Bayesian Computing with INLA: A Review

Håvard Rue,[1] Andrea Riebler,[1] Sigrunn H. Sørbye,[2] Janine B. Illian,[3] Daniel P. Simpson,[4] and Finn K. Lindgren[5]

[1]Department of Mathematical Sciences, Norwegian University of Science and Technology, N-7491 Trondheim, Norway; email: hrue@math.ntnu.no

[2]Department of Mathematics and Statistics, The Arctic University of Norway, 9037 Tromsø, Norway

[3]Centre for Research into Ecological and Environmental Modelling, School of Mathematics and Statistics, University of St. Andrews, KY16 9LZ Fife, United Kingdom

[4]Department of Mathematical Sciences, University of Bath, BA2 7AY Bath, United Kingdom

[5]School of Mathematics, The University of Edinburgh, EH9 3FD Edinburgh, United Kingdom

Annu. Rev. Stat. Appl. 2017. 4:395–421

First published online as a Review in Advance on December 23, 2016

The *Annual Review of Statistics and Its Application* is online at statistics.annualreviews.org

This article's doi:
10.1146/annurev-statistics-060116-054045

Keywords

Gaussian Markov random fields, Laplace approximations, approximate Bayesian inference, latent Gaussian models, numerical integration, sparse matrices

Abstract

The key operation in Bayesian inference is to compute high-dimensional integrals. An old approximate technique is the Laplace method or approximation, which dates back to Pierre-Simon Laplace (1774). This simple idea approximates the integrand with a second-order Taylor expansion around the mode and computes the integral analytically. By developing a nested version of this classical idea, combined with modern numerical techniques for sparse matrices, we obtain the approach of integrated nested Laplace approximations (INLA) to do approximate Bayesian inference for latent Gaussian models (LGMs). LGMs represent an important model abstraction for Bayesian inference and include a large proportion of the statistical models used today. In this review, we discuss the reasons for the success of the INLA approach, the `R-INLA` package, why it is so accurate, why the approximations are very quick to compute, and why LGMs make such a useful concept for Bayesian computing.

1. INTRODUCTION

A key obstacle in Bayesian statistics is to actually do the Bayesian inference. From a mathematical point of view, the inference step is easy, transparent, and defined by first principles: We simply update prior beliefs about the unknown parameters with available information in observed data, and obtain the posterior distribution for the parameters. Based on the posterior, we can compute relevant statistics for the parameters of interest, including marginal distributions, means, variances, quantiles, credibility intervals, et cetera. In practice, this is much easier said than done.

The introduction of simulation-based inference, through the idea of Markov chain Monte Carlo (MCMC) (Robert & Casella 1999), hit the statistical community in the early 1990s and represented a major breakthrough in Bayesian inference. MCMC provided a general recipe to generate samples from posteriors by constructing a Markov chain with the target posterior as the stationary distribution. This made it possible (in theory) to extract and compute whatever one could wish for. Additional major developments have paved the way for popular user-friendly MCMC tools, such as `WinBUGS` (Spiegelhalter et al. 1995), `JAGS` (Plummer 2016), and the new initiative `Stan` (Stan Development Team 2015), which uses Hamiltonian Monte Carlo. Armed with these and similar tools, Bayesian statistics has quickly grown in popularity, and it is now well represented in all the major research journals in all branches of statistics.

In our opinion, however, from the point of view of applied users, the impact of the Bayesian revolution has been less apparent. This is not a statement about how Bayesian statistics itself is viewed by that community, but about its rather cumbersome inference, which still requires a large amount of CPU—and hence human—time, as well as tweaking of simulation and model parameters to get it right. Re-running many alternative models becomes even more cumbersome, making the iterative process of model building in statistical analysis impossible (Box & Tiao 1973, section 1.1.4). For this reason, simulation-based inference (and hence, in most cases, also Bayesian statistics) has too often been avoided as being practically infeasible.

In this article, we review a different take on doing Bayesian inference that recently has facilitated the increased use of Bayesian modeling within the community of applied users. This approach is restricted to the specific class of latent Gaussian models (LGMs) which, as will be clear soon, includes a wide variety of commonly applied statistical models, making this restriction less limiting than it might appear at first sight. The crucial point here is that we can derive integrated nested Laplace approximation (INLA) methodology for LGMs, a deterministic approach to approximate Bayesian inference. INLA performs inference within a reasonable time frame and in most cases is both faster and more accurate than MCMC alternatives. This might seem like a contradiction to most readers, who are used to trading speed for accuracy. The corresponding R package (`R-INLA`, see **http://www.r-inla.org**) has turned out to be very popular in applied sciences and applied statistics, and has become a versatile tool for quick and reliable Bayesian inference.

Recent examples of applications using the `R-INLA` package for statistical analysis include disease mapping (Schrödle & Held 2011a,b; Ugarte et al. 2014, 2016; Papoila et al. 2014; Goicoa et al. 2016; Riebler et al. 2016); age-period-cohort models (Riebler & Held 2016); a study of the evolution of the Ebola virus (Santermans et al. 2016); the relationships between access to housing, health, and well-being in cities (Kandt et al. 2016); the prevalence and correlates of intimate partner violence against men in Africa (Tsiko 2016); a search for evidence of gene expression heterosis (Niemi et al. 2015); analysis of traffic pollution and hospital admissions in London (Halonen et al. 2016); early transcriptome changes in maize primary root tissues in response to moderate water deficit conditions by RNA sequencing (Opitz et al. 2016); performance of inbred and hybrid genotypes in plant breeding and genetics (Lithio & Nettleton 2015); a study of Norwegian emergency wards (Goth et al. 2014); effects of measurement errors (Muff et al. 2015, Muff & Keller 2015,

Kröger et al. 2016); network meta-analysis (Sauter & Held 2015); time-series analysis of genotyped human campylobacteriosis cases from the Manawatu region of New Zealand (Friedrich et al. 2016); modeling of parrotfish habitats (NC Roos et al. 2015); Bayesian outbreak detection (Salmon et al. 2015); long-term trends in the number of Monarch butterflies (Crewe & Mccracken 2015); long-term effects on hospital admission and mortality of road traffic noise (Halonen et al. 2015); spatio-temporal dynamics of brain tumors (Iulian et al. 2015); ovarian cancer mortality (García-Pérez et al. 2015); the effect of preferential sampling on phylodynamic inference (Karcher et al. 2016); analysis of the impact of climate change on abundance trends in central Europe (Bowler et al. 2015); investigation of drinking patterns in US counties from 2002 to 2012 (Dwyer-Lindgren et al. 2015); resistance and resilience of terrestrial birds in drying climates (Selwood et al. 2015); cluster analysis of population amyotrophic lateral sclerosis risk (Rooney et al. 2015); malaria infection in Africa (Noor et al. 2014); effects of fragmentation on infectious disease dynamics (Jousimo et al. 2014); soil-transmitted helminth infection in sub-Saharan Africa (Karagiannis-Voules et al. 2015); analysis of the effect of malaria control on *Plasmodium falciparum* in Africa between 2000 and 2015 (Bhatt et al. 2015); adaptive prior weighting in generalized regression (Held & Sauter 2016); analysis of hand, foot, and mouth disease surveillance data in China (Bauer et al. 2016); estimation of the biomass of anchovies in the coast of Perú (Quiroz et al. 2015); and many others.

We review the key components that make up INLA in Section 2, and in Section 3 we combine these to outline why, and in which situations, INLA works. In Section 4 we show some examples of the use of `R-INLA` and discuss some special features that expand the class of models that `R-INLA` can be applied to. In Section 5, we discuss the challenge of choosing appropriate priors in Bayesian methodology, and, in particular, reason why it is important to provide better suggestions for default priors. We conclude with a general discussion and outlook in Section 6.

2. BACKGROUND ON THE KEY COMPONENTS

In this section, we review the key components of the INLA approach to approximate Bayesian inference. We introduce these concepts using a top-down approach, starting with LGMs and what type of statistical models may be viewed as LGMs. We also discuss the types of Gaussians/Gaussian processes that are computationally efficient within this formulation, and illustrate Laplace approximation to perform integration—a method that has been around for a very long time and proves to be a key ingredient in the methodology we review here.

Owing to the top-down structure of this text, we occasionally have to mention specific concepts before properly introducing and/or defining them—we ask the reader to bear with us in these cases.

2.1. Latent Gaussian Models

The concept of LGMs represents a very useful abstraction subsuming a large class of statistical models, in the sense that the task of statistical inference can be unified for the entire class (Rue et al. 2009). The model class abstraction is obtained using a three-stage hierarchical model formulation, in which observations $\boldsymbol{y}$ can be assumed to be conditionally independent, given a latent Gaussian random field $\boldsymbol{x}$ and hyperparameters $\boldsymbol{\theta_1}$,

$$\boldsymbol{y} \mid \boldsymbol{x}, \boldsymbol{\theta}_1 \sim \prod_{i \in \mathcal{I}} \pi(y_i \mid x_i, \boldsymbol{\theta}_1).$$

The versatility of the model class relates to the specification of the latent Gaussian field:

$$\boldsymbol{x} \mid \boldsymbol{\theta}_2 \sim \mathcal{N}\left(\boldsymbol{\mu}(\boldsymbol{\theta}_2),\, \boldsymbol{Q}^{-1}(\boldsymbol{\theta}_2)\right),$$

which includes all random terms in a statistical model, describing the underlying dependence structure of the data. The hyperparameters $\boldsymbol{\theta} = (\boldsymbol{\theta}_1, \boldsymbol{\theta}_2)$ control the latent Gaussian field and/or the likelihood for the data, and the posterior reads

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

We make the following critical assumptions:

1. The number of hyperparameters $|\boldsymbol{\theta}|$ is small, typically 2 to 5, but not exceeding 20.
2. The distribution of the latent field $\boldsymbol{x}|\boldsymbol{\theta}$ is Gaussian and required to be a Gaussian Markov random field (GMRF) (or to be close to one) when the dimension n is high (10^3–10^5).
3. The data $\boldsymbol{y}$ are mutually conditionally independent of $\boldsymbol{x}$ and $\boldsymbol{\theta}$, implying that each observation y_i only depends on one component of the latent field, for example, x_i. Most components of $\boldsymbol{x}$ will not be observed.

These assumptions are required both for computational reasons and to ensure, with a high degree of certainty, that the approximations we describe below are accurate.

2.2. Additive Models

Now, how do LGMs relate to other better-known statistical models? Broadly speaking, they are an umbrella class generalizing the large number of related variants of additive and/or generalized (linear) models. For instance, interpreting the likelihood $\pi(y_i|x_i, \boldsymbol{\theta})$ so that y_i only depends on its linear predictor x_i yields the generalized linear model setup. We can interpret $\{x_i, i \in \mathcal{I}\}$ as η_i (the linear predictor), which itself is additive with respect to other effects,

$$\eta_i = \mu + \sum_j \beta_j z_{ij} + \sum_k f_{k,jk(i)}. \tag{2}$$

Here, μ is the overall intercept and $\boldsymbol{z}$ are fixed covariates with linear effects $\{\beta_j\}$. The difference between this formulation and an ordinary generalized linear model lies in the terms $\{f_k\}$, which are used to represent specific Gaussian processes. We label each f_k as a model component, in which element j contributes to the ith linear predictor. Examples of model components f_k include auto-regressive time-series models, stochastic spline models and models for smoothing, measurement error models, random effects models with different types of correlations, and spatial models. We assume that the model components are a priori independent, the fixed effects $(\mu, \boldsymbol{\beta})$ have a joint Gaussian prior, and the fixed effects are a priori independent of the model components.

The key is now that the model formulation in Equation 2 and LGMs relate to the same class of models when we assume Gaussian priors for the intercept and the parameters of the fixed effects. The joint distribution of

$$\boldsymbol{x} = (\boldsymbol{\eta}, \mu, \boldsymbol{\beta}, \boldsymbol{f}_1, \boldsymbol{f}_2, \ldots) \tag{3}$$

is then Gaussian, and also nonsingular if we add a tiny noise term in Equation 2. This yields the latent field $\boldsymbol{x}$ in the hierarchical LGM formulation. Clearly, $\dim(\boldsymbol{x}) = n$ can easily get large, as it equals the number of observations, plus the intercept(s) and fixed effects, plus the sum of the dimension of all the model components.

The hyperparameters $\boldsymbol{\theta}$ comprise the parameters of the likelihood and the model components. A likelihood family and each model component typically have between zero and two hyperparameters each. These parameters often include some kind of variance, scale, or correlation parameters. Conveniently, the number of hyperparameters is typically small and, further, does not depend on

the dimension of the latent field n or the number of observations. This is crucial for computational efficiency, as even with a big dataset, the number of hyperparameters remains constant and Assumption 1 still holds.

2.3. Gaussian Markov Random Fields

In practice, the latent field not only should be Gaussian, but also should be a (sparse) GMRF; the reader is directed to Rue & Held (2005, 2010) and Held & Rue (2010) for an introduction to GMRFs. A GMRF $\boldsymbol{x}$ is simply a Gaussian with additional conditional independence properties, meaning that x_i and x_j are conditionally independent given the remaining elements $\boldsymbol{x}_{-ij}$, for quite a few $\{i, j\}$s. The simplest nontrivial example is the first-order auto-regressive model, $x_t = \phi x_{t-1} + \epsilon_t$, $t = 1, 2, \ldots, m$, having Gaussian innovations $\boldsymbol{\epsilon}$. For this model, the correlation between x_t and x_s is $\phi^{|s-t|}$, and the resulting $m \times m$ covariance matrix is dense. However, x_s and x_t are conditionally independent given $\boldsymbol{x}_{-st}$, for all $|s - t| > 1$. In the Gaussian case, a very useful consequence of conditional independence is that this results in zeros for pairs of conditionally independent values in the precision matrix (the inverse of the covariance matrix). Considering GMRFs provides a huge computational benefit, as calculations involving a dense $m \times m$ matrix are much more costly than when a sparse matrix is used. In the auto-regressive example, the precision matrix is tridiagonal and can be factorized in $\mathcal{O}(m)$ time, whereas we need $\mathcal{O}(m^3)$ in the general dense case. Memory requirement is also reduced from $\mathcal{O}(m^2)$ to $\mathcal{O}(m)$, which makes it much easier to run larger models. For models with a spatial structure, the cost is $\mathcal{O}(m^{3/2})$ paired with a $\mathcal{O}(m \log(m))$ memory requirement. In general, the computational cost depends on the actual sparsity pattern in the precision matrix, hence it is hard to provide precise estimates.

2.4. Additive Models and Gaussian Markov Random Fields

In the construction of additive models including GMRFs, the following fact is one of the convenient results that are exploited in INLA:

> The joint distribution for $\boldsymbol{x}$ in Equation 3 is also a GMRF, and its precision matrix consists of sums of the precision matrices of the fixed effects and the other model components.

We will see below that we need to form the joint distribution of the latent field many times, as it depends on the hyperparameters $\boldsymbol{\theta}$. Hence, it is essential that this can be done efficiently, avoiding computationally costly matrix operations. Being able to simply treat the joint distribution as a GMRF with a precision matrix that is easy to compute is one of the key reasons why the INLA approach is so efficient. Also, the sparse structure of the precision matrix boosts computational efficiency, compared with operations on dense matrices.

To illustrate more clearly what happens, let us consider the following simple example,

$$\eta_i = \mu + \beta z_i + f_{1j_1(i)} + f_{2j_2(i)} + \epsilon_i, \quad i = 1, \ldots, n, \tag{4}$$

where we have added a small amount of noise ϵ_i. The two model components $f_{1j_1(i)}$ and $f_{2j_2(i)}$ have sparse precision matrices $\boldsymbol{Q}_1(\boldsymbol{\theta})$ and $\boldsymbol{Q}_2(\boldsymbol{\theta})$ of dimension $m_1 \times m_1$ and $m_2 \times m_2$, respectively. Let τ_μ and τ_β be the (fixed) prior precisions for μ and β. We can express Equation 4 using matrices

$$\boldsymbol{\eta} = \mu \mathbf{1} + \beta \boldsymbol{z} + \boldsymbol{A}_1 \boldsymbol{f}_1 + \boldsymbol{A}_2 \boldsymbol{f}_2 + \boldsymbol{\epsilon}.$$

Here, $\boldsymbol{A}_1$ (and similarly for $\boldsymbol{A}_2$) is a $n \times m_1$ sparse matrix, which is zero except for exactly one in each row. The joint precision matrix of $(\boldsymbol{\eta}, \boldsymbol{f}_1, \boldsymbol{f}_2, \beta, \mu)$ is straightforward to obtain by rewriting

$$\exp\Big(-\frac{\tau_\epsilon}{2}(\boldsymbol{\eta}-(\mu\mathbf{1}+\beta\boldsymbol{z}+\boldsymbol{A}_1\boldsymbol{f}_1+\boldsymbol{A}_2\boldsymbol{f}_2))^T(\boldsymbol{\eta}-(\mu\mathbf{1}+\beta\boldsymbol{z}+\boldsymbol{A}_1\boldsymbol{f}_1+\boldsymbol{A}_2\boldsymbol{f}_2))$$
$$-\frac{\tau_\mu}{2}\mu^2-\frac{\tau_\beta}{2}\beta^2-\frac{1}{2}\boldsymbol{f}_1^T\boldsymbol{Q}_1(\boldsymbol{\theta})\boldsymbol{f}_1-\frac{1}{2}\boldsymbol{f}_2^T\boldsymbol{Q}_2(\boldsymbol{\theta})\boldsymbol{f}_2\Big)$$

into

$$\exp\left(-\frac{1}{2}(\boldsymbol{\eta},\boldsymbol{f}_1,\boldsymbol{f}_2,\beta,\mu)^T\boldsymbol{Q}_{\text{joint}}(\boldsymbol{\theta})(\boldsymbol{\eta},\boldsymbol{f}_1,\boldsymbol{f}_2,\beta,\mu)\right)$$

where

$$\boldsymbol{Q}_{\text{joint}}(\boldsymbol{\theta})=\begin{bmatrix}\tau_\epsilon\boldsymbol{I} & \tau_\epsilon\boldsymbol{A}_1 & \tau_\epsilon\boldsymbol{A}_2 & \tau_\epsilon\boldsymbol{I}\boldsymbol{z} & \tau_\epsilon\boldsymbol{I}\mathbf{1}\\ & \boldsymbol{Q}_1(\boldsymbol{\theta})+\tau_\epsilon\boldsymbol{A}_1\boldsymbol{A}_1^T & \tau_\epsilon\boldsymbol{A}_1\boldsymbol{A}_2^T & \tau_\epsilon\boldsymbol{A}_1\boldsymbol{z} & \tau_\epsilon\boldsymbol{A}_1\mathbf{1}\\ & & \boldsymbol{Q}_2(\boldsymbol{\theta})+\tau_\epsilon\boldsymbol{A}_2\boldsymbol{A}_2^T & \tau_\epsilon\boldsymbol{A}_2\boldsymbol{z} & \tau_\epsilon\boldsymbol{A}_2\mathbf{1}\\ & \text{sym.} & & \tau_\beta+\tau_\epsilon\boldsymbol{z}^T\boldsymbol{z} & \tau_\epsilon\boldsymbol{z}^T\mathbf{1}\\ & & & & \tau_\mu+\tau_\epsilon\mathbf{1}^T\mathbf{1}\end{bmatrix}.$$

The dimension is $n+m_1+m_2+2$. Concretely, the convenient result mentioned above implies that the only matrices that need to be multiplied are the $\boldsymbol{A}$-matrices, which are extremely sparse and contain only one nonzero element in each row. These matrix products do not depend on $\boldsymbol{\theta}$ and hence only need to be computed once. The joint precision matrix only depends on $\boldsymbol{\theta}$ through $\boldsymbol{Q}_1(\boldsymbol{\theta})$ and $\boldsymbol{Q}_2(\boldsymbol{\theta})$, and as $\boldsymbol{\theta}$ change, the computational cost of recomputing $\boldsymbol{Q}_{\text{joint}}(\boldsymbol{\theta})$ is negligible.

The sparsity of $\boldsymbol{Q}_{\text{joint}}(\boldsymbol{\theta})$ illustrates how the additive structure of the model facilitates computational efficiency. For simplicity, assume $n=m_1=m_2$, and denote by e_1 and e_2 the average number of nonzero elements in a row of $\boldsymbol{Q}_1(\boldsymbol{\theta})$ and $\boldsymbol{Q}_2(\boldsymbol{\theta})$, respectively. An upper bound for the number of nonzero terms in $\boldsymbol{Q}_{\text{joint}}(\boldsymbol{\theta})$ is $n(19+e_1+e_2)+4$. Approximately, this gives on average only $(19+e_1+e_2)/3$ nonzero elements for a row in $\boldsymbol{Q}_{\text{joint}}(\boldsymbol{\theta})$, which is very sparse.

2.5. Laplace Approximations

The Laplace approximation or method is an old technique for the approximation of integrals; see Barndorff-Nielsen & Cox (1989, chapter 3.3) for a general introduction. The setting is as follows: The aim is to approximate the integral,

$$I_n=\int_x\exp(nf(x))\,\mathrm{d}x$$

as $n\to\infty$. Let x_0 be the point in which $f(x)$ has its maximum, then

$$I_n\approx\int_x\exp\left(n\left(f(x_0)+\frac{1}{2}(x-x_0)^2f''(x_0)\right)\right)\mathrm{d}x \tag{5}$$

$$=\exp(nf(x_0))\sqrt{\frac{2\pi}{-nf''(x_0)}}=\widetilde{I}_n. \tag{6}$$

The idea is simple but powerful: Approximate the target with a Gaussian, matching the mode and the curvature at the mode. By interpreting $nf(x)$ as the sum of log-likelihoods and x as the unknown parameter, the Gaussian approximation will be exact as $n\to\infty$, if the central limit theorem holds. The extension to higher-dimensional integrals is immediate, and the error turns out to be

$$I_n=\widetilde{I}_n\left(1+\mathcal{O}(n^{-1})\right).$$

This is a good result for two reasons. The error is relative and with rate n^{-1}, as opposed to an additive error and a rate $n^{-1/2}$, which are common in simulation-based inference.

The Laplace approximation used to be a key tool for doing high-dimensional integration in pre-MCMC times but quickly went out of fashion when MCMC entered the stage. But how does it relate to what we endeavor to do here? Let's assume that we would like to compute a marginal distribution $\pi(\gamma_1)$ from a joint distribution $\pi(\boldsymbol{\gamma})$:

$$\begin{aligned}\pi(\gamma_1) &= \frac{\pi(\boldsymbol{\gamma})}{\pi(\boldsymbol{\gamma}_{-1}|\gamma_1)} \\ &\approx \left.\frac{\pi(\boldsymbol{\gamma})}{\pi_G(\boldsymbol{\gamma}_{-1};\ \boldsymbol{\mu}(\gamma_1),\ \boldsymbol{Q}(\gamma_1))}\right|_{\boldsymbol{\gamma}_{-1}=\boldsymbol{\mu}(\gamma_1)},\end{aligned} \tag{7}$$

where we have exploited the fact that we approximate $\pi(\boldsymbol{\gamma}_{-1}|\gamma_1)$ with a Gaussian. In the context of the LGMs, we have $\boldsymbol{\gamma} = (\boldsymbol{x}, \boldsymbol{\theta})$. Tierney & Kadane (1986) show that if $\pi(\boldsymbol{\gamma}) \propto \exp(nf_n(\boldsymbol{\gamma}))$, that is, if $f_n(\boldsymbol{\gamma})$ is the average log likelihood, the relative error of the normalized approximation Equation 7, within a $\mathcal{O}(n^{-1/2})$ neighborhood of the mode, is $\mathcal{O}(n^{-3/2})$. In other words, if we have n replicated data from the same parameters $\boldsymbol{\gamma}$, we can compute posterior marginals with a relative error of $\mathcal{O}(n^{-3/2})$, assuming the numerical error to be negligible. This is an extremely positive result, but unfortunately the underlying assumptions usually do not hold:

1. Instead of replicated data from the same model, we may have one replicate from one model (as is common in spatial statistics) or several observations from similar models.
2. The implicit assumption in the above result is also that $|\boldsymbol{\gamma}|$ is fixed as $n \to \infty$. However, there is only one realization for each observation/location in the random effect(s) in the model, implying that $|\gamma|$ grows with n.

Is it still possible to gain insight into when the Laplace approximation would give good results, even if these assumptions do not hold? First, let us replace replicated observations from the same model with several observations from similar models, where we deliberately use the term "similar" in a loose sense. We can borrow strength across variables that we a priori assume to be similar, for example, by smoothing over time or over space. In this case, the resulting linear predictors for two observations could differ in only one realization of the random effect. In addition, borrowing strength and smoothing can reduce the effect of the model dimension growing with n, because the effective dimension can then grow much more slowly with n.

Another way to interpret the accuracy in computing posterior marginals using Laplace approximations is to not look at the error rate but rather at the implicit constant upfront. If the posterior is close to a Gaussian density, the results will be more accurate compared with a density that is very different from a Gaussian. This is similar to the convergence for the central limit theorem where convergence is faster if relevant properties such as unimodality, symmetry, and tail behavior are satisfied; see, for example, Baghishani & Mohammadzadeh (2012). Similarly, in the context here, unimodality is necessary because we approximate the integrand with a Gaussian. Symmetry helps because the Gaussian distribution is symmetric, and heavier tails will be missed by the Gaussian. For example, assume

$$\exp(nf_n(\boldsymbol{\gamma})) = \prod_i \text{Poisson}(y_i; \lambda = \exp(\gamma_1 + \gamma_2 z_i))$$

with centered covariates $\boldsymbol{z}$. We then expect better accuracy for $\pi(\gamma_1)$, having high values for y_i compared with low low values. With high counts, the Poisson distribution is approximately Gaussian and almost symmetric. Low counts are more challenging, because the likelihood for $y_i = 0$ and $z_i = 0$ is proportional to $\exp(-\exp(\gamma_1))$, which has a maximum value at $\gamma_1 = -\infty$. The situation is similar for binomial data of size m, where low values of m are more challenging than high values of m. Theoretical results for the current rather vague context are difficult to obtain and constitute a largely unsolved problem (see, for example, Shun & McCullagh 1995, Kauermann et al. 2009, Ogden 2016).

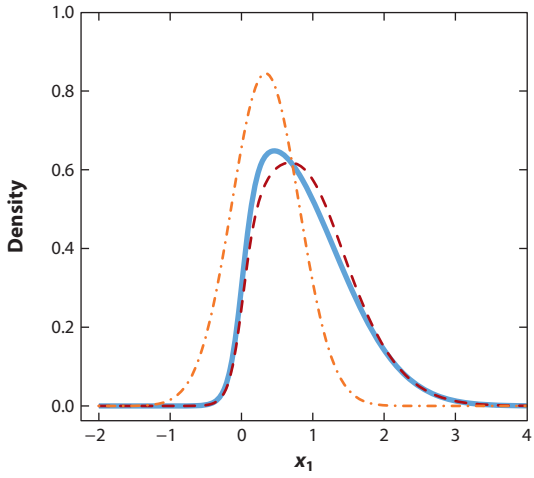

Figure 1

The true marginal (*solid blue line*), the Laplace approximation (*dashed red line*) and the Gaussian approximation (*dot-dashed orange line*).

Let us now discuss a simplistic, but realistic, model in two dimensions $\boldsymbol{x} = (x_1, x_2)^T$, where

$$\pi(\boldsymbol{x}) \propto \exp\left(-\frac{1}{2}\boldsymbol{x}^T \begin{bmatrix} 1 & \rho \\ \rho & 1 \end{bmatrix} \boldsymbol{x}\right) \prod_{i=1}^{2} \frac{\exp(c\, x_i)}{1 + \exp(c\, x_i)} \tag{8}$$

for a constant $c > 0$ and $\rho \geq 0$. This is the same functional form as we get from two Bernoulli successes, using a logit-link. Using the constant c is an alternative to scaling the Gaussian part, and the case where $\rho < 0$ is similar. The task now is to approximate $\pi(x_1) = \pi(x_1, x_2)/\pi(x_2|x_1)$, using Equation 7. Here, the Gaussian approximation is indexed by x_1, and we use one Laplace approximation for each value of x_1. The likelihood term has a mode at (∞, ∞), hence the posterior is a compromise between this and the Gaussian prior centered at $(0, 0)$.

We first demonstrate that even if the Gaussian approximation matching the mode of $\pi(\boldsymbol{x})$ is not so good, the Laplace approximation, which uses a sequence of Gaussian approximations, can do much better. Let $\rho = 1/2$ and $c = 10$ (which is an extreme value). The resulting marginal for x_1 (solid), the Laplace approximation of it (dashed) and Gaussian approximation (dot-dashed), are shown in **Figure 1**. The Gaussian approximation fails both to locate the marginal correctly and also, of course, it also fails to capture the skewness that is present. In spite of this, the sequence of Gaussian approximations used in the Laplace approximation performs much better and only seems to run into slight trouble where the curvature of the likelihood changes abruptly.

An important feature of Equation 7 is its properties in the limiting cases $\rho \to 0$ and $\rho \to 1$. When $\rho = 0$, x_1 and x_2 become independent and $\pi(x_2|x_1)$ does not depend on x_1. Hence, Equation 7 is exact up to a numerical approximation of the normalizing constant. In the other limiting case, $\rho \to 1$, $\pi(x_2|x_1)$ is the point mass at $x_2 = x_1$, and Equation 7 is again exact up to numerical error. This illustrates the convenient property of Equation 7, being exact in the two limiting cases of weak and strong dependence, respectively. This indicates that the approximation should not fail too badly for intermediate dependence. **Figure 2** illustrates the Laplace approximation and the true marginals, using $\rho = 0.05, 0.4, 0.8$ and 0.95, and $c = 10$. For $\rho = 0.05$ (**Figure 2*a***) and $\rho = 0.95$ (**Figure 2*d***), the approximation is almost perfect, whereas the error is largest for intermediate dependence where $\rho = 0.4$ (**Figure 2*b***) and $\rho = 0.8$ (**Figure 2*c***).

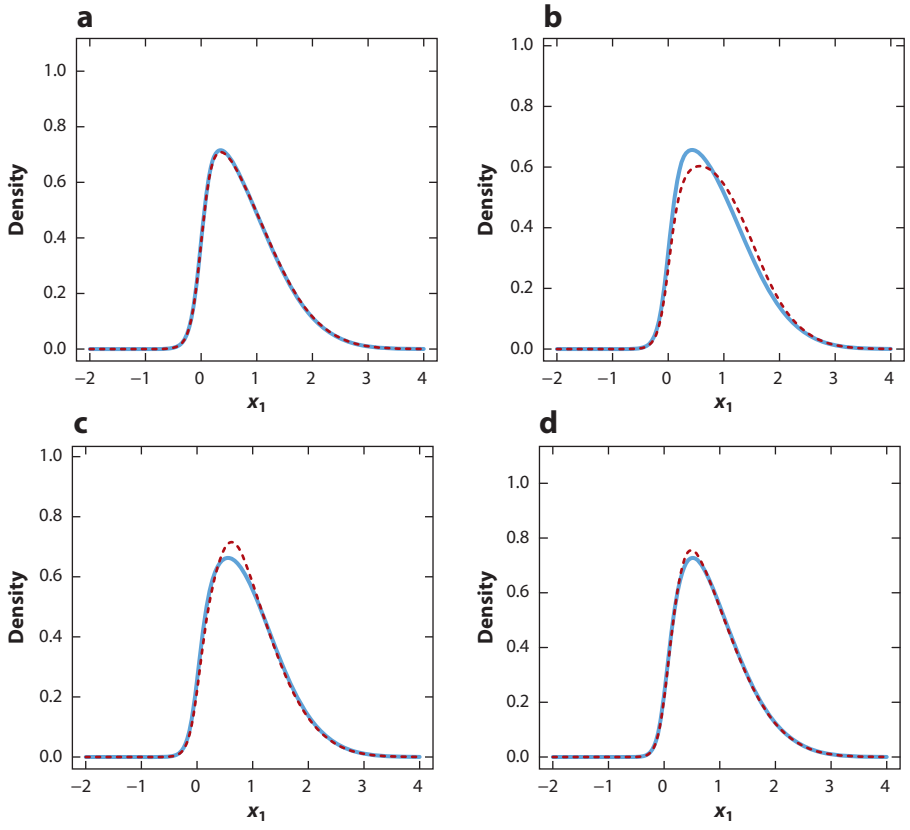

Figure 2

The true marginal (*solid blue line*) and the Laplace approximation (*dashed red line*), for $\rho = 0.05$ (*a*), 0.4 (*b*), 0.8 (*c*) and 0.95 (*d*).

3. PUTTING IT ALL TOGETHER: INLA

With all the key components at hand, we now can put all these together to illustrate how they are combined to form INLA. The main aim of Bayesian inference is to approximate the posterior marginals

$$\pi(\theta_j|\boldsymbol{y}), \quad j = 1, \ldots, |\boldsymbol{\theta}|, \qquad \pi(x_i|\boldsymbol{y}), \quad i = 1, \ldots, n. \tag{9}$$

Our approach is tailored to the structure of LGMs, where $|\boldsymbol{\theta}|$ is low-dimensional, $\boldsymbol{x}|\boldsymbol{\theta}$ is a GMRF, and the likelihood is conditional independent in the sense that y_i only depends on one x_i and $\boldsymbol{\theta}$. From the discussion in Section 2.5, we know that we should aim to apply Laplace approximation only to near-Gaussian densities. For LGMs, it turns out that we can reformulate our problem as series of subproblems that allows us to use Laplace approximations on these. To illustrate the general principal, consider an artificial model

$$\eta_i = g(\beta) u_{j(i)},$$

where $y_i|\eta_i \sim \text{Poisson}(\exp(\eta_i))$, $i = 1, \ldots, n$, $\beta \sim \mathcal{N}(0, 1)$, $g(\cdot)$ is some well-behaved monotone function, and $\boldsymbol{u} \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{Q}^{-1})$. The index mapping $j(i)$ is made such that the dimension of $\boldsymbol{u}$ is fixed and does not depend on n, and all u_js are observed roughly the same number of times.

Computation of the posterior marginals for β and all u_j is problematic, because we have a product of a Gaussian and a non-Gaussian (which is rather far from a Gaussian). Our strategy is to break down the approximation into smaller subproblems and only apply the Laplace approximation where the densities are almost Gaussian. The key idea is to use conditioning, here on β. Then

$$\pi(\beta|\boldsymbol{y}) \propto \pi(\beta) \int \prod_{i=1}^{n} \pi\left(y_i|\lambda_i = \exp\left(g(\beta)u_{j(i)}\right)\right) \times \pi(\boldsymbol{u})\, \mathrm{d}\boldsymbol{u}. \tag{10}$$

The integral we need to approximate should be close to Gaussian, because the integrand is a Poisson-count correction of a Gaussian prior. The marginals for each u_j can be expressed as

$$\pi(u_j|\boldsymbol{y}) = \int \pi(u_j|\beta, \boldsymbol{y}) \times \pi(\beta|\boldsymbol{y})\, \mathrm{d}\beta. \tag{11}$$

Note that we can compute the integral directly, because β is one-dimensional. Similar to Equation 10, we have that

$$\pi(\boldsymbol{u}|\beta, \boldsymbol{y}) \propto \prod_{i=1}^{n} \pi\left(y_i|\lambda_i = \exp\left(g(\beta)u_{j(i)}\right)\right) \times \pi(\boldsymbol{u}), \tag{12}$$

which should be close to a Gaussian. Approximating $\pi(u_j|\beta, \boldsymbol{y})$ involves approximation of the integral of this density in one dimension less, because u_j is fixed. Again, this is close to Gaussian.

The key lesson learned is that we can break down the problem into three subproblems.

1. Approximate $\pi(\beta|\boldsymbol{y})$ using Equation 10.
2. Approximate $\pi(u_j|\beta, \boldsymbol{y})$, for all j and for all required values of β, from Equation 12.
3. Compute $\pi(u_j|\boldsymbol{y})$ for all j using the results from the two first steps, combined with numerical integration Equation 11.

The price we have to pay for taking this approach is increased complexity; for example, step 2 needs to be computed for all values of β that are required. We also need to integrate out the βs in Equation 11, numerically. If we remain undeterred by the increased complexity, the benefit of this procedure is clear; we only apply Laplace approximations to densities that are near-Gaussian, replacing complex dependencies with conditioning and numerical integration.

The big question is whether we can pursue the same principle for LGMs, and whether we can make it computationally efficient by accepting appropriate trade-offs that allow us to still be sufficiently exact. The answer is yes in both cases. The strategy outlined above can be applied to LGMs by replacing β with $\boldsymbol{\theta}$, and $\boldsymbol{u}$ with $\boldsymbol{x}$, and then deriving approximations to the Laplace approximations and the numerical integration. The resulting approximation is fast to compute, with little loss of accuracy. We now discuss the main ideas for each step—skipping some practical and computational details that are somewhat involved but still relatively straightforward, using every trick in the book for GMRFs.

3.1. Approximating the Posterior Marginals for the Hyperparameters

Because the aim is to compute a posterior for each θ_j, it is tempting to use the Laplace approximation directly, which involves approximating the distribution of $(\boldsymbol{\theta}_{-j}, \boldsymbol{x})|(\boldsymbol{y}, \theta_j)$ with a Gaussian. Such an approach will not be very successful, because the target will never be very close to Gaussian; it will typically involve triplets such as $\tau x_i x_j$. Instead we can construct an approximation to

$$\pi(\boldsymbol{\theta}|\boldsymbol{y}) \propto \frac{\pi(\boldsymbol{\theta})\pi(\boldsymbol{x}|\boldsymbol{\theta})\pi(\boldsymbol{y}|\boldsymbol{x}, \boldsymbol{\theta})}{\pi(\boldsymbol{x}|\boldsymbol{\theta}, \boldsymbol{y})}, \tag{13}$$

in which the Laplace approximation requires a Gaussian approximation of the denominator

$$\pi(\boldsymbol{x}|\boldsymbol{y},\boldsymbol{\theta}) \propto \exp\left(-\frac{1}{2}\boldsymbol{x}^T\,\boldsymbol{Q}(\boldsymbol{\theta})\boldsymbol{x} + \sum_i \log\pi(y_i|x_i,\boldsymbol{\theta})\right) \tag{14}$$

$$= (2\pi)^{-n/2}|\boldsymbol{P}(\boldsymbol{\theta})|^{1/2}\exp\left(-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu}(\boldsymbol{\theta}))^T\,\boldsymbol{P}(\boldsymbol{\theta})(\boldsymbol{x}-\boldsymbol{\mu}(\boldsymbol{\theta}))\right). \tag{15}$$

Here, $\boldsymbol{P}(\boldsymbol{\theta}) = \boldsymbol{Q}(\boldsymbol{\theta}) + \text{diag}(\boldsymbol{c}(\boldsymbol{\theta}))$, and $\boldsymbol{\mu}(\boldsymbol{\theta})$ is the location of the mode. The vector $\boldsymbol{c}(\boldsymbol{\theta})$ contains the negative second derivatives of the log-likelihood at the mode, with respect to x_i. There are two important aspects of Equation 15.

1. It is a GMRF with respect to the same graph as from a model without observations $\boldsymbol{y}$, so computationally it does not cost anything to account for the observations because their impact is a shift in the mean and the diagonal of the precision matrix.
2. The approximation is likely to be quite accurate because the impact of conditioning on the observations is only on the diagonal; it shifts the mean, reduces the variance, and might introduce some skewness into the marginals etc. Importantly, the observations do not change the Gaussian dependency structure through the terms $x_i x_j Q_{ij}(\boldsymbol{\theta})$, as these are untouched.

Because $|\boldsymbol{\theta}|$ is of low dimension, we can derive marginals for $\theta_j|\boldsymbol{y}$ directly from the approximation to $\boldsymbol{\theta}|\boldsymbol{y}$. Thinking traditionally, this might be costly, because every new $\boldsymbol{\theta}$ would require an evaluation of Equation 15, and the cost of numerical integration would still be exponential in $|\boldsymbol{\theta}|$. Luckily, the problem is somewhat more well behaved, because the latent field $\boldsymbol{x}$ introduces quite some uncertainty and more smooth behavior on the $\boldsymbol{\theta}$ marginals.

In situations where the central limit theorem becomes applicable, $\pi(\boldsymbol{\theta}|\boldsymbol{y})$ will be close to a Gaussian. We can improve this approximation using variance-stabilizing transformations of $\boldsymbol{\theta}$, such as using log(precisions) instead of precisions, or the Fisher transform of correlations. Additionally, we can use the Hessian at the mode to construct almost independent linear combinations (or transformations) of $\boldsymbol{\theta}$. These transformations simplify the problem, as they tend to diminish long tails and reduce skewness, which gives much simpler and better-behaved posterior densities.

The task of finding a quick and reliable approach to deriving all the marginal distributions from an approximation to the posterior density (Equation 13), while keeping the number of evaluation points low, was a serious challenge. We did not succeed on this until several years after Rue et al. (2009), and after several failed attempts. It was hard to beat the simplicity and stability of using the (Gaussian) marginals derived from a Gaussian approximation at the mode. However, we needed to do better, as these Gaussian marginals were not sufficiently accurate. The default approach used now is outlined in Martins et al. (2013, section 3.2), and involves correction of local skewness (in terms of difference in scale) and an integration-free method to approximate marginals from a skewness-corrected Gaussian. How this is technically achieved is somewhat involved, and we refer to Martins et al. (2013) for details. In our experience, we now balance accuracy and computational speed well, with an improvement over Gaussian marginals while still being exact in the Gaussian limit.

In some situations, our approximation to Equation 13 can be a bit off. This typically happens in cases with little smoothing and/or no replications, for example, when $\eta_i = \mu + \beta_z z_i + u_i$, for a random-effect $\boldsymbol{u}$, and a binary likelihood (Sauter & Held 2016). With vague priors, models such as this verge on being improper. Ferkingstad & Rue (2015) discuss these cases and derive a correction term which clearly improves the approximation to $\pi(\boldsymbol{\theta}|\boldsymbol{y})$.

3.2. Approximating the Posterior Marginals for the Latent Field

We will now discuss how to approximate the posterior marginals for the latent field. For linear predictors with no attached observations, the posterior marginals are also the basis to derive the predictive densities, as the linear predictor itself is a component of the latent field. Similar to Equation 11, we can express the posterior marginals as

$$\pi(x_i|\boldsymbol{y}) = \int \pi(x_i|\boldsymbol{\theta}, \boldsymbol{y})\, \pi(\boldsymbol{\theta}|\boldsymbol{y})\, \mathrm{d}\boldsymbol{\theta}, \tag{16}$$

hence we are faced with two more challenges.

1. We need to integrate over $\pi(\boldsymbol{\theta}|\boldsymbol{y})$, but the computational cost of standard numerical integration is exponential in the dimension of $\boldsymbol{\theta}$. We have already ruled out such an approach in Section 3.1, because it was too costly computationally, except when the dimension is low.
2. We need to approximate $\pi(x_i|\boldsymbol{\theta}, \boldsymbol{y})$ for a subset of all $i = 1, \ldots, n$, where n can be (very) large, such as in the range of 10^3 to 10^5. A standard application of the Laplace approximation, which involves location of the mode and factorization of a $(n-1) \times (n-1)$ matrix many times for each i, will simply be too demanding.

The key to success is to come up with efficient approximate solutions for each of these problems.

Classical numerical integration is only feasible in lower dimensions. If we want to use 5 integration points in each dimension, the cost would be 5^k to cover all combinations in k dimensions, which is 125 ($k = 3$) and 625 ($k = 4$). Using only 3 integration points in each dimension, we get 81 ($k = 4$) and 729 ($k = 6$). This is close to the practical limits. Beyond these limits we cannot aim to do accurate integration, but should rather aim for something that is better than avoiding the integration step, such as an empirical Bayes approach which just uses the mode. In dimensions >2, we borrow ideas from central composite design (Box & Wilson 1951) and use integration points on a sphere around the center; **Figure 3** illustrates the procedure in dimension 2 (even though we do not suggest using this approach in dimensions 1 and 2). The integrand is approximately spherical (after rotation and scaling), and the integration points will approximately be located on an appropriate level set for the joint posterior of $\boldsymbol{\theta}$. We can weight the spherical integration points equally, and determine the relative weight with the central point requiring the correct expectation of $\boldsymbol{\theta}^T\boldsymbol{\theta}$, if the posterior is standard Gaussian (Rue et al. 2009, section 6.5). It is our experience that this approach balances computational costs and accuracy well, and it is applied as the default integration scheme. More complex integration schemes could be used with increased computational costs.

For the second challenge, we need to balance the need for improved approximations beyond the Gaussian for $\pi(x_i|\boldsymbol{\theta}, \boldsymbol{y})$ with the fact that we (potentially) need to do this n times. Because n can be large, we cannot afford to do heavy computations for each i to improve on the Gaussian approximations. The default approach is to compute a Taylor expansion around the mode of the Laplace approximation, which provides a linear and a cubic correction term to the (standardized) Gaussian approximation,

$$\log \pi(x_i|\boldsymbol{\theta}, \boldsymbol{y}) \approx -\frac{1}{2}x_i^2 + b_i(\boldsymbol{\theta})x_i + \frac{1}{6}c_i(\boldsymbol{\theta})x_i^3. \tag{17}$$

We match a skew-Normal distribution (Azzalini & Capitanio 1999) to Equation 17, such that the linear term provides a correction term for the mean, while the cubic term provides a correction for skewness. This means that we approximate Equation 16 with a mixture of skew-Normal

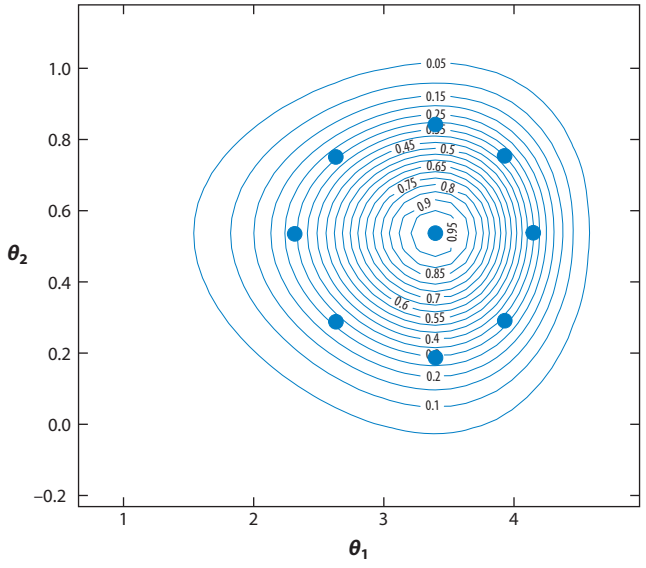

Figure 3

The contours of a posterior marginal for (θ_1, θ_2) and the associated integration points (*blue dots*).

distributions. This approach, termed simplified Laplace approximation, gives a very good trade-off between accuracy and computational speed.

Additional to posterior marginals, we can also provide estimates of the deviance information criterion (Spiegelhalter et al. 2002), Watanabe-Akaike information criterion (Wantanabe 2010, Gelman et al. 2014), and marginal likelihood and conditional predictive ordinates (Held et al. 2010). Other predictive criteria such as the ranked probability score or the Dawid-Sebastiani score (Gneiting & Raftery 2007) can also be derived in certain settings (Riebler et al. 2012, Schrödle et al. 2012). Martins & Rue (2014) discuss how the INLA-framework can be extended to a class of near-Gaussian latent models.

4. THE R-INLA PACKAGE: EXAMPLES

The `R-INLA` package (see **http://www.r-inla.org**) provides an implementation of the INLA approach, including standard and nonstandard tools to define models based on the `formula` concept in `R`. In this section, we present some examples of basic usage and some special features of `R-INLA`.

4.1. A Simple Example

We first show the usage of the package through a simple simulated example,

$$\boldsymbol{y}|\boldsymbol{\eta} \sim \text{Poisson}(\exp(\boldsymbol{\eta}))$$

where $\eta_i = \mu + \beta w_i + u_{j(i)}$, $i = 1, \ldots, n$, w are covariates, $\boldsymbol{u} \sim \mathcal{N}_m(\mathbf{0}, \tau^{-1}\boldsymbol{I})$, and $j(i)$ is a known mapping from $1:n$ to $1:m$. We generate data as follows:

```
set.seed(123456L)
n = 50; m = 10
w = rnorm(n, sd = 1/3)
u = rnorm(m, sd = 1/4)
intercept = 0; beta = 1
idx = sample(1:m, n, replace = TRUE)
y = rpois(n, lambda = exp(intercept + beta * w + u[idx]))
```

giving

```
> table(y, dnn=NULL)
 0 1 2 3 5
17 18 9 5 1
```

We use R-INLA to do the inference for this model, by

```
library(INLA)
my.data = data.frame(y, w, idx)
formula = y ~ 1 + w + f(idx, model="iid"),
r = inla(formula, data = my.data, family = "poisson")
```

The `formula` defines how the response depends on covariates, as usual, but the term `f(idx, model="iid")` is new. It corresponds to the function f that we introduced above in Equation 2, one of many implemented GMRF model components. The `iid` term refers to the $\mathcal{N}(\mathbf{0}, \tau^{-1}\boldsymbol{I})$ model, and `idx` is an index that specifies which elements of the model component go into the linear predictor.

Figure 4*a* shows three estimates of the posterior marginal of u_1. The solid line is the default estimate, the simplified Laplace approximation, as outlined in Section 3 (and with the R commands given above). The dashed line is the simpler Gaussian approximation which avoids integration over $\boldsymbol{\theta}$,

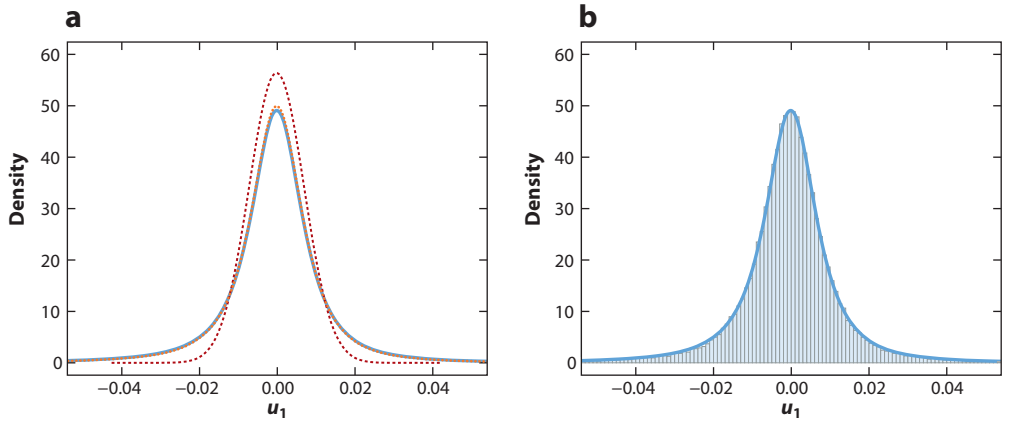

Figure 4

(*a*) The default estimate (simplified Laplace approximation) of the posterior marginal for u_1 (*blue solid*); a simplified estimate, i.e., the Gaussian approximation (*red dashed*); and the best possible Laplace approximation (*orange dotted*). (*b*) Histogram of u_1 using 10^5 samples produced using JAGS, together with the simplified Laplace approximation from panel *a*.

```
r.ga = inla(formula, data = my.data, family = "poisson",
 control.inla = list(strategy = "gaussian", int.strategy = "eb"))
```

The dotted line represents the (almost) true Laplace approximations and accurate integration over $\boldsymbol{\theta}$, and is the best approximation we can provide with the current software,

```
r.la = inla(formula, data = my.data, family = "poisson",
 control.inla = list(strategy = "laplace",
 int.strategy = "grid", dz=0.1, diff.logdens=20))
```

It is hard to see, as it is almost entirely covered by the solid line, meaning that our mixture of skew-Normals is very close to being exact in this example. We also note that by integrating out $\boldsymbol{\theta}$, the uncertainty increases, as it should. To compare the approximations with a simulation-based approach, **Figure 4*b*** shows the corresponding histogram for 10^5 samples using `JAGS`, together with the default estimate from **Figure 4*a***. The fit is quite accurate. The CPU time used by `R-INLA` with default options was approximately 0.16 seconds on a standard laptop, where 2/3 of this time was used for administration.

4.2. A Less Simple Example Including Measurement Error

We continue with a measurement error extension of the previous example, assuming that the covariate $\boldsymbol{w}$ is only observed indirectly through $\boldsymbol{z}$, where

$$z_i|\ldots \sim \texttt{Binomial}\left(m, \text{prob} = \frac{1}{1+\exp(-(\gamma + w_i))}\right), \quad i = 1, \ldots, n,$$

with intercept γ. In this case, the model needs to be specified using two likelihoods and also a special feature called `copy`. Each observation can have its own type of likelihood (i.e., family), which is coded using a matrix (or list) of observations, where each column represents one family. A linear predictor can only be associated with one observation. The `copy` feature allows us to have additional identical copies of the same model component in the formula, and we have the option to scale it as well. An index `NA` is used to indicate if there is no contribution to the linear predictor, and this is used to zero out contributions from model components. This is done in the code below:

```
## generate observations that we observe for 'w'
m = 2
z = rbinom(n, size = m, prob = 1/(1+exp(-(0 + w))))
## create the response. since we have two families, poisson and
## binomial, we use a matrix, one column for each family
Y = matrix(NA, 2*n, 2)
Y[1:n, 1] = y
Y[n + 1:n, 2] = z
## we need one intercept for each family. this is an easy way to achieve that
Intercept = as.factor(rep(1:2, each=n))
## say that we have 'beta*w' only for 'y' and 'w' only for 'z'. the formula
## defines the joint model for both the observations, 'y' and 'z'
NAs = rep(NA, n)
idx = c(NAs, 1:n)
idxx = c(1:n, NAs)
formula2 = Y ~ -1 + Intercept + f(idx, model="iid") +
```

```
                f(idxx, copy="idx", hyper = list(beta = list(fixed = FALSE)))
## need to use a 'list' since 'Y' is a matrix
my.data2 = list(Y=Y, Intercept = Intercept, idx = idx, idxx = idxx)
## we need to define two families and give the 'size' for the binomial
r2 = inla(formula2, data = my.data2, family = c("poisson", "binomial"),
                Ntrials = c(NAs, rep(m, n)))
```

We refer to Muff et al. (2015) for more details on measurement error models using INLA, and to the specific LGMs termed `mec` and `meb` that are available in `R-INLA` to facilitate the implementation of classical error models and Berkson error models, respectively.

4.3. A Spatial Example

The `R-INLA` package has extensive support for spatial Gaussian models, including intrinsic GMRF models on regions (often called CAR models; Hodges 2013, chapter 5.2) and a subclass of continuously indexed Gaussian field models. Of particular interest are Gaussian fields derived from stochastic partial differential equations (SPDEs). The simplest cases are Matérn fields in dimension d, which can be described as the solution to

$$(\kappa^2 - \Delta)^{\alpha/2}(\tau x(\boldsymbol{s})) = \mathcal{W}(\boldsymbol{s}), \tag{18}$$

where Δ is the Laplacian, $\kappa > 0$ is the spatial scale parameter, α controls the smoothness, τ controls the variance, and $\mathcal{W}(\boldsymbol{s})$ is a Gaussian spatial white noise process. Whittle (1954, 1963) shows that its solution is a Gaussian field with a Matérn covariance function having smoothness $\nu = \alpha - d/2$. The smoothness is usually kept fixed based on prior knowledge of the underlying process. A formulation of Matérn fields as solutions to Equation 18 might seem unnecessarily complicated, because we already know the solution. However, Lindgren et al. (2011) showed that by using a finite basis-function representation of the continuously indexed solution, one can derive (in analogy to the well-known finite element method) a local representation with Markov properties. This means that the joint distribution for the weights in the basis-function expansion is a GMRF, and the distribution follows directly from the basis functions and the triangulation of space. The main implication of this result is that it allows us to continue to think about and interpret the model using marginal properties such as covariances, but at the same time we can do fast computations because the Markov properties make the precision matrix very sparse. It also allows us to add this component in the `R-INLA` framework, like any other GMRF model component.

The dual interpretation of the Matérn field, both using covariances and also using its Markov properties, is very convenient from both a computational and also a statistical modeling point of view (Simpson et al. 2011, 2012; Lindgren & Rue 2015). The same ideas also apply to nonstationary Gaussian fields using nonhomogeneous versions of an appropriate SPDE (Lindgren et al. 2011; Yue et al. 2014; Fuglstad et al. 2015a,b), Gaussian fields that treat land as a barrier to spatial correlation (Bakka et al. 2016), multivariate random fields (Hu & Steinsland 2016), log-Gaussian Cox processes (Simpson et al. 2016a), and in the near future also to nonseparable space-time models.

We end this section with a simple example of spatial survival analysis taken from Henderson et al. (2002), studying spatial variation in leukemia survival data in northwest England in the period 1982–1998. The focus of the example is to see how, and how easily, the spatial model integrates into the model definition (Martino et al. 2010). We therefore omit further details about the dataset and refer to the original article.

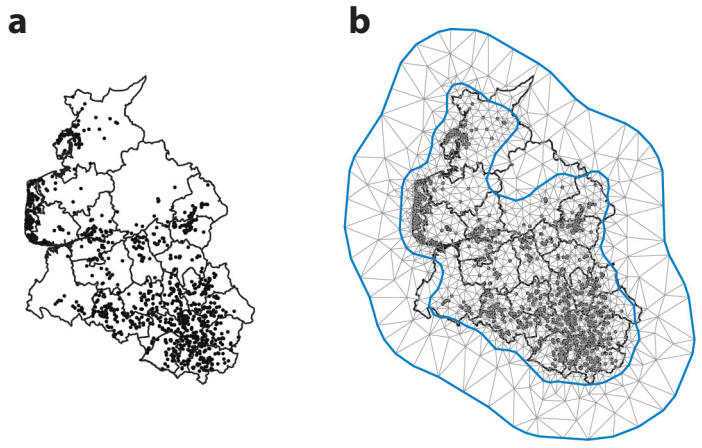

Figure 5

(*a*) The area of northwest England for the leukemia study, where the (post-code) locations of the events are shown as dots. (*b*) Overlay of the mesh used for the stochastic partial differential equation model.

First, we need to load the data and create the mesh, that is, a triangulation of the area of interest to represent the finite dimensional approximation to Equation 18.

```
library(INLA)
data(Leuk)
loc <- cbind(Leuk$xcoord, Leuk$ycoord)
bnd1 <- inla.nonconvex.hull(loc, convex=0.05)
bnd2 <- inla.nonconvex.hull(loc, convex=0.25)
mesh <- inla.mesh.2d(loc, boundary=list(bnd1, bnd2),
 max.edge=c(0.05, 0.2), cutoff=0.005)
```

Figure 5*a* displays the study area and the locations of the events, and **Figure 5*b*** shows the associated mesh with respect to which we define the SPDE model. We use an additional rougher mesh to reduce boundary effects. The next step is to create a mapping matrix from the mesh onto the locations where the data are observed. Then we define the SPDE model, to define the statistical model including covariates such as sex, age, white blood-cell counts (`wbc`), and the Townsend deprivation index (`tpi`), and call a bookkeeping function which keeps the indices in correct order. Finally, we call `inla()` to do the analysis, assuming a Weibull likelihood. Note that application of a Cox proportional hazard model will give similar results.

```
A <- inla.spde.make.A(mesh, loc)
spde <- inla.spde2.matern(mesh, alpha=2) ## alpha=2 is the default choice
formula <- inla.surv(time, cens) ~ 0 + a0 + sex + age + wbc + tpi +
                     f(spatial, model=spde)
stk <- inla.stack(data=list(time=Leuk$time, cens=Leuk$cens), A=list(A, 1),
                     effect=list(list(spatial=1:spde$n.spde),
                     data.frame(a0=1, Leuk[,-c(1:4)])))
r <- inla(formula, family="weibull", data=inla.stack.data(stk),
                     control.predictor=list(A=inla.stack.A(stk)))
```

Figure 6*a* shows the estimated spatial effect, with the posterior mean (*left panel*), and posterior standard deviation (*right panel*).

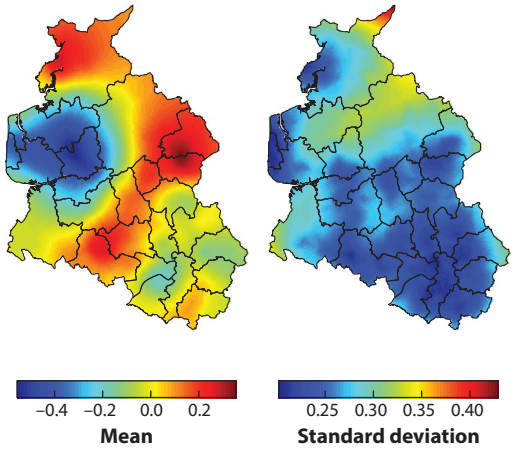

Figure 6

The spatial effect in the model: (*left panel*) mean, (*right panel)* standard deviation.

4.4. Special Features

In addition to standard analyses, the `R-INLA` package also contains nonstandard features that boost the complexity of models that can be specified and analyzed. Here, we give a short summary of these; for more details, the reader is directed to Martins et al. (2013).

- `replicate`: Each model component given as a `f()`-term can be replicated, creating `nrep` iid replications with shared hyperparameters. For example,

  ```
  f(time, model="ar1", replicate=person)
  ```

  defines one AR(1) model for each person sharing the same hyperparameters.
- `group`: Each model component given as a `f()`-term can be grouped, creating `ngroup` dependent replications with a separable correlation structure. To create a separable space-time model, with an AR(1) dependency in time, we can specify

  ```
  f(space, model=spde, group=time, control.group = list(model = "ar1"))
  ```

  Riebler et al. (2012) used grouped smoothing priors in `R-INLA` to impute missing mortality rates for a specific country by taking advantage from similar countries where these data are available. The authors provide the corresponding `R` code in the supplementary material. We can both group and replicate model components.
- A-matrix: We can create a second layer of linear predictors where $\boldsymbol{\eta}$ is defined by the `formula`, but $\boldsymbol{\eta}^* = \boldsymbol{A\eta}$ is connected to the observations. Here, $\boldsymbol{A}$ is a constant (sparse) matrix; see the above spatial example.
- Linear combinations: We can also compute posterior marginals of $\boldsymbol{v} = \boldsymbol{Bx}$, where $\boldsymbol{x}$ is the latent field and $\boldsymbol{B}$ is a fixed matrix. This could, for example, be $\beta_1 - \beta_2$ for two fixed effects, or any other linear combinations. Here is an example computing the posterior for the difference between two linear effects, $\beta_u - \beta_v$:

  ```
  lc = inla.make.lincomb(u=1, v=-1)
  r = inla(y ~ u + v, data = d, lincomb = lc)
  ```

- Remote server: It is easy to set up a remote Mac OSX/Linux server to host the computations while doing the `R` work at your local laptop. The job can be submitted and the results can be

retrieved later, or it can be used interactively. This is a very useful feature for larger models. It also ensures that computational servers will in fact be used, because we can work in a local R session but use a remote server for the computations. Here is an example running the computations on a remote server:

```
r = inla(formula, family, data = data, inla.call = "remote")
```

To submit a job we specify

```
r = inla(formula, family, data = data, inla.call = "submit")
```

and we can check the status and retrieve the results when the computations are done, by

```
inla.qstat(r)
r = inla.qget(r)
```

- R support: Although the core INLA program is written in C, it is possible to pass a user-defined latent model component written in R and use that as any other latent model component. The R code will be evaluated within the C program. This is very useful for more specialized model components or reparameterizations of existing ones, even though it will run more slowly than a proper implementation in C. As a simple example, the code below implements the model component iid, which is just independent Gaussian random effects $\mathcal{N}_n(\mathbf{0}, (\tau \boldsymbol{I})^{-1})$. The skeleton of the function is predefined and must return the graph, $\boldsymbol{Q}$-matrix, initial values, mean, log normalizing constant, and log prior for the hyperparameters.

```
iid.model = function(cmd = c("graph", "Q", "mu", "initial",
 "log.norm.const", "log.prior", "quit"),
 theta = NULL, args = NULL)
{
 interpret.theta = function(n, theta)
    return (list(prec = exp(theta[1L])))
 graph = function(n, theta)
    return (Diagonal(n, x= rep(1, n)))
 Q = function(n, theta) {
    prec = interpret.theta(n, theta)$prec
    return (Diagonal(n, x= rep(prec, n))) }
 mu = function(n, theta) return (numeric(0))
 log.norm.const = function(n, theta) {
    prec = interpret.theta(n, theta)$prec
    return (sum(dnorm(rep(0, n),
       sd = 1/sqrt(prec), log=TRUE))) }
 log.prior = function(n, theta) {
    prec = interpret.theta(n, theta)$prec
    return (dgamma(prec, shape = 1, rate = 5e-05, log=TRUE)
       + theta[1L]) }
 initial = function(n, theta) return (4.0)
 quit = function(n, theta) return (invisible())

 val = do.call(match.arg(cmd),
    args = list(n = as.integer(args$n), theta = theta))
 return (val)
}
```

```
n = 50 ## the dimension
my.iid = inla.rgeneric.define(iid.model, n=n)
```

Hence, we can replace `f(idx,model="iid")` with our own R implementation, using `f(idx, model=my.iid)`. For details on the format, see `inla.doc("rgeneric")` and `demo(rgeneric)`.

5. A CHALLENGE FOR THE FUTURE: PRIORS

Although the R-INLA project has been highly successful, it has also revealed some weak points in general Bayesian methodology from a practical point of view. In particular, our main concern is how we think about and specify priors in LGMs. We will now discuss this issue and our current plan to provide good, sensible default priors.

Bayesian statistical models require prior distributions for all the random elements of the model. Working within the class of LGMs, choosing prior distributions involves choosing priors for all the hyperparameters $\boldsymbol{\theta}$ in the model, because the latent field is by definition Gaussian. We deliberately write "priors" because it is common practice to define independent priors for each θ_j, whereas we really should aim for a joint prior for all $\boldsymbol{\theta}$, when appropriate.

The ability to incorporate prior knowledge in Bayesian statistics is a great tool and potentially very useful. However, except for cases where we do have real prior knowledge, for example, through results from previous experiments, it is often conceptually difficult to encode prior knowledge through probability distributions for all model parameters. Examples include priors for precision and overdispersion parameters, or the amount of t-ness in Student's *t*-distribution. Simpson et al. (2016b) discuss these aspects in great detail.

In R-INLA, we have chosen to provide default prior distributions for all parameters. We admit that currently these have been chosen partly based on the priors that are commonly used in the literature and partly out of the blue. It might be argued that this is not a good strategy, and that we should force the user to provide the complete model including the joint prior. This is a valid point, but all priors in R-INLA can easily be changed, allowing the user to define any arbitrary prior distribution. So the whole argument can be reduced to a question of convenience.

Do we have a "Houston, we have a problem" situation with priors? Looking at the current practice within the Bayesian society, we came to the conclusion that we do. In the following, we argue for this through a simple example, showing what can go wrong, how we can think about the problem, and how we can fix it. We only discuss proper priors.

Consider the problem of replacing a linear effect of the Townsend deprivation index `tpi` with a smooth effect of `tpi` in the leukemia example in Section 4.3. This is easily implemented by replacing `tpi` with `f(tpi, model="rw2")`. Here, `rw2` is a stochastic spline, simply saying that the second derivative is independent Gaussian noise (Rue & Held 2005, Lindgren & Rue 2008). By default, we constrain the smooth effect to also sum to zero, so that these two model formulations are the same in the limit as the precision parameter τ tends to infinity, and a vague Gaussian prior is used for the linear effect. The question is which prior should be used for τ. An overwhelming majority of cases in the literature use some kind of a Gamma(a, b) prior for τ, implying that $\pi(\tau) \propto \tau^{a-1} \exp(-b\tau)$ for some $a, b > 0$. This prior is flexible, conjugate with the Gaussian, and seems like a convenient choice. Since almost everyone else is using it, how wrong can it be?

If we rewind to the point where we replaced the linear effect with a smooth effect, we realize that we do this because we want a more flexible model than the linear effect, that is, we also want to capture deviations from the linear effect. Implicitly, if there is a linear effect, we do want to retrieve that with enough data. Measuring the distance between the straight line and the stochastic

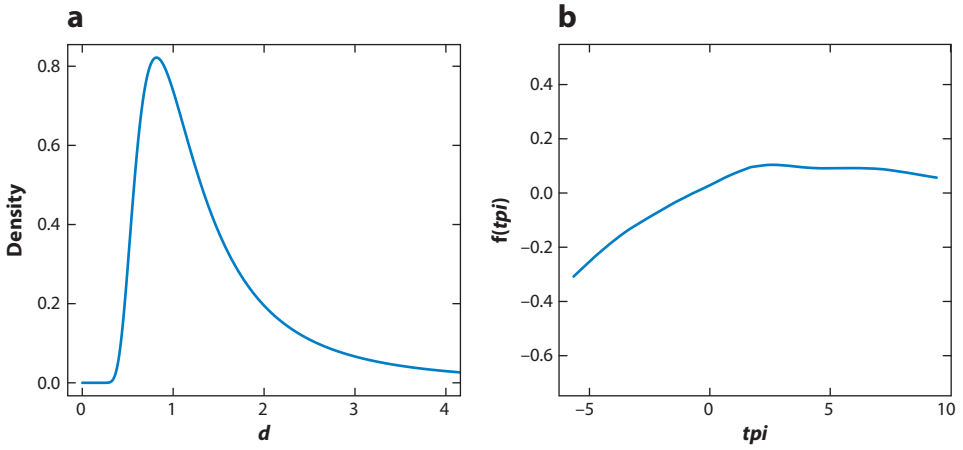

Figure 7

(*a*) The Gamma(1, 1) prior on the distance scale. (*b*) The smoothed effect of covariate `tpi` using the exponential prior on the distance scale $\lambda \exp(-\lambda)$.

spline using the Kullback-Leibler divergence (KLD), we find that KLD $\propto 1/\tau$ meaning that the (unidirectional) distance is $d \propto \sqrt{1/\tau}$. For simplicity, choose $a = b = 1$ in the Gamma-prior, then the derived prior for the distance d is

$$\pi(d) \propto \exp(-1/d^2)/d^3. \tag{19}$$

Figure 7*a* displays this prior on the distance scale, revealing two surprising features. First, the mode is approximately $d \approx 0.82$, and second, the prior appears to be zero for a range of positive distances. The second feature is serious, as it simply prevents the spline from getting too similar to the linear effect. It is clear from Equation 19 that the effect is severe, and in practice, $\pi(d) \approx 0$ even for positive d. This is an example of what Simpson et al. (2016b) call prior overfitting; the prior prevents the simpler model to be located, even when it is the true model. Choosing different parameters in the Gamma prior does not change the overfitting issue. For all $a, b > 0$, the corresponding prior for the distance tends to 0 as $d \to 0$. For a (well-behaved) prior to have $\pi(d = 0) > 0$, we need $E(\tau) = \infty$.

If we are concerned about the behavior of the distance between the more flexible and the simpler model component, we should define the prior directly on the distance, as proposed by Simpson et al. (2016b). A prior for the distance should be decaying with the mode at distance zero. This makes the simpler model central and the point of attraction. The exponential prior is recommended as a generic choice because it has a constant rate penalization, $\pi(d) = \lambda \exp(-\lambda d)$. The value of λ could be chosen by calibrating some property of the model component under consideration. Note that this way of defining the prior is invariant to reparameterizations, as it is defined on the distance and not for a particular parameterization.

Let us return to the stochastic spline example, assigning the exponential prior to the distance. The parameter λ can be calibrated by imposing the knowledge that the effect of `tpi` is not likely to be above 1 on the linear predictor scale,

```
..+ f(tpi, model="rw2", scale.model = TRUE,
   hyper = list(prec = list(prior="pc.prec", param=c(1, 0.01))))
```

Here, `scale.model` is required to ensure that the parameter τ represents the precision, not just a precision parameter (Sørbye & Rue 2014). The estimated results are given in **Figure 7*b***,

illustrating the point-wise posterior mean, median, and the 2.5% and 97.5% credibility intervals, for the effect of `tpi` on the mean survival time.

Here, we have only briefly addressed the important topic of constructing well-working priors, and currently we are focusing much activity on this issue to take the development further. In addition to other plans, we plan to integrate automatic tests for prior sensitivity, following the work of Roos & Held (2011) and M. Roos et al. (2015). The final goal is to use the above ideas to construct a joint default prior for LGMs, which can be easily understood and interpreted. A main issue is how to decompose and control the variance of the linear predictor, an issue we have not discussed here. For further information about this issue, please see Simpson et al. (2016b) for the original report that introduces the class of penalized complexity (PC) priors. Some examples of the application of these priors include disease mapping (Riebler et al. 2016), bivariate meta-analysis (Guo et al. 2015, Guo & Riebler 2015), age-period-cohort models (Riebler & Held 2016), Bayesian P-splines (Ventrucci & Rue 2016), structured additive distributional regression (Klein & Kneib 2016), Gaussian fields in spatial statistics (Fuglstad et al. 2016), modeling monthly maxima of instantaneous flow (Ferkingstad et al. 2016) and autoregressive processes (Sørbye & Rue 2016).

Interestingly, the framework and ideas behind PC priors are also useful for sensitivity analysis of model assumptions and developing robust models, but not enough work has yet been done in this area to go into detail here. Stay tuned!

6. DISCUSSION

We hope we have convinced the reader that the INLA approach to approximate Bayesian inference for LGMs is a useful addition to the applied statistician's toolbox; the key components just play so nicely together, providing a very exact approximation while reducing computational costs substantially. The key benefit of the INLA approach is that it is central to our long-term goal of making LGMs a class of models that we (as a community) can use and understand.

Developing, writing, and maintaining the code base for such a large open-source project is a huge job. Nearly all the `R/C/C++` code is written and maintained by F.K. Lindgren (20%) and H. Rue (80%), and is a result of a substantial amount of work over many years. Many more have contributed indirectly by challenging the current practice and implementation. The current version of this project is a result of the cumulative effort of the many users and their willingness to share, challenge, and question essentially everything. Documentation is something we could and should improve upon, but the recent book by Blangiardo & Cameletti (2015) does a very good job.

The current status of the package is good, but we have to account for the fact that the software has been developed over many years, and is basically the version we used while developing the methods. Hence, although the software works well, it is less streamlined and less easy to maintain than it ought to be. We are now at a stage where we know what we want the package to do and software to be, hence a proper rewrite by skilled people would be a useful project for the society. If this were to happen, we would be more than happy to share all our knowledge in such a "version 2.0" project.

Another use of `R-INLA` is to use it purely as computational back end. The generality of `R-INLA` results in increased complexity for the user, hence a simplified interface for a restricted set of models can be useful to improve accessibility for a specific target audience or provide additional tools that are mainly relevant for these models. Examples of such projects are `AnimalINLA` (Holand et al. 2013), `ShrinkBayes` (Van De Wiel et al. 2013a,b, 2014; Riebler et al. 2014), `meta4diag` (Guo & Riebler 2015), `BAPC` (Riebler & Held 2016), `diseasemapping` and `geostatp` (Brown 2015), and Bivand et al. (2015). Similarly, the `excursions` package for calculating joint exceedance

probabilities in GMRFs (Bolin & Lindgren 2015, 2016) includes an interface to analyze LGMs estimated by R-INLA. Recent work on methodology for filtered spatial point patterns in the context of distance sampling (Yuan et al. 2016) has initiated the construction of wrapper software for fitting other complex spatial models such as those resulting from plot sampling data or for point process models within R-INLA. There is also an interesting line of research using R-INLA to do approximate inference on a submodel within a larger model; the reader is directed to Guihenneuc-Jouyaux & Rousseau (2005) for a theoretical justification and Li et al. (2012) for an early application of this idea. One particular problem is how to handle missing data in cases where the joint model is not an LGM.

Please visit us at **http://www.r-inla.org**!

DISCLOSURE STATEMENT

The authors are not aware of any affiliations, memberships, funding, or financial holdings that might be perceived as affecting the objectivity of this review.

ACKNOWLEDGMENTS

We would like to acknowledge all the users of the R-INLA package, who have challenged and questioned essentially everything, and their willingness to share this with us.

LITERATURE CITED

Azzalini A, Capitanio A. 1999. Statistical applications of the multivariate skew-normal distribution. *J. R. Stat. Soc. B* 61:579–602

Baghishani H, Mohammadzadeh M. 2012. Asymptotic normality of posterior distributions for generalized linear mixed models. *J. Multivariate Anal.* 111:66–77

Bakka H, Vanhatalo J, Illian J, Simpson D, Rue H. 2016. Accounting for physical barriers in species distribution modeling with non-stationary spatial random effects. arXiv:1608.03787 [stat.AP]

Barndorff-Nielsen OE, Cox DR. 1989. *Asymptotic Techniques for Use in Statistics*. Boca Raton, FL: Chapman and Hall/CRC

Bauer C, Wakefield J, Rue H, Self S, Feng Z, Wang Y. 2016. Bayesian penalized spline models for the analysis of spatio-temporal count data. *Stat. Med.* 35:1848–65

Bhatt S, Weiss DJ, Cameron E, Bisanzio D, Mappin B, et al. 2015. The effect of malaria control on *Plasmodium falciparum* in Africa between 2000 and 2015. *Nature* 526:207–11

Bivand RS, Gómez-Rubio V, Rue H. 2015. Spatial data analysis with R-INLA with some extensions. *J. Stat. Softw.* 63:1–31

Blangiardo M, Cameletti M. 2015. *Spatial and Spatio-Temporal Bayesian Models with R-INLA*. New York: John Wiley & Sons

Bolin D, Lindgren F. 2015. Excursion and contour uncertainty regions for latent Gaussian models. *J. R. Stat. Soc. B* 77:85–106

Bolin D, Lindgren F. 2016. Quantifying the uncertainty of contour maps. *J. Comput. Graph. Stat.* arXiv:1507.01778

Bowler DE, Haase P, Kröncke I, Tackenberg O, Bauer HG, et al. 2015. A cross-taxon analysis of the impact of climate change on abundance trends in central Europe. *Biol. Conserv.* 187:41–50

Box GEP, Tiao GC. 1973. *Bayesian Inference in Statistical Analysis*. Reading, MA: Addison-Wesley

Box GEP, Wilson KB. 1951. On the experimental attainment of optimum conditions (with discussion). *J. R. Stat. Soc. B* 13:1–45

Brown PE. 2015. Model-based geostatistics the easy way. *J. Stat. Softw.* 63:1–24

Crewe TL, Mccracken JD. 2015. Long-term trends in the number of monarch butterflies (Lepidoptera: Nymphalidae) counted on fall migration at Long Point, Ontario, Canada (1995–2014). *Ann. Entomol. Soc. Am.* 105:707–17

Dwyer-Lindgren L, Flaxman AD, Ng M, Hansen GM, Murray CJ, Mokdad AH. 2015. Drinking patterns in US counties from 2002 to 2012. *Am. J. Public Health* 105:1120–27

Ferkingstad E, Geirsson OP, Hrafnkelsson B, Davidsson OB, Gardarsson SM. 2016. A Bayesian hierarchical model for monthly maxima of instantaneous flow. arXiv:1606.07667 [stat.AP]

Ferkingstad E, Rue H. 2015. Improving the INLA approach for approximate Bayesian inference for latent Gaussian models. *Electron. J. Stat.* 9:2706–31

Friedrich A, Marshall JC, Biggs PJ, Midwinter AC, French NP. 2016. Seasonality of *Campylobacter jejuni* isolates associated with human campylobacteriosis in the Manawatu region, New Zealand. *Epidemiol. Infect.* 144:820–28

Fuglstad GA, Lindgren F, Simpson D, Rue H. 2015a. Exploring a new class of non-stationary spatial Gaussian random fields with varying local anisotropy. *Stat. Sin.* 25:115–33

Fuglstad GA, Simpson D, Lindgren F, Rue H. 2015b. Does non-stationary spatial data always require non-stationary random fields? *Spat. Stat.* 14(C):505–31

Fuglstad GA, Simpson D, Lindgren F, Rue H. 2016. Constructing priors that penalize the complexity of Gaussian random fields. arXiv:1503.00256 [stat.ME]

García-Pérez J, Lope V, López-Abente G, González-Sánchez M, Fernández-Navarro P. 2015. Ovarian cancer mortality and industrial pollution. *Environ. Pollut.* 205:103–10

Gelman A, Hwang J, Vehtari A. 2014. Understanding predictive information criteria for Bayesian models. *Stat. Comput.* 24:997–1016

Gneiting T, Raftery AE. 2007. Strictly proper scoring rules, prediction, and estimation. *J. Am. Stat. Assoc.* 102:359–78

Goicoa T, Ugarte MD, Etxeberria J, Militino AF. 2016. Age-space-time CAR models in Bayesian disease mapping. *Stat. Med.* 35:2391–405

Goth US, Hammer HL, Claussen B. 2014. Utilization of Norway's emergency wards: the second 5 years after the introduction of the patient list system. *Int. J. Environ. Res. Public Health* 11:3375–86

Guihenneuc-Jouyaux C, Rousseau J. 2005. Laplace expansion in Markov chain Monte Carlo algorithms. *J. Comput. Graphical Stat.* 14:75–94

Guo J, Riebler A. 2015. Meta4diag: Bayesian bivariate meta-analysis of diagnostic test studies for routine practice. arXiv:1512.06220 [stat.AP]

Guo J, Rue H, Riebler A. 2015. Bayesian bivariate meta-analysis of diagnostic test studies with interpretable priors. arXiv:1512.06217 [stat.ME]

Halonen JI, Blangiardo M, Toledano MB, Fecht D, Gulliver J, et al. 2016. Long-term exposure to traffic pollution and hospital admissions in London. *Environ. Pollut.* 208(A):48–57

Halonen JI, Hansell AL, Gulliver J, Morley D, Blangiardo M, et al. 2015. Road traffic noise is associated with increased cardiovascular morbidity and mortality and all-cause mortality in London. *Eur. Heart J.* 36:2653–61

Held L, Rue H. 2010. Conditional and intrinsic autoregressions. In *Handbook of Spatial Statistics*, ed. A Gelfand, P Diggle, M Fuentes, P Guttorp, pp. 201–16. Boca Raton, FL: CRC/Chapman & Hall

Held L, Sauter R. 2016. Adaptive prior weighting in generalized regression. *Biometrics* doi:10.1111/biom.12541

Held L, Schrödle B, Rue H. 2010. Posterior and cross-validatory predictive checks: a comparison of MCMC and INLA. In *Statistical Modelling and Regression Structures—Festschrift in Honour of Ludwig Fahrmeir*, ed. T Kneib, G Tutz, pp. 91–110. Berlin: Springer Verlag

Henderson R, Shimakura S, Gorst D. 2002. Modeling spatial variation in leukemia survival data. *J. Am. Stat. Assoc.* 97:965–72

Hodges JS. 2013. *Richly Parameterized Linear Models: Additive, Time Series, and Spatial Models Using Random Effects.* Boca Raton, FL: Chapman and Hall/CRC

Holand AM, Steinsland I, Martino S, Jensen H. 2013. Animal models and integrated nested Laplace approximations. *G3* 3:1241–51

Hu X, Steinsland I. 2016. Spatial modeling with system of stochastic partial differential equations. *Wiley Interdiscip. Rev. Comput. Stat.* 8:112–25

Iulian TV, Juan P, Mateu J. 2015. Bayesian spatio-temporal prediction of cancer dynamics. *Comput. Math. Appl.* 70:857–68

Jousimo J, Tack AJM, Ovaskainen O, Mononen T, Susi H, et al. 2014. Ecological and evolutionary effects of fragmentation on infectious disease dynamics. *Science* 344:1289–93

Kandt J, Chang S, Yip P, Burdett R. 2016. The spatial pattern of premature mortality in Hong Kong: How does it relate to public housing? *Urban Stud.* doi: 10.1177/0042098015620341

Karagiannis-Voules DA, Biedermann P, Ekpo UF, Garba A, Langer E, et al. 2015. Spatial and temporal distribution of soil-transmitted helminth infection in sub-Saharan Africa: a systematic review and geostatistical meta-analysis. *Lancet Infect. Dis.* 15:74–84

Karcher MD, Palacios JA, Bedford T, Suchard MA, Minin VN. 2016. Quantifying and mitigating the effect of preferential sampling on phylodynamic inference. *PLOS Comput. Biol.* 12:1–19

Kauermann G, Krivobokova T, Fahrmeir L. 2009. Some asymptotic results on generalized penalized spline smoothing. *J. R. Stat. Soc. B* 71:487–503

Klein N, Kneib T. 2016. Scale-dependent priors for variance parameters in structured additive distributional regression. *Bayesian Anal.* 11:1071–1106

Kröger H, Hoffmann R, Pakpahan E. 2016. Consequences of measurement error for inference in cross-lagged panel design—the example of the reciprocal causal relationship between subjective health and socio-economic status. *J. R. Stat. Soc. A* 179:607–28

Li Y, Brown P, Rue H, al-Maini M, Fortin P. 2012. Spatial modelling of lupus incidence over 40 years with changes in census areas. *J. R. Stat. Soc. C* 61:99–115

Lindgren F, Rue H. 2008. A note on the second order random walk model for irregular locations. *Scand. J. Stat.* 35:691–700

Lindgren F, Rue H. 2015. Bayesian spatial modelling with R-INLA. *J. Stat. Softw.* 63:1–25

Lindgren F, Rue H, Lindström J. 2011. An explicit link between Gaussian fields and Gaussian Markov random fields: the SPDE approach (with discussion). *J. R. Stat. Soc. B* 73:423–98

Lithio A, Nettleton D. 2015. Hierarchical modeling and differential expression analysis for RNA-seq experiments with inbred and hybrid genotypes. *J. Agric. Biol. Environ. Stat.* 20:598–613

Martino S, Akerkar R, Rue H. 2010. Approximate Bayesian inference for survival models. *Scand. J. Stat.* 28:514–28

Martins TG, Rue H. 2014. Extending INLA to a class of near-Gaussian latent models. *Scand. J. Stat.* 41:893–912

Martins TG, Simpson D, Lindgren F, Rue H. 2013. Bayesian computing with INLA: new features. *Comput. Stat. Data Anal.* 67:68–83

Muff S, Keller LF. 2015. Reverse attenuation in interaction terms due to covariate measurement error. *Biometrical J.* 57:1068–83

Muff S, Riebler A, Rue H, Saner P, Held L. 2015. Bayesian analysis of measurement error models using integrated nested Laplace approximations. *J. R. Stat. Soc. C* 64:231–52

Niemi J, Mittman E, Landau W, Nettleton D. 2015. Empirical Bayes analysis of RNA-seq data for detection of gene expression heterosis. *J. Agric. Biol. Environ. Stat.* 20:614–28

Noor AM, Kinyoki DK, Mundia CW, Kabaria CW, Mutua JW, et al. 2014. The changing risk of *Plasmodium falciparum* malaria infection in Africa: 2000-10: a spatial and temporal analysis of transmission intensity. *Lancet* 383:1739–47

Ogden H. 2016. On asymptotic validity of approximate likelihood inference. arXiv:1601.07911 [math.ST]

Opitz N, Marcon C, Paschold A, Malik WA, Lithio A, et al. 2016. Extensive tissue-specific transcriptomic plasticity in maize primary roots upon water deficit. *J. Exp. Bot.* 67:1095–107

Papoila AL, Riebler A, Amaral-Turkman A, São-João R, Ribeiro C, et al. 2014. Stomach cancer incidence in Southern Portugal 1998–2006: a spatio-temporal analysis. *Biometrical J.* 56:403–15

Plummer M. 2016. Rjags: Bayesian graphical models using MCMC. *R Software Package for Graphical Models.* **https://cran.r-project.org/web/packages/rjags/index.html**

Quiroz Z, Prates MO, Rue H. 2015. A Bayesian approach to estimate the biomass of anchovies in the coast of Perú. *Biometrics* 71:208–17

Riebler A, Held L. 2016. Projecting the future burden of cancer: Bayesian age-period-cohort analysis with integrated nested Laplace approximations. *Biometrical J.* In press

Riebler A, Held L, Rue H. 2012. Estimation and extrapolation of time trends in registry data—borrowing strength from related populations. *Ann. Appl. Stat.* 6:304–33

Riebler A, Robinson M, van de Wiel M. 2014. Analysis of next generation sequencing data using integrated nested Laplace approximation (INLA). *Statistical Analysis of Next Generation Sequencing Data*, ed. S Datta, D Nettleton, pp. 75–91. New York: Springer

Riebler A, Sørbye SH, Simpson D, Rue H. 2016. An intuitive Bayesian spatial model for disease mapping that accounts for scaling. *Stat. Methods Med. Res.* 25:1145–65

Robert CP, Casella G. 1999. *Monte Carlo Statistical Methods*. New York: Springer-Verlag

Rooney J, Vajda A, Heverin M, Elamin M, Crampsie A, et al. 2015. Spatial cluster analysis of population amyotrophic lateral sclerosis risk in Ireland. *Neurology* 84:1537–44

Roos M, Held L. 2011. Sensitivity analysis in Bayesian generalized linear mixed models for binary data. *Bayesian Anal.* 6:259–78

Roos M, Martins TG, Held L, Rue H. 2015. Sensitivity analysis for Bayesian hierarchical models. *Bayesian Anal.* 10:321–49

Roos NC, Carvalho AR, Lopes PF, Pennino MG. 2015. Modeling sensitive parrotfish (Labridae: Scarini) habitats along the Brazilian coast. *Mar. Environ. Res.* 110:92–100

Rue H, Held L. 2005. *Gaussian Markov Random Fields: Theory and Applications*. Boca Raton, FL: CRC/Chapman and Hall

Rue H, Held L. 2010. Markov random fields. In *Handbook of Spatial Statistics*, ed. A Gelfand, P Diggle, M Fuentes, P Guttorp, pp. 171–200. Boca Raton, FL: CRC/Chapman and Hall

Rue H, Martino S, Chopin N. 2009. Approximate Bayesian inference for latent Gaussian models using integrated nested Laplace approximations (with discussion). *J. R. Stat. Soc. B* 71:319–92

Salmon M, Schumacher D, Stark K, Höhle M. 2015. Bayesian outbreak detection in the presence of reporting delays. *Biometrical J.* 57:1051–67

Santermans E, Robesyn E, Ganyani T, Sudre B, Faes C, et al. 2016. Spatiotemporal evolution of Ebola virus disease at sub-national level during the 2014 West Africa epidemic: model scrutiny and data meagreness. *PLOS ONE* 11:e0147172

Sauter R, Held L. 2015. Network meta-analysis with integrated nested Laplace approximations. *Biometrical J.* 57:1038–50

Sauter R, Held L. 2016. Quasi-complete separation in random effects of binary response mixed models. *J. Stat. Comput. Simul.* 86:2781–96

Schrödle B, Held L. 2011a. A primer on disease mapping and ecological regression using INLA. *Comput. Stat.* 26:241–58

Schrödle B, Held L. 2011b. Spatio-temporal disease mapping using INLA. *Environmetrics* 22:725–34

Schrödle B, Held L, Rue H. 2012. Assessing the impact of network data on the spatio-temporal spread of infectious diseases. *Biometrics* 68:736–44

Selwood KE, Thomson JR, Clarke RH, McGeoch MA, Mac Nally R. 2015. Resistance and resilience of terrestrial birds in drying climates: Do floodplains provide drought refugia? *Glob. Ecol. Biogeogr.* 24:838–48

Shun Z, McCullagh P. 1995. Laplace approximation of high dimensional integrals. *J. R. Stat. Soc. B* 57:749–60

Simpson D, Illian J, Lindgren F, Sørbye S, Rue H. 2016a. Going off grid: computational efficient inference for log-Gaussian Cox processes. *Biometrika* 103:1–22

Simpson DP, Lindgren F, Rue H. 2011. Think continuous: Markovian Gaussian models in spatial statistics. *Spat. Stat.* 1:16–29

Simpson D, Lindgren F, Rue H. 2012. In order to make spatial statistics computationally feasible, we need to forget about the covariance function. *Environmetrics* 23:65–74

Simpson DP, Rue H, Riebler A, Martins TG, Sørbye SH. 2016b. Penalising model component complexity: a principled, practical approach to constructing priors (with discussion). *Stat. Sci.* In press

Sørbye SH, Rue H. 2014. Scaling intrinsic Gaussian Markov random field priors in spatial modelling. *Spat. Stat.* 8:39–51

Sørbye SH, Rue H. 2016. Penalised complexity priors for stationary autoregressive processes. arXiv:1608.08941 [stat.ME]

Spiegelhalter DJ, Best NG, Carlin BP, van der Linde A. 2002. Bayesian measures of model complexity and fit (with discussion). *J. R. Stat. Soc. B* 64:583–639

Spiegelhalter DJ, Thomas A, Best NG, Gilks WR. 1995. BUGS: Bayesian inference using Gibbs sampling. *Software Package for Performing Bayesian Inference Using Markov Chain Monte Carlo.* **http://www.mrc-bsu.cam.ac.uk/software/bugs/**

Stan Development Team. 2015. Stan modeling language user's guide and reference manual. **http://www.uvm.edu/~bbeckage/Teaching/DataAnalysis/Manuals/stan-reference-2.8.0.pdf**

Tierney L, Kadane JB. 1986. Accurate approximations for posterior moments and marginal densities. *J. Am. Stat. Assoc.* 81:82–86

Tsiko RG. 2016. A spatial latent Gaussian model for intimate partner violence against men in Africa. *J. Fam. Violence* 41:443–59

Ugarte MD, Adin A, Goicoa T. 2016. Two-level spatially structured models in spatio-temporal disease mapping. *Stat. Methods Med. Res.* 25:1080–100

Ugarte MD, Adin A, Goicoa T, Militino AF. 2014. On fitting spatio-temporal disease mapping models using approximate Bayesian inference. *Stat. Methods Med. Res.* 23:507–30

Van De Wiel MA, De Menezes RX, Siebring E, Van Beusechem VW. 2013a. Analysis of small-sample clinical genomics studies using multi-parameter shrinkage: application to high-throughput RNA interference screening. *BMC Med. Genom.* 6:1–9

Van De Wiel MA, Leday GGR, Pardo L, Rue H, van der Vaart AW, van Wieringen WN. 2013b. Bayesian analysis of high-dimensional RNA sequencing data: estimating priors for shrinkage and multiplicity correction. *Biostatistics* 14:113–28

Van De Wiel MA, Neerincx M, Buffart TE, Sie D, Verheul HMW. 2014. ShrinkBayes: a versatile R-package for analysis of count-based sequencing data in complex study design. *BMC Bioinform.* 15:116

Ventrucci M, Rue H. 2016. Penalized complexity priors for degrees of freedom in Bayesian P-splines. *Stat. Model.* doi:10.1177/1471082X16659154. In press

Wantanabe S. 2010. Asymptotic equivalence of Bayes cross validation and widely applicable information criterion in singular learning theory. *J. Mach. Learn. Res.* 11:3571–94

Whittle P. 1954. On stationary processes in the plane. *Biometrika* 41:434–49

Whittle P. 1963. Stochastic processes in several dimensions. *Bull. Inst. Internat. Statist.* 40:974–94

Yuan Y, Bachl FE, Borchers DL, Lindgren F, Illian JB, et al. 2016. Point process models for spatio-temporal distance sampling data. arXiv:1604.06013 [stat.ME]

Yue YR, Simpson D, Lindgren F, Rue H. 2014. Bayesian adaptive smoothing spline using stochastic differential equations. *Bayesian Anal.* 9:397–424

Annual Review of Statistics and Its Application

Volume 4, 2017

Contents

Errata

An online log of corrections to *Annual Review of Statistics and Its Application* articles may be found at http://www.annualreviews.org/errata/statistics