Biostatistics (2007), **8**, 1, *pp.* 1–27
doi:10.1093/biostatistics/kxl000
Advance Access publication on May 15, 2006

Bayesian Inference for Generalized Linear Mixed Models

YOUYI FONG
Department of Biostatistics, University of Washington, Seattle, WA 98112, USA
HÅVARD RUE
The Norwegian University for Science and Technology, Trondheim, Norway
JON WAKEFIELD*
Departments of Statistics and Biostatistics, University of Washington, Seattle, WA 98112, USA
jonno@u.washington.edu

SUMMARY

Generalized linear mixed models (GLMMs) continue to grow in popularity due to their ability to directly acknowledge multiple levels of dependency, and model different data types. For small sample sizes especially, likelihood-based inference can be unreliable with variance components being particularly difficult to estimate. A Bayesian approach is appealing, but has been hampered by the lack of a fast implementation, and the difficulty in specifying prior distributions with variance components again being particularly problematic. Here we briefly review previous approaches to computation in Bayesian implementations of GLMMs, and illustrate in detail the use of integrated nested Laplace approximations in this context. We consider a number of examples, carefully specifying prior distributions on meaningful quantities in each case. The examples cover a wide range of data types including those requiring smoothing over time, and a relatively complicated spline model for which we examine our prior specification in terms of the

*To whom correspondence should be addressed.

implied degrees of freedom. We conclude that Bayesian inference is now practically feasible for GLMMs, and provides an attractive alternative to likelihood-based approaches such as penalized quasi-likelihood. As with likelihood-based approaches, great care is required in the analysis of clustered binary data, since approximation strategies may be less accurate for such data.

Keywords: Integrated nested Laplace approximations; Longitudinal data; Penalized quasi-likelihood; Prior specification; Spline models

1. INTRODUCTION

Generalized linear mixed models (GLMMs) combine a generalized linear model with normal random effects on the linear predictor scale, to give a rich family of models that have been used in a wide variety of applications, see for example Diggle et al. (2002), Verbeke and Molenberghs (2000, 2005), and McCulloch et al. (2008). This flexibility comes at a price, however, in terms of analytical tractability, which has a number of implications including computational complexity, and an unknown degree to which inference is dependent on modeling assumptions. Likelihood-based inference may be carried out relatively easily within many software platforms (except perhaps for binary responses), but inference is dependent on asymptotic sampling distributions of estimators, with few guidelines available as to when such theory will produce accurate inference. A Bayesian approach is attractive, but requires the specification of prior distributions which is not straightforward, in particular for variance components. Computation is also an issue since the usual implementation is via Markov chain Monte Carlo (MCMC), which carries a large computational overhead. The seminal article of Breslow and Clayton (1993) helped to popularize GLMMs and placed an emphasis on likelihood-based inference via penalized quasi-likelihood (PQL). It is the aim of this article to describe, through a series of examples (including all of those considered in Breslow and

Clayton, 1993), how Bayesian inference may be performed with computation via a fast implementation, and with guidance on prior specification.

The structure of this article is as follows. In Section 2 we define notation for the GLMM, and in Section 3 describe the integrated nested Laplace approximation (INLA) which has recently been proposed as a computationally convenient alternative to MCMC. Section 4 gives a number of prescriptions for prior specification. Three examples are considered in Section 5 (with additional examples being reported in the supplementary material, along with a simulation study that reports the performance of INLA in the binary response situation). We conclude the paper with a discussion in Section 6.

2. THE GENERALIZED LINEAR MIXED MODEL

GLMMs extend the generalized linear model, as proposed by Nelder and Wedderburn (1972) and comprehensively described in McCullagh and Nelder (1989), by adding normally distributed random effects on the linear predictor scale. Suppose Y_{ij} is of exponential family form: $Y_{ij}|\theta_{ij}, \phi_1 \sim p(\cdot)$ where $p(\cdot)$ is a member of the exponential family, that is

$$p(y_{ij}|\theta_{ij}, \phi_1) = \exp\left[\frac{y_{ij}\theta_{ij} - b(\theta_{ij})}{a(\phi_1)} + c(y_{ij}, \phi_1)\right],$$

for $i = 1, ..., m$ units (clusters), and $j = 1, ..., n_i$, measurements per unit, and where θ_{ij} is the (scalar) canonical parameter. Let $\mu_{ij} = \mathrm{E}[Y_{ij}|\boldsymbol{\beta}, \boldsymbol{b}_i, \phi_1] = b'(\theta_{ij})$ with

$$g(\mu_{ij}) = \eta_{ij} = \boldsymbol{x}_{ij}\boldsymbol{\beta} + \boldsymbol{z}_{ij}\boldsymbol{b}_i,$$

where $g(\cdot)$ is a monotonic *link* function, $\boldsymbol{x}_{ij}$ is $1 \times p$ and $\boldsymbol{z}_{ij}$ is $1 \times q$, with $\boldsymbol{\beta}$ a $p \times 1$ vector of fixed effects and $\boldsymbol{b}_i$ a $q \times 1$ vector of random effects, hence $\theta_{ij} = \theta_{ij}(\boldsymbol{\beta}, \boldsymbol{b}_i)$. Assume $\boldsymbol{b}_i|\boldsymbol{Q} \sim N(\boldsymbol{0}, \boldsymbol{Q}^{-1})$, where the precision matrix $\boldsymbol{Q} = \boldsymbol{Q}(\boldsymbol{\phi}_2)$ depends on parameters $\boldsymbol{\phi}_2$. For some choices of model the matrix $\boldsymbol{Q}$ is

singular; examples include random walk models (as considered in Section 5.2) and intrinsic conditional autoregressive (ICAR) models. We further assume that $\boldsymbol{\beta}$ is assigned a normal prior distribution. Let $\boldsymbol{\gamma} = (\boldsymbol{\beta}, \boldsymbol{b})$ denote the $G \times 1$ vector of parameters assigned Gaussian priors. We also require priors for ϕ_1 (if not a constant) and for $\boldsymbol{\phi}_2$. Let $\boldsymbol{\phi} = (\phi_1, \boldsymbol{\phi}_2)$ be the variance components for which non-Gaussian priors are assigned, with $V = \dim(\boldsymbol{\phi})$.

3. INTEGRATED NESTED LAPLACE APPROXIMATION

Before the MCMC revolution, there were few examples of the applications of Bayesian GLMMs since, outside of the linear mixed model, the models are analytically intractable. Kass and Steffey (1989) describe the use of Laplace approximations in Bayesian hierarchical models, while Skene and Wakefield (1990) used numerical integration in the context of a binary GLMM. The use of MCMC for GLMMs is particularly appealing since the conditional independencies of the model may be exploited when the required conditional distributions are calculated. Zeger and Karim (1991) described approximate Gibbs sampling for GLMMs, with non-standard conditional distributions being approximated by normal distributions. More general Metropolis-Hastings algorithms are straightforward to construct, see for example Clayton (1996) and Gamerman (1997). The `WinBUGs` (Spiegelhalter et al., 1998) software example manuals contain many GLMM examples. There are now a variety of additional software platforms for fitting GLMMs via MCMC including `JAGS` (Plummer, 2009) and `BayesX` (Fahrmeir et al., 2004). A large practical impediment to data analysis using MCMC is the large computational burden. For this reason we now briefly review the INLA computational approach upon which we concentrate. The method combines Laplace approximations and numerical integration in a very efficient manner, see Rue et al. (2009) for a

more extensive treatment. For the GLMM described in Section 2, the posterior is given by

$$\begin{aligned}\pi(\boldsymbol{\gamma},\boldsymbol{\phi}|\boldsymbol{y}) &\propto \pi(\boldsymbol{\gamma}|\boldsymbol{\phi})\pi(\boldsymbol{\phi})\prod_{i=1}^{m} p(\boldsymbol{y}_i \mid \boldsymbol{\gamma},\boldsymbol{\phi}) \\ &\propto \pi(\boldsymbol{\phi})\pi(\boldsymbol{\beta}) \mid \boldsymbol{Q}(\boldsymbol{\phi}_2) \mid^{1/2} \exp\left\{-\frac{1}{2}\boldsymbol{b}^T\boldsymbol{Q}(\boldsymbol{\phi}_2)\boldsymbol{b} + \sum_{i=1}^{m}\log p(\boldsymbol{y}_i \mid \boldsymbol{\gamma},\phi_1)\right\}\end{aligned}$$

where $\boldsymbol{y}_i = (y_{i1}, ..., y_{in_i})$ is the vector of observations on unit/cluster i. We wish to obtain the posterior marginals $\pi(\gamma_g \mid \boldsymbol{y})$, $g = 1, ..., G$, and $\pi(\phi_v|\boldsymbol{y})$, $v = 1, ..., V$. The number of variance components, V, should not be too large for accurate inference (since these components are integrated out via Cartesian product numerical integration, which does not scale well with dimension). We write

$$\pi(\gamma_g|\boldsymbol{y}) = \int \pi(\gamma_g|\boldsymbol{\phi},\boldsymbol{y}) \times \pi(\boldsymbol{\phi}|\boldsymbol{y})d\boldsymbol{\phi}$$

which may be evaluated via the approximation

$$\begin{aligned}\widetilde{\pi}(\gamma_g|\boldsymbol{y}) &= \int \widetilde{\pi}(\gamma_g|\boldsymbol{\phi},\boldsymbol{y}) \times \widetilde{\pi}(\boldsymbol{\phi}|\boldsymbol{y})d\boldsymbol{\phi} \\ &\approx \sum_{k=1}^{K} \widetilde{\pi}(\gamma_g \mid \boldsymbol{\phi}^k,\boldsymbol{y}) \times \widetilde{\pi}(\boldsymbol{\phi}^k \mid \boldsymbol{y}) \times \Delta_k \qquad (3.1)\end{aligned}$$

where Laplace (or other related analytical approximations) are applied to carry out the integrations required for evaluation of $\widetilde{\pi}(\gamma_g|\boldsymbol{\phi},\boldsymbol{y})$. To produce the grid of points $\{\boldsymbol{\phi}^k, k = 1, ..., K\}$ over which numerical integration is performed, the mode of $\widetilde{\pi}(\boldsymbol{\phi}|\boldsymbol{y})$ is located, and the Hessian is approximated, from which the grid is created and exploited in (3.1). The output of INLA consists of posterior marginal distributions, which can be summarized via means, variances and quantiles. Importantly for model comparison, the normalizing constant $p(\boldsymbol{y})$ is calculated. The evaluation of this quantity is not straightforward using MCMC (DiCiccio et al., 1997; Meng and Wong, 1996). The deviance information criterion (Spiegelhalter et al., 1998) is popular as a model selection tool, but in random effects models the implicit approximation in its

use is valid only when the effective number of parameters is much smaller than the number of independent observations, see Plummer (2008).

4. PRIOR DISTRIBUTIONS

4.1 *Fixed Effects*

Recall that we assume $\boldsymbol{\beta}$ is normally distributed. Often there will be sufficient information in the data for $\boldsymbol{\beta}$ to be well estimated with a normal prior with a large variance (of course there will be circumstances under which we would like to specify more informative priors, for example, when there are many correlated covariates). The use of an improper prior for $\boldsymbol{\beta}$ will often lead to a proper posterior though care should be taken. For example, Wakefield (2007) shows that a Poisson likelihood with a linear link can lead to an improper posterior if an improper prior is used. Hobert and Casella (1996) discuss the use of improper priors in linear mixed effects models.

If we wish to use informative priors we may specify independent normal priors with the parameters for each component being obtained via specification of two quantiles with associated probabilities. For logistic and log-linear models these quantiles may be given on the exponentiated scale since these are more interpretable (as the odds ratio and rate ratio, respectively). If θ_1, θ_2 are the quantiles on the exponentiated scale, and p_1, p_2 are the associated probabilities, then the parameters of the normal prior are given by:

$$\mu = \frac{z_2 \log(\theta_1) - z_1 \log(\theta_2)}{z_2 - z_1}$$
$$\sigma = \frac{\log(\theta_2) - \log(\theta_1)}{z_2 - z_1}$$

where z_1, z_2 are the p_1, p_2 quantiles of a standard normal random variable. For example, in an epidemiological context we may wish to specify a prior on a relative risk parameter, $\exp(\beta_1)$, which has a median

of 1 and a 95% point of 3 (if we think it is unlikely that the relative risk associated with a unit increase in exposure exceeds 3). These specifications lead to $\beta_1 \sim N(0, 0.668^2)$.

4.2 *Variance Components*

We begin by describing an approach for choosing a prior for a single random effect, based on Wakefield (2009). The basic idea is to specify a range for the more interpretable marginal distribution of b_i, and use this to drive specification of prior parameters. We state a trivial lemma upon which prior specification is based, but first define some notation. We write $\tau \sim \text{Ga}(a_1, a_2)$ for the gamma distribution with unnormalized density $\tau^{a_1-1}\exp(-a_2\tau)$. For q-dimensional $\boldsymbol{x}$ we write $\boldsymbol{x} \sim \text{T}_q(\boldsymbol{\mu}, \boldsymbol{\Omega}, d)$ for the Student's t distribution with unnormalized density $[1 + (\boldsymbol{x}-\boldsymbol{\mu})^T\boldsymbol{\Omega}^{-1}(\boldsymbol{x}-\boldsymbol{\mu})/d]^{-(d+q)/2}$. This distribution has location $\boldsymbol{\mu}$, scale matrix $\boldsymbol{\Omega}$ and degrees of freedom d.

Lemma 1: Let $b|\tau \sim N(0, \tau^{-1})$ and $\tau \sim \text{Ga}(a_1, a_2)$. Integration over τ gives the marginal distribution of b as $\text{T}_1(0, a_2/a_1, 2a_1)$.

To decide upon a prior we give a range for a generic random effect b, and specify the degrees of freedom, d, and then solve for a_1, a_2. For the range $(-R, R)$ we use the relationship $\pm t^d_{1-(1-q)/2}\sqrt{a_2/a_1} = \pm R$, where t^d_q is the $100 \times q$-th quantile of a Student t random variable with d degrees of freedom, to give $a_1 = d/2$, $a_2 = R^2 d/2(t^d_{1-(1-q)/2})^2$. In the linear mixed effects model, b is directly interpretable, while for binomial or Poisson models it is more appropriate to think in terms of the marginal distribution of $\exp(b)$, the residual odds and rate ratio, respectively, and this distribution is log Student's t. For example, if we choose $d = 1$ (to give a Cauchy marginal), and a 95% range of $[0.1, 10]$ we take $R = \log 10$ and obtain $a = 0.5, b = 0.0164$.

Another convenient choice is $d = 2$ to give the exponential distribution $\text{Exp}(a_2)$ (where the mean is

a_2^{-1}) for σ^{-2}. This leads to closed form expressions for the more interpretable quantiles of σ so that, for example, if we specify the median for σ as σ_m we obtain $a_2 = \sigma_m^2 \log 2$.

Unfortunately, the use of Ga(ϵ, ϵ) priors has become popular as a prior for σ^{-2} in a GLMM context, arising from their use in the `WinBUGS` examples manual. As has been pointed out many times (e.g. Kelsall and Wakefield, 1999, Gelman 2006, Crainiceanu et al. 2008) this choice places the majority of the prior mass away from zero, and leads to a marginal prior for the random effects which is Student's t with 2ϵ degrees of freedom (so that the tails are much heavier than even a Cauchy), and difficult to justify in any practical setting.

We now specify another trivial lemma, but first establish notation for the Wishart distribution. For the $q \times q$ non-singular matrix $\boldsymbol{z}$ we write $\boldsymbol{z} \sim \text{Wishart}_q(r, \boldsymbol{S})$ for the Wishart distribution with unnormalized density $|\boldsymbol{z}|^{(r-q-1)/2} \exp\left\{-\frac{1}{2}\text{tr}(\boldsymbol{z}\boldsymbol{S}^{-1})\right\}$. This distribution has $\text{E}[\boldsymbol{z}] = r\boldsymbol{S}$ and $\text{E}[\boldsymbol{z}^{-1}] = \boldsymbol{S}^{-1}/(r-q-1)$, and we require $r > q-1$ for a proper distribution.

Lemma: Let $\boldsymbol{b} = (b_1, ..., b_q)$, with $\boldsymbol{b}|\boldsymbol{Q} \sim_{iid} N_q(\boldsymbol{0}, \boldsymbol{Q}^{-1})$, $\boldsymbol{Q} \sim \text{Wishart}_q(r, \boldsymbol{S})$. Integration over $\boldsymbol{Q}$ gives the marginal distribution of $\boldsymbol{b}$ as $\text{T}_q(\boldsymbol{0}, [(r-q+1)\boldsymbol{S}]^{-1}, r-q+1)$.

The margins of a multivariate Student's t are t also, which allows r and $\boldsymbol{S}$ to be chosen as in the univariate case. Specifically, the k-th element of a generic random effect, b_k, follows a univariate Student t distribution with location 0, scale $S^{kk}/(r-q+1)$, and degrees of freedom $d = r-q+1$ where S^{kk} is element (k,k) of the inverse of $\boldsymbol{S}$. We obtain $r = d+q-1$ and $S^{kk} = (t^d_{1-(1-q)/2})^2/(dR^2)$. If *a priori* we have no reason to believe that elements of $\boldsymbol{b}$ are correlated we may specify $S_{jk} = 0$ for $j \neq k$ and $S^{kk} = 1/S_{kk}$, to recover the univariate specification, recognizing that with $q = 1$ the univariate Wishart has parameters $a_1 = r/2, a_2 = 1/(2S)$. If we believe that elements of $\boldsymbol{b}$ are dependent then we may specify the correlations and solve for the off-diagonal elements of $\boldsymbol{S}$. To ensure propriety of the posterior,

proper priors are required for $\mathbf{\Sigma}$; Zeger and Karim (1991) use an improper prior for $\mathbf{\Sigma}$, so that the posterior is improper also.

4.3 *Effective Degrees of Freedom Variance Components Prior*

In Section 5.3 we describe the GLMM representation of a spline model. A generic linear spline model is given by

$$y_i = \boldsymbol{x}_i\boldsymbol{\beta} + \sum_{k=1}^{K} z_{ik}b_k + \epsilon_i$$

where $\boldsymbol{x}_i$ is a $p \times 1$ vector of covariates with $p \times 1$ associated fixed effects $\boldsymbol{\beta}$, z_{ik} denote the spline basis, $b_k \sim_{iid} N(0, \sigma_b^2)$ and $\epsilon_i \sim_{iid} N(0, \sigma_\epsilon^2)$, with b_k and ϵ_i independent. Specification of a prior for σ_b^2 is not straightforward, but may be of great importance since it contributes to determining the amount of smoothing that is applied. Ruppert et al. (2003, p. 177) raise concerns, "about the instability of automatic smoothing parameter selection even for single predictor models", and continue, "Although we are attracted by the automatic nature of the mixed model-REML approach to fitting additive models, we discourage blind acceptance of whatever answer it provides and recommend looking at other amounts of smoothing". While we would echo this general advice, we believe that a Bayesian mixed model approach, with carefully chosen priors, can increase the stability of the mixed model representation. There has been some discussion of choice of prior for σ_b^2 in a spline context (Crainiceanu et al., 2005, 2008). More general discussion can be found in Natarajan and Kass (2000) and Gelman (2006).

In practice (e.g. Hastie and Tibshirani, 1990) smoothers are often applied with a fixed degrees of freedom. We extend this rationale by examining the prior degrees of freedom that is implied by the choice $\sigma_b^{-2} \sim \text{Ga}(a_1, a_2)$. For the general linear mixed model

$$\boldsymbol{y} = \boldsymbol{x}\boldsymbol{\beta} + \boldsymbol{z}\boldsymbol{b} + \boldsymbol{\epsilon}$$

we have

$$\widehat{\boldsymbol{y}} = \boldsymbol{x}\widehat{\boldsymbol{\beta}} + \boldsymbol{z}\widehat{\boldsymbol{b}} = \boldsymbol{C}(\boldsymbol{C}^T\boldsymbol{C} + \boldsymbol{\Lambda})^{-1}\boldsymbol{C}^T\boldsymbol{y}$$

where $\boldsymbol{C} = [\boldsymbol{x}|\boldsymbol{z}]$ is $n \times (p + K)$ and

$$\boldsymbol{\Lambda} = \begin{bmatrix} \mathbf{0}_{p\times p} & \mathbf{0}_{p\times K} \\ \mathbf{0}_{K\times p} & \sigma^2_\epsilon \text{cov}(\boldsymbol{b})^{-1} \end{bmatrix}$$

see, for example, Ruppert et al. (2003, Section 8.3). The total degrees of freedom associated with the model is

$$\text{df} = \text{tr}\{(\boldsymbol{C}^T\boldsymbol{C} + \boldsymbol{\Lambda})^{-1}\boldsymbol{C}^T\boldsymbol{C}\}$$

which may be decomposed into the degrees of freedom associated with $\boldsymbol{\beta}$ and $\boldsymbol{b}$, and extends easily to situations in which we have additional random effects, beyond those associated with the spline basis (such an example is considered in Section 5.3). In each of these situations the degrees of freedom associated with the respective parameter is obtained by summing the appropriate diagonal elements of $(\boldsymbol{C}^T\boldsymbol{C} + \boldsymbol{\Lambda})^{-1}\boldsymbol{C}^T\boldsymbol{C}$. Specifically, if we have $j = 1, ..., d$ sets of random effect parameters (there are $d = 2$ in the model considered in Section 5.3) then let $\boldsymbol{E}_j$ be the $(p + K) \times (p + K)$ diagonal matrix with ones in the diagonal positions corresponding to set j. Then the degrees of freedom associated with this set is $\text{df}_j = \text{tr}\{\boldsymbol{E}_j(\boldsymbol{C}^T\boldsymbol{C} + \boldsymbol{\Lambda})^{-1}\boldsymbol{C}^T\boldsymbol{C}$. Note that the effective degrees of freedom changes as a function of K, as expected. To evaluate $\boldsymbol{\Lambda}$, σ^2_ϵ is required. If we specify a proper prior for σ^2_ϵ then we may specify the joint prior as $\pi(\sigma^2_b, \sigma^2_\epsilon) = \pi(\sigma^2_\epsilon)\pi(\sigma^2_b|\sigma^2_\epsilon)$. Usually, however, we assume the improper prior $\pi(\sigma^2_\epsilon) \propto 1/\sigma^2_\epsilon$, since the data provide sufficient information with respect to σ^2_ϵ. Hence we have found the substitution of an estimate for σ^2_ϵ (for example, from the fitting of a spline model in a likelihood implementation) to be a practically reasonable strategy.

As a simple non-spline demonstration of the derived effective degrees of freedom, consider a one-way

ANOVA model

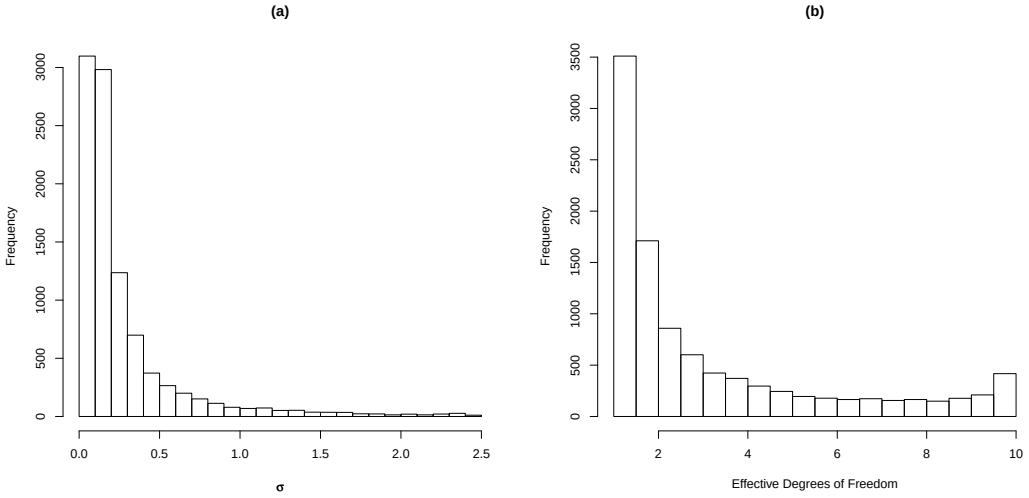
$$Y_{ij} = \beta_0 + b_i + \epsilon_{ij}$$

with $b_i \sim_{iid} N(0, \sigma_b^2)$, $\epsilon_{ij} \sim_{iid} N(0, \sigma_\epsilon^2)$ for $i = 1, ..., m = 10$ groups and $j = 1, ..., n = 5$ observations per group. For illustration we assume $\sigma_b^{-2} \sim \text{Ga}(0.5, 0.005)$. Figure 1 displays the prior distribution for σ, the implied prior distribution on the effective degrees of freedom, and the bivariate plot of these quantities. For clarity of plotting we exclude a small number of points beyond $\sigma > 2.5$ (4% of points). In panel (c) we have placed dashed horizontal lines at effective degrees of freedom equal to 1 (complete smoothing) and 10 (no smoothing). From panel (b) we conclude that here the prior choice favors quite strong smoothing. This may be contrasted with the Gamma prior with parameters $(0.001, 0.001)$, which, in this example, gives greater than 99% of the prior mass on an effective degrees of freedom greater than 9.9, again showing the inappropriateness of this prior.

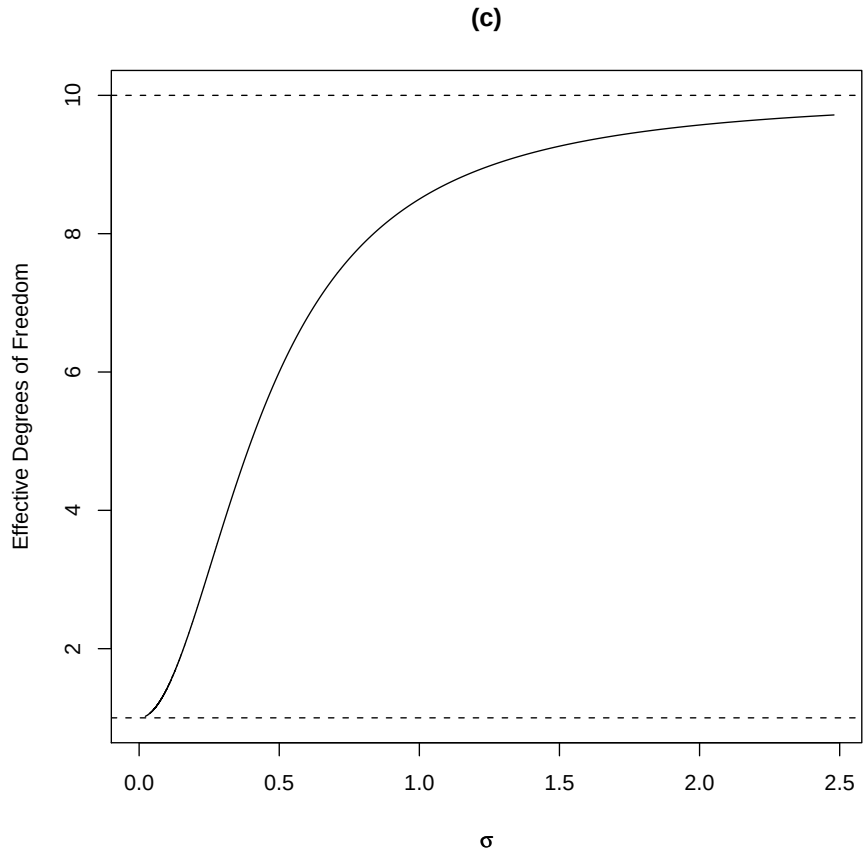
It is appealing to extend the above argument to non-linear models, but unfortunately this is not straightforward. For a non-linear model the degrees of freedom may be approximated by

$$\text{df} = \text{tr}\{(\boldsymbol{C}^T \boldsymbol{W} \boldsymbol{C} + \boldsymbol{\Lambda})^{-1} \boldsymbol{C}^T \boldsymbol{W} \boldsymbol{C}\}$$

where $\boldsymbol{W} = \text{diag}\left\{ V_i^{-1} \left(\frac{d\mu_i}{dh} \right)^2 \right\}$ and $h = g^{-1}$ denotes the inverse link function. Unfortunately this quantity depends on $\boldsymbol{\beta}$ and $\boldsymbol{b}$, which means that in practice we would have to use prior estimates for all of the parameters, which may not be practically possible. Fitting the model using likelihood and then substituting in estimates for $\boldsymbol{\beta}$ and $\boldsymbol{b}$ seems philosophically dubious.



(a) (b)



(c)

Fig. 1. Gamma prior for σ^{-2} with parameters 0.5 and 0.005, (a) implied prior for σ, (b) implied prior for the effective degrees of freedom, (c) effective degrees of freedom versus σ.

4.4 *Random Walk Models*

Conditionally represented smoothing models are popular for random effects in both temporal and spatial applications, see for example Besag et al. (1995) and Rue and Held (2005). For illustration, consider models of the form

$$p(\boldsymbol{u}|\sigma_u^2) = (2\pi)^{-(m-r)/2}|\boldsymbol{Q}^\star|^{1/2}\sigma_u^{-(m-r)} \exp\left(-\frac{1}{2\sigma_u^2}\boldsymbol{u}^T\boldsymbol{Q}\boldsymbol{u}\right) \tag{4.1}$$

where $\boldsymbol{u} = (u_1, ..., u_m)$ is the collection of random effects, $\boldsymbol{Q}$ is a (scaled) "precision" matrix of rank $m-r$, whose form is determined by the application at hand, and $|\boldsymbol{Q}^\star|$ is a generalized determinant which is the product over the $m-r$ non-zero eigenvalues of $\boldsymbol{Q}$. Picking a prior for σ_u is not straightforward because σ_u has an interpretation as the conditional standard deviation, where the elements that are conditioned upon depends on the application. We may simulate realizations from (4.1) to examine candidate prior distributions. Due to the rank deficiency, (4.1) does not define a probability density, and so we cannot directly simulate from this prior. However, Rue and Held (2005) give an algorithm for generating samples from (4.1):

1. Simulate $z_j \sim N(0, \lambda_j^{-1})$, for $j = m - r + 1, ..., m$, where λ_j are the eigenvalues of $\boldsymbol{Q}$ (there are $m - r$ non-zero eigenvalues as $\boldsymbol{Q}$ has rank $m - r$).

2. Return $\boldsymbol{u} = z_{m-r+1}\boldsymbol{e}_{n-r+1} + z_3\boldsymbol{e}_3 + ... + z_n\boldsymbol{e}_m = \boldsymbol{E}\boldsymbol{z}$ where $\boldsymbol{e}_j$ are the corresponding eigenvectors of $\boldsymbol{Q}$, $\boldsymbol{E}$ is the $m \times (m - r)$ matrix with these eigenvectors as columns, and $\boldsymbol{z}$ is the $(m - r) \times 1$ vector containing z_j, $j = m - r + 1, ..., m$.

The simulation algorithm is conditioned so that samples are zero in the null-space of $\boldsymbol{Q}$; if $\boldsymbol{u}$ is a sample and the null-space is spanned by $\boldsymbol{v}_1$ and $\boldsymbol{v}_2$, then $\boldsymbol{u}^T\boldsymbol{v}_1 = \boldsymbol{u}^T\boldsymbol{v}_2 = \boldsymbol{0}$. For example, suppose $\boldsymbol{Q}\boldsymbol{1} = \boldsymbol{0}$ so that the null space is spanned by $\boldsymbol{1}$, and the rank deficiency is 1. Then $\boldsymbol{Q}$ is improper since the eigenvalue

corresponding to $\mathbf{1}$ is zero, and samples $\boldsymbol{u}$ produced by the algorithm are such that $\boldsymbol{u}^T\mathbf{1} = \mathbf{0}$. In Section 5.2 we use this algorithm to evaluate different priors, via simulation. It is also useful to note that if we wish to compute the marginal variances only, simulation is not required, as they are available as the diagonal elements of the matrix $\sum_j \lambda_j^{-1}\boldsymbol{e}_j\boldsymbol{e}_j^T$.

5. EXAMPLES

Here we report three examples, with four others described in the supplementary material. Together these cover all of the examples in Breslow and Clayton (1993), along with an additional spline example. In the first example results using the INLA numerical/analytical approximation described in Section 3 were compared with MCMC as implemented in the `JAGS` software (Plummer, 2009) and, found to be accurate. For the models considered in the second and third examples the approximation was compared with the MCMC implementation contained in the INLA software.

5.1 *Longitudinal Data*

We consider the much analyzed epilepsy dataset of Thall and Vail (1990). These data concern the number of seizures, Y_{ij} for patient i on visit j, with $Y_{ij}|\boldsymbol{\beta}, \boldsymbol{b}_i \sim_{ind}$ Poisson(μ_{ij}), $i = 1, ..., 59$, $j = 1, ..., 4$. We concentrate on the three random effects models fitted by Breslow and Clayton (1993):

$$\log \mu_{ij} = \boldsymbol{x}_{ij}\boldsymbol{\beta} + b_{1i} \tag{5.1}$$

$$\log \mu_{ij} = \boldsymbol{x}_{ij}\boldsymbol{\beta} + b_{1i} + b_{0ij} \tag{5.2}$$

$$\log \mu_{ij} = \boldsymbol{x}_{ij}\boldsymbol{\beta} + b_{1i} + b_{2i}\mathrm{V}_j/10 \tag{5.3}$$

where $\boldsymbol{x}_{ij}$ is a 1×6 vector containing a 1 (representing the intercept), an indicator for baseline measurement, a treatment indicator, the baseline by treatment interaction, which is the parameter of interest, age, and either an indicator of the fourth visit (models (5.1) and (5.2) and denoted V_4), or visit number coded $-3, -1, +1, +3$ (model (5.3), and denoted $\mathrm{V}_j/10$); $\boldsymbol{\beta}$ is the associated fixed effect. All three models include patient-specific random effects $b_{1i} \sim N\left(0, \sigma_1^2\right)$, while in model (5.2) we introduce independent "measurement errors", $b_{0ij} \sim N(0, \sigma_0^2)$. Model (5.3) includes random effects on the slope associated with visit, b_{2i} with

$$\left[\begin{array}{c} b_{1i} \\ b_{2i} \end{array} \right] \sim N(\mathbf{0}, \boldsymbol{Q}^{-1}). \tag{5.4}$$

We assume $\boldsymbol{Q} \sim$ Wishart$(r, \boldsymbol{S})$ with $\boldsymbol{S} = \left[\begin{smallmatrix} S_{11} & S_{12} \\ S_{21} & S_{22} \end{smallmatrix} \right]$. For prior specification we begin with the bivariate model and assume that $\boldsymbol{S}$ is diagonal. We assume the upper 95% point of the priors for $\exp(b_{1i})$ and $\exp(b_{2i})$ are 5 and 4, respectively, and that the marginal distributions are t with 4 degrees of freedom. Following the procedure outlined in Section 4.2 we obtain $r = 5$ and $\boldsymbol{S} = \mathrm{diag}(0.439, 0.591)$. We take the prior for σ_1^{-2} in model (5.1) to be Ga(a_1, a_2) with $a_1 = (r-1)/2 = 2$ and $a_2 = 1/2S_{11} = 1.140$ (so that this prior coincides with the marginal prior obtained from the bivariate specification). In model (5.2) we assume b_{1i} and b_{0ij} are independent, and that σ_0^{-2} follows the same prior as σ_1^{-2}, i.e. Ga$(2, 1.140)$. We assume a flat prior on the intercept, and assume that the rate ratios, $\exp(\beta_j)$, $j = 1, ..., 5$, lie between 0.1 and 10 with probability 0.95 which gives, using the approach described in Section 4.1, a normal prior with mean 0 and variance 1.17^2.

Table 1 gives PQL and INLA summaries for models (5.1)–(5.3). There are some differences between the PQL and Bayesian analyses, with slightly larger standard deviations under the latter, which probably reflects that with $m = 59$ clusters a little accuracy is lost when using asymptotic inference. There are some differences in the point estimates which is at least partly due to the non-flat priors used — the priors

have relatively large variances, but here the data are not so abundant so there is sensitivity to the prior. Reassuringly under all three models inference for the baseline-treatment interaction of interest is virtually identical, and suggests no significant treatment effect. We may compare models using $\log p(\boldsymbol{y})$: for the three models we obtain values of -674.8, -638.9 and -665.5, so that the second model is strongly preferred.

	Model (5.1)		Model (5.2)		Model (5.3)	
Variable	PQL	INLA	PQL	INLA	PQL	INLA
Base	0.87 ± 0.14	0.88 ± 0.15	0.86 ± 0.13	0.88 ± 0.15	0.87 ± 0.14	0.88 ± 0.14
Trt	-0.91 ± 0.41	-0.94 ± 0.44	-0.93 ± 0.40	-0.96 ± 0.44	-0.91 ± 0.41	-0.94 ± 0.44
Base $\times$ Trt	0.33 ± 0.21	0.34 ± 0.22	0.34 ± 0.21	0.35 ± 0.23	0.33 ± 0.21	0.34 ± 0.22
Age	0.47 ± 0.36	0.47 ± 0.38	0.47 ± 0.35	0.48 ± 0.39	0.46 ± 0.36	0.47 ± 0.38
V4 or V/10	-0.16 ± 0.05	-0.16 ± 0.05	-0.10 ± 0.09	-0.10 ± 0.09	-0.26 ± 0.16	-0.27 ± 0.16
σ_0	—	—	0.36 ± 0.04	0.41 ± 0.04	—	—
σ_1	0.53 ± 0.06	0.56 ± 0.08	0.48 ± 0.06	0.53 ± 0.07	0.52 ± 0.06	0.56 ± 0.06
σ_2	—	—	—	—	0.74 ± 0.16	0.70 ± 0.14

Table 1. PQL and INLA summaries for the epilepsy data.

5.2 *Smoothing of Birth Cohort Effects in an Age-Cohort Model*

We analyze data from Breslow and Day (1975) on breast cancer rates in Iceland. Let Y_{jk} be the number of breast cancer of cases in age group j (20–24, ..., 80–84) and birth cohort k (1840–1849, ...,1940–1949) with $j = 1, ..., J = 13$ and $k = 1, ..., K = 11$. Following Breslow and Clayton (1993) we assume $Y_{jk}|\mu_{jk} \sim_{ind}$ Poisson(μ_{jk}) with

$$\log \mu_{jk} = \log n_{jk} + \beta_j + \beta k + v_k + u_k \tag{5.5}$$

and where n_{jk} is the person-years denominator, $\exp(\beta_j), j = 1, ..., J$, represent fixed effects for age relative risks, $\exp(\beta)$ is the relative risk associated with a one group increase in cohort group, $v_k \sim_{iid} N(0, \sigma_v^2)$ represent unstructured random effects associated with cohort k, with smooth cohort terms u_k following a second-order random effects model with $\mathrm{E}[u_k|\{u_i : i < k\}] = 2u_{k-1} - u_{k-2}$ and $\mathrm{Var}(u_k|\{u_i :$

$i < k\}) = \sigma_u^2$. This latter model is to allow the rates to vary smoothly with cohort. An equivalent representation of this model is, for $2 < k < K - 1$,

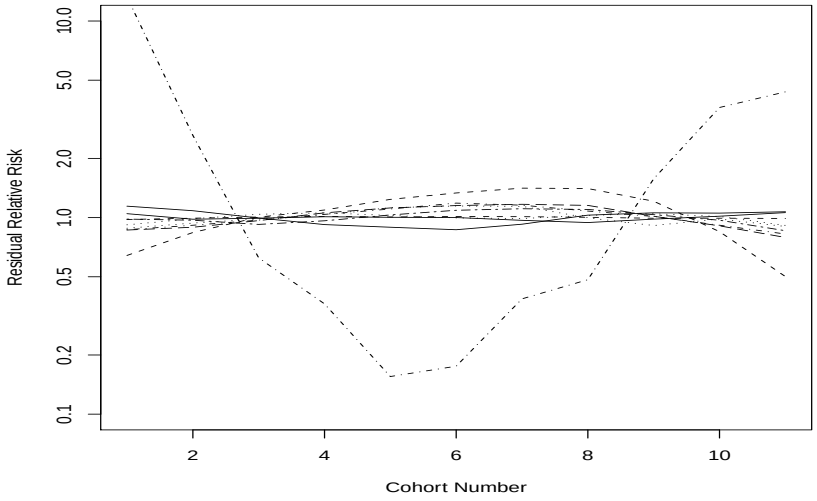
$$\mathrm{E}[u_k|\{u_l : l \neq k\}] = \frac{1}{6}(4u_{k-1} + 4u_{k+1} - u_{k-2} - u_{k+2})$$
$$\mathrm{Var}(u_k|\{u_l : l \neq k\}) = \frac{\sigma_u^2}{6}$$

The rank of $\boldsymbol{Q}$ in the (4.1) representation of this model is $K - 2$ reflecting that both the overall level and the overall trend are aliased (hence the appearance of β in (5.5)).

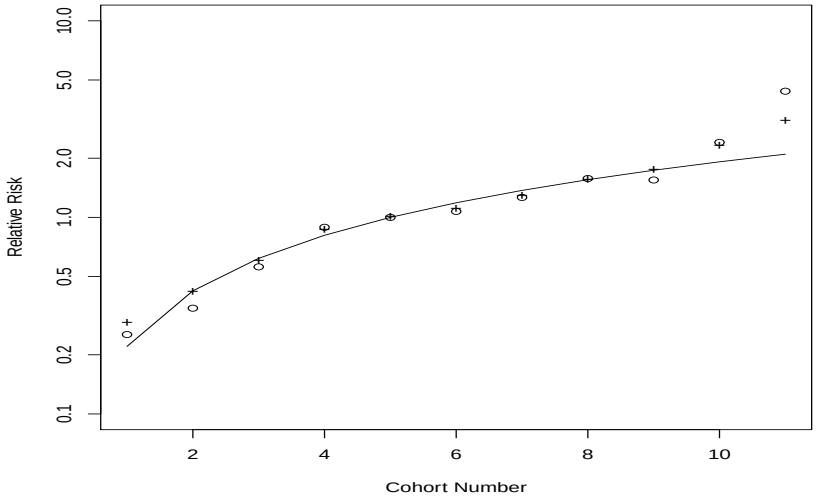
The term $\exp(v_k)$ reflects the unstructured residual relative risk and, following the argument in Section 4.2, we specify that this quantity should lie in [0.5,2.0] with probability 0.95, with a marginal log Cauchy distribution, to obtain the gamma prior $\sigma_v^{-2} \sim \mathrm{Ga}(0.5, 0.00149)$.

The term $\exp(u_k)$ reflects the smooth component of the residual relative risk, and the specification of a prior for the associated variance component σ_u^2 is more difficult, given its conditional interpretation. Using the algorithm described in Section 4.2 we examined simulations of $\boldsymbol{u}$ for different choices of gamma hyperparameters, and decided on the choice $\sigma_u^{-2} \sim \mathrm{Ga}(0.5, 0.001)$; Figure 2 shows 10 realizations from the prior. The rationale here is to examine realizations to see if they conform to our prior expectations and in particular exhibit the required amount of smoothing. All but one of the realizations vary smoothly across the 11 cohorts, as is desirable. Due to the tail of the gamma distribution, we will always have some extreme realizations.

The INLA results, summarized in graphical form, are presented in Figure 2(b), alongside likelihood fits in which the birth cohort effect is incorporated as a linear term and as a factor. We see that the smoothing model provides a smooth fit in birth cohort, as we would hope.



(a)



(b)

Fig. 2. (a) Ten realizations (on the relative risk scale) from the random effects second-order random walk model in which the prior on the random effects precision is Ga(0.5,0.001). (b) Summaries of fitted models: the solid line corresponds to a loglinear model in birth cohort, the circles to birth cohort as a factor, and "+" to the Bayesian smoothing model.

5.3 *B-Spline Nonparametric Regression*

We demonstrate the use of INLA for nonparametric smoothing using O'Sullivan splines, which are based on a B-spline basis. We illustrate using data from Bachrach et al. (1999) that concerns longitudinal measurements of spinal bone mineral density (SBMD) on 230 female subjects aged between 8 and 27, and of one of four ethnic groups: Asian, Black, Hispanic and White. Let y_{ij} denote the SBMD measure for subject i at occasion j, for $i = 1, ..., 230$ and $j = 1, ..., n_i$ with n_i being between 1 and 4. Figure 3 shows these data, with the grey lines indicating measurements on the same woman.

We assume the model

$$Y_{ij} = \boldsymbol{x}_i\boldsymbol{\beta}_1 + \text{age}_{ij}\beta_2 + \sum_{k=1}^{K} z_{ijk}b_{1k} + b_{2i} + \epsilon_{ij}$$

where $\boldsymbol{x}_i$ is a 1×4 vector containing an indicator for the ethnicity of individual i, with $\boldsymbol{\beta}_1$ the associated 4×1 vector of fixed effects, z_{ijk} is the k-th basis associated with age, with associated parameter $b_{1k} \sim N(0, \sigma_1^2)$, and $b_{2i} \sim N(0, \sigma_2^2)$ are woman-specific random effects, finally, $\epsilon_{ij} \sim_{iid} N(0, \sigma_\epsilon^2)$. All random terms are assumed independent. Note that the spline model is assumed common to all ethnic groups and all women, though it would be straightforward to allow a different spline for each ethnicity. Writing this model in the form

$$\boldsymbol{y} = \boldsymbol{x}\boldsymbol{\beta} + \boldsymbol{z}_1\boldsymbol{b}_1 + \boldsymbol{z}_2\boldsymbol{b}_2 + \boldsymbol{\epsilon} = \boldsymbol{C}\boldsymbol{\gamma} + \boldsymbol{\epsilon}$$

we use the method described in Section 4.3 to examine the effective number of parameters implied by different priors $\sigma_1^{-2} \sim \text{Ga}(a_1, a_2)$, $\sigma_2^{-2} \sim \text{Ga}(a_3, a_4)$.

To fit the model we first use the `R` code provided in Wand and Ormerod (2008) to construct the basis functions, which are then input to the INLA program. Running the REML version of the model we obtain $\widehat{\sigma}_\epsilon = 0.033$ which we use to evaluate the effective degrees of freedoms associated with priors for σ_1^2 and

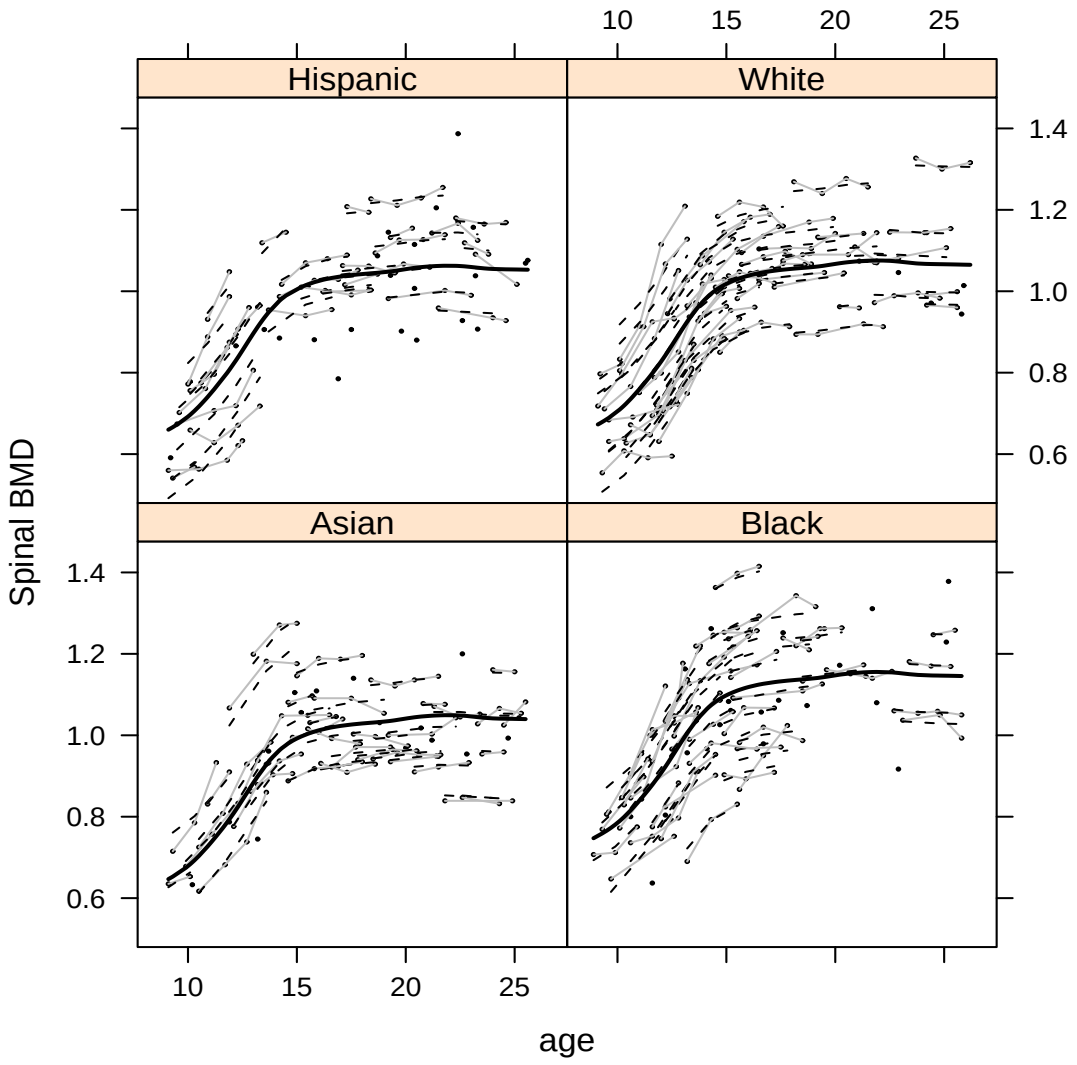


Fig. 3. Spinal bone mineral density versus age by ethnicity. Measurements on the same woman are joined with grey lines. The solid curve corresponds to the fitted spline, and the dashed lines to the individual fits.

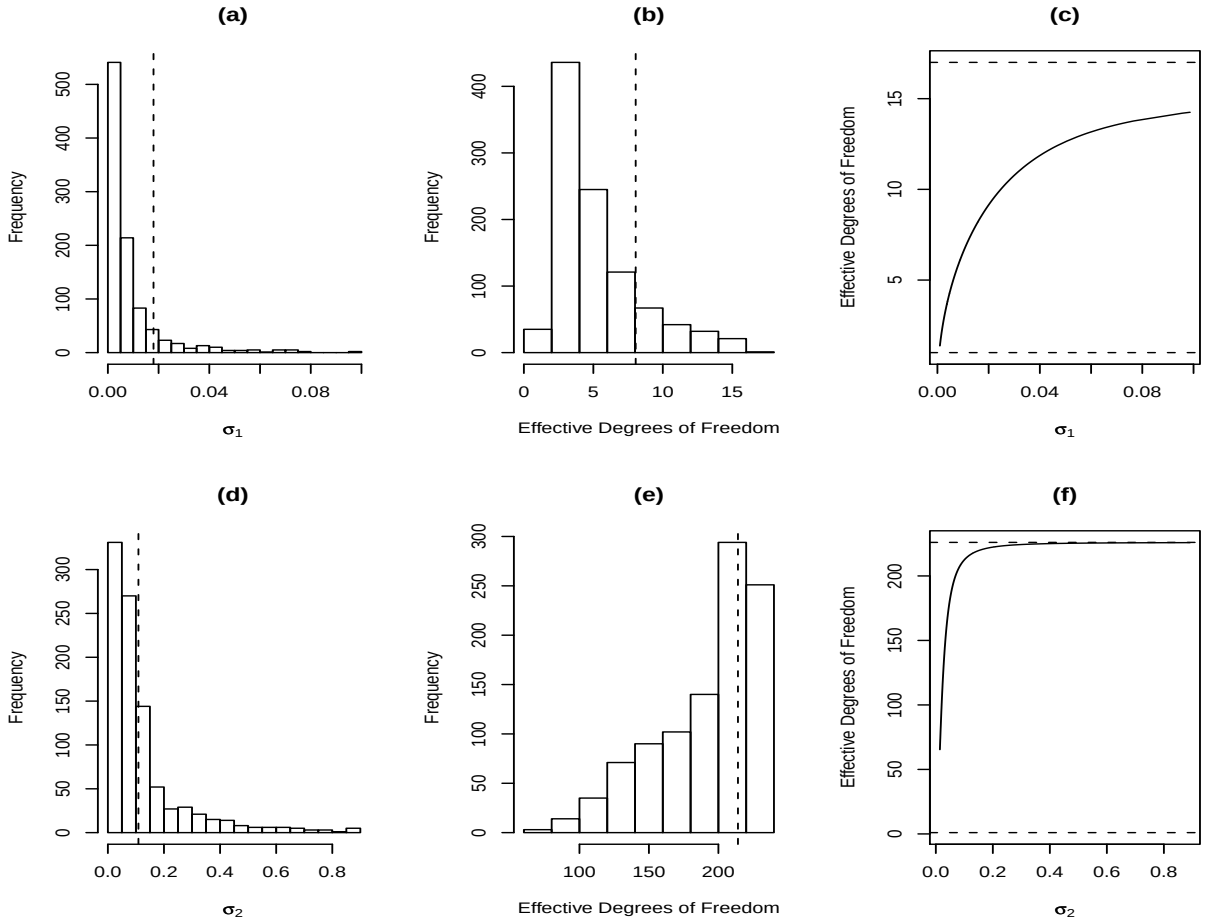


Fig. 4. Prior summaries: (a) σ_1, the standard deviation of the spline coefficients, (b) effective degrees of freedom associated with the prior for the spline coefficients, (c) effective degrees of freedom versus σ_1, (d) σ_2, the standard deviation of the between-individual random effects, (e) effective degrees of freedom associated with the individual random effects, (f) effective degrees of freedom versus σ_2. The vertical dashed lines on panels (a), (b), (d) and (e) correspond to the posterior medians.

σ_2^2. We assume the usual improper prior, $\pi(\sigma_\epsilon^2) \propto 1/\sigma_\epsilon^2$ for σ_ϵ^2. After some experimentation we settled on the prior $\sigma_1^{-2} \sim \text{Ga}(0.5, 5 \times 10^{-6})$. For σ_2^2 we wished to have a 90% interval for b_{2i} of ± 0.3 which, with 1 degree of freedom for the marginal distribution, leads to $\sigma_2^{-2} \sim \text{Ga}(0.5, 0.00113)$. Figure 4 shows the priors for σ_1 and σ_2, along with the implied effective degrees of freedom under the assumed priors. For the spline component the 90% prior interval for the effective degrees of freedom is [2.4,10].

Table 2 compares estimates from REML and INLA implementations of the model, and we see close correspondence between the two. Figure 4 also shows the posterior medians for σ_1, σ_2 and the two effective degrees of freedom. For the spline and random effects these correspond to 8 and 214, respectively. The latter figure shows that there is considerable variability between the 230 women here. This is confirmed in Figure 3 where we observe large vertical differences between the profiles. This figure also shows the fitted spline, which appears to mimic the trend in the data well.

Variable	REML	INLA
Intercept	0.560 ± 0.029	0.563 ± 0.031
Black	0.106 ± 0.021	0.106 ± 0.021
Hispanic	0.013 ± 0.022	0.013 ± 0.022
White	0.026 ± 0.022	0.026 ± 0.022
Age	0.021 ± 0.002	0.021 ± 0.002
σ_1	$0.018^\star$	0.024 ± 0.006
σ_2	$0.109^\star$	0.109 ± 0.006
σ_ϵ	$0.033^\star$	0.033 ± 0.002

Table 2. REML and INLA summaries for spinal bone data. Intercept corresponds to Asian group. For the entries marked with a * standard errors were unavailable.

5.4 *Timings*

For the three models in the longitudinal data example, INLA takes 1 to 2 seconds to run, using a single CPU. To get estimates with similar precision with MCMC, we ran `JAGS` for 100,000 iterations, which

took 4 to 6 minutes. For the model in the temporal smoothing example, INLA takes 45 seconds to run, using one CPU. Because part of the INLA procedure can be executed in a parallel manner, if there are two CPUs available, as is the case with today's prevalent INTEL Core 2 Duo processors, INLA only takes 27 seconds to run. It is not currently possible to implement this model in `JAGS`. We ran the MCMC utility built into the INLA software for 3.6 million iterations, to obtain estimates of comparable accuracy, which took 15 hours. For the model in the B-spline nonparametric regression example, INLA took 5 seconds to run, using a single CPU. We ran the MCMC utility built into the INLA software for 2.5 million iterations to obtain estimates of comparable accuracy, the analysis taking 40 hours.

6. DISCUSSION

In this paper we have demonstrated the use of the INLA computational method for GLMMs. We have found that the approximation strategy employed by INLA is accurate in general, but less accurate for binomial data with small denominators. The on-line supplementary material contains an extensive simulation study, replicating that presented in Breslow and Clayton (1993). There are some suggestions in the discussion of Rue et al. (2009) on how to construct an improved Gaussian approximation that does not use the mode and the curvature at the mode. It is likely that these suggestions will improve the results for binomial data with small denominators. There is an urgent need for diagnosis tools to flag when INLA is inaccurate. Conceptually, computation for non-linear mixed effects models (Davidian and Giltinan, 1995; Pinheiro and Bates, 2000) can also be handled by INLA, but this capability is not currently available.

The website `www.r-inla.org` contains all of the data and `R` scripts to perform the analyses and simulations reported in the paper. The latest release of software to implement INLA can also be found at this site. Recently, Breslow (2005) revisited PQL and concluded that, "PQL still performs remarkably

well in comparison with more elaborate procedures in many practical situations". We believe that INLA provides an attractive alternative to PQL for GLMMs, and we hope that this paper stimulates the greater use of Bayesian methods for this class.

ACKNOWLEDGMENTS

JW was supported by grant R01 CA095994 from the National Institutes of Health.

REFERENCES

Bachrach, L.K., Hastie, T., Wang, M.C., Narasimhan, B., and Marcus, R. (1999). Bone mineral acquisition in healthy Asian, Hispanic, Black and Caucasian youth. A longitudinal study. *Journal of Clinical Endocrinology and Metabolism 84*, 4702–4712.

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

Breslow, N.E. (2005). Whither PQL? In D. Lin and P.J. Heagerty (Eds.), *Proceedings of the Second Seattle Symposium*, New York, pp. 1–22. Springer-Verlag.

Breslow, N.E. and Clayton, D.G. (1993). Approximate inference in generalized linear mixed models. *Journal of the American Statistical Association 88*, 9–25.

Breslow, N.E. and Day, N.E. (1975). Indirect standardization and multiplicative models for rates, with reference to the age adjustment of cancer incidence and relative frequency data. *Journal of Chronic Diseases 28*, 289–301.

Clayton, D.G. (1996). Generalized linear mixed models. In W.R. Gilks, S. Richardson, and D.J. Spiegelhalter (Eds.), *Markov Chain Monte Carlo in Practice*, pp. 275–301. Chapman and Hall.

Crainiceanu, C.M., Diggle, P.J., and Rowlingson, B. (2008). Bayesian analysis for penalized spline regression using WinBUGS. *Journal of the American Statistical Association 102*, 21–37.

Crainiceanu, C.M., Ruppert, D., and Wand, M.P. (2005). Bayesian analysis for penalized spline regression using

WinBUGS. *Journal of Statistical Software 14*.

Davidian, M. and Giltinan, D.M. (1995). *Nonlinear Models for Repeated Measurement Data*. Boca Raton: CRC Press.

DiCiccio, T.J., Kass, R.E., Raftery, A., and Wasserman, L. (1997). Computing Bayes Factors by Combining Simulation and Asymptotic Approximations. *Journal of the American Statistical Association 92*, 903–915.

Diggle, P., Heagerty, P., Liang, K.Y., and Zeger, S. (2002). *Analysis of Longitudinal Data, Second Edition*. Oxford University Press.

Fahrmeir, L., Kneib, T., and Lang, S. (2004). Penalized structured additive regression for space-time data: A Bayesian perspective. *Statistica Sinica 14*, 715–745.

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

Gelman, A. (2006). Prior distributions for variance parameters in hierarchical models. *Bayesian Analysis 1*, 515–534.

Hastie, T.J. and Tibshirani, R.J. (1990). *Generalized additive models*. Chapman and Hall, London.

Hobert, J.P. and Casella, G. (1996). The effect of improper priors on Gibbs sampling in hierarchical linear mixed models. *Journal of the American Statistical Association 91*, 1461–1473.

Kass, R.E. and Steffey, D. (1989). Approximate Bayesian inference in conditionally independent hierarchical models (parametric empirical Bayes models). *Journal of the American Statistical Association 84*, 717–726.

Kelsall, J.E. and Wakefield, J.C. (1999). Discussion of "Bayesian models for spatially correlated disease and exposure data" by N. Best, L. Waller, A. Thomas, E. Conlon and R. Arnold. In J.M. Bernardo, J.O. Berger, A.P. Dawid, and A.F.M. Smith (Eds.), *Sixth Valencia international meeting on Bayesian statistics*, London. Oxford University Press.

McCullagh, P. and Nelder, J.A. (1989). *Generalized Linear Models, Second Edition*. London: Chapman and Hall.

McCulloch, C.E., Searle, S.R., and Neuhaus, J.M. (2008). *Generalized, Linear, and Mixed Models* (Second ed.). John Wiley and Sons.

Meng, X. and Wong, W. (1996). Simulating ratios of normalizing constants via a simple identity. *Statistical Sinica 6*, 831–860.

Natarajan, R. and Kass, R.E. (2000). Reference Bayesian methods for generalized linear mixed models. *Journal of the American Statistical Association 95*, 227–237.

Nelder, J. and Wedderburn, R. (1972). Generalized linear models. *Journal of the Royal Statistical Society, Series A 135*, 370–384.

Pinheiro, J.C. and Bates, D.M. (2000). *Mixed-Effects Models in S and S-plus*. New York: Springer.

Plummer, M. (2008). Penalized loss functions for Bayesian model comparison. *Biostatistics 9*, 523–539.

Plummer, M. (2009). JAGS Version 1.0.3 Manual. Technical report.

Rue, H. and Held, L. (2005). *Gaussian Markov Random Fields: Thoery and Application*. Chapman and Hall/CRC.

Rue, H., Martino, S., and Chopin, N. (2009). Approximte Bayesian inference for latent Gaussian models using integrated nested Laplace approximations (with discussion). *Journal of the Royal Statistical Society, Series B 71*, 319–392.

Ruppert, D.R., Wand, M.P., and Carroll, R.J. (2003). *Semiparametric Regression*. New York: Cambridge University Press.

Skene, A.M. and Wakefield, J.C. (1990). Hierarchical models for multi-centre binary response studies. *Statistics in Medicine 9*, 919–929.

Spiegelhalter, D., Best, N., Carlin, B., and van der Linde, A. (1998). Bayesian measures of model complexity and fit (with discussion). *Journal of the Royal Statistical Society, Series B 64*, 583–639.

Spiegelhalter, DJ., Thomas, A., and Best, NG. (1998). WinBUGS User Manual, version 1.1.1. Cambridge, UK.

Thall, P.F. and Vail, S.C. (1990). Some covariance models for longitudinal count data with overdispersion. *Biometrics 46*, 657–671.

Verbeke, G. and Molenberghs, G. (2000). *Linear Mixed Models for Longitudinal Data*. New York: Springer-Verlag.

Verbeke, G. and Molenberghs, G. (2005). *Models for Discrete Longitudinal Data*. New York: Springer-Verlag.

Wakefield, J.C. (2007). Disease mapping and spatial regression with count data. *Biostatistics 8*, 158–183.

Wakefield, J.C. (2009). Multi-level modelling, the ecologic fallacy, and hybrid study designs. *International Journal of Epidemiology 38*, 330–336.

Wand, M.P. and Ormerod, J.T. (2008). On semiparametric regression with O'Sullivan penalised splines. *Australian and New Zealand Journal of Statistics 50*, 179–198.

Zeger, S.L. and Karim, M.R. (1991). Generalized linear models with random effects: A Gibbs sampling approach. *Journal of the American Statistical Association 86*, 79–86.