Grid Based Variational Approximations

John T. Ormerod[1]

School of Mathematics and Statistics,
University of Sydney, NSW 2006, Australia

Abstract

Variational methods for approximate Bayesian inference provide fast, flexible, deterministic alternatives to Monte Carlo methods. Unfortunately, unlike Monte Carlo methods, variational approximations cannot, in general, be made to be arbitrarily accurate. This paper develops grid-based variational approximations which endeavor to approximate marginal posterior densities in a spirit similar to the Integrated Nested Laplace Approximation (INLA) of Rue, Martino & Chopin (2009) but may be applied in situations where INLA cannot be used. The method can greatly increase the accuracy of a base variational approximation, although not in general to arbitrary accuracy. The methodology developed is at least reasonably accurate on all of the examples considered in the paper.

Key words: Bayesian Inference, Variational approximation, Kullback-Liebler divergence, Markov chain Monte Carlo.

1. Introduction

Markov chain Monte Carlo (MCMC) methods are a mainstay in Bayesian inference but can be painfully slow from the end user's perspective. Variational methodology for approximate Bayesian inference is emerging as viable alternative to MCMC methods when such methods become computationally infeasible. Originating mainly from Computer Science with the work of MacKay (1995) relatively accessible summaries of variational approximations may be found in Jordan, Ghahramani, Jaakkola & Saul (1999), Bishop (2006, Chapter 10) and an introduction from a statistical perspective can be found in Ormerod & Wand (2010a). These approaches should become easier to use with the emergence of a variational approximation-based software package named `Infer.NET` (Minka, Winn, Guiver & Kannan, 2008) which claims to be capable of handling a wide variety of statistical problems.

Ormerod & Wand (2010a) distinguish between a number of different types of variational approximations which together constitute a richer class of techniques than Laplace's method and its extensions (Breslow & Clayton, 1993; Rue, Martino & Chopin, 2009). These methods include what is most commonly called Variational Bayes, which we call product variational approximations (PVA) since the same techniques can be used in frequentist settings (see, for instance, Hall, Humphreys & Titterington 2002; Ormerod & Wand 2009; Hall, Ormerod & Wand, 2010; Ormerod & Wand 2010b). Particular variational approximations, depending on the problem at hand, can be hundreds of times faster than MCMC methods. Furthermore, implementation of these methods is often elegant, adding to their practical appeal.

Unfortunately, unlike MCMC methods, variational approximations cannot, in general, be made to be arbitrarily accurate. In several contexts variational approximations can be shown to be overconfident, i.e. they underestimate posterior variances. This overconfidence has been observed both numerically and theoretically in different settings (see for instance Wang & Titterington, 2006; Bishop, 2006; Cassonni & Marin, 2006; Rue, Martino & Chopin, 2009). For some problems this potential inaccuracy can rule out their use in practice.

In this paper grid-based variational approximations (GBVA) are developed which endeavor to approximate marginal posterior densities in a spirit similar to the Integrated Nested Laplace Approximation (INLA) of Rue, Martino & Chopin (2009). However, GBVA uses variational approximations in place of the Laplace-based approximations used by INLA and so may be used

[☆] Telephone: +61 2 9351 5883
Email address: `john.ormerod@sydney.edu.au` (John T. Ormerod)

in situations where INLA (which focuses on Gaussian latent effect models) cannot be used. The method can greatly increase the accuracy of a base variational approximation, although not in general to arbitrary accuracy. The methodology developed also has the advantage that only the marginal posterior distributions of interest need to be calculated, potentially resulting huge efficiency gains over MCMC methods.

Section 2 gives an overview of GBVA. Section 3 briefly describes PVAs and their corresponding product-type GBVAs. Section 4 illustrates this methodology for a pathological example where PVAs exhibit extreme overconfidence and demonstrates that a product-type GBVA can avoid this issue. Section 5 illustrates parametric variational approximations and their corresponding parametric-type GBVAs on, perhaps the simplest model where INLA cannot be applied, a linear regression model where a binary covariate contains missing values. Section 6 describes a parametric-type GBVAs and shows for a particular example where GBVA and INLA are directly comparable that the GBVA method can attain slightly greater accuracy is a reasonable amount of time. Concluding remarks are given in Section 7.

1.1. Notation

Integrals without limits are assumed to be over the entire space of the integrand argument. We use $\mathbf{1}_d$ to denote the $d \times 1$ column vector with all entries equal to 1. For a $d \times 1$ vector $\mathbf{a}$ we let $\mathbf{a}_{-i}$ denote the $(d-1) \times 1$ vector $\mathbf{a}$ with the ith entry removed and diag($\mathbf{a}$) the $d \times d$ diagonal matrix containing the entries of $\mathbf{a}$ along the main diagonal. For square matrices $\mathbf{A}_1, \ldots, \mathbf{A}_k$ we let blockdiag($\mathbf{A}_1, \ldots, \mathbf{A}_k$) denote the block diagonal matrix with ith block equal to $\mathbf{A}_i$. The matrix $\mathbf{A}_{-i}$ denotes the matrix $\mathbf{A}$ with the ith column removed. Scalar functions applied to vectors are evaluated element-wise, e.g. $\exp([a_1, a_2]^T) \equiv [\exp(a_1), \exp(a_2)]^T$. The derivative vector $\mathsf{D}_{\mathbf{x}} f(\mathbf{x})$, is the $1 \times d$ matrix with ith entry $\partial f(\mathbf{x})/\partial x_i$. The corresponding Hessian matrix is given by $\mathsf{H}_{\mathbf{x}} f(\mathbf{x}) = \mathsf{D}_{\mathbf{x}}\{\mathsf{D}_{\mathbf{x}} f(\mathbf{x})\}^T$. For a $p \times q$ matrix $\mathbf{A}$ we define vec($\mathbf{A}$) to be the $pq \times 1$ vector obtained by stacking the columns of $\mathbf{A}$ underneath each other from left to right and vech($\mathbf{A}$) to be vec($\mathbf{A}$) with the columns above the diagonal deleted. We let $\mathbf{D}_d$ denote the *duplication matrix* of order d defined by the relationship $\text{vec}(\mathbf{A}) = \mathbf{D}_d \text{vech}(\mathbf{A})$ for a symmetric $d \times d$ matrix $\mathbf{A}$. The density function of a random vector $\mathbf{u}$ is denoted by $p(\mathbf{u})$. The conditional density of $\mathbf{u}$ given $\mathbf{v}$ is denoted by $p(\mathbf{u}|\mathbf{v})$. A random variable x has an inverse-gamma distribution with parameters $s > 0$ and $r > 0$ is denoted $x \sim \text{IG}(s, r)$ if its density function is $p(x) = r^s \Gamma(s)^{-1} x^{-s-1} \exp(-r/x)$, $x > 0$.

2. Grid Based Variational Approximations

Most of Bayesian inference is concerned with the calculation of the posterior density $p(\mathcal{H}|\mathcal{E}) = p(\mathcal{E}, \mathcal{H})/\int p(\mathcal{E}, \mathcal{H}) d\mathcal{H}$ where $\mathcal{E}$ denotes the set of all observed or evidence variables, for example the response variables and covariates, and $\mathcal{H}$ denotes all unobserved or hidden variables, for example model parameters, latent and auxiliary variables or missing data. Integrals are replaced by summands over all combinations of discrete values for discrete random variables. For all but the simplest of problems the direct calculation of the quantity $p(\mathcal{H}|\mathcal{E})$ is problematic due to the presence of analytically intractable integrals (or computationally intractable summands over discrete variables). The marginal posterior densities may be written as

$$p(\boldsymbol{\theta}_i|\mathcal{E}) = \frac{\int p(\mathcal{E}, \mathcal{H}) d\mathcal{H}_{-i}}{\int p(\mathcal{E}, \mathcal{H}) d\mathcal{H}} = \frac{p(\mathcal{E}, \boldsymbol{\theta}_i)}{\int p(\mathcal{E}, \boldsymbol{\theta}_i) d\boldsymbol{\theta}_i} \quad \text{where} \quad p(\mathcal{E}, \boldsymbol{\theta}_i) = \int p(\mathcal{E}, \mathcal{H}) d\mathcal{H}_{-i} \tag{1}$$

where $\boldsymbol{\theta}_i$ is a low-dimensional subset of $\mathcal{H}$ and $\mathcal{H}_{-i}$ denotes the subset of $\mathcal{H} \setminus \boldsymbol{\theta}_i$. Equation (1) demonstrates that calculation of $p(\boldsymbol{\theta}_i|\mathcal{E})$ can be broken into two steps: 1. *Marginalization*: requiring calculation of $p(\mathcal{E}, \boldsymbol{\theta}_i)$ and 2. *Normalization*: requiring calculation of $p(\mathcal{E}) = \int p(\mathcal{E}, \boldsymbol{\theta}_i) d\boldsymbol{\theta}_i$.

Suppose that θ_i is a scalar so that normalization can be accurately and efficiently performed using one-dimensional quadrature via $p(\theta_{ij}|\mathcal{E}) \approx p(\mathcal{E}, \theta_{ij})/\sum_{j=1}^N w_j p(\mathcal{E}, \theta_{ij})$ for some weights $w_1, \ldots, w_N$ and abscissae (or grid) $\mathcal{G} = \{\theta_{i1}, \ldots, \theta_{iN}\}$ over the effective domain of $p(\theta_i|\mathcal{E})$. Provided that $p(\theta_i|\mathcal{E})$ is "not too rough" we can then accurately approximate $p(\theta_i|\mathcal{E})$ via interpolation. The main computational hurdle for such an approach is that it involves N (typically difficult) marginalization steps.

Grid-based variational approximations overcome this computational hurdle via the steps:

1. *Select a Grid.* Select a grid $\mathcal{G} = \{\theta_{i1}, \ldots, \theta_{iN}\}$ covering the effective domain of $p(\theta_i|\mathcal{E})$.
2. *Approximate Marginalization.* For each θ_{ij} approximate $p(\mathcal{E}, \theta_{ij})$ by $\widetilde{p}(\mathcal{E}, \theta_{ij})$ via some type of variational approximation.

3. *Interpolation.* Find a function $\widetilde{q}(\theta_i)$ such that $\log(\widetilde{q}(\theta_{ij})) \approx \log(\widetilde{p}(\mathcal{E}, \theta_{ij}))$, $1 \leq j \leq N$.
4. *Approximate Normalization.* Approximate $p(\theta_i|\mathcal{E})$ by $q(\theta_i) \equiv \widetilde{q}(\theta_i)/\int \widetilde{q}(\theta_i)d\theta_i$ where $\int \widetilde{q}(\theta_i)d\theta_i$ is approximated via numerical quadrature.

An overview of this methodology is illustrated in Figure 1. The twofold motivation behind this procedure is (1) we anticipate that $p(\mathcal{E}, \theta_i)$ is "not too rough" so that accurate interpolation of $p(\mathcal{E}, \theta_i)$ can be made between adjacent $p(\mathcal{E}, \theta_{ij})$ values and (2) variational approximations are a fast, flexible, reasonably accurate alternative to accurate but slow marginalization methods.

For simplicity we select the grid $\mathcal{G}$ for the parameter θ_i based on an initial variational approximation, say $q(\theta_i)$ of $p(\theta_i|\mathcal{E})$, but compensate for the fact that sometimes such approximations can be overconfident. Often $q(\theta_i)$ takes the form of a known parametric density, e.g. normal with mean μ and covariance σ^2, in which case we would write $q(\theta_i) = N(\mu, \sigma^2)$ for short. If the initial variational approximation is $q(\theta_i) = N(\mu, \sigma^2)$ or $q(\theta_i) = \text{Beta}(a, b)$ then we set $\mathcal{G}$ to the N equally spaced points between $\langle\theta_i\rangle - 5\sqrt{\text{Var}_q(\theta_i)}$ and $\langle\theta_i\rangle + 5\sqrt{\text{Var}_q(\theta_i)}$ where $\langle\theta_i\rangle$ and $\text{Var}_q(\theta_i)$ denote the mean and variances of θ_i with respect to q. Similarly, if $q(\theta_i) \sim IG(a, b)$ or $q(\theta_i) \sim \text{Gamma}(a, b)$ then we set $\mathcal{G}$ to the N logarithmically-spaced points between $\max(\langle\theta_i\rangle - 5\sqrt{\text{Var}_q(\theta_i)}, 10^{-3})$ and $\langle\theta_i\rangle + 10\sqrt{\text{Var}_q(\theta_i)}$. Note that all expressions for the expectations $\langle \cdot \rangle$ used in this paper well known and are summarized in Appendix A for convenience.

The INLA method described in Rue *et al.* (2009) uses a similar approach with Laplace-like methods for approximate marginalization. However, their method is quite involved due to the fact that the authors strive both for speed and high accuracy. The grid based variational approximations offer a simpler approach to INLA which may be attractive to situations where the INLA cannot be applied, i.e. to non-latent effect models. The method proposed by Rue *et al.* (2009), and the one presented here, might be categorized as *nested* approximations.

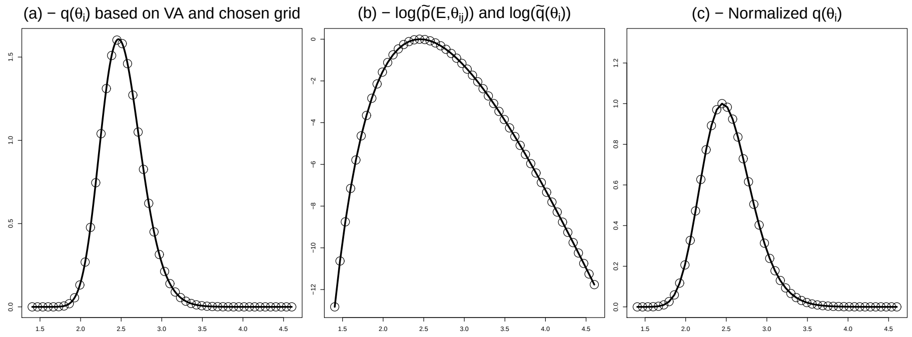

Figure 1: *An illustrative overview of grid-based variational approximations. (a) – A variational approximation is obtained. The variational approximation is used to select a grid of values over the approximate domain of the marginal posterior density. (b) – The unnormalized marginal posterior density is approximated via variational methods over the grid. (c) – An interpolant is found which is used to perform approximate normalization.*

2.1. Comparing Accuracy of Marginal Posterior Approximations

A comparison of the accuracies of various marginal posterior approximations requires calculation of highly accurate marginal posterior approximations. To this end, for each of the examples in the following sections, we used the `R` package `BRugs` (Ligges *et al.*, 2009) to obtain 505,000 MCMC samples for the model in question. The first 5,000 samples we used as burn-in and a thinning factor of 5 was used leaving 10^5 samples for inference. For such a high Monte Carlo sample size we would expect these MCMC based approximations to be fairly accurate.

Let $p(\theta_i|\mathcal{E})$ be the posterior density of interest, $q_{\text{MCMC}}(\theta_i)$ be the MCMC "gold standard" and $q(\theta_i)$ be an alternative approximation. Comparisons of accuracy between various methods were made via the L_2 or integrated square error, denoted $\text{ISE}(\theta_i) = \int [q(\theta_i) - p(\theta_i|\mathcal{E})]^2 d\theta_i$ which we approximate by $\int_{\theta_{i1}}^{\theta_{iN}} [q(\theta_i) - q_{\text{MCMC}}(\theta_i)]^2 d\theta_i$ via a composite Simpson's rule with 10,001 abscissa.

As with all comparisons some caveats need to be taken into account. Firstly, the MCMC, INLA, variational (VA) and GBVA results were computed using different programming languages. The MCMC model fits were obtained using the `BUGS` inference engine (Lunn *et al.* 2000) with interfacing via the package `BRugs` (Ligges, et al. 2009) in the `R` computing environment (R Development Core Team, 2008). The `INLA` package in `R` uses similar interfacing with `C` code (Martino, & Rue, 2009). The VA and GBVA methods described in this paper where implemented almost entirely in `R`.

Furthermore, no effort was made to tailor MCMC to the models at hand and each algorithm uses different termination conditions. Finally, any time recorded are merely indicative of the times each algorithm took on a typical 2010 computer. Comparisons of these methods should be made with these facts in mind.

3. Grid Based Product Variational Approximations

In the most common type of variational approximation, called amongst other things Product Variational Approximations (PVA), probability calculus is simplified by assuming selected sets of the model parameters are conditionally independent given the data. Such approximations do not take into account for all of the variability in the problem and tend to be overconfident. On the positive side PVA have elegant implementations, are reasonably flexible and are extremely efficient to use in practice. Furthermore, we can modify this type of variational method to be part of a GBVA approach with relative ease.

3.1. Product Variational Approximations

The PVA method is one way of dealing with the analytically intractable integrals required to calculate $p(\mathcal{H}|\mathcal{E})$ and (1). Like most variational approximations the PVA method uses the Kullback-Leibler distance between an arbitrary density $q(\mathcal{H})$ and the posterior density $p(\mathcal{H}|\mathcal{E})$ defined by $\text{KL}(q(\mathcal{H}), p(\mathcal{H}|\mathcal{E})) = \int q(\mathcal{H}) \log\{q(\mathcal{H})/p(\mathcal{H}|\mathcal{E})\}\, d\mathcal{H}$ noting that $\text{KL}(q(\mathcal{H}), p(\mathcal{H}|\mathcal{E}))$ is strictly positive and zero if and only if $q(\mathcal{H}) = p(\mathcal{H}|\mathcal{E})$ almost everywhere (Kullback & Leibler, 1951). The PVA of $p(\mathcal{H}|\mathcal{E})$ corresponds to minimizing KL subject to a factorization constraint on $q(\mathcal{H})$, i.e.

$$\min_{q(\mathcal{H})} \text{KL}(q(\mathcal{H}), p(\mathcal{H}|\mathcal{E})) \quad \text{subject to} \quad q(\mathcal{H}) = \prod_{i=1}^{M} q(\boldsymbol{\theta}_i) \tag{2}$$

where $\mathcal{H} = \{\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_M\}$ is some chosen partition of the hidden variables.

It can be shown, either using properties of the Kullback-Leibler distance or using calculus of variations, that the optimal $q(\boldsymbol{\theta}_i)$s for this problem, sometimes called q-densities, satisfy

$$q(\boldsymbol{\theta}_i) \propto \exp\{E_{-\boldsymbol{\theta}_i} \log p(\boldsymbol{\theta}_i|\mathcal{E}, \boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_{i-1}, \boldsymbol{\theta}_{i+1}, \ldots, \boldsymbol{\theta}_M)\}\,, \ 1 \le i \le M, \tag{3}$$

where $E_{-\boldsymbol{\theta}_i}$ denotes expectations with respect to the density $\prod_{j \neq i} q(\boldsymbol{\theta}_j)$.

A second useful result is the existence of a simple expressions for a lower bound on the marginal log-likelihood $\log p(\mathcal{E})$. It is easy derive from the Kullback-Leibler distance that

$$\log p(\mathcal{E}) \geq \log \underline{p}(\mathcal{E}) \equiv \int q(\mathcal{H}) \log\left\{\frac{p(\mathcal{E}, \mathcal{H})}{q(\mathcal{H})}\right\} d\mathcal{H}. \tag{4}$$

Combining (3) and (4) the following iterative scheme may be used to solve for the $q(\boldsymbol{\theta}_i)$s:

Algorithm 1: Product Variational Approximation

```
Initialize $q(\boldsymbol{\theta}_2), \ldots, q(\boldsymbol{\theta}_M)$.
Cycle:
    $q(\boldsymbol{\theta}_1) \leftarrow \dfrac{\exp\{E_{-\boldsymbol{\theta}_1} \log p(\boldsymbol{\theta}_1|\mathcal{E}, \boldsymbol{\theta}_2, \ldots, \boldsymbol{\theta}_M)\}}{\int \exp\{E_{-\boldsymbol{\theta}_1} \log p(\boldsymbol{\theta}_1|\mathcal{E}, \boldsymbol{\theta}_2, \ldots, \boldsymbol{\theta}_M)\}\, d\boldsymbol{\theta}_1}$
    $\vdots$
    $q(\boldsymbol{\theta}_M) \leftarrow \dfrac{\exp\{E_{-\boldsymbol{\theta}_M} \log p(\boldsymbol{\theta}_M|\mathcal{E}, \boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_{M-1})\}}{\int \exp\{E_{-\boldsymbol{\theta}_M} \log p(\boldsymbol{\theta}_M|\mathcal{E}, \boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_{M-1})\}\, d\boldsymbol{\theta}_M}$
until the change in $\log \underline{p}(\mathcal{E})$ becomes negligible.
```

This scheme has the advantage that $\log \underline{p}(\mathcal{E})$ is guaranteed to increase with every update and the $q(\boldsymbol{\theta}_i)$s converge to at least a local maximiser of $\log \underline{p}(\mathcal{E})$. Upon convergence the $q(\boldsymbol{\theta}_i)$s can used to approximate the marginal posteriors $p(\boldsymbol{\theta}_i|\mathcal{E})$. If conjugate priors are used the $q(\boldsymbol{\theta}_i)$s will belong to recognizable distributions and the updates in Algorithm 1 reduce to updating the parameters of the q-densities. In Computer Science, such an approach has become known as variational message passing (Winn & Bishop, 2005).

3.2. Grid Based Variational Approximations

The main element missing in the approach described in Section 2 is how to approximate $p(\mathcal{E}, \theta_{ij})$ for fixed θ_{ij}. To this end we calculate a lower bound $\underline{p}(\mathcal{E}, \boldsymbol{\theta}_{ij})$ for $p(\mathcal{E}, \theta_{ij})$ which can be efficiently calculated by replacing $\mathcal{E}$ and $\mathcal{H}$ with $\{\mathcal{E}, \theta_{ij}\}$ and $\mathcal{H}_{-i}$ respectively in (2)–(4). Algorithm 1 can be altered in a similar manner all the time treating θ_{ij} as a fixed observed value. Note that computational times can be significantly reduced by using the q-densities from θ_{ij} as initial densities for $\theta_{i,j+1}$. Examples of product-type GBVAs are described in Sections 4 and 5.

4. A Pathological Case for Product Variational Approximations

Consider the model

$$\mathbf{y}|\mathbf{u} \sim N(\mathbf{u}, \sigma^2\mathbf{I}), \quad \mathbf{u} \sim N(\mathbf{0}, \tau^{-1}\mathbf{I}) \quad \text{and} \quad \tau \sim \text{Gamma}(s, r) \tag{5}$$

where $\mathbf{y}$ is the observed data and we will use the prior hyperparameters $\sigma^2 = 100$, $s = 0.01$ and $r = 0.01$. For this model $\mathcal{E} = \{\mathbf{y}\}$ and $\mathcal{H} = \{\mathbf{u}, \tau\}$. This model is important because this example has been used to criticize PVA (see for example, Rue *et al.*, 2009) and is illustrative of both when PVA performs badly and where their GBVA-based modifications can perform well.

4.1. Product Variational Approximation

Consider the PVA to the model (5) corresponding to the partition $\mathcal{H} = \{\mathbf{u}, \tau\}$ so that $q(\mathcal{H}) = q(\mathbf{u})q(\tau)$. Using (3) we find the q-densities are of the form

$$q(\mathbf{u}) = N\{\left[\langle\tau\rangle + \sigma^{-2}\right]^{-1}\sigma^{-2}\mathbf{y}, \left[\langle\tau\rangle + \sigma^{-2}\right]^{-1}\mathbf{I}\} \quad \text{and} \quad q(\tau) = \text{Gamma}\left\{s + \tfrac{n}{2}, r + \tfrac{1}{2}\left\langle\|\boldsymbol{\nu}\|^2\right\rangle\right\} \tag{6}$$

where $\langle \cdot \rangle$ denotes expectations with respect to the q-densities and using (4) the corresponding lower bound for $\log p(\mathbf{y})$ is given by

$$\begin{aligned}\log \underline{p}(\mathbf{y}) \;&= -\tfrac{n}{2}\log(2\pi\sigma^2) - \tfrac{1}{2\sigma^2}\left[\|\mathbf{y}\|^2 - 2\mathbf{y}^T\langle\mathbf{u}\rangle + \langle\|\mathbf{u}\|^2\rangle\right] + \tfrac{n}{2}\left\langle \log\left(\tfrac{\tau}{2\pi}\right)\right\rangle - \tfrac{1}{2}\langle\tau\rangle\langle\|\mathbf{u}\|^2\rangle \\ &\quad + s\log(r) - \log\Gamma(s) + (s-1)\left\langle\log(\tau)\right\rangle - r\langle\tau\rangle - \left\langle\log q(\tau)\right\rangle - \left\langle\log q(\mathbf{u})\right\rangle.\end{aligned} \tag{7}$$

Using (6) and (7) Algorithm 1 can be used to fit the q-densities. Often Algorithm 1 is expressed as a sequence of update equations. For example, let us denote the mean and covariance of $q(\mathbf{u})$ as $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ respectively and the shape and rate of $q(\tau)$ by S and R respectively. Then $(\boldsymbol{\mu}, \boldsymbol{\Sigma}, S, R)$ are updated via

$$\begin{gathered}\boldsymbol{\Sigma}^{(t+1)} \leftarrow \left[\langle\tau\rangle + \sigma^{-2}\right]^{-1}\mathbf{I}, \qquad \boldsymbol{\mu}^{(t+1)} \leftarrow \sigma^{-2}\boldsymbol{\Sigma}^{(t)}\mathbf{y} \\ S^{(t+1)} \leftarrow s + \tfrac{n}{2} \quad \text{and} \quad R^{(t+1)} \leftarrow r + \tfrac{1}{2}\left[\|\boldsymbol{\mu}^{(t)}\|^2 + \text{tr}(\boldsymbol{\Sigma}^{(t)})\right]\end{gathered} \tag{8}$$

where $\boldsymbol{\mu}^{(0)}$, $\boldsymbol{\Sigma}^{(0)}$, $S^{(0)}$ and $R^{(0)}$ are chosen initial values. Defining new variables corresponding to the arguments of each q-density can be quite cumbersome, especially for more complicated examples. Henceforth, we will only derive the q-densities themselves, (e.g. (6)), from which it is trivial to find the appropriate update formula (e.g. (8)). Rue *et al.* (2009) used the updates (8) to show that as $\sigma^2 \to \infty$ the densities $q(\tau)$ and $p(\tau|\mathbf{y})$ have the same means but that the variance of $q(\tau)$ is too small by a factor of $O(n)$. This suggests that this variational approximation will be crude for large values of σ^2 or n.

4.2. Grid Based Variational Approximation for $p(\tau|\mathbf{y})$

Consider a product-type GBVA for $p(\tau|\mathbf{y})$ where we apply (2)–(4) with $\mathcal{E} = \{\mathbf{y}, \tau\}$ and $\mathcal{H} = \{\mathbf{u}\}$. Under this situation only the updates for $q(\mathbf{u})$ are applied since τ is observed. Since $\mathbf{u}$ is the only unobserved variable these updates converge in one iteration. Also it can easily be shown that $q(\mathbf{u}) = N(\sigma^{-2}(\tau + \sigma^{-2})^{-1}\mathbf{y}, (\tau + \sigma^{-2})^{-1}\mathbf{I}) = p(\mathbf{u}|\mathbf{y}, \tau)$ and

$$\log \underline{p}(\mathbf{y}, \tau) = -\tfrac{1}{2}\left\{n\log\left(2\pi(\sigma^2 + \tau^{-1})\right) + \tfrac{\|\mathbf{y}\|^2}{\sigma^2 + \tau^{-1}}\right\} + \log p(\tau) = \log p(\mathbf{y}, \tau) \tag{9}$$

for all $\mathbf{y}$, τ and σ^2. Since, in this case, $\underline{p}(\mathbf{y}, \tau) = p(\mathbf{y}, \tau)$ we can evaluate $\underline{p}(\mathbf{y}, \tau)$ over a grid of τ values and approximate $p(\mathbf{y})$, and hence $p(\tau|\mathbf{y})$, with high precision using one-dimensional quadrature.

4.3. Grid Based Variational Approximation for $p(u_i|\mathbf{y})$

Consider a product-type GBVA for $p(u_i|\mathbf{y})$ where $\mathcal{E} = \{\mathbf{y}, u_i\}$ and we apply (2)–(4) with the partition $\mathcal{H} = \{\mathbf{u}_{-i}, \tau\}$ (so that $q(\mathcal{H}) = q(\mathbf{u}_{-i})q(\tau)$). These choices give rise to the q-densities

$$q(\mathbf{u}_{-i}) = N\left\{\tfrac{\mathbf{y}_{-i}}{1+\sigma^2\langle\tau\rangle}, \left[\langle\tau\rangle + \sigma^{-2}\right]^{-1}\mathbf{I}\right\} \quad \text{and} \quad q(\tau) = \text{Gamma}\left\{s + \tfrac{n}{2}, r + \tfrac{1}{2}\left[u_i^2 + \langle\|\mathbf{u}_{-i}\|^2\rangle\right]\right\}$$

and $p(\mathbf{y}, u_i)$ can be approximated by $\underline{p}(\mathbf{y}, u_i)$ by replacing $\underline{p}(\mathbf{y})$ and $\langle\log q(\mathbf{u})\rangle$ with $\underline{p}(\mathbf{y}, u_i)$ and $\langle\log q(\mathbf{u}_{-i})\rangle$ respectively in (7). Following the methodology outlined in Section 2, we can evaluate $\underline{p}(\mathbf{y}, u_i)$ over a grid of u_is and then use interpolation and quadrature to approximate $p(u_i|\mathbf{y})$.

4.4. Comparisons

To illustrate the relative computational times and accuracies of the posterior approximations of $p(\tau|\mathbf{y})$ and some of the $p(u_i|\mathbf{y})$s we simulated 20 dataset from $\mathbf{y}|\mathbf{u} \sim N(\mathbf{u}, \sigma^2\mathbf{I})$ and $\mathbf{u} \sim N(\mathbf{0}, 10\mathbf{I})$ (so that the true value of τ was 1/10) with sample sizes $n = 100$, $n = 200$ and $n = 400$. We compared the times and ISEs of the variational approximation described in Section 4.1 (VA), the corresponding product-type GBVAs described in Section 4.2 and 4.3 (using 10, 30 and 50 grid points) and MCMC methods (using 10^3 and 10^4 samples). The results are illustrated in Figures 2 and 3 while Figure 4 illustrates some typical approximations of various marginal posterior approximations.

From Figures 2, 3 and 4 a number of conclusions can be entertained. Firstly, although VA is fast and its approximation of the $p(u_i|\mathbf{y})$s are accurate, its approximation of $p(\tau|\mathbf{y})$ is relatively crude. Furthermore, as is consistent with the theory of Rue *et al.* (2009), $q(\tau)$ becomes increasingly inaccurate as n increases. Secondly, GBVA is, on average, more accurate than MCMC approximations using 10^3 and 10^4 samples, regardless of the number of grid points used by GBVA. However, for the $p(u_i|\mathbf{y})$s little accuracy is gained by using GBVA over VA.

In terms of speed, if all of the posterior densities are required, GBVA is around 20 times faster than MCMC approximations using 10^5 samples. However, if we were willing to settle for MCMC approximations using only 10^3 samples then MCMC methods become faster than GBVAs, albeit with a loss in accuracy. On the other hand, if only 10 posterior densities where of interest, then GBVA would be hundreds of times faster than the MCMC method. Furthermore, if we were to only calculate $p(\tau|\mathbf{y})$ using GBVA and VA to calculate the $p(u_i|\mathbf{y})$s then accurate approximation of all of the posterior densities can be calculated in a way hundreds of times faster than MCMC methods, with little, if any, loss of accuracy.

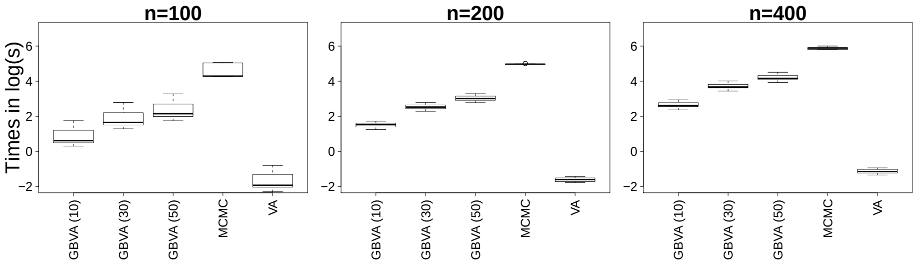

Figure 2: *Boxplots of times (in* $\log(seconds)$*) for* 30 *simulations of model (5) using VA, GBVA (with* 10, 30 *and* 50 *grid points) and MCMC with* 10^5 *methods with simulation settings described in Section 4.4.*

5. Linear Regression with Missing Binary Covariate

Consider the following model which falls outside the scope of the models INLA can handle but where GBVA can be used. Suppose that $y_i = \beta_0 + \beta_1 b_i + \varepsilon_i$, $1 \leq i \leq n$, where $\varepsilon_i \overset{\text{ind.}}{\sim} N(0, \sigma_\varepsilon^2)$, $1 \leq i \leq n$, and b is binary covariate (some of which are missing). For simplicity, using the terminology of Rubin (1976), we will assume that the b_is are missing completely at random (so that the missing data mechanism is ignorable). Furthermore, suppose that the b_is are 0 or 1 with a fixed probability ρ. Thus, we might consider the model

$$y_i|b_i, \boldsymbol{\beta}, \sigma_\varepsilon^2 \sim N((\mathbf{X}\boldsymbol{\beta})_i, \sigma_\varepsilon^2) \quad \text{and} \quad b_i|\rho \sim \text{Bernoulli}(\rho), \quad 1 \leq i \leq n \tag{10}$$

where $\mathbf{X} = [\mathbf{1}_n, \mathbf{b}]$, $\mathbf{b} = [b_1, \ldots, b_n]$ and $\boldsymbol{\beta} = [\beta_0, \beta_1]^T$. We will also use the priors $\boldsymbol{\beta} \sim N(\mathbf{0}, \sigma_\beta^2\mathbf{I})$, $\sigma_\varepsilon^2 \sim \text{IG}(s, r)$ and $\rho \sim \text{Unif}(0, 1)$ where $\sigma_\beta^2 = 10^8$, $s = 0.01$ and $r = 0.01$ are constants. Finally, we denote the vector of observed b_is as $\mathbf{b}_{\text{obs}}$ and the vector of missing b_is as $\mathbf{b}_{\text{mis}}$.

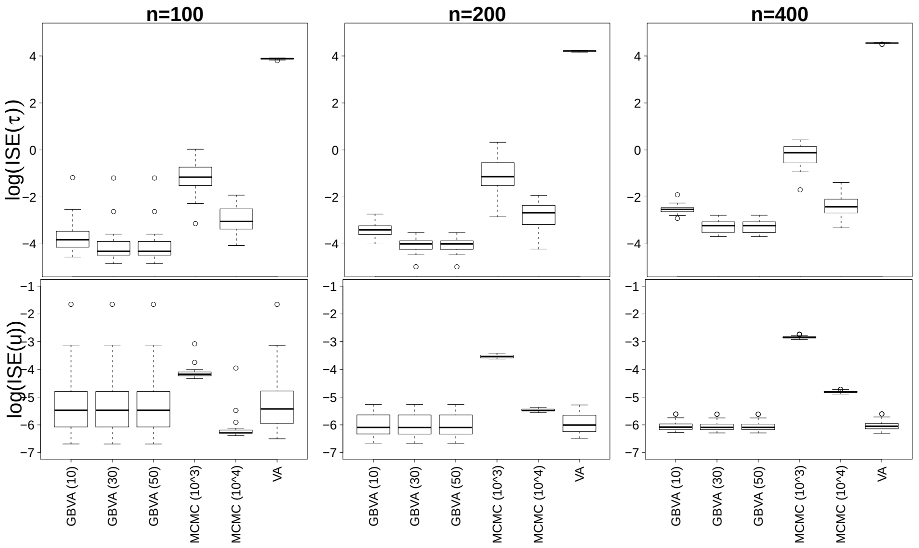

Figure 3: *Boxplots of approximate errors (in* $\log(ISE)$ *based on MCMC with* 10^5 *samples) for* 30 *simulations of model (5) using VA, GBVA (with* 10*,* 30 *and* 50 *grid points) and MCMC (with* 10^3 *and* 10^4 *samples) methods with simulation settings described in Section 4.4.*

5.1. Product Variational Approximation

Following the methodology described in Section 3 the q-densities corresponding to the partition $\mathcal{H} = \{\boldsymbol{\beta}, \sigma_\varepsilon^2, \rho, \mathbf{b}_{\text{mis}}\}$ are given by

$$\begin{array}{rl} q(\boldsymbol{\beta}) = N\left\{\left[\langle\sigma_\varepsilon^{-2}\rangle\langle\mathbf{X}^T\mathbf{X}\rangle + \sigma_\beta^{-2}\mathbf{I}\right]^{-1}\langle\sigma_\varepsilon^{-2}\rangle\langle\mathbf{X}\rangle^T\mathbf{y}, \left[\langle\sigma_\varepsilon^{-2}\rangle\langle\mathbf{X}^T\mathbf{X}\rangle + \sigma_\beta^{-2}\mathbf{I}\right]^{-1}\right\}, \\ q(\sigma_\varepsilon^2) = \text{IG}\left\{s + \frac{n}{2}, r + \frac{1}{2}\left[\|\mathbf{y}\|^2 - 2\mathbf{y}^T\langle\mathbf{X}\rangle\langle\boldsymbol{\beta}\rangle + \text{tr}(\langle\mathbf{X}^T\mathbf{X}\rangle\langle\boldsymbol{\beta}\boldsymbol{\beta}^T\rangle)\right]\right\}, \\ q(\rho) = \text{Beta}\left\{1 + \mathbf{1}_n^T\langle\mathbf{b}\rangle, 1 + n - \mathbf{1}_n^T\langle\mathbf{b}\rangle\right\} \quad \text{and} \quad q(b_i) = \text{Bernoulli}\{\eta_i\} \quad (\text{if } b_i \in \mathcal{H}) \end{array} \tag{11}$$

where the ith row of $\langle\mathbf{X}\rangle$ is $[1, b_i]$ if $b_i \in \mathcal{E}$ and $[1, \eta_i]$ if $b_i \in \mathcal{H}$, $\boldsymbol{\eta} = [\eta_1, \ldots, \eta_n]$,

$$\langle\mathbf{X}^T\mathbf{X}\rangle = \left[\begin{array}{cc} n & \mathbf{1}_n^T\langle\mathbf{b}\rangle \\ \mathbf{1}_n^T\langle\mathbf{b}\rangle & \mathbf{1}_n^T\langle\mathbf{b}\rangle \end{array}\right], \quad \mathbf{1}_n^T\langle\mathbf{b}\rangle = \mathbf{1}_{n_{\text{obs}}}^T\mathbf{b}_{\text{obs}} + \mathbf{1}_{n_{\text{mis}}}^T\boldsymbol{\eta} \quad \text{and}$$

$$\text{logit}(\eta_i) \quad = \langle\sigma_\varepsilon^{-2}\rangle\left[\langle\beta_1\rangle y_i - \langle\beta_0\beta_1\rangle - \tfrac{1}{2}\langle\beta_1^2\rangle\right] + \left\{\begin{array}{ll} \psi(1 + \mathbf{1}_n^T\langle\mathbf{b}\rangle) - \psi(1 + n - \mathbf{1}_n^T\langle\mathbf{b}\rangle), & \text{if} \quad \rho \in \mathcal{H}, \\ \log(\rho) - \log(1-\rho), & \text{if} \quad \rho \in \mathcal{E}. \end{array}\right.$$

We can now use (4) to derive a lower bound $\log \underline{p}(\mathbf{y}, \mathbf{b}_{\text{obs}})$ on $\log p(\mathbf{y}, \mathbf{b}_{\text{obs}})$ where

$$\begin{array}{rl} \log \underline{p}(\mathbf{y}, \mathbf{b}_{\text{obs}}) & = -\frac{n}{2}\langle \log(2\pi\sigma_\varepsilon^2)\rangle - \frac{1}{2}\langle\sigma_\varepsilon^{-2}\rangle\left[\|\mathbf{y}\|^2 - 2\mathbf{y}^T\langle\mathbf{X}\rangle\langle\boldsymbol{\beta}\rangle + \text{tr}(\langle\mathbf{X}^T\mathbf{X}\rangle\langle\boldsymbol{\beta}\boldsymbol{\beta}^T\rangle)\right] \\ & \quad -\frac{p}{2}\log(2\pi\sigma_\beta^2) - \frac{1}{2\sigma_\beta^2}\langle\|\boldsymbol{\beta}\|^2\rangle + \mathbf{1}_n^T\langle\mathbf{b}\rangle\langle\log(\rho)\rangle + (n - \mathbf{1}_n^T\langle\mathbf{b}\rangle)\langle\log(1-\rho)\rangle \\ & \quad +s\log(r) - \log\Gamma(s) - (s+1)\langle\log(\sigma_\varepsilon^2)\rangle - r\langle\sigma_\varepsilon^{-2}\rangle \\ & \quad -\langle\log q(\boldsymbol{\beta})\rangle - \langle\log q(\sigma_\varepsilon^2)\rangle - \langle\log q(\rho)\rangle - \langle\log q(\mathbf{b}_{\text{mis}})\rangle. \end{array} \tag{12}$$

Now that we have (11) and (12) we can use Algorithm 1 to fit the q-densities in an efficient and robust manner.

5.2. Grid Based Variational Approximations

The product-type GBVA corresponding to the PVA described in Section 5.1 can be obtained with a few modifications. In order to find the GBVA for parameter $\theta \in \{\beta_0, \beta_1, \sigma_\varepsilon^2, \rho\}$ say, we need to approximate $p(\mathbf{y}, \mathbf{x}_{\text{obs}}, \theta)$. Using the partition $\mathcal{H} \setminus \theta$ defined in Section 5.1 the q-densities are identical to those in (11) except when θ is β_i for $i = \{0, 1\}$. In this case the $q(\beta_{-i})$ is given by $q(\beta_{-i}) \sim N\{[\langle\sigma_\varepsilon^{-2}\rangle\langle\mathbf{X}_{-i}^T\mathbf{X}_{-i}\rangle + \sigma_\beta^{-2}]^{-1}\langle\sigma_\varepsilon^{-2}\rangle[\langle\mathbf{X}_{-i}\rangle^T\mathbf{y} - \langle\mathbf{X}_i^T\mathbf{X}_{-i}\rangle\beta_i], [\langle\sigma_\varepsilon^{-2}\rangle\langle\mathbf{X}_{-i}^T\mathbf{X}_{-i}\rangle + \sigma_\beta^{-2}]^{-1}\}$. The corresponding lower bound $\log \underline{p}(\mathbf{y}, \mathbf{x}_{\text{obs}}, \theta)$ can be easily calculated by using $\mathcal{E} = \{\mathbf{y}, \mathbf{x}_{\text{obs}}, \theta\}$ and $\mathcal{H} = \{\boldsymbol{\beta}, \sigma_\varepsilon^2, \rho, \mathbf{b}_{\text{mis}}\} \setminus \theta$. The q-densities can then be sequentially updated until the difference in $\log \underline{p}(\mathbf{y}, \mathbf{x}_{\text{obs}}, \theta)$ is negligible. The final value of $\log \underline{p}(\mathbf{y}, \mathbf{x}_{\text{obs}}, \theta)$ can be used to approximate $\log p(\mathbf{y}, \mathbf{x}_{\text{obs}}, \theta)$. Then using the methodology described in Section 2 we can approximate $\log p(\theta|\mathbf{y}, \mathbf{x}_{\text{obs}})$.

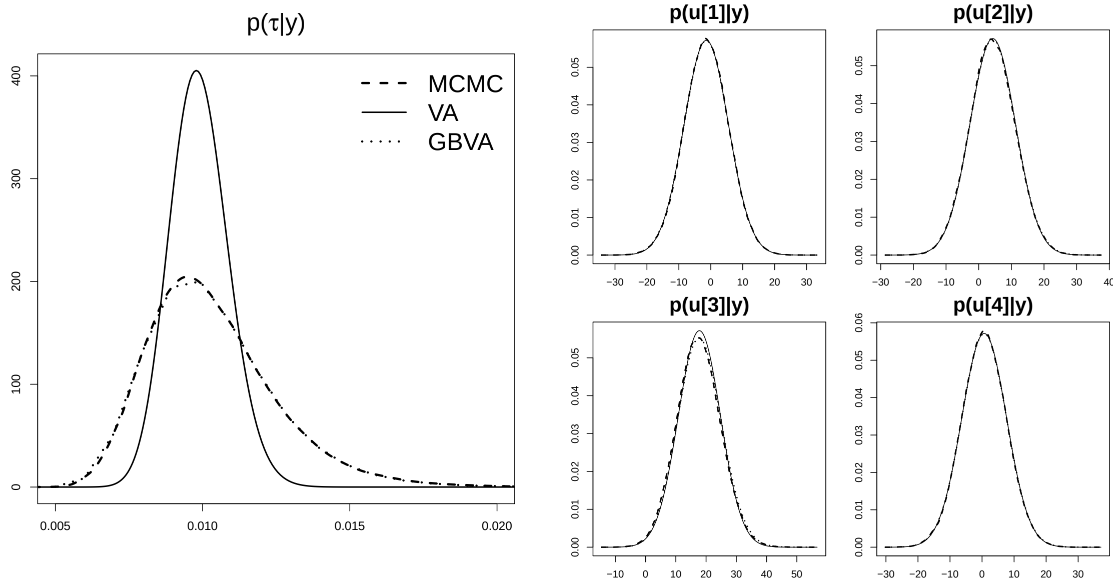

Figure 4: *Approximate posterior densities based on VA, MCMC and GBVA methods for simulated data (based on settings described in Section 4.4) for τ and the first 4 u_is for model (5). Note that all of the approximate posterior densities corresponding to the u_is are almost indistinguishable.*

5.3. Comparisons

To compare the accuracy of the PVA described in Section 5.1 and the corresponding product-type GBVA described in Section 5.2 we consider a simulation from the model (10) where 30 datasets with $n = 200$ points are generated with $\beta_0 = -1$, $\beta_1 = 2$, $\sigma_\varepsilon^2 = 2$, $\rho = 1/2$ and 50% points are removed completely at random. Box plots of times and ISEs of the variational approximation (VA), GBVA (with 10, 30 and 50 grid points) and MCMC methods (using 10^3 and 10^4 samples) are illustrated in Figure 5 while the most relevant approximate posterior densities are illustrated in Figure 6. We also considered different combinations of n, σ_ε^2 and percentage of points removed at random. However, the results from these simulations were very similar to the results illustrated in Figure 5 and are not shown.

Form Figure 5 we see that the proposed GBVA is both faster more accurate than the MCMC method. However, the times and ISEs for the GBVA methods do not include the calculation of the posterior distributions of the missing $\mathbf{b}_{\text{mis}}$ because these are probably not of interest. In Figure 6 we see that the proposed GBVA method gives posterior distributions which are indistinguishable to those given by the MCMC method with 10^5 samples.

Finally we note that the VA and GBVA algorithms scale very well to large n values. For $n = 10^6$ with 50% of xs randomly removed the VA, GBVA and MCMC (with 10^4 posterior samples) methods took 15 seconds, 10 minutes and 44 hours of computing time respectively.

6. Logistic Random Intercept Model

So far we have only considered models where product variational approximations give rise to q-densities which take the form of known distributions, i.e. conjugate models. In many situations, for example some latent effect models, this is not the case. Thus, in order to compare the INLA method (which deals solely with Gaussian latent effect models) with a GBVA alternative, a different variational method to the product type needs to be pursued.

Ormerod & Wand (2010a) described several such approaches including tangent and parametric variational approximations. Tangent-type GBVAs, based on the work of Jaakkola & Jordan (2000) where first considered in Ormerod & Wand (2008) with limited success and will not be pursued again here.

Parametric variational approximations, similarly to PVA methods minimize (2) with the additional constraint that one or more of the q-densities take a known parametric form, i.e.

$$\min_{q(\mathcal{H})} \ \text{KL}(q(\mathcal{H}), p(\mathcal{H}|\mathcal{E})) \quad \text{subject to} \quad q(\mathcal{H}) = \prod_{i=1}^{M} q(\boldsymbol{\theta}_i; \boldsymbol{\xi}_i) \tag{13}$$

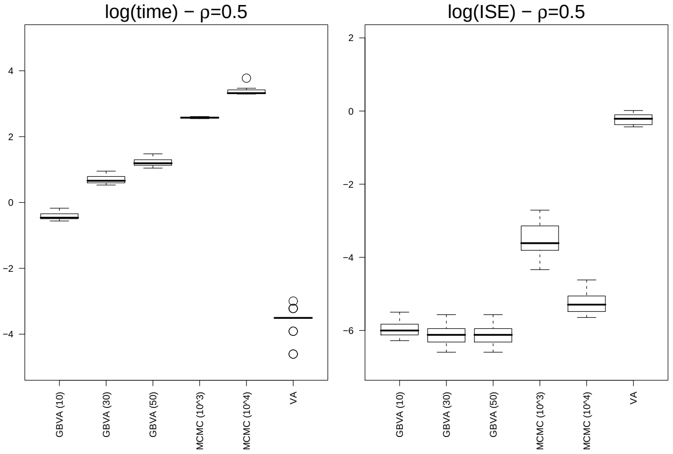

Figure 5: *A boxplot of the times (in* $\log(seconds)$*) and total approximate errors (in* $\log(ISE)$ *based on MCMC with* 10^5 *samples) for* 30 *simulations of model (10) with setting described in Section 5.3 using VA, GBVA (with* 10, 30 *and* 50 *grid points) and MCMC (with* 10^3 *and* 10^4 *samples) methods using simulation settings described in Section 5.3.*

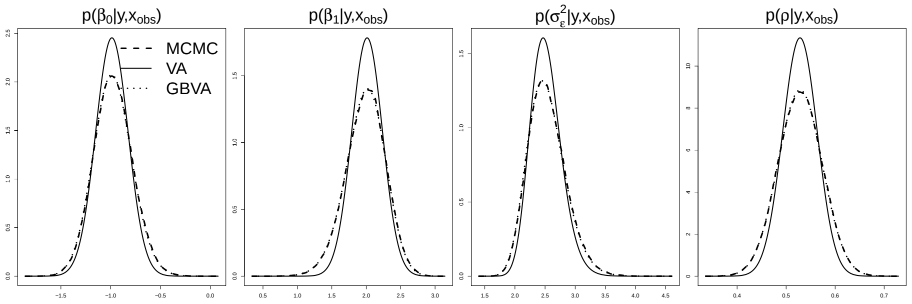

Figure 6: *Approximate posterior densities based on VA, MCMC and GBVA methods for* β_0*,* β_1*,* σ^2_ε *and* ρ *for model (10) with setting described in Section 5.3.*

where the $q(\boldsymbol{\theta}_i;\boldsymbol{\xi}_i)$s are now some conveniently chosen densities parameterized by $\boldsymbol{\xi}_i$. This approach originated from the work of Barber & Bishop (1998) where $q(\mathcal{H})$ was restricted to be a multivariate Gaussian density and appeared shortly after in Seeger (2000), and more recently in Opper & Archambeau (2009) and Ormerod & Wand (2010a-b) who called these Gaussian Variational Approximations (GVA).

To fix ideas consider the `bacteria` dataset which may be obtained from the `R` package `MASS` (Venables & Ripley, 2008). The *bacteria* datasets records tests of the presence of the bacteria H. influenzae in children with a history of otitis media in Northern Territory, Australia (Leach, 2000). The children were randomized into three groups: placebo, drug, and drug with active encouragement to comply. The presence of H. influenzae was checked at weeks 0, 2, 4, 6 and 11 to 30 and recorded as $\texttt{week}_{ij}$. If a particular check was missed no data was recorded for that week. High or low compliance of the patient in taking the treatment are indicated by the variables $\texttt{drugHi}_{ij}$ and $\texttt{drugLo}_{ij}$ respectively. The logistic random intercept model we considered for this dataset is described by

$$\text{logit}\left\{P(y_{ij}=1|\boldsymbol{\beta},u_i)\right\} = \beta_0 + \beta_1\texttt{drugLo}_{ij} + \beta_2\texttt{drugHi}_{ij} + \beta_3\texttt{week}_{ij} + u_i$$

for $1 \le i \le 50$, $1 \le j \le n_i$, where n_i takes values from 2 to 5. We assume that the u_is are normally distributed random effects with mean 0 and precision τ. We place the priors $\boldsymbol{\beta} \sim N(\mathbf{0},\sigma^2\mathbf{I})$ and $\tau \sim \Gamma(s,r)$ on $\boldsymbol{\beta}$ and τ respectively for known constants $\sigma^2 = 10^8$, $s = 0.01$ and $r = 0.01$.

6.1. Parametric Variational Approximation

Before considering the parametric variational approximation for this model first consider the PVA corresponding to the partition $q(\boldsymbol{\nu},\tau) = q(\boldsymbol{\nu})q(\tau)$ where $\boldsymbol{\nu} = [\boldsymbol{\beta}^T,\mathbf{u}^T]^T$. Applying (3) we find $q(\boldsymbol{\nu}) \propto \exp\{\mathbf{y}^T\mathbf{C}\boldsymbol{\nu} - \mathbf{1}_n^T b(\mathbf{C}\boldsymbol{\nu}) - \frac{1}{2}\boldsymbol{\nu}^T\mathbf{B}\boldsymbol{\nu}\}$ and $q(\tau) = \text{Gamma}\{s + \frac{m}{2}, r + \frac{1}{2}\big\langle\|\mathbf{u}\|^2\big\rangle\}$ where $b(x) = \log(1+e^x)$, $\mathbf{B} = \text{blockdiag}(\sigma^{-2}\mathbf{I}, \langle\tau\rangle\mathbf{I})$, $\mathbf{y} = [y_{11}, \ldots, y_{1n_1}, y_{21}, \ldots, y_{mn_m}]^T$ and

$$\mathbf{C} = \begin{bmatrix} 1 & \texttt{drugLo}_{11} & \texttt{drugHi}_{11} & \texttt{week}_{11} & 1 & 0 & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & \texttt{drugLo}_{1n_1} & \texttt{drugHi}_{1n_1} & \texttt{week}_{1n_1} & 1 & 0 & \cdots & 0 \\ 1 & \texttt{drugLo}_{21} & \texttt{drugHi}_{21} & \texttt{week}_{21} & 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & \texttt{drugLo}_{mn_m} & \texttt{drugHi}_{mn_m} & \texttt{week}_{mn_m} & 0 & 0 & \cdots & 1 \end{bmatrix}.$$

Unlike the previous examples not all q-densities, in this case $q(\boldsymbol{\nu})$, take the form of known parametric densities. Instead, using the optimal q-densities as a guide, we calculate the lower bound $\log\underline{p}(\mathbf{y};\boldsymbol{\xi})$ on $\log p(\mathbf{y})$ via (4) with $q(\boldsymbol{\nu}) = N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ and $q(\tau) = \text{Gamma}(S, R)$ in order to obtain

$$\begin{aligned} \log\underline{p}(\mathbf{y};\boldsymbol{\xi}) &= \mathbf{y}^T\mathbf{C}\big\langle\boldsymbol{\nu}\big\rangle - \mathbf{1}_n^T\big\langle b(\mathbf{C}\boldsymbol{\nu})\big\rangle - \tfrac{1}{2}\text{tr}\left(\mathbf{B}\big\langle\boldsymbol{\nu}\boldsymbol{\nu}^T\big\rangle\right) - \tfrac{m+p}{2}\log(2\pi) - \tfrac{p}{2}\log(\sigma^2) + \tfrac{m}{2}\big\langle\log(\tau)\big\rangle \\ &\quad + s\log(r) - \log\Gamma(s) + (s-1)\,\langle\log(\tau)\rangle - r\big\langle\tau\big\rangle - \big\langle\log q(\boldsymbol{\nu})\big\rangle - \big\langle\log q(\tau)\big\rangle \end{aligned} \tag{14}$$

where $\boldsymbol{\xi} = (\boldsymbol{\mu}, \text{vech}(\boldsymbol{\Sigma}), S, R)$ are additional variational parameters. Note that the calculation of $\big\langle b(\mathbf{C}\boldsymbol{\nu})\big\rangle$ and its derivatives requires approximation of one dimensional integrals, the details of which are summarized in the Appendix. We call $q(\boldsymbol{\nu}) = N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ the Gaussian Variational Approximation of $p(\boldsymbol{\nu}|\mathbf{y})$.

Note that $p(\mathbf{y}) \geq \underline{p}(\mathbf{y};\boldsymbol{\xi})$ for all $\boldsymbol{\xi}$. Hence, maximizing $\underline{p}(\mathbf{y};\boldsymbol{\xi})$ with respect to $\boldsymbol{\xi}$ and reduces the gap between $p(\mathbf{y})$ and $\underline{p}(\mathbf{y};\boldsymbol{\xi})$. Differentiating $\log\underline{p} \equiv \log\underline{p}(\mathbf{y};\boldsymbol{\xi})$ with respect to $\boldsymbol{\xi}$ we obtain

$$\begin{aligned} \mathsf{D}_{\boldsymbol{\mu}}\log\underline{p} &= \mathbf{C}^T\left[\mathbf{y} - \big\langle b'(\mathbf{C}\boldsymbol{\nu})\big\rangle\right] - \mathbf{B}\boldsymbol{\mu} \\ \mathsf{D}_{\text{vech}(\boldsymbol{\Sigma})}\log\underline{p} &= \tfrac{1}{2}\text{vec}\big[\boldsymbol{\Sigma}^{-1} - \mathbf{C}^T\text{diag}(\big\langle b''(\mathbf{C}\boldsymbol{\nu})\big\rangle)\mathbf{C} - \mathbf{B}\big]^T\mathbf{D}_m. \end{aligned} \tag{15}$$

The Hessian matrix corresponding to $\boldsymbol{\mu}$ is $\mathsf{H}_{\boldsymbol{\mu}\boldsymbol{\mu}}\log\underline{p} = -\mathbf{C}^T\text{diag}(\big\langle b''(\mathbf{C}\boldsymbol{\nu})\big\rangle)\mathbf{C} - \mathbf{B}$. We see from the second line of (15) that the first order optimality conditions for $\boldsymbol{\Sigma}$ are satisfied when $\boldsymbol{\Sigma} = [\mathbf{C}^T\text{diag}(\big\langle b''(\mathbf{C}\boldsymbol{\nu})\big\rangle)\mathbf{C} + \mathbf{B}]^{-1}$. We use this condition as a fixed point update for $\boldsymbol{\Sigma}$ by setting $\boldsymbol{\Sigma}$ to $[\mathbf{C}^T\text{diag}(\big\langle b''(\mathbf{C}\boldsymbol{\nu})\big\rangle)\mathbf{C} + \mathbf{B}]^{-1}$ at the beginning of each iteration. This guarantees that $\boldsymbol{\Sigma}$ is symmetric positive definite and can then be used in a Newton-Raphson update of $\boldsymbol{\mu}$. The optimal values for S and R are the same as those for the PVA. This leads to Algorithm 2 below.

Algorithm 2: Variational Approximation for the Logistic Random Intercept Model

```
Initialize $\boldsymbol{\mu}$, $\boldsymbol{\Sigma}$, $S = s + \frac{m}{2}$ and $R$.
Cycle:
    $\boldsymbol{\Sigma} \leftarrow \left[\mathbf{C}^T\text{diag}(\langle b''(\mathbf{C}\boldsymbol{\nu})\rangle)\mathbf{C} + \mathbf{B}\right]^{-1}$
    $\boldsymbol{\mu} \leftarrow \boldsymbol{\mu} + \boldsymbol{\Sigma}\left\{\mathbf{C}^T\left[\mathbf{y} - \langle b'(\mathbf{C}\boldsymbol{\nu})\rangle\right] - \mathbf{B}\boldsymbol{\mu}\right\}$;    $R \leftarrow r + \frac{1}{2}\langle\|\mathbf{u}\|^2\rangle$
until the change in $\underline{p}(\mathbf{y};\boldsymbol{\xi})$ becomes negligible.
```

Note that GVA is similar in spirit to the Laplace's method, however, as noted in Bishop (2006) and Ormerod & Wand (2010b), empirical evidence suggests that GVA is typically better at approximating posterior means than Laplace's method. This in turn suggests that a GBVA based on a GVA may perform better than INLA in this context.

6.2. Grid Based Variational Approximations

The GBVA for $p(\tau|\mathbf{y})$ requires only a little modification of Algorithm 2. In this case we treat τ as observed so that we do not apply the update for R. Furthermore, the only modification required to calculate $\log\underline{p}(\mathbf{y}, \tau;\boldsymbol{\xi})$ is to omit the term $\big\langle\log q(\tau)\big\rangle$ in (14).

Finally, consider the GBVA to $p(\beta_i|\mathbf{y})$ where we approximate $p(\boldsymbol{\beta}_{-i}|\mathbf{y}, \beta_i)$ by $q(\boldsymbol{\beta}_{-i})$, a multivariate Gaussian density with mean $\boldsymbol{\mu}$ and covariance $\boldsymbol{\Sigma}$. Using (4) with $\mathcal{H} = \{\boldsymbol{\beta}_{-i}\}$ and $\mathcal{E} = \{\mathbf{y}, \beta_i\}$ we can derive $\underline{p}(\mathbf{y}, \beta_i;\boldsymbol{\xi})$, a lower bound on $p(\mathbf{y}, \beta_i)$, is given by (14) where we replace $\log\underline{p}(\mathbf{y};\boldsymbol{\xi})$ with $\log\underline{p}(\mathbf{y}, \beta_i;\boldsymbol{\xi})$ and treat the appropriate element of $\boldsymbol{\nu}$ as observed. Following a similar line or argument to that described in Section 6.1 using a fixed value for β_i leads to Algorithm 2 with $\mathbf{C}$ replaced with $\mathbf{C}_{-i}$ and $\mathbf{B}$ replaced with $\mathbf{B}_{-i}$.

The final value for $\underline{p}(\mathbf{y}, \beta_i;\boldsymbol{\xi})$ obtained from Algorithm 3 is used as an approximation to $p(\mathbf{y}, \beta_i)$ which can be used as part of a grid-based approximation as described in Section 2.

6.3. *Comparisons*

For the `bacteria` dataset described in Section 6.1 the main parameters of interest are the β_is and τ. Figure 7 illustrates the approximate posterior densities and values for these parameters based on VA (based on the variational approximation described in Section 6.1), MCMC (based on 10^5 samples), GBVA (based on the grid-based variational approximations described in Section 6.2 with 10 grid points per parameter) and the `INLA-R` implementation of the INLA method. Table 1 displays the ISEs for the parameters of interest based on VA, GBVA and INLA methods and the times that each of these methods took.

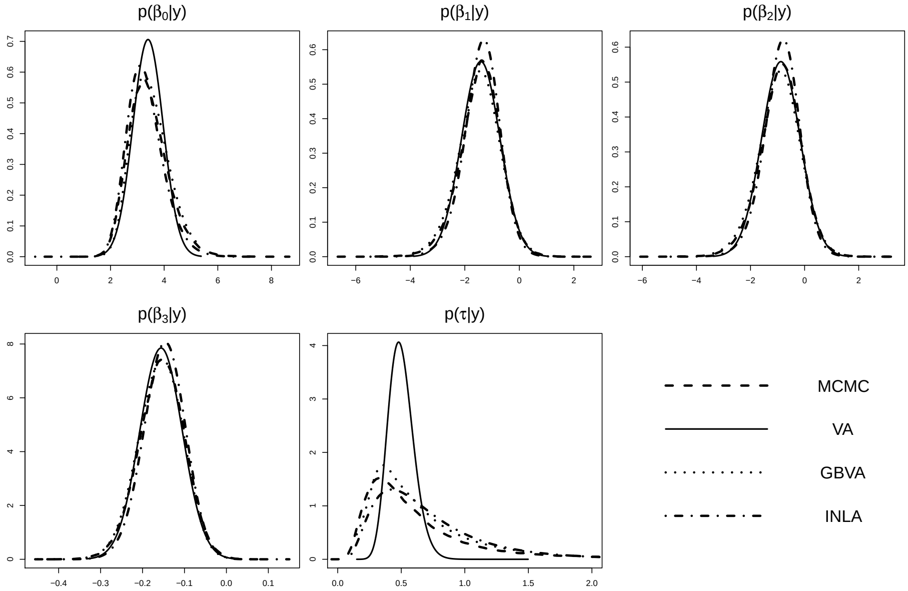

Figure 7: *Approximate posterior densities based on VA, MCMC, GBVA and INLA methods for* $\beta_0, \ldots, \beta_3$ *and* τ *for model (10) with setting described in Section 5.3.*

Method	β_0	β_1	β_2	β_3	τ	Time (s)
VA	0.023	0.002	0.001	0.025	1.396	0.09
GBVA	0.003	0.002	0.001	0.008	0.029	7.08
INLA	0.006	0.004	0.005	0.074	0.057	1.95
MCMC (10^3)	0.004	0.003	0.001	0.018	0.008	43.16

Table 1: Integrated square error (ISE) values based on MCMC with 10^5 samples for the main parameters of interest and times for the random intercept model of the `Bacteria` dataset using the VA, GBVA, INLA and MCMC methods.

From Table 1 we note that GBVA offers a slight improvement over both VA and INLA for all of the β_is, although each of the methods is quite accurate for these parameters. The GBVA method significantly improves over VA for the τ parameter, but perhaps only offers a slight improvement over INLA for this parameter. This suggests that GBVA may offer a slight improvement over INLA in an acceptable amount of time. Furthermore, if $q(\beta_i)$s is fit using VA and $q(\tau)$ is fit using GBVA the total time taken is 0.4 seconds. This combination gives a slight improvement in accuracy over INLA and less than half the time with little loss of accuracy when compared to INLA. Finally, we see that there is a speed versus accuracy tradeoff when comparing these methods to MCMC. While the VA, GBVA and INLA methods are faster, for this particular example, they do not offer an advantage in terms of accuracy when compared to the MCMC method with 10^3 samples.

The reduced accuracy of both approximations of $p(\tau|\mathbf{y})$ stems from the fact that both methods use similar but different Gaussian approximations to $p(u_i|\mathbf{y})$. Ormerod & Wand (2010b) demonstrated that for a frequentist model of a similar dataset that some of the $p(u_i|\mathbf{y})$s are severely skewed so that a Gaussian approximation is of the $p(u_i|\mathbf{y})$s is not sufficiently accurate. A similar inaccuracy in the precision of INLA has also been encountered in similar models by Fong, Rue & Wakefield (2010).

The GBVA and INLA methods for this example are very similar but differ in some important aspects which account for GBVA being about four times slower than INLA. The fact that INLA is faster than GBVA stems from three main differences between the two methods. Firstly, GBVA was programmed in `R` while `INLA`, implemented in the low-level programming language `C`. Secondly, Algorithm 2 requires $3\sum_{i=1}^m n_i$ one-dimensional integral approximations each iteration.

7. Discussion

This paper describes a new method called grid-based variational approximations for calculating approximate marginal posterior densities. We have shown that the method is extremely fast and reasonably accurate on all of the models considered in this paper, and is a particularly attractive alternative to MCMC methods, especially when only a few marginal posterior densities are of interest and the model cannot be fit using INLA.

Several modifications can be envisaged for superior performance. Firstly, grid points could be chosen adaptively. If the grid could be chosen in such a way that as few as possible grid points are chosen then computational times might be greatly reduced. Secondly, the algorithm could be modified for approximation of higher dimensional marginal posterior densities. Such a modification, perhaps along the lines of the INLA method of Rue *et al.* (2009), might also yield superior speed and accuracy. These modifications could lead to an attractive alternative method to INLA for problems where the INLA method cannot be used.

Appendix

Expectations with respect to the q-densities are denoted by $\langle\boldsymbol{\theta}\rangle$ in this paper, for some vector $\boldsymbol{\theta}$. In an effort to reduce notation we summarize the expectations appearing in this paper here. Note that if $\boldsymbol{\theta}$ is observed, i.e. $\boldsymbol{\theta}\in\mathcal{E}$, then $\langle f(\boldsymbol{\theta})\rangle = f(\boldsymbol{\theta})$ for any vector valued function f, otherwise, i.e. when $\boldsymbol{\theta}\in\mathcal{H}$, the expectations are summarized below.

If $q(\boldsymbol{\nu}) = N(\boldsymbol{\mu},\boldsymbol{\Sigma})$ then $\big\langle\boldsymbol{\nu}\big\rangle = \boldsymbol{\mu}$ and $\big\langle\|\boldsymbol{\nu}\|\big\rangle = \|\boldsymbol{\mu}\| + \text{tr}(\boldsymbol{\Sigma})$. Similarly, if $q(\boldsymbol{\nu}_{-i}) = N(\boldsymbol{\mu},\boldsymbol{\Sigma})$ the ith element of $\big\langle\boldsymbol{\nu}\big\rangle$ is ν_i and the remaining elements are $\boldsymbol{\mu}$ and $\big\langle\|\boldsymbol{\nu}\|\big\rangle = \nu_i^2 + \|\boldsymbol{\mu}\| + \text{tr}(\boldsymbol{\Sigma})$. If $q(\sigma^2)\sim\text{IG}(S,R)$ then $\big\langle\sigma^{-2}\big\rangle = S/R$ and $\big\langle\log(\sigma^2)\big\rangle = \log(R)-\psi(S)$ where $\psi(x) = [d\log\Gamma(t)/dt]_{t=x}$ is the digamma function and $\Gamma(x)$ is the gamma function (see Abramowitz & Stegun, 1964, Chapter 6 for details). Let $\tau\sim\text{Gamma}(S,R)$ then $\big\langle\tau\big\rangle = S/R$ and $\big\langle\log(\tau)\big\rangle = \psi(S)-\log(R)$. Let $\rho\sim\text{Beta}(\alpha,\beta)$ then $\big\langle\rho\big\rangle = \frac{\alpha}{\alpha+\beta}$, $\big\langle\log(\rho)\big\rangle = \psi(\alpha)-\psi(\alpha+\beta)$ and $\big\langle\log(1-\rho)\big\rangle = \psi(\beta)-\psi(\alpha+\beta)$. Finally, if $b_i\sim\text{Bernoulli}(\eta)$ then $\big\langle b_i\big\rangle = \eta$.

Shannon's entropy (or simply entropy) of a density $q(\boldsymbol{\theta})$ is defined as $-\int q(\boldsymbol{\theta})\log q(\boldsymbol{\theta})d\boldsymbol{\theta} = -\langle\log q(\boldsymbol{\theta})\rangle$ where larger entropy values characterize great randomness of a random vector with density $q(\boldsymbol{\theta})$. These entropies are frequently used in this paper to calculate values of $\log\underline{p}(\mathcal{E})$ for different configurations of $\mathcal{E}$. Nadarajah & Zografos (2003) contains entropy expressions for all the distributions used in this paper.

In Section 6.1 expectations of the form $\big\langle b^{(r)}(\mathbf{C}\boldsymbol{\nu})\big\rangle$, $r=0,1,2$, were encountered where $b(x) = \log(1+\exp(x))$ and $q(\boldsymbol{\nu}) = N(\boldsymbol{\mu},\boldsymbol{\Sigma})$ or $q(\boldsymbol{\nu}_{-i}) = N(\boldsymbol{\mu},\boldsymbol{\Sigma})$. Firstly let,

$$B^{(r)}(\mu,\sigma^2) = \int b^{(r)}(x)(2\pi\sigma^2)^{-1/2}\exp\left\{-\tfrac{1}{2\sigma^2}(x-\mu)^2\right\}dx.$$

Then it is fairly easy to show that $\big\langle b^{(r)}(\mathbf{c}^T\boldsymbol{\nu})\big\rangle = B^{(r)}(\mathbf{c}^T\boldsymbol{\mu},\mathbf{c}^T\boldsymbol{\Sigma}\mathbf{c})$ if $q(\boldsymbol{\nu}) = N(\boldsymbol{\mu},\boldsymbol{\Sigma})$ and $\big\langle b^{(r)}(\mathbf{c}^T\boldsymbol{\nu})\big\rangle = B^{(r)}(c_i\nu_i + \mathbf{c}_{-i}^T\boldsymbol{\mu},\mathbf{c}_{-i}^T\boldsymbol{\Sigma}\mathbf{c}_{-i})$ if ν_i is fixed and $q(\boldsymbol{\nu}_{-i}) = N(\boldsymbol{\mu},\boldsymbol{\Sigma})$. Secondly, note that $\frac{\partial}{\partial\mu}B^{(r)}(\mu,\sigma^2) = B^{(r+1)}(\mu,\sigma^2)$ and $\frac{\partial}{\partial\sigma^2}B^{(r)}(\mu,\sigma^2) = \frac{1}{2}B^{(r+2)}(\mu,\sigma^2)$. Gauss-Hermite quadrature can then be use to approximate $B^{(r)}(\mu,\sigma^2)$ (see Ormerod & Wand, 2010b, for details).

Acknowledgment

This research was partially supported by Australian Research Council Discovery Project DP0877055.

References

Barber, D., Bishop, C.M., 1998. Ensemble learning for multi-layer networks. In Jordan, M.I. Kearns, K.J. and Solla, S.A. (Eds.), Advances in Neural Information Processing Systems 10, 395–401.

Bishop, C.M., 2006. Pattern Recognition and Machine Learning. New York: Springer.

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

Consonni, G., Marin, J.-M. 2007. Mean-field variational approximate Bayesian inference for latent variable models. Computational Statistics and Data Analysis 52, 790–798.

Fong Y., Rue H., Wakefield J., 2010. Bayesian inference for Generalized Linear Mixed Models. Biostatistics, (to appear).

Hall, P., Humphreys, K., Titterington, D.M., 2002. On the adequacy of variational lower bound functions for likelihood-based inference in Markovian models with missing values. Journal of the Royal Statistical Society Series B 64, 549–564.

Hall, P., Ormerod, J.T., Wand, M.P. 2010. Theory of Gaussian variational approximation for a generalised linear mixed model. Statistica Sinica, (to appear).

Jaakkola, T.S., Jordan, M.I. 2000. Bayesian parameter estimation via variational methods. Statistics and Computing 10, 25–37.

Jordan, M.I., Ghahramani, Z., Jaakkola, T.S., Saul, L.K. (1999). An introduction to variational methods for graphical models. Machine Learning 37, 183–233.

Kullback, S., Leibler, R.A., 1951. On information and sufficiency. The Annals of Mathematical Statistics 22, 79–86.

Leach, A., 2000. Menzies School of Health Research 1999–2000 Annual Report, pp. 18–21.

Ligges, U., Thomas, A., Spiegelhalter, D., Best, N., Lunn, D., Rice, K., Sturtz, S., 2007. OpenBUGS and its R/S-PLUS interface BRugs. BRugs 0.4-2.

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.

Martino, S., Rue, H., 2009. Implementing approximate Bayesian inference using Integrated Nested Laplace Approximation: A manual for the inla program.
Available from: http://www.math.ntnu.no/˜hrue/GMRFLib.

Minka, T., Winn, J., Guiver, G., Kannan, A., 2008. Infer.Net. Microsoft Research Cambridge, Cambridge, UK.

Nadarajah, S., Zografos, K., 2003. Formulas for Renyi information and related measures for univariate distributions. Information Sciences 55, 119–138.

Opper, M., Archambeau, C., 2009. Variational Gaussian approximation revisited. Neural Computation 21, 786–792.

Ormerod, J.T., Wand, M.P., 2008. Variational approximations for logistic mixed models. Proceedings of the Ninth Iranian Statistics Conference, Department of Statistics, University of Isfahan, Isfahan, Iran, pp. 450–467.

Ormerod, J.T., Wand, M.P., 2009. Comment on paper by Rue, Martino and Chopin. Journal of the Royal Statistical Society, Series B 71, 377–378.

Ormerod, J.T. Wand, M.P. 2010a. Explaining variational approximations. The American Statistician, (to appear).

Ormerod, J.T. Wand, M.P., 2010b. Gaussian variational approximate inference for generalized linear mixed models. Unpublished manuscript.

R Development Core Team, 2008. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3–900051–07–0. http://www.R-project.org.

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 7, 319–392.

Seeger, M., 2000. Bayesian Model Selection for Support Vector Machines, Gaussian Processes and Other Kernel Classifiers. Neural Information Processing Systems 12, 603–609.

Venables, W.N., Ripley, B.D., 2009. MASS: functions and datasets to support Venables and Ripley, 'Modern Applied Statistics with S' (4th edition). R package version 7.2-48.

Wang, B., Titterington, D.M., 2006. Convergence properties of a general algorithm for calculating variational Bayesian estimates for a normal mixture model. Bayesian Analysis 1, 625–650.

Winn, J., Bishop, C.M., 2005. Variational message passing. Journal of Machine Learning Research 6, 661–694.