Variational approximations for Logistic Mixed Models

J.T. Ormerod, and M.P. Wand

School of Mathematics and Applied Statistics, University of Wollongong,
Northfields Avenue, Wollongong 2522, AUSTRALIA

Abstract: Variational approximations have been used extensively in Statistical Physics and Computer Science as a means of simplifying probability calculus. We investigate the transferral and adaptation of variational approximation methodology for the logistic mixed model setting, which is hindered by intractable multivariate integrals.Comparisons with accurate approximations based on Markov Chain Monte Carlo sampling shows that our methodology is comparable, while being significantly faster.

Keywords: Logistic Mixed Models; Variational Approximations; Penalized Splines.

1 Introduction

Logistic mixed models are an extremely useful class of models for binary data. They are a fundamental in longitudinal data analysis, a common analysis in biomedical statistics, where they can be used to model correlation in grouped data and include simple, hierarchical, crossed and nested random effect models (McCulloch & Searle, 2001; Zhao, Staudenmayer, Coull & Wand, 2006). They also can be used for function estimation including scatterplot smoothing, random coefficient and kriging models (Ruppert, Wand & Carrol, 2003; Zhao, Staudenmayer, Coull & Wand, 2006).

Unfortunately the analysis of logistic mixed models is hindered by the presence of analytically intractable integrals and approximations must be made. Approximations include Laplace-like approximations such as Penalized Quasi-Likelihood (PQL, Breslow & Clayton, 1993), Gauss-Hermite quadrature (Naylor & Smith, 1982) and

Monte Carlo methods (Robert & Casella, 1999; Gilks, Richardson & Spiegelhalter, 1996; McCullagh, 1997). Each of these methods have computational shortcomings associated with them. PQL approximations can be severely biased (Breslow & Lin, 1995; Lin & Breslow, 1996), Gauss-Hermite quadrature does not scale well to high dimensional integrals and Monte Carlo methods suffer from the problems of the slowness and difficulties accessing convergence (although some progress has been made, see Rosenthal, 1995 and Cowles & Rosenthal, 1998 for instance). Excellent summaries of existing approximations for generalized linear mixed models (GLMM), of which logistic mixed models are a special case, may be found in McCulloch & Searle (2001, Chapter 10) and Tuerlinckx, Rijmen, Verbeke & de Boeck (2006). However these overviews do not include variational approximations.

Variational methods are a class of analytic approximations which offer a fresh approach to many statistical problems. These are a class of analytic approximations with origins in physics which have recently been applied by computer scientists to a variety of statistical models (MacKay, 1995; Attias, 2000; Beal & Ghahramani, 2002; Beal, 2003; Bishop & Winn, 2003; Winn & Bishop, 2005; Consonni & Marin, 2007). In this article we describe a variational approach, similar to Rijmen & Vomlel (2007), for approximation of integrals arising in logistic mixed models and Bayesian versions thereof based on the ideas of Jaakkola & Jordan (2000). We also develop *grid-based* variational approximations for calculating marginal poster densities. We illustrate these methods with a case study.

2 Logistic Mixed Models

Consider the logistic mixed model with normally distributed random effects given by

$$\begin{aligned} [\mathbf{y}|\boldsymbol{\beta},\mathbf{u}] &= \exp\left\{\mathbf{y}^T(\mathbf{X}\boldsymbol{\beta}+\mathbf{Z}\mathbf{u}) - \mathbf{1}^T\log(\mathbf{1}+e^{\mathbf{X}\boldsymbol{\beta}+\mathbf{Z}\mathbf{u}})\right\} \\ [\mathbf{u}|\boldsymbol{\sigma}^2] &= |2\pi\mathbf{G}_{\boldsymbol{\sigma}^2}|^{-1/2}\exp\left\{-\tfrac{1}{2}\mathbf{u}^T\mathbf{G}_{\boldsymbol{\sigma}^2}^{-1}\mathbf{u}\right\} \end{aligned} \tag{1}$$

where $\mathbf{y} \equiv (y_1,\ldots,y_n)$ is a response vector, $\mathbf{X}$ and $\mathbf{Z}$ are $n \times p$ and

$n \times q$ matrices, $\boldsymbol{\beta}$ is a vector corresponding to the fixed effects of length p, $\mathbf{u}$ are random effects of length q with covariance matrix $\mathbf{G}_{\boldsymbol{\sigma}^2}$ which is parameterized by variance components $\boldsymbol{\sigma}^2 \equiv (\sigma_1^2, \ldots, \sigma_v^2)$.

This model arises in a number of common applications including scatterplot smoothing, additive models and longitudinal models (Ruppert *et al.*, 2003; Zhao *et al.*, 2006) which specify the particular values for the matrices $\mathbf{X}$, $\mathbf{Z}$ and $\mathbf{G}_{\boldsymbol{\sigma}^2}$.

The likelihood for $\boldsymbol{\beta}$ and $\boldsymbol{\sigma}^2$ is obtained by integrating out the random effects and is given by

$$\ell(\boldsymbol{\beta}, \boldsymbol{\sigma}^2) = \int_{\mathbb{R}^q} \exp\left\{\mathbf{y}^T(\mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{u}) - \mathbf{1}^T \log(\mathbf{1} + e^{\mathbf{X}\boldsymbol{\beta}+\mathbf{Z}\mathbf{u}})\right\} \frac{\exp\left(-\frac{1}{2}\mathbf{u}^T\mathbf{G}_{\boldsymbol{\sigma}^2}^{-1}\mathbf{u}\right)}{|2\pi\mathbf{G}_{\boldsymbol{\sigma}^2}|^{q/2}} d\mathbf{u}. \quad (2)$$

Bayesian logistic mixed models are also of considerable interest. A common Bayesian approach is to place vague priors on the parameters $\boldsymbol{\beta}$ and $\boldsymbol{\sigma}^2$. A common choice is

$$\boldsymbol{\beta} \sim N(\mathbf{0}, \sigma_\beta^2 \mathbf{I}) \quad \text{and} \quad \sigma_i^2 \sim IG(\alpha_i, \alpha_i) \quad (3)$$

where σ_β^2 is chosen to be a large positive constant, α_i is chosen to be a small positive constant and $IG(\alpha_i, \alpha_i)$ denotes the inverse-gamma distribution parameterized such that $\mathbb{E}(\sigma_i^2) = 1$. Gelman (2006) advocates the use of other priors, particularly for the σ_i^2s, but for simplicity we will focus on these.

Often the marginal posterior distributions for the β_is are of interest. These require calculation of integrals of the form

$$[\beta_i|\mathbf{y}] = \frac{\int [\mathbf{y}|\boldsymbol{\beta}, \mathbf{u}][\boldsymbol{\beta}][\mathbf{u}|\boldsymbol{\sigma}^2][\boldsymbol{\sigma}^2] d\boldsymbol{\beta}_{-i} d\mathbf{u} d\boldsymbol{\sigma}^2}{\int [\mathbf{y}|\boldsymbol{\beta}, \mathbf{u}][\boldsymbol{\beta}][\mathbf{u}|\boldsymbol{\sigma}^2][\boldsymbol{\sigma}^2] d\boldsymbol{\beta} d\mathbf{u} d\boldsymbol{\sigma}^2} \quad (4)$$

where $\boldsymbol{\beta}_{-i}$ is the vector $\boldsymbol{\beta}$ with the ith element removed.

Unfortunately there is no known closed form for (2) or (4) and so we must pursue approximations. We will now pursue variational approximations to these integrals.

3 Variational Approximations

There are two standard ways of finding such parameterized bounds. One method of deriving parameterized lower bounds exploits convexity properties of the integrand, we call *tangent transforms* (see Jaakkola & Jordan, 2000), while the other uses Jensen's inequality and leads to EM like algorithms (Hinton & van Camp 1993; MacKay, 1995; Attias, 2000; Beal & Ghahramani, 2002; Beal, 2003; Bishop & Winn, 2003; Winn & Bishop, 2005). For simplicity we will only focus on the first of these.

Tangent transforms exploit the fact that for any convex function $f : \mathbb{R}^d \to \mathbb{R}$ that

$$f(\mathbf{x}) \geq f(\boldsymbol{\xi}) + \mathsf{D}_{\mathbf{x}} f(\boldsymbol{\xi})(\mathbf{x} - \boldsymbol{\xi}) \quad \text{for all } \mathbf{x}, \boldsymbol{\xi} \in \mathbb{R}^d \tag{5}$$

and also

$$f(\mathbf{x}) = \max_{\boldsymbol{\xi}} f(\boldsymbol{\xi}) + \mathsf{D}_{\mathbf{x}} f(\boldsymbol{\xi})(\mathbf{x} - \boldsymbol{\xi}). \tag{6}$$

where $\mathsf{D}_{\mathbf{x}} f(\boldsymbol{\xi}) = (\partial f(\mathbf{x})/\partial x_i|_{\mathbf{x}=\boldsymbol{\xi}})_{1 \leq i \leq d}$ (e.g. Rockafellar, 1972). Similarly if f is concave we replace $\geq$ with $\leq$ in (5) and "max" with "min" in (6). While the approximation appears to be simplistic, since it only relies on first derivative information, often this type of approximation is adequate.

For logistic mixed models the difficult in calculating the likelihood (2) stems from the non-quadratic term $b(\mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{u}) = \log(\mathbf{1} + \exp(\mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{u}))$. Jaakkola & Jordan (1997) noticed that an upper bound for $b(x)$ can be obtained due to the fact that $g(x) = \log(e^{-x/2} + e^{x/2}) = b(x) - x/2$ is a concave function of x^2 and hence

$$b(x) \leq b_U(x, \xi) = \frac{x}{2} + \log(e^{-\xi/2} + e^{\xi/2}) + \frac{\tanh(\xi/2)}{4\xi}(x^2 - \xi^2) \tag{7}$$

which holds for all $x, \xi \in \mathbb{R}$. This bound is illustrated in Figure 1.

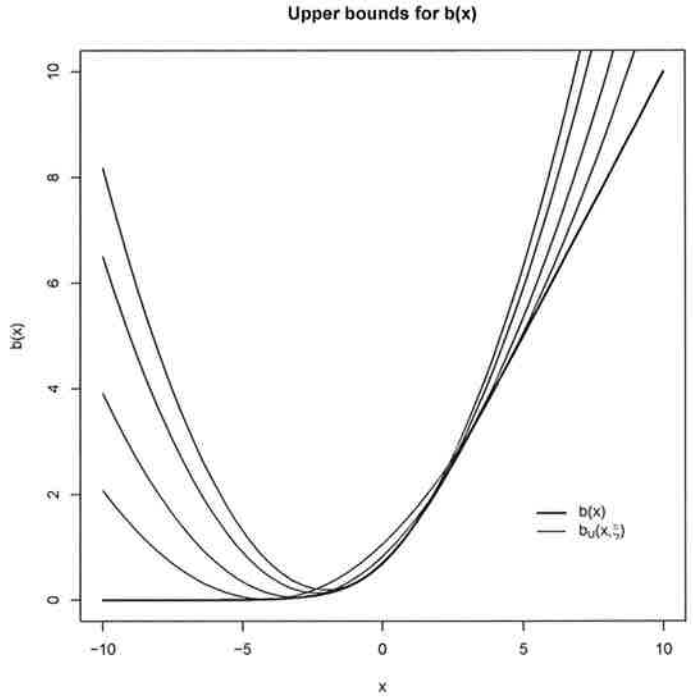

Figure 1: *Upper bounds for* $b(x)$.

Now we can use (7) to find a lower bound for the joint log-likelihood of $\mathbf{y}$ and $\mathbf{u}$, ignoring additive constants,

$$\log[\mathbf{y}, \mathbf{u}] \geq \left(\mathbf{y} - \tfrac{1}{2}\right)^T (\mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{u}) - \tfrac{1}{2} (\mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{u})^T \boldsymbol{\Lambda} (\mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{u}) - \tfrac{1}{2}\mathbf{u}^T \mathbf{G}_{\boldsymbol{\sigma}^2}^{-1} \mathbf{u} \quad (8)$$

where $\boldsymbol{\Lambda} = \text{diag}\left\{\tanh(\xi_i/2)\xi_i/2\right\}_{1\leq i\leq n}$ and $\boldsymbol{\xi} = (\xi_1, \ldots, \xi_n)$ are additional parameters, called *variational parameters*. Noting that since the right hand side of (8) is, ignoring additive constants, the log of a multivariate Gaussian function of $\mathbf{u}$ we might approximate the posterior distribution $\mathbf{u}|\mathbf{y}$ by $[\mathbf{u}|\mathbf{y}] \approx \delta(\mathbf{u}) = \phi_{\boldsymbol{\Sigma}}(\mathbf{u}-\boldsymbol{\mu})$ where $\phi_{\boldsymbol{\Sigma}}(\mathbf{u}-\boldsymbol{\mu}) = |2\pi\boldsymbol{\Sigma}|^{-1/2}\exp(-(\mathbf{u}-\boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\mathbf{u}-\boldsymbol{\mu})/2)$,

$$\boldsymbol{\mu} = \boldsymbol{\Sigma}\mathbf{Z}^T\left(\mathbf{y}-\tfrac{1}{2}\right) \quad \text{and} \quad \boldsymbol{\Sigma} = \left(\mathbf{Z}^T\boldsymbol{\Lambda}\mathbf{Z} + \mathbf{G}_{\boldsymbol{\sigma}^2}^{-1}\right)^{-1}. \tag{9}$$

We can now use this as the basis for an approximate EM algorithm.

Let $\boldsymbol{\mu}_{\text{old}}$ and $\boldsymbol{\Sigma}_{\text{old}}$ be the values for $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ evaluated at the current values for $\boldsymbol{\beta}$, $\boldsymbol{\sigma}^2$ and $\boldsymbol{\xi}$. The expectation step of the EM approach uses, ignoring additive constants,

$$\begin{aligned} Q(\boldsymbol{\beta},\boldsymbol{\sigma},\boldsymbol{\xi}|\boldsymbol{\mu}_{\text{old}},\boldsymbol{\Sigma}_{\text{old}}) &= \mathbb{E}_\delta\left\{\log[\mathbf{y},\mathbf{u}]\right\} \\ &\geq \log(\mathbf{1}^T g(\boldsymbol{\xi})) - \tfrac{1}{2}\log|\mathbf{G}_{\boldsymbol{\sigma}^2}| + \left(\mathbf{y}-\tfrac{1}{2}\right)^T(\mathbf{X}\boldsymbol{\beta}+\mathbf{Z}\boldsymbol{\mu}_{\text{old}}) \\ &\quad -\tfrac{1}{2}(\mathbf{X}\boldsymbol{\beta}+\mathbf{Z}\boldsymbol{\mu}_{\text{old}})^T\boldsymbol{\Lambda}(\mathbf{X}\boldsymbol{\beta}+\mathbf{Z}\boldsymbol{\mu}_{\text{old}}) - \tfrac{1}{2}\boldsymbol{\mu}_{\text{old}}^T\mathbf{G}_{\boldsymbol{\sigma}^2}^{-1}\boldsymbol{\mu}_{\text{old}} \\ &\quad -\tfrac{1}{2}\text{tr}\left(\boldsymbol{\Sigma}_{\text{old}}(\mathbf{Z}^T\boldsymbol{\Lambda}\mathbf{Z}+\mathbf{G}_{\boldsymbol{\sigma}^2}^{-1})\right). \end{aligned} \tag{10}$$

where $\mathbb{E}_\delta$ denotes expectations with respect to the approximate posterior distribution δ.

The maximisation step then maximizes Q with respect to $\boldsymbol{\beta}$, $\boldsymbol{\sigma}^2$ and $\boldsymbol{\xi}$. In order to optimize Q with respect to $\boldsymbol{\sigma}^2$ we need to specify a particular structure for $\mathbf{G}_{\boldsymbol{\sigma}^2}$. For simplicity we will use $\mathbf{G}_{\boldsymbol{\sigma}^2} = \text{blockdiag}_{1\leq i\leq v}(\sigma_i^2\boldsymbol{\Omega}_i)$. Other covariance structures for $\mathbf{u}$ are desirable in some contexts (see for example Verbeke & Molenberghs, 2000; McCulloch & Searle, 2001; Zhao *et al.*, 2006), however the covariance structure we have chosen here is quite general.

Differentiating Q with respect to $\boldsymbol{\beta}$, $\boldsymbol{\sigma}^2$ and $\boldsymbol{\xi}$, we find that first order optimality conditions require

$$\begin{aligned} \boldsymbol{\beta} &:= (\mathbf{X}^T\boldsymbol{\Lambda}\mathbf{X})^{-1}\mathbf{X}\left(\mathbf{y}-\tfrac{1}{2}-\boldsymbol{\Lambda}\mathbf{Z}\boldsymbol{\mu}_{\text{old}}\right) \\ \xi_i &:= \sqrt{(\mathbf{X}\boldsymbol{\beta}+\mathbf{Z}\boldsymbol{\mu}_{\text{old}})_i^2 + \tfrac{1}{2}(\mathbf{Z}\boldsymbol{\Sigma}_{\text{old}}\mathbf{Z}^T)_{ii}},\ 1\leq i\leq n \\ \sigma_i^2 &:= \left(\boldsymbol{\mu}_{\text{old}}^T\mathbf{D}_i\boldsymbol{\mu}_{\text{old}} + \text{tr}\left(\boldsymbol{\Sigma}_{\text{old}}\mathbf{D}_i\right)\right)/q_i,\ 1\leq i\leq v \end{aligned} \tag{11}$$

where $\mathbf{D}_i = \text{blockdiag}_{1\leq j\leq v}(\boldsymbol{\Omega}_j \mathbb{I}_{j=i})$, the q_is are the sizes of the square matrices $\boldsymbol{\Omega}_i$ and $\mathbb{I}_{\{x\}}$ denotes the indicator variable which takes the value 1 if x is true and 0 otherwise. We use (11) as the update equations for the EM algorithm and (9) to update the approximation of the posterior distribution $\mathbf{u}|\mathbf{y}$. Similar updates we first developed by Jaakkola & Jordan (1997) in the context of Bayesian logistic linear models and later by Rijmen & Vomlel (2007) for logistic random effects models. Our approach is an extension of these approaches to a more general model.

A similar approach may be used for Bayesian logistic mixed models. Suppose that we apply the priors (3) to our logistic mixed model. Integrating out the variance components $\boldsymbol{\sigma}^2$ we obtain, ignoring additive constants,

$$\begin{array}{l} \log[\mathbf{y}, \boldsymbol{\nu}] = \mathbf{y}^T\mathbf{C}\boldsymbol{\nu} - \mathbf{1}^T b(\mathbf{C}\boldsymbol{\nu}) - \\ \frac{\|\boldsymbol{\beta}\|^2}{2\sigma_\beta^2} - \sum_{i=1}^v (\alpha_i + q_i/2)\log\left(\alpha_i + \frac{\mathbf{u}^T\mathbf{D}_i\mathbf{u}}{2}\right) \end{array} \tag{12}$$

where $\mathbf{C} \equiv [\mathbf{X}, \mathbf{Z}]$ and $\boldsymbol{\nu} = (\boldsymbol{\beta}, \mathbf{u})$.

Similar to the logistic mixed model case $[\mathbf{y}, \boldsymbol{\nu}]$ is not multivariate Gaussian in shape. However, noting that $-\log(x) \geq -\xi x + 1 + \log(\xi)$, see for example Jordan, Ghahramani, Jaakkola & Saul (1999), and again using (7) we obtain the following lower bound for (12), again ignoring additive constants,

$$\log[\mathbf{y}, \boldsymbol{\nu}] \geq h(\widetilde{\boldsymbol{\xi}}) + \left(\mathbf{y} - \tfrac{1}{2}\right)^T \mathbf{C}\boldsymbol{\nu} - \tfrac{1}{2}\boldsymbol{\nu}^T\left(\mathbf{C}^T\boldsymbol{\Lambda}\mathbf{C} + \mathbf{B}\right)\boldsymbol{\nu} \tag{13}$$

where $h(\widetilde{\boldsymbol{\xi}}) = \mathbf{1}^T g(\boldsymbol{\xi}) + \sum_{i=1}^v (\alpha_i + q_i/2)(\log(\xi_{n+i}) - \alpha_i\xi_{n+i})$,

$$\mathbf{B} = \begin{bmatrix} \sigma_\beta^{-2}\mathbf{I} & \mathbf{0} \\ \mathbf{0} & \sum_{i=1}^v (\alpha_i + q_i/2)\xi_{n+i}\mathbf{D}_i \end{bmatrix} \tag{14}$$

and $\widetilde{\boldsymbol{\xi}} = (\boldsymbol{\xi}, \xi_{n+1}, \ldots, \xi_{n+v})$ are additional variational parameters. Now the right hand side of (13) is, up to additive constants, the log of a

multivariate Gaussian distribution in $\boldsymbol{\nu}$. We use this to approximate the posterior $\boldsymbol{\nu}|\mathbf{y}$ by $\delta(\boldsymbol{\nu}) = \phi_{\boldsymbol{\Sigma}}(\boldsymbol{\nu} - \boldsymbol{\mu})$ where

$$\boldsymbol{\Sigma} = \left(\mathbf{C}^T\boldsymbol{\Lambda}\mathbf{C} + \mathbf{B}\right)^{-1} \quad \text{and} \quad \boldsymbol{\mu} = \boldsymbol{\Sigma}\mathbf{C}^T\left(\mathbf{y} - \tfrac{1}{2}\right) \tag{15}$$

noting that $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ depend on $\widetilde{\boldsymbol{\xi}}$ through $\boldsymbol{\Lambda}$ and $\mathbf{B}$.

We select $\widetilde{\boldsymbol{\xi}}$ by maximising a lower bound for the marginal likelihood. As before this is most elegantly done via the EM algorithm. Let $\boldsymbol{\mu}_{\text{old}}$ and $\boldsymbol{\Sigma}_{\text{old}}$ be the values for $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ evaluated at the current values for $\widetilde{\boldsymbol{\xi}}$. Then the Q function is given by

$$\begin{aligned} Q(\widetilde{\boldsymbol{\xi}}|\boldsymbol{\mu}_{\text{old}}, \boldsymbol{\Sigma}_{\text{old}}) &= \mathbb{E}_\delta\left\{\log\left[\mathbf{y}, \boldsymbol{\nu}\right]\right\} \\ &\leq h(\widetilde{\boldsymbol{\xi}}) + \left(\mathbf{y} - \tfrac{1}{2}\right)^T\mathbf{C}\boldsymbol{\mu}_{\text{old}} - \tfrac{1}{2}\boldsymbol{\mu}_{\text{old}}^T\left(\mathbf{C}^T\boldsymbol{\Lambda}\mathbf{C} + \mathbf{B}\right)\boldsymbol{\mu}_{\text{old}} \\ &+\text{tr}\left[\boldsymbol{\Sigma}_{\text{old}}\left(\mathbf{C}^T\boldsymbol{\Lambda}\mathbf{C} + \mathbf{B}\right)\right]. \end{aligned} \tag{16}$$

The maximisation step then maximizes Q with respect to $\widetilde{\boldsymbol{\xi}}$. Differentiating Q with respect to $\widetilde{\boldsymbol{\xi}}$ and solving the first order optimality conditions we obtain

$$\begin{aligned} \xi_i &:= \sqrt{(\mathbf{C}\boldsymbol{\mu}_{\text{old}})_i^2 + \tfrac{1}{2}(\mathbf{C}\boldsymbol{\Sigma}_{\text{old}}\mathbf{C}^T)_{ii}},\ 1 \leq i \leq n \\ \xi_{n+i} &:= 2/(2\alpha_i + \left(\boldsymbol{\mu}_{\text{old}}^T\mathbf{B}_i\boldsymbol{\mu}_{\text{old}} + \text{tr}\left(\boldsymbol{\Sigma}_{\text{old}}\mathbf{B}_i\right)\right)),\ 1 \leq i \leq v \end{aligned} \tag{17}$$

where $\mathbf{B}_i = \text{blockdiag}(\mathbf{0}_p, \mathbf{D}_i)$, $1 \leq i \leq v$. We then iterate over (15) and (17) until convergence.

4 Grid-Based Posterior Density Approximations

Approximating marginal posterior densities is a common problem in Bayesian statistics. Unfortunately, as we will see in Section 5, using (15) can underestimate the variances of the marginal posterior densities. This tendency of variational approximations has been shown to arise in various settings (Humphreys & Titterington, 2000; Wang & Titterington, 2005; Consonni & Marin, 2007). Here we consider

approximating the marginal posterior densities by directly approximating (4).

Firstly, the joint distribution of $\mathbf{y}$ and $\boldsymbol{\nu}_{-i}$ is, ignoring additive constants,

$$\begin{aligned}\log[\mathbf{y}, \boldsymbol{\nu}_{-i}] \; &\geq (\mathbf{C}_{-i}^T\left(\mathbf{y} - \tfrac{1}{2}\right) - \left[\mathbf{C}^T\boldsymbol{\Lambda}\mathbf{C} + \mathbf{B}\right]_{-i,i}\nu_i)\boldsymbol{\nu}_{-i} \\ &\quad -\tfrac{1}{2}\boldsymbol{\nu}_{-i}^T\left[\mathbf{C}^T\boldsymbol{\Lambda}\mathbf{C} + \mathbf{B}\right]_{-i,-i}\boldsymbol{\nu}_{-i}\end{aligned} \tag{18}$$

where $\mathbf{C}_i$ is the ith column of $\mathbf{C}$, $\mathbf{C}_{-i}$ is the matrix $\mathbf{C}$ with the ith column removed, $\boldsymbol{\nu}_{-i}$ is the vector with the ith element removed, $[\mathbf{C}^T\boldsymbol{\Lambda}\mathbf{C} + \mathbf{B}]_{-i,-i}$ is the matrix $\mathbf{C}^T\boldsymbol{\Lambda}\mathbf{C} + \mathbf{B}$ with the ith row and column removed and $[\mathbf{C}^T\boldsymbol{\Lambda}\mathbf{C} + \mathbf{B}]_{-i,i}$ is the ith column of $\mathbf{C}^T\boldsymbol{\Lambda}\mathbf{C} + \mathbf{B}$ with the ith row removed. As before, the right hand side of (18) is, ignoring additive constants, the log of a multivariate Gaussian we make the approximation $[\boldsymbol{\nu}_{-i}|\mathbf{y}] \approx \delta(\boldsymbol{\nu}_{-i}) = \phi_{\boldsymbol{\Sigma}}(\boldsymbol{\nu}_{-i} - \boldsymbol{\mu})$ where

$$\boldsymbol{\Sigma} = \left[\mathbf{C}^T\boldsymbol{\Lambda}\mathbf{C} + \mathbf{B}\right]_{-i,-i}^{-1} \quad \text{and} \quad \boldsymbol{\mu} = \boldsymbol{\Sigma}(\mathbf{C}_{-i}^T\mathbf{y} - \left[\mathbf{C}^T\boldsymbol{\Lambda}\mathbf{C} + \mathbf{B}\right]_{-i,i}\nu_i). \tag{19}$$

Let $\boldsymbol{\mu}_{\text{old}}$ and $\boldsymbol{\Sigma}_{\text{old}}$ be the values for $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ evaluated at the current values for $\widetilde{\boldsymbol{\xi}}$. It can be shown (e.g. Hinton & van Camp 1993; MacKay, 1995; Attias, 2000), for any density δ, that

$$\begin{aligned}\log[\mathbf{y}, \nu_i] &\geq Q(\widetilde{\boldsymbol{\xi}}|\boldsymbol{\mu}_{\text{old}}, \boldsymbol{\Sigma}_{\text{old}}) + \mathcal{H}_\delta \\ &\geq h(\widetilde{\boldsymbol{\xi}}) + \left(\mathbf{y} - \tfrac{1}{2}\right)^T\mathbf{C}\boldsymbol{\mu}_{\text{old}} \\ &\quad -\tfrac{1}{2}\boldsymbol{\mu}_{\text{old}}^T\left(\mathbf{C}^T\boldsymbol{\Lambda}\mathbf{C} + \mathbf{B}\right)\boldsymbol{\mu}_{\text{old}} + \text{tr}\left[\boldsymbol{\Sigma}_{\text{old}}\left(\mathbf{C}^T\boldsymbol{\Lambda}\mathbf{C} + \mathbf{B}\right)\right] \\ &\quad +\tfrac{1}{2}\log|2e\pi\boldsymbol{\Sigma}_{\text{old}}| = \log[\mathbf{y}, \nu_i]_L\end{aligned} \tag{20}$$

where $Q(\widetilde{\boldsymbol{\xi}}|\boldsymbol{\mu}_{\text{old}}, \boldsymbol{\Sigma}_{\text{old}}) = \mathbb{E}_\delta\{\log[\mathbf{y}, \boldsymbol{\nu}]\}$ and $\mathcal{H}_\delta = \frac{1}{2}\log|2e\pi\boldsymbol{\Sigma}_{\text{old}}|$ is the entropy function for δ. Maximising for $\widetilde{\boldsymbol{\xi}}$ we obtain

$$\begin{aligned}\xi_i &:= \sqrt{(\mathbf{C}\boldsymbol{\mu}_{\text{old}})_i^2 + \tfrac{1}{2}(\mathbf{C}\boldsymbol{\Sigma}_{\text{old}}\mathbf{C}^T)_{ii}}, 1 \leq i \leq n \\ \xi_{n+i} &:= 2/(2\alpha_i + \left(\boldsymbol{\mu}_{\text{old}}^T\mathbf{B}_i\boldsymbol{\mu}_{\text{old}} + \text{tr}\left(\boldsymbol{\Sigma}_{\text{old}}\mathbf{B}_i\right)\right)), 1 \leq i \leq v.\end{aligned}$$

In order to tighten the bound (20) we maximise $\log[\mathbf{y}, \nu_i; \boldsymbol{\xi}]_L$ with respect to $\boldsymbol{\xi}$. Let

$$\widetilde{\boldsymbol{\xi}}^*(\nu_i) = \underset{\boldsymbol{\xi}}{\text{argmax}}\{\log[\mathbf{y}, \nu_i; \boldsymbol{\xi}]_L\} \tag{21}$$

so that $[\mathbf{y}, \nu_i; \widetilde{\boldsymbol{\xi}}^*]_L$ is also a tight lower bound for $[\mathbf{y}, \nu_i]$ noting we write $\widetilde{\boldsymbol{\xi}}^*(\nu_i)$ because of the fact that $\widetilde{\boldsymbol{\xi}}^*$ depends implicitly on ν_i via the optimisation problem (21).

Given $[\mathbf{y}, \nu_i; \widetilde{\boldsymbol{\xi}}^*(\nu_i)]_L$ we could approximate the marginal likelihood by $[\mathbf{y}]_L \equiv \int [\mathbf{y}, \nu_i; \widetilde{\boldsymbol{\xi}}^*(\nu_i)]_L d\nu_i$ The complicated dependency of $\widetilde{\boldsymbol{\xi}}^*$ on ν_i means that it may be impossible to find a closed form expression for $[\mathbf{y}, \nu_i; \widetilde{\boldsymbol{\xi}}^*(\nu_i)]_L$. Instead we evaluate $\widetilde{\boldsymbol{\xi}}_j^* = \max_{\boldsymbol{\xi}}[\mathbf{y}, \widehat{\nu}_{ij}; \widetilde{\boldsymbol{\xi}}]_L$ for a grid of values $(\widehat{\nu}_{i1}, \ldots, \widehat{\nu}_{iN})$ for some integer N. We then approximate $\log[\mathbf{y}, \nu_i; \widetilde{\boldsymbol{\xi}}^*(\nu_i)]_L$ by some curve $\log[\mathbf{y}, \nu_i]_G$ (where the subscript G denotes a grid based approximation) such that

$$\log[\mathbf{y}, \widehat{\nu}_{ij}]_G = \log[\mathbf{y}, \widehat{\nu}_{ij}; \widetilde{\boldsymbol{\xi}}_j^*]_L \quad \text{for} \quad 1 \leq j \leq N, \tag{22}$$

i.e. $\log[\mathbf{y}, \widehat{\nu}_{ij}]_G$ interpolates the points $(\widehat{\nu}_{ij}, \log[\mathbf{y}, \widehat{\nu}_{ij}; \widehat{\boldsymbol{\xi}}_j]_L)$ for $1 \leq j \leq N$. Finally a grid based variational posterior approximation (GBVPA) for $[\nu_i|\mathbf{y}]$ is given by

$$[\nu_i|\mathbf{y}]_G \equiv [\mathbf{y}, \nu_i]_G / [\mathbf{y}]_G \tag{23}$$

where the one dimensional integral $[\mathbf{y}]_G \equiv \int [\mathbf{y}, \nu_i]_G d\nu_i$ is evaluated numerically.

There are a number of choice to be made. These include the choice and number of grid values, type of interpolation used to approximate $\log[\mathbf{y}, \nu_i; \widetilde{\boldsymbol{\xi}}^*(\nu_i)]_L$ and quadrature method to approximate $[\mathbf{y}, \nu_i]_G$. The choices we have made in the following examples are as follows: 1) Suppose that (ν_{iL}, ν_{iR}) is a 95% highest posterior density credible region for ν_i based on (15). Then we let $(\widehat{\nu}_{i1}, \ldots, \widehat{\nu}_{iN})$

be equally spaced on the interval $\nu_i \in (\nu_{iL} - \triangle/2, \nu_{iR} + \triangle/2)$ where $\triangle = \nu_{iR} - \nu_{iL}$. 2) We used the `R` function `spline()` for the interpolater $[\mathbf{y}, \nu_i]_G$. 3) A 5,000 point trapezoid rule was used to approximate $[\mathbf{y}]_G \equiv \int [\mathbf{y}, \nu_i]_G d\nu_i$ on the interval $\nu_i \in (\nu_{iL} - \triangle/2, \nu_{iR} + \triangle/2)$.

One possible downside of GBVPA is that N optimisation problems of the form (21) need to be solved for each marginal posterior density. Thus, in practice, we seek to choose the grid $(\widehat{\nu}_{i1}, \ldots, \widehat{\nu}_{iN})$ with as few points as possible but enough points to ensure that we have a reasonable approximation for $[\nu_i|\mathbf{y}]_G$. We note that GBVPA could potentially be easily improved however we propose GBVPA as a starting place for such improvements.

5 Numerical Experience

As outlined in the introduction there are many applications to logistic mixed models. For simplicity we will examine the effectiveness of variational approximations for logistic mixed models for the context of additive penalized spline smoothing (e.g. Eilers & Marx, 1996; Ruppert *et al.*, 2003; Wood, 2003; Welham, Cullis, Kenward & Thompson, 2007; Wand & Ormerod, 2008). Wand & Ormerod (2008) Section 5 considered a penalised spline analysis of union membership for a sample of 534 U.S. workers (source: Berndt, 1991) with the subset of covariates

$$\mathbf{x}_i = [\texttt{south}_i,\ \texttt{female}_i,\ \texttt{married}_i,\ \texttt{years.educ}_i,\ \texttt{wage}_i,\ \texttt{age}_i].$$

The variables `years.educ`, `years.experience`, `wage` and `age` are continuous and the variables `south`, `female` and `married` are binary. We consider a model of the form

$$\begin{aligned}\text{logit}\left\{\mathbb{P}\left(\texttt{union.member}_i = 1|\mathbf{x}_i\right)\right\} &= \beta_0 + \beta_1 \texttt{south}_i + \beta_2 \texttt{female}_i \\ &+ \beta_3 \texttt{married}_i + f_{\texttt{years.educ}}(\texttt{years.educ}_i) + f_{\texttt{wage}}(\texttt{wage}_i) \\ &+ f_{\texttt{age}}(\texttt{age}_i) = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{u}\end{aligned}$$

and use the mixed model formulation of cubic O'Sullivan splines (Wand & Ormerod 2008, Section 4) to model $f_{\texttt{years.educ}}$, $f_{\texttt{wage}}$ and $f_{\texttt{age}}$. We

used $K = 25$ quantile spaced inner knots for each of the continuous variables. Let $\mathbf{Z}_{\texttt{years.educ}}$, $\mathbf{Z}_{\texttt{wage}}$ and $\mathbf{Z}_{\texttt{age}}$ be the spline matrices for `years.educ`, `wage` and `age` respectively. Each of these matrices has $q_i = K + 2$ columns and

$$\mathbf{X} = [1, \texttt{years.educ}_i, \texttt{wage}_i, \texttt{age}_i]_{1 \leq i \leq n}, \quad \mathbf{Z} = [\mathbf{Z}_{\texttt{years.educ}}, \mathbf{Z}_{\texttt{wage}}, \mathbf{Z}_{\texttt{age}}]$$

and $\mathbf{D}_{\sigma^2} = \text{blockdiag}\left\{\sigma_i^{-2}\mathbf{I}_{q_i}\right\}_{1 \leq i \leq 3}$.

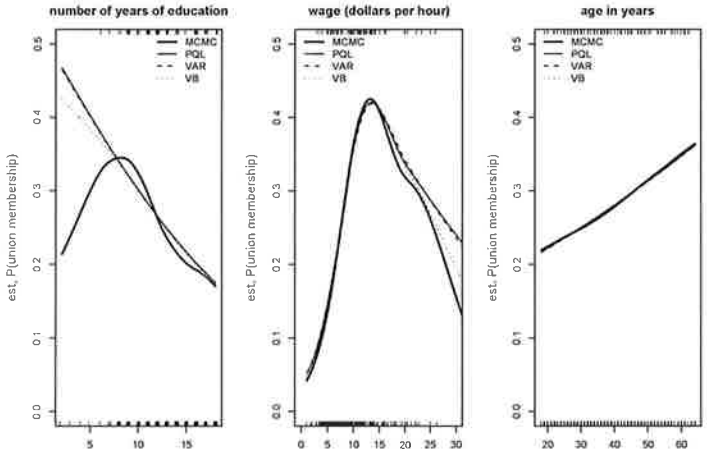

Figure 2: *Smooth function fits for the* `Trade Union` *model using the MCMC, PQL, VAR and VB approximations.*

From Figure 2 we see that the VAR and VB approximations produce fits which are quite similar to PQL. The approximations for the VB were, to the eye, slightly closer to the MCMC fit than the other approximations. While the analytic approximations are less accurate than MCMC methods they are extremely fast; taking about 4 minutes to fit each model. In contrast, for this example, the MCMC

approximation using `WinBUGS` (Spiegelhalter, Thomas & Best, 2000) took a little over an hour.

Finally, for Bayesian logistic mixed we can compare posterior density approximations for `south`, `female` and `married` using (9) and grid-based posterior approximations described in Section 4 with kernel density estimates (Scott, 1992; Wand & Jones, 1995) of posterior samples obtained via MCMC. The kernel density estimates use the Gaussian kernel with the bandwidth chosen via a direct plug-in method (Wand & Jones, 1995, Section 3.6) using the `R` package `KernSmooth`. Alternatively, the Sheather-Jones method (Sheather & Jones, 1991) can deliver excellent results.

It has been well-established in kernel smoothing literature that the choice of kernel has little effect on density estimates (e.g. Marron & Nolan, 1988, Wand & Jones, Chapter 2). However, the how the bandwidth is chosen does matter. Extensive simulation studies (e.g. Park & Turlach, 1992; Cao, Cuevas & Gonzalez-Manteiga, 1994; Jones, Marron & Sheather, 1996) have shown that, for large sample sizes and densities that are Gaussian in shape, automatic bandwidth methods such as the direct plug-in methods and the Sheather-Jones method lead to quite accurate density estimates.

For the MCMC fit we used `WinBUGS` to generate chains of length 50,000 after a burn-in of 5,000 and applied a thinning factor of 5, resulting in posterior samples of size 10,000 and then used the `R` package `KernSmooth` to estimate the densities of these posterior samples. Figure 3 illustrates these estimates. From this figure we notice that densities approximations based on (9) underestimate the amount of variance of the posteriors. The GBVPA approximations, on the other hand, were significantly better. Each GBVPA approximation took roughly 5 minutes to compute while the MCMC approach via `WinBUGS` took a little over 6 hours.

6 Conclusion

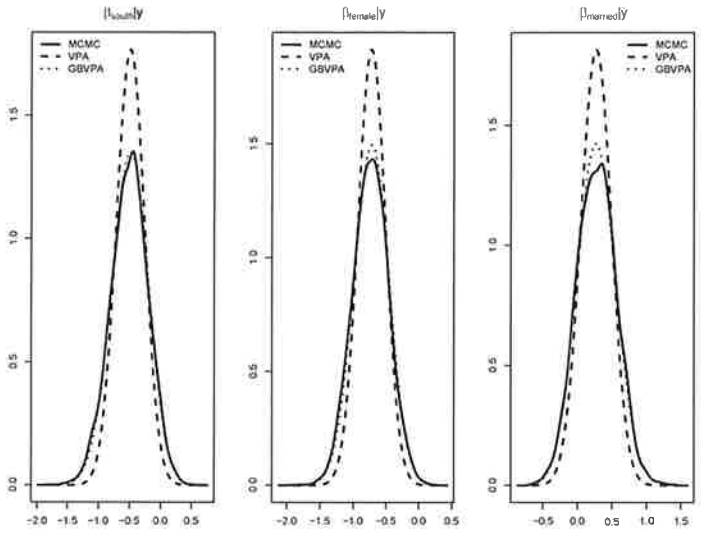

Figure 3: *Illustration of the kernel density estimates of MCMC posterior samples, variational posterior approximations (VPA) and grid-based variational posterior approximations (GBVPA) for* `south`, `female` *and* `married` *coefficients.*

In this article we focused on the logistic mixed model case. The extension to Generalised linear mixed models requires additional ideas which we do not have the space to cover here. In an upcoming paper Ormerod & Wand (2008) cover these cases for general responses including Poisson, gamma and inverse-Gaussian response types.

References

[1] Attias, H (2000). A variational Bayesian framework for graphical models. In Solla, S., Leen, T., and Muller, K.-R. (Eds.), *Advances in Neural Information*

Processing Systems, Volume 12, 209-215. Cambridge, MA: MIT press.

[2] Beal, M. (2003). *Variational algorithms for approximate Bayesian inference.* Ph.D. thesis, Gatsby Computational Neuroscience Unit.

[3] Beal, M. and Ghahramani, Z. (2002). The variational Bayesian EM algorithm for incomplete data: with application to scoring graphical model structures. In *Bayesian Statistics*, Volume 7. Oxford: Oxford University Press.

[4] Berndt, E.R. (1991). *The Practice of Econometrics: Classical and Contemporary.* Reading, Massachusetts: Addison-Wesley.

[5] Bishop, C.M. and Winn, J. (2003). Structural variational distributions in VIBES. In Bishop, C.M. and Frey, B. (Eds.), *Proceedings of Artificial Intelligence*, Florida, USA.

[6] Breslow, N.E. and Clayton, D.G. (1993). Approximate inference in generalized linear mixed models. *Journal American Statistical Association* 88(1), 9-25.

[7] Breslow, N.E. and Lin, X. (1995). Bias correction in generalized linear mixed models with a single component of dispersion. *Biometrika* 82(1), 81-91.

[8] Cao, R., Cuevas, A., and González Manteiga, W. (1994). A comparative study of several smoothing methods in density estimation. *Computational Statistics & Data Analysis* 17(2), 153-176.

[9] Consonni, G. and Marin, J.-M. (2007). Mean-field variational approximate Bayesian inference for latent variable models. *Computational Statistics Data Analysis* 52(2), 790-798.

[10] Cowles, M.K. and Carlin, B.P. (1996). Markov chain Monte Carlo convergence diagnostics: A comparative review. *Journal of American Statistics Association* 91, 883-904.

[11] Cowles, M.K. and Rosenthal, J.S. (1998). A simulation approach to convergence rates for markov chain Monte Carlo. *Statistics and Computing* 8, 115-124.

[12] Dempster, A.P., Laird, N.M., and Rubin, D.B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society*, Series B 339, 1-38.

[13] Eilers, P. and Marx, B.D. (1996). Flexible smoothing with B-splines and penalties (with comments and rejoinder). *Statistical Science* 11(2), 89-121.

[14] Gelman, A. (2006). Prior distributions for variance parameters in hierarchical models. (Comment on an article by Browne & Draper, 2006). *Bayesian Analysis* 1(3), 515-534.

[15] Gilks,W. R., Richardson, S., and Spiegelhalter, D.J. (1996). *Markov Chain Monte Carlo in Practice.* London: Chapman & Hall.

[16] Hastie, T.J. and Tibshirani, R.J. (1990). *Generalized Additive Models.* London: Chapman & Hall.

[17] Hastings,W.K. (1970). Monte Carlo sampling methods using Markov chains and their applications. *Biometrika* 57(1), 97-109.

[18] Hinton, G.E. and van Camp, D. (1993). Keeping the neural networks simple by minimizing the description length of the weights. In *COLT 93: Proceedings of the sixth annual conference on Computational learning theory*, New York, USA, 5-13. Association for Computing Machinery.

[19] Humphreys, K. and Titterington, D. (2001). Some examples of recursive variational approximations for Bayesian inference. In Opper, M. and Saad, D. (Eds.), *Advances Mean Field Methods: Theory and Practice.* Cambridge, MA: MIT Press.

[20] Jaakkola, T.S. (2001). Tutorial on variational approximation methods. In Opper, M. and Saad, D. (Eds.), *Advanced Mean Field Methods: Theory and Practice.* Cambridge, MA: MIT Press.

[21] Jaakkola, T.S. and Jordan, M.I. (2000). Bayesian parameter estimation via variational methods. *Statistics and Computing* 10, 25-37.

[22] Jones, M. C., Marron, J.S., and Sheather, S.J. (1996). Progress in data-based bandwidth selection for kernel density estimation. *Computational Statistics* 11(3), 337-381.

[23] Jordan, M.I., Ghahramani, Z., Jaakkola T.S., and Saul, L.K. (1999). An introduction to variational methods for graphical models. *Machine Learning* 37(2), 183-233.

[24] Lin, X. and Breslow, N.E. (1996). Bias correction in generalized linear mixed models with multiple components of dispersion. *Journal of the American Statistical Association* 91(435), 1007-1016.

[25] MacKay, D.J.C. (1995). Developments in probabilistic modelling with neural networks - ensemble learning. In Kappen, B. and Gielen, S. (Eds.), *Neural Networks: Artificial Intelligence and Industrial Applications. Proceedings of the 3rd Annual Symposium on Neural Networks*, Netherlands. Nijmegen.

[26] Marron, J.S. and Nolan, D. (1988). Canonical kernels for density estimation. *Statistics & Probability Letters* 7(3), 195-199.

[27] McCullagh, P. (1997). Maximum likelihood algorithms for generalized linear mixed models. *Journal of the American Statistical Association* 92, 162-170.

[28] McCullagh, P. and Nelder, J. (1989). *Generalized Linear Models* (2nd ed.). London: Chapman & Hall.

[29] McCulloch, C.E. and Searle, S.R. (2001). *Generalized, Linear, and Mixed Models.* New York: John, Wiley & Sons.

[30] Neal, R.M. and Hinton, G.E. (1998). A new view of the EM algorithm that justifies incremental, sparse and other variants. In Jordan, M.I. (Ed.), *Learning in Graphical Models*, 355-368. Kluwer Academic Publishers.

[31] Park, B.U. and Turlach, B.A. (1992). Practical performance of several data driven bandwidth selectors (with discussion). *Computational Statistics* 7(3), 251-270. Correction in Vol. 9, p. 79.

[32] Rijmen, F. and Vomlel, J. (2007). Assessing the performance of variational methods for mixed logistic regression models. *Journal of Statistical Computation and Simulation.* (accepted).

[33] Robert, C. and Casella, G. (1999). *Monte Carlo Statistical Methods* (2nd ed.). New York: Springer-Verlag.

[34] Rockafellar, R. (1972). *Convex Analysis.* Princeton, New Jersey: Princeton University Press.

[35] Rosenthal, J.S. (1995). Minorization conditions and convergence rates for Markov chain Monte Carlo. *Journal of American Statistics Association* 90, 558-566.

[36] Ruppert, D., Wand, M.P., and Carroll, R.J. (2003). *Semiparametric Regression.* Cambridge Series in Statistical and Probabilistic Mathematics. New York: Cambridge University Press.

[37] Scott, D.W. (1992). *Multivariate Density Estimation: Theory, Practice, and Visualization.* New York, Chichester: John Wiley & Sons.

[38] Sheather, S.J. and Jones, M.C. (1991). A reliable data-based bandwidth selection method for kernel density estimation. *Journal of the Royal Statistical Society, Series B* 53, 683-690.

[39] Spiegelhalter, D., Thomas, A., and Best, N. (2000). WinBUGS Version 1.3 User Manual. www.hrc-bsu.cam.ac.uk/bugs.

[40] Tuerlinckx, F., Rijmen, F., Verbeke, G., and de Boeck, P. (2006). Statistical inference in generalized linear mixed models: A review. *British Journal of Mathematical and Statistical Psychology* 59, 225-255.

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

[42] Wand, M.P. and Jones, M.C. (1995). *Kernel Smoothing.* London: Chapman & Hall.

[43] Wand, M.P. and Ormerod, J.T. (2008). On semiparametric regression with O'Sullivan penalised splines. *Australian & New Zealand Journal of Statistics.* (to appear).

[44] Wang, B. and Titterington, D.M. (2005). Inadequacy of interval estimates corresponding to variational Bayesian approximations. In Cowell, R.G. and Ghahramani, Z. (Eds.), *Proceedings of the Tenth International Workshop on Artificial Intelligence and Statistics*, 373-380. Society for AISTATS.

[45] Welham, S.J., Cullis, B.R., Kenward, M.G., and Thompson, R. (2007). A comparison of mixed model splines for curve fitting. *Australian & New Zealand Journal of Statistics* 49, 1-23.

[46] Winn, J. and Bishop, C. M. (2005). Variational message passing. *Journal of Machine Learning Research* 6, 661-694.

[47] Wood, S.N. (2003). Thin-plate regression splines. *Journal of the Royal Statistical Society, Series B* 65, 95-114.

[48] Zhao, Y., Staudenmayer, J., Coull, B.A., and Wand, M.P. (2006). General Design Bayesian Generalized Linear Mixed Models. *Statistical Science* 21, 35-51.