Stat

The ISI's Journal for the Rapid
Dissemination of Statistics Research

(wileyonlinelibrary.com) DOI: 10.1002/sta4.18

Variational inference for marginal longitudinal semiparametric regression

Marianne Menictas[a] **and Matt P. Wand**[a,*]

Received 10 December 2012; Accepted 3 January 2013

We derive a variational inference procedure for approximate Bayesian inference in marginal longitudinal semiparametric regression. Fitting and inference is much faster than existing Markov chain Monte Carlo approaches. Numerical studies indicate that the new methodology is very accurate for the class of models under consideration.
Copyright © 2013 John Wiley & Sons, Ltd.

Keywords: Bayesian computing; longitudinal data analysis; mean field variational Bayes

1 Introduction

We study marginal longitudinal semiparametric regression analysis and develop variational approximation methods that enable significantly faster fitting and inference with little degradation of accuracy.

The defining features of marginal longitudinal semiparametric regression analysis are (a) flexible estimation of regression functions, and (b) marginal estimation of the covariance matrix of the response vector for each subject. There are numerous contributions to the topic. We focus on the Bayesian penalized spline approach of Al Kadiri, Carroll & Wand (2010). References to related methodology are provided there.

Al Kadiri et al. (2010) settle upon Markov chain Monte Carlo (MCMC) as a convenient and effective means of fitting their class of models. However, they admit that such an approach is quite slow. Their real data example takes almost an hour to run on a contemporary laptop when using BUGS (Spiegelhalter et al., 2003) for the MCMC. The variational algorithm developed in the present article, based on the mean field variational Bayes (MFVB) paradigm, fits the data in seconds with very similar results. The algorithm is also one of the first variational algorithms involving estimation of an unstructured covariance matrix.

Section 2 provides a brief recap of the marginal longitudinal semiparametric regression models of Al Kadiri et al. (2010). A variational inference algorithm for a Bayesian hierarchical version of the models is developed in Section 3. Section 4 provides simulation evidence of impressive speed and accuracy attributes of MFVB for the Al Kadiri et al. (2010) marginal longitudinal semiparametric regression models. An appendix provides some algebraic details tied to the new methodology.

[a]School of Mathematical Sciences, University of Technology, Sydney, P.O. Box 123, Broadway, 2007, Australia
*Email: matt.wand@uts.edu.au

Copyright © 2013 John Wiley & Sons Ltd

2 Model description

Al Kadiri et al. (2010) developed a class of marginal longitudinal semiparametric regression models based on penalized splines. An example is the additive model

$$E(y_{ij}|\boldsymbol{u}) = \beta_0 + f_1(x_{1ij}) + f_2(x_{2ij}),\ \mathrm{Cov}(\boldsymbol{y}_i|\boldsymbol{u}) = \boldsymbol{\Sigma},\ 1 \leqslant i \leqslant m,\ 1 \leqslant j \leqslant n \tag{1}$$

where $(x_{1ij}, x_{2ij}, y_{ij})$ denotes the jth measurement for subject i of the predictor/response triple (x_1, x_2, y). Also, $\boldsymbol{y}_i$ denotes the $n \times 1$ vector containing the responses for subject i and $\boldsymbol{u}$ contains all spline coefficients that, as explained below, are random effects. Here f_1 and f_2 are smooth functions of predictors x_1 and x_2. Penalized spline models for f_1 and f_2 take the form

$$f_1(x_1) = \beta_{11}x_1 + \sum_{k=1}^{K_1} u_{1k}z_{1k}(x_1) \quad \text{and} \quad f_2(x_2) = \beta_{21}x_2 + \sum_{k=1}^{K_2} u_{2k}z_{2k}(x_2), \tag{2}$$

where the z_{1k} and z_{2k} are appropriate spline basis functions and their coefficients are independently subject to

$$u_{1k} \text{ i.i.d. } N\left(0, \sigma_1^2\right) \quad \text{and} \quad u_{2k} \text{ i.i.d. } N\left(0, \sigma_2^2\right).$$

Sections 2.1, 2.2 and 2.3 in Al Kadiri et al. (2010) provide semiparametric variants of model (1). Each of these marginal longitudinal semiparametric regression models, and their extensions to d smooth functions, can be handled using the following Gaussian linear mixed model:

$$\boldsymbol{y}\,|\,\boldsymbol{u} \sim N\left(\boldsymbol{X\beta} + \boldsymbol{Zu},\ \boldsymbol{I}_m \otimes \boldsymbol{\Sigma}\right), \quad \boldsymbol{u}\,|\,\boldsymbol{\sigma}^2 \sim N\left(\boldsymbol{0}, \underset{1 \leqslant \ell \leqslant d}{\text{blockdiag}}\left(\sigma_\ell^2\, \boldsymbol{I}_{K_\ell}\right)\right), \tag{3}$$

where $\boldsymbol{\sigma}^2 \equiv \left(\sigma_1^2, \ldots, \sigma_d^2\right)$ and K_ℓ corresponds to the number of spline basis functions used in the ℓth smooth function estimate. Also, let $\boldsymbol{u}_\ell$ denote the $K_\ell \times 1$ sub-vector of $\boldsymbol{u}$ corresponding to the ℓth block in (3). Then $\boldsymbol{u}_\ell|\sigma_\ell^2 \sim N\left(\boldsymbol{0}, \sigma_\ell^2\, \boldsymbol{I}_{K_\ell}\right)$ for $1 \leqslant \ell \leqslant d$.

As pointed out in Al Kadiri et al. (2010), (3) is either difficult or impossible to fit using standard mixed model software so the authors opt for a Bayesian approach and employ Markov chain Monte Carlo for achieving approximate fitting and inference. We take a Bayesian approach here as well, although we use less informative prior distributions for the variance components, σ_ℓ^2, $1 \leqslant \ell \leqslant d$, and the covariance matrix $\boldsymbol{\Sigma}$. Specifically, we impose Half-t priors on the σ_ℓ and the square-rooted diagonal entries of $\boldsymbol{\Sigma}$. Our prior specification also allows for correlation parameters within $\boldsymbol{\Sigma}$ to have uniform distributions on $(-1, 1)$. Such priors are achieved via auxiliary variable constructions embodied in Result 5 of Wand et al. (2011), Proposition 1 of Armagan et al. (2011) and Properties 2–4 of Huang & Wand (2013).

In full, the Bayesian hierarchical model that we treat here is

$$\begin{gathered}
\boldsymbol{y}\,|\,\boldsymbol{\beta}, \boldsymbol{u}, \boldsymbol{\Sigma} \sim N(\boldsymbol{X}\,\boldsymbol{\beta} + \boldsymbol{Z}\,\boldsymbol{u}, \boldsymbol{I}_m \otimes \boldsymbol{\Sigma}), \quad \boldsymbol{u}\,|\,\boldsymbol{\sigma}^2 \sim N\left(\boldsymbol{0}, \underset{1 \leqslant \ell \leqslant d}{\text{blockdiag}}\left(\sigma_\ell^2\, \boldsymbol{I}_{K_\ell}\right)\right) \\
{\sigma^2}_\ell|\,a_\ell \overset{\text{ind.}}{\sim} \text{Inverse-Gamma}\left(\frac{1}{2}, 1/a_\ell\right), \quad a_\ell \overset{\text{ind.}}{\sim} \text{Inverse-Gamma}\left(\frac{1}{2}, 1/{A_\ell}^2\right), \quad 1 \leqslant \ell \leqslant d, \\
\boldsymbol{\Sigma}\,|\,a_{\Sigma,1}, \ldots, a_{\Sigma,n} \sim \text{Inverse-Wishart}\left(\nu + n - 1, 2\nu\, \text{diag}\left(1/a_{\Sigma,1}, \ldots, 1/a_{\Sigma,n}\right)\right), \\
a_{\Sigma,j} \overset{\text{ind.}}{\sim} \text{Inverse-Gamma}\left(\frac{1}{2}, 1/A_{\Sigma,j}^2\right),\ 1 \leqslant j \leqslant n,\ \text{and } \boldsymbol{\beta} \sim N\left(\boldsymbol{0}, \sigma_\beta^2 \boldsymbol{I}_p\right)
\end{gathered} \tag{4}$$

(wileyonlinelibrary.com) DOI: 10.1002/sta4.18

with $\overset{\text{ind.}}{\sim}$ denoting "distributed independently as" and p denoting the dimension of $\boldsymbol{\beta}$. The auxiliary variables that impose the aforementioned noninformative priors on $\boldsymbol{\sigma}^2$ and $\boldsymbol{\Sigma}$ are denoted, respectively, by a_ℓ and $a_{\Sigma,j}$. The notation $x \sim \text{Inverse-Gamma}(A, B)$ for $A, B > 0$ means that the random variable x has density function

$$p(x) = \Gamma(A)^{-1} B^A x^{-A-1} \exp(-B/x), \quad x > 0.$$

If $\boldsymbol{M}$ is an $n \times n$ random matrix then the notation $\boldsymbol{M} \sim \text{Inverse-Wishart}(A, \boldsymbol{B})$, for $A > 0$ and $\boldsymbol{B}$ positive definite, means that

$$p(\boldsymbol{M}) = C_{n,A}^{-1} |\boldsymbol{B}|^{\frac{A}{2}} |\boldsymbol{M}|^{-(A+n+1)/2} \exp\left\{-\frac{1}{2}\,\text{tr}(\boldsymbol{B}\boldsymbol{M}^{-1})\right\}, \quad \boldsymbol{M} \text{ positive definite} \qquad (5)$$

where $C_{n,A} \equiv 2^{An/2}\, \pi^{n(n-1)/4} \prod_{i=1}^{n} \Gamma\left(\frac{A+1-i}{2}\right)$.

Lastly, we note that (4) entails the setting of hyperparameters ν, σ_β, $A_1, \ldots, A_d$ and $A_{\Sigma,1}, \ldots, A_{\Sigma,n}$, all of which are constrained to be positive. We recommend setting $\nu = 2$ so that Property 3 of Huang & Wand (2013) applies. For the remaining hyperparameters, non-informativity corresponds to them being set to very large values. Assuming that the data have been transformed to have mean zero and unit standard deviation, our default setting is 10^5.

We close this section with Figure 1, a directed acyclic graph representation of model (4). Here, $\boldsymbol{a}$ is the vector containing the a_ℓ and $\boldsymbol{a}_{\boldsymbol{\Sigma}}$ is the vector containing the $a_{\Sigma,j}$. Such graphical representation is quite useful. Apart from aiding digestion of (4), it assists the calculations required for mean field variational approximate inference – the topic of the next section.

3 Variational inference algorithm

The essence of our MFVB approach to variational inference is to replace the joint posterior density function of the hidden nodes (i.e. parameters, random effects and auxiliary variables) in Figure 1:

$$p\left(\boldsymbol{\beta}, \boldsymbol{u}, \boldsymbol{a}, \boldsymbol{\sigma}^2, \boldsymbol{\Sigma}, \boldsymbol{a}_\Sigma | \boldsymbol{y}\right)$$

with an approximating density function $q(\boldsymbol{\beta}, \boldsymbol{u}, \boldsymbol{a}, \boldsymbol{\sigma}^2, \boldsymbol{\Sigma}, \boldsymbol{a}_\Sigma)$ that assumes the product density form

$$q\left(\boldsymbol{\beta}, \boldsymbol{u}, \boldsymbol{a}, \boldsymbol{\sigma}^2, \boldsymbol{\Sigma}, \boldsymbol{a}_\Sigma\right) = q(\boldsymbol{\beta}, \boldsymbol{u}, \boldsymbol{a}, \boldsymbol{a}_\Sigma)\, q(\boldsymbol{\sigma}^2, \boldsymbol{\Sigma}). \qquad (6)$$

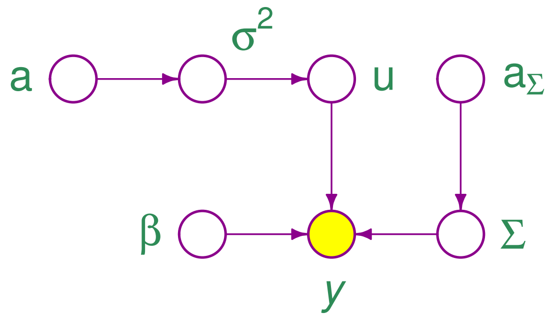

Figure 1. Directed acyclic graph corresponding to the Bayesian hierarchical model (4). The shaded node corresponds to the observed data.

Copyright © 2013 John Wiley & Sons Ltd

The q-densities are then chosen to minimize the Kullback-Leibler distance between the two density functions:

$$\int \log \left\{ \frac{p\left(\boldsymbol{\beta}, \boldsymbol{u}, \boldsymbol{a}, \boldsymbol{a}_{\Sigma}, \boldsymbol{\sigma}^2, \boldsymbol{\Sigma} | \boldsymbol{y}\right)}{q\left(\boldsymbol{\beta}, \boldsymbol{u}, \boldsymbol{a}, \boldsymbol{a}_{\Sigma}, \boldsymbol{\sigma}^2, \boldsymbol{\Sigma}\right)} \right\} q(\boldsymbol{\beta}, \boldsymbol{u}, \boldsymbol{a}, \boldsymbol{a}_{\Sigma}, \boldsymbol{\sigma}^2, \boldsymbol{\Sigma})\, d\left(\boldsymbol{\beta}, \boldsymbol{u}, \boldsymbol{a}, \boldsymbol{a}_{\Sigma}, \boldsymbol{\sigma}^2, \boldsymbol{\Sigma}\right)$$

subject to (6). As explained in, for example, Chapter 10 of Bishop (2006) and Section 2.2 of Ormerod & Wand (2010) the optimal q-densities, denoted here by q^*, can be obtained via a coordinate ascent algorithm that arises from relationships such as

$$q^*(\boldsymbol{\Sigma}) \propto \exp E_q\{\log p(\boldsymbol{\Sigma}|\text{rest})\}$$

where 'rest' denotes all of the random variables except for $\boldsymbol{\Sigma}$. The graphical structure conveyed by Figure 1 can simplify the requisite calculations since

$$p(\boldsymbol{\Sigma}|\text{rest}) = p(\boldsymbol{\Sigma}|\text{Markov blanket of } \boldsymbol{\Sigma}) = p(\boldsymbol{\Sigma}|\boldsymbol{\beta}, \boldsymbol{u}, \boldsymbol{a}_{\Sigma}, \boldsymbol{y})$$

leads to a more localized calculation around $\boldsymbol{\Sigma}$. The Markov blanket of a node on a directed acyclic graph is the set of children, parents and co-parents (parents sharing the same child) of the node. Details are given in an appendix.

The graphical structure can also be used to determine additional factorizations, called *induced factorizations*, in the MFVB solution (Bishop, 2006, Section 10.2.5). Such induced factorizations can be detected using simple graphical tests such as *d-separation*. These lead to the q^*-densities satisfying

$$q^*(\boldsymbol{\beta}, \boldsymbol{u}, \boldsymbol{\sigma}^2, \boldsymbol{a}, \boldsymbol{a}_{\Sigma}, \boldsymbol{\Sigma}) = q^*(\boldsymbol{\beta}, \boldsymbol{u}) \left\{ \prod_{\ell=1}^{d} q^*\left(\sigma_\ell^2\right) q^*(a_\ell) \right\} \left\{ \prod_{j=1}^{n} q^*(a_{\Sigma,j}) \right\} q^*(\boldsymbol{\Sigma}). \qquad (7)$$

The factors on the right-hand side of (7) have the following forms:

$$\begin{aligned}
&q^*(\boldsymbol{\beta}, \boldsymbol{u}) \text{ is the } N(\boldsymbol{\mu}_{q(\boldsymbol{\beta},u)}, \boldsymbol{\Sigma}_{q(\boldsymbol{\beta},u)}) \text{ density function,}\\
&q^*\left(\sigma_\ell^2\right) \text{ is the Inverse-Gamma} \left(\tfrac{1}{2}(K_\ell + 1), B_{q(\sigma_\ell^2)}\right) \text{ density function,}\\
&q^*(a_\ell) \text{ is the Inverse-Gamma} \left(1, B_{q(a_\ell)}\right) \text{ density function,}\\
&q^*(a_{\Sigma,j}) \text{ is the Inverse-Gamma} \left(\tfrac{1}{2}(\nu + n), B_{q(a_{\Sigma,j})}\right) \text{ density function and}\\
&q^*(\boldsymbol{\Sigma}) \text{ is the Inverse-Wishart}(\nu + m + n - 1, \boldsymbol{B}_{q(\boldsymbol{\Sigma})}) \text{ density function,}
\end{aligned} \qquad (8)$$

for parameters $\boldsymbol{\mu}_{q(\boldsymbol{\beta},\boldsymbol{u})}$ and $\boldsymbol{\Sigma}_{q(\boldsymbol{\beta},\boldsymbol{u})}$, the mean vector and covariance matrix of $q^*(\boldsymbol{\beta}, \boldsymbol{u})$, $B_{q(\sigma_\ell^2)}$, the rate parameter of $q^*\left(\sigma_\ell^2\right)$, $B_{q(a_\ell)}$, the rate parameter of $q^*(a_\ell)$, $B_{q(a_{\Sigma,j})}$, the rate parameter of $q^*(a_{\Sigma,j})$ and $\boldsymbol{B}_{q(\boldsymbol{\Sigma})}$, the rate matrix of $q^*(\boldsymbol{\Sigma})$. Details on the derivation of (8) are given in an appendix. The parameters in these q^*-densities are obtained via Algorithm 1. The algorithm uses some additional notation. Firstly, $\boldsymbol{C} \equiv [\boldsymbol{X}\ \boldsymbol{Z}]$. Also, $\boldsymbol{y}_i$ denotes the $n \times 1$ vector containing the entries of $\boldsymbol{y}$ corresponding to the ith subject. An analogous definition applies to $\boldsymbol{C}_i$. Note that, in the special case of $n = 1$, Algorithm 1 matches Algorithm 3 of Wand & Ormerod (2011).

Copyright © 2013 John Wiley & Sons Ltd

Convergence in Algorithm 1 is assessed using the variational lower bound on the marginal log-likelihood, denoted by $\underline{p}\,(\boldsymbol{y};q)$, and admits the following explicit expression:

$$
\begin{aligned}
\log \underline{p}(\boldsymbol{y};q) = & -\frac{1}{2} n\,(\nu+n-1)\log(2\nu) - \frac{3}{2}\log(\pi) - \frac{1}{2} m \log(2\pi) - \frac{1}{2} p \log\left(\sigma^2_{\boldsymbol{\beta}}\right) \\
& - \frac{1}{2\sigma^2_{\boldsymbol{\beta}}}\,\{\|\,\mu_{q(\boldsymbol{\beta})}\,\|^2 + \mathrm{tr}(\boldsymbol{\Sigma}_{q(\boldsymbol{\beta})})\} + \frac{1}{2}\log|\boldsymbol{\Sigma}_{q(\boldsymbol{\beta},\boldsymbol{u})}| + \frac{1}{2}\left(\sum\nolimits_{\ell=1}^{d} K_\ell + p\right) \\
& + \sum\nolimits_{\ell=1}^{d} \log\Gamma\left(\frac{1}{2}(K_\ell+1)\right) - \frac{1}{2}\sum\nolimits_{\ell=1}^{d}(K_\ell+1)\log\left(B_{q(\sigma^2_\ell)}\right) - \sum\nolimits_{\ell=1}^{d}\log(A_\ell) \\
& - \sum\nolimits_{\ell=1}^{d}\log\left(B_{q(a_\ell)}\right) + \sum\nolimits_{\ell=1}^{d}\mu_{q(1/\sigma^2_\ell)}\mu_{q(1/a_\ell)} - \log(C_{n,\,\nu+n-1}) \\
& + \log(C_{n,\,\nu+m+n-1}) - \sum\nolimits_{j=1}^{n}\log(A_j) + \log\Gamma\left(\frac{1}{2}(\nu+n)\right) - \frac{1}{2}(\nu+n)\sum\nolimits_{j=1}^{n}\log\left(B_{q(a_{\Sigma,j})}\right) \\
& + \sum\nolimits_{j=1}^{n}\nu\left(M_{q(\boldsymbol{\Sigma}^{-1})}\right)_{jj}\mu_{q(1/a_{\Sigma,j})} - \frac{1}{2}(\nu+m+n-1)\log|\boldsymbol{B}_{q(\boldsymbol{\Sigma})}|.
\end{aligned}
\tag{9}
$$

Algorithm 1 *Mean field variational Bayes algorithm for the determination of the optimal parameters in $q^*(\boldsymbol{\beta},\boldsymbol{u})$, $q^*(\sigma^2_\ell)$, $q^*(a_\ell)$, $q^*(a_{\Sigma,j})$ and $q^*(\boldsymbol{\Sigma})$ from data modeled according to (4).*

```
Initialize: $\mu_{q(1/\sigma^2_\ell)} > 0,\ \ \mu_{q(1/a_\ell)} > 0,\ 1 \leqslant \ell \leqslant d,\ \ \boldsymbol{M}_{q(\boldsymbol{\Sigma}^{-1})}$ positive definite.
Cycle:
    $\boldsymbol{\Sigma}_{q(\boldsymbol{\beta},\boldsymbol{u})} \leftarrow \left\{\boldsymbol{C}^T\left(\mathbf{I}_m \otimes \mathbf{M}_{q(\boldsymbol{\Sigma}^{-1})}\right)\boldsymbol{C} + \mathrm{blockdiag}\left(\sigma_{\boldsymbol{\beta}}^{-2}\boldsymbol{I}_p, \mu_{q(1/\sigma^2_1)}\boldsymbol{I}_{K_1}, \ldots, \mu_{q(1/\sigma^2_d)}\boldsymbol{I}_{K_d}\right)\right\}^{-1}$
    $\boldsymbol{\mu}_{q(\boldsymbol{\beta},\boldsymbol{u})} \leftarrow \boldsymbol{\Sigma}_{q(\boldsymbol{\beta},\boldsymbol{u})}\boldsymbol{C}^T\left(\boldsymbol{I}_m \otimes \boldsymbol{M}_{q(\boldsymbol{\Sigma}^{-1})}\right)\boldsymbol{y}$
    For $\ell = 1,\ldots,d$:
        $B_{q(a_\ell)} \leftarrow \mu_{q(1/\sigma^2_\ell)} + A_\ell^{-2}$   ;   $\mu_{q(1/a_\ell)} \leftarrow 1/B_{q(a_\ell)}$
        $B_{q(\sigma^2_\ell)} \leftarrow \frac{1}{2}\left(\|\boldsymbol{\mu}_{q(\boldsymbol{u}_\ell)}\|^2 + \mathrm{tr}(\boldsymbol{\Sigma}_{q(\boldsymbol{u}_\ell)})\right) + \mu_{q(1/a_\ell)}$
        $\mu_{q(1/\sigma^2_\ell)} \leftarrow \frac{1}{2}(K_\ell+1)/B_{q(\sigma^2_\ell)}$
    $\boldsymbol{B}_{q(\boldsymbol{\Sigma})} \leftarrow \sum_{i=1}^{m}\left\{\left(\boldsymbol{y}_i - \boldsymbol{C}_i\mu_{q(\boldsymbol{\beta},\boldsymbol{u})}\right)\left(\boldsymbol{y}_i - \boldsymbol{C}_i\mu_{q(\boldsymbol{\beta},\boldsymbol{u})}\right)^T + \boldsymbol{C}_i\boldsymbol{\Sigma}_{q(\boldsymbol{\beta},\boldsymbol{u})}\boldsymbol{C}_i^T\right\}$
                $+2\nu\,\mathrm{diag}\left(1/\mu_{q(a_{\Sigma,1})},\ldots,1/\mu_{q(a_{\Sigma,n})}\right)$
    $\boldsymbol{M}_{q(\boldsymbol{\Sigma}^{-1})} \leftarrow (\nu+m+n-1)\boldsymbol{B}^{-1}_{q(\boldsymbol{\Sigma})}$
    For $j = 1,\ldots,n$:
        $B_{q(a_{\Sigma,j})} \leftarrow \nu\left(\boldsymbol{M}_{q(\boldsymbol{\Sigma}^{-1})}\right)_{jj} + 1/A^2_{\Sigma,j}$   ;   $\mu_{q(1/a_{\Sigma,j})} \leftarrow \frac{1}{2}(\nu+n)/B_{q(a_{\Sigma,j})}$
until the increase in $\underline{p}(\boldsymbol{y};q)$ is negligible.
```

Convergence of Algorithm 1 is monotone in $\underline{p}(\boldsymbol{y};q)$ and usually quite rapid. Section 4 provides some illustrations.

(wileyonlinelibrary.com) DOI: 10.1002/sta4.18

3.1. Heuristic justification of mean field approximation

We briefly provide some heuristic arguments that justify the mean field approximation conveyed by (6) for the Bayesian version of model (3) with noninformative priors on the model parameters.

First consider the marginal longitudinal *parametric* regression model

$$\boldsymbol{y}|\boldsymbol{\beta},\boldsymbol{\Sigma} \sim N(\boldsymbol{X\beta}, \boldsymbol{I}_m \otimes \boldsymbol{\Sigma})$$

where the priors on $\boldsymbol{\beta}$ and $\boldsymbol{\Sigma}$ are noninformative. Under such noninformativity, the Bayes estimates and credible sets of the entries of $\boldsymbol{\beta}$ and $\boldsymbol{\Sigma}$ will be close to the likelihood-based estimates and confidence intervals in the frequentist model

$$\boldsymbol{y} \sim N(\boldsymbol{X\beta}, \boldsymbol{I}_m \otimes \boldsymbol{\Sigma}).$$

For this latter model it is well-known that the Fisher information matrix admits the following block-diagonal form:

$$\begin{bmatrix} \boldsymbol{I}_{\boldsymbol{\beta\beta}} & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{I}_{\boldsymbol{\Sigma\Sigma}} \end{bmatrix}.$$

This implies that the maximum likelihood estimators of $\boldsymbol{\beta}$ and $\boldsymbol{\Sigma}$ are asymptotically independent, sometimes referred to as *parameter orthogonality*. In the Bayesian case, such asymptotic independence corresponds to the approximate product form:

$$p(\boldsymbol{\beta},\boldsymbol{\Sigma}|\boldsymbol{y}) \approx p(\boldsymbol{\beta}|\boldsymbol{y})\,p(\boldsymbol{\Sigma}|\boldsymbol{y}).$$

But replacement of approximate with exact equality in this expression coincides with the mean field assumption $q(\boldsymbol{\beta},\boldsymbol{\Sigma}) = q(\boldsymbol{\beta})\,q(\boldsymbol{\Sigma})$. Hence, we would expect this assumption to be reasonable and, therefore, should incur little loss in accuracy.

Frequentist inference for (3) requires estimation of the random effects vector $\boldsymbol{u}$ and approaches such as best prediction (e.g. Robinson, 1991) or h-likelihood theory (e.g. Lee, Nelder & Pawitan, 2006) are required instead. However, orthogonality between the mean and covariance components of the model is maintained. This provides a heuristic justification of (6).

4 Numerical results

We have applied Algorithm 1 to both simulated and actual data and, in particular, assessed the accuracy of its Bayesian inference against that of MCMC. The results are summarized here.

4.1. Simulation study

We replicated 1000 data-sets corresponding to the simulation setting described in Section 5.1 of Al Kadiri et al. (2010) involving the nonparametric regression model

$$E(y_{ij}|\boldsymbol{u}) = f(x_{ij}), \quad \text{Cov}(\boldsymbol{y}_i|\boldsymbol{u}) = \boldsymbol{\Sigma}, \quad 1 \leqslant i \leqslant 100,\ 1 \leqslant j \leqslant 5 \tag{10}$$

with

$$f(x) = 1 + \frac{1}{2}\Phi((2x-36)/5) \qquad \text{and} \qquad \boldsymbol{\Sigma} = \begin{bmatrix} 0.122 & 0.098 & 0.078 & 0.063 & 0.050 \\ 0.098 & 0.122 & 0.098 & 0.078 & 0.063 \\ 0.078 & 0.098 & 0.122 & 0.098 & 0.078 \\ 0.063 & 0.078 & 0.098 & 0.122 & 0.098 \\ 0.050 & 0.063 & 0.078 & 0.098 & 0.122 \end{bmatrix}.$$

Copyright © 2013 John Wiley & Sons Ltd

Note that (10) corresponds to the special case of the Bayesian hierarchical model (4) with $d = 1$. For each replication, the model was fitted using MFVB via Algorithm 1 and MCMC. As we mentioned earlier, Al Kadiri et al. (2010) used BUGS for MCMC fitting of (10). However the covariance matrix prior of Huang & Wand (2013), used in (4), is not implementable in BUGS. Fortunately, as shown in the appendix, the full conditionals for model (4) are all standard forms and MCMC reduces to Gibbs sampling and is readily implementable in R. Since we also used R for our implementation of Algorithm 1, this allows a fairer comparison of computational times for each approach. Complete fairness is a tall order, though, since convergence is assessed differently for MCMC and MFVB. For MCMC we used a burnin of size 5000 followed by the generation of 5000 samples, with a thinning factor of 5. This resulted in MCMC samples of size 1000 being retained for inference. The MFVB iterations were terminated when the relative change in $\log \underline{p}(\boldsymbol{y}; q)$ fell below 10^{-4}. All computations were performed on the first author's laptop computer (Mac OS X; 2.8 GHz processor, 16 GBytes of random access memory). Table I summarizes the computation times for each approach. The average computing time for MFVB is about $2\frac{1}{2}$ seconds — considerably faster than the $3\frac{1}{4}$ minutes taken by MCMC. The speed gains of MFVB need to be traded off against accuracy losses incurred by restriction (6). Figure 2 displays side-by-side boxplots of accuracy scores defined by

$$\text{accuracy}(q^*) = 100\left(1 - \frac{1}{2}\int_{-\infty}^{\infty} |\, q^*(\theta) - p(\theta|\boldsymbol{y})\,|\, d\theta\right)\% \qquad (11)$$

for a generic parameter θ, and with $p(\theta|\boldsymbol{y})$ replaced by a kernel density estimate based on the MCMC sample. Faes et al. (2010) provides justification for this accuracy score. The parameters monitored are $f(H_k)$, $1 \leqslant k \leqslant 5$, where H_k is the kth sample hexile of the x_{ij} data, and the entries of $\boldsymbol{\Sigma}$. The boxplots show that the majority of accuracy scores exceed 95%, and that they rarely drop below 85%. Such results are in keeping with the heuristics given in Section 3.1.

Table I. Average (standard deviation) times in seconds for MCMC and MFVB fitting of (10) in the simulation study described in the text.

MCMC	MFVB
197.5 (16.13)	2.550 (0.08314)

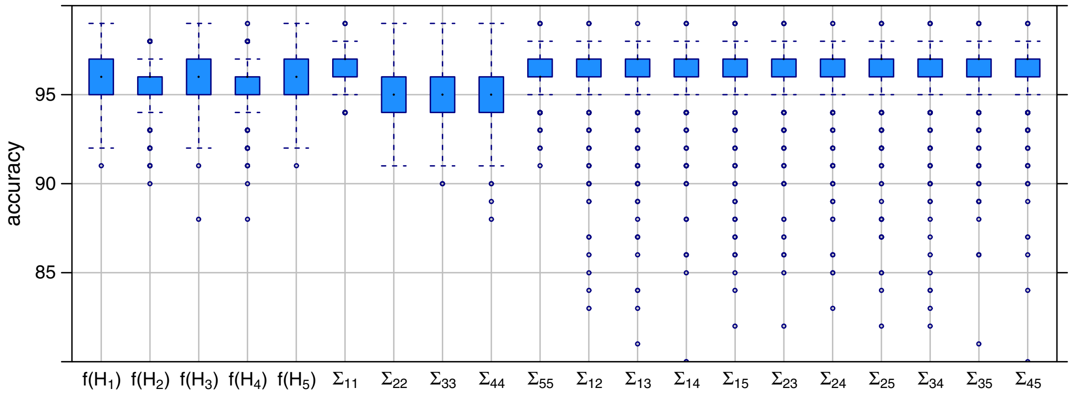

Figure 2. Side-by-side boxplots of accuracy values for MFVB against an MCMC benchmark. Each boxplot corresponds to a parameter in model (10).

Copyright © 2013 John Wiley & Sons Ltd

(wileyonlinelibrary.com) DOI: 10.1002/sta4.18

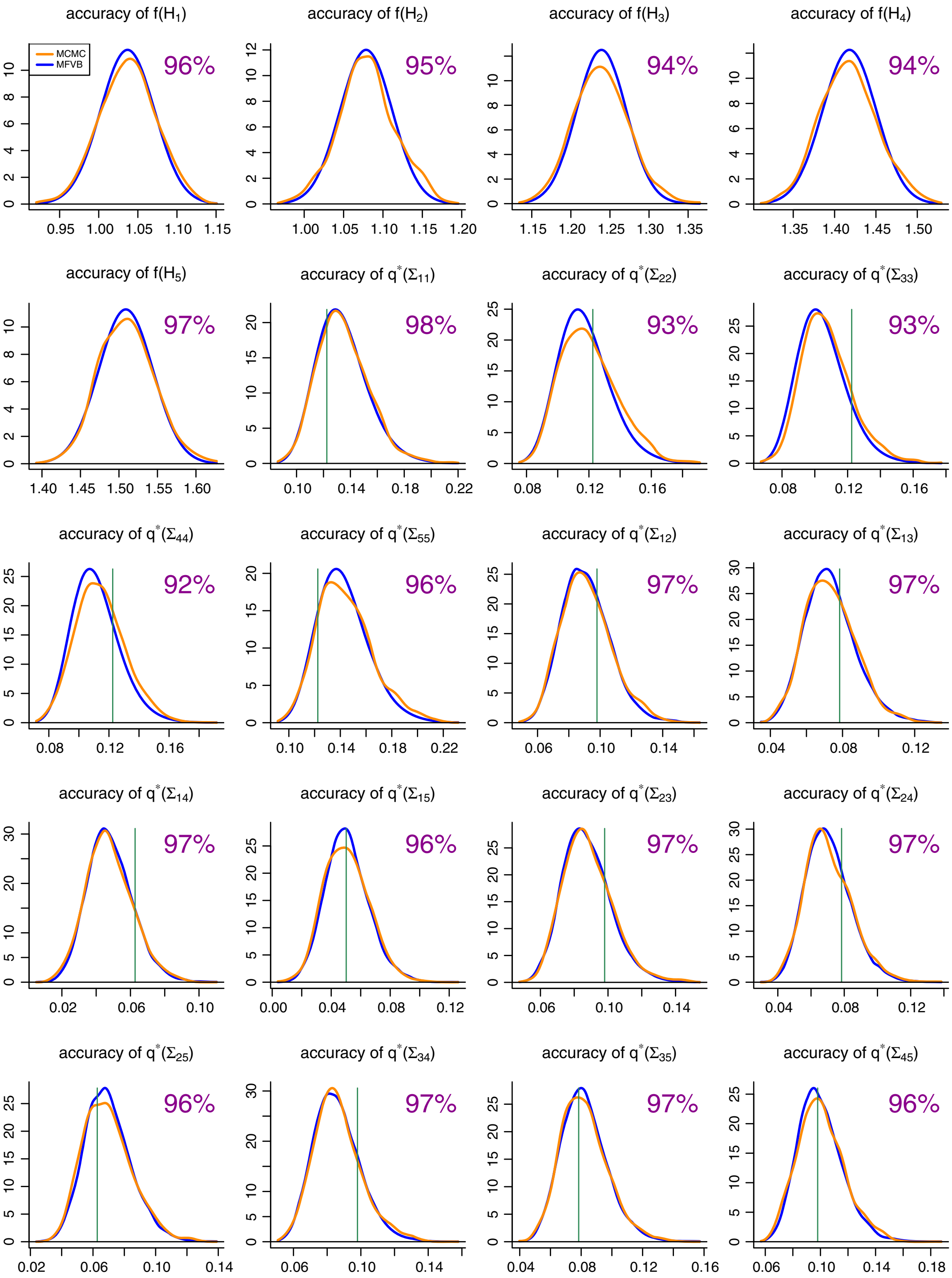

Figure 3. Approximate posterior density functions obtained via MCMC and MFVB for a single replication of the simulation study described in the text. Each pair of density functions corresponds to a parameter in model (10). MFVB accuracy scores, as defined by (11), are also displayed.

Copyright © 2013 John Wiley & Sons Ltd

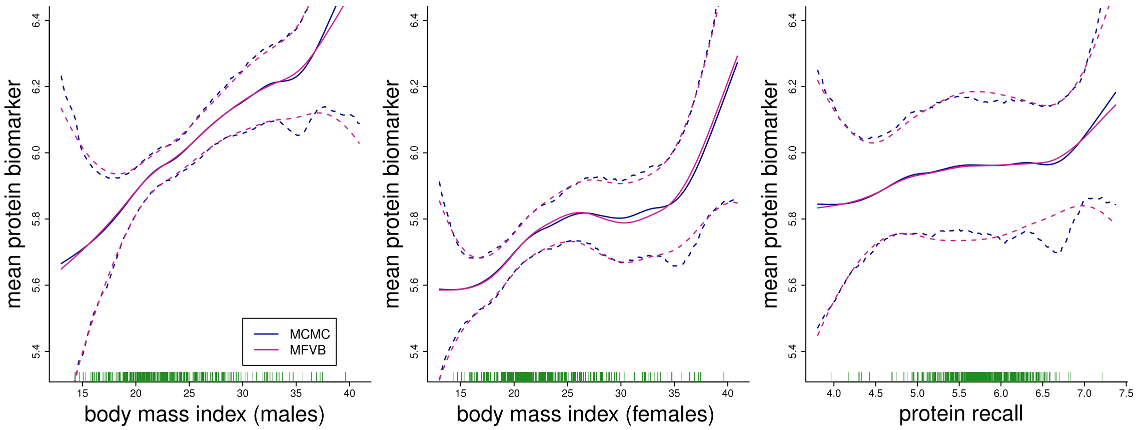

Figure 4. Comparison of MCMC and MFVB fitted functions for the additive/interaction model described in Section 5.2 of Al Kadiri et al. (2010). The solid curves are posterior means and the dashed curves are pointwise 95% credible sets.

Figure 3 allows visual assessment of the MFVB-based approximate posterior density functions against the MCMC-based benchmark for a single replication of the simulation study. MFVB accuracy is seen to be excellent.

4.2. Application

We repeated the additive/interaction model analysis of data from a nutritional epidemiology study presented in Section 5.2 of Al Kadiri et al. (2010). Full details are provided there, although the third panel in its Figure 5 has a slight error which we have corrected here. Figure 4 shows excellent agreement between the MCMC and MFVB fits. For MCMC we used the same sample sizes as in the simulation described in Section 4.1. but for MFVB, we found that a tolerance of 10^{-6} for the relative change in $\log \underline{p}(\boldsymbol{y};q)$ was required. The MCMC fit took 14.4 minutes, whilst the MFVB fit took 19 seconds.

Appendix: Derivation of q^* density functions

Standard manipulations lead to the following full conditional distributions:

$$\begin{aligned}
\begin{bmatrix}\boldsymbol{\beta}\\ \boldsymbol{u}\end{bmatrix}\Bigg|\,\text{rest} \sim N\Bigg(&\left\{\boldsymbol{C}^T(\boldsymbol{I}_m\otimes\boldsymbol{\Sigma}^{-1})\boldsymbol{C}+\text{blockdiag}(\sigma_\beta^{-2}\boldsymbol{I}_p,\sigma_1^{-2}\boldsymbol{I}_{K_1},\ldots,\sigma_d^{-2}\boldsymbol{I}_{K_d})\right\}^{-1}\boldsymbol{C}^T\left(\boldsymbol{I}_m\otimes\boldsymbol{\Sigma}^{-1}\right)\boldsymbol{y},\\
&\left\{\boldsymbol{C}^T(\boldsymbol{I}_m\otimes\boldsymbol{\Sigma}^{-1})\,\boldsymbol{C}+\text{blockdiag}\left(\sigma_\beta^{-2}\boldsymbol{I}_p,\sigma_1^{-2}\boldsymbol{I}_{K_1},\ldots,\sigma_d^{-2}\boldsymbol{I}_{K_d}\right)\right\}^{-1}\Bigg),\\
\sigma_\ell^2|\,\text{rest} &\sim \text{Inverse-Gamma}\left(\tfrac{1}{2}(K_\ell+1),\tfrac{1}{2}\|\boldsymbol{u}_\ell\|^2+a_\ell^{-1}\right),\\
a_\ell|\text{rest} &\sim \text{Inverse-Gamma}\left(1,\sigma_\ell^{-2}+A_\ell^{-2}\right),\quad 1\leqslant\ell\leqslant d,\\
\boldsymbol{\Sigma}|\,\text{rest} &\sim \text{Inverse-Wishart}\Bigg(\nu+m+n-1,\sum_{i=1}^{m}\left(\boldsymbol{y}_i-\boldsymbol{C}_i\begin{bmatrix}\boldsymbol{\beta}\\ \boldsymbol{u}\end{bmatrix}\right)\left(\boldsymbol{y}_i-\boldsymbol{C}_i\begin{bmatrix}\boldsymbol{\beta}\\ \boldsymbol{u}\end{bmatrix}\right)^T\\
&\qquad+2\nu\,\text{diag}\left(1/a_{\Sigma,1},\ldots,1/a_{\Sigma,n}\right)\Bigg)\quad\text{and}\\
a_{\Sigma,j}|\,\text{rest} &\overset{\text{ind.}}{\sim} \text{Inverse-Gamma}\left(\frac{\nu+n}{2},\nu(\boldsymbol{\Sigma}^{-1})_{jj}\;+1/A_{\Sigma,j}^2\right),\;1\leqslant j\leqslant n.
\end{aligned}$$

Copyright © 2013 John Wiley & Sons Ltd

We will restrict attention to $q^*(\boldsymbol{\Sigma})$. The remaining q^* densities involve straightforward adaptation of the derivations given in Appendix C of Wand & Ormerod (2011). Firstly,

$$\log p(\boldsymbol{\Sigma}\,|\,\text{rest}) = -\frac{1}{2}(\nu+m+2n)\log|\boldsymbol{\Sigma}| - \frac{1}{2}\text{tr}\left[\left\{\sum_{i=1}^m\left(\boldsymbol{y}_i-\boldsymbol{C}_i\begin{bmatrix}\boldsymbol{\beta}\\\boldsymbol{u}\end{bmatrix}\right)\left(\boldsymbol{y}_i-\boldsymbol{C}_i\begin{bmatrix}\boldsymbol{\beta}\\\boldsymbol{u}\end{bmatrix}\right)^T\right.\right.$$
$$\left.\left.+2\nu\,\text{diag}(1/a_{\Sigma,1},\ldots,1/a_{\Sigma,n})\right\}\boldsymbol{\Sigma}^{-1}\right] + \text{const}.$$

where 'const' denotes terms not depending on $\boldsymbol{\Sigma}$. Therefore,

$$\begin{aligned}\log q^*(\boldsymbol{\Sigma}) &= E_q\{\log p(\boldsymbol{\Sigma}\,|\,\text{rest})\} + \text{const} = E_q\{p(\boldsymbol{\Sigma}|\boldsymbol{\beta},\boldsymbol{u},\boldsymbol{a}_\Sigma,\boldsymbol{y})\} + \text{const}\\ &= -\frac{1}{2}(\nu+m+2n)\log|\boldsymbol{\Sigma}| - \frac{1}{2}\text{tr}\left[\left\{\sum_{i=1}^m(\boldsymbol{y}_i-\boldsymbol{C}_i\boldsymbol{\mu}_{q(\boldsymbol{\beta},\boldsymbol{u})})(\boldsymbol{y}_i-\boldsymbol{C}_i\boldsymbol{\mu}_{q(\boldsymbol{\beta},\boldsymbol{u})})^T\right.\right.\\ &\left.\left.\quad+\boldsymbol{C}_i\boldsymbol{\Sigma}_{q(\boldsymbol{\beta},\boldsymbol{u})}\boldsymbol{C}_i^T + 2\nu\,\text{diag}\left(\mu_{q(1/a_{\Sigma,1})},\ldots,\mu_{q(1/a_{\Sigma,n})}\right)\right\}\boldsymbol{\Sigma}^{-1}\right] + \text{const}.\end{aligned}$$

It immediately follows that $q^*(\boldsymbol{\Sigma})$ is an Inverse-Wishart$(\nu+m+n-1, \boldsymbol{B}_{q(\boldsymbol{\Sigma})})$ density function with rate matrix $\boldsymbol{B}_{q(\boldsymbol{\Sigma})}$ as given in Algorithm 1. Finally, note that $q^*(\boldsymbol{\Sigma}^{-1})$ is the density function of the Wishart distribution with shape parameter $\nu+m+n-1$ and rate matrix $\boldsymbol{B}_{q(\boldsymbol{\Sigma})}$ and, hence,

$$\boldsymbol{M}_{q(\boldsymbol{\Sigma}^{-1})} \equiv E_q(\boldsymbol{\Sigma}^{-1}) = (\nu+m+n-1)\,\boldsymbol{B}_{q(\boldsymbol{\Sigma})}^{-1}.$$

Acknowledgement

This research was partially supported by an Australian Postgraduate Award and Australian Research Council Discovery Project DP110100061.

References

Al Kadiri, M, Carroll, RJ & Wand, MP (2010), 'Marginal longitudinal semiparametric regression via penalized splines', *Statistics and Probability Letters*, **80**, 1242–1252.

Armagan, A, Dunson, DB & Clyde, M. (2011), *Generalized Beta Mixtures of Gaussians*, Advances in Neural Information Processing Systems 24 in Shawe-Taylor, J, Zemel, RS, Bartlett, P, Pereira, F & Weinberger, KQ (eds.), pp. 523–531.

Bishop, C M (2006), *Pattern Recognition and Machine Learning*, *Springer*, New York.

Faes, C, Ormerod, JT & Wand, MP (2010), 'Variational Bayesian inference for parametric and nonparametric regression with missing data', *Journal of the American Statistical Association*, **106**, 959–971.

Huang, A & Wand, MP (2013), 'Simple marginally noninformative prior distributions for covariance matrices', *Bayesian Analysis*. in press.

Lee, Y, Nelder, J A & Pawitan, Y (2006), *Generalized Linear Models with Random Effects*, *Chapman & Hall/CRC*, Boca Raton, USA.

Copyright © 2013 John Wiley & Sons Ltd

(wileyonlinelibrary.com) DOI: 10.1002/sta4.18

Ormerod, JT & Wand, MP (2010), 'Explaining variational approximations', *The American Statistician*, **64**, 140–153.

Robinson, GK (1991), 'That BLUP is a good thing: the estimation of random effects (with discussion)', *Statistical Science*, **6**, 15–51.

Spiegelhalter, DJ, Thomas, A, Best, NG, Gilks, WR & Lunn, D (2003), *BUGS: Bayesian inference using Gibbs sampling*, *Medical Research Council Biostatistics Unit*, Cambridge, UK. http://www.mrc-bsu.cam.ac.uk/bugs.

Wand, MP & Ormerod, JT (2011), 'Penalized wavelets: embedding wavelets into semiparametric regression', *Electronic Journal of Statistics*, **5**, 1654–1717.

Wand, MP, Ormerod, JT, Padoan, SA & Frühwirth, R (2011), 'Mean field variational Bayes for elaborate distributions', *Bayesian Analysis*, **64**, 847–900.

Copyright © 2013 John Wiley & Sons Ltd