Improving the INLA approach for approximate Bayesian inference for latent Gaussian models

Egil Ferkingstad*1,2 and Håvard Rue1

1Department of Mathematical Sciences, NTNU, Norway
2Science Institute, University of Iceland, Reykjavik, Iceland

April 8, 2015

Abstract

We introduce a new copula-based correction for generalized linear mixed models (GLMMs) within the integrated nested Laplace approximation (INLA) approach for approximate Bayesian inference for latent Gaussian models. GLMMs for Binomial and Poisson data with many zeroes and low counts have been a somewhat difficult case for INLA, and the case of binary data has been particularly problematic. Our new correction has been implemented as part of the R-INLA package, and adds only negligible computational cost. Empirical evaluations on both real and simulated data indicate that the method works well. Keywords: Bayesian computation; copulas; generalized linear mixed models; integrated nested Laplace approximation; latent Gaussian models

1 Introduction

Integrated Nested Laplace Approximations (INLA) were introduced by Rue et al. (2009) as a tool to do approximate Bayesian inference in latent Gaussian models (LGMs). The class of LGMs covers a large part of models used today, and the INLA approach has been shown to be very accurate and extremely fast in most cases. Software is provided through the R-INLA package, see http://www.r-inla.org.

An important subclass of LGMs is the rich family of generalized linear mixed models (GLMMs) with Gaussian priors on fixed and random effects (Breslow & Clayton, 1993; McCulloch et al., 2008). The use of INLA for Bayesian inference for GLMMs was investigated by Fong et al. (2010), who reanalyzed all of the examples from Breslow & Clayton (1993). Fong et al. (2010) found that INLA is somewhat inaccurate for binary data with few or no replications. In this paper, we introduce a new correction term for INLA, significantly improving accuracy while adding negligibly to the overall computational cost.

To set the scene, we consider a minimal simulated example illustrating the problem (postponing more thorough empirical evaluations until Section 3). Consider the following model: For \( i = 1, 2, \ldots, n \), let \( \text{Prob}(y_i = 0) = 1 - p_i \), \( \text{Prob}(y_i = 1) = p_i \), and

\[
\logit(p_i) = \beta + u_i,
\]

\( \star \)egil.ferkingstad@gmail.com
where $u_i \sim N(0, \sigma^2)$, iid. Let the precision $\sigma^{-2}$ have a Gamma(1, 1) prior, while the prior for $\beta$ is $N(0, 1)$. We simulated data from this model, setting $n = 100$, $\sigma^2 = 1$ and $\beta = 2$. Figure 1 shows the resulting posterior distributions for the intercept $\beta$ and for the log precision, $\log(\sigma^{-2})$, where the histograms show results from long MCMC runs using JAGS (Plummer, 2013), the black curves show posteriors from INLA without any correction, and the red curves show results using the new correction defined in Section 2. While some of our later examples show more dramatic differences between INLA and long MCMC runs, these results exemplify quite well our general experience with using INLA for “difficult” binary response GLMMs: Variances of both random and fixed effects tends to be underestimated, while the means of the fixed effects are reasonably well estimated.

One part of the problem is that the usual assumptions ensuring asymptotic validity of the Laplace approximation do not hold here (for details on asymptotic results, see the discussion in Section 4 of Rue et al. (2009)). The independence of the random effects make the effective number of parameters (Spiegelhalter et al., 2002) on the order of the number of data points. In more realistic models, there is often some amount of smoothing or replication that alleviates the problem, but it may still occur. Except in the case of spline smoothing models (Kauermann et al., 2009), there is a lack of strong asymptotic results for random effects models with a large effective number of parameters. In the simulation from model (1), the data provide little information about the parameters, with the shape of the likelihood function adding to the problem. Figure 2 illustrates the general problem. Here, the top panel shows the log-likelihood of a single Bernoulli observation $Y$ as a function of the linear predictor $\eta$, i.e. $\log(\text{Prob}(Y = 1)) = \log(p)$ where $\logit(p) = \eta$. The bottom panel shows the corresponding derivative. We see that the log-likelihood gets very flat (and the derivative near zero) for high values of $\eta$, so inference will get difficult.

Bayesian and frequentist estimation for GLMMs with binary outcomes has been given some attention in the recent literature (Capau et al., 2013; Grilli et al., 2014; Sauter & Held, 2015), but a computationally efficient Bayesian solution appropriate for the INLA approach has been lacking. An alternative to our new approach would be to consider higher-order Laplace approximations (Shun & McCullagh, 1995; Raudenbush et al., 2000; Evangelou et al., 2011), other
Figure 2: Log-likelihood (top panel) and derivative of log-likelihood (bottom panel) for a single Bernoulli observation as a function of the linear predictor in a logistic model.
modifications to the Laplace approximation (Ruli et al., 2015; Ogden, 2015), or expectation propagation-type solutions (Cseke & Heskes, 2010), but we view them as too costly to be applicable for general use in INLA. The motivation for using INLA is speed, so we see it as a design requirement for any correction that it should add minimally to the running time of the algorithm.

We proceed as follows. In Section 2, we present a derivation of our new correction method. Section 3 presents empirical studies, both on real and simulated data, showing that the method works well in practice. Finally Section 4 gives a brief discussion and some concluding remarks.

2 Methodology

Consider a latent Gaussian model (Rue et al., 2009), with hyperparameters $\theta = (\theta_1, \ldots, \theta_p)'$, latent field $x = (x_1, \ldots, x_n)'$ and observed data $y = \{y_i : i \in I\}$ (for $I \subseteq \{1, \ldots, n\}$), where the joint distribution may be written as

$$
\pi(x, \theta, y) = \pi(\theta)\pi(x|\theta)\prod_{i \in I} \pi(y_i|x_i, \theta)
$$

where $\pi(x|\theta)$ is a multivariate Gaussian density. We want to approximate the posterior marginals $\pi(x_i|y)$ and $\pi(\theta_j|y)$. The Laplace approximation of $\tilde{\pi}(\theta|y)$ is

$$
\tilde{\pi}(\theta|y) \propto \frac{\pi(x, \theta, y)}{\pi_G(x|\theta, y)|_{x=\mu(\theta)}}
$$

(2)

where $\pi_G(x|\theta, y)$ is a Gaussian approximation found by matching the mode and the curvature at the mode of $\pi(x|\theta, y)$, and $\mu(\theta)$ is the mean of the Gaussian approximation.

Given $\tilde{\pi}(\theta|y)$ and some approximation $\tilde{\pi}(x_i|\theta, y)$ (see below), the posterior marginals of interest are calculated using numerical integration:

$$
\tilde{\pi}(\theta_j|y) = \int \tilde{\pi}(\theta|y)d\theta_{-j},
$$

$$
\tilde{\pi}(x_i|y) = \int \tilde{\pi}(x_i|\theta, y)\tilde{\pi}(\theta|y)d\theta.
$$

(3)

In the current implementation of INLA, the $\tilde{\pi}(x_i|\theta, y)$ used in (3) are approximated using skew normal densities $\tilde{\pi}_{SN}(x_i|\theta, y)$ (Azzalini & Capitanio, 1999), based on a second Laplace approximation; see Section 3.2.3 of Rue et al. (2009) for details. Notice that, in Equation (2) we use a Gaussian approximation $\pi_G(x|\theta, y)$, with marginals $\pi_G(x_i|\theta, y), i = 1, \ldots, n$. Thus, both $\pi_{SN}(x_i|\theta, y)$ and $\pi_G(x_i|\theta, y)$ are approximations to the marginals $\pi(x_i|\theta, y)$, but the $\pi_{SN}(x_i|\theta, y)$ are more accurate since they are based on a second Laplace approximation. In (2) we need to approximate the full joint distribution $\pi(x|\theta, y)$. Our basic idea is to use the improved approximations $\tilde{\pi}_{SN}(x_i|\theta, y)$ in order to construct a better approximation to the joint distribution $\pi(x|\theta, y)$. We aim for an approximation of $\pi(x|\theta, y)$ that retains the dependence structure of the Gaussian approximation $\pi_G(x|\theta, y)$, while having the improved marginals $\tilde{\pi}_{SN}(x_i|\theta, y)$. This can be achieved by using a Gaussian copula.

Before we describe the copula construction, we need to define some notation. First, for $i = 1, \ldots, n$, let $\tilde{\mu}_i(\theta)$ and $\tilde{\sigma}_i^2(\theta)$ denote the mean and variance of each marginal $\tilde{\pi}_{SN}(x_i|\theta, y)$, and let $F_i$ be the cumulative distribution function corresponding to $\tilde{\pi}_{SN}(x_i|\theta, y)$. Second, let $\tilde{x}_i \sim F_i$ and assume that $F_i$ is the distribution of $\tilde{z}_i = (\tilde{x}_i - \tilde{\mu}_i(\theta))/\tilde{\sigma}_i(\theta)$. As usual, $\Phi$ denotes the cumulative standard Gaussian distribution function. Furthermore, let $\mu_i(\theta)$ and $\sigma_i^2(\theta)$ denote the marginal means and variances of the Gaussian approximation $\pi_G(x_i|\theta, y)$, let $Q(\theta)$ be the
Comparing Equations (5), (6), and (7), we see that the copula approximation can be implemented as follows:

\[ \tilde{x}_i = \tilde{\sigma}_i(\theta) \tilde{F}_i^{-1} \left[ \Phi \left( \frac{x_i - \mu_i(\theta)}{\sigma_i(\theta)} \right) \right] + \tilde{\mu}_i(\theta). \]  

(4)

We may simplify the construction above by replacing the \( F_i \) in (4) by \( \Phi \) and assuming \( \tilde{\sigma}_i(\theta) = \sigma_i(\theta) \). This means that we do not correct for skewness, but we take advantage of the improved mean \( \tilde{\mu}_i(\theta) \) from \( \tilde{\pi}_{SN}(x_i|\theta,y) \). We denote this as the “mean only” correction. (We will later discuss the possibility of retaining \( F_i \) as a skew normal, while still assuming \( \tilde{\sigma}_i(\theta) = \sigma_i(\theta) \); this we denote as the “mean and skewness” correction.) In the simple “mean only” case, the transformation reduces to a shift in mean:

\[ \tilde{x}_i = x_i - \mu_i(\theta) + \tilde{\mu}_i(\theta), \]

the Jacobian is equal to one, and the transformed joint density function is a multivariate normal with mean \( \tilde{\mu} \) and precision matrix \( Q \), i.e.

\[ \log \tilde{\pi}(\tilde{x}|\theta,y) = \frac{1}{2} \log |Q(\theta)| - \frac{1}{2}(\tilde{x} - \tilde{\mu}(\theta))^T Q(\theta)(\tilde{x} - \tilde{\mu}(\theta)) + \text{constant} \]  

(5)

In the Laplace approximation defined in Equation (2), the density functions above should be evaluated in the point \( \tilde{x} = \mu(\theta) \), where \( \mu(\theta) = (\mu_1(\theta), \ldots, \mu_n(\theta))^T \) is the mean of the Gaussian approximation \( \tilde{\pi}_G(x|\theta,y) \). From Equation (2), the original (uncorrected) log posterior is

\[ \log \tilde{\pi}(\theta|y) = \log \pi(x,\theta,y) - \log \tilde{\pi}_G(x|\theta,y) + \text{constant} \]  

(6)

evaluated at \( x = \mu(\theta) \), where

\[ \log \tilde{\pi}_G(x|\theta,y)|_{x=\mu(\theta)} = \frac{1}{2} \log |Q(\theta)| + \text{constant}, \]  

(7)

Comparing Equations (5), (6), and (7), we see that the copula approximation can be implemented by adding the term \( C(\theta) \) to the already calculated log posterior evaluated at \( \mu(\theta) \), where

\[ C(\theta) = \frac{1}{2}(\mu(\theta) - \tilde{\mu}(\theta))^T Q(\theta)(\mu(\theta) - \tilde{\mu}(\theta)). \]

The addition of the term \( C(\theta) \) does not add significantly to the computational cost of INLA — this simple operation is essentially free.

For the INLA implementation of the copula correction, we have found that it is sufficient to only include fixed effects (including any random effects of length one) in the calculation of \( C(\theta) \). This gives better numerical stability and also seems to provide a more accurate approximation. Conceptually, this step involves finding \( \Sigma(\theta) = Q(\theta)^{-1} \), and then again finding...
$Q_J(\theta) = \Sigma_J(\theta)^{-1}$, where $J$ is the index set of the fixed effects, so this might seem computationally costly. However, it can be done cheaply by using the linear combination feature described in Section 4.4 of Martins et al. (2013): If $n_f$ is the number of fixed effects, only the (parallel) solution of a $n_f$-dimensional linear system is needed.

Additionally, to guard against over-correction, we perform a soft thresholding on $C(\theta)$, as follows: First we define a sigmoidal function $f(t)$:

$$f(t) = \frac{2}{1 + \exp(-2t)} - 1,$$

which is increasing, has derivative equal to one at the point $t = 0$, and where $f(t) \to 1$ as $t \to \infty$. Then we replace $C(\theta)$ by $C_t(\theta)$, where

$$C_t(\theta) = uf(C(\theta))/u,$$

for $u = n_f \xi$ and with the “correction factor” parameter $\xi > 0$ determining the degree of shrinkage (more shrinkage for smaller values of $\xi$). Since the function $f(t)$ is approximately linear with unit slope around zero, $C_t(\theta)$ will be close to $C(\theta)$ for small and moderate values of $C(\theta)$, while larger values will be increasingly shrunk toward zero. Note that since $f(t) < 1$ for all $t$, $C_t(\theta) < u$. We have found that $\xi = 10$ is a good choice, letting the correction do its job while guarding against too large changes, and we have used this value for all of the examples.

As mentioned, we have also investigated a more general case of the copula construction, where we retain $F_i$ as a skew normal distribution, i.e. the CDF of $\tilde{\pi}_{SN}(x_i|\theta, y)$ (and $\tilde{\pi}_i(\theta) = \sigma_i(\theta)$ as before). This results in a more complicated correction term $C_{\text{skew}}(\theta)$, derived in Appendix A. We have not found any appreciable differences in the accuracy compared to the simpler case without skewness, so we have concluded that the non-skew version is preferable due to its simplicity. We will show both the skew and the non-skew correction for the toenail data discussed in Section 3.2, but otherwise we show only results from the simpler non-skew version. We have tried both corrections on many (both real and simulated) data sets, and never seen a significant difference in the results.

### 3 Empirical results

#### 3.1 Have we solved the problems detected by Fong et al. (2010)?

As mentioned in the Introduction, Fong et al. (2010) studied the use of INLA for binary valued GLMMs, and they showed that the approximations were inaccurate in some cases. We have redone the full simulation experiment described on pages 10–14 of the Supplementary Material of Fong et al. (2010), both for INLA without any correction, and INLA with the “mean only” correction described in Section 2.

In the original simulation study by Fong et al. (2010), $Y_{ij}$ are iid $\text{Binomial}(m, p_{ij})$, with $i = 1, \ldots, 100$ clusters, $j = 1, \ldots, 7$ observations per cluster, and $m \in \{1, 2, 4, 8\}$. Given $x_i = 0$ for half of the sample and $x_i = 1$ for the rest, and sampling times $t = (t_1, \ldots, t_T)' = (-3, -2, -1, 0, 1, 2, 3)'$, the following two models were considered:

\begin{align*}
\text{logit } p_{ij} & = \beta_0 + \beta_1 t_j + \beta_2 x_i + \beta_3 t_j x_i + b_{0i} \\
\text{logit } p_{ij} & = \beta_0 + \beta_1 t_j + \beta_2 x_i + \beta_3 t_j x_i + b_{0i} + b_{1i} t_j,
\end{align*}

which corresponds to models (0.7) and (0.8) on page 11 of the Supplementary Material of Fong et al. (2010).
We only show the results from the $m = 1$ (i.e. binary data) version of the model (8), as this is the most difficult case with the largest errors in the approximation. The correction also works well for $m > 1$ and the more complex model (9), but this model is easier to deal with for INLA. This is seen empirically, and is also as expected based on considering the asymptotic properties of the Laplace approximation: for $m > 1$ and model (9) there is more “borrowing of strength”/replication, so the original approximation should work better.

We use the same settings as Fong et al. (2010): $b_{0i} \simiid N(0, \sigma_0^2)$ where $\sigma_0^2 = 1$, the prior Gamma$(0.5, 0.0164)$ for $\sigma_0^{-2}$ and $N(0, 1000)$ priors for the $\beta_i$. The true values of the fixed effects are $\beta = (\beta_0, \beta_1, \beta_2, \beta_3)' = (-2.5, 1.0, -1.0, -0.5)'$. We made 100 simulated data sets, running INLA both with and without the new correction, as well as very long MCMC chains using JAGS (each of the 100 datasets were run with one million MCMC samples after a burn-in of one million).

Results from the simulation study are shown in Table 1. Note that the aim here is to be as close as possible to the MCMC results, not the true values. The upper part of the table shows averages of the posterior means over the 100 simulations. We see that INLA gets much closer to the MCMC results for all parameters except $\beta_3$, which is in any case reasonably close to the MCMC value. The improvement is particularly large for the variance parameter. This is also seen in the middle panel, which shows $(E(\theta_{\text{INLA}}|y) - E(\theta_{\text{MCMC}}|y))/sd(\theta_{\text{MCMC}}|y)$ for each parameter $\theta$ (averaged over the 100 simulations), i.e. the difference in INLA and MCMC estimates scaled by the MCMC standard deviation. Here, the random effects variance and the fixed effects except $\beta_3$ are also more accurately estimated. Finally, the lower panel shows the ratios $\text{Var}(\theta_{\text{INLA}}|y)/\text{Var}(\theta_{\text{MCMC}}|y)$; here the correction improves the estimation of the variance for all parameters. For $\sigma_0^2$, $\sigma_0$ and $\log \sigma_0^{-2}$ we get very close to a ratio of one, and there are also major improvements for the fixed effects variances.

### Averages of posterior means:

| Parameter | True values | Uncorrected INLA | Corrected INLA | MCMC |
|-----------|-------------|------------------|---------------|------|
| $\sigma_0^2$ | 1.000 | 0.724 | 0.969 | 0.960 |
| $\sigma_0$ | 1.000 | 0.726 | 0.857 | 0.853 |
| $\log \sigma_0^{-2}$ | 0.000 | 1.164 | 0.782 | 0.782 |
| $\beta_0$ | -2.500 | -2.420 | 1.010 | 1.004 |
| $\beta_1$ | 1.000 | 0.985 | -1.200 | -1.205 |
| $\beta_2$ | -1.000 | -1.169 | -0.462 | -0.456 |
| $\beta_3$ | -0.500 | -0.462 | -0.479 | -0.456 |

### Comparison between INLA and MCMC, $(E(\text{INLA})-E(\text{MCMC}))/sd(\text{MCMC})$:

| Parameter | Uncorrected INLA | Corrected INLA |
|-----------|------------------|---------------|
| $\sigma_0^2$ | -0.386 | 0.008 |
| $\sigma_0$ | -0.413 | 0.011 |
| $\log \sigma_0^{-2}$ | 0.412 | -0.007 |
| $\beta_0$ | -2.420 | -0.67 |
| $\beta_1$ | 0.985 | 0.043 |
| $\beta_2$ | -1.169 | 0.004 |
| $\beta_3$ | -1.169 | -0.103 |

### Ratio of variances, $\text{Var}(\text{INLA})/\text{Var}(\text{MCMC})$:

| Parameter | Uncorrected INLA | Corrected INLA |
|-----------|------------------|---------------|
| $\sigma_0^2$ | 0.575 | 0.943 |
| $\sigma_0$ | 0.791 | 0.961 |
| $\log \sigma_0^{-2}$ | 1.140 | 0.999 |
| $\beta_0$ | 0.825 | 0.908 |
| $\beta_1$ | 0.836 | 0.873 |
| $\beta_2$ | 0.880 | 0.942 |
| $\beta_3$ | 0.891 | 0.910 |

Table 1: Results from simulation study

3.2 What happens when the random effects variance increases?

We start by discussing the toenail data, which is a classical data set with a binary response and repeated measures (De Backer et al., 1998; Lesaffre & Spiessens, 2001). The data are
from a clinical trial comparing two competing treatments for toenail infection (dermatophyte onychomycosis). The 294 patients were randomized to receive either itraconazole or terbinafine, and the outcome (either “not severe” infection, coded as 0, or “severe” infection, coded as 1) was recorded at seven follow-up visits. Not all patients attended all the follow-up visits, and the patients did not always appear at the scheduled time. The exact time of the visits (in months since baseline) was recorded. For individual \( i \), visit \( j \), with outcome \( y_{ij} \), treatment \( \text{Trt}_i \) and time \( \text{Time}_{ij} \) our model is then

\[
y_{ij} \sim \text{Bernoulli}(p_{ij})
\]

\[
\logit(p_{ij}) = \alpha_0 + \alpha_{\text{Trt}_i} + \alpha_{\text{Time}_{ij}} + \alpha_{\text{Trt}_i \text{Time}_{ij}} + b_i
\]

Notice that this is the same model as model (8), except that the time variable here varies over individuals. INLA underestimates \( \sigma^2 \) quite severely. The top panel of Figure 3 shows the different estimates of the posterior distribution of the log precision, \( \log \sigma^{-2} \). The histogram shows the results from a long MCMC run using JAGS, the black curve shows the posterior from INLA without the correction, the red curve shows the simple (non-skew) version of the INLA correction, while the green curve shows the INLA correction accounting for skewness (as discussed in Appendix A). The bottom panel of Figure 3 shows the additive correction to log posterior density, as a function of the hyperparameter (log precision). We see that there is very little difference between the two corrections.

For the toenail data, the estimated random effects standard deviation is approximately \( \sigma = 4 \), which is very high. To investigate how the copula correction works as \( \sigma \) increases, we studied simulated data sets from the model above, where we set \( \sigma \) to different values, and where the \( \alpha \) parameters were fixed to the values from a long MCMC run using the real toenail data (i.e., we simulate only the outcome, keeping the covariates unchanged). Results are shown in Figure 4 for different values of \( \sigma \) ranging from \( \sigma = 1 \) to \( \sigma = 16 \). We clearly get very accurate corrected posteriors for \( \sigma < 4 \). For \( \sigma \geq 4 \), we gradually get a tendency of under-correction.

### 3.3 Simulated example with Poisson likelihood

As our final example, we study the case where the data are Poisson distributed. We consider a simple simulated (extreme) example in order to investigate how well the correction works in the Poisson case. For \( i = 1, \ldots, n \) we generated iid \( y_i \sim \text{Poisson}(\mu_i) \) where

\[
\log(\mu_i) = \beta + u_i,
\]

with \( u_i \sim N(0, \sigma^2) \). We chose \( n = 300 \) and \( \sigma^2 = 1 \), and a Gamma(1, 1) prior for the precision \( \sigma^{-1} \). Figure 5 shows the results for different values of the intercept \( \beta \). Each histogram is based on ten parallel MCMC runs using JAGS, each with 200,000 iterations after a burn-in of 100,000. Here, reduction of \( \beta \) implies that estimation is more difficult, since negative \( \beta \) with large absolute value will imply that the counts \( y_i \) are very low, with many zeroes, and the data are uninformative. We see that uncorrected INLA tends to get less accurate as \( \beta \) moves towards more extreme values, while the correction seems to work well.

### 4 Discussion

The binary (and, more generally, binomial) GLMMs discussed in Sections 3.1 and 3.2 are are important in many applications, particularly for biomedical data. Poisson GLMMs are also of
Figure 3: The top panel shows the posterior density of the hyperparameter (log precision) for the toenail data. The histogram is from one million MCMC samples from JAGS, the blue curve is from uncorrected INLA, the red curve from INLA with the "mean only" correction, and the green curve from INLA with the "mean and skewness" correction. The bottom panel shows the additive corrections to the log posterior for the toenail data for the "mean only" and the "mean and skewness" correction.
Figure 4: Posteriors densities of log precision from the toenail simulation experiment for different values of the random effects standard deviation $\sigma$ (the value of $\sigma$ is shown above each histogram). The histograms are from long MCMC runs, uncorrected INLA are shown as black curves, while the red curves shows INLA with the correction. Since the goal here is to study the difference between MCMC and INLA, we omit axes – the relevant scale is given by the MCMC variances, which are evident from the histograms.
Figure 5: Posteriors densities of log precision from the Poisson simulation experiment for different values of the intercept $\beta$ (the value of $\beta$ is shown above each histogram). The histograms are from long MCMC runs, uncorrected INLA posteriors are shown as black curves, while the red curves shows INLA posteriors with the correction.
Letting $\tilde{x}_i - \tilde{\mu}_i(\theta)$ denote the "standardized" skew normal CDF corresponding to $\tilde{\pi}_{SN}(x_i|\theta, y)$. We start by finding the Jacobian of the transformation defined in Equation (4). Note first that, immediately from (4) (setting $\tilde{\sigma}_i(\theta) = \sigma_i(\theta)$),

$$\tilde{F}_i \left( \frac{x_i - \mu_i(\theta)}{\sigma_i(\theta)} \right) = \Phi \left( \frac{x_i - \mu_i(\theta)}{\sigma_i(\theta)} \right).$$

Letting $\tilde{f}_i$ and $\phi$ denote the density functions corresponding to $\tilde{F}_i$ and $\Phi$, differentiating with respect to $\tilde{x}_i$, then gives

$$\frac{1}{\sigma_i(\theta)} \phi \left( \frac{x_i - \mu_i(\theta)}{\sigma_i(\theta)} \right) \frac{\partial x_i}{\partial \tilde{x}_i} = \tilde{f}_i \left( \frac{x_i - \mu_i(\theta)}{\sigma_i(\theta)} \right) \frac{\partial x_i}{\partial \tilde{x}_i}$$

so

$$\frac{\partial x_i}{\partial \tilde{x}_i} = \frac{\tilde{f}_i \left( \frac{x_i - \mu_i(\theta)}{\sigma_i(\theta)} \right)}{\frac{1}{\sigma_i(\theta)} \phi \left( \frac{x_i - \mu_i(\theta)}{\sigma_i(\theta)} \right)} = \frac{\tilde{f}_i \left( \frac{x_i - \mu_i(\theta)}{\sigma_i(\theta)} \right)}{\phi \left( \frac{\tilde{F}_i \left( \frac{x_i - \mu_i(\theta)}{\sigma_i(\theta)} \right) - \tilde{\pi}_{SN}(x_i|\theta, y)}{\Phi(\tilde{F}_i \left( \frac{x_i - \mu_i(\theta)}{\sigma_i(\theta)} \right) - \tilde{\pi}_{SN}(x_i|\theta, y))} \right)}$$

and the Jacobian of the transformation is $\prod_{i=1}^{n} \frac{\partial x_i}{\partial \tilde{x}_i}$ since $\frac{\partial x_i}{\partial \tilde{x}_i} = 0$ for $i \neq j$ and $\frac{\partial x_i}{\partial \tilde{x}_i} \geq 0$ for all $i$.

Note that $(x - \mu(\theta))'Q(\theta)(x - \mu(\theta)) = \sum_{i=1}^{n} \sum_{j=1}^{n} Q_{ij}(\theta)(x_i - \mu_i(\theta))(x_j - \mu_j(\theta))$ where $Q(\theta) = (Q_{ij}(\theta))$. Collecting the different terms and again substituting

$$\frac{x_i - \mu_i(\theta)}{\sigma_i(\theta)} = \Phi^{-1} \left[ \tilde{F}_i \left( \frac{\tilde{x}_i - \tilde{\mu}_i(\theta)}{\sigma_i(\theta)} \right) \right],$$

the transformed joint log density $\tilde{\pi}(\tilde{x}|\theta, y)$ is therefore

$$\log \tilde{\pi}(\tilde{x}|\theta, y) = \frac{1}{2} \log |Q(\theta)| - \sum_{i=1}^{n} \sum_{j=1}^{n} Q_{ij}(\theta)\sigma_i(\theta)\sigma_j(\theta) \Phi^{-1} \left[ \tilde{F}_i \left( \frac{\tilde{x}_i - \tilde{\mu}_i(\theta)}{\sigma_i(\theta)} \right) \right] \Phi^{-1} \left[ \tilde{F}_j \left( \frac{\tilde{x}_j - \tilde{\mu}_j(\theta)}{\sigma_j(\theta)} \right) \right]$$

$$+ \sum_{i=1}^{n} \log \tilde{f}_i \left( \frac{\tilde{x}_i - \tilde{\mu}_i(\theta)}{\sigma_i(\theta)} \right) - \sum_{i=1}^{n} \log \phi \left( \Phi^{-1} \left[ \tilde{F}_i \left( \frac{\tilde{x}_i - \tilde{\mu}_i(\theta)}{\sigma_i(\theta)} \right) \right] \right) + \text{constant}$$
The original (uncorrected) log posterior is
\[
\log \tilde{\pi}(\theta|y) = \log \pi(x, \theta, y) - \log \tilde{\pi}_F(x|\theta, y) + \text{constant},
\]
evaluated at \(x = \mu(\theta)\) where \(\log \tilde{\pi}_F(x|\theta, y)\big|_{x=\mu(\theta)} = \frac{1}{2} \log |Q(\theta)| + \text{constant}\). Therefore, the version of the copula correction accounting for skewness amounts to adding a term \(C_{\text{skew}}(\theta)\) to the original log joint posterior, where

\[
C_{\text{skew}}(\theta) = \frac{1}{2} \sum_{i=1}^{n} \sum_{j=1}^{n} Q_{ij}(\theta) \sigma_i(\theta) \sigma_j(\theta) \Phi^{-1} \left( \tilde{F}_i \left( \frac{\mu_i(\theta) - \tilde{\mu}_i(\theta)}{\sigma_i(\theta)} \right) \right) \Phi^{-1} \left( \tilde{F}_j \left( \frac{\mu_j(\theta) - \tilde{\mu}_j(\theta)}{\sigma_j(\theta)} \right) \right) + \sum_{i=1}^{n} \log \tilde{f}_i \left( \frac{\mu_i(\theta) - \tilde{\mu}_i(\theta)}{\sigma_i(\theta)} \right) - \sum_{i=1}^{n} \log \phi \left( \Phi^{-1} \left( \tilde{F}_i \left( \frac{\mu_i(\theta) - \tilde{\mu}_i(\theta)}{\sigma_i(\theta)} \right) \right) \right).
\]

Acknowledgments

We thank Youyi Fong for providing R code relating to the simulation study described in Section 3.1. We are also very grateful to Leonard Held and Rafael Sauter for providing us with a copy of their unpublished paper along with R code relevant for the analysis of toenail data described in Section 3.2. We thank Janine Illian, Geir-Arne Fuglstad and Dan Simpson for helpful comments. This research was supported by the Norwegian Research Council.

References

Azzalini, A & Capitanio, A (1999), ‘Statistical applications of the multivariate skew normal distribution,’ Journal of the Royal Statistical Society: Series B (Statistical Methodology), 61(3), pp. 579–602, doi:10.1111/1467-9868.00194.

Breslow, NE & Clayton, DG (1993), ‘Approximate inference in generalized linear mixed models,’ Journal of the American Statistical Association, 88(421), pp. 9–25, doi:10.1080/01621459.1993.10594284.

Capanu, M, Gönen, M & Begg, CB (2013), ‘An assessment of estimation methods for generalized linear mixed models with binary outcomes,’ Statistics in Medicine, 32(26), pp. 4550–4566, doi:10.1002/sim.5866.

Cseke, B & Heskes, T (2010), ‘Improving posterior marginal approximations in latent Gaussian models,’ in International conference on Artificial Intelligence and Statistics, pp. 121–128.

De Backer, M, De Vroey, C, Lesaffre, E, Scheys, I & De Keyser, P (1998), ‘Twelve weeks of continuous oral therapy for toenail onychomycosis caused by dermatophytes: a double-blind comparative trial of terbinafine 250 mg/day versus itraconazole 200 mg/day,’ Journal of the American Academy of Dermatology, 38(5), pp. S57–S63, doi:10.1016/S0190-9622(98)70486-4.

Evangelou, E, Zhu, Z & Smith, RL (2011), ‘Estimation and prediction for spatial generalized linear mixed models using high order Laplace approximation,’ Journal of Statistical Planning and Inference, 141(11), pp. 3564–3577, doi:10.1016/j.jspi.2011.05.008.

Fong, Y, Rue, H & Wakefield, J (2010), ‘Bayesian inference for generalized linear mixed models,’ Biostatistics, 11(3), pp. 397–412, doi:10.1093/biostatistics/kxp053.
Grilli, L, Metelli, S & Rampichini, C (2014), ‘Bayesian estimation with integrated nested Laplace approximation for binary logit mixed models,’ Journal of Statistical Computation and Simulation, doi:10.1080/00949655.2014.935377.

Illian, JB, Sørbye, SH & Rue, H (2012), ‘A toolbox for fitting complex spatial point process models using integrated nested Laplace approximation (INLA),’ The Annals of Applied Statistics, 6(4), pp. 1499–1530, doi:10.1214/11-AOAS530.

Kauermann, G, Krivobokova, T & Fahrmeir, L (2009), ‘Some asymptotic results on generalized penalized spline smoothing,’ Journal of the Royal Statistical Society. Series B (Statistical Methodology), 71(2), pp. 487–503, doi:10.1111/j.1467-9868.2008.00691.x.

Lesaffre, E & Spiessens, B (2001), ‘On the effect of the number of quadrature points in a logistic random effects model: an example,’ Journal of the Royal Statistical Society: Series C (Applied Statistics), 50(3), pp. 325–335, doi:10.1111/1467-9876.00237.

Martins, TG, Simpson, D, Lindgren, F & Rue, H (2013), ‘Bayesian computing with INLA: New features,’ Computational Statistics & Data Analysis, 67, pp. 68 – 83, doi:10.1016/j.csda.2013.04.014.

McCulloch, CE, Searle, SR & Neuhaus, JM (2008), Generalized, Linear, and Mixed Models, John Wiley and sons, New York, 2nd edn.

Nelsen, RB (2007), An Introduction to Copulas, Springer Science & Business Media, New York, 2nd edn.

Ogden, HE (2015), ‘A sequential reduction method for inference in generalized linear mixed models,’ Electronic Journal of Statistics, 9, pp. 135–152, doi:10.1214/15-EJS991.

Plummer, M (2013), ‘JAGS version 3.4.0,’ http://mcmc-jags.sourceforge.net.

Raudenbush, SW, Yang, ML & Yosef, M (2000), ‘Maximum likelihood for generalized linear models with nested random effects via high-order, multivariate Laplace approximation,’ Journal of Computational and Graphical Statistics, 9(1), pp. 141–157, doi:10.1080/10618600.2000.10474870.

Rue, H, Martino, S & 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), pp. 319–392, doi:10.1111/j.1467-9868.2008.00700.x.

Ruli, E, Sartori, N & Ventura, L (2015), ‘Improved Laplace approximation for marginal likelihoods,’ arXiv preprint arXiv:1502.06440.

Sauter, R & Held, L (2015), ‘Quasi-complete separation in random effects of binary response mixed models: Integrated nested Laplace approximations vs. MCMC,’ Unpublished technical report.

Shun, Z & McCullagh, P (1995), ‘Laplace approximation of high dimensional integrals,’ Journal of the Royal Statistical Society. Series B (Methodological), pp. 749–760.

Simpson, D, Illian, J, Lindgren, F, Sørbye, S & Rue, H (2013), ‘Going off grid: Computationally efficient inference for log-Gaussian Cox processes,’ arXiv preprint arXiv:1111.0641.
Spiegelhalter, DJ, Best, NG, Carlin, BP & Van Der Linde, A (2002), ‘Bayesian measures of model complexity and fit,’ *Journal of the Royal Statistical Society: Series B (Statistical Methodology),* 64(4), pp. 583–639, doi:10.1111/1467-9868.00353.