A technical report on Integrated Nested Laplace Approximation

Aaron Zimmerman

Department of Statistics, University of Washington Seattle

June 18, 2013

The Paper

This report is based on *Approximate Bayesian Inference for latent Gaussian Models by using integrated nested Laplace Approximations* written by Håvard Rue, Sara Martino, and Nicolas Chopin in 2009 and published in the Journal of the Royal Statistical Society (Rue et al., 2009).

1 Introduction

The science of statistical inference is often broken into two distinct philosophical approaches: that of *frequentist* inference and that of *Bayesian* inference. While the argument for or against either approach is certainly not within the scope of this paper (nor does the author wish to claim proclivity to either approach at this time) there has been a relatively recent explosion of interest in Bayesian approaches to inference. Within the field of statistics, as measured by the articles published in 23 journals, the percent of articles referencing "Bayes" or "Bayesian" nearly tripled from 0.11 to 0.32 in the 30 year stretch between 1970 and 2000 (Poirier, 2006). Poirier further notes that the increase is seen most drastically after 1995, most likely due to the increasing influence and use of the Monte Carlo Markov Chain (MCMC).

One of the historical drawbacks to Bayesian inference has been the intractability of necessary equations, in particular the evaluation of the integrals needed to obtain posterior densities. This happens in many interesting applications such as, for example, analysis of spatial point processes (Plagnol and Tavaré, 2004) or image analysis (Jun et al., 2008). The problem can generally be formulated by assuming that we have some dataset y (from a sample space $\mathcal{Y}$), which has been generated from a statistical model with density $f(y|\theta)$. In a Bayesian setting we assume that the parameter of the model θ has a prior density $\pi(\theta)$ and our goal is to solve for the posterior density $\pi(\theta|y) = \frac{\pi(\theta)f(y|\theta)}{f(y)} = \frac{\pi(\theta)f(y|\theta)}{\int \pi(\theta)f(y|\theta)d\theta}$.

The issue is often that the normalizing constant, $\int \pi(\theta)f(y|\theta)d\theta$, can't be determined in a closed form. Sometimes, as in the case of conjugate priors, we can find the form of the normalizing constant and proceed. Unfortunately, that's the exception to the rule and without the normalizing constant, we're not able to perform any number of evaluations on the posterior density that we've gone through this work to find. The problem of determining the posterior density has motivated a number of solutions of both stochastic and deterministic nature. Integrated nested Laplace approximations (INLA) provide one deterministic approach to finding (approximating) posterior distributions. Throughout this paper will work through and explore the

formulation of the INLA methodology and provide examples of the algorithm in practice and as compared to MCMC methods.

The rest of the paper is organized as follows. Section 2 presents a formal specification of the problem, section 3 discusses some previous solutions to the problem, section 4 develops the Laplace expansion, section 5 develops the INLA methodology, section 6 provides example problems and their solutions in INLA, and section 7 includes a discussion of the method.

2 The Problem

The paper proposes to develop methods that are efficient and accurate within the context of Bayesian inference among latent Gaussian models, a subset of structured additive regression models.

The class of structured additive regression models are widely used in statistics and other fields, and we start the problem formulation by assuming that the observation variables of interest, y_i, belong to an exponential family. The model assumes that the mean of the y_i observations, the μ_is, are linked to some η_is through a link function such that $g(\mu_i) = \eta_i$, where η_i is a structured additive predictor which accounts for the effects of covariates additively through the model:

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

In this formulation, we define the $\{f^{(j)}(\cdot)\}$s to be unknown functions of the covariates $\mathbf{u}$ (which are possibly unmeasured), the β_ks are the linear effects of the remainder of the covariates $\mathbf{z}$, and the ϵ_is are in the model as unstructured terms. By imposing the restriction to (1) that the priors for α, $\{f^{(j)}\}$s, $\{\beta_k\}$s, and $\{\epsilon_i\}$s are all Gaussian, we reduce the the model to a latent Gaussian model.

For the remainder of the discussion we assume the following notation and model specification:

- The $\mathbf{y}s$ are our observed data with n observations from an exponential family model. We assume that the data come from a distribution which depends on parameters $\boldsymbol{\theta}$ and possible hyperparameters $\boldsymbol{\gamma}_2$.
- The $\boldsymbol{\theta}$s are our parameters of interest. Our primary goal is to determine their marginal posterior distributions, and if desired, posterior distributions for the hyperparameters. In particular, in terms of (1), we define $\boldsymbol{\theta} = \{\alpha, \boldsymbol{f^{(j)}}, \boldsymbol{\beta_k}, \boldsymbol{\eta_i}\}$ for $j = 1, \ldots n_f$, $k = 1, \ldots, n_\beta$, and $i = 1, \ldots, n$. We further assume some prior normal distributions on the $\boldsymbol{\theta}$s such that $\pi(\boldsymbol{\theta}|\boldsymbol{\gamma}_2) \sim \mathcal{N}(\mathbf{0}, Q(\boldsymbol{\gamma}_2))$ for some hyperparameters $\boldsymbol{\gamma}_2$ which determine the precision matrix Q.
- Then $\boldsymbol{\gamma} \equiv \{\boldsymbol{\gamma}_1, \boldsymbol{\gamma}_2\}$ is our collection of hyperparameters for which we can assume another layer of priors which need not be Gaussian in nature.

Latent Gaussian models, as described, form a class of very flexible and adaptable models which are involved in a host of applications. Regression models can be formulated as Bayesian generalized linear

models which fit within this framework (Dey et al., 2000), possibly with random effects or using penalized spline models (Lang and Brezger, 2004). By allowing the $f(\cdot)$s to be indexed by a time index t, these models can then be used to fit either discrete or continuous time series models with temporal dependence or even the underlying latent process of a time series model (Kitagawa and Gersch, 1984). By further extending the $f(\cdot)$s to be indexed by a spatial or spatial and temporal index, latent Gaussian models can be adapted into the spatial-temporal model setting where we use the $f(\cdot)$s to promote smoothing across space or across space and time. The Besag-York-Mollie model (Besag et al., 1991) is an example of a spatial model that fits this framework, and there are many examples stretching the family to include spatial-temporal dependencies such as in Cressie and Johannesson (2008).

3 Proposed and Existing Solutions

There are a number of other methods that work to find the posterior marginals in similar or identical settings. The most well known in the statistics literature are Monte Carlo Markov Chain (MCMC) methods. MCMC is sometimes known to perform poorly in latent Gaussian models due to the strong dependence in the latent field $\boldsymbol{\theta}$ and the possibly strong dependence between $\boldsymbol{\gamma}$ and $\boldsymbol{\theta}$. While there is work on MCMC methods to overcome these difficulties by constructing joint proposal densities from approximations to the full conditional of $\boldsymbol{\theta}$ (e.g. Gamerman (1997); Rue and Martino (2007)) or by blocking the parameters (e.g. see Chapter 4 of Rue and Held (2005)), it can be difficult and the computational time can still be large enough to deter use.

At its simplest, Monte Carlo methods can be used to determine a mean or probability of independent identically distributed (*iid*) random variables by implementing the Law of Large Numbers through computer simulation. That is, given *iid* random variables $X_1, X_2, \ldots$, we can approximate $\mathrm{E}[X_i]$ by simulating realizations $X_1 = x_1, X_2 = x_2, \ldots$. Then, by leveraging the law of large numbers, we approximate $\mathrm{E}[X_i] \approx \frac{1}{n}\sum_{i=1}^{n} x_i$ and for large enough n, we will closely approximate the truth. MCMC methods first construct a (dependent) chain of simulated observations $X_1, X_2, \ldots$ on a state-space ($\mathcal{X}$) that are set up to converge to a target stationary distribution of interest, π. Then given a real-valued function $f : \mathcal{X} \to \mathbb{R}$ with $\sum_{x\in\mathcal{X}} f(x)\pi(x) = \mathrm{E}(X)$, we still have that $\mathrm{E}[X] \approx \frac{1}{n}\sum_{i=1}^{n} x_i$, even though our observations are no longer independent. MCMC methods can be used to simulate observations from posterior distributions and these simulations can then determine posterior estimates for the parameters of interest (e.g. the mean or median of the sample) as well as posterior intervals (using quantiles of the sample). One notable feature of MCMC simulation is the promise of convergence to the truth if the simulation is allowed to run "long enough." Unfortunately, long enough could be very long, and determining if you've allowed the chain to run until convergence can also be challenging.

In addition to MCMC methods, two other methods for approximate inference, Variational Bayes (VB) (Hinton and Van Camp, 1993) and Expectation-Propagation (EP) (Minka, 2001) have also been applied to

this set of problems. VB uses an approximation to the joint density of $\pi(\boldsymbol{\theta}, \boldsymbol{\gamma})$ that minimizes the Kullback-Leibler contrast of $\pi(\boldsymbol{\theta}, \boldsymbol{\gamma}|\mathbf{y})$ relative to $q(\boldsymbol{\theta}, \boldsymbol{\gamma})$ subject to the (often not realistic) constraint that $\boldsymbol{\theta}$ and $\boldsymbol{\gamma}$ are independent. EP works similarly, but contrasts $q(\boldsymbol{\theta}, \boldsymbol{\gamma})$ to $\pi(\boldsymbol{\theta}, \boldsymbol{\gamma}|\mathbf{y})$ using KL. In EP, the posterior variance can be highly underestimated, and due to the reversal of contrasts, EP can overestimate the posterior variance.

In order to respond to the shortcomings of the existing methods, Rue et al. (2009) describe a quick and (we hope to show) accurate method to determine some or all of the posteriors for the components of the latent Gaussian vector, $\boldsymbol{\theta}$, and in addition, some of the posteriors for the hyperparameters, $\boldsymbol{\gamma}$, used in the Gaussian priors of the model. The authors demonstrate the utility of their method by finding accurate approximations for the posterior marginals in a number of applications. Their results indicate that their estimated posteriors are accurate and the computational time (when compared to MCMC methods) can drop significantly. For example, while coding up the main effects model from Knorr-Held (1999), my peer, Laina Mercer, found improvements of over 1000-fold from 2.4 hours for $250,000$ MCMC iterations down to 7.1 seconds using R-INLA.

We describe the authors' steps for a solution after making two assumptions which are often validated in latent Gaussian Model problems. We assume the latent field $\boldsymbol{\theta}$ admits conditional independence properties and also that the number of hyperparameters is relatively small (the authors suggest $|\boldsymbol{\gamma}| \leq 6$).

1. Approximate the posterior marginal for the hyperparameters, $\pi(\boldsymbol{\gamma}|\mathbf{y})$, with the equivalent of a Laplace approximation: $\tilde{\pi}(\boldsymbol{\gamma}|\mathbf{y}) \propto \frac{\pi(\boldsymbol{\theta},\boldsymbol{\gamma},\mathbf{y})}{\pi_G(\boldsymbol{\theta}|\boldsymbol{\gamma},\mathbf{y})}|_{\boldsymbol{\theta}=\boldsymbol{\theta}^*(\boldsymbol{\gamma})}$, where $\pi_G(\boldsymbol{\theta}|\boldsymbol{\gamma},\mathbf{y})$ is a Gaussian approximation to the full conditional of $\boldsymbol{\theta}$ and $\boldsymbol{\theta}^*(\boldsymbol{\gamma})$ is the mode of the full conditional.

2. Determine the Laplace approximation $\tilde{\pi}_{\mathcal{L}}(\theta_i|\boldsymbol{\gamma},\mathbf{y})$ for $\pi(\theta_i|\boldsymbol{\gamma},\mathbf{y})$ at selected values of $\boldsymbol{\gamma}$.

3. Numerically integrate out the hyperparameters using approximations from steps 1 and 2 in order to find an approximation for the posterior marginals:

$$\tilde{\pi}(\theta_i|\mathbf{y}) = \sum_k \tilde{\pi}(\theta_i|\boldsymbol{\gamma}_k,\mathbf{y})\tilde{\pi}(\boldsymbol{\gamma}_k|\mathbf{y})\Delta_k.$$

These three steps make up the core of the INLA algorithm and will be discussed in further detail in sections 5.3, 5.4, and 5.5 respectively. We will refer back to these three steps throughout this report.

4 The Laplace Expansion and Approximation

One of the central features in the proposed INLA method is the repeated use of Laplace expansions to quickly obtain accurate approximations of posterior and marginal posterior densities. As an important recurring technique, we take the time here to elaborate in detail the Laplace approximation as used by Rue, Martino, and Chopin.

The Laplace expansion is a general technique which may be found in numerous reference sources (see e.g. De Bruijn (1970); Erdélyi (1956)). For a random variable y, the problem at hand is to find an approximation for the expectation of a function $q(\cdot)$ of the random variable, i.e. $E[q(y)]$. In particular, we are looking for a simple approximation of the form $E[q(y)] \approx q(\hat{y})$, where $\hat{y}$ is a value of y chosen to ensure a reasonable approximation. We know that there exists a $y^\star$ such that $E[q(y^\star)] = q(y^\star)$ by the mean value theorem. Unfortunately, solving for the exact $y^\star$ is no easier than finding $E[q(y)]$ directly, which we're trying to avoid. That leaves us with the opportunity to find accurate, and hopefully simple, approximations to $y^\star$.

Generally, we assume that the random variable y has a density of the following form:

$$p(y) \propto e^{-m \times h(y)} b(y),$$

where the constant of proportionality, which ensures that the density integrates to 1, may depend on m and is unknown (thus the proportionality sign). In an exponential family setting, we can think of $h(y)$ as the log-likelihood. We'll use the representation of h as a log-likelihood in section 5 when we further describe the INLA methodology. To make it explicit, we would like to find an approximation for

$$E[q(y)] = \frac{\int q(y) b(y) e^{-mh(y)} dy}{\int b(y) e^{-mh(y)} dy}. \tag{2}$$

We proceed with the Laplace expansion of (2) assuming that h has an absolute minimum at some $\hat{y}$, with $\frac{dh}{dy}\big|_{y=\hat{y}} = 0$, and $\frac{d^2h}{dy^2}|_{y=\hat{y}} > 0$. To develop the Laplace expansion, we first expand the functions q, b, and h about $\hat{y}$. Letting $x = \sqrt{m}(y - \hat{y})$ we then obtain the following three expansions:

$$\begin{aligned}
q(y) &= q(\hat{y}) + \frac{xq'(\hat{y})}{\sqrt{m}} + \frac{x^2 q''(\hat{y})}{2!m} + \dots \\
b(y) &= b(\hat{y}) + \frac{xb'(\hat{y})}{\sqrt{m}} + \frac{x^2 b''(\hat{y})}{2!m} + \dots \\
m \times h(y) &= mh(\hat{y}) + 0 + \frac{h''(\hat{y})x^2}{2!} + \frac{h'''(\hat{y})x^3}{3!\sqrt{m}} + \frac{h''''(\hat{y})x^4}{4!m} + \dots \quad .
\end{aligned}$$

Substituting these expansions into (2), canceling the first terms in the expansion of $m \times h(y)$ as well as the change of variable coefficients, pulling out out the third term in $m \times h(y)$, and noting that the fourth term in the series dominates the expression lets us rewrite $E[q(y)]$ as:

$$E[q(y)] = \frac{\int e^{-\frac{h''(\hat{y})x^2}{2}} [(q(\hat{y}) + \frac{xq'(\hat{y})}{\sqrt{m}} + \dots)(b(\hat{y}) + \frac{xb'(\hat{y})}{\sqrt{m}} + \dots) exp\{\frac{-h'''(\hat{y})x^3}{6\sqrt{m}} + \dots\}] dx}{\int e^{-\frac{h''(\hat{y})x^2}{2}} [(b(\hat{y}) + \frac{xb'(\hat{y})}{\sqrt{m}} + \dots) exp\{\frac{-h'''(\hat{y})x^3}{6\sqrt{m}} + \dots\}] dx}. \tag{3}$$

The initial terms in both the numerator and the denominator can be seen to be kernels of a normal distribution with 0 mean and variance equal to $[h''(\hat{y})]^{-1}$. In an exponential family model, this is analogous to using the inverse of the Hessian as an approximation to the covariance in our Gaussian approximation. Furthermore, we could use a power series in $\frac{1}{\sqrt{m}}$ to continue the expansion of the terms within the square braces in both the numerator and the denominator, leaving the Gaussian kernel unchanged. The expansion results in the ratio of two power series from which the odd terms (the $\frac{1}{\sqrt{m}}$ terms) vanish. The remaining

ratio of power series in $\frac{1}{m}$ can be further simplified, and after cancellation the first order approximation is found to be

$$E[q(y)] \approx q(\hat{y}) + \frac{1}{m}\left(\frac{q'(\hat{y})}{h''(\hat{y})}\left[\frac{b'(\hat{y})}{b(\hat{y})} - \frac{h'''(\hat{y})}{2h''(\hat{y})}\right] + \frac{q'(\hat{y})}{h''(\hat{y})}\right). \tag{4}$$

The use of the Laplace expansion in the scope of the INLA paper is limited to the basic approximation $E[q(y)] \approx q(\hat{y})$. Inclusion of the first order correction term would increase the quality of the approximation, though it's not always clear that higher order approximations are worth the added complication.

In much a similar way, we can use the Laplace expansion technique to approximate posterior densities in a Bayesian setting. By expanding the log of the posterior density (or at least a function proportional to the posterior density) up to quadratic terms we can approximate the log posterior with a Gaussian approximation. By applying Laplace's method to both the numerator and denominator of a Bayesian marginal posterior distribution, $\pi(\theta_1|y) = \frac{\int \pi(\theta_1,\boldsymbol{\theta}_{-1})f(y|\theta)d\boldsymbol{\theta}_{-1}}{f(y)} = \frac{\int \pi(\theta_1,\boldsymbol{\theta}_{-1})f(y|\theta)d\boldsymbol{\theta}_{-1}}{\int \pi(\theta)f(y|\theta)d\theta}$, we can find Tierney and Kadane's Laplace approximation (Tierney and Kadane, 1986) which is equivalent to the Laplace approximation used throughout INLA.

5 INLA Methodology

5.1 Gaussian Markov Random Field (GMRF) Calculations

While the general techniques and approximations described in this report could be applied to other model settings, a large part of INLAs recent success is due to the speed at which this approximation can be calculated. The speed of many of the computations leverages the GMRF framework. The R-INLA package works in conjunction with Rue's GMRFLib of C-routines which was built to perform fast and exact simulation of GMRFs on graphs. For these reasons, we start off the methodology section with a brief description of GMRFs.

Markov random fields are a natural extension of Markov Chains (see section 3). Specifically, a Gaussian Markov random field is a multivariate Gaussian random variable, $\mathbf{y} = (y_1, \ldots, y_n)$, with conditional independences encoded in the precision matrix. If we define the mean of $\mathbf{y}$ to be μ and the precision (inverse of the covariance) to be $\mathbf{Q}$, then we can write the density of our GMRF as

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

By construction, we have that two random variables in the field, y_i and y_j, are conditionally independent given $\mathbf{y}_{-ij}$ (all random variables in the $\mathbf{y}$ vector except for y_i and y_j) if and only if $Q_{ij} = 0$ for the precision matrix $\mathbf{Q}$. Marginal means are simply $\mathrm{E}[y_i] = \mu_i$, and if the precision matrix is easily and quickly inverted, then marginal variances maybe found by taking diagonal elements from the covariance matrix such that $\mathrm{Var}[y_i] = [\mathbf{Q}^{-1}]_{ii} = \Sigma_{ii}$. In some applications, the dimension of $\mathbf{Q}$ may be large enough to desire avoiding the matrix inversion to find marginal variances. In these cases, marginal variances can be computed more efficiently by using the Cholesky decomposition of $\mathbf{Q}$ and a recursion which we describe here.

Using the Cholesky decomposition, we factor $\mathbf{Q}$ into a lower Cholesky triangular matrix $\mathbf{L}$ such that $\mathbf{Q} = \mathbf{L}\mathbf{L}^T = \mathbf{V}\mathbf{D}\mathbf{V}^T$ and $\mathbf{L} = \mathbf{V}\mathbf{D}^{1/2}$. By noting that $\mathbf{Q}\boldsymbol{\Sigma} = \mathbf{V}\mathbf{D}\mathbf{V}^T\boldsymbol{\Sigma} = \mathbf{I}$, multiplying on the left by $(\mathbf{V}\mathbf{D})^{-1} = (\mathbf{D}^{-1}\mathbf{V}^{-1})$, adding $\boldsymbol{\Sigma}$ to both sides, and rearrangement, we find that

$$\mathbf{D}^{-1}\mathbf{V}^{-1} + (\mathbf{I} - \mathbf{V}^T)\boldsymbol{\Sigma} = \boldsymbol{\Sigma}. \tag{5}$$

The recursion described in section 2.1 of Rue et al. (2009) describes the upper triangular matrix of (5) with the following form:

$$\Sigma_{ij} = \left(\frac{\delta_{ij}}{L_{ii}}\right)^2 - \frac{1}{L_{ii}} \sum_{k=i+1}^{n} L_{ki}\Sigma_{kj},$$

where $\delta_{ij} = 0$ if $i \neq j$ and $\delta_{ij} = 1$ if $i = j$. By looping on i from n to 1 while an inner loop for j runs from n to i, we can efficiently compute the marginal variances. This can reduce the cost of inverting $\mathbf{Q}$ in high dimension latent fields. See Rue and Martino (2007) for more information.

We need one other property of GMRFs before we can continue. In many situations, the model specification of the GMRF may include linear constraints such that $\mathbf{M}\mathbf{y} = \mathbf{v}$ for a constant vector $\mathbf{v}$ and rank-k, $k \times n$, matrix $\mathbf{M}$. By starting with an unconstrained sample $\mathbf{y}$, we can construct a constrained sample through conditioning on kriging (see section 3.6.2 of Cressie (1992)), such that

$$\mathbf{y}_{\text{constrained}} = \mathbf{y} - \mathbf{Q}^{-1}\mathbf{M}^T(\mathbf{M}\mathbf{Q}^{-1}\mathbf{M}^T)^{-1}(\mathbf{M}\mathbf{y} - \mathbf{v}). \tag{6}$$

This construction will be used in the following section to correctly perform Gaussian approximations.

5.2 Model Reparameterization

In order to fit our Laplace approximations, we need to perform Gaussian approximations to posterior densities. The approximation we desire in section 5.3 assumes the following form for the full conditional density:

$$\pi(\boldsymbol{\theta}|\mathbf{y}, \boldsymbol{\gamma}) \propto \exp\left(-\frac{1}{2}\boldsymbol{\theta}^T\mathbf{Q}\boldsymbol{\theta} + \sum_{i=1}^{n} g_i(\theta_i)\right). \tag{7}$$

We take the summation over the indices of our data, and for the full conditional densities that we need to approximate, we have that $g_i(\theta_i)$ are the log-likelihoods for our data, $\log[\pi(y_i|\theta_i, \boldsymbol{\gamma})]$. If the dimension of the parameters is the same as the dimension of the data, there is no need to reparameterize. Otherwise, we need to transform $\boldsymbol{\theta}$ to meet this parameterization. A linear transformation is all that is needed, but the transformation often results in a non-sparse precision matrix $\mathbf{Q}$, even if the original parameterization had a sparse precision. As an example, consider a random effects model from a seeds study found in Crowder (1978). We assume that the number of seeds which germinate on 21 different experimental plates are binomial random variables with probabilities logit-linked to to an additive model such that $\text{logit}(\mathbf{p}) = \mathbf{X}\boldsymbol{\alpha}+\mathbf{b}$ for design matrix $\mathbf{X}$, regression coefficients $\boldsymbol{\alpha}$ and random effects $\mathbf{b}$. With this parameterization, the log-likelihood for each y_i depends on all $\boldsymbol{\alpha}$s and b_i. To reparameterize, we use a transformation matrix $\mathbf{A}$ designed to take $\boldsymbol{\theta}^0 = \{\boldsymbol{\alpha}, \mathbf{b}\}$ to $\boldsymbol{\theta}^1 = \{\mathbf{X}\boldsymbol{\alpha} + \mathbf{b}, \boldsymbol{\alpha}\}$. Now the log-likelihood for each y_i depends on exactly one θ_i^1 as we

desired. Furthermore, the transformed precision can be found by transforming the original precision matrix such that $\mathbf{Q}^1 = [\mathbf{A}^T]^{-1}\mathbf{Q}^0[\mathbf{A}]^{-1}$. From here on out, we assume that we have the necessary transformed parameterization of $\boldsymbol{\theta}$ and the transformed precision $\mathbf{Q}$.

With the reparameterization, we proceed with the INLA method using approximations to the densities described in (7). Using the Fisher-Scoring algorithm (Newton-Raphson on log-likelihoods with expected information) we can quickly compute Gaussian approximations to the densities in (7) which will take the form

$$\tilde{\pi}_G(\boldsymbol{\theta}|\mathbf{y},\boldsymbol{\gamma}) \propto \exp\left(-\frac{1}{2}(\boldsymbol{\theta}-\boldsymbol{\mu}^*)^T\mathbf{Q}^*(\boldsymbol{\theta}-\boldsymbol{\mu}^*)\right), \tag{8}$$

where $\mathbf{Q}^*$ and $\boldsymbol{\mu}^*$ are the converged solutions from the Fisher-Scoring algorithm which we now describe.

We Taylor expand the $g_i(\theta_i)$ functions from (7) up to the second order around some initial guess for the mean, $\boldsymbol{\mu}_0$ such that

$$g_i(\theta_i) \approx g_i(\mu_{0_i}) + b_i\theta_i - \frac{1}{2}c_i\theta_i^2,$$

where $b_i = g_i'(\mu_{0_i}) - g_i''(\mu_{0_i})\mu_{0_i}$ and $c_i = -g_i''(\mu_{0_i})$ from the Taylor expansion. The expansion leads to an initial Gaussian approximation with a new precision matrix $\mathbf{Q}_1 = \mathbf{Q}+\text{diag}(\mathbf{c})$, and mode $\boldsymbol{\mu}_1 = \mathbf{Q}_1^{-1}\mathbf{b}$. This process is repeated until convergence is reached (e.g. by looking at mean square errors between different iterations of the mode) which yields an approximation of the form in (8). If there are linear constraints in the GMRF, the mean must be corrected at every iteration using (6). Note that under the reparameterization, there may be fewer functions g_i than $|\boldsymbol{\theta}|$. In this case, we fill in the vectors $\mathbf{c}$ and $\mathbf{b}$ with zeros for all θ_j that have zero function $g_j(\theta_j)$.

5.3 Step 1: Build and explore the approximation $\tilde{\pi}_G(\boldsymbol{\gamma}|\mathbf{y})$

As mentioned, we use a Laplace approximation for the posterior distribution of the hyperparameters which takes the following form

$$\tilde{\pi}(\boldsymbol{\gamma}|\mathbf{y}) \propto \left.\frac{\pi(\boldsymbol{\gamma},\boldsymbol{\theta},\mathbf{y})}{\tilde{\pi}_G(\boldsymbol{\theta}|\boldsymbol{\gamma},\mathbf{y})}\right|_{\boldsymbol{\theta}=\boldsymbol{\theta}^*(\boldsymbol{\gamma})}. \tag{9}$$

In order to explore our approximation to the posterior distribution of the hyperparameters, we have to reevaluate the approximation for the full conditional in the denominator for each set of hyperparameter values. In order to make this process more efficient, we first locate the mode of $\tilde{\pi}_G(\boldsymbol{\gamma}|\mathbf{y})$ through numerical methods (e.g. the authors recommend using a quasi-Newton method which first builds up numerical gradients using finite differences and then builds up the Hessian through differences in gradient vectors). We denote our estimated modal value as $\boldsymbol{\gamma}^*$. At the modal point, we (again) numerically determine the Hessian, and as we saw in section 4, we can use the inverse of the Hessian to approximate the covariance structure. Let the inverse of the Hessian be defined as $\mathbf{C}^*$, which contains information about the curvature of $\tilde{\pi}_G(\boldsymbol{\gamma}|\mathbf{y})$ at $\boldsymbol{\gamma}^*$. We decompose $\mathbf{C}^*$ using an eigendecomposition such that $\mathbf{C}^* = \mathbf{V}\mathbf{D}\mathbf{V}^T$ in order to explore $\tilde{\pi}_G(\boldsymbol{\gamma}|\mathbf{y})$ along

standardized variable axes $\mathbf{z}$ where

$$\boldsymbol{\gamma} = \boldsymbol{\gamma}^* + \mathbf{V}\mathbf{D}^{1/2}\mathbf{z}.$$

We note, as the authors note, that if $\tilde{\pi}_G(\boldsymbol{\gamma}|\mathbf{y})$ is Gaussian, then $\mathbf{C}^*$ will be the covariance, and $\mathbf{z}$ will be a standard normal random variable.

To explore $\tilde{\pi}_G(\boldsymbol{\gamma}|\mathbf{y})$, we start at $\boldsymbol{\gamma}^* \equiv \mathbf{z} = \mathbf{0}$ and we step out along each of the standardized axes (in the positive direction and then in the negative direction) until a stopping criterion, (10), has been reached.

$$\log\left[\tilde{\pi}_G\left(\boldsymbol{\gamma}(\mathbf{0})|\mathbf{y}\right)\right] - \log[\tilde{\pi}_G\left(\boldsymbol{\gamma}(\mathbf{z})|\mathbf{y}\right)] < \delta \tag{10}$$

The authors recommend taking steps along the standardized axes of size $\Delta_z = 1$ and setting the stopping criterion tolerance to be $\delta = 2.5$. If it is desired to explore the posterior hyperparameter density in more detail (e.g. to approximate the hyperparameter density instead of using it only for integration), the step sizes Δ_z can be reduced, and our stopping criterion tolerance δ can be increased.

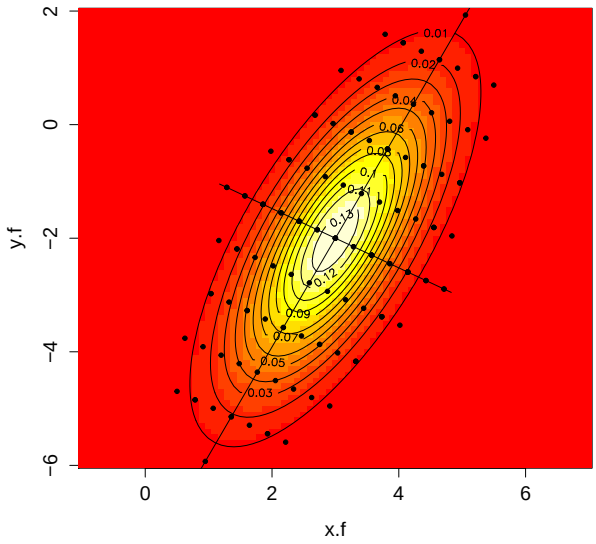


Figure 1: Exploration in 2-D. First the standardized axes are explored, and then the rectangle of grid points is filled in. Note that the rectangle isn't completely filled because the stopping criterion was hit and exploration in that direction was stopped. The plot axes are the original axes, the axes drawn in the plot are the standardized axes, and the points are the explored grid locations.

The final step in each direction before stopping defines a hyperrectangle that contains what we hope is the bulk of the mass of the hyperparameter density. We now explore the interior of the hyperrectangle in a similar fashion. At each step we check to see if the stopping criterion has been met, and if it has, we discontinue exploration in that direction. The hyperrectangle defines the furthest that we may have to explore, but we may not need explore it in totality as shown in Figure 1. Note that at each grid point we must evaluate $\tilde{\pi}_G(\boldsymbol{\gamma}|\mathbf{y})$, and we store the values to later use in numerical integration. We call the stored grid points $\boldsymbol{\gamma}_k$. Once we've finished exploring the grid, we standardize our evaluated values such that $\sum_k \tilde{\pi}_G(\boldsymbol{\gamma}_k|\mathbf{y}) = 1$.

Marginals for the hyperparemters can be computed either by numerical integration from the standardized grid, or by fitting a smooth interpolant to the stored values of $\tilde{\pi}_G(\boldsymbol{\gamma}|\mathbf{y})$ at $\boldsymbol{\gamma}_k$ and then numerically integrating with respect to the interpolant.

5.4 STEP 2: Calculate the approximations $\tilde{\pi}_{\mathcal{L}}(\theta_i|\mathbf{y},\boldsymbol{\gamma})$

We discuss three methods of varying accuracy for this approximation.

Gaussian Approximation

In the precursor to this main INLA paper (Rue and Martino, 2007), the authors use a rough Gaussian approximation to $\pi(\theta_i|\mathbf{y},\boldsymbol{\gamma})$. The Gaussian approximation can suffer in some situations because it forces a symmetric density and cannot model skewness that may be present. That said, this approximation is very quick due to the GMRF structure of the approximation to the full conditional as in (8). Approximate marginal expectations are just values from the mean of $\tilde{\pi}_G(\boldsymbol{\theta}|\mathbf{y},\boldsymbol{\gamma})$, $\mathrm{E}[\theta_i|\mathbf{y},\boldsymbol{\gamma}] = \mu_i^*$. Furthermore, marginal variances are items on the diagonal of the inverse of the precision matrix of $\tilde{\pi}_G(\boldsymbol{\theta}|\mathbf{y},\boldsymbol{\gamma})$ such that $\mathrm{Var}[\theta_i|\mathbf{y},\boldsymbol{\gamma}] = \Sigma_{ii}^* = [\mathbf{Q}^*]_{ii}^{-1}$. Instead of inverting the precision matrix to find marginal variances, the recursion defined in Section 5.1 can be used to speed up the computation.

Laplace Approximation

To correct for the inflexibility of the Gaussian approximation, we could instead use Laplace approximations. These approximations will often yield reasonable results and conveniently allows us to correct for skew (as compared to the Gaussian approximation). Unfortunately, the approximation takes the form

$$\tilde{\pi}_{LA}(\theta_i|\boldsymbol{\gamma},\mathbf{y}) \propto \left.\frac{\pi(\boldsymbol{\gamma},\boldsymbol{\theta},\mathbf{y})}{\tilde{\pi}_G(\boldsymbol{\theta}_{-i}|\theta_i,\boldsymbol{\gamma},\mathbf{y})}\right|_{\boldsymbol{\theta}_{-i}=\boldsymbol{\theta}^*_{-i}(\gamma)}. \tag{11}$$

While it is possible to use (11) as our approximation, we would need to reevaluate the Gaussian approximation in the denominator for each different value of each θ_i at all of the hyperparameter grid points $\boldsymbol{\gamma}_k$. For even moderate dimensionality of hyperparmeters (this is one of the reasons to limit $|\boldsymbol{\gamma}| \leq 6$), proceeding with this version of the Laplace approximation can quickly become prohibitively expensive. The authors propose two modifications to this "naive" Laplace approximation to increase the speed of the computation. In the first change, they assume a further approximation for the modal configuration of the denominator in (11), and in the second change they assume that the influence of some θ_j on θ_i will be negligible and can be ignored.

In the first modification, we evaluate the right hand side of (11) under the modal approximation

$$\boldsymbol{\theta}^*_{-i}(\gamma) \approx \mathrm{E}_{\tilde{\pi}_G}[\boldsymbol{\theta}_{-i}|\theta_i,\mathbf{y},\boldsymbol{\gamma}]$$

which is obtained from the Gaussian approximation to $\pi(\boldsymbol{\theta}|\mathbf{y},\boldsymbol{\gamma})$ and is calculated using standard multivariate Gaussian techniques (e.g. see Casella and Berger (1990)). The authors note that this may be calculated efficiently on a computer using rank 1 updates from the unconditional mean by applying (6).

In the second modification, we assume that there exists some radius of influence, R_i, around each θ_i and any θ_j outside of this radius will not affect the marginal of θ_i. The first modification leads to

$$\mathrm{E}[x_j|x_i] - \mu_{G_j}(\boldsymbol{\gamma}) = r_{ij}(\boldsymbol{\gamma})\frac{\sigma_{G_j}(\boldsymbol{\gamma})}{\sigma_{G_i}(\boldsymbol{\gamma})}\left(\theta_i - \mu_{G_i}(\boldsymbol{\gamma})\right)$$

for $r_{ij}(\boldsymbol{\gamma}) = \rho_{G_{ij}}(\boldsymbol{\gamma}) * \sigma_{G_i}(\boldsymbol{\gamma})$, where $\mu_{G.}, \sigma_{G.}, \rho_{G..}$ are all from the Gaussian approximation to the full conditional found in (8) or equivalently in the denominator of (9). The radius of influence is then defined by the author by a straightforward set construction

$$R_i(\boldsymbol{\gamma}) = \{j : |r_{ij}(\boldsymbol{\gamma})| > 0.001\}.$$

Using these two simplifications still requires us to evaluate (11) for different values of θ_i. If we standardize θ_i using a Gaussian approximation as described above such that

$$\theta_i^s = \frac{1}{\sigma_{G_i}(\boldsymbol{\gamma})} \times \left(\theta_i - \mu_{G_i}(\boldsymbol{\gamma})\right),$$

then we can select points θ_i to use in the evaluation of (11) by using a Gauss-Hermite quadrature rule to select the standardized evaluation points.

Finally, we represent the density by the semi-parametric form

$$\tilde{\pi}_{\mathcal{L}}(\theta_i|\mathbf{y},\boldsymbol{\gamma}) \propto \mathcal{N}\Big(\mu_{G_i}(\boldsymbol{\gamma}), \sigma^2_{G_i}(\boldsymbol{\gamma})\Big) \times \exp[\text{cubic spline}(\theta_i)]. \tag{12}$$

The cubic spline is fitted to the difference between the log-density of the Gaussian and the Laplace approximations to $\pi(\theta_i|\mathbf{y},\boldsymbol{\gamma})$ at the points chosen from the Gauss-Hermite quadrature rule.

The largest computational benefits result from the need to only factorize a matrix of dimension $|R_i(\boldsymbol{\gamma})|\mathrm{x}|R_i(\boldsymbol{\gamma})|$ in the denominator for each evaluation of (11).

Simplified Laplace Approximation

Finally, the authors propose a simplified Laplace approximation (SLA) which is derived by expanding the Laplace approximation, $\tilde{\pi}_{\mathcal{L}}(\theta_i|\mathbf{y},\boldsymbol{\gamma})$ around $\mu_{G_i}(\boldsymbol{\gamma})$ up to third order terms. Inclusion of terms higher than the second order allow the approximation to correct for skewness which a second order (Gaussian) approximation can not do.

After expansion and simplification, the resulting approximation yields

$$\log\left[\tilde{\pi}_{\mathcal{SL}}(\theta_i^s|\boldsymbol{\gamma},\mathbf{y})\right] = \text{constant} + [\theta_i^s]\xi_i(\boldsymbol{\gamma}) - \frac{1}{2}[\theta_i^s]^2 + \frac{1}{6}[\theta_i^s]^3\zeta_i(\boldsymbol{\gamma}) + \cdots \tag{13}$$

where we define

$$\begin{aligned}
\xi_i(\boldsymbol{\gamma}) &= \frac{1}{2}\sum_{j\in\mathcal{I}:j\neq i} \sigma^2_{G_j} * (\boldsymbol{\gamma})\left(1 - corr_{\tilde{\pi}_G}(\theta_i,\theta_j)^2\right) * \dddot{d}_j(\mu_{G_i}(\boldsymbol{\gamma}),\boldsymbol{\gamma}) * \sigma_{G_j}(\boldsymbol{\gamma}) * r_{ij}(\boldsymbol{\gamma}),\\
\zeta_i(\boldsymbol{\gamma}) &= \sum_{j\in\mathcal{I}:j\neq i} \dddot{d}_j(\mu_{G_i}(\boldsymbol{\gamma}),\boldsymbol{\gamma}) * [\sigma_{G_j}(\boldsymbol{\gamma}) * r_{ij}(\boldsymbol{\gamma})]^3, \text{ and}\\
\dddot{d}_j(\theta_i,\boldsymbol{\gamma}) &= \frac{\partial^3}{\partial x_j^3}\log\left[\pi(y_j|\theta_j,\boldsymbol{\gamma})\right]\Big|_{\theta_j=\mathrm{E}[\theta_j|\theta_i]}.
\end{aligned}$$

If we compare the form of (13) to the quadratic kernel of a normal density ($\frac{-1}{2\sigma^2}x^2 + \frac{\mu}{\sigma^2}x - \frac{mu^2}{2\sigma^2}$), we see that the variance is 1 and $\xi_i(\cdot)$ plays the role of something like the mean. This leaves $\zeta_i(\cdot)$ as the non-Gaussian term which we can use to correct skewness. While (13) does not define a density, we can use this derivation to fit a skew normal density of the form

$$\pi_{\text{skew}} = \frac{2}{\omega}\phi\left(\frac{z-\lambda}{\omega}\right)\Phi\left(\psi\frac{z-\lambda}{\omega}\right). \tag{14}$$

Here we have that $\omega > 0$ is the scale parameter, λ is the location parameter, and ψ is the skewness parameters for $\phi(\cdot)$ and $\Phi(\cdot)$ the density and distribution of the standard normal (see Azzalini and Capitanio (1999) for information on the skew normal). By setting the mean of the skew normal, $\omega^2(1 - \frac{2\psi^2}{\pi(1+\psi^2)})$, equal to $\xi_i(\boldsymbol{\gamma})$ and setting the variance of the skew normal, $\lambda + \omega\frac{\psi}{\sqrt{1+\psi^2}}\sqrt{2/\pi}$, equal to 1 (we noted this above) we leave only $\zeta_i(\boldsymbol{\gamma})$ to contribute to the skewness parameter ψ. Using Taylor expansions on the log-skew normal density about the location parameter, we can first find an approximate mode and then by expanding the third order derivative of the log-skew normal density about the modal approximation, we achieve the third necessary equation, $(4-\pi)\sqrt{2}(\frac{\psi}{\sqrt{\pi}\omega})^3$, which we set equal to $\zeta_i(\boldsymbol{\gamma})$. Using the system of three equations, we can then fit the skew normal approximation which we call our simplified Laplace approximation. That is, we fit the skew normal by solving the following system of equations

$$\begin{aligned}
\xi_i(\boldsymbol{\gamma}) &= \omega^2\left(1 - \frac{2\psi^2}{\pi(1+\psi^2)}\right) \\
1 &= \lambda + \omega\frac{\psi}{\sqrt{1+\psi^2}}\sqrt{2/\pi} \\
\zeta_i(\boldsymbol{\gamma}) &= (4-\pi)\sqrt{2}\left(\frac{\psi}{\sqrt{\pi}\omega}\right)^3.
\end{aligned} \tag{15}$$

While these equations are for a standardized variable θ_i^s, they may be tweaked for a non-standardized variable by replacing θ_i^s by $\frac{\theta_i - \mu}{\sigma}$ in (13) and reworking (15). See Appendix A for the details.

5.5 STEP 3: Numerically integrate to find the approximations $\tilde{\pi}(\theta_i|\mathbf{y})$

In the last step, we simply implement numerical integration to determine the semi-parametric forms of the posterior marginals $\tilde{\pi}(\theta_i|\mathbf{y})$. Since we explored the density of the hyperparameter using a standardized grid size, numerical integration is easily implemented because all the weights are the same. As a result, we perform the calculation in (16) to find the posterior marginal density approximations using Δ_k equal to the grid exploration step size raised to the the dimensionality of our hyperparameter, i.e. $\Delta_k = \Delta_z^{|\boldsymbol{\gamma}|}$.

$$\tilde{\pi}(\theta_i|\mathbf{y}) = \sum_k \tilde{\pi}(\theta_i|\boldsymbol{\gamma}_k, \mathbf{y})\tilde{\pi}(\boldsymbol{\gamma}_k|\mathbf{y})\Delta_k. \tag{16}$$

5.6 Assessing the Approximation

The authors recommend two methods for assessing the approximations for the posterior marginal densities that are found with the INLA methodology using the three different approximations described in section 5.4.

Symmetric Kullback-Leibler Divergence

One of the recommended methods is to use the symmetric Kullback-Leibler divergence (SKLD) to measure how far off the successive approximations are from each other. For two probability distributions P and Q, the Kullback-Leibler divergence of Q from P is defined as

$$D_{KL}(P||Q) = \int ln\left(\frac{p(x)}{q(x)}\right) p(x)dx,$$

with the obvious analogue in discrete cases. This is not a symmetric measure, and the SKLD is the average $D_{KL}(P||Q)$ and $D_{KL}(Q||P)$. If the Gaussian approximation is close to the SLA (i.e. SKLD between the two posterior density estimates is small), then the user can proceed assuming that either approximation is fairly accurate. If not, the authors propose to check how far off the SLA is from the Laplace approximation. Again, if the two estiamtes are close, the SLA and Laplace are assumed to be valid. If the Laplace approximation is relatively dissimilar to the SLA, then they recommend using the Laplace approximation but R-INLA delivers a warning suggesting that the approximations may be 'problematic.'

Effective Sample Size

The second strategy uses the effective number of parameters $p_D(\theta) \approx n - \text{tr}[\mathbf{Q}(\theta)\mathbf{Q}^*(\theta)^{-1}]$ as in Spiegelhalter et al. (2002). If $p_D(\theta) = 0$, then the data is non-informative and the approximation is exact because our posteriors will be exactly Gaussian. As $p_D(\theta)$ increases, the data becomes more informative, and $p_D(\theta)$ may be seen to represent the extent to which we've departed from our Gaussian prior and the dependence structure therein encoded. This is related to the authors quick description of aymptotic results which I leave the reader to find in Rue et al. (2009).

6 Examples and Results

In this section, we show two examples implementing the INLA approach. In the first example, we apply the method to reconstruct a latent field which has been used to generate observed Bernoulli (which is the hard case) and Binomial ($n = 20$, which is an easier case) data. In the second example, the INLA method is applied to a Tokyo rain series dataset (Kitagawa, 1987) to reconstruct a dependent latent field logit-linked to the daily probabilities of rain throughout the year. Although not in the paper, this example is one of the prominent examples on the authors website ⟨http://www.r-inla.org⟩.

6.1 Bernoulli and Binomial Observations from a First-Order Auto-Regressive Latent Field

The simulation study applies the INLA methodology to the task of finding posterior estimates for an auto-regressive latent field which generates discrete count data. As in Rue et al. (2009), for this initial example,

the hyperparameters are fixed in the following model setup. For the simulation, we first construct an auto-regressive order-one latent field with an unknown mean, μ, such that

$$\begin{aligned}
\mu &\sim \mathcal{N}(0,1) \\
f_1 - \mu &\sim \mathcal{N}(0,1) \\
f_t - \mu | f_1, \ldots, f_{t-1}, \mu &\sim \mathcal{N}\left(\phi(f_{t-1} - \mu), \sigma^2\right) \text{ for } t = 2, \ldots, 50 \ ,
\end{aligned}$$

for $\phi = 0.85$, and $\text{var}(f_t - \mu | \mu) = 1$. The field is simulated by first drawing μ, then $f_1 - \mu$, $f_2 - \mu$ and so on. Once the field is generated, we draw observations from

$$y_t | (\eta, \mu) \sim \text{Binomial}\left(n, \text{logit}^{-1}(\eta_t)\right) \text{ for } t = 1, \ldots, 50 \ ,$$

where we take $n = \{1, 20\}$ and in this case our structured additive predictor is simply $\eta_t = f_t$.

Fig. 2 displays a comparison of the Gaussian, Laplace, and Simplified Laplace approximations for a randomly selected node (node 3). Table 1 compares Symmetric Kullback-Leibler Divergence between the three approximations for the two simulations. Fig. 3(a) and (b) shows latent field simulations as well as the Bernoulli (left-hand side) and Binomial (right-hand side) observations simulated using the latent field. Given the (sum of) binary results, the goal is to determine the posterior estimates for some underlying latent field. Fig. 3(c) and (d) display my posterior modal/median estimates for the latent field with 95% posterior quantiles, and Fig. 3(e) and (f) display my posterior modal estimates for the latent field using R-INLAs results for the posterior distribution of μ. This is discussed in more detail below. All INLA results in Fig. 3 are from using the Gaussian approximation for the marginal conditionals. This was chosen for the computational efficiency, and based on Fig. 2 and Table 1, there seems to be very little extra error due to using the Gaussian approximation.

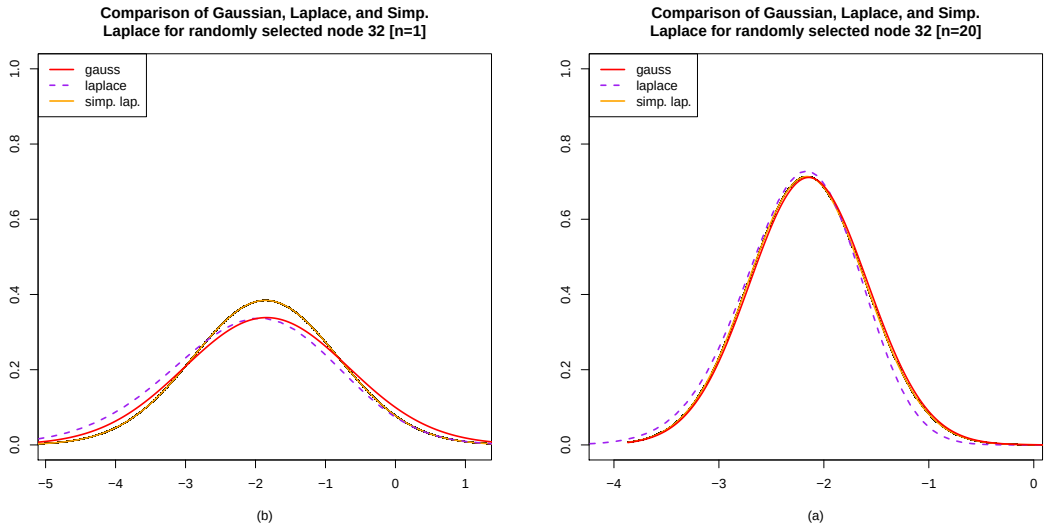


Figure 2: This figure demonstrates the differences between the Gaussian, the Laplace, and the Simplified Laplace approximations in both the Bernoulli and Binomial($n = 20$) cases for the same node.

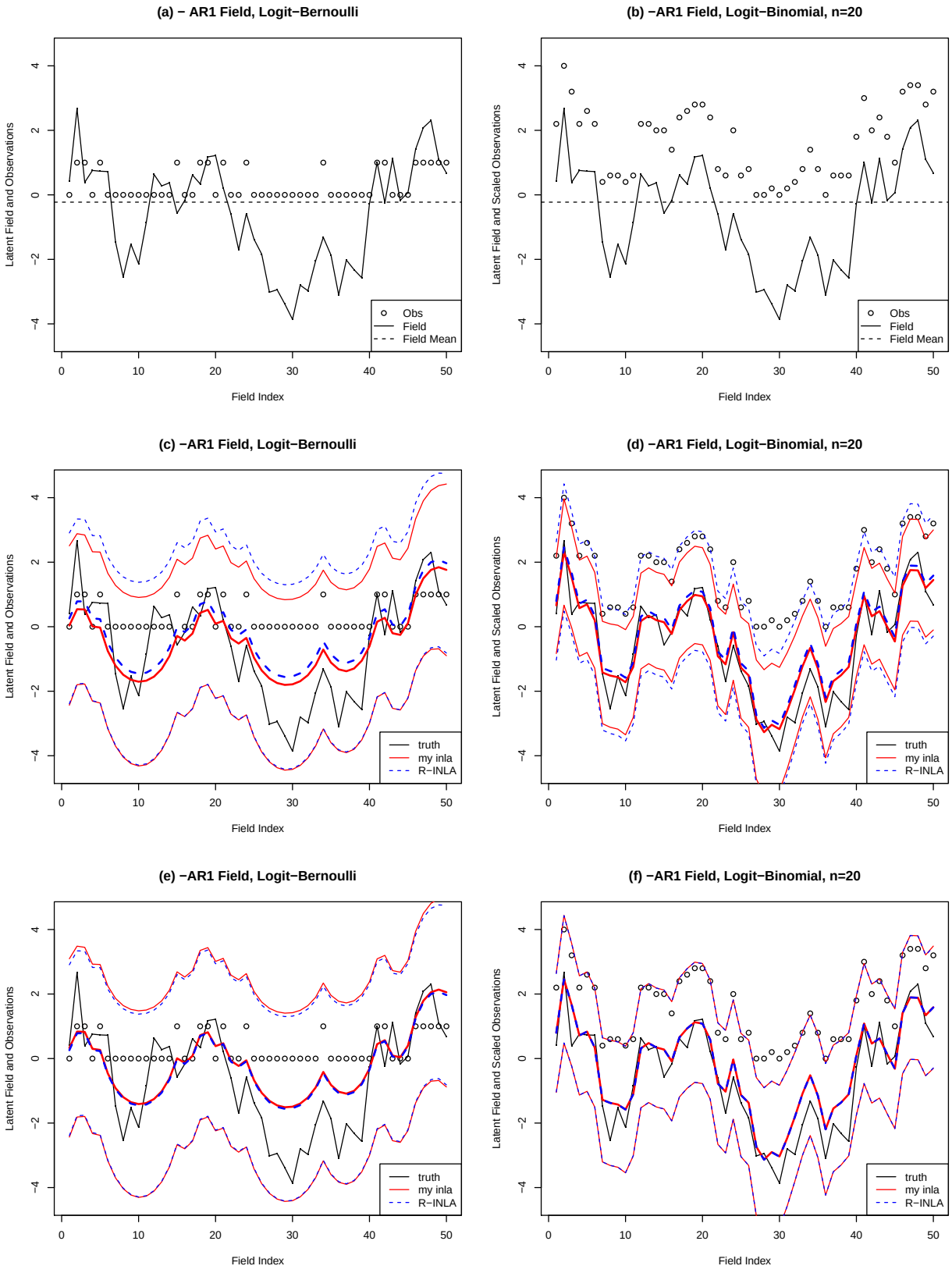


Figure 3: (a) and (b) : Latent field with Bernoulli and scaled Binomial(n=20) observations. (c) and (d) : My INLA posterior modes and 2.5% and 97.5% quantiles contrasted against equivalent R-INLA results. (e) and (f) : My INLA posterior f_i modes and 2.5% and 97.5% quantiles but using R-INLA posterior mode and variance for μ contrasted against equivalent R-INLA results.

	(a) $n = 20$			(b) $n = 1$		
	g	sl	lap	g	sl	lap
Gauss	0.0000	0.0403	0.0393	0.0000	0.0227	0.0215
Simp. Laplace	•	0.0000	0.0002	•	0.0000	0.0002
Laplace	•	•	0.0000	•	•	0.0000

Table 1: Symmetric Kullback-Leibler Divergence values between the three posterior approximations for node 35 shown in each plot in Fig. 2.

In this simulation study, I was able to exactly match the R-INLA results if I used their estimate for the mean parameter μ, but I was slightly off if I used my own estimates for μ as shown in Fig. 3. That is, my field estimates were exact up to a small constant shift deviation as seen in Fig. 3. This deviation in nuisance parameter estimation is further discussed at the end of this section. From Table 1, we see that the Gaussian, Simplified Laplace, and Laplace posterior density estimates are all very similar (and are similar for all nodes). In this case, we can proceed with the Gaussian approximation for computational efficiency without much degradation of the approximation. This is further seen in the effective sample size calculations shown in Table 3.

6.2 Tokyo Rainfall Data

For the second example, we attempt to find posterior estimates for a dependent latent field which governs the daily probability of rain in Tokyo. The time series is a binomial series, recorded over 1983 and 1984 (Kitagawa, 1987) which consists of binary indicators for the 731 (1984 was a leap year) days indicating whether or not at least 1 mm of rain was observed. Fig. 5 has a plot of the data in background. We assume a *circular* random walk of order 2 for our underlying latent GMRF which has $n = 366$ dimensions. The model is constructed such that

$$\begin{aligned}
\kappa &\sim \text{Gamma}(1, 0.0001) \\
\boldsymbol{\theta} &\sim \mathcal{N}(\mathbf{0}, \kappa\mathbf{Q}), \text{ with } |\boldsymbol{\theta}| = 366 \\
y_t|\theta_t &\sim \text{Binomial}\left(n_t, p_t = \text{logit}^{-1}(\theta_t)\right) \text{ for } t = 1, \ldots, 366.
\end{aligned}$$

In terms of our problem formulation in (1), we have for our parameters that $\eta_t = f_t = \theta_t$ and for our hyperparameters $\gamma = \kappa$. The precision matrix for a circular random walk of order 2 may be found in Appendix C. Unlike the first example, this problem has a (nuisance) hyperparemter κ which must first be explored. Fig. 4 plots my INLA results contrasted against MCMC results for the estimated posterior densities of the hyperparameter $\frac{1}{\sqrt{\kappa}}$. My results are displayed for two different grid exploration step-sizes ($\Delta_z = \{0.1, 0.01\}$) and stopping criterion tolerance $\delta = 5$ to demonstrate the exploration. Numerical comparisons between my INLA, R-INLA, and MCMC results may be found in Table 2.

Once the hyperparameter space has been explored, we can continue through INLA steps two and three to obtain posterior estimates and confidence intervals for out latent variables (or since the logit link is monotonic,

we can easily transform these into posterior estimates for the latent probabilities). Fig. 5 contains the data series, as well as posterior probability estimates from my INLA, R-INLA, and MCMC results.

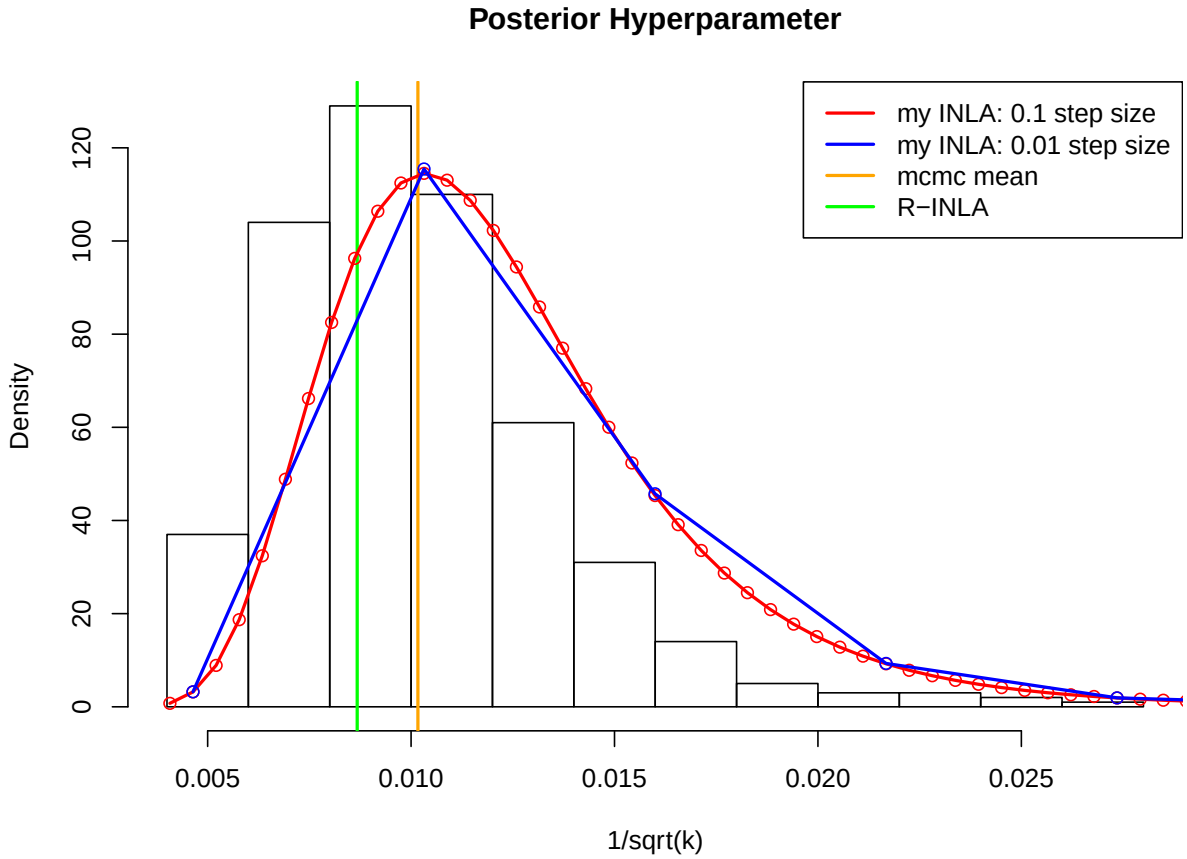


Figure 4: Hyperparameter (standard deviation) posterior density found by MCMC, my INLA with two different grid exploration step sizes and the posterior estimate from R-INLA. Vertical lines (and the mode in my density estimates) represent the posterior estimates.

Finally, Laplace and Simplified Laplace approximations were implemented. Fig. 6 contrasts the Gaussian, Laplace, and Simplified Laplace approximations for randomly selected node 108 against each other and also compares my results against the respective R-INLA estimates and MCMC results.

In this Tokyo rainfall example, my posterior estimates were relatively robust to the differences in hyperparameter estimates that my INLA and R-INLA produced. Fig. 4 shows the differences between my estimates of the posterior hyperparameter density contrasted against R-INLA and MCMC results. My estimates are actually closer to the MCMC simulation results and my standard deviation is slightly larger (or my precision is smaller). The differences in κ estimates is later on reflected in both Figures 5 and 6. The precision determines how quickly the latent field probabilities can move around the state-space in the random walk of order 2. Since my posterior hyperparameter estimates yield larger standard deviations, my field estimates have larger "peak and valley" values because more "movement" is permitted.

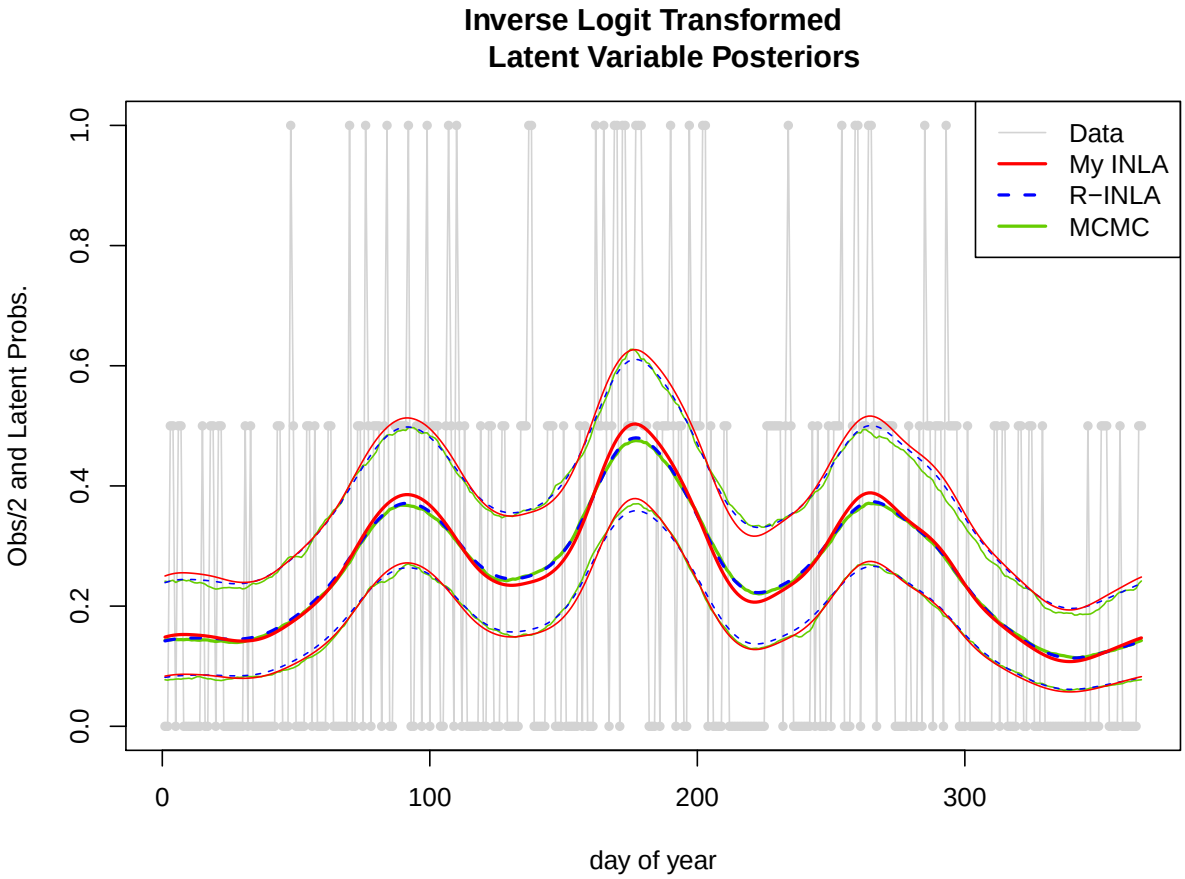


Figure 5: The grey lines in the background are scaled versions of the Binomial dataset for each day over the 1983-1984 two year period. On top of the dataset are posterior estimates, 2.5% and 97.5% posterior quantiles found using my INLA, R-INLA and MCMC simulation.

	MCMC	My INLA	R-INLA
κ est.	12978.21	13308.64	13287.47
κ SD	9971.059	-	8962.27
$\frac{1}{\sqrt{\kappa}}$ est.	0.008675	0.008777	0.008668
$\frac{1}{\sqrt{\kappa}}$ SD	0.003432	0.002449	-
total time (sec)	57.14	17.23	1.87

Table 2: Comparison of different hyperparameter estimates. Times are for the duration of the entire process (not just hyperparameter estimation) using Gaussian approximations.

In both of these examples, I demonstrate the ability to reproduce R-INLA results up to a reasonable accuracy. My estimates for the latent field parameters seem very accurate up to differences in estimating the nuisance parameters (μ in the simulations and κ in the Tokyo example). Some of these differences come from steps in the algorithm (tolerances to pick in recursions, grid exploration methods, ...), some come from numerical precision difficulties I ran into in R (e.g. exponentiating log-likelihood values), and some variations may be due to newer algorithm implementation for posterior hyperparameter estimation in R-INLA (Martins et al., 2012). In particular, I know that the my Laplace approximation in Fig. 6 suffered

from precision errors which may be part of the reason it's further off from my Gauss and Simplified Laplace approximations.

	Simulation		
	$n = 1$	$n = 20$	Tokyo Rainfal
My INLA	10.67	34.34	10.18
R-INLA	10.64	34.27	9.79

Table 3: Effective sample size calculations for my INLA compared to R-INLA results in the two examples.

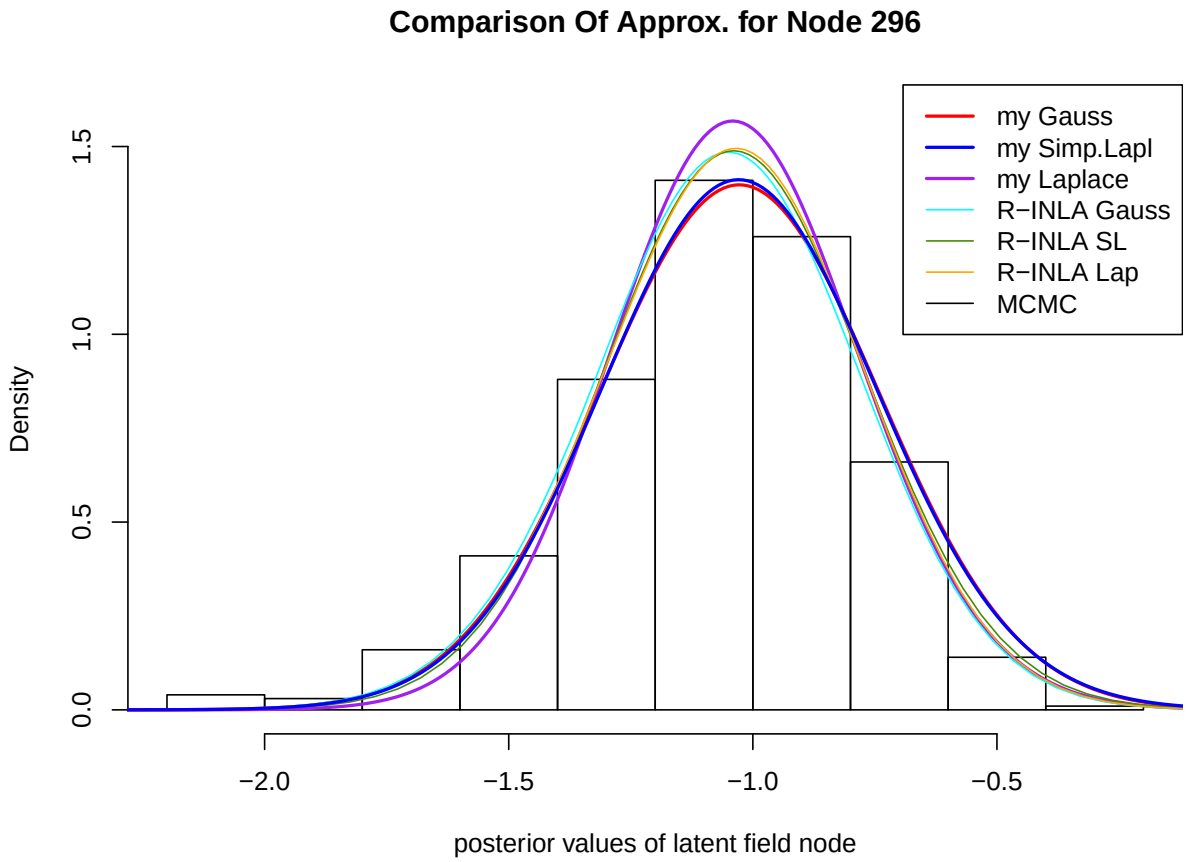


Figure 6: For randomly selected node 108 (April 17^{th}), we display my INLA results using the Gaussian, Laplace, and Simplified Laplace examples as well as the Laplace approximation from R-INLA (which is nearly identical with the R-INLA Gaussian and Simplified Laplace approximations).

7 Discussion

I congratulate Rue, Martino and Chopin on developing an innovative and highly successful algorithm which filled and still fills a sorely needed niche in the statistics and machine learning literature, namely that of a quick and accurate alternative to MCMC sampling. In Besag et al. (1995), Besag states that he believes MCMC methods are putting probability back into statistics. In a discussion of Rue et al. (2009), Knorr-Held

and Riebler see the INLA methodology as putting the "numerics back into statistics," and I would agree with this statement. That said, in the field of statistics, numerics cannot be used as a complete substitute for probability but instead the two methods should be used in conjunction. For instance, given a dataset that took (on the conservative side) weeks to collect, is it a serious extra inconvenience to allow an MCMC chain to run for a few days until convergence if you believe that you're implementing a reasonable model? On the other hand, to allow your chain to run until convergence over a few days only to decide it was misspecified is not an efficient use of your time. I think that the INLA methodology really shines for initial exploration of a dataset. After the modelers have a good grasp on the situation, they can then "bring back the probability" and allow an MCMC chain to yield results as close to the truth as patience will allow.

Of course, INLA has its drawbacks. For example, the computation time is exponential in the number of hyperparameters, and the Laplace approximation used to explore the hyperparameter density relies on the unimodality of the density. The first drawback is a feature of the method, and researchers should strive to avoid model specification chosen to fit within the INLA framework if it is chosen first to allow use of the method and second because it is reasonable. While there are models with more hyperparameters, many problems can be fit within the INLA framework as described in this report. As for the second drawback, it may be possible to alleviate the situation by introducing other approximations, for example a multi-modal Laplace approximation (the IterLap package in CRAN has one method implemented). Also of concern is the detailed implementation of the method. While the R-INLA package has been impressively optimized, for most users it is a "black box" technique which may hide limitations of the method for the uninquisitive user.

While the authors recommend two methods for assessing the approximation error, the only true way is to compare the results against a long MCMC chain. Their first strategy uses the effective number of parameters $p_D(\theta)$. This is reasonable, but the authors give no method to interpret $p_D(\theta)$ except to say that the approximation will be better if the value is small compared to the number of observed datapoints. In my simulation example, the effective sample size for the binomial case is three times the effective sample size for the bernoulli case (Table 3), even though we have twenty times the datapoints. Yet, the Gaussian approximation seems valid in both cases (Fig. 2). Their second technique uses the symmetric Kullback-Leibler divergence (SKLD) to measure how far off the successive approximations (Gaussian$\rightarrow$ Simplified Laplace$\rightarrow$Laplace) are from each other. Again, while a rule of thumb is proposed, there is no way to quantify the validity of the approximation. Either way, a more rigorous exploration of approximation error is needed before the INLA methodology can be used without a final comparison against MCMC results.

In conclusion, in this report we've explored the INLA methodology and elaborated on some of the details of the implementation. My results from the simulation study and the Tokyo rainfall data demonstrate my ability to reproduce the R-INLA results and highlight some difficulties in the implementation, in particular the instability of nuisance parameter estimation.

Acknowledgments

I would like to thank Jon Wakefield for his help and constant encouragement this quarter and both Jon and Patrick Heagerty for their advice, criticisms and comments throughout the quarter in 518. In addition, I would like to thank Vladimir Minin for recommending the paper and my peers in 518 and 572 for their attention and feedback this quarter. Finally, I would like to thank the authors of the paper. Their work has enabled me to enjoy a thought-provoking and exciting project this quarter.

References

Azzalini, A. and Capitanio, A. (1999). Statistical applications of the multivariate skew normal distribution. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 61(3):579–602.

Besag, J., Green, P., Higdon, D., and Mengersen, K. (1995). Bayesian computation and stochastic systems. *Statistical Science*, pages 3–41.

Besag, J., York, J., and Mollié, A. (1991). Bayesian image restoration, with two applications in spatial statistics. *Annals of the Institute of Statistical Mathematics*, 43(1):1–20.

Cameletti, M., Lindgren, F., Simpson, D., and Rue, H. (2011). Spatio-temporal modeling of particulate matter concentration through the spde approach. *AStA Advances in Statistical Analysis*, pages 1–23.

Casella, G. and Berger, R. L. (1990). *Statistical inference*, volume 70. Duxbury Press Belmont, CA.

Cressie, N. (1992). Statistics for spatial data. *Terra Nova*, 4(5):613–617.

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

Crowder, M. J. (1978). Beta-binomial anova for proportions. *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, 27(1):34–37.

De Bruijn, N. G. (1970). *Asymptotic methods in analysis*, volume 4. Courier Dover Publications.

Dey, D., Ghosh, S. K., and Mallick, B. K. (2000). *Generalized linear models: A Bayesian perspective*, volume 5. CRC Press.

Erdélyi, A. (1956). *Asymptotic expansions*. Number 3. Courier Dover Publications.

Gamerman, D. (1997). Sampling from the posterior distribution in generalized linear mixed models. *Statistics and Computing*, 7(1):57–68.

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

Jun, S. C., George, J. S., Kim, W., Paré-Blagoev, J., Plis, S., Ranken, D. M., and Schmidt, D. M. (2008). Bayesian brain source imaging based on combined meg/eeg and fmri using mcmc. *Neuroimage*, 40(4):1581.

Kitagawa, G. (1987). Non-gaussian state—space modeling of nonstationary time series. *Journal of the American statistical association*, 82(400):1032–1041.

Kitagawa, G. and Gersch, W. (1984). A smoothness priors–state space modeling of time series with trend and seasonality. *Journal of the American Statistical Association*, 79(386):378–389.

Knorr-Held, L. (1999). Bayesian modelling of inseparable space-time variation in disease risk.

Lang, S. and Brezger, A. (2004). Bayesian p-splines. *Journal of computational and graphical statistics*, 13(1):183–212.

Lindgren, F., Lindström, J., and Rue, H. (2010). *An explicit link between Gaussian fields and Gaussian Markov random fields: The SPDE approach*. Mathematical Statistics, Centre for Mathematical Sciences, Faculty of Engineering, Lund University.

Martins, T. G., Simpson, D., Lindgren, F., and Rue, H. (2012). Bayesian computing with inla: new features. *arXiv preprint arXiv:1210.0333*.

Minka, T. P. (2001). Expectation propagation for approximate bayesian inference. In *Proceedings of the Seventeenth conference on Uncertainty in artificial intelligence*, pages 362–369. Morgan Kaufmann Publishers Inc.

Plagnol, V. and Tavaré, S. (2004). Approximate bayesian computation and mcmc. In *Monte Carlo and Quasi-Monte Carlo Methods 2002*, pages 99–113. Springer.

Poirier, D. J. (2006). The growth of bayesian methods in statistics and economics since 1970. *Bayesian Analysis*, 1(4):969–979.

Rue, H. and Held, L. (2005). *Gaussian Markov random fields: theory and applications*, volume 104. Chapman & Hall.

Rue, H. and Martino, S. (2007). Approximate bayesian inference for hierarchical gaussian markov random field models. *Journal of statistical planning and inference*, 137(10):3177–3192.

Rue, H., Martino, S., and Chopin, N. (2009). Approximate bayesian inference for latent gaussian models by using integrated nested laplace approximations. *Journal of the royal statistical society: Series b (statistical methodology)*, 71(2):319–392.

Spiegelhalter, D. J., Best, N. G., Carlin, B. P., and Van Der Linde, A. (2002). Bayesian measures of model complexity and fit. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 64(4):583–639.

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

A Non-standardized Skew Details

By substituting $\frac{\theta_i-\mu}{\sigma} = \theta_i^{(s)}$ into (13), expanding and consolidating terms, we find equivalent forms for non-standardized random variables θ_i that we show here:

$$\begin{aligned}\left(\frac{\xi_i}{\sigma}+\frac{nu}{\sigma^2}+\frac{\zeta_i\mu^2}{2\sigma^2}\right)\times\left(\frac{1}{\sigma^2}+\frac{\mu\zeta_i}{\sigma^2}\right)^{-1} &= \omega^2\left(1-\frac{2\psi^2}{\pi(1+\psi^2)}\right)\\ \left(\frac{1}{\sigma^2}+\frac{\mu\zeta_i}{\sigma^2}\right)^{-1} &= \lambda+\omega\frac{\psi}{\sqrt{1+\psi^2}}\sqrt{2/\pi}\\ \frac{\zeta_i}{\sigma^3} &= (4-\pi)\sqrt{2}\left(\frac{\psi}{\sqrt{\pi}\omega}\right)^3.\end{aligned} \tag{17}$$

These were the equations that I implemented to fit my skew-normal approximations in the simplified Laplace approximation method. Note that ζ_i and ξ_i each depend on the choice of the hyperparameter values.

B Extensions

Since this paper was published, there have been some successor papers (e.g. Martins et al. (2012); Cameletti et al. (2011)) that have attempted to extend and further the methodology.

Martins et al. (2012) discusses some new methods already implemented in INLA with a focus on new work for approximating the posterior hyperparameter density. In particular, they discuss an asymmetric Gaussian interpolation which had been implemented in INLA with good results for a while and further discuss a new numerical integration free algorithm which is now the default in R-INLA.

Cameletti et al. (2011) use R-INLA to implement a Bayesian model for a Gaussian Markov Random Field with a Matérn covariance function represented through the Stochastic Partial Differential Equations (SPDE) approach. More information may also be found in Lindgren et al. (2010).

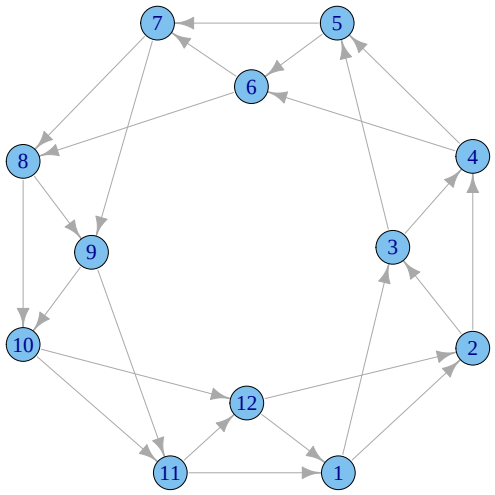
C Example Details

Tokyo Rainfall

The precision matrix for the a circular random walk of order two takes the form:

$$Q = \begin{pmatrix} 6 & -4 & 1 & * & * & * & * & \cdots & * & * & 1 & -4 \\ -4 & 6 & -4 & 1 & * & * & * & \cdots & * & * & * & 1 \\ 1 & -4 & 6 & -4 & 1 & * & * & \cdots & * & * & * & * \\ * & 1 & -4 & 6 & -4 & 1 & * & \cdots & * & * & * & * \\ * & * & 1 & -4 & 6 & -4 & 1 & \cdots & * & * & * & * \\ & \vdots & & & & \ddots & & \cdots & & & \vdots & \\ -4 & 1 & * & * & * & * & * & \cdots & * & 1 & -4 & 6 \\ 1 & * & * & * & * & * & * & \cdots & 1 & -4 & 6 & -4 \end{pmatrix}$$

If the dimension of the field were 12, the following graph would represent the GMRF:



Conditional means and precision are easily calculated as follows:

$$\mathrm{E}[\theta_i|\boldsymbol{\theta}_{-i}, \kappa] = \frac{4}{6}(\theta_{i-1} + \theta_{i+1}) - \frac{1}{6}(\theta_{i-2} + \theta_{i+2})$$

$$\mathrm{Prec}[\theta_i|\boldsymbol{\theta}_{-i}, \kappa] = 6 * \kappa.$$