Journal of the Royal Statistical Society
SERIES B Statistical Methodology

J. R. Statist. Soc. B (2009)
71, *Part* 2, *pp.* 319–392

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 geoadditive 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

 1369–7412/09/71319

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

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

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

The main aim of this paper is twofold:

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

1.2. Latent Gaussian models: applications

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

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

(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; Gschløssl and Czado, 2008; 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 (Kammann 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 $\boldsymbol{\theta} = (\boldsymbol{\theta}_1^{\mathrm{T}}, \boldsymbol{\theta}_2^{\mathrm{T}})^{\mathrm{T}}$ with $\dim(\boldsymbol{\theta}) = m$. The posterior then reads (for a non-singular $\mathbf{Q}(\boldsymbol{\theta})$)

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

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 $m \leqslant 6$. Both properties are usually required to produce fast inference, but exceptions exist (Eidsvik *et al.*, 2009).

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 $\mathbf{x}$ (Gamerman, 1997, 1998; Carter and Kohn, 1994; Knorr-Held, 1999; Knorr-Held and Rue, 2002; Rue *et al.*, 2004). The second problem requires, at least partially, a joint update of both $\boldsymbol{\theta}$ and $\mathbf{x}$. One suggestion is to 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 what 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

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

1.6. Approximation methods in machine learning

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

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

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

Another approximation scheme that is popular in machine learning is the expectation–propagation (EP) approach (Minka, 2001); see for example Zoeter and Heskes (2005) and Kuss and Rasmussen (2005) for applications of EP to latent Gaussian models. EP follows principles which are quite similar to VB methods, i.e. it minimizes iteratively some pseudodistance between $\pi(\mathbf{x}, \boldsymbol{\theta}|\mathbf{y})$ and 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 parametric factors, each involving a single component of $(\mathbf{x}, \boldsymbol{\theta})$. However, the pseudodistance 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 other way around (as in VB methods). Because of this, EP usually overestimates the posterior variance (Bishop (2006), chapter 10). Kuss and Rasmussen (2005) derived an EP approximation scheme for classification problems involving Gaussian processes that seems to be accurate and 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 clear how to extend this approach to a fully Bayesian analysis. More importantly, deriving an efficient EP algorithm seems to require specific efforts for each class of models. With respect to computational cost, VB and EP methods are both designed to be faster than exact MCMC methods, but, owing to their iterative nature, they are (much)

slower than analytic approximations (such as those developed in this paper); see Section 5.3 for an illustration of this in one of our examples. Also, it is not clear whether EP and VB 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

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

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

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

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 $\boldsymbol{\theta}$:

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

where $\tilde{\pi}_{\mathrm{G}}(\mathbf{x}|\boldsymbol{\theta}, \mathbf{y})$ 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 $\tilde{\pi}_{\mathrm{G}}(\mathbf{x}|\boldsymbol{\theta}, \mathbf{y})$ and to

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

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

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

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

The sum is over values of $\boldsymbol{\theta}$ with area weights Δ_k. Rue and Martino (2007) showed that the approximate posterior marginals for $\boldsymbol{\theta}$ were accurate, 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_is 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 ith element and by $\Gamma(\tau; a, b)$ the $\Gamma(a, b)$ density (with mean a/b) at point τ.

2.1. Gaussian Markov random fields

A GMRF is a Gaussian random variable $\mathbf{x} = (x_1, \ldots, x_n)$ with Markov properties: for some $i \neq j$s, 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

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

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 $\mathbf{L}\mathbf{L}^{\mathrm{T}}$, 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 $F(i, j) = \{i+1, \ldots, j-1, j+1, \ldots, n\}$ 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 $\mathbf{L}\mathbf{L}^{\mathrm{T}}$ 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 $\mathbf{L}^{\mathrm{T}}\mathbf{x} = \mathbf{v}$. If $\mathbf{z} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ then the solution of $\mathbf{L}^{\mathrm{T}}\mathbf{x} = \mathbf{z}$ 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 $\log|\mathbf{Q}| = 2\Sigma_i \log(L_{ii})$.

Marginal variances can also be computed efficiently. To see this, we can start with the equation $\mathbf{L}^{\mathrm{T}}\mathbf{x} = \mathbf{z}$ 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 $L_{ii}x_i = z_i - \Sigma_{k=i+1}^{n} L_{ki}x_k$ for $i = n, \ldots, 1$. Multiplying each side with x_j, $j \geqslant i$, and taking the expectation, we obtain

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

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

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

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

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

where $g_i(x_i)$ is $\log\{\pi(y_i|x_i,\boldsymbol{\theta})\}$ in our setting. The Gaussian approximation $\tilde{\pi}_G(\mathbf{x})$ 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,

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

where $\{b_i\}$ and $\{c_i\}$ depend on $\boldsymbol{\mu}^{(0)}$. A Gaussian approximation is obtained, with precision matrix $\mathbf{Q}+\mathrm{diag}(\mathbf{c})$ and mode given by the solution of $\{\mathbf{Q}+\mathrm{diag}(\mathbf{c})\}\boldsymbol{\mu}^{(1)} = \mathbf{b}$. This process is repeated until it converges to a Gaussian distribution with, say, mean $\mathbf{x}^*$ and precision matrix $\mathbf{Q}^* = \mathbf{Q}+\mathrm{diag}(\mathbf{c}^*)$. 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_i and x_j, say, the precision matrix of the Gaussian approximation is of the form $\mathbf{Q}+\mathrm{diag}(\mathbf{c})$. 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, $\pi(x_i|\mathbf{y})$, $i = 1,\ldots,n$. 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 $\boldsymbol{\theta} = (\theta_1,\ldots,\theta_m) \in \mathbb{R}^m$, 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 $\boldsymbol{\Sigma} = \mathbf{V}\boldsymbol{\Lambda}\mathbf{V}^{\mathrm{T}}$ 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

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

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

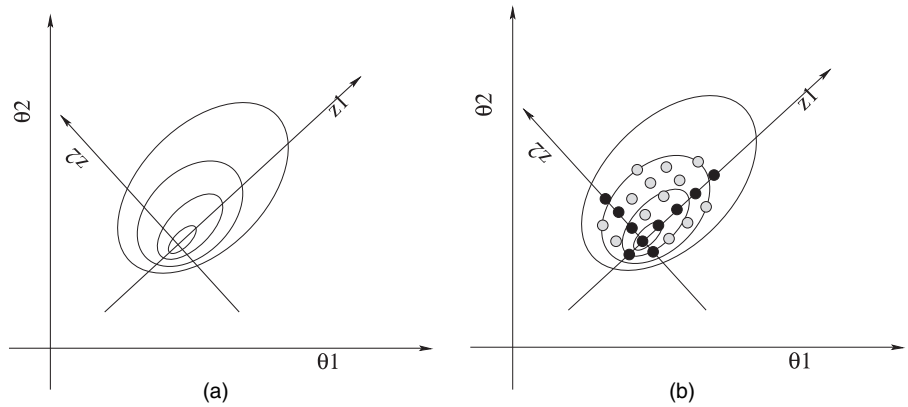

Fig. 1. Illustration of the exploration of the posterior marginal for θ: in (a) the mode is located and the Hessian and the co-ordinate system for $\mathbf{z}$ are computed; in (b) each co-ordinate direction is explored (●) until the log-density drops below a certain limit; finally the new points (◉) 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 (e.g. $\delta_{\mathbf{z}} = \frac{1}{2}$ 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_is, 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 $\tilde{\pi}_{\mathrm{G}}(x_i|\boldsymbol{\theta}, \mathbf{y})$, 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 $\tilde{\pi}_{\mathrm{G}}(\mathbf{x}|\boldsymbol{\theta}, \mathbf{y})$, 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

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

Here, $\tilde{\pi}_{\mathrm{GG}}$ 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 $\tilde{\pi}_{\mathrm{GG}}$ is different from the conditional density corresponding to $\tilde{\pi}_{\mathrm{G}}(\mathbf{x}|\boldsymbol{\theta}, \mathbf{y})$.

Unfortunately, expression (12) implies that $\tilde{\pi}_{\mathrm{GG}}$ must be recomputed for each value of x_i and $\boldsymbol{\theta}$, since its precision matrix depends on x_i and $\boldsymbol{\theta}$. This is far too expensive, as it requires n 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 $\tilde{\pi}_{\mathrm{GG}}(\mathbf{x}_{-i}|x_i, \boldsymbol{\theta}, \mathbf{y})$ by approximating the modal configuration,

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

The right-hand side is evaluated under the conditional density that is derived from the Gaussian approximation $\tilde{\pi}_{\mathrm{G}}(\mathbf{x}|\boldsymbol{\theta}, \mathbf{y})$. 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_js 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

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

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

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

The most important computational saving using $R_i(\boldsymbol{\theta})$ comes from the calculation of the denominator of 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

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

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

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

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

$$\mathrm{d}_j^{(3)}(x_i,\boldsymbol{\theta})=\frac{\partial^3}{\partial x_j^3}\log\{\pi(y_j|x_j,\boldsymbol{\theta})\}\bigg|_{x_j=E_{\tilde{\pi}_{\mathrm{G}}}(x_j|x_i)},$$

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

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

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

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

$$\log\{\pi(\mathbf{x},\boldsymbol{\theta},\mathbf{y})\}\bigg|_{\mathbf{x}_{-i}=E_{\tilde{\pi}_{\mathrm{G}}}(\mathbf{x}_{-i}|x_i)}=-\frac{1}{2}(x_i^{(s)})^2+\frac{1}{6}(x_i^{(s)})^3\sum_{j\in\mathcal{I}\setminus i}\mathrm{d}_j^{(3)}\{\mu_i(\boldsymbol{\theta}),\boldsymbol{\theta}\}\{\sigma_j(\boldsymbol{\theta})a_{ij}(\boldsymbol{\theta})\}^3+\ldots. \tag{18}$$

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

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

where $\mathbf{H}$ is the prior precision matrix of the GMRF with ith 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 $x_j = E_{\tilde{\pi}_{\mathrm{G}}}(x_j|x_i)$; see equation (14). Using that

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

we obtain

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

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

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

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

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

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

then

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

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)

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

where $\phi(\cdot)$ and $\Phi(\cdot)$ are the density and distribution function of the standard normal distribution, and ξ, $\omega>0$ and a are respectively the location, scale and skewness parameters. We fit a skew normal density to 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 $a_{i\cdot}(\boldsymbol{\theta})$, 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 $\tilde{\pi}(x_i|\mathbf{y})$, $i = 1, \ldots, n$, 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 $n_d \to \infty$; 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 McCullagh (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 $n/n_d \to 0$ 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 $n_d \to \infty$, 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*, $p_{\text{D}}(\boldsymbol{\theta})$, as defined by Spiegelhalter *et al.* (2002). Since $\mathbf{x}$ given $\mathbf{y}$ and $\boldsymbol{\theta}$ is roughly Gaussian, $p_{\text{D}}(\boldsymbol{\theta})$ is conveniently approximated by

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

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 $p_{\text{D}}(\boldsymbol{\theta})$ is

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

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

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

4.2. Assessing the approximation error

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

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

$$\frac{\pi(\boldsymbol{\theta}|\mathbf{y})}{\tilde{\pi}(\boldsymbol{\theta}|\mathbf{y})} \propto |\mathbf{Q}^*(\boldsymbol{\theta})|^{1/2} \int \exp[-\tfrac{1}{2}(\mathbf{x}-\mathbf{x}^*(\boldsymbol{\theta}))^{\mathrm{T}}\mathbf{Q}^*(\boldsymbol{\theta})(\mathbf{x}-\mathbf{x}^*(\boldsymbol{\theta})) + r(\mathbf{x};\boldsymbol{\theta},\mathbf{y})]\,\mathrm{d}\mathbf{x}$$
$$= E_{\tilde{\pi}_{\mathrm{G}}}[\exp\{r(\mathbf{x};\boldsymbol{\theta},\mathbf{y})\}],$$

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

Our second strategy is based on the simple idea of comparing elements of a sequence of increasingly accurate approximations. In our case, this sequence consists of the Gaussian approximation (4), followed by the simplified Laplace approximation (22), then by the Laplace approximation (12). Specifically we compute the integrated marginal (5) on the basis of both the Gaussian approximation and the simplified Laplace approximation, and compute their symmetric Kullback–Leibler divergence (SKLD). If the divergence is small then both approximations are considered as acceptable. Otherwise, compute equation (5) by using the Laplace approximation (12) and compute the divergence with that based on the simplified Laplace approximation. Again, if the divergence is small, simplified Laplace and Laplace approximations appear to be acceptable; otherwise, the Laplace approximation is our best estimate but the label 'problematic' should be attached to the approximation to warn the user. (This last option has not yet happened to us.)

To assess the error due to the numerical integration (5), we can compare the SKLD between the posterior marginals 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 $\boldsymbol{\theta}$ in Section 5.1, to verify the (simplified) Laplace approximation for $x_i|\boldsymbol{\theta}, \mathbf{y}$. 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 $\pi(x_i|\boldsymbol{\theta}, \mathbf{y})$ in two quite challenging examples. The first model is based on a first-order auto-regressive latent field with unknown mean,

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

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

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

for $t = 1, \ldots, 50$, 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 $(\mathbf{f}^{\mathrm{T}}, \mu)^{\mathrm{T}}$ from the prior, and then simulated the observations. We computed $\tilde{\pi}(f_t|\boldsymbol{\theta}, \mathbf{y})$ for $t = 1, \ldots, 50$ and $\tilde{\pi}(\mu|\boldsymbol{\theta}, \mathbf{y})$ 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 2 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

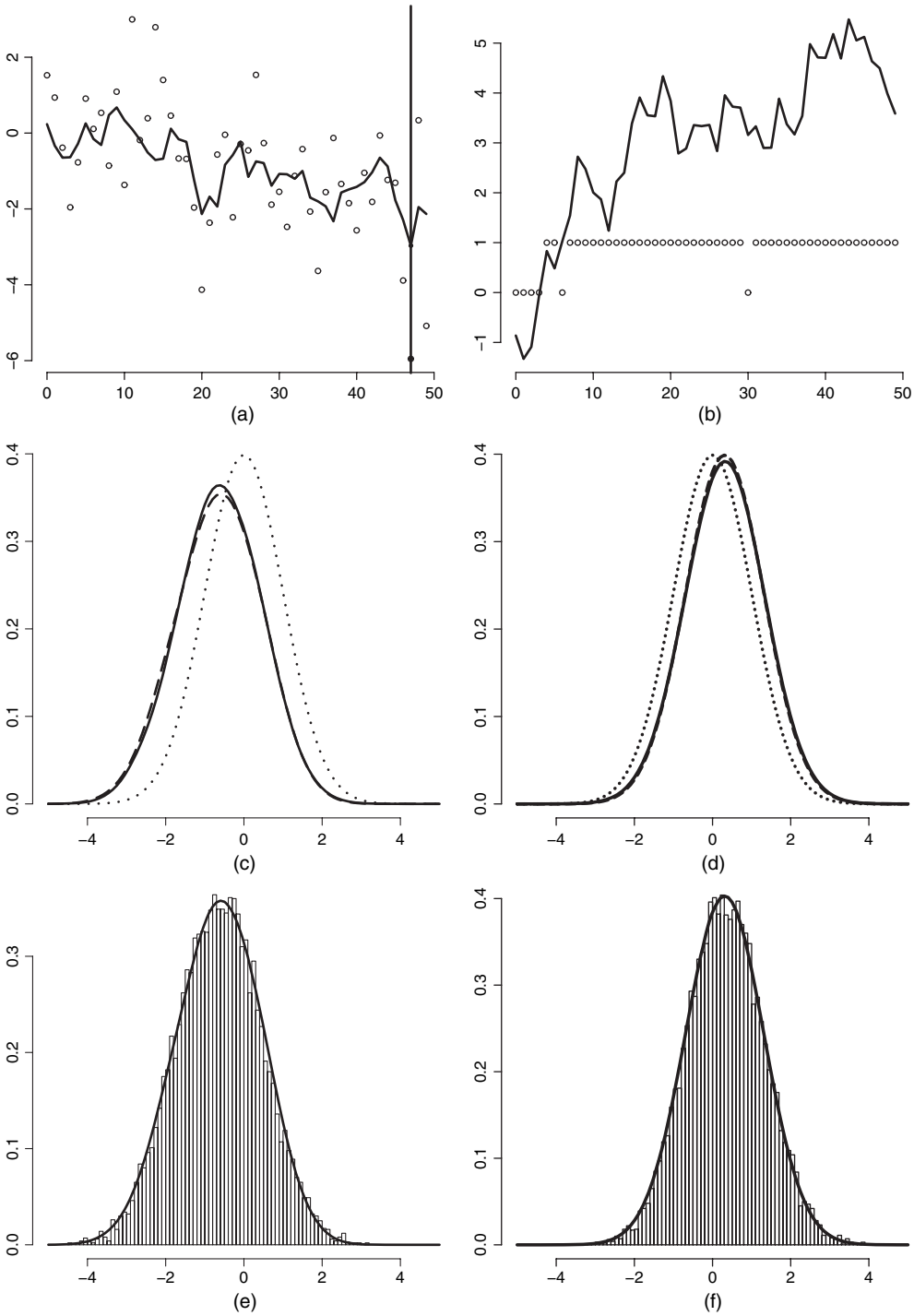

Fig. 2. (a), (b) True latent Gaussian field (———), observed Student t_3-data and Bernoulli data (∘), (c), (d) approximate marginal for a selected node by using various approximations (·····, Gaussian; – – –, simplified Laplace; ———, Laplace) and (e), (f) comparison of samples from a long MCMC chain with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

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

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

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

5.2. A generalized linear mixed model for longitudinal data

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

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

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

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

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

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

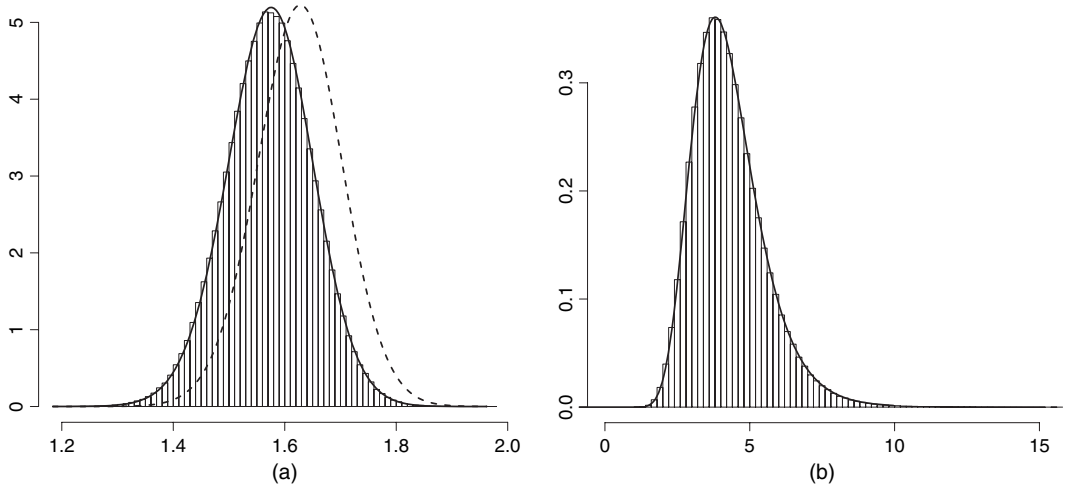

Fig. 3. Posterior marginal for (a) β_0 (———, simplified Laplace approximation; -------, Gaussian approximation) and (b) τ_ε (———, after integrating out τ_ν) for the example in Section 5.2: the histograms result from a long MCMC run using OpenBUGS

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

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

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

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

and μ is a Gaussian mean value. In this example, $\mathbf{x} = (\mu, \eta_1, \ldots, \eta_T)^{\mathrm{T}}$ and $\boldsymbol{\theta} = (\phi, \tau)^{\mathrm{T}}$. The log-likelihood (with respect to η_t) is quite far from being Gaussian and is non-symmetric. There is some evidence that financial data have heavier tails than the Gaussian 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).

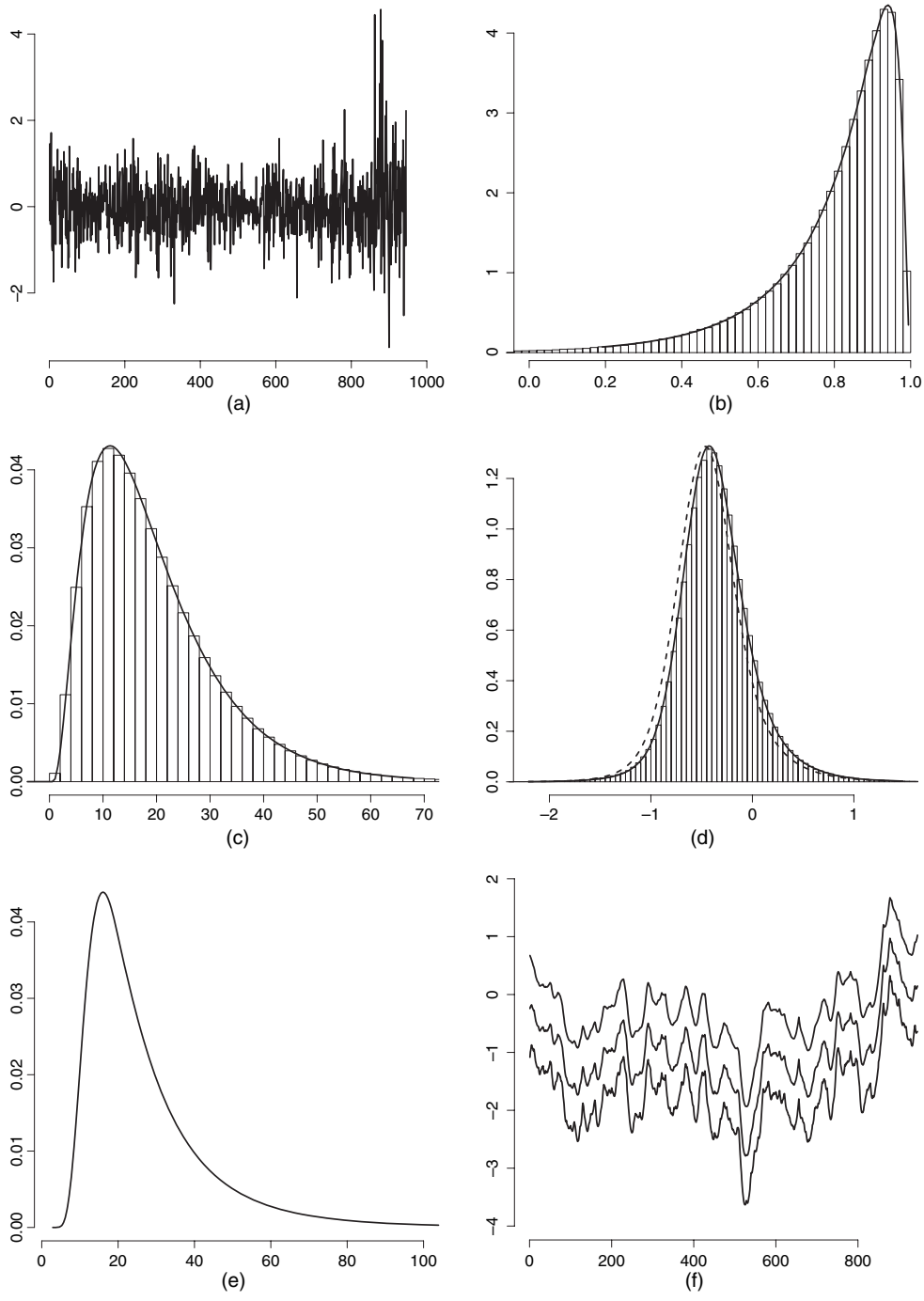

Fig. 4. (a) Log-daily-difference of the pound–dollar exchange rate from October 1st, 1981, to June 28th, 1985, (b), (c) approximated posterior marginals for ϕ and τ by using only the first $n = 50$ observations in (a) (overlaid are the histograms that were obtained from a long MCMC run using OpenBUGS), (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 and Heskes (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 and Heskes (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 $\boldsymbol{\theta}^*$. The effective number of parameters (27) was about 63, which is small compared with n_d. A 95% interval for the remainder $r(\mathbf{x}; \boldsymbol{\theta}^*, \mathbf{y})/n_d$ 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 $i = 1, \ldots, 6990$. There are 216 districts and 15 age groups (15–19, 20–24, ..., > 84 years). We follow Rue and Held (2005), section 4.3.5, and use a logistic binary regression model:

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

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

$$\pi(\mathbf{f}^{(\text{a})} | \kappa^{(\text{a})}) \propto (\kappa^{(\text{a})})^{(15-2)/2} \exp\left\{ -\frac{\kappa^{(\text{a})}}{2} \sum_{j=3}^{15} (f^{(\text{a})}_j - 2f^{(\text{a})}_{j-1} + f^{(\text{a})}_{j-2})^2 \right\}. \tag{27}$$

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

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

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 an $\mathcal{N}(0, 0.1)$ 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 $\boldsymbol{\theta}^*$. The effective number of parameters (24) was about 101, which is small compared with n_d. A 95% interval for the remainder $r(\mathbf{x}; \boldsymbol{\theta}^*, \mathbf{y})/n_d$ 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 $\mathbf{y} = (y_1, \ldots, y_N)^T$. Let η_i be the random variable $Z(\boldsymbol{\xi}_i)$. Clearly $\pi(\mathbf{y}|\boldsymbol{\eta}) = \Pi_i \pi(y_i|\eta_i)$ 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

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

where $\mathbf{f}^{(s)}$ represents 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).

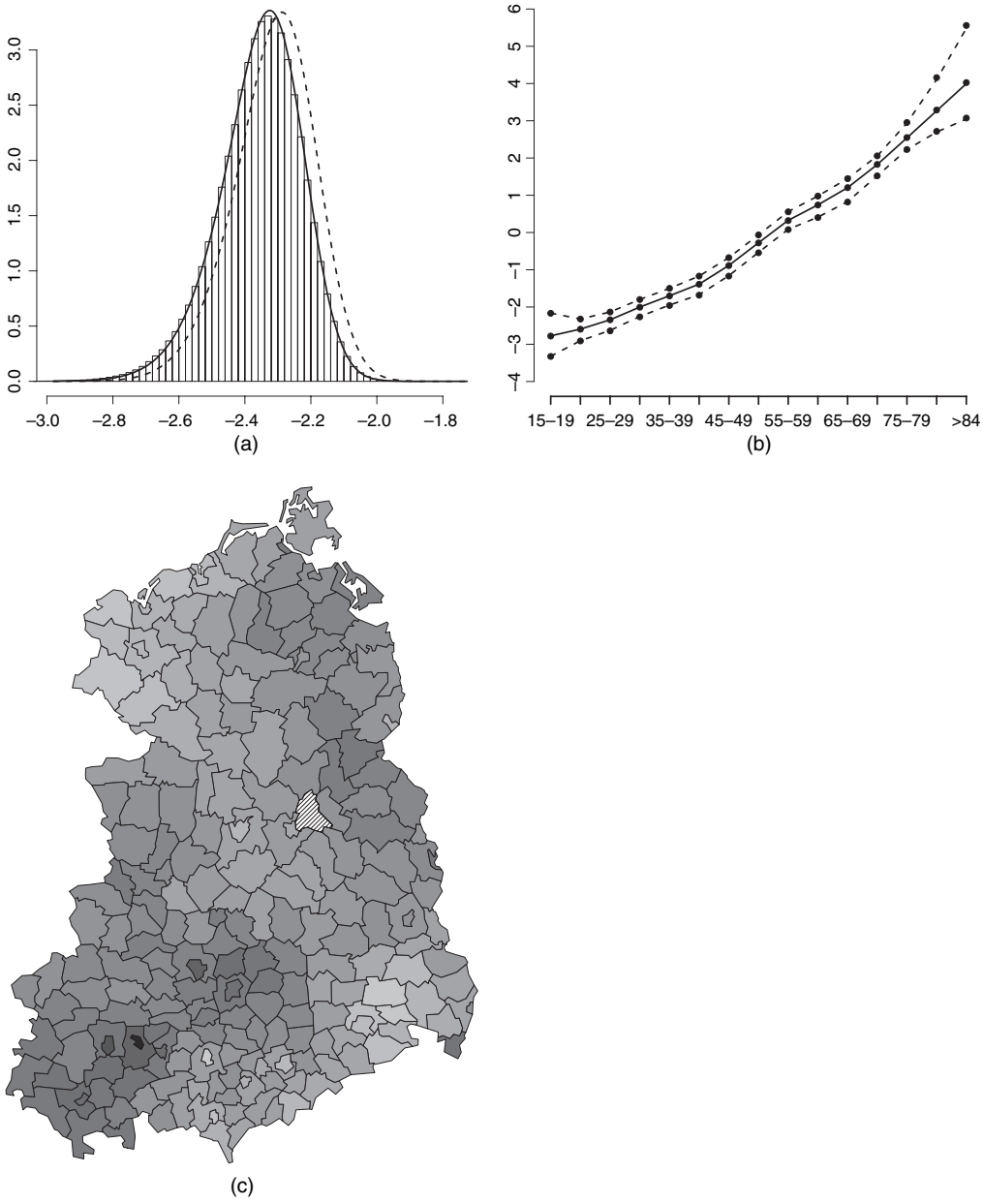

Fig. 5. Results for the cancer incidence example: (a) posterior marginal for $f_3^{(a)}$ by using simplified Laplace approximations (———), Gaussian approximations (-------) and samples from a long MCMC run (□); (b) posterior median (———) and 0.025- and 0.975-quantiles (-------) of the age–class effect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

We start by dividing the area of interest into a 200×100 regular lattice, where each square pixel of the lattice represents 25 m^2. This makes $n_d = 20\,000$. 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 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)

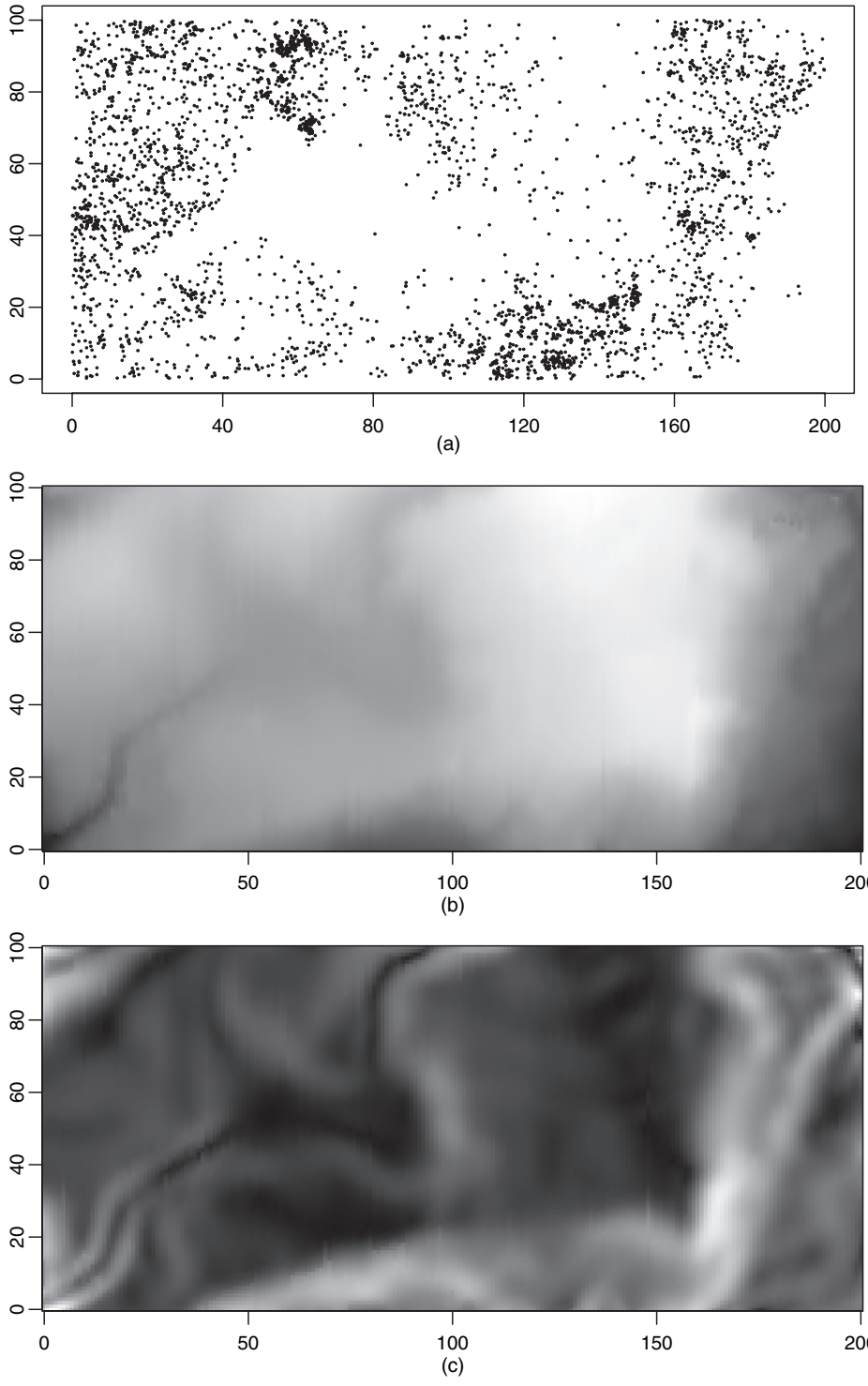

Fig. 6. Data and covariates from the log-Gaussian Cox process example: (a) locations of the 3605 trees, (b) altitude and (c) norm of the gradient

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

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

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 β_0, β_{Alt} and β_{Grad}. The latent field is $\mathbf{x} = (\boldsymbol{\eta}^{\text{T}}, (\mathbf{f}^{(s)})^{\text{T}}, \beta_0, \beta_{\text{Alt}}, \beta_{\text{Grad}})^{\text{T}})$ with dimension 40003, and $\boldsymbol{\theta} = (\log(\kappa^{(s)}), \log(\kappa^{(u)}))$ with dimension 2.

We computed the approximation for 20003 posterior marginals $\mathbf{f}^{(s)}$, β_0, β_{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 become 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 β_0, β_{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 $p_{\text{D}}(\boldsymbol{\theta}^*) \approx 1714$ 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 approximation does not work well in this case. Owing to the size of the GMRF, the comparisons 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

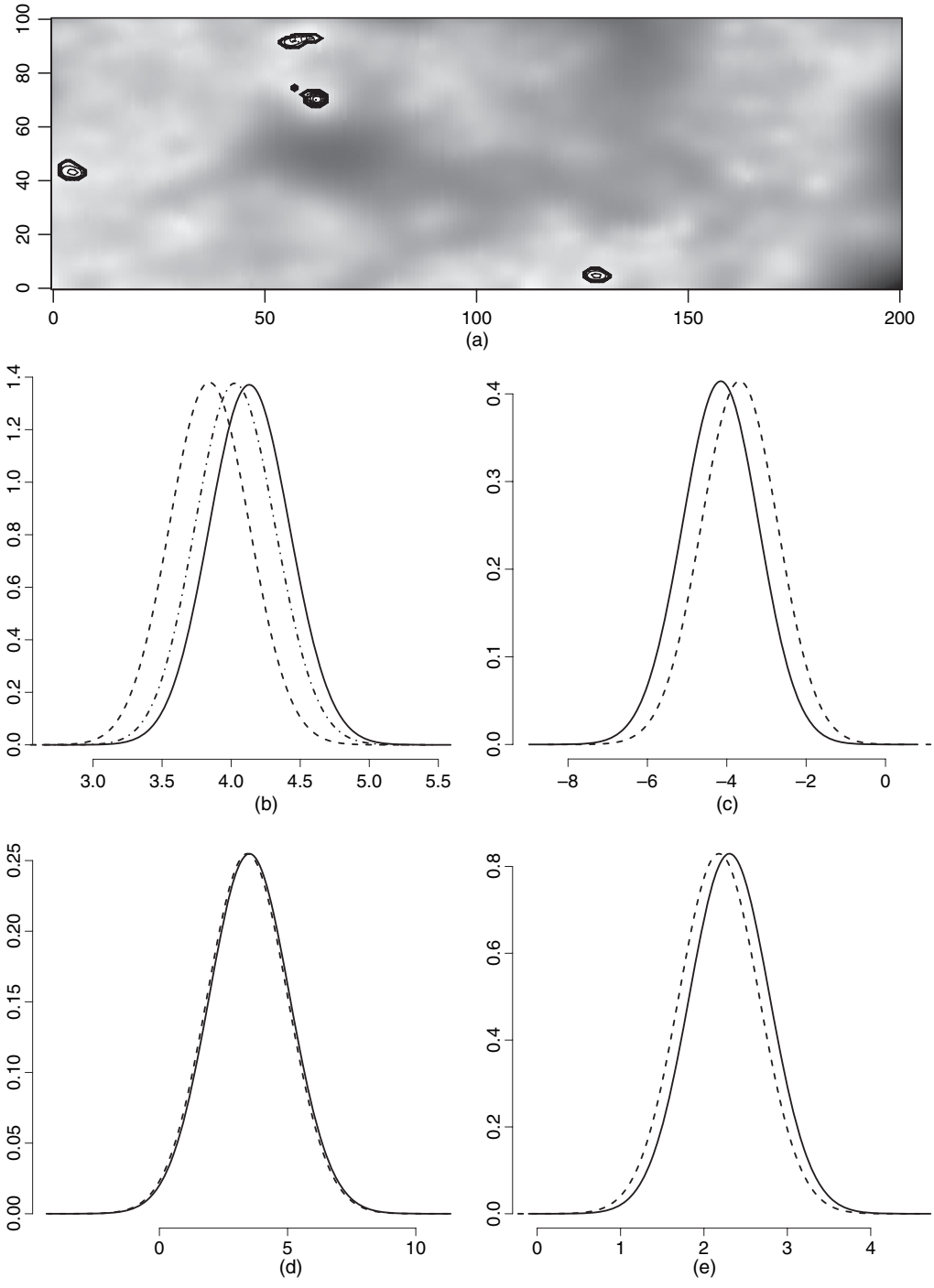

Fig. 7. Log-Gaussian Cox process example: (a) posterior mean of the spatial component with contour indicating an SKLD above 0.25, (b) marginals for node (61,73) in the spatial component with maximum SKLD 0.50, Gaussian (-------), simplified Laplace (———) and Laplace approximations (· - - - - -) and (c)–(e) 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 $x_i|\boldsymbol{\theta}, y$ and $x_j|\boldsymbol{\theta}, y$. From the Gaussian approximation $\tilde{\pi}_{\mathrm{G}}(\mathbf{x}|\boldsymbol{\theta}, \mathbf{y})$ 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 $\tilde{\pi}(\boldsymbol{\theta}|\mathbf{y})$,

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

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 to 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. $\pi(y_i|\mathbf{y}_{-i})$. 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

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

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 $\mathrm{PIT}_i = \mathrm{Prob}(y_i^{\mathrm{new}} \leqslant y_i|\mathbf{y}_{-i})$ 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_is 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_is are then about $(1\text{–}5)\times 10^{-4}$.

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_is. 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, $\delta_{\mathbf{z}} = 1$ 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

$(1,1,1,1,1,1)$,	$(-1,1,1,1,-1)$,	$(1,-1,1,1,-1)$,	$(-1,-1,1,1,1)$,
$(1,1,-1,1,-1)$,	$(-1,1,-1,1,1)$,	$(1,-1,-1,1,1)$,	$(-1,-1,-1,1,-1)$,
$(1,1,1,-1,-1)$,	$(-1,1,1,-1,1)$,	$(1,-1,1,-1,1)$,	$(-1,-1,1,-1,-1)$,
$(1,1,-1,-1,1)$,	$(-1,1,-1,-1,-1)$,	$(1,-1,-1,-1,-1)$,	$(-1,-1,-1,-1,1)$.

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 $\boldsymbol{\theta}^{\mathrm{T}}\boldsymbol{\theta}$ 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 $x_i|\boldsymbol{\theta}, \mathbf{y}$ independently for all i for fixed $\boldsymbol{\theta}$. Both the `inla` program and `GMRFLib` use the OpenMP (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

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

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

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

When $\kappa \to 0$, $\pi(\theta|\mathbf{y}) \to \Gamma(\theta; a, b)$ 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)):

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

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 $b_{t+1} = b + \frac{1}{2}\{\mathbf{m}_{t+1}^{\mathrm{T}}\mathbf{R}\mathbf{m}_{t+1} + \mathrm{tr}(\mathbf{R}\mathbf{Q}_{t+1}^{-1})\}$. The limit b_{∞} of b_t is defined implicitly by the equation

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

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

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

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 (23) equals $\gamma^{(3)}$. This gives three equations to determine (ξ, ω, a). The modal configuration is not available analytically, but a series expansion of the log-skew-normal density around $x = \xi$ gives

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

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

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

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

References

Ainsworth, L. M. and Dean, C. B. (2006) Approximate inference for disease mapping. *Computnl Statist. Data Anal.*, **50**, 2552–2570.

Albert, J. H. and Chib, S. (1993) Bayesian analysis of binary and polychotomous response data. *J. Am. Statist. Ass.*, **88**, 669–679.

Attias, H. (1999) Inferring parameters and structure of latent variable models by variational Bayes. In *Proc. 15th Conf. Uncertainty in Artificial Intelligence*, vol. 2, pp. 21–30. San Francisco: Morgan Kaufmann.

Attias, H. (2000) A variational Bayesian framework for graphical models. *Adv. Neur. Informn Process. Syst.*, **12**, 209–215.

Azzalini, A. and Capitanio, A. (1999) Statistical applications of the multivariate skew normal distribution. *J. R. Statist. Soc.* B, **61**, 579–602.

Banerjee, S., Carlin, B. P. and Gelfand, A. E. (2004) *Hierarchical Modeling and Analysis for Spatial Data*. London: Chapman and Hall.

Banerjee, S., Gelfand, A. E., Finley, A. O. and Sang, H. (2008) Gaussian predictive process models for large spatial data sets. *J. R. Statist. Soc.* B, **70**, 825–848.

Beal, M. J. (2003) Variational algorithms for approximate Bayesian inference. *PhD Thesis*. University College London, London.

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

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

Beskos, A., Papaspiliopoulos, O., Roberts, G. O. and Fearnhead, P. (2006) Exact and computationally efficient likelihood-based estimation for discretely observed diffusion processes (with discussion). *J. R. Statist. Soc.* B, **68**, 333–382.

Bishop, C. M. (2006) *Pattern Recognition and Machine Learning*. New York: Springer.

Box, G. E. P. and Wilson, K. B. (1951) On the experimental attainment of optimum conditions (with discussion). *J. R. Statist. Soc.* B, **13**, 1–45.

Breslow, N. E. and Clayton, D. G. (1993) Approximate inference in generalized linear mixed models. *J. Am. Statist. Ass.*, **88**, 9–25.

Brix, A. and Diggle, P. J. (2001) Spatiotemporal prediction for log-Gaussian Cox processes. *J. R. Statist. Soc.* B, **63**, 823–841.

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

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

Chib, S., Nardari, F. and Shephard, N. (2002) Markov chain Monte Carlo methods for stochastic volatility models. *J. Econometr.*, **108**, 281–316.

Chu, W. and Ghahramani, Z. (2005) Gaussian processes for ordinal regression. *J. Mach. Learn. Res.*, **6**, 1019–1041.

Cressie, N. A. C. (1993) *Statistics for Spatial Data*, revised edn. New York: Wiley.

Cressie, N. and Johannesson, G. (2008) Fixed rank kriging for very large spatial data sets. *J. R. Statist. Soc.* B, **70**, 209–226.

Czado, C., Gneiting, T. and Held, L. (2007) Predictive model assessment for count data. *Technical Report 518*. Department of Statistics, University of Washington, Seattle.

Dey, D. K., Ghosh, S. K. and Mallick, B. K. (eds) (2000) *Generalized Linear Models: a Bayesian Perspective*. Boca Raton: Chapman and Hall–CRC.

Diggle, P. J. and Ribeiro, P. J. (2006) *Model-based Geostatistics*. New York: Springer.

Durbin, J. and Koopman, S. J. (2000) Time series analysis of non-Gaussian observations based on state space models from both classical and Bayesian perspectives (with discussion). *J. R. Statist. Soc.* B, **62**, 3–56.
Eidsvik, J., Martino, S. and Rue, H. (2009) Approximate Bayesian inference in spatial generalized linear mixed models. *Scand. J. Statist.*, to be published.
Fahrmeir, L. and Lang, S. (2001) Bayesian inference for generalized additive mixed models based on Markov random field priors. *Appl. Statist.*, **50**, 201–220.
Fahrmeir, L. and Tutz, G. (2001) *Multivariate Statistical Modelling based on Generalized Linear Models*, 2nd edn. Berlin: Springer.
Finkenstadt, B., Held, L. and Isham, V. (eds) (2006) *Statistical Methods for Spatio-temporal Systems*. Boca Raton: Chapman and Hall–CRC.
Friel, N. and Rue, H. (2007) Recursive computing and simulation-free inference for general factorizable models. *Biometrika*, **94**, 661–672.
Frühwirth-Schnatter, S. and Frühwirth, R. (2007) Auxiliary mixture sampling with applications to logistic models. *Computnl Statist. Data Anal.*, **51**, 3509–3528.
Frühwirth-Schnatter, S. and Wagner, H. (2006) Auxiliary mixture sampling for parameter-driven models of time series of small counts with applications to state space modelling. *Biometrika*, **93**, 827–841.
Gamerman, D. (1997) Sampling from the posterior distribution in generalized linear mixed models. *Statist. Comput.*, **7**, 57–68.
Gamerman, D. (1998) Markov chain Monte Carlo for dynamic generalised linear models. *Biometrika*, **85**, 215–227.
Gelfand, A. E. (1996) Model determination using sampling-based methods. In *Markov Chain Monte Carlo in Practice* (eds W. R. Gilks, S. Richardson and D. J. Spiegelhalter), pp. 145–161. London: Chapman and Hall.
Gelman, A., Carlin, J. B., Stern, H. S. and Rubin, D. B. (2004) *Bayesian Data Analysis*, 2nd edn. Boca Raton: Chapman and Hall–CRC.
Gneiting, T. (2002) Nonseparable, stationary covariance functions for space-time data. *J. Am. Statist. Ass.*, **97**, 590–600.
Gneiting, T. and Raftery, A. E. (2007) Strictly proper scoring rules, prediction, and estimation. *J. Am. Statist. Ass.*, **102**, 359–378.
Gschløssl, S. and Czado, C. (2008) Modelling count data with overdispersion and spatial effects. *Statist. Pap.*, **49**, 531–552.
Held, L., Natario, I., Fenton, S., Rue, H. and Becker, N. (2005) Towards joint disease mapping. *Statist. Meth. Med. Res.*, **14**, 61–82.
Hinton, G. E. and van Camp, D. (1993) Keeping the neural networks simple by minimizing the description length of the weights. In *Proc. 6th A. Conf. Computational Learning Theory, Santa Cruz*, pp. 5–13. New York: Association for Computing Machinery Press.
Holmes, C. C. and Held, L. (2006) Bayesian auxiliary variable models for binary and multinomial regression. *Bayes. Anal.*, **1**, 145–168.
Hsiao, C. K., Huang, S. Y. and Chang, C. W. (2004) Bayesian marginal inference via candidate's formula. *Statist. Comput.*, **14**, 59–66.
Humphreys, K. and Titterington, D. M. (2000) Approximate Bayesian inference for simple mixtures. In *Proc. Computational Statistics 2000* (eds J. G. Bethlehem and P. G. M. van der Heijden), pp. 331–336. Heidelberg: Physica.
Jordan, M. I. (2004) Graphical models. *Statist. Sci.*, **19**, 140–155.
Kammann, E. E. and Wand, M. P. (2003) Geoadditive models. *Appl. Statist.*, **52**, 1–18.
Kass, R. E., Tierney, L. and Kadane, J. B. (1999) The validity of posterior expansions based on Laplace's method. In *Essays in Honor of George Bernard* (eds S. Geisser, J. S. Hodges, S. J. Press and A. Zellner), pp. 473–488. Amsterdam: North-Holland.
Kass, R. E. and Vaidyanathan, S. K. (1992) Approximate Bayes factors and orthogonal parameters, with application to testing equality of two binomial proportions. *J. R. Statist. Soc.* B, **54**, 129–144.
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. *Biometrics*, **58**, 492–501.
Knorr-Held, L. and Rue, H. (2002) On block updating in Markov random field models for disease mapping. *Scand. J. 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**, 183–212.
Mackay, D. J. C. (1995) Ensemble learning and evidence maximization. *Technical Report*. Cavendish Laboratory University of Cambridge, Cambridge.
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.
Martino, S. (2007) Approximate Bayesian inference for latent Gaussian models. *PhD Thesis*. Department of Mathematical Sciences, Norwegian University of Science and Technology, Trondheim.
Martino, S. and Rue, H. (2008) Implementing approximate Bayesian inference for latent Gaussian models using integrated nested Laplace approximations: a manual for the `inla`-program. *Technical Report 2*. Department of Mathematical Sciences, Norwegian University of Science and Technology, Trondheim.
Minka, T. P. (2001) Expectation propagation for approximate Bayesian inference. *Uncertnty Artif. Intell.*, **17**, 362–369.
Møller, J., Syversveen, A. R. and Waagepetersen, R. P. (1998) Log Gaussian Cox processes. *Scand. J. Statist.*, **25**, 451–482.
Møller, J. and Waagepetersen, R. (2003) *Statistical Inference and Simulation for Spatial Point Processes*. London: Chapman and Hall.
Neal, R. M. (1998) Regression and classification using Gaussian process priors. In *Bayesian Statistics 6* (eds J. M. Bernardo, J. O. Berger, A. P. Dawid and A. F. M. Smith), pp. 475–501. New York: Oxford University Press.
O'Hagan, A. (1978) Curve fitting and optimal design for prediction (with discussion). *J. R. Statist. Soc.* B, **40**, 1–42.
Papaspiliopoulos, O., Roberts, G. O. and Sköld, M. (2007) A general framework for the parameterization of hierarchical models. *Statist. Sci.*, **22**, 59–73.
Pettit, L. I. (1990) The conditional predictive ordinate for the normal distribution. *J. R. Statist. Soc.* B, **52**, 175–184.
Rasmussen, C. E. and Williams, C. K. I. (2006) *Gaussian Processes for Machine Learning*. Cambridge: MIT Press.
R Development Core Team (2007) *R: a Language and Environment for Statistical Computing*. Vienna: R Foundation for Statistical Computing.
Rellier, G., Descombes, X., Zerubia, J. and Falzon, F. (2002) A Gauss-Markov model for hyperspectral texture analysis of urban areas. In *Proc. 16th Int. Conf. Pattern Recognition*, pp. 692–695. Washington DC: Institute of Electrical and Electronics Engineers Computer Society.
Robert, C. P. and Casella, G. (1999) *Monte Carlo Statistical Methods*. New York: Springer.
Rue, H. (2001) Fast sampling of Gaussian Markov random fields. *J. R. Statist. Soc.* B, **63**, 325–338.
Rue, H. and Held, L. (2005) *Gaussian Markov Random Fields: Theory and Applications*. London: Chapman and Hall–CRC Press.
Rue, H. and Martino, S. (2007) Approximate Bayesian inference for hierarchical Gaussian Markov random fields models. *J. Statist. Planng Inf.*, **137**, 3177–3192.
Rue, H., Steinsland, I. and Erland, S. (2004) Approximating hidden Gaussian Markov random fields. *J. R. Statist. Soc.* B, **66**, 877–892.
Sanchez, S. M. and Sanchez, P. J. (2005) Very large fractional factorials and central composite designs. *ACM Trans. Modlng Comput. Simuln*, **15**, 362–377.
Schervish, M. J. (1995) *Theory of Statistics*, 2nd edn. New York: Springer.
Shephard, N. (1994) Partial non-Gaussian state space. *Biometrika*, **81**, 115–131.
Shephard, N. and Pitt, M. K. (1997) Likelihood analysis of non-Gaussian measurement time series. *Biometrika*, **84**, 653–667.
Shun, Z. and McCullagh, P. (1995) Laplace approximation of high dimensional integrals. *J. R. Statist. Soc.* B, **57**, 749–760.
Smith, A. F. M., Skene, A. M., Shaw, J. E. H. and Naylor, J. C. (1987) Progress with numerical and graphical methods for practical Bayesian statistics. *Statistician*, **36**, 75–82.
Spiegelhalter, D. J., Best, N. G., Carlin, B. P. and van der Linde, A. (2002) Bayesian measures of model complexity and fit (with discussion). *J. R. Statist. Soc.* B, **64**, 583–639.
Terzopoulos, D. (1988) The computation of visible-surface representations. *IEEE Trans. Pattn Anal. Mach. Intell.*, **10**, 417–438.
Thall, P. F. and Vail, S. C. (1990) Some covariance models for longitudinal count data with overdispersion. *Biometrics*, **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.
Titterington, 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 regression. *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 Epidemiology: 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 Titterington, 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. Society for Artificial Intelligence and Statistics.
Wang, B. and Titterington, D. M. (2006) Convergence properties of a general algorithm for calculating variational Bayesian estimates for a normal mixture model. *Bayes. Anal.*, **1**, 625–650.
Wecker, W. E. and Ansley, C. F. (1983) The signal extraction approach to nonlinear regression and spline smoothing. *J. Am. Statist. Ass.*, **78**, 81–89.
Weir, I. S. and Pettitt, A. N. (2000) Binary probability maps using a hidden conditional autoregressive Gaussian process with an application to Finnish common toad data. *Appl. Statist.*, **49**, 473–484.
West, M. and Harrison, J. (1997) *Bayesian Forecasting and Dynamic Models*, 2nd edn. New York: Springer.
Williams, C. K. I. and Barber, D. (1998) Bayesian classification with Gaussian processes. *IEEE Trans. Pattn Anal. Mach. Intell.*, **20**, 1342–1351.
Zoeter, O. and Heskes, T. (2005) Gaussian quadrature based expectation propagation. In *Proc. 10th Int. Wrkshp Artificial Intelligence and Statistics* (eds R. G. Cowell and Z. Ghahramani), pp. 445–452. Society for Artificial Intelligence and Statistics.

Discussion on the paper by Rue, Martino and Chopin

Gareth Roberts (*University of Warwick, Coventry*)

The authors are to be congratulated on a thought-provoking and challenging contribution, which is the culmination of an impressive body of work on analytic approximation alternatives to Markov chain Monte Carlo (MCMC) methods in Bayesian computation. The ease and speed of implementation are appealing and the empirical results seem impressive. The idea of focusing on large generic families of problems with characteristic posterior distributions seems to me to be very sensible. Furthermore, the availability of R software will greatly assist the assimilation of this methodology into statistical practice.

I particularly support the view that MCMC methods need to be challenged and improved, and the existence of alternative methodologies can only serve to encourage the development of improved methods, based on both analytic approximations and Monte Carlo methods.

Some asymptotics

I shall make some remarks on the general problem of simulating in high dimensional space, with a view to shedding some light on the reasons why the authors' methodology is so rapid in comparison with off-the-shelf MCMC methods, and to explain how some MCMC methods can improve considerably on standard MCMC methods, though fail to approach the efficiency of the authors' approach. The theory supporting the attachments below can be found, for example, in Roberts and Rosenthal (2001).

Therefore consider the generic problem of simulating, approximating or calculating moments in n-dimensional space. The computational complexity of the methods that are proposed in the paper can be as good as $O(n)$. How does this compare with MCMC algorithms?

Computational complexity of implementing one iteration of an MCMC sampler is at least $O(n)$. However, it is also necessary to consider the mixing time of MCMC sampling. It is now well established that for random-walk Metropolis methods (suitably optimized) the *best* that we can hope for is $O(n)$. Actually the situation could be considerably worse for strongly dependent distributions or where algorithms are not scaled appropriately. However, this is a realistic goal for common statistical applications. This makes the most optimistic prediction of overall computing cost for the MCMC approach to be $O(n^2)$. Therefore this explains why the methods in the paper are computationally much less demanding than standard MCMC approaches.

Can Markov chain Monte Carlo methods do better?

There are simple modifications of random-walk Metropolis methods which greatly enhance their performance. I shall concentrate on perhaps the simplest. The Metropolis-adjusted Langevin algorithm MALA (Besag, 1994; Roberts and Tweedie, 1996a) uses proposals something like

$$\mathbf{Y} \sim N[\mathbf{X}_t + \nabla \log\{\pi(\mathbf{X}_t)\}h^2/2, h^2]$$

where π is the target density.

Implementation of MALA is typically $O(n)$ and its mixing time (suitably optimized) is $n^{1/3}$, making an overall cost of the order $n^{4/3}$.

The following example is taken from Christensen *et al.* (2006) and Möller *et al.* (1998). The data consist of the locations of 126 Scots pine saplings in a Finnish forest (Fig. 8). The observed point pattern modelled as a Poisson point process X with intensity

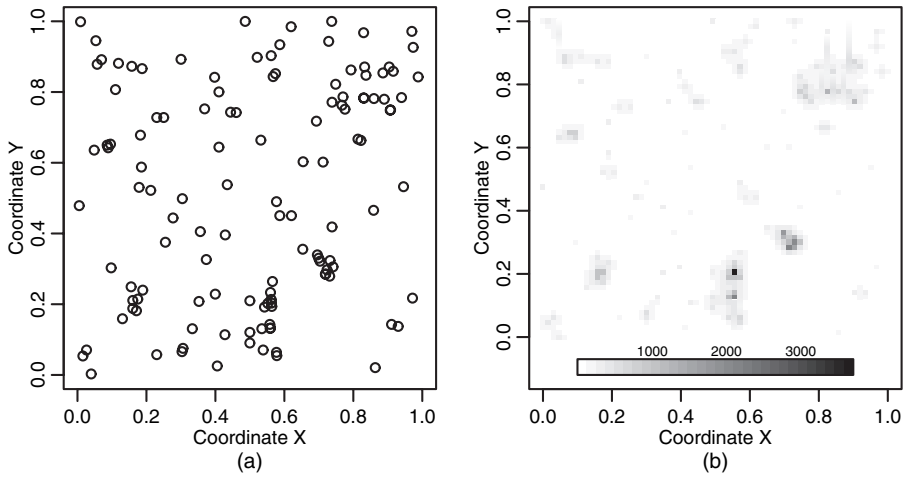

Fig. 8. Scots pine saplings: (a) locations of trees; (b) estimated intensity $E\{\Lambda(s)|x\}$

$$\Lambda(s) = \exp\{Y(s)\},$$

where $Y(\cdot) = \{Y(s)|s \in \mathbf{R}^2\}$ is a Gaussian process with mean $E\{Y(s)\} = \mu$ and covariance

$$\mathrm{cov}\{Y(s), Y(s')\} = \sigma^2 \exp(-\|s - s'\|\beta).$$

The latent Gaussian (prior) process is discretized on a 64×64 regular grid.

In this example, updating the latent field requires MALA updates since random-walk Metropolis methods are prohibitively slow in mixing. Here we compare the performance of the algorithm for three different starting values. The starting values expressed in terms of Y (which must be transformed to starting values for Γ) are

(a) $Y_{i,j} = \mu$ for $i, j = 1, \ldots, 64$,
(b) a random starting value, simulated from the prior $Y \sim N(\mu, \Sigma)$, and
(c) a starting value near the posterior mode. Let $Y_{i,j}$ solve the equation $0 = x_{i,j} - \exp(Y_{i,j}) - (Y_{i,j} - \beta)/\sigma^2$.

In all three cases we use the scaling $l^2/4096^{1/3} = 0.16$ where $l = 1.6$ is derived by using 'optimal scaling' criteria; see for example Roberts and Rosenthal (2007) (Fig. 9).

The instability of MALA algorithms in their transient phase is well known; see Roberts and Tweedie (1996b) and Christensen *et al.* (2005). However, there is an improved methodology which is robust to bad starting values. Now, using the scaling, $l^2/4096^{1/2} = 0.034$ (Fig. 10).

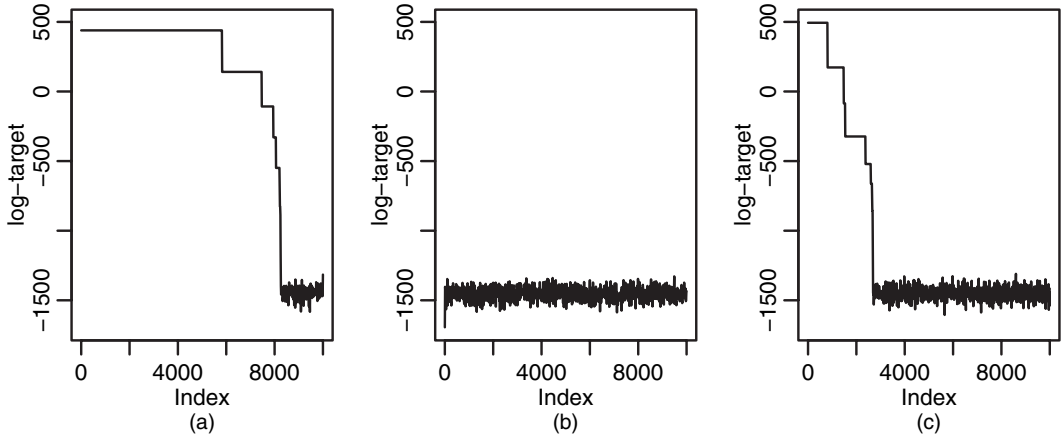

Fig. 9. Scots pine saplings—trace plots $\log(\gamma|x)$ when using the scaling 0.16: (a) starting value I; (b) starting value II; (c) starting value III

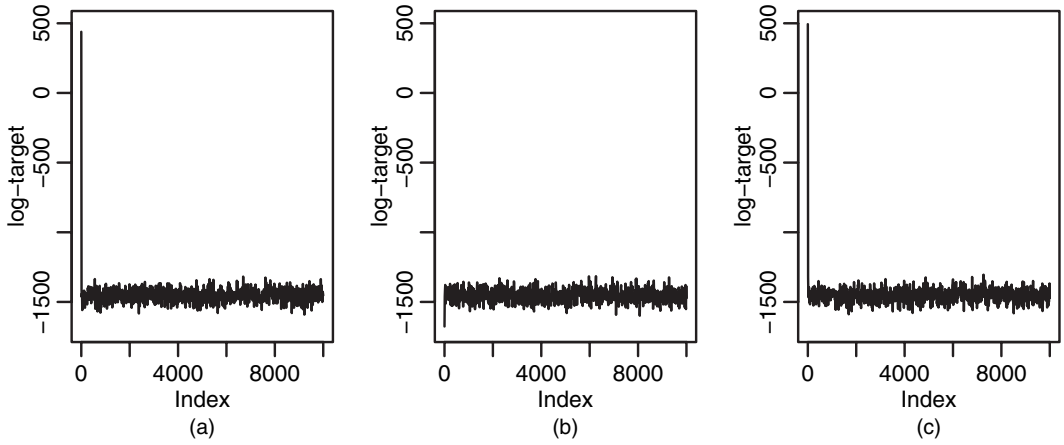

Fig. 10. Scots pine saplings—trace plots log($\gamma|x$) when using the scaling 0.34: (a) starting value I; (b) starting value II; (c) starting value III

Stability of these methods to poor starting values therefore requires care. One would expect similar problems to afflict all competing numerical methods, and it would be interesting to find out whether similar issues affect the methods that were introduced in the paper.

Further thoughts about asymptotics

Section 4.1 looks at several asymptotic regimes. For the type of problems that are considered in the paper, it is important to consider a number of different regimes. For MCMC, 'outfill' asymptotics where $n_d/n = O(1)$, methods typically achieve the *best possible* rates for convergence times above (subject to some kind of *no phase transition* problem). However, for MCMC, 'infill' asymptotics where $n_d/n = O(1)$, methods are typically much worse. Do the same issues aflict the methods of this paper in another form, for instance instability due to almost singularity of covariance matrices?

Finally I add a remark about errors in MCMC analysis. The paper alludes to the fact that Monte Carlo methodology (with T iterations) has error $O(T^{-1/2})$. Actually MCMC implementational methodology typically relies on this in its empirical assessment of Monte Carlo error. It is important to recognize that for MCMC methods the truth can be much worse than that, and there are innocuous looking situations where $\sqrt{n}$–central limit theorems fail; see for example Roberts and Rosenthal (2007). Here, we just give an informal statement of the main result in Roberts and Rosenthal (2007).

Theorem 1 (Roberts and Rosenthal, 2007). For reversible Markov chains, the following are equivalent:

(a) the chain is *variance bounding* (essentially the same as *geometrically ergodic*);
(b) for all functions f such that $\mathbf{E}_\pi\{f^2(X)\} < \propto$ a $\sqrt{N}$–central limit theorem holds:

$$\lim_{N\to\infty}\left[N^{1/2}\left\{N^{-1}\sum_{t=1}^{N} f(X_t) - \pi(f)\right\}\right] \to N(0, \sigma_f^2).$$

Conclusions

MCMC methods can work well on problems that are described in the paper, i.e. the computational gap can be reduced (but not removed completely), but this requires careful algorithm construction. (Perhaps there is a role here for *adaptive MCMC* methods.) The authors' work offers an appealing and rapid approximation which provides a pragmatic alternative to full posterior exploration. I think it is imperative that we learn more about the accuracy of the approximations that are used in the paper; in particular, how this is affected by the data. However, it is clear that this is exciting and stimulating work. I am therefore very pleased to be able to propose the vote of thanks to the authors for their work.

Christian P. Robert (*Université Paris Dauphine and Centre de Recherche en Economie et Statistique, Malakoff*)

Rue, Martino and Chopin are to be congratulated on their impressive and wide-ranging attempt at overcoming the difficulties in handling latent Gaussian structures. In time series as well as spatial problems, the

explosion in the dimension of the latent variable is indeed a stumbling block for Markov chain Monte Carlo (MCMC) implementations and convergence, and recent solutions all resort to approximations of sorts whose effect has not always been completely assessed (see, for example, Polson *et al.* (2008)). The double Laplace approximation that is proposed by the authors is therefore an interesting and competing alternative in these areas.

Nonetheless, as much as I respect the other major contributions to (Bayesian) statistics, mathematics, and other branches of science, of my fellow Norman, Pierre Simon de Laplace, and despite earlier uses in Bayesian statistics (Tierney *et al.*, 1989), I have always (Robert, 1992, 1994) been wary of Laplace's approximation because

(a) it is not parameterization invariant,
(b) it requires some analytic computation or/and some black box numerical differentiation, while being based on a standard second-order Taylor approximation to the log-density, and
(c) it misses a clean evaluation of the associated error.

In the present paper the amount of recourse to this approximation is particularly intense since both $\pi(\mathbf{x}|\theta, \mathbf{y})$ and $\pi(\mathbf{x}_i|\theta, \mathbf{x}_{-i}, \mathbf{y})$ are approximated by multilevel (nested) Laplace approximations. I, however, contend that a less radical interpretation of approximation (3) as a proposal could lead to a manageable MCMC implementation, at least in some settings.

My first reservation is that the calibration of those Laplace approximations seems to require a high level of expertise that conflicts with the off-the-shelf features that are advertised in the final section of the paper. Designing the approximation then represents an overwhelming portion of the analysis time, whereas computation indeed becomes close to instantaneous, unless one abandons the analytical derivation for a numerical version that is difficult to trust entirely. After reading the paper, I certainly feel less than confident in implementing this type of approximation, although I did not try to use the generic open source software that has been carefully developed by Rue and Martino. Attempting to apply this approach to the standard stochastic volatility model using our own programming thus took us quite a while, even though it produces decent approximations to the marginal $\pi(\theta|\mathbf{y})$, as Casarin and Robert will discuss here later. I am, however, wondering whether or not approximation (3) is a proper density for any and every model, since it is written as $|\mathbf{Q}(\boldsymbol{\theta})|^{1/2}|\mathbf{Q}(\boldsymbol{\theta}) + \mathrm{diag}\{c(\boldsymbol{\theta})\}|^{-1/2}\,\pi\{\mathbf{x}^*(\boldsymbol{\theta}), \boldsymbol{\theta}, \mathbf{y}\}$, where the dependence of both $\mathbf{x}^*$ and c on $\boldsymbol{\theta}$ is quite obscure.

My second reservation is that, although the pseudomarginal (3) seems to be an acceptably manageable version of the marginal posterior of θ, the additional Laplace approximations in the derivation of the marginals of the x_is do not appear as crucial or necessary. Indeed, once approximation (3) is available as a (numerical) approximation to $\pi(\boldsymbol{\theta}|\mathbf{y})$, given that $\boldsymbol{\theta}$ has a limited number of components, as hinted at on several occurrences in the paper, a regular MCMC strategy targeted at this distribution is likely to work. This would result in a much smaller cost than the discretization underlying Fig. 1 (which cannot resist the curse of dimensionality, unless cruder approximations as in Section 6.5 are used, but at a cost in accuracy). Once simulations of the $\boldsymbol{\theta}$s are available, simulations of the $\mathbf{x}$s can also be produced by using the Gaussian approximation as a proposal and the true target as a Metropolis–Hastings correction. (It is thus possible to envision the whole simulation of the pair $(\theta, \mathbf{x})$.) Indeed, the derivation of $\tilde{\pi}(x_i|\theta, y)$ is extremely complex and thus difficult to assess. In particular, the construction is partly numeric and must be repeated for a large number of θs, even though I understand that this does not induce high computing costs. Given that the $\tilde{\pi}(x_i|\theta, y)$s are averaged over several values of θ, it somehow is delicate to become convinced that this complex construction is worthwhile, when compared with an original Gaussian approximation coupled with an MCMC simulation. Simulating a single x_i or the whole vector $\mathbf{x}$ from the Gaussian approximation has the same precision (in x_i); therefore the dimension of $\mathbf{x}$ cannot be advanced as having a negative effect on the convergence of the MCMC algorithm. Furthermore, a single run of the chain produces approximations for all x_is.

My last reservation is that the error resulting from this approximation is not, despite the authors' efforts, properly assessed. We can check on the simpler examples that the error resulting from one of the schemes is indeed minimal but the $\mathcal{O}(n_d^{-1})$ error rates do little for my reassurance as

(a) they involve the sample size, even though we are dealing with a fixed sample, and not a computational effort that seems to be restricted to the size of the grid in the θ-space, and
(b) they do not calibrate the error resulting from the Bayesian inference based on this approximation.

Using and comparing different approximations that are all based on the same principle (Laplace's!) does not provide a clear indicator of the performances of those approximations. Furthermore, resorting to the

measure of effective dimension of Spiegelhalter *et al.* (2002) does not bring additional validation to the approximation.

A minor side remark is that, from a Bayesian point of view, I find rather annoying that latent variables and parameters or hyperparameters with Gaussian priors are mixed together in $\mathbf{x}$ (as in the stochastic volatility example) and that $\boldsymbol{\theta}$ coalesces all left-overs without paying any attention to the model hierarchy (as in Section 1.3 with θ_1 *versus* θ_2). Of course, this does not impact the sampling performances of the method, but it still feels awkward. In addition, this may push towards a preference for Gaussian priors, since, the more (hyper)parameters with Gaussian priors, the smaller m and the less costly the numerical integration step.

Given the potential advances resulting from (as well as the challenges that are posed by) this paper, both in terms of modelling and of numerical approximation, I am unreservedly glad to second the vote of thanks.

The vote of thanks was passed by acclamation.

Peter J. Diggle (*Lancaster University*)

The rediscovery of Markov chain Monte Carlo (MCMC) methods in the 1980s has had a profound effect on statistical practice, enabling the apparently routine fitting of complex models to experimental and observational data. The benefits of this are obvious, but they come at a price.

The authors of this paper raise one often-voiced concern with MCMC methods, namely that model fitting can take hours, or even days, of computing time. I discount this objection when the model in question is not in doubt—waiting days for the answer seems unexceptionable when data may have taken months or years to collect—but it carries some force at an exploratory stage, when many competing models may need to be fitted and their results compared.

Another familiar concern is the difficulty of asessing convergence of MCMC algorithms. This seems to me a more important problem, and one that is fundamentally insoluble in complete generality unless and until exact algorithms can be found for a usefully wide class of models. Until then, I see a fundamental distinction between MCMC and direct Monte Carlo methods of inference, namely that, for the latter, convergence is not an issue; hence the assessment of Monte Carlo error is straightforward by using elementary methods applied to independent simulations from the target distribution.

Finally, there is an ever-present danger that the ability to fit overcomplicated models to sparse data sets will encourage people to do precisely that. And, within the Bayesian paradigm at least, a flat likelihood is no bar to inference and the fact that apparently innocuous prior specifications can be highly informative in some dimensions of the multivariate posterior distribution can easily go unnoticed.

For all these reasons and more, reliable alternatives to MCMC methods would be of enormous value, and I consider this paper to be a very important development.

I would like to be able to use the authors' methods for geostatistical prediction. In this context X represents an unobserved, spatially continuous stochastic process and a common requirement is to be able to draw samples from the predictive distribution of one or more non-linear functionals of X, e.g. its maximum value over a spatial domain of interest. The authors mention in Section 6.1 the possiblity of extending their methods to enable evaluation of the joint predictive distribution for multiple elements of X, but they warn us that the accuracy of their methods 'may possibly decrease' as the dimensionality of X increases. Do they have any more tangible results on this issue?

Leonhard Held and Andrea Riebler (*University of Zurich*)

In a wide-ranging paper, Besag *et al.* (1995) saw Markov chain Monte Carlo (MCMC) methods as 'putting probability back into statistics'. The authors of the paper discussed tonight are to be congratulated on bringing 'numerics back into statistics'. Their work proposes important extensions on numerical approximations to posterior distributions (Tierney and Kadane, 1986). The accuracy and computational speed of the integrated nested Laplace approximation (INLA), as illustrated in the examples, is remarkable and—without doubt—INLA will replace MCMC sampling in routine applications of structured additive regression models.

One advantage of MCMC methods is that functions of high dimensional posterior distributions can be estimated easily. For example, Besag *et al.* (1995), section 6.3, proposed an algorithm to compute *simultaneous credible bands*, a useful tool for Bayesian data analysis (Held, 2004). Such bands would address the question whether a linear age group effect is plausible in example 5.4. Similarly, there is often interest in the *posterior ranking* of a large number of random effects (Goldstein and Spiegelhalter, 1996). In Section

6.1, Rue and his colleagues consider approximations of multivariate marginal distributions of small size. For up to what number of parameters can we compute simultaneous credible bands and posterior ranks with INLA?

Going back to example 5.4, the question arises where exactly INLA adjusts for the sum-to-zero constraint on $\mathbf{f}^{(a)}$ and $\mathbf{f}^{(s)}$. Martino and Rue (2008) mention a small positive scalar c to be added to the diagonal of the precision matrix of the unconstrained model to ensure positive definiteness. Does this include an additional approximation error depending on c? A similar problem occurs in MCMC sampling (Knorr-Held and Rue (2002), page 608); however, a Metropolis–Hastings step automatically corrects for c, so the algorithm remains valid exactly.

An exciting feature of INLA is the ease with which the marginal likelihood and cross-validated probability integral transform values can be calculated. There is a large literature on how to estimate these quantities through MCMC methods; however, some of the estimates proposed have unduly large Monte Carlo error. A multivariate extension of the probability integral transform value, the Box (1980) density ordinate transform has recently been suggested by Gneiting *et al.* (2008) as

$$1 - P\{\pi(\mathbf{Y}_S|y_{-S}) \leqslant \pi(\mathbf{y}_S|\mathbf{y}_{-S})\},$$

where $\mathbf{y}_S = \{y_i : i \in S\}$ and $\mathbf{Y}_S$ is the associated random vector with density $\pi(\mathbf{y}_S|\mathbf{y}_{-S})$. We would be interested to learn from the authors whether this quantity can be computed via INLA for moderate dimension of S.

Finally we are wondering whether a *sequential* version of INLA can be developed for dynamic models with non-Gaussian observations to improve the accuracy of the extended Kalman filter and smoother (Fahrmeir, 1992).

Adam M. Johansen (*University of Warwick, Coventry*)

I congratulate the authors on this interesting and thought-provoking paper which presents a potentially widely applicable and practically important contribution to the field of Bayesian inference. It has been clear from the outset, even to Monte Carlo specialists, that whenever it is possible to avoid the use of Monte Carlo methods it is desirable to do so (Trotter and Tukey, 1956). This paper expands the class of problems which can be addressed without recourse to such techniques.

I have two comments.

(a) An exciting aspect of this work is the impressive performance which is obtained when $p/n_d \sim 1$. In their setting, Shen and McCullagh (1995) suggested that a Laplace approximation can be expected to be reliable (asymptotically) if p is $o(n_d^{1/3})$ and that *peculiar characteristics of individual problems* must be exploited outside this regime. The integrated nested Laplace approximation (INLA) can clearly perform well here. Further investigation seems warranted: a clear theoretical characterization would provide confidence in the robustness of the INLA approach; additionally, INLA may provide some insight into what properties are required for Laplace approximation techniques to be justified in this type of problem, perhaps using novel arguments, and this may be more generally important.

(b) Hidden Markov models, such as the stochastic volatility model of Section 5.3, in which an unobserved, discrete time Markov process $\{X_n\}$ is imperfectly observed through a second process $\{Y_n\}$ such that, conditional on X_n, Y_n is independent of the remainder of X and Y, are common in time series analysis. A small number of unknown parameters are often present. It is challenging to design good Markov chain Monte Carlo algorithms to perform inference in such models. Inference for these models is often restricted to the estimation of low dimensional marginals, often using sequential Monte Carlo (Doucet *et al.*, 2001) methods. Sometimes this choice is due to a requirement that computation can be performed in an 'on-line' fashion (and one might ask whether it is possible to develop 'sequential' versions of INLA); in other cases it is simply an efficient computational technique—see, for example, Gaetan and Grigoletto (2004). Smoothing (the problem of estimating the marginal posterior distribution of $X_k, k \leqslant t$, given $y_{1:t}$) as well as parameter estimation as touched on in Section 5.3 (which is non-trivial with sequential Monte Carlo methods and computationally intensive with Markov chain Monte Carlo methods) are of considerable interest. Existing analytic techniques are very limited: this seems to be a natural application for INLA.

Sujit K. Sahu (*University of Southampton*)

In this very impressive and stimulating paper it is nice to see the triumph of integrated nested Laplace approximations for making Bayesian inference. The paper develops easy-to-implement and ready-to-use software routines for approximations as alternatives to Markov chain Monte Carlo (MCMC) methods for many relatively low dimensional Bayesian inference problems. Thus, these methods have the potential

to provide an independent check for the MCMC code written for at least a simple version of the more complex general problem. However, the conditional independence assumption of the latent Gaussian field in Section 1.3 seems to limit the use of the methods in many space–time data modelling problems. The following example shows that it may be possible to relax the assumption.

Suppose that

$$\mathbf{Y}|\mathbf{x}, \boldsymbol{\theta} \sim \mathbf{N}(\mathbf{Z}\boldsymbol{\theta} + \mathbf{x}, \mathbf{Q}_{\mathbf{y}|\mathbf{x}}^{-1})$$

where $\mathbf{Z}$ is the design matrix and $\mathbf{Q}_{\mathbf{y}|\mathbf{x}}^{-1}$ is a known covariance matrix. Assume further that the latent field $\mathbf{x}$ follows the multivariate normal distribution with zero mean and precision matrix $\mathbf{Q}_{\mathrm{x}}$. A flat prior on $\boldsymbol{\theta}$ completes the model specification. As in Section 1.3, we have

$$\begin{aligned}\pi(\mathbf{x}, \boldsymbol{\theta}|\mathbf{y}) &\propto \exp[-\tfrac{1}{2}\{\mathbf{x}^{\mathrm{T}}\mathbf{Q}_{\mathbf{x}}\mathbf{x} + (\mathbf{y} - \mathbf{Z}\boldsymbol{\theta} - \mathbf{x})^{\mathrm{T}}\mathbf{Q}_{\mathbf{y}|\mathbf{x}}(\mathbf{y} - \mathbf{Z}\boldsymbol{\theta} - \mathbf{x})\}] \\ &= \exp[-\tfrac{1}{2}\{\mathbf{x}^{\mathrm{T}}(\mathbf{Q}_{\mathbf{y}|\mathbf{x}} + \mathbf{Q}_{\mathbf{x}})\mathbf{x} - 2\mathbf{x}^{\mathrm{T}}\mathbf{Q}_{\mathbf{y}|\mathbf{x}}(\mathbf{y} - \mathbf{Z}\boldsymbol{\theta}) + (\mathbf{y} - \mathbf{Z}\boldsymbol{\theta})^{\mathrm{T}}\mathbf{Q}_{\mathbf{y}|\mathbf{x}}(\mathbf{y} - \mathbf{Z}\boldsymbol{\theta})\}].\end{aligned}$$

From this we see that

$$\mathbf{x}|\boldsymbol{\theta} \sim \mathbf{N}\{(\mathbf{Q}_{\mathbf{y}|\mathbf{x}} + \mathbf{Q}_{\mathbf{x}})^{-1}\mathbf{Q}_{\mathbf{y}|\mathbf{x}}(\mathbf{y} - \mathbf{Z}\boldsymbol{\theta}), (\mathbf{Q}_{\mathbf{y}|\mathbf{x}} + \mathbf{Q}_{\mathbf{x}})^{-1}\}.$$

Thus, there is no need to obtain a Gaussian approximation, and we have

$$\mathbf{x}^*(\boldsymbol{\theta}) = (\mathbf{Q}_{\mathbf{y}|\mathbf{x}} + \mathbf{Q}_{\mathbf{x}})^{-1}\mathbf{Q}_{\mathbf{y}|\mathbf{x}}(\mathbf{y} - \mathbf{Z}\boldsymbol{\theta}).$$

Because of this exact result $\tilde{\pi}(\mathbf{x}|\boldsymbol{\theta}, \mathbf{y})$ at $\mathbf{x} = \mathbf{x}^*(\boldsymbol{\theta})$ will be free of $\boldsymbol{\theta}$ and according to equation (3) of the paper

$$\tilde{\pi}(\boldsymbol{\theta}|\mathbf{y}) \propto \exp[-\tfrac{1}{2}(\mathbf{y} - \mathbf{Z}\boldsymbol{\theta})^{\mathrm{T}}\{\mathbf{Q}_{\mathbf{y}|\mathbf{x}} - \mathbf{Q}_{\mathbf{y}|\mathbf{x}}(\mathbf{Q}_{\mathbf{y}|\mathbf{x}} + \mathbf{Q}_{\mathbf{x}})^{-1}\mathbf{Q}_{\mathbf{y}|\mathbf{x}}\}(\mathbf{y} - \mathbf{Z}\boldsymbol{\theta})],$$

which implies that

$$\tilde{\pi}(\boldsymbol{\theta}|\mathbf{y}) = \mathbf{N}\{(\mathbf{Z}^{\mathrm{T}}\mathbf{A}\mathbf{Z})^{-1}\mathbf{Z}^{\mathrm{T}}\mathbf{A}\mathbf{y}, (\mathbf{Z}^{\mathrm{T}}\mathbf{A}\mathbf{Z})^{-1}\} \tag{33}$$

where

$$\mathbf{A} = \mathbf{Q}_{\mathbf{y}|\mathbf{x}} - \mathbf{Q}_{\mathbf{y}|\mathbf{x}}(\mathbf{Q}_{\mathbf{y}|\mathbf{x}} + \mathbf{Q}_{\mathbf{x}})^{-1}\mathbf{Q}_{\mathbf{y}|\mathbf{x}},$$

again requiring no approximations.

The marginal model here is given by

$$\mathbf{Y}|\boldsymbol{\theta} \sim \mathbf{N}(\mathbf{Z}\boldsymbol{\theta}, \mathbf{Q}_{\mathbf{y}}^{-1} = \mathbf{Q}_{\mathbf{y}|\mathbf{x}}^{-1} + \mathbf{Q}_{\mathbf{x}}^{-1}),$$

with the exact posterior distribution given by

$$\pi(\boldsymbol{\theta}|\mathbf{y}) = \mathbf{N}\{(\mathbf{Z}^{\mathrm{T}}\mathbf{Q}_{\mathbf{y}}\mathbf{Z})^{-1}\mathbf{Z}^{\mathrm{T}}\mathbf{Q}_{\mathbf{y}}, (\mathbf{Z}^{\mathrm{T}}\mathbf{Q}_{\mathbf{y}}\mathbf{Z})^{-1}\}. \tag{34}$$

The nested approximation $\tilde{\pi}(\boldsymbol{\theta}|\mathbf{y})$ in equation (33) and the exact marginalization $\tilde{\pi}(\boldsymbol{\theta}|\mathbf{y})$ in equation (34) are in fact identical since it can be shown that $\mathbf{A} = \mathbf{Q}_{\mathbf{y}}$. Thus we see that the proposed nested approximation method, INLA, works even for spatially coloured covariance matrices for the random field as well as the data. Thus, is it possible to relax the conditional independence assumption in general? In practice, however, being a spatially coloured matrix, $\mathbf{Q}_{\mathbf{x}}$ will depend on several unknown parameters describing smoothness and rate of spatial decay. It is likely that Gaussian approximations for the posterior distributions of those parameters will be more challenging, since MCMC sampling algorithms often behave very poorly owing to weak identifiability; see for example Sahu *et al.* (2007).

Roberto Casarin and Christian P. Robert (*Université Paris Dauphine*)

To evaluate the effect of the Gaussian approximation on the marginal posterior on $\boldsymbol{\theta}$, we consider here a slightly different albeit standard stochastic volatility model

$$\mathbf{x}, \mathbf{y}|\boldsymbol{\theta} \sim \frac{\sigma^{-T-1}}{\sqrt{(1-\rho^2)}} \exp\left(-\frac{1}{2}\left[\frac{x_0^2}{1-\rho^2} + \sum_{t=1}^{T}(x_t - \rho x_{t-1})^2 + \sum_{t=0}^{T}\left\{y_t^2 \exp(-x_t)\sigma^{-2} + x_t\right\}\right]\right).$$

(The difference from the authors' expression is that the variance of the x_ts is set to 1 and that we use the notations ρ instead of ϕ and σ^2 instead of $\exp(\mu)$.) If we look at the second-order approximation of the non-linear term, we have

$$y_t^2 \exp(-x_t)\sigma^{-2} + x_t \approx y_t^2 \exp(-x_t^*)\sigma^{-2} + x_t^* + \frac{y_t^2 \exp(-x_t^*)}{2\sigma^2}(x_t - x_t^*)^2 = 1 + x_t^* + \frac{1}{2}(x_t - x_t^*)^2$$

where $x_t^* = \log(y_t^2, \sigma^2)$. A Gaussian approximation to the stochastic volatility model is thus

$$\mathbf{x}|\mathbf{y}, \boldsymbol{\theta} \sim |\mathbf{Q}(\boldsymbol{\theta})|^{-1/2} \exp\left[-\frac{1}{2}\left\{\frac{x_0^2}{1-\rho^2} + \sum_{t=1}^{T}(x_t - \rho x_{t-1})^2 + \frac{1}{2}\sum_{t=0}^{T}\sigma^{-2}(x_t - x_t^*)^2\right\}\right],$$

where the Gaussian precision matrix $\mathbf{Q}(\boldsymbol{\theta})^{-1}$ has $3/2 + \rho^2$ on its diagonal, $-\rho$ on its first subdiagonal and sup-diagonals, and 0 elsewhere. Therefore the approximation (3) of the marginal posterior of $\boldsymbol{\theta}$ is equal to

$$\begin{aligned}\tilde{\pi}(\boldsymbol{\theta}|\mathbf{y}) &\propto \frac{\pi(\mathbf{x}, \boldsymbol{\theta}|\mathbf{y})}{\pi_{\mathrm{G}}(\mathbf{x}|\boldsymbol{\theta}, \mathbf{y})} \\ &\propto \frac{\sigma^{-T-1}|\mathbf{Q}(\boldsymbol{\theta})|^{1/2}}{\sqrt{(1-\rho^2)}} \exp\left[-\frac{1}{2}\sum_{t=0}^{T}\left\{y_t^2 \exp(-x_t)\sigma^{-2} + x_t - \frac{1}{2}(x_t - x_t^*)^2\right\}\right]\pi(\theta),\end{aligned}$$

for a specific plug-in value of $\mathbf{x}$.

Using for this plug-in value the mode (and mean) $\mathbf{x}^{\mathrm{M}}$ of the Gaussian approximation, as it is readily available, in contrast with the mode of the full conditional of $\mathbf{x}$ given $\mathbf{y}$ and $\boldsymbol{\theta}$ as suggested by the authors, we obtain a straightforward recurrence relation on the components of $\mathbf{x}^{\mathrm{M}}$

$$-(\rho x_{t+1}^{\mathrm{M}} - \rho x_t^{\mathrm{M}}) + x_t^{\mathrm{M}} - \rho x_{t-1}^{\mathrm{M}} + \tfrac{1}{2}(x_t^{\mathrm{M}} - x_t^*) = 0,$$

with appropriate modifications for $t = 0, T$. We thus obtain the recurrence ($t > 0$)

$$x_t^{\mathrm{M}} = \alpha_t x_{t-1}^{\mathrm{M}} + \beta_t,$$

with

$$\begin{cases} \alpha_T = 2\rho/3, \\ \beta_T = x_T^*/3, \\ \alpha_t = \dfrac{\rho}{3/2 + \rho^2 - \rho\alpha_{t+1}}, \\ \beta_t = \dfrac{\rho\beta_{t+1} + x_t^*/2}{3/2 + \rho^2 - \rho\alpha_{t+1}}, \qquad 1 \leqslant t < T, \end{cases}$$

and

$$x_0^{\mathrm{M}} = \frac{\rho\beta_1 + x_t^*/2}{(1-\rho^2)^{-1} + \rho^2 - \alpha_1\rho + \frac{1}{2}}.$$

This choice of $\mathbf{x}^{\mathrm{M}}$ as a plug-in value for the approximation to $\pi(\theta|y)$ gives quite accurate results, when compared with the 'true' likelihood that is obtained by a regular (and unrealistic) importance sampling approximation. Fig. 11 shows the correspondence between both approximations, indicating that the Gaussian approximation (3) can be used as a good proxy to the true marginal.

Jean-Michel Marin (*Université Montpellier 2 and Centre de Recherche en Economie et Statistique, Malakoff*), **Roberto Casarin** (*University of Brescia*) **and Christian P. Robert** (*Université Paris Dauphine and Centre de Recherche en Economie et Statistique, Malakoff*)

The authors state that for a given computational cost the use of Markov chain Monte Carlo (MCMC) sampling for the latent Gaussian models rarely makes sense in practice. We wish to argue against this point: MCMC sampling, when combined with other Monte Carlo techniques, can still be an efficient methodology.

To assess this point, we consider the same univariate stochastic volatility model as that studied by the authors in Section 5.3. We assume that $\mu = 0$, $\beta^2 = 1/\tau$ and $f_0 \sim \mathcal{N}\{0, \beta^2/(1-\phi^2)\}$. For $\boldsymbol{\theta} = (\phi, \beta^2)$, we use a uniform prior distribution on $\Theta = [0, 1] \times [0, 1]$.

We combine sequential Monte Carlo and Metropolis–Hastings (MH) algorithms. Using standard particle filters, we generate a weighted random sample of size N: $S_t^N = \{\mathbf{z}_t^i, w_t^i\}_{i=1}^N$ where $\mathbf{z}_t^i = (f_0, f_1, \ldots, f_t, \boldsymbol{\theta})^i$. This weighted random sample is targeting $\pi_t(\mathbf{z}_t|\mathbf{y}_t)$ where $\mathbf{y}_t = (y_1, \ldots, y_t)$ is the observed process. At time t, given $\mathcal{S}_{t-1}^N$, we obtain a new set $\mathcal{S}_t^N$ by applying sequentially a propagation and a self-avoiding MH step.

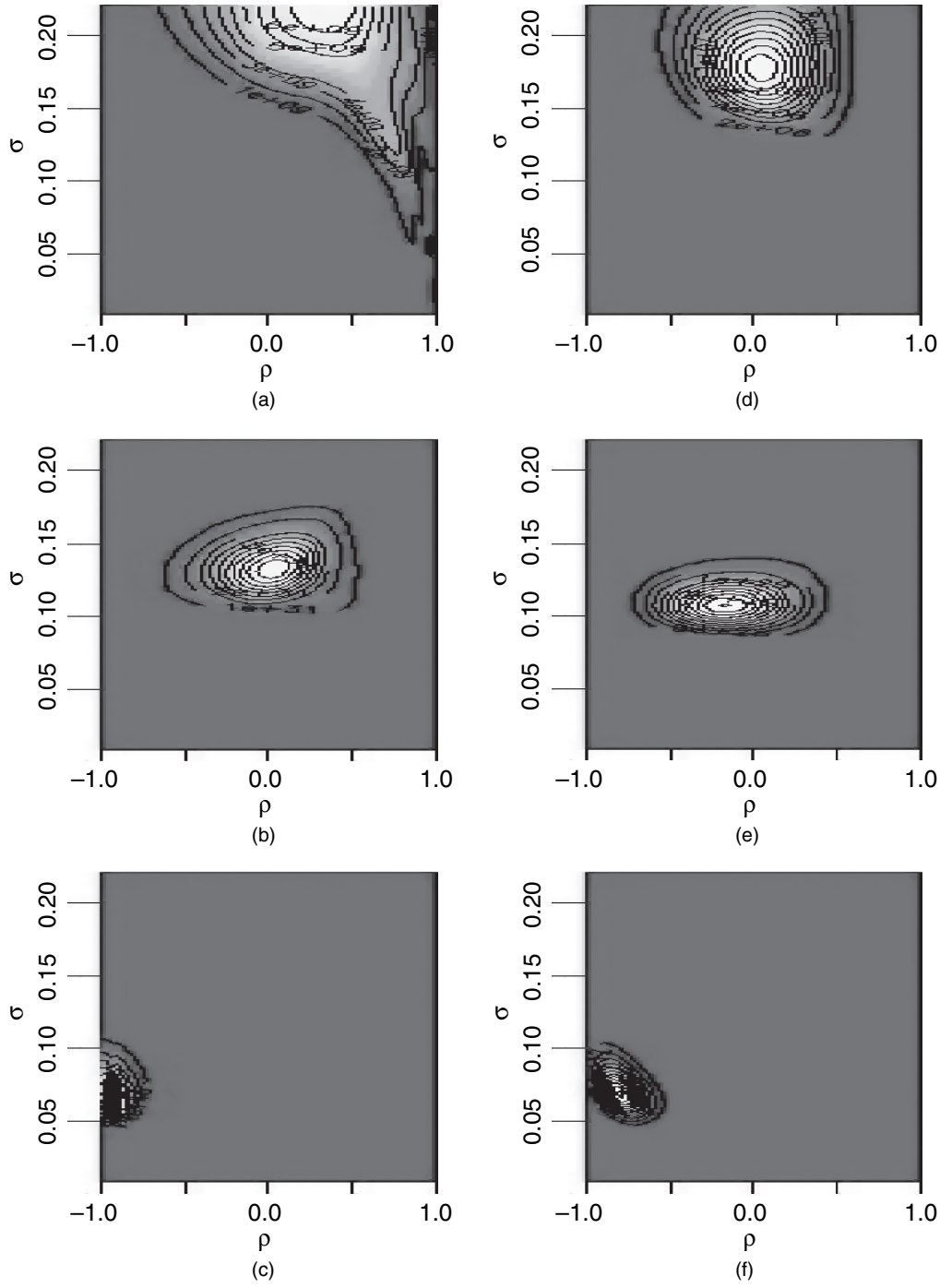

Fig. 11. Comparison of an importance sampling approximation (with 10^3 simulations) to the likelihood of a stochastic volatility model (a)–(c) with the approximation based on the Gaussian approximation of Rue and his colleagues and (d)–(f) when centred at $\mathbf{x}^{\mathrm{M}}$, the mode of the Gaussian approximation: this likelihood is associated with (a), (d) 25 simulated values with $\sigma = 0.1$ and $\rho = 0.9$, (b), (e) 50 simulated values with $\sigma = 0.1$ and $\rho = -0.3$ and (c), (f) 20 simulated values with $\sigma = 0.1$ and $\rho = -0.9$

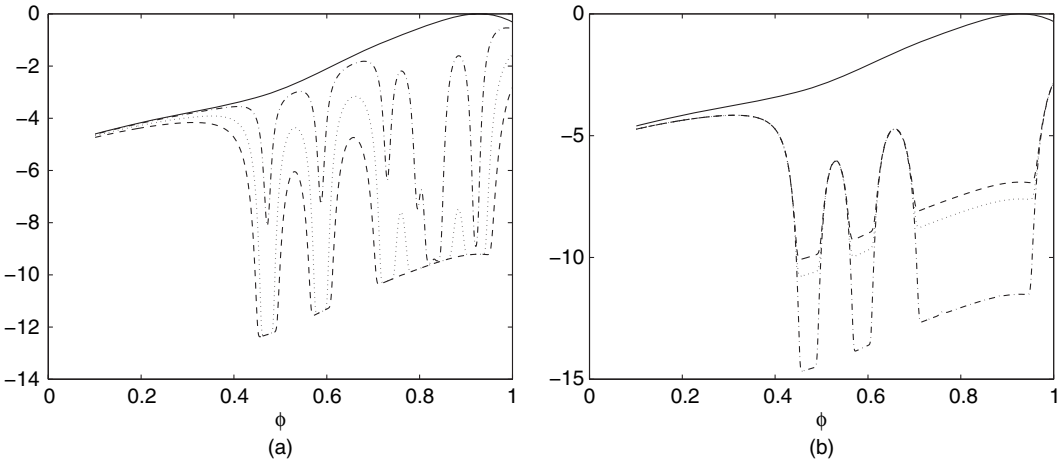

Fig. 12. Illustrative example ($N = 10$ particles) of the effect of (a) ξ_t ($\alpha = 10^{-4}$; ———, $\pi(\phi|\beta^2)$; - - - - - -, $\pi^{\mathrm{R}}(\phi|\beta^2)$, $\xi = 10^{-1}$; · · · · · · ·, $\pi^{\mathrm{R}}(\phi|\beta^2)$, $\xi = 10^{-2}$; · - · - · - ·, $\pi^{\mathrm{R}}(\phi|\beta^2)$, $\xi = 10^{-3}$) and (b) α_t ($\xi = 10^{-2}$; ———, $\pi(\phi|\beta^2)$, $\alpha = 1$; – – –, $\pi^{\mathrm{R}}(\phi|\beta^2)$, $\alpha = 10^{-2}$; · · · · · · ·, $\pi^{\mathrm{R}}(\phi|\beta^2)$, $\alpha = 10^{-3}$; · - · - · - ·, $\pi^{\mathrm{R}}(\phi|\beta^2)$, $\alpha = 10^{-4}$) on the density $\log\{\pi_t^{\mathrm{R}}(\phi|\beta^2)\}$

The self-avoiding MH step is applied particlewise to the particle set considered as a whole, $\mathbf{Z}_t = (\mathbf{z}_t^1, \ldots, \mathbf{z}_t^N)$ and is targeting the modified N-product of the posterior

$$\pi_t^{\mathrm{R}}(\mathbf{Z}_t|\mathbf{y}_t) \propto \prod_{k=1}^{N} \pi_t(\mathbf{z}_t^k|\mathbf{y}_t)\left\{\alpha_t + (1-\alpha_t)\exp\left(-\sum_{j\neq k}\frac{\xi_t}{\|\mathbf{z}_t^k - \mathbf{z}_t^j\|^2}\right)\right\}$$

(Mengersen and Robert, 2003). The second term in the product forces the particles to explore the posterior. Fig. 12 presents the way that this term acts on a part of the posterior distribution. Increasing ξ_t induces a heavier mass reduction in correspondence to the simulated values. Increasing α_t reduces the repulsion. After the move step from $\mathbf{z}_t^i$ to $\tilde{\mathbf{z}}_t^i$, the weights are updated as follows:

$$\tilde{w}_t^i \propto w_t^i \frac{\pi_{i,t}^{\mathrm{R}}(\tilde{\mathbf{z}}_t^i|\mathbf{y}_t)}{\pi_t(\tilde{\mathbf{z}}_t^i|\mathbf{y}_t)} \frac{\pi_t(\tilde{\mathbf{z}}_t^i|\mathbf{y}_t)}{\pi_{i,t}^{\mathrm{R}}(\tilde{\mathbf{z}}_t^i|\mathbf{y}_t)}$$

with

$$\pi_{k,t}^{\mathrm{R}}(\mathbf{z}_t^k|\mathbf{y}_t) \propto \pi_t(\mathbf{z}_t^k|\mathbf{y}_t)\left\{\alpha_t + (1-\alpha_t)\exp\left(-\sum_{j\neq k}\frac{\xi_t}{\|\mathbf{z}_t^k - \mathbf{z}_t^j\|^2}\right)\right\}.$$

Fig. 13 relates to the exploration of the parameter space and shows that a strong repulsion factor, when compared with a weaker repulsion, causes a higher number of the particles to explore the tails of $\pi_t(\boldsymbol{\theta}|\mathbf{y}_t)$.

In contrast with the two-step deterministic procedure that is described in the paper, a Monte Carlo hybrid algorithm can thus address simultaneously the estimation and the parameter space exploration problems. The algorithm proposed produces samples from a sequence of distributions and is thus suitable for on-line applications. Particularly welcome would be a future extension of the Laplace approximation framework in the context of sequential inference problems.

Sylvia Richardson (*Imperial College London*) **and Arnoldo Frigessi** (*University of Oslo*)
Given the identity

$$\pi(\theta|y) = \frac{\pi(y|x,\theta)\,\pi(x|\theta)\,\pi(\theta)\{1/\pi(y)\}}{\exp[-\frac{1}{2}x^{\mathrm{T}}Q_x + \sum_i \log\{\pi(y_i|x_i,\theta)\}](1/Z_{\theta,y})}$$

the strategy in this paper is to produce an approximating model for which the normalizing function is easy to compute as the square-root determinant of Q plus a diagonal matrix. Alternative approaches have

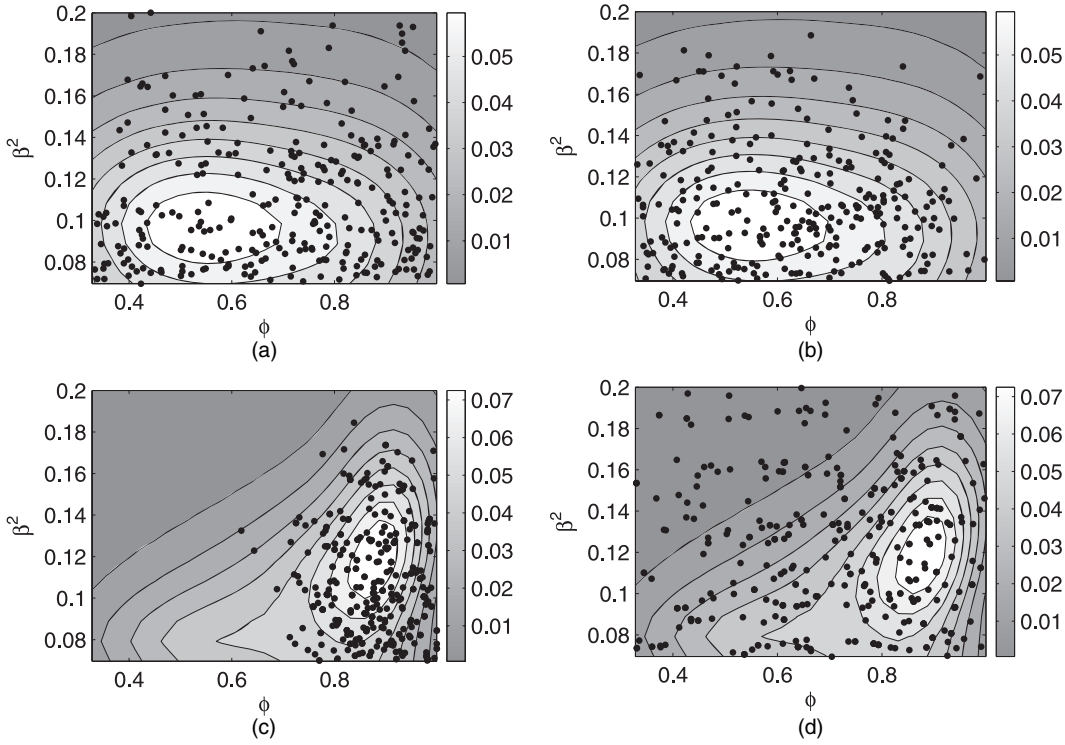

Fig. 13. Level sets $\pi_t(\boldsymbol{\theta}|\mathbf{y}_t)$ at (a), (b) time $t = 26$ and (c), (d) time $t = 50$, approximated by a Monte Carlo integration of $(f_0,\ldots,f_t)$ (we generated $(y_1,\ldots,y_{50})$ with $\phi = 0.9$ and $\beta^2 = 0.1$; the factors ξ_t and α_t can be adapted over iterations to address both exploration and estimation issues): •, particle set ($N = 500$) after the self-avoiding move, with (a), (c) $\alpha_t = 0.95$ and $\xi_t = 0.01$ and (b), (d) $\alpha_t = 0.1$ and $\xi_t = 0.1$

aimed at approximating the partition function $Z_{\theta,y}$ directly, starting with Geyer and Thompson (1992). Have the authors given any thought to understanding the link between the two types of approximation?

Given $\tilde{\pi}_{\mathrm{G}}(x|\theta, y)$ and $\tilde{\pi}(\theta|y)$ as defined in approximation (3), define $\tilde{\pi}(x, \theta|y) \propto \tilde{\pi}_{\mathrm{G}}(x|\theta, y)\,\tilde{\pi}(\theta|y)$ and rewrite this as

$$\tilde{\pi}(x, \theta|y) \propto \tilde{L}(y|\theta, x)\,\pi(x|\theta)\,\pi(\theta).$$

In this form, the approximation that is used by the authors is equivalent to an implicit approximation $\tilde{L}(y|\theta, x)$ for the likelihood term of the original model. Could the authors work out which approximate likelihood is used for some examples? Since all parameters retain their original interpretation, could we simply use this approximate (non-Gaussian) model and forget the original model?

The emphasis is put on evaluating *marginal* posterior distributions in latent Gaussian models in a range of examples with few hyperparameters. You have demonstrated that the method that you propose is computationally efficient, very accurate and beneficial in this respect in comparison with alternative standard implementations of Markov chain Monte Carlo (MCMC) algorithms. The *generic nature* of your computational approach is less obvious. MCMC sampling is (much) easier to implement and extend; for example, if we wanted to depart from the latent Gaussian Markov random field, the integrated nested Laplace approximation (INLA) involves complex numerical optimization and programming, and users will mostly rely on existing software, rather than develop their own extensions. INLA requires the user to bring his model into the form of equation (1). This might not be easy, and reparameterization may be necessary, whereas MCMC methods would work with the natural parameterization. We have also experienced some difficulties with using INLA with 3000 latent variables, which could indicate that the optimization requires much random-access memory storage.

One crucial benefit of MCMC algorithms is their capacity to compute easily *any joint posterior distribution* for parameters of interest. In Section 6.1, you discussed the possibility of obtaining bivariate (and

low dimensional multivariate) posterior distributions by using INLA-generated posterior marginals and a copula-like approximation for the joint distribution. If this works well, this would be an important step towards embedding INLA within more general hierarchical structures and to retain inferential flexibility that is typical of MCMC outputs.

Omiros Papaspiliopoulos (*Universitat Pompeu Fabra, Barcelona*)

I congratulate the authors for an excellent paper. Contrary to what is hinted in Section 7 I think that the paper shares certain principles with Beskos *et al.* (2006) by devising efficient computational methods for a particular class of models as opposed to appealing to generic Monte Carlo methods. The efficiency comes at the cost of it not being straightforward to use the methodology outside the class.

I have two comments on the methodology: first on the use of the ratio identity for the marginal likelihood

$$p(y|\theta) = \frac{p(y, x|\theta)}{p(x|y, \theta)} \tag{35}$$

as opposed to the integral identity

$$p(y|\theta) = \int p(y|x|\theta)\, p(x|\theta)\, \mathrm{d}x. \tag{36}$$

Equation (36) has been intensively used in conjunction with importance sampling. Here the authors approximate with a Gaussian density the denominator and use equation (35) to approximate $p(y|\theta)$ 'semi-parametrically'. It is interesting to understand the full potential of equation (35). In some cases it can be used for the *exact* computation of $p(y|\theta)$, although this is not obvious from equation (36). An example is the so-called Matérn hard core process (Matérn, 1986), which is obtained as a thinning of an initial Poisson marked point process. A point of the process is selected if its mark is larger than all marks in a radius of range $\delta > 0$. Here $x \cup y$ are the original marked points, y are the selected points and θ are intensity parameters.

An alternative use of equation (35) was proposed in Beskos *et al.* (2006) for the simultaneous *unbiased estimation* of $p(y|\theta)$. Suppose that we can simulate from $p(x|y, \theta)$ by using rejection sampling with proposal $q(x|y)$; hence

$$\sup_x \left\{ \frac{p(y, x|\theta)}{q(x|y)} \right\} = K(y, \theta) < \infty.$$

Then we have the identity

$$p(y|\theta) = K(y, \theta)\, \alpha(y, \theta)$$

where $\alpha(y, \theta)$ is the acceptance probability of the algorithm and can be estimated unbiasedly with simulations. In Beskos *et al.* (2006) the simulation was performed without even computing $p(y, x|\theta)$ by using retrospective sampling techniques.

My second comment relates to the application of the proposed methodology to sequential estimation. Suppose concretely that we deal with state space models with Gaussian state dynamics and data $y_{1:T}$ (the stochastic volatility example that the authors consider is such an instance). The approach of the paper directly covers the problem of computing $p(y_{1:T}|\theta)$. Suppose instead that we wish to compute $p(y_{1:t}|\theta)$, as t varies from 1 to T. It seems to me that in general the Laplace approximation would require a computational cost of order T^2 to achieve this. I would generally be very interested in the authors' perspectives on how the integrated nested Laplace approximation fits in with problems in estimation of state space models.

The following contributions were received in writing after the meeting.

Philos Kjersti Aas (*Norwegian Computing Center, Oslo*)

I congratulate the authors on presenting us with a most excellent and interesting paper. I am most interested in the possibilities that integrated nested Laplace approximations (INLAs) may open in this field of financial econometrics. Volatility in financial time series is mainly analysed through two classes of models; generalized auto-regressive conditional heteroscedastic (GARCH) models (Bollerslev, 1986) and stochastic volatility (SV) models (Taylor, 1982). Whereas GARCH models are relatively straightforward to estimate by using maximum likelihood techniques, the likelihood function in SV models does not have a closed form, Hence, they require more complex inferential and computational tools. Much attention has been devoted to the development of efficient Markov chain Monte Carlo (MCMC) algorithms for SV

models; see for example Shephard and Pitt (1997) and Chib *et al.* (2002). However, the latent field and strong correlation structures which are often found in SV models make even well-constructed MCMC algorithms slow and their convergence dubious to assure. Hence, even though SV models in general are recognized to be more flexible than GARCH models (Kim *et al.*, 1998), the latter are still by far the most popular in terms of real life applications.

It is my opinion that the INLA approach that is suggested by Rue and his colleagues may help SV models to exit the academic world and to reach the financial industry. The most basic SV model can be written as

$$r_t = \exp(h_t/2)\varepsilon_t, \tag{37}$$

$$h_t = \nu + \phi(h_{t-1} - \nu) + \sigma\eta_t, \tag{38}$$

where r_t and h_t are the logarithmic return and log-variance at time t respectively, ε_t and η_t are independent and identically distributed $N(0, 1)$ and ν, ϕ and σ are parameters to be estimated. Hence, this model belongs to the class of latent Gaussian models, for which Rue and his colleagues have shown that MCMC calculations can be substituted by accurate deterministic approximations. One may therefore obtain accurate approximations for the posterior marginals of the latent log-variances as well as of the model parameters in a fraction of the time that is used by well-designed MCMC algorithms.

In financial applications, a Gaussian distribution for ε_t usually is too restrictive. Using the INLA framework, the Gaussian distribution may be replaced by heavy-tailed and skew distributions like the Student or the normal inverse Gaussian distribution (Barndorff-Nielsen, 1997) without any additional computational cost. Another feature that is often observed for financial data is that volatility responds asymmetrically to positive and negative return shocks (so-called leverage effects). Leverage effects may be incorporated in SV models by letting the noise terms ε_t and η_t be correlated (Harvey and Shephard, 1996). Does the INLA approach allow for performing approximate inference even for this kind of model?

Sudipto Banerjee (*University of Minnesota, Minneapolis*)

I congratulate the authors for a fine paper on Bayesian computation for a very general class of models. The integrated nested Laplace approximations (INLAs) that are developed here can potentially abate the computational concerns that are associated most conspicuously with Bayesian methods. The benefits of these 'fast and accurate' approximations will be more pronounced for non-Gaussian likelihoods, where marginalization of the latent effects does not render easily tractable distributions and Markov chain Monte Carlo sampling suffers from high auto-correlations and slow convergence.

However, the conditional independence assumption of the latent effects considerably restricts its scope. The INLA uses conditional independence to achieve a sparse $\mathbf{Q}(\boldsymbol{\theta})$ in expression (9), which is crucial in speeding up computations. Yet, this assumption precludes several more challenging Bayesian models. In particular, I am dubious about the effectiveness of the INLA with the spatial and spatiotemporal models that are alluded to in part (c) of Section 1.2—especially for high dimensional spatial models. Here the random effects often arise as realizations of a Gaussian process with a correlation function and, in general, the Markovian dependence is lost. Relaxing the assumption of conditional independence would significantly detract from the computational benefits of the INLA and $\pi(\mathbf{x})$ will now need to be *approximated* by a Gaussian Markov random field. Such approximations have been explored elsewhere (Rue and Tjelmeland, 2002; Rue and Held, 2005), but this approach is best suited for spatial locations on a regular grid. With irregular locations, realignment to a grid or a torus is required, which is done by an algorithm, possibly introducing unquantifiable errors in precision. Adapting these approaches to richer hierarchical spatial models is potentially problematic.

Another brief remark: the INLA does not deliver posterior samples. The power of Bayesian analysis lies in the vast array of model assessment and model choice techniques facilitated by sampling-based inference. For instance, the posterior samples from θ will immediately yield the posterior of $g(\boldsymbol{\theta})$ for a function $g(\cdot)$. Although the authors have discussed computing posterior predictive distributions and model choice measures such as the deviance information criterion, these (and many others) may not be as easily obtained as from Markov chain Monte Carlo sampling.

Finally, the `GMRFLib` C libraries have been useful to me in my research and I welcome the proposal to translate this technology into an R package. This will considerably enhance its accessibility to researchers in diverse fields and only then will the INLA be put to its true test.

George Casella (*University of Florida, Gainesville*)
It was a pleasure to hear the presentation of the authors, and I join in congratulating them for a stimulating paper and a new approach to the problem. In this discussion I would like to point out one fact about the approximation that should be noted. The approximation in equation (3) starts from the fact that, from the calculus of probabilities (Bayes's rule),

$$\pi(\theta|\mathbf{y}) = \frac{\pi(\mathbf{x}, \theta, \mathbf{y})}{\pi(\mathbf{x}|\theta, \mathbf{y})\,\pi(\mathbf{y})},$$

but ignores the factor $\pi(\mathbf{y})$, which the authors call the normalizing constant. Hence, approximation (3) suffers from both an approximation error (due to the Laplace approximation) and the error from ignoring $\pi(\mathbf{y})$, the marginal distribution of $\mathbf{y}$. Conditions under which the error in the Laplace approximation will improve, as discussed by the authors in Section 4, will not fix the normalizing constant, which can only be taken care of by renormalization. As an example, consider a very special case of equation (l) where

$$\begin{aligned} Y &\sim N(X, \sigma^2 I), \\ X &\sim N(\mathbf{1}\theta, r^2 I), \\ \theta &\sim N(0, \delta^2), \end{aligned}$$

where Y and X are p-dimensional vectors, $\mathbf{1}$ is a vector of 1s and θ is a scalar. In this case all approximations are exact, and the correct $\pi(\mathbf{x}|\mathbf{y})$, using equation (5) exactly, is

$$\pi(\mathbf{x}|\mathbf{y}) = N\left(\mathbf{x}; \frac{\tau^2}{\sigma^2+\tau^2} A^{-1} Y, \frac{\tau^2\sigma^2}{\sigma^2+\tau^2} A^{-1}\right),$$

with $A = I - \delta^2\sigma^2 \mathbf{1}\mathbf{1}'/\{(\sigma^2+\tau^2)(p\delta^2+\tau^2)\}$. However, if equation (5) is used with the approximation (3), we obtain an expression for $\tilde{\pi}(\mathbf{x}|\mathbf{y})$, the approximation, to be

$$\tilde{\pi}(\mathbf{x}|\mathbf{y}) = N\left(\mathbf{x}; \frac{\tau^2}{\sigma^2+\tau^2} A^{-1} Y, \frac{\sigma^2\tau^2}{\sigma^2+\tau^2} A^{-1}\right) N\left\{Y; 0, \sigma^2\left(I - \frac{\tau^2}{\sigma^2+\tau^2} A^{-1}\right)^{-1}\right\}.$$

Thus, without renormalization there is potential for large errors, even if the Laplace approximation is accurate, so equation (5) should never be used without normalization. The authors, of course, know this, as it is discussed in Section 4.1. Moreover, in implementing the approximation in Section 3.2.1, they explicitly normalize (see the discussion after expression (17)), and in the implementation in Section 3.2.2 there is implicit normalization.

However, this discussion highlights another difference between the Markov chain Monte Carlo approach and the approximation approach. The more computationally intensive Markov chain Monte Carlo approach will often circumvent the calculation of the normalizing constant. But this must always be done in using the approximation, and computation of the normalizing constant can sometimes, in itself, be a difficult task.

Nikolaos Demiris (*Medical Research Council Biostatistics Unit, Cambridge*)
The authors are to be congratulated for an excellent paper that could enable the routine use of this wide class of models among practitioners. I have two questions that the authors may be able to shed some light on. From a practical perspective, interest often focuses on, possibly non-linear, functions (or functionals) of the parameters θ and/or x. Inference for such functions is typically trivial in a simulation-based approach such as Markov chain Monte Carlo sampling. Matters do not appear as straightforward in the new approach, and some generic guidance to the functions that may or may not be covered with methods based on integrated nested Laplace approximations would be particularly valuable. From a theoretical viewpoint, I wonder whether the authors considered a non-Bayesian likelihood-based approach based on their methods. The formula $\pi(\theta|y) = \pi(\theta, x|y)/\pi(x|\theta, y)$ may also be written as $\pi(y|\theta) = \pi(y, x|\theta)/\pi(x|\theta, y)$. In view of the similarity between Laplace approximations and the p^*-formula (Barndorff-Nielsen, 1983), it appears that there may be some scope for further exploration of this issue.

Richard Everitt (*University of Bristol*)
The authors are to be congratulated on their work for two reasons: firstly, for the interesting technical approach and, secondly, for highlighting why Monte Carlo methods should not always be the first resort

when tackling complex inference problems. As is mentioned in the paper, the use of an approximation to the true posterior has received much attention in machine learning but is largely out of favour in statistics. Although the theoretical guarantees regarding the Markov chain Monte Carlo errors are appealing, for practical applications there is certainly a place for approximation-based methods. Even if Monte Carlo algorithms are truly the only viable approaches, it seems likely that approximation-based methods could be used to improve the efficiency of these algorithms.

An important aspect of the integrated nested Laplace approximation approach is its generality. The ability to obtain accurate results easily for a wide variety of models is important in itself, but the fact that this can be achieved with minimal tuning is highly significant. The design time that is involved in an inference method is frequently overlooked but makes a large difference to the user of a technique.

As noted in the paper, the clear limitation of the method as it stands is in cases where the dimensionality of the hyperparameters is high. Do the authors believe that there is any scope to tackle this problem through making assumptions about the distributions or conditional independences of the hyperparameters?

It is interesting that in the discussion the authors mention the parallel implementation of the algorithm. Given the recent shift towards parallel computing in home and office computers (using either multicore processors or graphics cards) it seems that parallel implementation of algorithms is an important topic in computational statistics and should receive more attention than it does at present.

I look forward to seeing this method used to tackle applications in signal and image processing.

Ludwig Fahrmeir (*Ludwig-Maximilians-Universität, Munich*) **and Thomas Kneib** (*Georg-August-Universität, Göttingen, and Ludwig-Maximilians-Universität, Munich*)

Our first comment relates to intrinsic Gaussian Markov random field (GMRF) priors. Whereas the precision matrix Q is assumed to be non-singular in Sections 1 and 2, examples 5.4 and 5.5 employ intrinsic GMRF priors with singular Q. The difference between singular and regular prior precision may be ignorable for model estimation, but proper latent Gaussian models seem to be necessary to guarantee that marginal likelihoods and Bayes factors are well defined in Section 6.2. In contrast, the predictive measure (Section 6.3) and the deviance information criterion (Section 6.2) only require propriety of the posterior which, under regularity conditions, can be guaranteed for a wide range of structured additive regression models even when considering partially improper priors (Fahrmeir and Kneib, 2009). Generally the question arises, which situations require regular precision matrices Q?

The authors compare variants of the integrated nested Laplace approximation (INLA) procedure with fully Bayesian inference based on Markov chain Monte Carlo sampling. The simplest variant of INLA seems to be closely related to mixed-model-based inference which is based on Laplace approximations and restricted maximum likelihood estimation for smoothing parameters. Mixed-model-based inference in semiparametric regression has gained much attention in recent years, so it would be interesting to contrast it with the INLA procedure. Whereas Ruppert *et al.* (2003) or Kauermann (2005) took a frequentist perspective on mixed-model-based inference, it can also be interpreted as an empirical Bayes approach (Fahrmeir *et al.*, 2004; Kneib and Fahrmeir, 2006, 2007).

Our final comment relates to the extensibility of the INLA procedure. The general framework in Sections 1.3, 2.1 and 2.2 suggests that there is only one stage in the model hierarchy involving GMRFs. However, models with more than one latent Gaussian hierarchy may be of interest when local adaptivity of the smoothing parameter will be achieved (see for example Lang *et al.* (2002), Lang and Brezger (2004), Baladandayuthapani *et al.* (2005) and Crainiceanu *et al.* (2007) for adaptive regression function smoothing or Brezger *et al.* (2007) for adaptive surface smoothing). In random-walk smoothing, the variance could itself be assumed to follow a random-walk prior. Recently, Krivobokova *et al.* (2008) proposed an approach based on Laplace approximations for locally adaptive penalized spline estimation. Although presented in the spirit of penalized likelihood estimation, we speculate that their Laplace approximations could successfully be improved by adapting the INLA approach.

Marco A. R. Ferreira (*University of Missouri, Columbia*)

I congratulate Professor Rue and his colleagues for their important contribution to the development of fast and accurate computations for the analysis of latent Gaussian models. In addition, I applaud the authors for making available on the Internet open source software for the implementation of their methods.

Their methodology makes use of two properties that are usually valid for latent Gaussian models: first, the latent Gaussian process is a Markov random field; second, the hyperparameters vector is low dimensional. In addition, the methodology implicitly assumes that the mode of the full conditional

distribution for $\mathbf{x}$ for a given $\boldsymbol{\theta}$ and the mode of the approximate marginal posterior for $\boldsymbol{\theta}$ are unique. As these assumptions are satisfied by many of the latent Gaussian models that are currently used in practice, their methodology will bring substantial savings in terms of computational time for end-users.

However, I wonder what happens when the posterior distribution has multiple modes. In that case, quasi-Newton methods will find only one local mode; then all the computations will be based on that specific local mode. More critically, most likely the proposed method for assessing the approximation error will not be able to detect the inadequacy of the analysis: the simplified Laplace and the Laplace approximations based on the same specific local mode will probably have small symmetric Kullback–Leibler divergence. Thus, to assess how appropriate is the use of their method in a specific problem, it seems extremely important to verify the uniqueness of the posterior mode. Nevertheless, this is usually a daunting task for most highly complex latent Gaussian models.

Montserrat Fuentes (*North Carolina State University, Raleigh*)

I congratulate Rue, Martino and Chopin for bringing forward new critical approaches to perform approximate Bayesian inference for latent Gaussian models. My comments focus on some of the limitations of the proposed Laplace approximation methods in the context of making Bayesian inference for spatial or spatiotemporal data. Nowadays, statisticians are frequently involved in the spatial analysis of huge data sets. One of the main challenges when analysing continuous spatial processes and making Bayesian spatial inference is calculating the likelihood function of the covariance parameters. For large data sets, calculating the determinants that we have in the likelihood function can often be infeasible. Spectral methods could be used to approximate the likelihood and to obtain the maximum likelihood estimates of the covariance parameters (e.g. Fuentes (2007)). Stein *et al.* (2004) proposed another spatial likelihood approximation method to reduce the computation of Vecchia's (1988) approach and to improve the efficiency of the estimated covariance parameters. Banerjee *et al.* (2008) have introduced some methodology for rank reduction. One of the main constraints of the elegant methods proposed by the authors is that they rely on having the maximum likelihood estimates of the covariance or correlation parameters. Therefore, in most practical settings when working with large continuous spatial processes, their approach could not be implemented unless it is combined with some other approximation approach (like those mentioned above) to obtain those maximum likelihood estimates.

In a spatial setting, often the observations are spatially correlated even after conditioning on a latent spatial process. In those situations, the residual correlation can be introduced via a spatial (or spatiotemporal) copula approach. For instance, if the interest is in spatial extremes a generalized Pareto distribution could be used such that the parameters of this distribution vary according to a latent Gaussian spatial model capturing spatial dependence. However, it is likely that there is spatial dependence which is unexplained by the latent spatial specifications for the distribution parameters. In these common situations with observations that are not independent given the spatially varying parameters, if the number of observations is large, then the Laplace approximations would not facilitate the inference, because the main computational challenge is due to the copula and the large covariance matrices in the likelihood function. A spectral approximation to the likelihood of the spatial correlation parameters would again facilitate the computational burden.

The Laplace approximation would not work well in situations where a good representative of the probability mass is not a local maximum, which could be easily the case in spatial settings with complex nonstationary patterns.

Alan Gelfand (*Duke University, Durham*)

The authors are to be congratulated on a valuable contribution for Bayesian computation. This work shows the full maturation of effort that Rue and his colleagues have expended for nearly a decade. Over time has come an increased appreciation of the subtleties that arise in posterior inference for hierarchical models along with associated computational tricks to best effect approximation at various places. However, a less sophisticated user will struggle to appreciate quantitatively when more sophisticated approximations are needed, much less to assess their value for any particular application. Unless the authors' technology becomes a 'black box', I am sceptical of its widespread usage.

I was struck by the underlying presumption of a conditionally independent hierarchical model specification. It reminded me that, arguably, this setting was the most successful application of Laplace approximation in the 1980s (see, for example, Kass and Steffey (1989)). It also reminded me of the need to build a new approximation for each posterior of interest as well as the need for high dimensional posterior mode evaluation, both of which are still part of the authors' new technology.

My main point, as one who now works primarily with space and space–time data, is that I believe that the authors' approximate inference strategy will still suffer some of the problems that plague Markov chain Monte Carlo (MCMC) methods in fitting customary models for such data. Following the authors' setting, suppose a spatiotemporal process yielding a conceptual binary observation at an arbitrary location and time. As a first comment, conditional independence may be unsuitable here since even a latent Gaussian process model yielding smooth realizations for the binary probabilities need not reveal spatial pattern in the realized binary outcomes. But, suppose that we seek to learn about the parameters of the covariance function of the Gaussian process as well as to infer about the binary probabilities over space and time. It is well known that the variance of the Gaussian process model and the space and time decay parameters are weakly identified. Hence, MCMC implementations struggle. But these parameters fall into the authors' $\boldsymbol{\theta}$ and I would be concerned about how well $\tilde{\pi}(\boldsymbol{\theta}|\mathbf{y})$ works in this case. The key issue here is prior specification, with very informative priors needed for some parameters. On attending to this, we have found (Banerjee *et al.*, 2008) that MCMC sampling with a predictive process approximation, introducing a small amount of white noise, handles these models very effectively and is essentially off the shelf.

Andrew Gelman (*Columbia University, New York*)

Statisticians often discuss the virtues of simple models and procedures for extracting a simple signal from messy noise. But in my own applied research I constantly find myself in the opposite situation: fitting models that are simpler than I would like—models that clearly miss important features—because of limitations of computing speed and memory.

But, after decades of Moore's law, it is only fair to describe these as limitations on our computational procedures. I routinely want to fit models that cannot be fitted by using existing software, even though I know that a sufficiently simple algorithm must be out there to fit it using much less than the capabilities of a modern desktop computer.

Examples include hierarchical models for parallel time series (e.g. trends in public opinion in each of 50 states, or models for stochastically aligning tree ring data) and varying-intercept, varying-slope logistic regressions (in which case a covariance matrix needs to be modelled for the group level structure).

When fitting such models I lurch between various approximate methods based on point estimates and full Gibbs–Metropolis steps which can be slow if not guided well. These two approaches meet in the middle: approximations can be iteratively adjusted, leading ultimately to a Gibbs-like stochastic procedure, and Markov chain Monte Carlo (MCMC) sampling becomes more efficient when guided by approximations that have been tailored to the problem at hand.

I welcome the paper under discussion because it provides a more general way to construct these approximations. I suspect that, in addition to being a competitor to Gibbs and Metropolis algorithms, this approach ultimately can be used to make these stochastic algorithms more efficient.

As the authors note, a challenge remains with problems with many hyperparameters. It might help to model the hyperparameters explicitly with a hierarchical model rather than to consider them as unconstrained in some potentially large space.

I conclude with some history. 20 years ago, importance sampling was commonly viewed as an exact method, with MCMC sampling as a sometimes necessary but unfortunate approximation. For example, it was sometimes proposed to start a computation with MCMC sampling and then to finish with importance sampling to obtain an exact result. Eventually, though, statisticians realized that actually existing importance sampling is not exact but can instead be viewed as just another iterative simulation method, and one that has no particular advantages over the Metropolis algorithm or other more clearly iterative approaches (Gelman, 1991). As noted by Rue and his colleagues, now MCMC sampling is often perceived to be 'exact', but in practice it is not.

John Haslett, Michael Salter-Townshend and Nial Friel (*Trinity College Dublin*)

With the assistance of Håvard Rue, we have been using integrated nested Laplace approximations (INLAs) in inverting a multivariate non-parametric regression, i.e. given training data data $= \{(y_i, c_i); i = 1, \ldots n\}$ we study the distribution of c given y_{new}; in our application, c is two dimensional. With INLAs, it is now possible to do fast, rather general, Bayesian, multivariate, non-parametric regression without Markov chain Monte Carlo (MCMC) methods. Further, by an extension of their identities in Section 6.3, it is now possible to do fast *inverse* cross-validation. For example, if y is multivariate—and there are cases within the applied literature where $\dim(y) > 50$—then model evaluation in the inverse sense is more natural. The MCMC difficulties that were addressed in Haslett *et al.* (2006) and Bhattacharya and Haslett (2007)—in particular those of model evaluation—can now be completely circumvented.

The simplest case is as follows. A univariate count y is related to c via the density $\pi\{y;\theta_y,x(c)\}$ to a smooth latent scalar function $x(c)$ itself modelled as a stochastic process, with smoothness parameters θ_y; in fact y and $x(c)$ can be multivariate. One simple version leads to

$$\pi(c|y_{\text{new}},\text{data}) \approx k\,\pi(c)\int_{x(c)} \pi\{y_{\text{new}}|x(c);\theta\}\,\pi\{x(c)|\text{data};\theta\}\,\mathrm{d}x(c)$$

where k is a normalizing constant. With Rue and his colleagues, we model $x(c)$ as an *a priori* Gaussian Markov random field on a finite lattice C; invoking the INLA, $x(c)$ is approximately Gaussian *a posteriori*, and, using the simplest of their approximations, $\pi\{x(c)|\text{data}\} \approx \pi_{\text{G}}(\hat{\theta})$, the parameter here denoting $(\hat{\mu},\hat{\tau})$ the mean and variance of $\pi_{\text{G}}(\cdot)$ evaluated at the mode. The evaluation of the integral by quadrature is fast and adequately accurate. Since C is finite, we can normalize by evaluating the right-hand side for all $c \in C$.

It is possible to go further with INLA, for fast approximate cross-validation within the given data is now possible, i.e. we can evaluate $\pi(c|y_i,\text{data}_{-i})$ and compare with the known c_i for every i. But, as we require for *each i* evaluation at *all c*, the updates in Section 6.3 no longer suffice. However, approximate fast updates $(\hat{\mu}_{-i},\hat{\tau}_{-i})$ are available, in the same spirit of the rank 1 constraint in their equation (8); see Salter-Townshend (2008) for details. A very powerful Bayesian tool is thus available without MCMC methods.

One limitation is the dimensionality of C. When C is two dimensional on, for example, a 50-lattice, the `GMRFLib` routines are more than adequately fast. But even for a $30\times30\times30$ three-dimensional lattice we encounter problems. What might the authors recommend?

Tom Heskes and Botond Cseke (*Radboud University, Nijmegen*)

The authors are to be congratulated for a very interesting and stimulating paper. For the special case of sparse Gaussian processes with a small number of hyperparameters, the authors provide an automated procedure for approximate inference, producing very accurate results, which is orders of magnitude faster than Markov chain Monte Carlo methods.

We deeply appreciate the authors' efforts to relate their own approach to the deterministic approximations that have been developed in the machine learning literature. Following up on that, we shall attempt to shed some light on the link to expectation–propagation (EP) and discuss whether it could be used as an alternative to the Laplace approximation.

Computing posterior marginals for $\boldsymbol{\theta}$

The authors essentially apply numerical integration with the 'standard' Laplace approximation to evaluate the posterior marginal at the grid points, which they claim to be sufficiently accurate. This is somewhat contrary to the observations in Kuss and Rasmussen (2005) (see Fig. 5), where the marginal likelihood that was obtained by using the Laplace approximation seriously deviates from the truth. EP here does a much better job. See Minka (2001) for other comparisons between the Laplace approximation and EP and Seeger (2008) for an example of a Gaussian process model on which EP can still be applied but where Laplace approximation fails.

Slightly changing its standard implementation, EP can exploit the sparsity of the precision matrix in much the same way as the Laplace approximation. The computational bottleneck then also becomes the computation of the Cholesky decomposition and the marginal variances (see Section 2.1). We implemented EP for the stochastic volatility example in Section 5.3 and obtained results that were indistinguishable from a Laplace implementation with roughly the same computational complexity.

Computing posterior marginals for the latent variables x_i

For accurate approximations of the posterior marginals for the latent variables the authors have to go beyond the Gaussian approximation that they used for computing the posterior marginals for the hyperparameters. A full nested Laplace approximation is (way) too expensive, as would be a full nested EP approximation, and the authors introduce several clever tricks to obtain faster approximations thereof. Although it is not clear to us how generally applicable these approximations are, this appears to be the most important contribution of the paper. It is fair to say that (to the best of our knowledge) there are no deterministic approximations in the machine learning literature that even attempt to reach the same level of accuracy. A recent interpretation of EP as a series expansion (Opper *et al.*, 2009) may be turned into an alternative approach.

Nils Lid Hjort (*University of Oslo*) **and D. M. Titterington** (*University of Glasgow*)

The authors are to be congratulated for what promises to be a very influential contribution to practical Bayesian analysis. The methodology is very well thought out and the examples are convincing. Of course,

there are many models involving latent variables that do not fit the Gaussian framework that is considered in this paper, so exploration of approximations such as variational Bayes (VB) methods and expectation–propagation is still appropriate. As mentioned in Section 1.6, VB methods tend to underestimate posterior variance, and experience in references such as Wang and Titterington (2005) is that the VB posterior distribution is typically 'similar' to that corresponding to complete-data analysis. Indeed, for the scenario that is considered in Appendix A, if $\mathbf{x}$ were known, then

$$\theta|\mathbf{x} \sim \Gamma(a + n/2, b + \mathbf{x}^{\mathrm{T}}\mathbf{R}\mathbf{x}/2).$$

If now we note that $E(\mathbf{x}^{\mathrm{T}}\mathbf{R}\mathbf{x}) = n/\theta$, assume as in the paper that the data are not very informative and, for a rough-and-ready calculation, substitute θ by the mean a/b of a $\Gamma(a, b)$ distribution, then the complete-data posterior distribution becomes the same as the VB approximation that is stated in the paper.

Although not noted at the time, this tie-up between the VB approximation and the complete-data result is manifest in the numerical illustration of a mixture of two known densities in Humphreys and Titterington (2000). The best of the approximate methods that were illustrated there was the so-called probabilistic editor, a recursive method based on matching first and second moments; see for example section 6.2.1 of Titterington *et al.* (1985). Again in hindsight, the corresponding empirical posterior variance can be seen to be very close to that from the gold standard Gibbs sampler. In fact, the probabilistic editor can be regarded as a recursive version of expectation–propagation. Similar empirical results are available in Stephens (1997) and Minka (2001) and, much more recently, we have exploited this link to establish that, at least for this and some other simple mixture problems, expectation–propagation gets the posterior variance right, asymptotically.

Finally, we have two questions. First, is there any hope of a version of the authors' approach that at some level handles scenarios, like mixtures, in which the latent variables are anything but Gaussian? Secondly, we wish to point to the 'model builder' methodology and software that was developed by Skaug and Fournier (2002) and others, which make it possible to fit quite general non-Gaussian hierarchical models with latent variables, also using Laplace approximations. Are there connections to the present paper, and are there classes of models where both approaches may be used?

Jim Hodges (*University of Minnesota, Minneapolis*)

Faster computing methods are always good news, and I congratulate the authors on a brave, interesting body of work. I have a big concern here, however: the authors' approach appears to require unimodal posteriors, and the authors seem to think that, in the relevant set of problems, multimodal posteriors are so rare that they can be ignored. Specifically, they say in Section 1.5, 'For most real problems and data sets, the conditional posterior of $\mathbf{x}$ is typically well behaved and looks "almost" Gaussian' and, again, in Section 6.2, 'Fortunately, latent Gaussian models generate unimodal posterior distributions in most cases'. They present no evidence to support this sanguine view, because no such evidence exists. The only systematic evidence that I know of is Liu and Hodges (2003), which showed that, in the simplest possible case, the balanced one-way random-effects model, the joint marginal posterior of the two variances becomes bimodal quite readily. Moreover, since the restricted likelihood is always unimodal in this problem, the bimodality arises entirely because of the prior, which suggests that the prior can create bimodality in any problem. Further examples of bimodal posteriors for simple models and real data sets include a two-level normal errors model (Wakefield, 1998) and a conditional auto-regressive model with two classes of neighbour relations (Reich *et al.*, 2007). And these are just from my own work; there is no reason to think that I am a magnet for freak problems.

I would argue, therefore, that the authors need to take multimodality more seriously. Perhaps they can handle multimodality with a modest extension of their method, which would be good news indeed, but, until they show that they can detect bimodality reliably and then work their approximation in spite of it, they really need to tone down the sales pitch.

Borus Jungbacker and Siem Jan Koopman (*VU University Amsterdam*)

We congratulate the authors for an interesting paper. Their main proposal is to rely on Laplace approximations for the observation density $\pi(\mathbf{y}|\mathbf{x}; \theta)$, with respect to both the latent process, represented by $\mathbf{x}$, and the parameters that drive the latent processes (which are sometimes referred to as hyperparameters), represented by θ. Although it is not recognized by the authors, likelihood-based methods have been introduced in the time series literature in papers such as Shephard and Pitt (1997) and Durbin and Koopman (1997). They used approximating Gaussian models based on Laplace approximations similar to those

which are described in Section 2.2. The generality of this approach for the inference of θ is made evident in Jungbacker and Koopman (2007).

The authors' proposal is closely related to the importance sampling approach in which the likelihood function is evaluated by Monte Carlo integration. The Monte Carlo likelihood function is given by

$$\hat{l}(\theta) = \tilde{l}_{\mathrm{G}}(\theta) M^{-1} \sum_{i=1}^{M} w(\mathbf{x}^{(i)}), \qquad w(\mathbf{x}) = \frac{\pi(\mathbf{y}|\mathbf{x};\theta)}{\tilde{\pi}_{\mathrm{G}}(\mathbf{y}|\mathbf{x};\theta)},$$

for some integer M, where $\tilde{l}_{\mathrm{G}}(\mathbf{y})$ is the likelihood function of the approximating Gaussian model and $\mathbf{x}^{(i)}$ is simulated from the importance density $\tilde{\pi}_{\mathrm{G}}(\mathbf{x}|\mathbf{y};\theta)$. The approximating model and the importance density are both related to the same Laplace approximation of $\pi(\mathbf{x}|\mathbf{y};\theta)$. It was proposed by Durbin and Koopman (2000) to initialize the numerical maximization of $\hat{l}(\theta)$ with respect to θ, by taking starting values $\theta = \tilde{\theta}$ where $\tilde{\theta}$ is obtained from the maximization of $\tilde{l}_{\mathrm{G}}(\theta)\, w(\hat{\mathbf{x}})$ where $\hat{\mathbf{x}}$ represents the value of $\mathbf{x}$ at the mode of $\pi(\mathbf{x}|\mathbf{y};\theta)$. This suggestion is equivalent to step (a) in Section 3.1, in which simulations are avoided and full reliance is given to the location of the mode $\hat{\mathbf{x}}$. The remaining steps (b)–(d) can be regarded as a numerical approximation of the Bayesian approach that was advocated in Durbin and Koopman (2000) and not based on Markov chain Monte Carlo sampling.

Many different inference procedures have been developed for the stochastic volatility model and they are usually quite successful. It would, however, be more convincing when the methods are also successful for a stochastic volatility model with leverage. It requires (negative) correlation between the standard normal sequences ε_t and η_t in the model

$$y_t = \exp(\tfrac{1}{2} h_t)\, \varepsilon_t, \qquad h_{t+1} = \mu + \phi h_t + \sigma_\eta \eta_t, \quad t = 1, \ldots, T,$$

where $\mu, 0 < \phi < 1$ and $\sigma_\eta > 0$ are fixed unknown coefficients; see Jungbacker and Koopman (2007) for further details concerning estimation based on Laplace approximations.

Andrew Lawson (*Medical University of South Carolina, Charleston*)

The authors are to be commended on a very interesting and innovative paper. There is a need for alternatives to sometimes time-consuming Markov chain Monte Carlo (MCMC) sampling for posterior sampling. Although the authors provide several convincing examples of the efficiency of the approach, I have a general concern about the flexibility of the approach and the need for tuning. In the disease mapping example (5.4) the authors assume particular priors, such as $N(0,0.01)$ for an intercept, whereas in other examples they use diffuse Gaussian priors for regression parameters. The authors also always use gamma priors for precisions. There are issues arising from these choices.

(a) Are the informative priors needed, for the disease mapping example? In which case what effect does this have on the modelling approach?
(b) A more general question is how sensitive is the approximation to changes in prior distribution *and* how easy is it to implement these changes? For example, if I chose not to use gamma priors for precisions and used the half-Cauchy priors of Gelman (2006) what would the effect be?
(c) It would also be interesting to find out how well posterior functionals can be computed (e.g., in the disease mapping example, $\Pr(p_i > 0.5)$). I would guess that there could be a greater problem where tail probabilities are important and shifts of probability mass are found.
(d) Is the computation of the deviance information criterion more stable than under MCMC sampling?, i.e. can you obtain negative p_{D}s?

Currently WinBUGS easily implements prior specification changes and can model spatial data with convolution models (to convergence in a reasonably small number of seconds).

Finally I take issue with two statements:

(a) 'Despite all these developments, MCMC sampling remains painfully slow from the end-user point of view'.
(b) In the analysis of a log-Gaussian Cox process, it is common to discretize the observation window…' (Section 5.5).

In relation to the first of these, for a wide range of models MCMC sampling is reasonably fast (try the WinBUGS examples on line!). In relation to the second: although this is convenient for the authors'

purpose, I do not think that this is common practice at all. It would usually be more usual to analyse the point locations as a point process.

I guess that implementing the integrated nested Laplace approximations approach for any given model and flexibly changing that model would not be easy (without easy-to-use software that is graphical user interface based, which is not currently available). Hence WinBUGS will remain the package of choice for its ease (even though MCMC sampling might be painfully slow in some *complex* cases).

Youngjo Lee (*Seoul National University*)

In the Bayesian approach, simulation techniques such as Markov chain Monte Carlo sampling have been often used. Likelihood is informative as well. To reduce the computational burden the authors propose to use the Laplace approximation to various marginal posterior distributions $\pi(\theta_i|\mathbf{y})$, $\pi(x_i|\mathbf{y})$ etc. Here if we take $\pi(\theta) = 1$ we have h-likelihood and we (Lee and Nelder, 1996, 2001a) proposed to use the Laplace approximation to the marginal likelihood

$$\int \pi(\mathbf{x}|\theta) \prod_{i \in I} \pi(y_i|x_i, \theta)\, \mathrm{d}\mathbf{x}.$$

We (Lee and Nelder, 2001a) showed that the form that is used in Laplace approximation is identical to the Cox and Reid (1987) adjusted profile likelihood to eliminate fixed parameters. Thus, we can use this form for the Laplace approximation to eliminate both fixed and random parameters simultaneously, by eliminating fixed parameters by conditioning on their maximum likelihood estimators and random parameters by integration. This allows various adjusted profile h-likelihoods (APHLs), e.g. the generalization of the restricted maximum likelihood estimators (Lee and Nelder, 2001a). Consider the Epil example in Section 5.2. Fig. 14 shows various marginal posteriors from OpenBUGS and corresponding APHLs. They show almost identical plots for both random and fixed effects. However, plots for dispersion components can be different because the authors' inverse gamma prior is informative. This leads to biases when the prior is not right, e.g. when dispersion parameters are not random but fixed unknowns (Jang *et al.*, 2007). APHL could be used for sensitivity analysis of the choice of priors. If the marginal posterior is somewhat different from the corresponding APHL the prior could be informative.

Finn Lindgren (*Lund University*)

A point that may not be apparent from the paper is the usefulness of the integrated nested Laplace approximation (lNLA) method in the context of latent spatial Gaussian or geostatistical models. Such models are often specified in terms of covariance models (Diggle and Ribeiro, 2006), but computationally efficient Gaussian Markov random-field (GMRF) alternatives exist (Rue and Held, 2005). The INLA approach promises to be an invaluable tool for practical inference for these types of GMRF models, with a wide range of applications, e.g. in environmetrics and epidemiology. Previously, numerical optimization has been used (Rue and Tjelmeland, 2002) to find GMRF models approximating given covariance models. However, recent advances (Lindgren and Rue, 2007) in methods for obtaining explicit expressions for the precision matrix can be combined with the INLA approach, fully exploiting the efficiency of sparse matrix calculations and direct approximation of the posterior distributions.

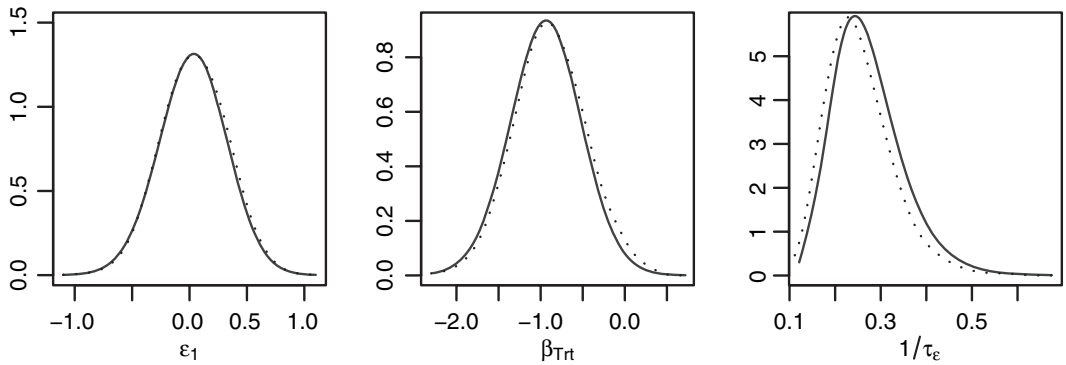

Fig. 14. Marginal posteriors (·····) *versus* APHLs (———)

Spatial random fields on $\mathbb{R}^d$ with Matérn covariance functions

$$\mathrm{cov}\{x(\mathbf{s}), x(\mathbf{s}+\mathbf{t})\} \propto (\kappa\|\mathbf{t}\|)^{\nu} K_{\nu}(\kappa\|\mathbf{t}\|), \qquad \nu, \kappa > 0,$$

where K_{ν} is the modified Bessel function of the second kind, are solutions to a fractional stochastic partial differential equation (Whittle, 1954). The stochastic partial differential equation is

$$(\beta - \Delta)^{\alpha/2} x(\mathbf{s}) = \mathcal{E}(\mathbf{s}), \qquad \beta = \kappa^2, \quad \alpha = \nu + d/2 \geqslant 0,$$

where Δ is the Laplace operator and $\mathcal{E}(\mathbf{s})$ is spatial Gaussian white noise. For integer values of α, an explicit GMRF approximation can be derived (Lindgren and Rue, 2007) from a finite element construction with local basis functions. For example, we can derive explicitly that, for a two-dimensional regular lattice, a GMRF with precision matrix with elements from

$$\begin{pmatrix} & & b^2 & & \\ & 2ab & -2bc & 2ab & \\ a^2 & -2ac & c^2 + 2(a^2+b^2) & -2ac & a^2 \\ & 2ab & -2bc & 2ab & \\ & & b^2 & & \end{pmatrix}, \qquad c = \beta + 2(a+b),$$

corresponds to an anisotropic Matérn covariance function with $\nu = 1$, and ranges $\sqrt{(8a/\beta)}$ and $\sqrt{(8b/\beta)}$. The resulting covariance approximation is shown in Fig. 15.

Generalizations include construction of (possibly oscillating) fields on non-Euclidean manifolds, such as geostatistical models on a globe, inclusion of non-homogeneous differential operators, specified either directly, via spatial deformation (Sampson and Guttorp, 1992), or via covariates. If the extensions are parameterized by using a few parameters, the INLA approach can be used for inference.

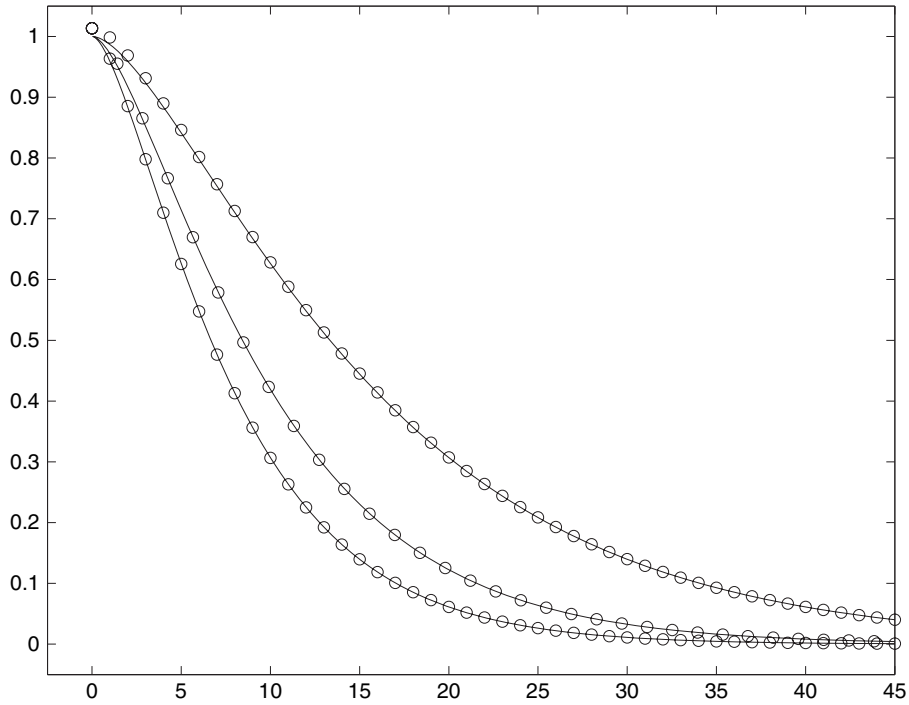

Fig. 15. Theoretical (———) and approximate covariances (○) along the two main axes as well as diagonally, normalized by the theoretical stochastic partial differential equation variance, $1/4\pi\beta\sqrt{(ab)}$: the approximate ranges are 15, 30 and 19

William J. McCausland (*University of Montreal*)

I congratulate the authors on a paper which may be remembered as a key paper marking a transition away from Markov chain Monte Carlo and importance sampling methods. To me, the most obvious and pressing need is a more convincing way to assess approximation error. The authors assess error by comparing two similar approximations of the target distribution with each other; it seems that respective approximations of the target distribution have unknown properties. I hope that the authors or others can find stronger results sanctioning the use of integrated nested Laplace approximations.

I shall also comment on the computational complexity of the simplified Laplace approximation $\tilde{\pi}_{\mathrm{SLA}}(x_i|\theta, \mathbf{y})$. The cost of computing expression (22) for a given i 'is the same order as the number of non-zero elements of the Cholesky triangle'. The authors conjecture that the computational complexity of doing this separately for all i is 'close to the lower limit for any general algorithm'. Although this may turn out to be true, I offer an example of a special case where similar approximations have lower order cost. I believe that there is some promise for extension to more general latent Gaussian models.

In McCausland (2008), I consider models where the latent Gaussian vector $(x_1, \ldots, x_n)$ has a band diagonal precision matrix with bandwidth 3. State space models with univariate states are important examples. I first express the derivative of the logarithm of the integration factor $c_1(x_i)$ in

$$\pi(x_i|\theta, x_{i+1}, \ldots, x_n, \mathbf{y}) \propto c_1(x_i)\, \pi(y_i|\theta, x_i)\, \pi(x_{i+1}|\theta, x_i)$$

in terms of the function $f(x_i) = E[x_{i-1}|\theta, x_i, \mathbf{y}]$. A handful of first-order difference equations then give coefficients of cubic approximations for all n of these functions, which leads to cubic approximations of the $\log\{\pi(x_i|\theta, x_{i+1}, \ldots, x_n)\}$. This leads in turn to an $O(n)$ approximation of $\pi(\mathbf{x}|\theta, \mathbf{y})$ that is extremely close for the example that is considered in the paper. Inspired by Rue, Martino and Chopin, I discovered that a simple extension gives the derivative of $\log\{c_2(x_i)\}$, where

$$\pi(x_i|\theta, \mathbf{y}) \propto c_2(x_i)\, \pi(y_i|\theta, x_i),$$

in terms of $E[x_{i-1}|\theta, x_i, \mathbf{y}]$ and $E[x_{i+1}|\theta, x_i, \mathbf{y}]$. The latter can be approximated in the same way as the former, using the same difference equations and the time-reversed series. One can use these to compute an approximation of *all* the $\pi(x_i|\theta, \mathbf{y})$ at cost $O(n)$.

C. McGrory (*Queensland University of Technology, Brisbane*), **J. Marriott** (*Nottingham Trent University*) **and A. N. Pettitt** (*Queensland University of Technology, Brisbane*)

We very much enjoyed reading this paper and are very impressed by the computational speed which the approximate methods enjoy, even for the large data set problem that is described in Section 5.5. We are reminded of the early work on Bayesian computation that was carried out at Nottingham in the 1980s. Naylor and Smith (1982) used iterative Gauss–Hermite quadrature to produce posterior expected values for functions of the model parameters. This relies on the posterior density being approximated by a product of a Gaussian density and a polynomial. They showed how the basic iterative Gauss–Hermite algorithm could be applied to multiple integrals as a Cartesian product rule and subsequently incorporated spherical quadrature in their Bayes4 software (Smith *et al.*, 1987). The Bayes4 software routinely provides univariate and bivariate posterior densities but there is an effective limit on the number of parameters at about 12. Naylor *et al.* (2008) show how, in some circumstances, this number can be considerably extended; nonetheless we note the contrast that this paper makes.

In the spatial models of Sections 5.4 and 5.5, the Gaussian Markov random field (GMRF) is chosen to be the intrinsic second-order model which has the tendency to oversmooth. In an attempt to be less smooth, Pettitt *et al.* (2002) introduced a full rank GMRF which could be used in the integrated nested Laplace approximations context and we believe that this would be a worthwhile alternative to the intrinsic GMRFs. A further attempt to be less smooth is to use an autologistic or Ising-type model for spatial data with a hidden binary variable (e.g. McGrory *et al.* (2008)), whereas Weir and Pettitt (2000) used a thresholded GMRF.

Another possibility (see Woolrich and Behrens (2006)) is to use a k-multivariate GMRF prior from which a hidden k categorical variable can be approximated, to imitate a Potts model. This can be achieved by introducing a prior distribution on the continuous weights vectors for the GMRF which results in a posterior probability of membership of one of K discrete classes for each observation. We would like to know whether or not the approximations of Section 3.2.2 would be sufficiently good to find good overall approximations to the posteriors in this case.

Debashis Mondal (*University of Chicago*)

To consolidate the ideas of this paper further, we indicate an empirical Bayes approach in which hyperparameters are estimated by using a set of empirical measurements and then posterior density of covariate and underlying spatial effects are obtained by using the approximations described in Section 3.2, without resorting to Markov chain Monte Carlo computations. Consider the longitudinal model discussed in Section 5.2. Assume, for convenience,

$$\beta \sim N\{0, \tau_\beta^{-1}(Z^{\mathrm{T}}Z)^{-1}\},$$

where Z forms the matrix of covariates. From the theoretical expected values of $Y_{j,k}(Y_{j,k}-1)$ and $Y_{j,k}Y_{j,k'}$ for $k \neq k'$, the moment estimate of τ_v^{-1} is

$$\log\left\{\frac{1}{177}\sum_{j=1}^{59}\sum_{k=1}^{3} Y_{j,k}(Y_{j,k}-1)\right\} - \log\left(\frac{1}{177}\sum_{j=1}^{59}\sum_{1\leqslant k<k'\leqslant 3} Y_{j,k}Y_{j,k'}\right) = 0.1426.$$

We then minimize the sum of squares of deviances between the observed and the expected values of $Y_{j,k}$ to obtain 0.8073 and 105.296 as respective estimates of τ_ε^{-1} and τ_β^{-1}. With this knowledge of hyperparameters, we can now advance calculations as outlined in Section 3.2. Although we sacrifice variability of hyperparameters, the empirical Bayes procedure simplifies the computations.

In spatial mixed linear models on regular arrays, the precision parameters can be estimated by equating the theoretical variogram with the observed variogram at small lags. Once we obtain knowledge of precision parameters, the marginal posterior means and variances for covariate and spatial effects can be computed by using the results of Rue and Martino (2007). My on-going project extends the parametric empirical Bayes estimation of precision parameters to Poisson and binomial spatial models. When the latent Gaussian field has a non-singular Q, we can invoke equation (7) to evaluate marginal moments of the observed data. However, when the latent Gaussian field is 'first order' intrinsic, differences of X rather than X itself have proper distribution. Consider $\gamma_{i,j} = \frac{1}{2}\,\mathrm{var}(X_i - X_j)$ to derive an extension of expression (7) as

$$\gamma_{i,j} = \frac{1}{2L_{i,i}^2} - \frac{1}{L_{i,i}}\sum_{k>i}\gamma_{k,i}L_{k,i} - \frac{1}{2L_{i,i}^2}\sum_{k>i}\sum_{k'\neq k>i}\gamma_{k,k'}L_{k,i}L_{k',i}, \qquad j>i, \quad i=n-1,\ldots,1.$$

This can be used to evaluate $\gamma_{i,j}$ to deduce empirical Bayes estimates of precision parameters after computing appropriate expectations of empirical measurements.

The gain in the empirical Bayes approach can be taken to derive frequentist inference of spatial models, which is a long-standing difficult problem. In particular, estimation of precision parameters can be done by moment (or variogram) expansions of marginal densities, and inference for random-treatment and spatial effects can be based on conditional distributions or point predictors (e.g. best linear unbiased predictors in mixed spatial models). Approximations in Section 3.2 can be used in computations of these point predictors.

In practice, it is important to obtain marginal posterior means, variances and select number of quantiles for covariate or spatial effects. We presume that the skew normal approximation in Section 3.2.3 provides an efficient way to compute these quantities when the log-likelihood is skewed; how does this compare with the general Gauss–Hermite quadrature formula (17)?

John Nelder (*Imperial College London*)

Since Lee and Nelder (1996) introduced the model class of hierarchical generalized linear models it has been further extended (Lee and Nelder, 2001b, 2006). Random effects can appear in the linear predictors not only for the mean but also for various variance components. With this new class heavy-tailed distributions can be used for the various components of the model, allowing robust modelling against misspecification of the distribution of random effects (Noh *et al.*, 2005) and various data contaminations (Noh and Lee, 2007a); it also allows modelling of abrupt changes (Yun and Lee, 2006) and modelling of parametric Lévy processes for stochastic volatility in finance data (Castillo and Lee, 2008). The model class (1) in this paper assumes normal random effects only in the linear predictor for the mean, but puts priors on hyperparameters θ. For inferences about extended models we proposed to use h-likelihood and introduced various adjusted profile h-likelihoods for inferences about various components of the model. To approximate the necessary integration we have proposed to use the Laplace approximation and for non-normal random effects the second-order Laplace approximation has been recommended (Lee and Nelder, 2001a, 2007b). With h-likelihood we can make inferences without inventing priors for hyperparameters.

David Nott (*National University of Singapore*) **and Robert Kohn** (*University of New South Wales, Sydney*)
We congratulate Rue, Martino and Chopin on their paper which addresses the important issue of effective computation for Bayesian inference. The authors demonstrate that the class of latent models that they consider makes fast computations possible and provides important insights and solutions. The results will be used by applied researchers and will also generate new research in Bayesian computation.

Our first comment concerns Section 6.1 where they suggest copula approximations for marginals of subsets of $\mathbf{x}$ based on the univariate marginals. We have also recently developed some approximate Bayesian computational methods using copulas, in particular for marginal likelihood computation in Bayesian model comparison. The starting point for this is the so-called candidate's formula, similar to expression (3). Writing the set of all unknowns including any latent variables now as simply θ,

$$p(\mathbf{y}) = p(\theta)\, p(\mathbf{y}|\theta)/p(\theta|\mathbf{y}),$$

which holds for any value of θ and clearly approximation of the posterior at a point allows approximation of the marginal likelihood. A subset of the parameters could be handled non-parametrically as in the present paper. Laplace approximation corresponds to use of a Gaussian approximation for $p(\theta|\mathbf{y})$ evaluated at the posterior mode. As an extension it is natural to approximate $p(\theta|\mathbf{y})$ with a Gaussian copula and this can be done both with and without simulation (Nott *et al.*, 2008). Various extensions such as the use of copulas with importance-sampling-based methods and the use of the t-copula instead of a Gaussian copula are possible. It would be interesting to apply copula approximations to posterior distributions based on methods for the current paper in models where those methods can be applied.

A second comment concerns the class of models that was considered. Although this class of models is quite broad, the use of a Gaussian latent variable, a small number of hyperparameters and perhaps most importantly the conditional independence assumptions for $\mathbf{y}$ given $\mathbf{x}$ are serious restrictions for many applications. Future developments of the work that is described in Section 6.5 are important, we feel, as are methods for combining the approximations that are proposed with simulation-based methods.

J. T. Ormerod and M. P. Wand (*University of Wollongong*)
We concur with the authors that good analytic approximations, as an accurate alternative to Markov chain Monte Carlo methods, are worth pursuing. These early results on integrated nested Laplace approximations are impressive and we look forward to seeing how this methodology progresses. In particular we are interested in the advertised interface from R and eventually in giving integrated nested Laplace approximations a 'test drive'.

Our recent research has involved work in variational approximation for similar models. Most of the discussion in Section 1.6 pertains to a particular version of variational approximation where $q(\mathbf{x}, \boldsymbol{\theta}) = q_{\mathbf{x}}(\mathbf{x})\, q_{\theta}(\boldsymbol{\theta})$. The phrase 'the variational Bayes approach is not without potential problems' and subsequent discussion actually correspond to this one type of variational approximation, even though $q(\mathbf{x}, \boldsymbol{\theta})$ can be constrained in other ways. Indeed, some variational approximations, such as those developed in Jaakkola and Jordan (2000), do not involve Kullback–Leibler contrast. Lastly, the name 'variational Bayes' gives the impression of variational approximation being specific to Bayesian approaches, which is not so.

Recently, we have explored some other approaches to variational approximations that exhibit improved accuracy in our test examples. One approach involves applying the Jaakkola and Jordan (2000) tangent transform idea in a gridwise fashion (Ormerod, 2008; Ormerod and Wand, 2008). Another takes the Kullback–Leibler contrast route but restricts q to be in a parametric family, such as the Gaussian distribution. We close with some details on the latter approach, which we call *Gaussian variational approximation*, for frequentist Poisson mixed models with a single variance component:

$$y_{ij}|u_i \overset{\text{ind}}{\sim} \text{Poisson}\{\exp(\boldsymbol{\beta}^{\mathrm{T}}\mathbf{x}_{ij} + u_i)\}, \qquad u_i \overset{\text{ind}}{\sim} N(0, \sigma^2), \quad 1 \leqslant j \leqslant n_i, \quad 1 \leqslant i \leqslant m. \tag{39}$$

The log-likelihood of $(\boldsymbol{\beta}, \sigma^2)$ is

$$l(\boldsymbol{\beta}, \sigma^2) = \sum_{i=1}^{m} \sum_{j=1}^{n_i} \{y_{ij}(\boldsymbol{\beta}^{\mathrm{T}}\mathbf{x}_{ij}) - \log(y_{ij}!)\} - \frac{m}{2}\log(2\pi\sigma^2) + \sum_{i=1}^{m} \log\left[\int_{-\infty}^{\infty} \exp\left\{\sum_{j=1}^{n_i} y_{ij}u - \exp(\boldsymbol{\beta}^{\mathrm{T}}\mathbf{x}_{ij} + u) - \frac{u^2}{2\sigma^2}\right\} \mathrm{d}u\right].$$

Table 1. Estimates and approximate 95% confidence intervals for Gaussian variational approximation corresponding to the example in Section 5.2 with the ν_{ij}-term omitted†

Parameter	*Gaussian variational approximation*	*Exact*
β_0	1.924 (1.767, 2.081)	1.924 (1.766, 2.082)
β_{Base}	0.165 (−0.128, 0.458)	0.165 (−0.128, 0.459)
β_{Trt}	0.842 (0.013, 1.671)	0.842 (0.014, 1.673)
β_{BT}	−0.366 (−0.805, 0.072)	−0.366 (−1.806, 0.073)
β_{Age}	−0.328 (−1.072, 0.416)	−0.328 (1.074, 0.418)
β_{V4}	0.236 (0.138, 0.333)	0.236 (0.138, 0.333)
$\tau_\varepsilon^{-1/2}$	0.580 (0.466, 0.723)	0.581 (0.461, 0.700)

†Exact answers (obtained via adaptive Gauss–Hermite quadrature are given for comparison.

A variational approach to handling the m intractable integrals is to multiply the integrand by the quotient of the $N(\mu_i, \lambda_i)$ density function with itself and to invoke Jensen's inequality:

$$\log\left[\int_{-\infty}^{\infty} \exp\left\{\sum_{j=1}^{n_i} y_{ij}u - \exp(\boldsymbol{\beta}^{\mathrm{T}}\mathbf{x}_{ij} + u) - \frac{u^2}{2\sigma^2}\right\} \frac{\exp\{-(u-\mu_i)^2/2\lambda_i\}/\sqrt{(2\pi\lambda_i)}}{\exp\{-(u-\mu_i)^2/2\lambda_i\}/\sqrt{(2\pi\lambda_i)}}\,\mathrm{d}u\right]$$
$$\geqslant E_{U\sim N(\mu_i,\lambda_i)}\left[\left\{\sum_{j=1}^{n_i} y_{ij}U - \exp(\boldsymbol{\beta}^{\mathrm{T}}\mathbf{x}_{ij} + U) - \frac{U^2}{2\sigma^2}\right\} + \frac{(U-\mu_i)^2}{2\lambda_i} + \frac{1}{2}\log(2\pi\lambda_i)\right].$$

After simplification we obtain the following lower bound on $l(\boldsymbol{\beta}, \sigma^2)$:

$$\underline{l(\boldsymbol{\beta}, \sigma^2, \boldsymbol{\mu}, \boldsymbol{\lambda})} = \sum_{i=1}^{m}\sum_{j=1}^{n_i}\{y_{ij}\boldsymbol{\beta}^{\mathrm{T}}\mathbf{x}_{ij} - \log(y_{ij}!)\} + \frac{m}{2}\{1 - \log(\sigma^2)\}$$
$$+ \sum_{i=1}^{m}\sum_{j=1}^{n_i}\left\{y_{ij}\mu_i - \exp\left(\boldsymbol{\beta}^{\mathrm{T}}\mathbf{x}_{ij} + \mu_i + \frac{1}{2}\lambda_i\right)\right\} + \frac{1}{2}\sum_{i=1}^{m}\left\{\log(\lambda_i) - \frac{\mu_i^2 + \lambda_i}{\sigma^2}\right\}$$

for all values of the *variational* parameters $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_m)$ and $\boldsymbol{\lambda} = (\lambda_1, \ldots, \lambda_m)$. Maximizing over these parameters narrows the gap between $\underline{l(\boldsymbol{\beta}, \sigma^2, \boldsymbol{\mu}, \boldsymbol{\lambda})}$ and $l(\boldsymbol{\beta}, \sigma^2)$ and so sensible estimators of the model parameters are

$$(\underline{\hat{\boldsymbol{\beta}}}, \underline{\hat{\sigma}}^2) = (\boldsymbol{\beta}, \sigma^2)\text{component of}\arg\max_{\boldsymbol{\beta}, \sigma^2, \boldsymbol{\mu}, \boldsymbol{\lambda}}\{\underline{l(\boldsymbol{\beta}, \sigma^2, \boldsymbol{\mu}, \boldsymbol{\lambda})}\}.$$

Table 1 conveys excellent performance of Gaussian variational approximation when expression (39) is applied to the data that were used in Section 5.2. Early theoretical exploration looks promising.

Carl Edward Rasmussen (*University of Cambridge*)

I congratulate Professor Rue and his colleagues for their contribution to developing efficient analytic approximation methods for a wide and practically important class of models.

I am concerned, however, about the extent to which the shortcomings of the Laplace approximation may have been treated too lightly when advocating it as a generally applicable tool. The Achilles heel of the Laplace approximation is expansion around the *mode* of the distribution. In high dimensions, for non-Gaussian, non-symmetric posterior distributions, the mode may not be typical of the distribution; for a skew distribution the majority of the mass may lie far to one side of the mode. This is true even for unimodal, log-concave and otherwise fairly harmless distributions. As the Laplace approximation is symmetric around the mode, this may seriously hamper its accuracy.

Gaussian latent variable models with a logistic likelihood is an example in point which has been studied carefully in the machine learning community, where it is known as Gaussian process classification (Rasmussen and Williams, 2006). Careful comparisons between the Laplace approximation and other analytical approximations as well as a Markov chain Monte Carlo gold standard (Kuss and Rasmussen,

2005) document exactly this widespread failure mode. The expectation–propagation (EP) method (Minka, 2001) is an alternative analytical approximation method which does not suffer from this problem, as it is based on (approximate) matching of marginal moments. The EP approximation is found to be in close agreement with Markov chain Monte Carlo results and much more accurate than the Laplace approximation in the difficult cases where the posterior deviates significantly from Gaussianity, in terms of both the quality of approximation to the marginal likelihood and the predictive distribution for test cases.

Kuss and Rasmussen (2005) focused on conditional distributions given the hyperparameters, but the same numerical method suggested by Rue and his colleagues can also be used with EP to treat hyperparameters. Although a simple implementation of both methods scales cubically with the number of latent variables, comparable implementations of both algorithms (available at `www.gaussianprocess.org/gpml`) show that the Laplace approximation is typically about 10 times faster than EP (and not 8000 times faster as implied in Section 5.4). Additionally, both methods can be *sparsified* to achieve a further speed-up. Whereas the Laplace approximation clearly has its merits, an exclusive reliance on it would appear hazardous.

R. A. Rigby (*London Metropolitan University*)

I congratulate the authors for their general approach and also for their computational achievements.

A key equation in the paper is the approximation that is used for the marginal posterior of the hyperparameters θ given by equation (3).

The denominator in equation (3), $\tilde{\pi}_{\mathrm{G}}(\mathbf{x}|\boldsymbol{\theta}, \mathbf{y})$, is a Gaussian approximation to $\pi(\mathbf{x}|\boldsymbol{\theta}, \mathbf{y})$ centred at the posterior mode $\mathbf{x}^*(\boldsymbol{\theta})$ with precision matrix $\mathbf{D}^*(\boldsymbol{\theta}) = \mathbf{Q} + \mathrm{diag}(\mathbf{c})$ derived from the curvature at the mode. Equation (3) is evaluated by the authors at $\mathbf{x} = \mathbf{x}^*(\boldsymbol{\theta})$, giving

$$\tilde{\pi}(\boldsymbol{\theta}|\mathbf{y}) = \frac{\pi\{\mathbf{x}^*(\boldsymbol{\theta}), \boldsymbol{\theta}, \mathbf{y}\}}{|\mathbf{D}^*(\boldsymbol{\theta})/2\pi|^{1/2}}.$$

This Laplace approximation is equivalent to the formula that was used for inference about hyperparameters θ by Lee and Nelder (1996), equation (4.6), in their hierarchical generalized linear models and, among others, by Rigby and Stasinopoulos (2005), equation (12), in their generalized additive models for location, scale and shape.

Matthias Seeger (*University of Saarbrücken*)

As a machine learning researcher, I welcome the publication of this paper, and the discussion that it might initiate. Bayesian statisticians restrict themselves to a single approved approximate inference methodology: Markov chain Monte Carlo (MCMC) sampling. For smooth generalized linear models with provably unimodal (log-concave) posteriors, but strong couplings, there are solid arguments that, on a reasonable time horizon, random sampling should be less accurate in general than deterministic approximations. Modern statistics supports decision making in new fields of science and engineering, where high dimensional, complex models are used under time constraints. I agree with the authors that the argument of accuracy after reasonable time cannot be taken lightly: in many applications, well served by rigorous Bayesian statistical analysis, it is more important than vanishing error in the limit.

The authors could have presented the considerable amount of work done on variational approximate inference in machine learning, information theory and statistical physics over the last 15 years a little more carefully; Section 1.6 seems somewhat dismissive though uninformed. Variational (mean field) Bayes methods tend to underestimate variances, but their comment about expectation–propagation seems ill informed. I am not convinced that their own method should fare consistently better, when compared on an equal footing (their second-order Gaussian approximation at the conditional mode being replaced by expectation–propagation or variational Bayes methods), and they do not present a comparison. In my experience, each of the major deterministic approximations comes with strengths and weaknesses, as does MCMC sampling. The technique presented here seems no exception. For example, sparse linear models, which are of substantial interest right now, are latent Gaussian models, but the second-order Gaussian approximation at the posterior mode is not well defined. A good Gaussian approximation can be found by expectation–propagation or other variational approximations. Also, for provably skew posteriors (logistic regression), it seems suboptimal to place a Gaussian approximation at the mode.

For many posteriors, MCMC sampling is still the only useful option. Often easy to implement, it can be overly slow, and it is very difficult to use properly for non-experts. For some models, deterministic approximations are nowadays used in many statistics applications and deserve attention in the field. Most concepts that they rely on are grounded in statistics and probability, and they can be analysed by similar mathematics to point estimation techniques. A wider interest among statisticians may help their properties

to become better understood, so that practitioners can use them with the confidence that is required for sound statistical analysis.

D. P. Simpson (*Queensland University of Technology, Brisbane*)

The analysis and implementation of integrated nested Laplace approximations described in this paper is grounded in the fast, exact Cholesky algorithms for sampling from a Gaussian Markov random field introduced by Rue (2001). However, as integrated nested Laplace approximation is an approximate method, further computational savings may possibly be made through the judicious use of inexact methods. This comment will focus on the computation of Gaussian approximations described in Section 2.2 and carry through the formulation of integrated nested Laplace approximation in Section 3 in the obvious way.

Two key computations, the mode and the marginal variances, are required to determine the Gaussian approximation to $\pi(\mathbf{x})$. The computation of the mode requires the solution of the sequence of linear systems

$$\{\mathbf{Q} + \text{diag}(\mathbf{c}^{(i)})\}\boldsymbol{\mu}^{(i+1)} = \mathbf{c}^{(i)}.$$

These systems can be solved quickly and inexactly by the preconditioned conjugate gradient method (Saad, 2003).

Consider the mth step of an orthogonal tridiagonalization

$$\{\mathbf{Q} + \text{diag}(\mathbf{c}^*)\}\mathbf{V}_m = \mathbf{V}_m\mathbf{T}_m + \mathbf{E}_m$$

which can be performed by using the symmetric Lanczos algorithm with a random starting vector (Saad, 2003; Stewart, 2001). This leads to the approximation to the marginal variances

$$\{\mathbf{Q} + \text{diag}(\mathbf{c}^*)\}^{-1} \approx \mathbf{V}_m\mathbf{T}_m^{-1}\mathbf{V}_m^{\mathrm{T}}.$$

For reasonable values of m, this approximation has two correct digits! These values can be refined by applying the preconditioned conjugate gradient to $\mathbf{1}_i^{\mathrm{T}}\{\mathbf{Q} + \text{diag}(\mathbf{c}^*)\}^{-1}\mathbf{1}_i$. This refinement can be performed in parallel. We note that a similar procedure does not lead to an efficient method for approximating the mode.

The computational saving can be demonstrated by considering the simple binomial regression model

$$y_i|x_i \sim \text{Bin}\{n, \text{logit}^{-1}(x_i)\},$$
$$\mathbf{x} \sim \text{MVN}(\mathbf{0}, 5\mathbf{Q}),$$

where $\mathbf{Q}$ is a second-order random walk on a 50×50 grid. The data were generated from the surface $z = \sin(3x)\cos(7y) + \cos(7x)\sin(4y)$. All computations were performed by using MATLAB 7.4 on a 2.3-GHz Macintosh Book Pro with 2 Gbytes of random-access memory. When the reduced model was run with $m = 20$, there was a saving of over a second compared with the exact Gaussian approximation (1.785 s compared with 2.867 s), and

$$\max_{1 \leqslant i \leqslant n} |\sigma_i^2 - \tilde{\sigma}_i^2| = 0.0189.$$

We note that, as m increases, the error in the marginal variances (Fig. 16) does not decay quickly. This validates our assertion that this approximation requires only relatively small values of m.

Hans J. Skaug (*University of Bergen*)

I congratulate the authors on an innovative paper on the use of Laplace-type approximations in modern Bayesian modelling. Apart from the computational techniques, I find the proposal that is made to assess the accuracy of the Laplace approximation by using asymptotics in terms of $p_{\mathrm{D}}(\theta)/n_d$ to be very promising. However, the interpretation of p_{D} as the *effective dimension* of $\mathbf{x}$ can be misleading in the current context. As the authors point out, for non-informative data we have $p_{\mathrm{D}} = 0$, without this meaning that the dimension of $\mathbf{x}$ is 0 in any sense. It is more appropriate to think of $(n - p_{\mathrm{D}})/n = \text{tr}\{\mathbf{Q}(\boldsymbol{\theta})\mathbf{Q}^*(\boldsymbol{\theta})^{-1}\}/n$ as the relative influence of the prior on the posterior. When the (Gaussian) prior dominates, the Laplace approximation becomes increasingly accurate.

My main comment is that several aspects of the computational machinery that is presented by Rue and his colleagues could benefit from the use of a numerical technique known as automatic differentiation (AD) (see Griewank (2000)). Simply put, AD is compiler-generated code for evaluating (numerically) the first- and higher order derivatives of a mathematical function. The method outperforms finite difference-based derivatives with respect to both accuracy and computational cost, and also differs from symbolic differ-

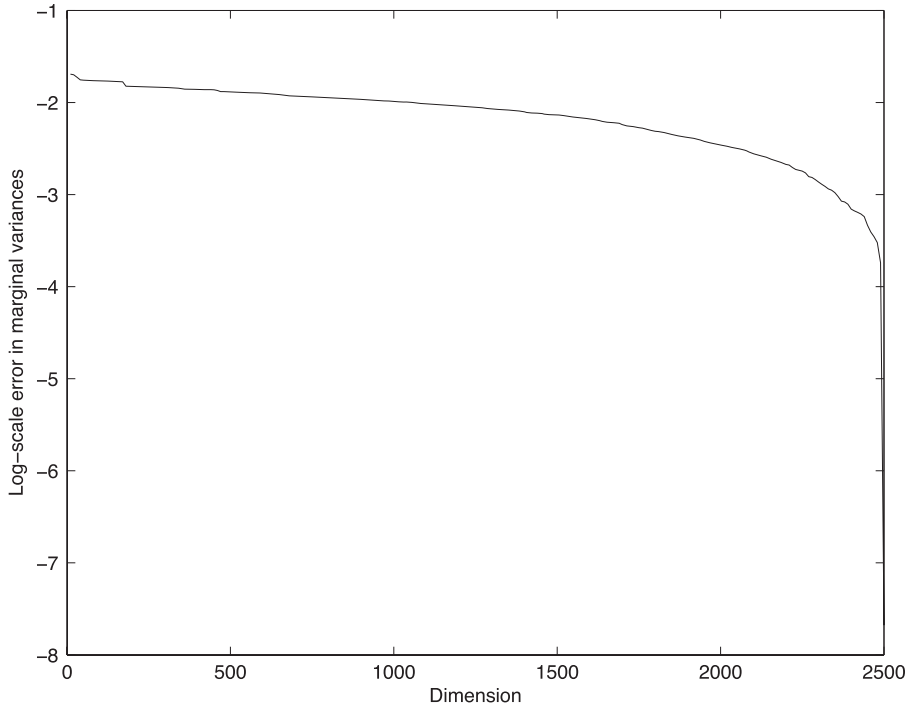

Fig. 16. Maximum error in the marginal variances for varying subspace sizes: there appears to be no benefit in taking a large m

entiation. Skaug and Fournier (2006) argued that by evaluating $\mathbf{Q}^*$ using AD the Laplace approximation becomes 'automatic', relieving the statistician of the burden of calculating second-order derivatives. Further, they presented a formula for the gradient with respect to $\boldsymbol{\theta}$ of $\tilde{\pi}(\boldsymbol{\theta}|\mathbf{y})$, involving up to third-order derivatives, that would be useful under (a) in Section 3.1 as well. We may view the authors' framework as a back-bone, where the user only has to specify a few parametric components: $\pi(\boldsymbol{\theta})$, $\mathbf{Q}(\boldsymbol{\theta})$ and $\pi(y_i|x_i, \boldsymbol{\theta})$, or more generally $\pi(\mathbf{y}|\mathbf{x}, \boldsymbol{\theta})$. By the use of AD one could obtain a system that is automatic from a user's perspective, almost to the same extent as with BUGS. The benefit would be a fast, flexible and easy-to-use system for doing Bayesian analysis in models with Gaussian latent variables.

I. L. Solis-Trapala (*Lancaster University*)

We propose a procedure to conduct likelihood-based inference based on the marginal density of the observational variables $\mathbf{y}$. The idea, in tune with the authors' strategy, is to explore specific features of the models under consideration to reduce the computational effort without compromising accuracy. A potentially useful procedure to achieve this is to implement the method of maximization by parts that was proposed by Song *et al.* (2005).

The maximization-by-parts procedure involves the decomposition of the log-likelihood function into two components,

$$l(\boldsymbol{\theta}) = l_{\mathrm{w}}(\boldsymbol{\theta}) + l_{\mathrm{e}}(\boldsymbol{\theta}),$$

where l_{w} is chosen as a 'working' model so that its likelihood score equation is easy to solve, but the second derivatives of l_{e} may be challenging. To find the maximum likelihood estimate of $\boldsymbol{\theta}$, the algorithm solves the working score equation $\dot{l}_{\mathrm{w}}(\boldsymbol{\theta}^1) = 0$ and then recursively solves $\dot{l}_{\mathrm{w}}(\boldsymbol{\theta}^{k+1}) = -\dot{l}_{\mathrm{e}}(\boldsymbol{\theta}^k)$, where $\dot{l}$ denotes the vector of first-order derivatives of l. The algorithm is attractive because, under certain conditions, its convergence is relatively fast and stable with little loss of efficiency relative to using the full likelihood function.

Let $\boldsymbol{\theta}_2$ denote the vector of parameters that index the conditional density of the observational variables $\mathbf{y} = \{y_i : i \in \mathcal{I}\}$ given a latent Gaussian prooess $\mathbf{x}'$ with mean $\mathbf{0}$ and precision matrix $\mathbf{Q}(\boldsymbol{\theta}_1)$. Note that $\mathbf{x}'$ does not contain any parameters, in contrast with the latent process $\mathbf{x}$ that is defined by the authors. The target

of inference is $\boldsymbol{\theta} = (\boldsymbol{\theta}_1, \boldsymbol{\theta}_2)$. Under the assumption of conditional independence of the variables $\{y_i : i \in \mathcal{I}\}$ given $\mathbf{x}'$, the full likelihood function of $\boldsymbol{\theta}$ is proportional to

$$\pi(\mathbf{y}; \boldsymbol{\theta}) = \int \pi(\mathbf{x}'; \boldsymbol{\theta}_1) \prod_{i \in \mathcal{I}} \pi(y_i | x_i'; \boldsymbol{\theta}_2) \, \mathrm{d}\mathbf{x}',$$

where $\pi(\mathbf{x}'; \boldsymbol{\theta}_1)$ is a multivariate Gaussian density function with mean $\mathbf{0}$ and precision matrix $\mathbf{Q}(\boldsymbol{\theta}_1)$.

To conduct likelihood inference by using maximization by parts we propose a working model, where we assume that the variables $\{x_i' : i \in \mathcal{I}\}$ are marginally independent normal random variables; and the working posterior density $\pi_{\mathrm{w}}(\mathbf{x}'|\mathbf{y}; \boldsymbol{\theta})$ is Gaussian with mean and precision as derived in Section 2.2. Under this choice for the working model, the log-likelihood function can be expressed as

$$l(\boldsymbol{\theta}; \mathbf{y}) = \log\{\pi_{\mathrm{w}}(\mathbf{y}; \boldsymbol{\theta})\} + \log\left\{ \int \frac{\pi(\mathbf{x}'; \boldsymbol{\theta})}{\pi_{\mathrm{w}}(\mathbf{x}'; \boldsymbol{\theta})} \pi_{\mathrm{w}}(\mathbf{x}'|\mathbf{y}; \boldsymbol{\theta}) \, \mathrm{d}\mathbf{x}' \right\}, \tag{40}$$

where

$$\pi_{\mathrm{w}}(\mathbf{y}; \boldsymbol{\theta}) \equiv \int \pi_{\mathrm{w}}(\mathbf{x}'; \boldsymbol{\theta}) \prod_{i \in \mathcal{I}} \pi(y_i | x_i; \boldsymbol{\theta}_2) \, \mathrm{d}\mathbf{x}'$$

and $\pi_{\mathrm{w}}(\mathbf{x}'; \boldsymbol{\theta})$ is a multivariate normal density with mean $\mathbf{0}$ and a diagonal covariance matrix. The first component in equation (40) involves the evaluation of $|\mathcal{I}|$ one-dimensional integrals that can be computed through Gaussian quadrature; and the integral in the second component, with the Gaussian working model, can be easily evaluated by using Monte Carlo methods even when the dimension of $\mathbf{x}'$ is large. The properties of this procedure are under current investigation.

S. H. Sørbye, F. Godtliebsen and K. Hindberg (*University of Tromsø*), **T. A. Øigård** (*Institute of Marine Research, Tromsø, and Norwegian Centre for Telemedicine, Tromsø*), **V. Hadziavdic** (*Discover Petroleum, Tromsø*) **and K. Thon** (*Norwegian Centre for Telemedicine, Tromsø*)

First of all, we congratulate the authors on a very important paper that we believe will have a huge influence in many areas of statistics. The authors have an impressive list of potential applications for their new methodology. In this discussion we shall add to their list.

Our focus is how to utilize the integrated nested Laplace approximation (INLA) approach within scale–space analysis, where an observed process that evolves in time and/or space is studied as a function of both location and scale. Scale–space ideas were introduced to non-parametric curve estimation by Chaudhuri and Marron (1999), presenting the 'SiZer' methodology. The main idea of SiZer is to perform inference for several scales or smoothing levels simultaneously, detecting and visualizing significant trends of the true curve viewed at different resolutions. Originally, inference in SiZer was based on Gaussian distributional quantile assumptions or bootstrapping. Computational improvements based on extreme value and asymptotic theory were developed in Hannig and Marron (2006).

For Gaussian observational models, the ideas of SiZer can be combined with fast and exact Bayesian inference, utilizing the special conditional independence structure of Gaussian Markov random fields; see Øigård *et al.* (2006) and Hadziavdic *et al.* (2008). By applying INLA, these ideas can be extended to non-Gaussian observational models. One example is given in Sørbye *et al.* (2009), where the simplified Laplace approximation was used to extend the ideas of SiZer to spectral density estimation. Applying an integrated Wiener process as a prior, posterior marginals of both the true curve and its first derivative can be estimated for different locations and scales. Tests for significant trends are then evaluated straightforwardly.

Scale–space methods might be rather computer intensive for large data sets. Applying INLA, fast Bayesian inference can now be obtained for the wide class of latent Gaussian models. Compared with Markov chain Monte Carlo alternatives, inference using INLA is also very accurate. Of special importance within scale–space applications is the fact that INLA produces posterior marginal estimates having a relative error, implying that tail probabilities are also estimated accurately. In comparison, Markov chain Monte Carlo techniques would produce marginal estimates with absolute error.

A very nice feature of INLA is that the methodology is easily accessed by using the `inla` program that is described in Martino and Rue (2008). Different regression models are easily specified, e.g. including covariates, unstructured effects and linear constraints. Interesting scale–space applications could involve for example image analysis and analysis of climate data sets.

Ingelin Steinsland and Henrik Jensen (*Norwegian University of Science and Technology, Trondheim*)
First we thank the authors for their contributions and also for making the methodology easily available to others through the integrated nested Laplace approximation (INLA) program (Martino and Rue, 2008). Our interest is in quantitative genetics, and we have two comments:

(a) the INLA methodology can be used for the animal model;
(b) issues about non-Gaussian hyperparameters.

The animal model
For the full Bayesian single-trial and multitrial animal models see Steinsland and Jensen (2005) or Sorensen and Gianola (2002). The model is also known as the additive genetic model and is used both in evolutionary biology and in animal breeding. It assumes that animal i's trait y_i can be divided into a genetic part (or inherited part) $u_i \sim N(0, \sigma_u^2)$ and an environmental part $\varepsilon_i \sim N(0, \sigma_\varepsilon^2)$. u_i is the breeding value of individual i. The breeding values in a population are assumed Gaussian with a dependence that is given by the pedigree. This gives a Gaussian Markov random-field structure and computational benefits (Steinsland and Jensen, 2005). Hence the animal model is a structured latent Gaussian model. This is also true when the observed trait y_i is non-Gaussian, e.g. binary with a logit link to the 'latent trait' η_i, $\eta_i = u_i + \varepsilon_i$, and observations $y_i = \text{logit}(\eta_i)$.

The INLA methodology enables us to find the posterior breeding values of individual animals. This is useful for choosing the best breeding animals, and also of interest when predicting response of selection.

Non-Gaussian hyperparameters
An important quantitative genetic parameter is the *heritability* h^2 of a trait, how much of the phenotypic (observed trait) variance in a population that can be explained by additive genetic effects:

$$h^2 = \frac{\sigma_u^2}{\sigma_\varepsilon^2 + \sigma_u^2}.$$

This is a function of two non-Gaussian hyperparameters, and we need $\pi(\sigma_\varepsilon^2, \sigma_u^2|y)$, which was found in Steinsland and Jensen (2005).

To obtain reasonable accuracy of the posterior heritability $\pi(h^2|y)$, we have experienced that the number of configurations of $(\sigma_\varepsilon^2, \sigma_u^2)$ for which we need to calculate $\pi(\sigma_\varepsilon^2, \sigma_u^2|y)$ is large and the computation time increases rapidly.

Often it is of interest to consider several traits simultaneously. With only two traits we then have a minimum of six non-Gaussian hyperparameters: the environmental and genetic variances of both traits and the genetic and environmental correlations. The INLA methodology is valid for *a small number* (5 or fewer) of non-Gaussian hyperparameters.

We find that the small number of non-Gaussian parameters is the most restrictive requirement of the methodology. To loosen this restriction should be a topic for further research.

Jarno Vanhatalo and Aki Vehtari (*Helsinki University of Technology, Espoo*)
There are two major challenges in using latent Gaussian models:

(a) the inference is analytically intractable, which leads to a need for approximate inference schemes, and
(b) the inference on the model requires inversion of the covariance matrix, which scales as $O(n^3)$ as a function of data size n.

Rue, Martino and Chopin tackle these problems in an elegant way and we are pleased to welcome their paper.

In the paper, the solution for problem (b) relies on Markov random-field approximation for the latent Gaussian field, where the conditional distribution of a latent variable depends only on a small number of its neighbours. This leads to a sparse precision matrix, from which the authors can efficiently evaluate all the marginal variances needed. The trade-off with this is that the prior is tied to a specified set of locations rather than being defined in a spatially continuous way. This might be a problem, for example, in high accuracy spatial data, which have large unpopulated areas. In such a case, it would be more desirable to define the correlations through distance by using the covariance function. The covariance function also gives more flexibility in defining the correlation structure and, for example, non-stationarity.

In Section 1.6 it is mentioned about expectation–propagation (EP) that '... it is not clear how to extend this approach to fully Bayesian analysis'. However, the numerical integration scheme over hyperparameters

θ that is presented in Sections 1.7 and 3.1 can also be applied for EP. Nickisch and Rasmussen (2008) have shown that EP gives very accurate approximations for the marginal likelihood, which can be used to find an approximation for the marginal posterior $\pi(\theta|\mathbf{y}) \propto \pi(\mathbf{y}|\theta)\,\pi(\theta)$. This, in turn, can be used to search the values θ_k that are needed in equation (5) of the paper.

Kuss and Rasmussen (2005) showed that EP outperforms Gaussian approximation (or Laplace approximation as it is called in their paper) in classification problems. In classification, the marginal posteriors of latent values are highly skewed, for which reason Gaussian approximation at the posterior mode is not effective. However, in our experiments, we have found that EP and a Gaussian approximation give the same answer in disease mapping problems with Poisson count data and a Gaussian prior for the log-relative-risk. The reason for this is that the marginal posteriors of latent values are very close to Gaussian in these models. With EP, it is attractive that it approximates the full posterior whereas integrated nested Laplace approximation approximates only marginals. Thus, it would be interesting to compare EP and integrated nested Laplace approximation in detail.

Christopher K. Wikle and Scott H. Holan (*University of Missouri, Columbia*)

We thank the authors for this timely paper describing a very useful and practical computational procedure. We have a variety of experience with latent Gaussian models and agree that the computational issues can be frustrating. The authors have done a very nice job of building on and improving the approach that was described in Rue and Martino (2007). Although the approach that is presented here is fairly straightforward, its implementation from scratch would require significantly more effort than a similar Markov chain Monte Carlo program (from scratch). However, our experience is that Markov chain Monte Carlo runs for these models must be tailored to the given model and data set to gain the maximum efficiency. Thus, the primary value of the methodology that is discussed here is that it provides a practical approach that can be implemented in standard software packages. This, along with the interface to the `inla` program for R, will probably lead to a much wider application of this very important and useful class of models by subject-matter scientists and practitioners.

We feel that the two primary model assumptions concerning the latent Gaussian field (the Markov random field and very small number of hyperparameters) are somewhat limiting for many real world problems of interest to statisticians in research collaborations with scientists. Indeed, from our perspective, the simple low dimensional hyperparameter and Markov random-field assumptions for spatiotemporal models are fairly unrealistic for many physical and biological processes. Hierarchical models for such processes are often significantly more complicated and require multiple levels of dependent processes and complicated process interactions. Nevertheless, we believe that the methods that are presented here provide a solid foundation on which to base extensions and will surely lead to additional computational innovations that will assist in providing efficient solutions to these problems as well.

Chris Williams (*University of Edinburgh*)

I congratulate Professor Rue and his colleagues for their line of work on deterministic approximation methods for latent Gaussian models. In statistics Markov chain Monte Carlo approximations have been prevalent but, as the authors point out, these approximations can be very expensive computationally, and their goal is to popularize deterministic approximation methods. My comments focus mainly on work from the machine learning literature, particularly the expectation–propagation (EP) method of Minka (2001) (see also Opper and Winther (2000)) and how it relates to the current paper. I note that both EP and mode finding for $\mathbf{x}^*$ (as discussed in Section 2.2) are iterative algorithms, although EP is somewhat more computationally expensive. I believe that EP is well behaved for latent Gaussian models with log-concave likelihood.

Firstly, using EP we can obtain an approximation for $\pi(x_i|\boldsymbol{\theta}, \mathbf{y})$ by combining the approximate Gaussian 'cavity' distribution for x_i given the ys for all sites excluding y_i, and the exact likelihood $\pi(y_i|x_i)$. It would be interesting to know how this compares with the approximations that are used in Section 3.2.

Secondly, in the paper a Laplace approximation is used for $\tilde{\pi}(\mathbf{y}|\boldsymbol{\theta})$, as (implicitly) in equation (3). The EP method also produces an approximation to this quantity. Kuss and Rasmussen (2005) have made a nice comparison of the fidelity of these two approximations for a Gaussian process classification problem, using 'ground truth' computed by the annealed importance sampling method (Neal, 2001). The results show that the EP method provides a much more accurate approximation than the Laplace approximation. This issue is also discussed in section 3.7 of Rasmussen and Williams (2006).

Finally, in Section 4.1 the authors discuss asymptotics for $\tilde{\pi}(\boldsymbol{\theta}|\mathbf{y})$. These can be somewhat complicated when the model dimensionality n grows along with the number of observations n_d. One attractive frame-

work to study this issue is for Gaussian process models where essentially n is infinite for non-degenerate covariance functions (but with summability conditions on the eigenvalues), but n_d increases steadily. Here $\boldsymbol{\theta}$ indexes the parameters in the covariance function. In this case Stein (1999) contrasted the regimes of fixed domain asymptotics (where one obtains increasingly dense observations within some region) with increasing domain asymptotics (where the size of the observation region grows with n_d). Increasing domain asymptotics are a natural choice in a time series context but fixed domain asymptotics seem more natural in spatial settings. For a discussion of the asymptotic theory here see Stein (1999), section 6.4.

The **authors** replied later, in writing, as follows.

We thank all the discussants for their insightful and supportive comments. Below we give our response to the main issues that are raised in the discussion; for brevity, many of the more detailed questions are unfortunately left unanswered. We would like to start with a general response before answering more in detail.

We consider latent Gaussian models (LGMs) to be a very important class of models, and we think that the unification into LGMs allows a unified approach for Bayesian inference and, possibly, also for theoretical analysis.

Among all generic algorithms for LGMs, the *performance* of the (approximate or not) inference algorithm is the main measure of success. By performance, we mean quality of the results compared with the exact results, when used central processor unit time is accounted for. We consider the complexity of implementation of the algorithm to be less important, as the algorithm is *generic* and then highly reusable.

We consider the integrated nested Laplace approximation (INLA) approach to be a good candidate for a generic algorithm for LGMs (under the assumptions that were stated earlier on) when the task is to compute posterior marginals. The performance of INLAs for joint inference is still unclear.

Our INLA approach does not imply that further research on Markov chain Monte Carlo (MCMC) methods is no longer needed; nor should it be seen as a *competitor* to MCMC methods. The overall goal is to make Bayesian inference in a reliable and computationally efficient way; an issue which seems somewhat neglected these days within the statistical community.

We now give a more detailed response.

The complexity of integrated nested Laplace approximations

Gelfand, Robert, Richardson and Frigessi express concerns about the complexity of INLAs. We agree that our approach is complex to implement, but at least it provides Bayesians with their omelet while it is still hot. (*Haslett* and his colleagues report that with INLAs they can finally taste the Bayesian omelet!) Moreover, owing to the simple form of LGMs, much of the code is reusable and a generic implementation of INLAs is already available. We do not understand the *high level of expertise* argument from the user's point of view, as our approach is essentially *off the shelf* as announced. An R interface is already available and we invite *Ormerod and Wand*, and others, to give INLA a 'test drive'; see also the supportive comment by *Banerjee, Ferreira, Haslett and his colleagues* and *Sørbye and her colleagues*.

Comparison with Markov chain Monte Carlo methods

We agree with *Diggle, Gelman, Roberts* and *Williams*, who mention the high computational cost of MCMC sampling in large dimensional problems and the general difficulty of assessing convergence of such algorithms. (Assessment of convergence is difficult for the same reason that makes MCMC methods work.) The high cost is especially prominent in situations where multiple models and priors must be considered. Precise estimates, like those reported in the paper, require much computational effort even for simple models. Less precise estimates can be provided in shorter time; see *Lawson*. See also *Roberts* for an interesting discussion on the theoretical limit of the performance of MCMC sampling with respect to the dimension of the sampling space.

Everitt, Gelman and *Robert* suggest that some of our insights could be used to derive better MCMC schemes, like using $\tilde{\pi}(\boldsymbol{\theta}|\mathbf{y})$ and $\tilde{\pi}_{\mathrm{G}}(\mathbf{x}|\boldsymbol{\theta}, \mathbf{y})$ to define block sampling schemes. Such approaches are indeed possible; see Rue *et al.* (2004) and Rue and Held (2005) for a discussion of these and related approaches for LGMs in general, and the references that were mentioned by *Jungbacker and Koopman* for dynamic models. Some insight from INLAs can be used to derive the following semi-MCMC scheme. Assume that we want to infer some function $g(\mathbf{x}, \boldsymbol{\theta})$. For each configuration of $\boldsymbol{\theta}_k$ (see Section 3.1), sample $\mathbf{x}^{(1)}, \mathbf{x}^{(2)}, \ldots$ from $\pi(\mathbf{x}|\boldsymbol{\theta}_k, \mathbf{y})$ by using for example the Gaussian or some improved approximation, and estimate the conditional density from $\{g(\mathbf{x}^{(k)}, \boldsymbol{\theta}_k)\}$. Finally, integrate these densities over $\boldsymbol{\theta}_k$ by using expression (5).

The conditioning on $\boldsymbol{\theta}_k$ will grossly simplify the MCMC algorithm as the strong dependence between $\boldsymbol{\theta}$ and $\mathbf{x}$ is broken, and the samples computed for each $\boldsymbol{\theta}_k$ can be run in parallel. Like INLA, this scheme assumes few hyperparameters and is not feasible otherwise.

As we argued earlier, the scope of MCMC-based inference for LGMs depends on the type of statistics that we want to infer, so the effort of *Marin and his colleagues* must be seen in that context. As long as the task is to obtain posterior marginals of $\{\theta_k\}$ and/or $\{x_i\}$, the INLA seems to us superior as long as the assumptions are valid.

Banerjee, Demiris, Diggle, Held and Riebler, and *Nott and Kohn* ask whether we can do joint inference within the INLA framework, like computing joint credibility intervals (*Held and Riebler*) or infer maximum values (*Diggle*). (Although INLAs provide posterior marginals, it can be extended as discussed in Section 6.1.) These issues will require more work; joint credibility intervals might be within reach as they are known analytically for the Gaussian distribution, whereas maximum values are not. See also *Nott and Kohn* for an interesting connection with their own research on a similar issue.

Gelfand is concerned about how well INLA behaves for weakly identified parameters; the accuracy of the INLA approach depends on the quality of the Gaussian approximation, not whether some parameters are weakly identified or not. *Banerjee* is concerned about the requirement that the latent field is a Gaussian Markov random field with sparse precision matrix; but any Gaussian distribution will do at a potentially higher computational cost. The new results of *Lindgren* (which we believe are very important) provide explicit results about concerns that are raised by *Banerjee* and *Gelfand*.

Comparison with machine learning methods

Williams, Heskes and Cseke, Hjort and Titterington, Rasmussen, Seeger and *Vanhatalo and Vehtari* mention interest in comparing the INLA approach with *expectation–propagation* (EP). In short, the idea in EP is to fit a Gaussian distribution by matching the marginals instead of matching the mode and the curvature at the mode. The INLA approach can obviously be modified by using other definitions of 'Gaussian approximation' and we welcome any study of this kind!

A note to the reader: in the machine learning literature, the expression *Laplace approximation* is also used for what we denote the Gaussian approximation $\tilde{\pi}_{\mathrm{G}}(\mathbf{x}|\boldsymbol{\theta}, \mathbf{y})$, which is unfortunately a source of confusion. We reserve that expression for situations where the Gaussian distribution is used to marginalize out some variables like expression (3).

EP requires more calculation than the Gaussian approximation as it is more iterative in nature. By using sparsity (as noted by *Heskes and Cseke* and *Rasmussen*), these calculations can be speeded up. It is likely that this sparse (matrix) variant of EP can deal with linear constraints at a slightly increasing cost. EP also requires some adaptation to the likelihood under study; see Kuss and Rasmussen (2005). If these extra efforts are required for difficult problems, so be it.

We believe that one reason for the comment by *Heskes and Cseke* about approximating $\pi(x_i|\mathbf{y})$,

> 'it is fair to say that there are no deterministic approximations in the machine learning literature that even attempt to reach the same level of accuracy',

is the *repeated* use of the Gaussian approximation. Instead of trying to approximate $\pi(x_i|\boldsymbol{\theta}_k, \mathbf{y})$ directly like EP, we repeatedly make use of the Gaussian approximation to $\pi(\mathbf{x}_{-i}|x_i, \boldsymbol{\theta}_k, \mathbf{y})$ as a function of x_i. This is the underlying idea of the Laplace approximation, and it allows for improved accuracy using a sequence of potentially lower quality approximations. In the context of binary classification that was discussed by Kuss and Rasmussen (2005), we would like to illustrate this difference on a simple example that mimics their Fig. 7. Let $n = 2, \boldsymbol{\theta} = \emptyset$ and

$$\pi(\mathbf{x}|\mathbf{y}) \propto \exp\left\{-\frac{1}{2}\mathbf{x}^{\mathrm{T}}\begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix}^{-1}\mathbf{x}\right\} \Phi(cy_1x_1)\,\Phi(cy_2x_2),$$

where $\Phi(\cdot)$ is the cumulative distribution function for a standard normal distribution and c is a scaling factor. Various marginals for x_1 are plotted in Fig. 17(a) for $y_1 = y_2 = -1, c = 3$ and $\rho = 0.8$; the true marginal (the full curve), the Gaussian approximation (the chain curve) and the Laplace approximation (the broken curve). Although the Gaussian approximation is quite far off, its *repeated use* provides a good approximation. By increasing the constant c the skewness increases and the Laplace approximation becomes less precise (narrower) but it is still better than Fig. 7a in Kuss and Rasmussen (2005). The repeated use of simpler approximations is also evident in our suggestion (30) for approximating the marginal likelihood with unknown hyperparameters. We believe that this approximation will improve on that used in Kuss and Rasmussen (2005) and Kass and Vaidyanathan (1992).

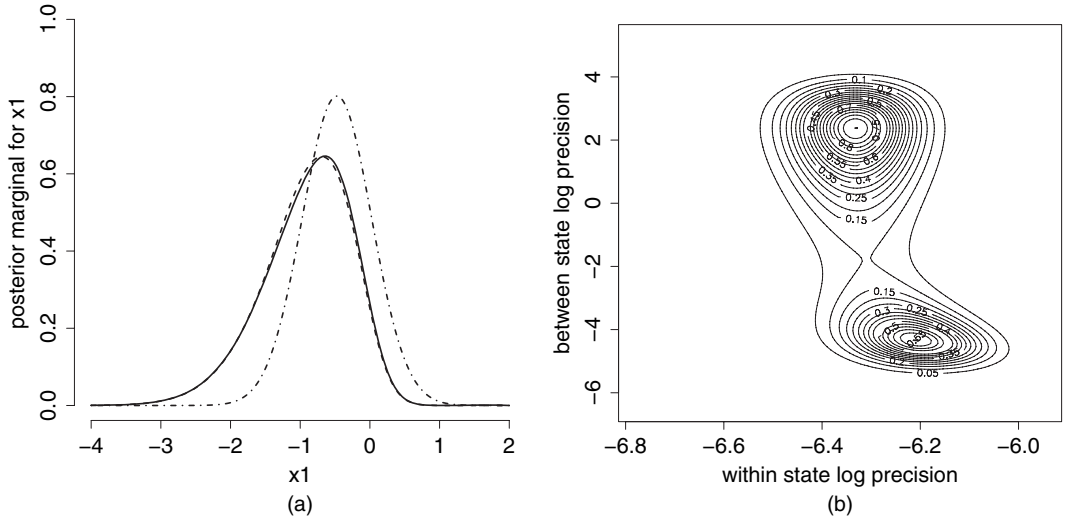

Fig. 17. (a) Posterior marginals for the binary classification example and (b) joint posterior marginal in the multimodality example

We welcome the preview of the promising variational Bayes approach of *Ormerod and Wand*, and we thank *Hjort and Titterington* for an interesting interpretation of variational Bayes methods as an approximation of the complete-data likelihood.

Properties of Laplace approximations

We agree with *Johansen, McCausland, Roberts, Robert* and *Williams*, who call for a better asymptotic assessment of the approximation error. This is a difficult issue, as it is always possible to consider the model of interest as an element of a well-chosen sequence of models that ensures that the approximation error vanishes as required, but it might not be very helpful to assess the error for the model of interest. It might be somehow contradictory to formalize properly the good performance of our approximations using asymptotic arguments, as the good performance seems directly related to how much the Gaussian prior departs from Gaussianity when conditioned on the observations. We agree that more work in this direction is welcome, even if our own attempts seem to suggest that this may be arduous. In connection with this discussion, *Roberts* notes that the performance of MCMC methods is better under outfill than infill asymptotics, whereas for the INLA approach we conjecture that it is the opposite for the approximation of $\pi(x_i|\boldsymbol{\theta}_k, \mathbf{y})$: in the outfill regime, additional observations are added further away, thus preventing the approximation error from vanishing to 0.

Robert recalls that the Laplace approximation is not parameterization invariant. Although this is true in general, we note that INLA is invariant (up to numerical integration) to one-to-one transforms of $\boldsymbol{\theta}$ and invariant to linear transforms of the latent Gaussian $\mathbf{x}$. Further, *Robert* finds it 'rather annoying' that a hyperparameter can appear in the latent field, e.g. the mean of the auto-regressive process in the stochastic volatility example. We can assure *Robert* that this is just a computational trick and that this particular hyperparameter does not lose its interpretation of being a hyperparameter in any sense; sometimes it is beneficial to separate the modelling and the inference steps.

Part of the inaccuracy that was reported by *Casarin and Robert* seems due to their use of the Taylor series expansion around the fixed point $x_t^* = \log(y_t/\sigma^2)$, which can be far off the modal configuration for large $|\rho|$.

Fahrmeir and Kneib, Lee, Nelder, Rigby and *Solis-Trapala* rightly point out that Laplace approximations are also used in frequentist contexts; see for example the references that are cited in their discussions. In return, it may be possible to reuse some of our ideas in frequentist situations, as mentioned by *Demiris. Papaspiliopoulos* nicely discusses a possible implication of what Besag (1989) denoted a 'candidate's formula', which is indeed curious!

Practical aspects and extensions

Ferreira and *Hodges* mention the issue of multimodality. In short, we cannot deal with multimodality of $\pi(\mathbf{x}|\boldsymbol{\theta}, \mathbf{y})$ but we can, to some extent, deal with multimodality (when the modes are sufficiently close) of

$\pi(\boldsymbol{\theta}|\mathbf{y})$. To illustrate this issue, we have reanalysed model (5.1) in Hodges (1998); our approximation for the joint marginal for the two log-precisions is displayed in Fig. 17(b). Our integration strategy will, with some more conservative settings, detect also the other mode in this particular example. It might be useful to modify the integration strategy to be aware of the events 'increasing $\tilde{\pi}_{\mathrm{G}}(\boldsymbol{\theta}|\mathbf{y})$ with increasing distance from the modal configuration', as this is an indicator for a second mode. Since we explore $\tilde{\pi}_{\mathrm{G}}(\boldsymbol{\theta}|\mathbf{y})$ directly, it is *easier* for INLA to detect multimodality than for an MCMC scheme where the exploration of $\boldsymbol{\theta}$ also depends on the current state of $\mathbf{x}$.

We thank *Simpson* for his comments on approximate linear algebra which is also better suited for parallel implementation; the approach for approximating the marginals is particularly interesting. *Skaug* suggests the use of tools for automatic differentiation which will simplify the implementation.

We thank *Aas, Haslett and his colleagues Lindgren, Mondal, Steinsland and Jensen*, and *Sørbye and her colleagues* for their supportive comments and their nice examples of application of our approach in their respective fields.

We agree with *Skaug*'s alternative interpretation of p_{D}.

We agree with *Everitt, Fahrmeir and Kneib, Gelman, Steinsland and Jensen*, and *Wikle and Holan* that there are several applications where the number of hyperparameters is large(r), e.g. adaptive smoothing as mentioned by *Fahrmeir and Kneib*. As *Everitt* and *Gelman* point out, adding structure through an additional hierarchical level may be useful. Alternatively, one may use the extension that we propose in Section 7 or, more expediently, one may ignore the uncertainty with respect to the hyperparameters $\boldsymbol{\theta}$, by approximating $\pi(\boldsymbol{\theta}|\mathbf{y})$ by a Dirac mass centred at either the posterior mode of $\boldsymbol{\theta}$ or some other estimate; see *Mondal* for an example based on empirical Bayes methods. (Note that this approach removes the exponential complexity of the integration!) In that respect, the posterior mode of $\tilde{\pi}(\boldsymbol{\theta}|\mathbf{y})$ can also be obtained (for some models) by an (approximate) EM algorithm, where the expectation step involves the Gaussian approximation $\tilde{\pi}_{\mathrm{G}}(\mathbf{x}|\boldsymbol{\theta}, \mathbf{y})$. This can be an interesting alternative, since EM algorithms can be quite efficient and are typically quite robust to bad starting points.

We agree with *Papaspiliopoulos* and *Roberts* on (quoting *Roberts*): it is by 'focusing on large generic families of problems with characteristic posterior distributions' that we can provide approximations that are both fast and accurate. Conversely, extending our approach to models that differ more or less markedly from the class that we currently consider may be difficult, in particular for models (and situations) that were mentioned by *Fuentes, Hjort and Titterington* and *McGrory and her colleagues*. We mention one possible class of extensions, which seems to be within reach. Consider $u \sim f(u)$, where u is univariate. Rewrite $f(\cdot)$ as

$$f(u) = \exp(-\tfrac{1}{2}u^2 + [\log\{f(u)\} + \tfrac{1}{2}u^2]).$$

Using this alternative interpretation, a non-Gaussian u can be considered as a Gaussian variable observed with a particular 'likelihood'. Extensions of this idea are immediate. However, it will require that we allow observation y_i to depend on $\{x_j : j \in \mathcal{S}_i\}$ (but still conditional independent), where $\mathrm{card}(\mathcal{S}_i) \geqslant 1$; an extension that is also required for stochastic volatility models with leverage (see *Aas*, and *Jungbacker and Koopman*). Although this is conceptually straightforward, our current implementation does not support this. Related are also the concerns of *Banerjee, Gelfand, Nott and Kohn* and *Sahu*, that conditional independence of the observations conditioned on $(\mathbf{x}, \boldsymbol{\theta})$ is a strong assumption in several applications. It is perhaps less strong than it might appear at first; see the example by *Sahu*.

Held and Riebler, Johansen and *Papaspiliopoulos* mention the possibility of extending INLA to sequential problems, where the size of the latent field grows linearly with time. Such an extension would surely be of interest, as it could improve on approximations such as the extended Kalman filter which can be interpreted as the sequential version of the simple Gaussian approximation.

References in the discussion

Baladandayuthapani, V., Mallick, B. K. and Carroll, R. J. (2005) Spatially adaptive Bayesian penalized regression splines (P-splines). *J. Computnl Graph. Statist.*, **14**, 378–394.

Banerjee, S., Gelfand, A. E., Finley, A. O. and Sang, H. (2008) Gaussian predictive process models for large spatial data sets. *J. R. Statist. Soc.* B, **70**, 825–848.

Barndorff-Nielsen, O. E. (1983) On a formula for the distribution of the maximum likelihood estimator. *Biometrika*, **70**, 343–365.

Barndorff-Nielsen, O. E. (1997) Normal inverse Gaussian distributions and stochastic volatility modelling. *Scand. J. Statist.*, **24**, 1–13.

Besag, J. (1989) A candidate's formula: a curious result in Bayesian prediction. *Biometrika*, **76**, 183.
Besag, J. (1994) Discussion on 'Representations of knowledge in complex systems' (by U. Grenander and M. I. Miller). *J. R. Statist. Soc.* B, **56**, 591–592.
Besag, J., Green, P., Higdon, D. and Mengersen, K. (1995) Bayesian computation and stochastic systems. *Statist. Sci.*, **10**, 3–41.
Beskos, A., Papaspiliopoulos, O., Roberts, G. O. and Fearnhead, P. (2006) Exact and computationally efficient likelihood-based inference for discretely observed diffusions (with discussion). *J. R. Statist. Soc.* B, **68**, 333–382.
Bhattacharya, S. and Haslett, J. (2007) Importance re-sampling MCMC for cross-validation in inverse problems. *Bayes. Anal.*, **2**, 385–408.
Bollerslev, T. (1986) Generalized autoregressive conditional heteroskedasticity. *J. Econometr.*, **31**, 307–327.
Box, G. E. P. (1980) Sampling and Bayes' inference in scientific modelling and robustness (with discussion). *J. R. Statist. Soc.* A, **143**, 383–430.
Brezger, A., Fahrmeir, L. and Hennerfeind, A. (2007) Adaptive Gaussian Markov random fields with applications in human brain mapping. *Appl. Statist.*, **56**, 327–345.
Castillo, J. and Lee, Y. (2008) GLM-methods for volatility models. *Statist. Modllng*, to be published.
Chaudhuri, P. and Marron, J. S. (1999) SiZer for exploration of structures in curves. *J. Am. Statist. Ass.*, **94**, 807–823.
Chib, S., Nardari, F. and Shephard, N. (2002) Markov Chain Monte Carlo methods for stochastic volatility models. *J. Econometr.*, **108**, 281–316.
Christensen, O., Roberts, G. and Rosenthal, J. (2005) Scaling limits for the transient phase of local Metropolis–Hastings algorithms. *J. R. Statist. Soc.* B, **67**, 253–268.
Christensen, O., Roberts, G. and Skold, M. (2006) Bayesian analysis of spatial GLMM using partially non-centered MCMC method. *J. Computnl Graph. Statist.*, **15**, 1–17.
Cox, D. R. and Reid, N. (1987) Parameter orthogonality and approximate conditional inference (with discussion). *J. R. Statist. Soc.* B, **49**, 1–39.
Crainiceanu, C. M., Ruppert, D., Carroll, R. J., Joshi, A. and Goodner, B. (2007) Spatially adaptive Bayesian penalized splines with heteroscedastic errors. *J. Computnl Graph. Statist.*, **16**, 265–288.
Diggle, P. J. and Ribeiro, P. J. (2006) *Model-based Geostatistics*. New York: Springer.
Doucet, A., de Freitas, N. and Gordon, N. (eds) (2001) *Sequential Monte Carlo Methods in Practice*. New York: Springer.
Durbin, J. and Koopman, S. J. (1997) Monte Carlo maximum likelihood estimation for non-Gaussian state space model. *Biometrika*, **84**, 669–684.
Durbin, J. and Koopman, S. J. (2000) Time series analysis of non-Gaussian observations based on state space models from both classical and Bayesian perspectives (with discussion). *J. R. Statist. Soc.* B, **62**, 3–56.
Fahrmeir, L. (1992) Posterior mode estimation by extended Kalman filtering for multivariate dynamic generalized linear models. *J. Am. Statist. Ass.*, **87**, 501–509.
Fahrmeir, L. and Kneib, T. (2009) Propriety of posteriors in structured additive regression models: theory and empirical evidence. *J. Statist. Plannng Inf.*, **139**, 843–859.
Fahrmeir, L., Kneib, T. and Lang, S. (2004) Penalized structured additive regression: a Bayesian perspective. *Statist. Sin.*, **14**, 731–761.
Fuentes, M. (2007) Approximate likelihood for large irregularly spaced spatial data. *J. Am. Statist. Ass.*, **102**, 321–331.
Gaetan, C. and Grigoletto, M. (2004) Smoothing sample extremes with dynamic models. *Extremes*, **7**, 221–236.
Gelman, A. (1991) Iterative and non-iterative simulation algorithms. *Comput. Sci. Statist.*, **24**, 433–438.
Gelman, A. (2006) Prior distributions for variance parameters in hierarchical models. *Bayes. Anal.*, **1**, 1–19.
Geyer, C. J. and Thompson, E. A. (1992) Constrained Monte Carlo maximum likelihood for dependent data (with discussion). *J. R. Statist. Soc.* B, **54**, 657–699.
Gneiting, T., Stanberry, L. I., Grimit, E. P., Held, L. and Johnson, N. A. (2008) Assessing probabilistic forecasts of multivariate quantities, with applications to ensemble predictions of surface winds. *Test*, **17**, 211–235.
Goldstein, H. and Spiegelhalter, D. J. (1996) League tables and their limitations: statistical issues in comparisons of institutional performance. *J. R. Statist. Soc.* A, **159**, 385–409.
Griewank, A. (2000) *Evaluating Derivatives: Principles and Techniques of Algorithmic Differentiation*. Philadelphia: Society for Industrial and Applied Mathematics.
Hadziavdic, V., Thon, K., Øigård, T. A. and Godtliebsen, F. (2008) Bayesian multiscale analysis for digital images. To be published.
Hannig, J. and Marron, J. S. (2006) Advanced distribution theory for SiZer. *J. Am. Statist. Ass.*, **101**, 484–499.
Harvey, A. C. and Shephard, N. (1996) Estimation of an asymmetric stochastic volatility model for asset returns. *J. Bus. Econ. Statist.*, **14**, 429–434.
Haslett, J., Whiley, M., Bhattacharya, S., Salter-Townshend, M., Wilson, S. P., Allen, J. R. M., Huntley, B. and Mitchell, F. J. G. (2006) Bayesian palaeoclimate reconstruction (with discussion). *J. R. Statist. Soc.* A, **169**, 395–438.

Held, L. (2004) Simultaneous posterior probability statements from Monte Carlo output. *J. Computnl Graph. Statist.*, **13**, 20–35.
Humphreys, K. and Titterington, D. M. (2000) Approximate Bayesian inference for simple mixtures. In *Proc. Computational Statistics 2000* (eds J. G. Bethlehem and P. G. M. van der Heijden), pp. 331–336. Heidelberg: Physica.
Jaakkola, T. S. and Jordan, M. I. (2000) Bayesian parameter estimation via variational methods. *Statist. Comput.*, **10**, 25–37.
Jang, M., Lee, Y., Lawson, A. and Browne, W. (2007) A comparison of the hierarchical likelihood and Bayesian approaches to spatial epidemiological modelling. *Environmetrics*, **18**, 809–821.
Jungbacker, B. and Koopman, S. J. (2007) Monte Carlo estimation for nonlinear non-Gaussian state space models. *Biometrika*, **94**, 827–839.
Kass, R. E. and Steffey, D. (1989) Approximate Bayesian inference in conditionally independent hierarchical models (parametric empirical Bayes). *J. Am. Statist. Ass.*, **84**, 717–726.
Kass, R. E. and Vaidyanathan, S. K. (1992) Approximate Bayes factors and orthogonal parameters, with application to testing equality of two binomial proportions. *J. R. Statist. Soc.* B, **54**, 129–144.
Kauermann, G. (2005) Penalised spline in multivariable survival models with varying coefficients. *Computnl Statist. Data Anal.*, **49**, 169–186.
Kim, S., Shephard, N. and Chib, S. (1998) Stochastic volatility: likelihood inference and comparison with arch models. *Rev. Econ. Stud.*, **65**, 361–393.
Kneib, T. and Fahrmeir, L. (2006) Structured additive regression for categorical space-time data: a mixed model approach. *Biometrics*, **62**, 109–118.
Kneib, T. and Fahrmeir, L. (2007) A mixed model approach for geoadditive hazard regression. *Scand. J. Statist.*, **34**, 207–228.
Knorr-Held, L. and Rue, H. (2002) On block updating in Markov random field models for disease mapping. *Scand. J. Statist.*, **29**, 597–614.
Krivobokova, T., Crainiceanu, C. M. and Kauermann, G. (2008) Fast adaptive penalized splines. *J. Computnl Graph. Statist.*, **17**, 1–20.
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**, 183–212.
Lang, S., Fronk, E.-M. and Fahrmeir, L. (2002) Function estimation with locally adaptive dynamic models. *Computnl Statist.*, **17**, 479–499.
Lee, Y. and Nelder, J. A. (1996) Hierarchical generalized linear models (with discussion). *J. R. Statist. Soc.* B, **58**, 619–678.
Lee, Y. and Nelder, J. A. (2001a) Hierarchical generalized linear models: a synthesis of generalized linear models, random-effect model and structured dispersion. *Biometrika*, **88**, 987–1006.
Lee, Y. and Nelder, J. A. (2001b) Modelling and analysing correlated non-normal data. *Statist. Modlling*, **1**, 3–16.
Lee, Y. and Nelder, J. A. (2006) Double hierarchical generalized linear models (with discussion). *Appl. Statist.*, **55**, 139–185.
Lindgren, F. and Rue, H. (2007) Explicit construction of GMRF approximations to generalised Matérn fields on irregular grids. *Scand. J. Statist.*, **35**, 691–700.
Liu, J. and Hodges, J. S. (2003) Posterior bimodality in the balanced one-way random-effects model. *J. R. Statist. Soc.* B, **65**, 247–255.
Martino, S. and Rue, H. (2008) Implementing approximate Bayesian inference for latent Gaussian models using integrated nested Laplace approximations; a manual for the `inla`-program. *Technical Report 2*. Department of Mathematical Sciences, Norwegian University of Science and Technology, Trondheim.
Matérn, B. (1986) Spatial variation. *Lect. Notes Statist.*, **36**.
McCausland, W. J. (2008) The Hessian method (highly efficient state smoothing, in a nutshell). *Report 2008-03*. Département de Sciences Économiques, Université de Montréal, Montréal.
McGrory, C. A., Titterington, D. M., Reeves, R. W. and Pettitt, A. N. (2008) Variational Bayes for estimating the parameters of a hidden Potts model. *Statist. Comput.*, to be published.
Mengersen, K. L. and Robert, C. P. (2003) The pinball sampler. In *Bayesian Statistics 7* (eds J. M. Bernardo, A. P. Dawid, J. O. Berger and M. West). Oxford: Oxford University Press.
Minka, T. (2001) A family of algorithms for approximate Bayesian inference. *PhD Dissertation*. Massachusetts Institute of Technology, Cambridge.
Möller, J., Syversveen, A. R. and Waagepetersen, R. P. (1998) Log-Gaussian Cox processes. *Scand. J. Statist.*, **25**, 451–482.
Naylor, J. C. and Smith, A. F. M. (1982) Applications of a method for the efficient computation of posterior distributions. *Appl. Statist.*, **31**, 214–225.
Naylor, J. C., Tremayne, A. and Marriott, J. (2008) Exploratory data analysis and model criticism with posterior plots. *Technical Report*. Royal Statistical Society Centre for Statistical Education, Nottingham.
Neal, R. M. (2001) Annealed importance sampling. *Statist. Comput.*, **11**, 125–139.

Nickisch, H. and Rasmussen, C. E. (2008) Approximations for binary gaussian process classification. *J. Mach. Learn. Res.*, **9**, 2035–2078.
Noh, M. and Lee, Y. (2007a) Robust modelling for inference from generalized linear model classes. *J. Am. Statist. Ass.*, **102**, 1059–1072.
Noh, M. and Lee, Y. (2007b) REML estimation for binary data in GLMMs. *J. Multiv. Anal.*, **98**, 896–915.
Noh, M., Lee, Y. and Pawitan, Y. (2005) Robust ascertainment-adjusted parameter estimation. *Genet. Epidem.*, **29**, 68–75.
Nott, D. J., Kohn, R. and Fielding, M. Approximating the marginal likelihood using copula. *Working Paper*. (Available from `http://arxiv.org/PS_cache/arxiv/pdf/0810/0810.5474v1.pdf`.)
Øigård, T. A., Rue, H. and Godtliebsen, F. (2006) Bayesian multiscale analysis for time series data. *Computnl Statist. Data Anal.*, **51**, 1719–1730.
Opper, M., Paquet, N. and Winther, O. (2009) Improving on expectation propagation. In *Advances in Neural Information Processing Systems*, vol. 21. Cambridge: MIT Press. To be published. (Available from `http://ulrichpaquet.com/Papers/ImprovingOnEP.pdf`.)
Opper, M. and Winther, O. (2000) Gaussian processes for classification: mean-field algorithms. *Neur. Computn*, **12**, 2665–2684.
Ormerod, J. T. (2008) On semiparametric regression and data mining. *PhD Thesis*. School of Mathematics and Statistics, University of New South Wales, Sydney.
Ormerod, J. T. and Wand, M. P. (2008) Variational approximations for logistic mixed models. In *Proc. 9th Iranian Statistics Conf., Isfahan*, pp. 450–467.
Pettitt, A. N., Weir, I. S. and Hart, A. (2002) A conditional autoregressive Gaussian process for irregularly spaced multivariate data. *Statist. Comput.*, **12**, 353–367.
Polson, N. G., Stroud, J. R. and Müller, P. (2008) Practical filtering with sequential parameter learning. *J. R. Statist. Soc.* B, **70**, 413–428.
Rasmussen, C. E. and Williams, C. K. I. (2006) *Gaussian Processes for Machine Learning*. Cambridge: MIT Press.
Reich, B. J., Hodges, J. S. and Carlin, B. P. (2007) Spatial analyses of periodontal data using conditionally autoregressive priors having two types of neighbor relationships. *J. Am. Statist. Ass.*, **102**, 44–55.
Rigby, R. A. and Stasinopoulos, D. M. (2005) Generalized additive models for location, scale and shape (with discussion). *Appl. Statist.*, **54**, 507–554.
Robert, C. (1992) *L'Analyse Statistique Bayésienne*. Paris: Economica.
Robert, C. (1994) *The Bayesian Choice*. New York: Springer.
Roberts, G. and Rosenthal, J. (2001) Optimal scaling for various Metropolis-Hastings algorithms. *Statist. Sci.*, **16**, 351–367.
Roberts, G. and Rosenthal, J. (2007) Variance-bounding Markov chains. *Ann. Appl. Probab.*, **18**, 1201–1214.
Roberts, G. and Tweedie, R. (1996a) Geometric convergence and central limit theorems for multidimensional Hastings and Metropolis algorithms. *Biometrika*, **83**, 95–110.
Roberts, G. and Tweedie, R. (1996b) Exponential convergence of Langevin diffusions and their discrete approximations. *Bernoulli*, **2**, 341–364.
Rue, H. (2001) Fast sampling of Gaussian Markov random fields. *J. R. Statist. Soc.* B, **63**, 325–338.
Rue, H. and Held, L. (2005) *Gaussian Markov Random Fields: Theory and Applications*. London: Chapman and Hall–CRC Press.
Rue, H. and Martino, S. (2007) Approximate Bayesian inference for hierarchical Gaussian Markov random fields models. *J. Statist. Planng Inf.*, **137**, 3177–3192.
Rue, H. Steinsland, I. and Erland, S. (2004) Approximating hidden Gaussian Markov random fields. *J. R. Statist. Soc.* B, **66**, 877–892.
Rue, H. and Tjelmeland, H. (2002) Fitting Gaussian Markov random fields to Gaussian fields. *Scand. J. Statist.*, **29**, 31–49.
Ruppert, D., Wand, M. P. and Carroll, R. J. (2003) *Semiparametric Regression*. New York: Cambridge University Press.
Saad, Y. (2003) *Iterative Methods for Sparse Linear Systems*. Philadelphia: Society for Industrial and Applied Mathematics.
Sahu, S. K., Gelfand, A. E. and Holland, D. M. (2007) High resolution space-time ozone modeling for assessing trends. *J. Am. Statist. Ass.*, **102**, 1221–1234.
Salter-Townshend, M. (2008) Fast approximate inverse Bayesian inference in non-parametric multivariate regression. *PhD Thesis*. Trinity College, Dublin.
Sampson, P. D. and Guttorp, P. (1992) Nonparametric estimation of nonstationary spatial covariance structure. *J. Am. Statist. Ass.*, **87**, 108–119.
Seeger, M. W. (2008) Bayesian inference and optimal design for the sparse linear model. *J. Mach. Learn. Res.*, **9**, 759–813.
Shen, Z. and McCullagh, P. (1995) Laplace approximation of high dimensional integrals. *J. R. Statist. Soc.* B, **57**, 749–760.
Shephard, N. and Pitt, M. K. (1997) Likelihood analysis of non-Gaussian measurement time series. *Biometrika*, **84**, 653–667.

Skaug, H. J. and Fournier, D. A. (2006) Automatic approximation of the marginal likelihood in non-Gaussian hierarchical models. *Computnl Statist. Data Anal.*, **5**, 699–709.
Smith, A. F. M., Skene, A. M., Shaw, J. E. H. and Naylor, J. C. (1987) Progress with numerical and graphical methods for practical Bayesian statistics. *Statistician*, **36**, 75–82.
Song, P. X.-K., Fan, Y. and Kalbfleisch, J. D. (2005) Maximization by parts in likelihood inference. *J. Am. Statist. Ass.*, **100**, 1145–1167.
Sørbye, S. H., Hindberg, K., Olsen, L. R. and Rue, H. (2009) Bayesian multiscale feature detection of log-spectral densities. Submitted to *Computnl Statist. Data Anal.*
Sorensen, D. and Gianola, D. (2002) *Likelihood, Bayesian and MCMC Methods in Quantitative Genetics*. New York: Springer.
Spiegelhalter, D. J., Best, N. G., Carlin, B. P. and der Linde, A. (2002) Bayesian measures of model complexity and fit (with discussion). *J. R. Statist. Soc.* B, **64**, 583–639.
Stein, M. L. (1999) *Interpolation of Spatial Data*. New York: Springer.
Stein, M. L., Chi, Z and Welty, L. J. (2004) Approximating likelihoods for large spatial data sets. *J. R. Statist. Soc.* B, **66**, 275–296.
Steinsland, I. and Jensen, H. (2005) Making inference from Bayesian animal models utilising Gaussian Markov random field properties. *Statistics Preprint 10/205*. Norwegian University of Science and Technology, Trondheim. (Available from `www.math.ntnu.no/~ingelins`.)
Stephens, M. (1997) Bayesian methods for mixtures of normal distributions. *DPhil Thesis*. University of Oxford, Oxford.
Stewart, G. W. (2001) *Matrix Algorithms*, vol. 2, *Eigensystems*. Philadelphia: Society for Industrial and Applied Mathematics.
Taylor, S. (1982) Financial returns modelled by the product of two stochastic processes: a study of daily sugar prices 1961-1979. In *Time Series Analysis: Theory and Practice 1*, pp. 203–226. Amsterdam: Elsevier.
Tierney, L. and Kadane, J. B. (1986) Accurate approximations for posterior moments and marginal densities. *J. Am. Statist. Ass.*, **81**, 82–86.
Tierney, L., Kass, R. and Kadane, J. (1989) Fully exponential Laplace approximations to expectations and variances of non-positive functions. *J. Am. Statist. Ass.*, **84**, 710–716.
Titterington, D. M., Smith, A. F. M. and Makov, U. E. (1985) *Statistical Analysis of Finite Mixture Distributions*. Chichester: Wiley.
Trotter, H. F. and Tukey, J. W. (1956) Conditional Monte Carlo for normal samples. In *Proc. Symp. Monte Carlo Methods* (ed. H. A. Meyer), pp. 64–79. New York: Wiley.
Vecchia, A. V. (1988) Estimation and model identification for continuous spatial processes. *J. R. Statist. Soc.* B, **50**, 297–312.
Wakefield, J. (1998) Discussion on 'Some algebra and geometry for hierarchical models, applied to diagnostics' (by J. S. Hodges). *J. R. Statist. Soc.* B, **60**, 523–526.
Wang, B. and Titterington, 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. Society for Artificial Intelligence and Statistics.
Weir, I. S. and Pettitt, A. N. (2000) Binary probability maps using a hidden conditional autoregressive Gaussian process with an application to Finnish common toad data. *Appl. Statist.*, **49**, 473–484.
Whittle, P. (1954) On stationary processes in the plane. *Biometrika*, **41**, 434–449.
Woolrich, M. W. and Behrens, T. E. (2006) Variational Bayes inference of spatial mixture models for segmentation. *IEEE Trans. Med. Imgng*, **25**, 1380–1391.
Yun, S. and Lee, Y. (2006) Robust estimation in mixed linear models with non-monotone missingness. *Statist. Med.*, **25**, 3877–3892.