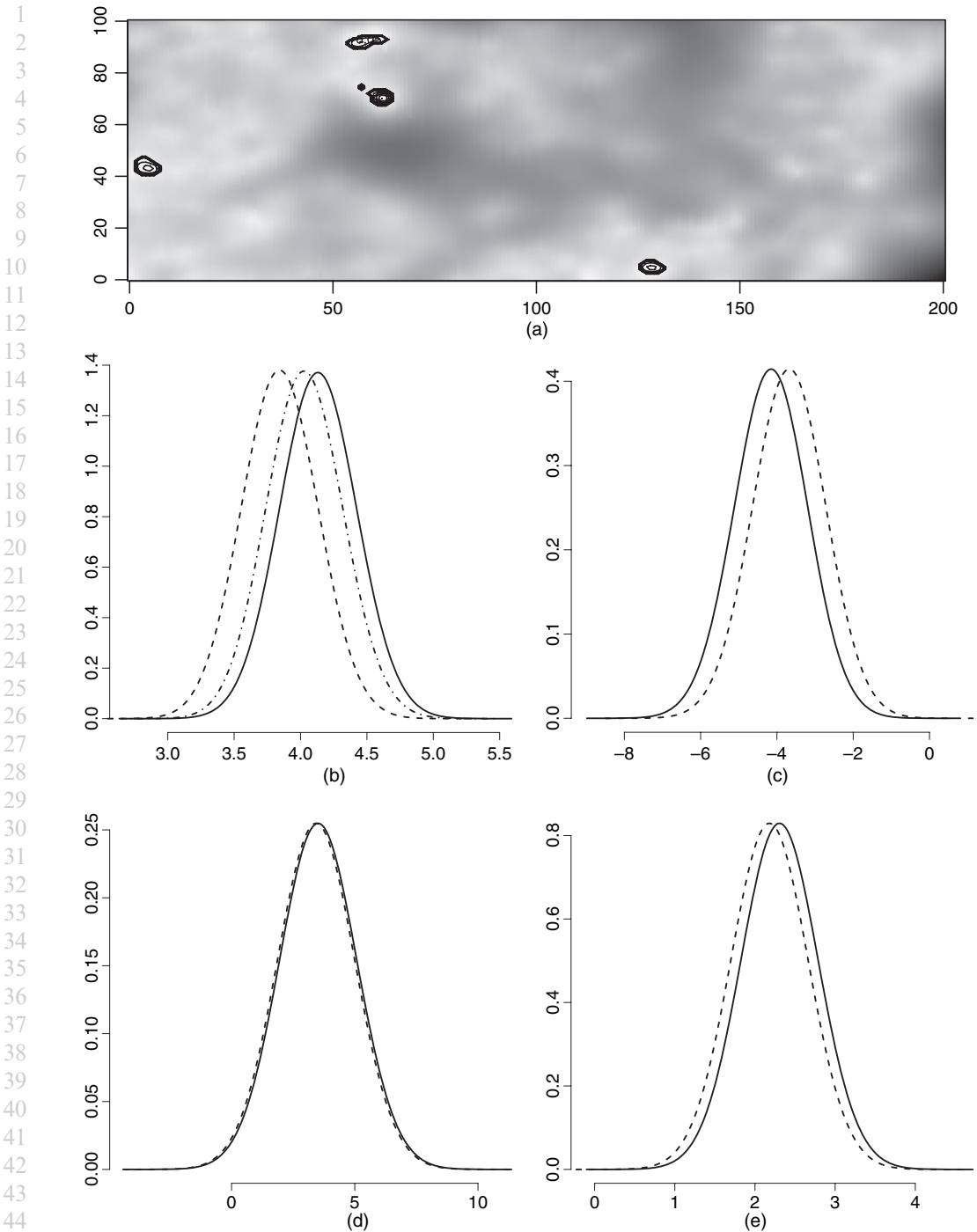


Fig. 7. Log-Gaussian Cox process example: (a) posterior mean of the spatial component with contour indicating an SKLD above 0.25, (b) marginals for node (61,73) in the spatial component with maximum SKLD 0.50, Gaussian (-----), simplified Laplace (—) and Laplace approximations (-·-·-) and (c)–(e) posterior marginals of β_0 , β_{Alt} and β_{Grad} by using the simplified Laplace (—) and Gaussian approximations (-----)

the Laplace approximation (12) can still be applied, replacing x_i with \mathbf{x}_S , and \mathbf{x}_{-i} with \mathbf{x}_{-S} , the practicalities become 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 $x_i|\boldsymbol{\theta}, \mathbf{y}$ and $x_j|\boldsymbol{\theta}, \mathbf{y}$. From the Gaussian approximation $\tilde{\pi}_G(\mathbf{x}|\boldsymbol{\theta}, \mathbf{y})$ we know the Gaussian marginal distribution for $x_i, x_j|\boldsymbol{\theta}, \mathbf{y}$. 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 approximating $x_i, x_j|\boldsymbol{\theta}, \mathbf{y}$ by 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 expression (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 $\pi(\mathbf{y})$ 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 expression (3) that the natural approximation to the marginal likelihood is the normalizing constant of $\tilde{\pi}(\boldsymbol{\theta}|\mathbf{y})$,

$$\tilde{\pi}(\mathbf{y}) = \int \frac{\pi(\boldsymbol{\theta}, \mathbf{x}, \mathbf{y})}{\tilde{\pi}_G(\mathbf{x}|\boldsymbol{\theta}, \mathbf{y})} \Big|_{\mathbf{x}=\mathbf{x}^*(\boldsymbol{\theta})} d\boldsymbol{\theta}. \tag{30}$$

where $\pi(\boldsymbol{\theta}, \mathbf{x}, \mathbf{y}) = \pi(\boldsymbol{\theta})\pi(\mathbf{x}|\boldsymbol{\theta})\pi(\mathbf{y}|\mathbf{x}, \boldsymbol{\theta})$. An alternative, cruder, estimate of the marginal likelihood is obtained by assuming that $\boldsymbol{\theta}|\mathbf{y}$ is Gaussian; then equation (30) turns into some known constant times $|\mathbf{H}|^{-1/2}$, where \mathbf{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 $\tilde{\pi}(\boldsymbol{\theta}|\mathbf{y})$ in a ‘non-parametric’ way. This allows for taking into account the departure from Gaussianity which, for instance, appears clearly in Fig. 4. Friel and Rue (2007) used a similar expression to formula (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 expression (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 if the posterior marginal $\pi(\boldsymbol{\theta}|\mathbf{y})$ is multimodal (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 generate 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. $\pi(y_i|y_{-i})$. We now explain how to approximate this quantity simply, without reanalysing the model. First, note that removing y_i from the data set affects the marginals of x_i and $\boldsymbol{\theta}$ as follows:

$$\pi(x_i|\mathbf{y}_{-i}, \boldsymbol{\theta}) \propto \frac{\pi(x_i|\mathbf{y}, \boldsymbol{\theta})}{\pi(y_i|x_i, \boldsymbol{\theta})},$$

$$\pi(\boldsymbol{\theta}|\mathbf{y}_{-i}) \propto \frac{\pi(\boldsymbol{\theta}|\mathbf{y})}{\pi(y_i|\mathbf{y}_{-i}, \boldsymbol{\theta})}$$

where a one-dimensional integral is required to compute

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

The effect of $\boldsymbol{\theta}$ can then be integrated out from $\pi(y_i|\mathbf{y}_{-i}, \boldsymbol{\theta})$, in the same way as equation (5). Unusually small values of $\pi(y_i|\mathbf{y}_{-i})$ indicate surprising observations, but what is meant by ‘small’ must be calibrated with the level of x_i . Pettit (1990) suggested calibrating with the maximum value of $\pi(\cdot|\mathbf{y}_{-i})$, but an alternative is to compute the probability integral transform $\text{PIT}_i = \text{Prob}(y_i^{\text{new}} \leq y_i|\mathbf{y}_{-i})$ by using the same device as above. (See also Gneiting and Raftery (2007) for 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 distribution, 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 1 for $i = 331, 651, 862$. These observations are less surprising under the Student t_ν observation model: i.e. the same PIT_i s are then about $(1 - 5) \times 10^{-4}$.

6.4. Deviance information criteria

The DIC (Spiegelhalter *et al.*, 2002) is a popular information criterion that was 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 approximation (24). In our context, the deviance is

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

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 expression (24) is a good approximation. The mean of the deviance can be computed in two steps: first, compute the conditional mean conditioned on $\boldsymbol{\theta}$ by using univariate numerical integration for each $i \in \mathcal{I}$; second, integrate out $\boldsymbol{\theta}$ with respect to $\pi(\boldsymbol{\theta}|\mathbf{y})$. 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 Fig. 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 6–12. Using, for example, $\delta_z = 1$ and $\delta_\pi = 2.5$, the integration scheme that was proposed in Section 3.1 will require, if $\boldsymbol{\theta}|\mathbf{y}$ is Gaussian, $\mathcal{O}(5^m)$ evaluation points. Even if we restrict ourselves to three eval-

uation 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 $\tilde{\pi}(\boldsymbol{\theta}|\mathbf{y})$. 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 more difficult, often increasing the number of hyperparameters increases also the variability and the regularity, and makes the integrand increasingly Gaussian. Meaningful results can be obtained even using an extreme choice akin to empirical Bayes, i.e. using only the modal configuration to integrate over $\pi(\boldsymbol{\theta}|\mathbf{y})$. This ‘plug-in’ approach will obviously underestimate variability, but it will provide reasonable results provided that the uncertainty in 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 lay out some ‘points’ in an m -dimensional space. On the basis of 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

$$\begin{array}{cccc} (1, 1, 1, 1, 1), & (-1, 1, 1, 1, -1), & (1, -1, 1, 1, -1), & (-1, -1, 1, 1, 1), \\ (1, 1, -1, 1, -1), & (-1, 1, -1, 1, 1), & (1, -1, -1, 1, 1), & (-1, -1, -1, 1, -1), \\ (1, 1, 1, -1, -1), & (-1, 1, 1, -1, 1), & (1, -1, 1, -1, 1), & (-1, -1, 1, -1, -1), \\ (1, 1, -1, -1, 1), & (-1, 1, -1, -1, -1), & (1, -1, -1, -1, -1), & (-1, -1, -1, -1, 1). \end{array}$$

They are all on the surface of the m -dimensional sphere with radius \sqrt{m} . The star points consist of $2m$ points along each axis at distance $\pm\sqrt{m}$ and the central point at the origin. For $m = 5$ this makes $n_p = 27$ points in total, which is small compared with $5^5 = 3125$ or $3^5 = 243$. The number of design points is 8 for $m = 3$, 16 for $m = 4$ and $m = 5$, 32 for $m = 6$, 64 for $m = 7$ and $m = 8$, 128 for $m = 9, 10, 11$, and 256 from $m = 12 - 17$; 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 equation (5) and the scaling of the points, assume for simplicity that $\boldsymbol{\theta}|\mathbf{y}$ is standard Gaussian. We require that the integral of 1 equals 1, and that the integral of $\boldsymbol{\theta}^T \boldsymbol{\theta}$ equals m . This gives the integration weight for the points on the sphere with radius $f_0 \sqrt{m}$

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

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 that 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 $\boldsymbol{\theta}$ must be done more accurately, but these can be detected by comparing the results that are obtained by using the empirical Bayes and the CCD approach. For these cases, the CCD integration seems to provide results that are halfway 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 INLAs. The results that were 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 can 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 Beskos *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 that are 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. one dimensional, two dimensional, three dimensional and so on. All the examples that were considered in this paper were computed by 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 by using this program. `inla` is built on 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 $x_i|\theta, \mathbf{y}$ independently for all i for fixed θ . Both the `inla` program and `GMRFLib` use the OpenMP API (see <http://www.openmp.org>) to speed up the computations for shared memory machines (i.e. multicore processors); however, we have not focused on these computational issues and speed-ups in this paper. Parallel computing is particularly important for spatial or spatiotemporal latent Gaussian models, but also smaller models enjoy good speed-ups.

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 than it appears at first glance: the CCD approach seems promising and provides reasonable results when m is not small, in the case where the user does 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 an MCMC algorithm, notably if he or she must analyse many different data sets 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 validate assumptions easily like a linear or smooth effect of a covariate is important, and our approach also gives access to Bayes factors, various predictive measures and the DIC, which are useful tools to compare models and to challenge the model under study.

Acknowledgements

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. Haslett, 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.

Appendix A: Variational Bayes methods for latent Gaussian models: an example

We consider a simple latent Gaussian model that is defined by

$$\begin{aligned}\theta &\sim \Gamma(a, b), \\ \mathbf{x}|\theta &\sim \mathcal{N}\left(\mathbf{0}, \frac{1}{\theta} \mathbf{R}^{-1}\right), \\ \mathbf{y}|\mathbf{x}, \theta &\sim \mathcal{N}\left(\mathbf{x}, \frac{1}{\kappa} \mathbf{I}\right)\end{aligned}$$

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

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

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

$$\begin{aligned}q_{\mathbf{x}}^{(t+1)}(\mathbf{x}) &\propto \exp[E_{q_{\theta}^{(t)}} \log\{\pi(\mathbf{x}, \mathbf{y}|\theta)\}], \\ q_{\theta}^{(t+1)}(\theta) &\propto \pi(\theta) \exp[E_{q_{\mathbf{x}}^{(t)}} \log\{\pi(\mathbf{x}, \mathbf{y}|\theta)\}].\end{aligned}$$

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

$$b_{\infty} = b + \kappa^2 \mathbf{y}^T \left(\frac{a+n/2}{b_{\infty}} \mathbf{R} + \kappa \mathbf{I}\right)^{-1} \mathbf{R} \left(\frac{a+n/2}{b_{\infty}} \mathbf{R} + \kappa \mathbf{I}\right)^{-1} \mathbf{y} + \text{tr}\left\{\left(\frac{a+n/2}{b_{\infty}} \mathbf{I} + \kappa \mathbf{R}^{-1}\right)^{-1}\right\},$$

which is not tractable. However, when $\kappa \rightarrow 0$, this transforms into $b_{\infty} = b + nb_{\infty}/\{2(a+n/2)\}$; hence $\lim_{\kappa \rightarrow 0} (b_{\infty}) = b(a+n/2)/a$. 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 that is estimated by VB methods is significantly smaller than the true posterior variance of θ . For non-Gaussian data we obtained similar empirical results.

Appendix B: Fitting the skew normal distribution

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

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

To second order, equation (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 that the third derivative at the mode in distribution

- 1 Diggle, P. J. and Ribeiro, P. J. (2006) *Model-based Geostatistics*. New York: Springer.
- 2 Durbin, J. and Koopman, S. J. (2000) Time series analysis of non-Gaussian observations based on state space
3 models from both classical and Bayesian perspectives (with discussion). *J. R. Statist. Soc. B*, **62**, 3–56.
- 4 Eidsvik, J., Martino, S. and Rue, H. (2008) Approximate Bayesian inference in spatial generalized linear mixed
5 models. *Scand. J. Statist.*, to be published.
- 6 Fahrmeir, L. and Lang, S. (2001) Bayesian inference for generalized additive mixed models based on Markov
7 random field priors. *Appl. Statist.*, **50**, 201–220.
- 8 Fahrmeir, L. and Tutz, G. (2001) *Multivariate Statistical Modelling based on Generalized Linear Models*, 2nd edn.
9 Berlin: Springer.
- 10 Finkenstadt, B., Held, L. and Isham, V. (eds) (2006) *Statistical Methods for Spatio-temporal Systems*. Boca Raton:
11 Chapman and Hall–CRC.
- 12 Friel, N. and Rue, H. (2007) Recursive computing and simulation-free inference for general factorizable models.
13 *Biometrika*, **94**, 661–672.
- 14 Frühwirth-Schnatter, S. and Frühwirth, R. (2007) Auxiliary mixture sampling with applications to logistic models.
15 *Computnl Statist. Data Anal.*, **51**, 3509–3528.
- 16 Frühwirth-Schnatter, S. and Wagner, H. (2006) Auxiliary mixture sampling for Parameter-driven models of time
17 series of small counts with applications to state space modelling. *Biometrika*, **93**, 827–841.
- 18 Gamerman, D. (1997) Sampling from the posterior distribution in generalized linear mixed models. *Statist.*
19 *Comput.*, **7**, 57–68.
- 20 Gamerman, D. (1998) Markov chain Monte Carlo for dynamic generalised linear models. *Biometrika*, **85**, 215–227.
- 21 Gelfand, A. E. (1996) Model determination using sampling-based methods. In *Markov Chain Monte Carlo in*
22 *Practice* (eds W. R. Gilks, S. Richardson and D. J. Spiegelhalter), pp. 145–161. London: Chapman and Hall.
- 23 Gelman, A., Carlin, J. B., Stern, H. S. and Rubin, D. B. (2004) *Bayesian Data Analysis*, 2nd edn. Boca Raton:
24 Chapman and Hall–CRC.
- 25 Gneiting, T. (2002) Nonseparable, stationary covariance functions for space-time data. *J. Am. Statist. Ass.*, **97**,
26 590–600.
- 27 Gneiting, T. and Raftery, A. E. (2007) Strictly proper scoring rules, prediction, and estimation. *J. Am. Statist.*
28 *Ass.*, **102**, 359–378.
- 29 Gschlößl, S. and Czado, C. (2007) Modelling count data with overdispersion and spatial effects. *Statist. Pap.*
30 (Available from <http://dx.doi.org/10.1007/s00362-006-0031-6>.)
- 31 Held, L., Natario, I., Fenton, S., Rue, H. and Becker, N. (2005) Towards joint disease mapping. *Statist. Meth.*
32 *Med. Res.*, **14**, 61–82.
- 33 Hinton, G. E. and van Camp, D. (1993) Keeping the neural networks simple by minimizing the description length
34 of the weights. In *Proc. 6th A. Conf. Computational Learning Theory*, pp. 5–13.
- 35 Holmes, C. C. and Held, L. (2006) Bayesian auxiliary variable models for binary and multinomial regression.
36 *Bayes. Anal.*, **1**, 145–168.
- 37 Hsiao, C. K., Huang, S. Y. and Chang, C. W. (2004) Bayesian marginal inference via candidate's formula. *Statist.*
38 *Comput.*, **14**, 59–66.
- 39 Humphreys, K. and Titterton, D. M. (2000) Approximate Bayesian inference for simple mixtures. In *Compstat:*
40 *Proc. 14th Symp. Computational Statistics, Utrecht*. Physica.
- 41 Jordan, M. I. (2004) Graphical models. *Statist. Sci.*, **19**, 140–155.
- 42 Kamman, E. E. and Wand, M. P. (2003) Geoadditive models. *Appl. Statist.*, **52**, 1–18.
- 43 Kass, R. E., Tierney, L. and Kadane, J. B. (1999) the validity of posterior expansions based on Laplace's method. In
44 *Essays in Honor of George Bernard* (eds S. Geisser, J. S. Hodges, S. J. Press and A. Z.), pp. 473–488. Amsterdam:
45 North-Holland.
- 46 Kass, R. E. and Vaidyanathan, S. K. (1992) Approximate Bayes factors and orthogonal parameters, with appli-
47 cation to testing equality of two binomial proportions. *J. R. Statist. Soc. B*, **54**, 129–144.
- 48 Kitagawa, G. and Gersch, W. (1996) Smoothness priors analysis of time series. *Lect. Notes Statist.*, **116**.
- Knorr-Held, L. (1999) Conditional prior proposals in dynamic models. *Scand. J. Statist.*, **26**, 129–144.
- Knorr-Held, L., Raßer, G. and Becker, N. (2002) Disease mapping of stage-specific cancer incidence data. *Bio-*
metrics, **58**, 492–501.
- Knorr-Held, L. and Rue, H. (2002) On block updating in Markov random field models for disease mapping.
Scand. J. Scient. Statist., **29**, 597–614.
- Kohn, R. and Ansley, C. F. (1987) A new algorithm for spline smoothing based on smoothing a stochastic process.
SIAM J. Scient. Statist. Comput., **8**, 33–48.
- Kuss, M. and Rasmussen, C. E. (2005) Assessing approximate inference for binary Gaussian process classification.
J. Mach. Learn. Res., **6**, 1679–1704.
- Lang S. and Brezger A. (2004) Bayesian P-splines. *J. Computnl Graph. Statist.*, **13**.
- Mackay, D. J. C. (1995) Ensemble learning and evidence maximization. In *Proc. NIPS Conf.*
- Mackay, D. J. C. (1997) Ensemble learning for hidden Markov models. *Technical Report*. Cavendish Laboratory,
University of Cambridge, Cambridge.
- Marroquin, J. L., Velasco, F. A., Rivera, M. and Nakamura, M. (2001) Gauss-Markov measure field models for
low-level vision. *IEEE Trans. Pattn Anal. Mach. Intell.*, **23**, 337–348.

- 1 Martino, S. (2007) Approximate Bayesian inference for latent Gaussian models. *PhD Thesis*. Department of
 2 Mathematical Sciences, Norwegian University of Science and Technology, Trondheim.
- 3 Martino, S. and Rue, H. (2008) Implementing approximate Bayesian inference for latent Gaussian models using
 4 integrated nested Laplace approximations: a manual for the `inla`-program. *Technical Report 2*. Department
 5 of Mathematical Sciences, Norwegian University of Science and Technology, Trondheim.
- 6 Minka, T. P. (2001) Expectation propagation for approximate Bayesian inference. *Uncertainty Artif. Intell.*, **17**,
 7 362–369.
- 8 Møller, J., Syversveen, A. R. and Waagepetersen, R. P. (1998) Log Gaussian Cox processes. *Scand. J. Statist.*, **25**,
 9 451–482.
- 10 Møller, J. and Waagepetersen, R. (2003) *Statistical Inference and Simulation for Spatial Point Processes*. London:
 11 Chapman and Hall.
- 12 Neal, R. M. (1998) Regression and classification using Gaussian Process priors. In *Bayesian Statistics 6*,
 13 pp. 475–501. New York: Oxford University Press.
- 14 O’Hagan, A. (1978) Curve fitting and optimal design for prediction (with discussion). *J. R. Statist. Soc. B*, **40**,
 15 1–42.
- 16 Papaspiliopoulos, O., Roberts, G. O. and Sköld, M. (2007) A general framework for the parameterization of
 17 hierarchical models. *Statist. Sci.*, **22**, 59–73.
- 18 Pettit, L. I. (1990) The conditional predictive ordinate for the normal distribution. *J. R. Statist. Soc. B*, **52**,
 19 175–184.
- 20 Rasmussen, C. E. and Williams, C. K. I. (2006) *Gaussian Processes for Machine Learning*. Cambridge: MIT Press.
- 21 R Development Core Team (2007) *R: a Language and Environment for Statistical Computing*. Vienna: R Founda-
 22 tion for Statistical Computing.
- 23 Rellier G., Descombes, X., Zerubia, J. and Falzon, F. (2002) A Gauss-Markov model for hyperspectral texture
 24 analysis of urban areas. In *Proc. 16th Int. Conf. Pattern Recognition*, pp. 692–695.
- 25 Robert, C. P. and Casella, G. (1999) *Monte Carlo Statistical Methods*. New York: Springer.
- 26 Rue, H. (2001) Fast sampling of Gaussian Markov random fields. *J. R. Statist. Soc. B*, **63**, 325–338.
- 27 Rue, H. and Held, L. (2005) *Gaussian Markov Random Fields: Theory and Applications*. London: Chapman and
 28 Hall.
- 29 Rue, H. and Martino, S. (2007) Approximate Bayesian inference for hierarchical Gaussian Markov random fields
 30 models. *J. Statist. Planning Inf.*, **137**, 3177–3192.
- 31 Rue, H., Steinsland, I. and Erland, S. (2004) Approximating hidden Gaussian Markov random fields. *J. R. Statist.*
 32 *Soc. B*, **66**, 877–892.
- 33 Sanchez, S. M. and Sanchez, P. J. (2005) Very large fractional factorials and central composite designs. *ACM*
 34 *Trans. Modelling Comput. Simuln.*, **15**, 362–377.
- 35 Schervish, M. J. (1995) *Theory of Statistics*, 2nd edn. New York: Springer.
- 36 Shephard, N. (1994) Partial non-Gaussian state space. *Biometrika*, **81**, 115–131.
- 37 Shephard, N. and Pitt, M. K. (1997) Likelihood analysis of non-Gaussian measurement time series. *Biometrika*,
 38 **84**, 653–667.
- 39 Shun, Z. and McCullagh, P. (1995) Laplace approximation of high dimensional integrals. *J. R. Statist. Soc. B*, **57**,
 40 749–760.
- 41 Smith, A. F. M., Skene, A. M., Shaw, J. E. H. and Naylor, J. C. (1987) Progress with numerical and graphical
 42 methods for practical Bayesian statistics. *Statistician*, **36**, 75–82.
- 43 Spiegelhalter, D. J., Best, N. G., Carlin, B. P. and van der Linde, A. (2002) Bayesian measures of model complexity
 44 and fit (with discussion). *J. R. Statist. Soc. B*, **64**, 583–639.
- 45 Terzopoulos, D. (1988) The computation of visible-surface representations. *IEEE Trans. Pattn Anal. Mach. Intell.*,
 46 **10**, 417–438.
- 47 Thall, P. F. Vail, S. C. (1990) Some covariance models for longitudinal count data with overdispersion. *Biometrics*,
 48 **46**, 657–671.
- Thomas, A., O’Hara, B., Ligges, U. and Sturtz, S. (2006) Making BUGS Open. *R News*, **6**, 12–16.
- Tierney, L. and Kadane, J. B. (1986) Accurate approximations for posterior moments and marginal densities.
J. Am. Statist. Ass., **81**, 82–86.
- Titterton, D. M. (2004) Bayesian methods for neural networks and related models. *Statist. Sci.*, **19**, 128–139.
- Waagepetersen, R. P. (2007) An estimating function approach to inference for inhomogeneous Neyman-Scott
 processes. *Biometrics*, **63**, 252–258.
- Wahba, G. (1978) Improper priors, spline smoothing and the problem of guarding against model errors in regres-
 sion. *J. R. Statist. Soc. B*, **40**, 364–372.
- Wakefield, J. (2007) Disease mapping and spatial regression with count data. *Biostatistics*, **8**, 158–183.
- Wakefield, J. C., Best, N. G. and Waller, L. A. (2000) Bayesian approaches to disease mapping. In *Spatial Epi-
 demiology: Methods and Applications* (eds P. Elliot, J. C. Wakefield, N. G. Best and D. J. Briggs) pp. 104–107.
 Oxford: Oxford University Press.
- Wang, B. and Titterton, D. M. (2005) Inadequacy of interval estimates corresponding to variational Bayesian
 approximations. In *Proc. 10th Int. Wrkshp Artificial Intelligence and Statistics* (eds R. G. Cowell and
 Z. Ghahramani), pp. 373–380.

- 1 Wang, B. and Titterton, D. M. (2006) Convergence properties of a general algorithm for calculating variational
2 Bayesian estimates for a normal mixture model. *Bayes. Anal.*, **1**, 625–650.
- 3 Wecker, W. E. and Ansley, C. F. (1983) The signal extraction approach to nonlinear regression and spline smooth-
4 ing. *J. Am. Statist. Ass.*, **78**, 81–89.
- 5 Weir, I. S. and Pettitt, A. N. (2000) Binary probability maps using a hidden conditional autoregressive Gaussian
6 process with an application to Finnish common toad data. *Appl. Statist.*, **49**, 473–484.
- 7 West, M. and Harrison, J. (1997) *Bayesian Forecasting and Dynamic Models*, 2nd edn. New York: Springer.
- 8 Williams, C. K. I. and Barber, D. (1998) Bayesian classification with Gaussian processes. *IEEE Trans. Pattn Anal.*
9 *Mach. Intell.*, **20**, 1342–1351.
- 10 Zoeter, O., Heskes, T. and Kappen, B. (2005) Gaussian quadrature based expectation propagation. In *Proc.*
11 *AISTATS* (eds Z. Ghahramani and R. Cowell).
- 12
- 13
- 14
- 15
- 16
- 17
- 18
- 19
- 20
- 21
- 22
- 23
- 24
- 25
- 26
- 27
- 28
- 29
- 30
- 31
- 32
- 33
- 34
- 35
- 36
- 37
- 38
- 39
- 40
- 41
- 42
- 43
- 44
- 45
- 46
- 47
- 48