Mean Field Variational Bayesian Inference for Nonparametric Regression with Measurement Error

Tung H. Pham[a], John T. Ormerod[b,*], M.P. Wand[c]

[a] *Institut de Mathématiques, École Polytechnique Fédérale de Lausanne, Station 8, CH-1015, Lausanne, Switzerland*
[b] *School of Mathematics and Statistics, University of Sydney, Sydney 2006, Australia*
[c] *School of Mathematical Sciences, University of Technology, Sydney Broadway 2007, Australia*

Abstract

A fast mean field variational Bayes (MFVB) approach to nonparametric regression when the predictors are subject to classical measurement error is investigated. It is shown that the use of such technology to the measurement error setting achieves reasonable accuracy. In tandem with the methodological development, a customized Markov chain Monte Carlo method is developed to facilitate the evaluation of accuracy of the MFVB method.

Keywords: Penalized splines, classical measurement error, Markov chain Monte Carlo, variational approximations.

1. Introduction

Flexible regression where the predictors are subject to measurement error continues to be an active area of research in the 2000s (Mallick et al., 2002; Liang et al., 2003; Carroll et al., 2004; Ganguli et al., 2005; Carroll et al., 2008) and is likely to be so in the 2010s. Carroll et al. (2006) offers a recent and comprehensive summary of the area.

Fitting and inference in such models is notoriously challenging. Berry et al. (2002) devised an elegant hierarchical Bayes approach to the simplest version of the problem and described Markov chain Monte Carlo (MCMC) based inference. Extensions have been considered by Carroll et al. (2004) and Ganguli et al. (2005). However, inference based on MCMC can be very slow for such models and may take hours if using `BUGS` (Lunn et al., 2000).

In this paper we investigate a faster mean field variational Bayes (MFVB) alternative to the problem. For an introduction to such techniques see Bishop (2006), Ormerod and Wand (2010) or Wand et al. (2011). We show that the transference of such technology to the measurement error setting achieves reasonable accuracy while being hundreds of times faster than MCMC. MFVB approximations to nonparametric regression problems

*Corresponding author
Email addresses: `tung.pham@epfl.ch` (Tung H. Pham), `john.ormerod@sydney.edu.au` (John T. Ormerod), `Matt.Wand@uts.edu.au` (M.P. Wand)

with measurement error in the predictors is challenging due to spline basis functions entering the approximate posterior densities of the unobserved predictor. A streamlined discretization of these approximate posterior densities on a grid across the domain of the predictor is utilized to achieve computational efficiency.

In tandem with the methodological development a customized MCMC is developed to facilitate the evaluation of accuracy of the MFVB method. Both MCMC and MFVB are straightforward for all components of nonparametric regression measurement error model, with the exception of the unobserved predictors. Approximate sampling from the full conditionals for the unobserved predictors can be performed efficiently using griddy-Gibbs sampling steps (Ritter and Tanner, 1992). Note that our MCMC and MFVB methods use an analogous approximation to the posterior distributions of the unobserved predictors.

After a brief introduction to MFVB methods (Section 2) we will develop these methods from the simplest case, simple linear regression (Section 3), and then extend these ideas to the more complex case of nonparametric regression with measurement error (Section 4) which could lay the foundation for more elaborate models such as additive models (see for example, Richardson and Green, 2002; Ganguli et al., 2005). The methodology will be illustrated using a mix of simulated and real world examples (Section 5) and conclusions will be drawn (Section 6).

1.1. Notation

Throughout this paper *i.i.d.* is an abbreviation for *independent and identically distributed*. The notation $\mathbf{x} \sim N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ means that $\mathbf{x}$ has a Multivariate Normal density with mean $\boldsymbol{\mu}$ and covariance $\boldsymbol{\Sigma}$. If x has an Inverse Gamma distribution, denoted $x \sim \text{IG}(A, B)$, if and only if it has density $p(x) = B^A\Gamma(A)^{-1}x^{-A-1}\exp(-B/x)$, $x, A, B > 0$.

2. Elements of Mean Field Variational Bayes

Let $\mathbf{D}$ be a vector of observed data, $\boldsymbol{\theta}$ be a parameter vector with joint distribution $p(\mathbf{D}, \boldsymbol{\theta})$. In the Bayesian inferential paradigm decisions are made based on the posterior distribution $p(\boldsymbol{\theta}|\mathbf{D}) \equiv p(\mathbf{D}, \boldsymbol{\theta})/p(\mathbf{D})$ where $p(\mathbf{D}) \equiv \int p(\mathbf{D}, \boldsymbol{\theta})d\boldsymbol{\theta}$. Let $\{\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_M\}$ be a partition of the parameter vector $\boldsymbol{\theta}$. Then mean field variational Bayes approximates $p(\boldsymbol{\theta}|\mathbf{D})$ by $q(\boldsymbol{\theta}) = \prod_{j=1}^{M} q(\boldsymbol{\theta}_j)$. It can be shown (see for example, Bishop, 2006; Ormerod and Wand, 2010) that the $q(\boldsymbol{\theta}_j)$s, often called q-densities, which minimize the Kullback-Leibler distance between $q(\boldsymbol{\theta})$ and $p(\boldsymbol{\theta}|\mathbf{D})$ defined by

$$\text{KL}(q, p) = \int q(\boldsymbol{\theta}) \log\left\{\frac{q(\boldsymbol{\theta})}{p(\boldsymbol{\theta}|\mathbf{D})}\right\} d\boldsymbol{\theta} \tag{1}$$

are given by

$$q^*(\boldsymbol{\theta}_j) \propto \exp\left[E_{-q(\boldsymbol{\theta}_j)}\left\{p(\boldsymbol{\theta}_j|\text{rest})\right\}\right], \qquad 1 \le j \le M, \tag{2}$$

where $E_{-q(\boldsymbol{\theta}_j)}$ denotes expectation with respect to $\prod_{k\neq j} q(\boldsymbol{\theta}_k)$. Note that only when (2) holds for each $q^*(\boldsymbol{\theta}_j)$, $1 \le j \le M$, is optimality obtained. Furthermore, a lower bound on the marginal log-likelihood is given by

$$\log p(\mathbf{D}) \ge \log \underline{p}(\mathbf{D}; q) = \int q(\boldsymbol{\theta}) \log\left\{\frac{p(\mathbf{D}, \boldsymbol{\theta})}{q(\boldsymbol{\theta})}\right\} d\boldsymbol{\theta}. \tag{3}$$

It can be shown that the calculation of $q^*(\boldsymbol{\theta}_j)$ in (2) for fixed $\{q^*(\boldsymbol{\theta}_k)\}_{k\neq j}$ guarantees a monotonic increase in the lower bound (3) or equivalently a monotonic decrease in the Kullback-Leibler distance (1). Thus, an at least locally optimal $\{q(\boldsymbol{\theta}_j)\}_{1\leq j\leq M}$ can be found by updating the $q^*(\boldsymbol{\theta}_j)$ in (2) sequentially until the lower bound (3) is judged to cease increasing.

To avoid notational clutter for a generic random variable v and density function $q(v)$ let

$$\mu_{q(v)} \equiv E_q(v) \qquad \text{and} \qquad \sigma_v^2 \equiv \text{Var}_q(v).$$

Also, in the special case that $q(v)$ is an Inverse Gamma density function we let $A_{q(v)}$ and $B_{q(v)}$ be the shape and rate parameters of $q(v)$ respectively, i.e. $v \sim \text{IG}(A_{q(v)}, B_{q(v)})$. Note $\mu_{q(1/v)} = A_{q(v)}/B_{q(v)}$. For a generic random vector $\mathbf{v}$ and density function $q(\mathbf{v})$ let

$$\boldsymbol{\mu}_{q(\mathbf{v})} \equiv E_q(\mathbf{v}) \qquad \text{and} \qquad \boldsymbol{\Sigma}_{q(\mathbf{v})} \equiv \text{Cov}_q(\mathbf{v}) = \text{covariance matrix of } \mathbf{v} \text{ under density } q(\mathbf{v}).$$

3. Simple Linear Regression with Measurement Error

We start with the simplest example of a measurement error model, where we want to perform a simple linear regression and the predictor is observed with error. Let

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i, \quad 1 \leq i \leq n, \tag{4}$$

where ε_i are i.i.d. $N(0, \sigma_\varepsilon^2)$. Here the responses, the y_is, are observed, but instead of observing $x_i \sim N(\mu_x, \sigma_x^2)$ we observe a corrupted version of x_i, w_i such that $w_i = x_i + v_i$, where v_i are i.i.d. $N(0, \sigma_v^2)$ random variables with σ_v^2 known.

For convenience we use independent priors with

$$\beta_0, \beta_1 \overset{\text{ind.}}{\sim} N(0, \sigma_\beta^2), \quad \mu_x \sim N(0, \sigma_{\mu_x}^2), \qquad \sigma_x^2 \sim \text{IG}(A_x, B_x), \quad \sigma_\varepsilon^2 \sim \text{IG}(A_\varepsilon, B_\varepsilon),$$

where $\sigma_\beta^2, A_\varepsilon, B_\varepsilon, A_x, B_x$ and $\sigma_{\mu_x}^2$ are positive hyperparameters.

Let $\mathbf{X}$ be the $n \times 2$ matrix with 1 in the first column and x_i in the ith row of the second column. It can be shown via standard algebraic manipulations that the full conditionals for this model are given by

$$\begin{aligned}
\boldsymbol{\beta}|\text{rest} &\sim N\left(\left(\sigma_\varepsilon^{-2}\mathbf{X}^T\mathbf{X} + \sigma_\beta^{-2}\mathbf{I}\right)^{-1}\mathbf{X}^T\mathbf{y}\sigma_\varepsilon^{-2}, \left(\sigma_\varepsilon^{-2}\mathbf{X}^T\mathbf{X} + \sigma_\beta^{-2}\mathbf{I}\right)^{-1}\right), \\
\sigma_\varepsilon^2|\text{rest} &\sim \text{IG}\left(A_\varepsilon + \tfrac{n}{2}, B_\varepsilon + \tfrac{1}{2}\|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|^2\right), \\
\mu_x|\text{rest} &\sim N\left(\frac{\mathbf{1}^T\mathbf{x}/\sigma_x^2}{n\sigma_x^{-2} + \sigma_{\mu_x}^{-2}}, \frac{1}{n\sigma_x^{-2} + \sigma_{\mu_x}^{-2}}\right), \\
\sigma_x^2|\text{rest} &\sim \text{IG}\left(A_x + \tfrac{n}{2}, B_x + \tfrac{1}{2}\|\mathbf{x} - \mu_x\mathbf{1}\|^2\right) \quad \text{and} \\
x_i|\text{rest} &\overset{\text{ind.}}{\sim} N\left(\frac{\beta_1(y_i - \beta_0)/\sigma_\varepsilon^2 + w_i/\sigma_v^2 + \mu_x/\sigma_x^2}{\beta_1^2/\sigma_\varepsilon^2 + 1/\sigma_v^2 + 1/\sigma_x^2}, \frac{1}{\beta_1^2/\sigma_\varepsilon^2 + 1/\sigma_v^2 + 1/\sigma_x^2}\right), \; 1 \leq i \leq n,
\end{aligned} \tag{5}$$

from which Gibbs sampling can be easily implemented.

We now consider a MFVB approximation based on the following factorization

$$q(\boldsymbol{\beta}, \mu_x, \sigma_x^2, \sigma_\varepsilon^2, \mathbf{x}) = q(\boldsymbol{\beta}, \mu_x)q(\sigma_x^2, \sigma_\varepsilon^2)q(\mathbf{x})$$

which leads to the induced factorization $q(\boldsymbol{\beta})q(\mu_x)q(\sigma_x^2)q(\sigma_\varepsilon^2)\prod_{i=1}^n q(x_i)$ (see Bishop, 2006, Section 10.2.5 for the notion of induced factorizations). This leads to the following forms of the optimal q-densities

$$
\begin{aligned}
&q^*(\boldsymbol{\beta}) \text{ is the } N(\boldsymbol{\mu}_{q(\boldsymbol{\beta})}, \boldsymbol{\Sigma}_{q(\boldsymbol{\beta})}) \text{ density function,}\\
&q^*(\sigma_\varepsilon^2) \text{ is the IG}(A_\varepsilon + \tfrac{n}{2}, B_{q(\sigma_\epsilon^2)}) \text{ density function,}\\
&q^*(\mu_x) \text{ is the } N(\mu_{q(\mu_x)}, \sigma^2_{q(\mu_x)}) \text{ density function,}\\
&q^*(\sigma_x^2) \text{ is the IG}(A_x + \tfrac{n}{2}, B_{q(\sigma_x^2)}) \text{ density function, and}\\
&q^*(x_i) \text{ are independent } N(\mu_{q(x_i)}, \sigma^2_{q(x)}) \text{ density functions, } 1 \le i \le n,
\end{aligned}
$$

where the parameters are updated according to Algorithm 1. The variational lower bound on the marginal log-likelihood at the bottom of the main loop is derived in Appendix B. The performance of Algorithm 1 will be analyzed empirically in Section 5.

Initialize: $\mu_{q(1/\sigma_\varepsilon^2)} > 0, \mu_{q(1/\sigma_x^2)} > 0, \mu_{q(\mu_x)}$, $\boldsymbol{\mu}_{q(\boldsymbol{\beta})}$ (2×1) and $\boldsymbol{\Sigma}_{q(\boldsymbol{\beta})}$ (2×2) positive definite.

Cycle:

$$
\begin{aligned}
&\sigma^2_{q(x_i)} \leftarrow 1/\left[\mu_{q(1/\sigma_\varepsilon^2)}\left(\mu^2_{q(\beta_1)} + \sigma^2_{q(\beta_1)}\right) + 1/\sigma_v^2 + \mu_{q(1/\sigma_x^2)}\right]\\
&\text{for} \quad i = 1, \ldots, n:\\
&\qquad \mu_{q(x_i)} \leftarrow \sigma^2_{q(x_i)}\left[\left(y_i\mu_{q(\beta_1)} - \mu_{q(\beta_0\beta_1)}\right)\mu_{q(1/\sigma_\varepsilon^2)} + w_i/\sigma_v^2 + \mu_{q(\mu_x)}\mu_{q(1/\sigma_x^2)}\right]\\
&E_q(\mathbf{X}) \leftarrow \left[\mathbf{1}, \boldsymbol{\mu}_{q(\mathbf{x})}\right] \quad ; \quad E_q\left(\mathbf{X}^T\mathbf{X}\right) \leftarrow \begin{bmatrix} n & \mathbf{1}^T\boldsymbol{\mu}_{q(\mathbf{x})} \\ \mathbf{1}^T\boldsymbol{\mu}_{q(\mathbf{x})} & \|\boldsymbol{\mu}_{q(\mathbf{x})}\|^2 + \mathbf{1}^T\boldsymbol{\sigma}^2_{q(\mathbf{x})} \end{bmatrix}\\
&\boldsymbol{\Sigma}_{q(\boldsymbol{\beta})} \leftarrow \left[\mu_{q(1/\sigma_\varepsilon^2)}E_q\left(\mathbf{X}^T\mathbf{X}\right) + \sigma_\beta^{-2}\mathbf{I}\right]^{-1}; \quad \boldsymbol{\mu}_{q(\boldsymbol{\beta})} \leftarrow \boldsymbol{\Sigma}_{q(\boldsymbol{\beta})}\mu_{q(1/\sigma_\varepsilon^2)}E_q(\mathbf{X})^T\mathbf{y}\\
&\sigma^2_{q(\mu_x)} \leftarrow 1\Big/\left[n\mu_{q(1/\sigma_x^2)} + 1/\sigma^2_{\mu_x}\right] \quad ; \quad \mu_{q(\mu_x)} \leftarrow \sigma^2_{q(\mu_x)}\mu_{q(1/\sigma_x^2)}\boldsymbol{\mu}^T_{q(\mathbf{x})}\mathbf{1}\\
&B_{q(\sigma_\varepsilon^2)} \leftarrow B_\varepsilon + \tfrac{1}{2}\left[\|\mathbf{y}\|^2 - 2\mathbf{y}^T E_q(\mathbf{X})\boldsymbol{\mu}_{q(\boldsymbol{\beta})} + \text{tr}\left(E_q(\mathbf{X}^T\mathbf{X})(\boldsymbol{\Sigma}_{q(\boldsymbol{\beta})} + \boldsymbol{\mu}_{q(\boldsymbol{\beta})}\boldsymbol{\mu}^T_{q(\boldsymbol{\beta})})\right)\right]\\
&B_{q(\sigma_x^2)} \leftarrow B_x + \tfrac{1}{2}\left[\|\boldsymbol{\mu}_{q(\mathbf{x})} - \mu_{q(\mu_x)}\mathbf{1}\|^2 + \mathbf{1}^T\boldsymbol{\sigma}^2_{q(\mathbf{x})} + n\sigma^2_{q(\mu_x)}\right]\\
&\mu_{q(1/\sigma_\varepsilon^2)} \leftarrow \left(A_\varepsilon + \tfrac{n}{2}\right)/B_{q(\sigma_\varepsilon^2)} \quad ; \quad \mu_{q(1/\sigma_x^2)} \leftarrow \left(A_x + \tfrac{n}{2}\right)/B_{q(\sigma_x^2)}
\end{aligned}
$$

until the increase in $\underline{p}(\mathbf{y}, \mathbf{w}; q)$ is negligible.

Algorithm 1: Iterative scheme for obtaining the parameters in the optimal densities $q^*(\boldsymbol{\beta})$, $q^*(\sigma_\varepsilon^2)$, $q^*(x_i)$, $q^*(\mu_x)$ and $q^*(\sigma_x^2)$ for the simple linear regression with measurement error on the predictor data.

4. Nonparametric Regression with Measurement Error

In this section, we consider a nonparametric extension to the previous model via penalized splines:

$$y_i = f(x_i) + \varepsilon_i, \quad 1 \le i \le n,$$

where again ε_i are i.i.d. $N(0, \sigma_\varepsilon^2)$. We model $f(x_i)$ using a typical random effects based spline model where

$$f(x_i) = \beta_0 + \beta_1 x_i + \sum_{k=1}^K u_k z_k(x_i) \quad \text{with} \quad \beta_0, \beta_1 \overset{\text{ind.}}{\sim} N(0, \sigma_\beta^2), \quad u_k|\sigma_u^2 \overset{\text{ind.}}{\sim} N(0, \sigma_u^2)$$

and use $\sigma_u^2 \sim \text{IG}(A_u, B_u)$ as a prior for σ_u^2. The functions $z_k(x)$ are spline basis functions and can take a variety of forms. We use the mixed model representation of O'Sullivan splines (or O-splines) described in Wand and Ormerod (2008). For conciseness of forthcoming expressions we adopt the notation $\mathbf{c}(x) = [1, x, z_1(x), \ldots, z_K(x)]$, $\boldsymbol{\nu} = [\boldsymbol{\beta}^T, \mathbf{u}^T]^T$ and

$$\mathbf{C} = \begin{bmatrix} 1 & x_1 & z_1(x_1) & \ldots & z_K(x_1) \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_n & z_1(x_n) & \ldots & z_K(x_n) \end{bmatrix} = \begin{bmatrix} \mathbf{c}(x_1) \\ \vdots \\ \mathbf{c}(x_n) \end{bmatrix}.$$

The full conditionals for $\boldsymbol{\nu}$, σ_ε^2 and σ_u^2 are given by

$$\begin{aligned} \boldsymbol{\nu}|\text{rest} &\sim N\Big(\sigma_\varepsilon^{-2}\left(\sigma_\varepsilon^{-2}\mathbf{C}^T\mathbf{C} + \text{blockdiag}(\sigma_\beta^{-2}\mathbf{I}, \sigma_u^{-2}\mathbf{I})\right)^{-1}\mathbf{C}^T\mathbf{y}, \\ &\qquad \left(\sigma_\varepsilon^{-2}\mathbf{C}^T\mathbf{C} + \text{blockdiag}(\sigma_\beta^{-2}\mathbf{I}, \sigma_u^{-2}\mathbf{I})\right)^{-1}\Big), \\ \sigma_\varepsilon^2|\text{rest} &\sim \text{IG}\left(A_\varepsilon + \tfrac{n}{2}, B_\varepsilon + \tfrac{1}{2}\|\mathbf{y} - \mathbf{C}\boldsymbol{\nu}\|^2\right) \quad \text{and} \\ \sigma_u^2|\text{rest} &\sim \text{IG}\left(A_u + \tfrac{K}{2}, B_u + \tfrac{1}{2}\|\mathbf{u}\|^2\right) \end{aligned}$$

while the full conditionals for μ_x and σ_x^2 are identical to those in (5). Thus, Gibbs sampling steps can easily be implemented for the model parameters $\boldsymbol{\nu}$, σ_ε^2, σ_u^2, μ_x and σ_x^2.

The full conditional for x_i is

$$p(x_i|\text{rest}) \propto \exp\left[-\frac{1}{2}\left\{\frac{(\mathbf{c}(x_i)^T\boldsymbol{\nu})^2}{\sigma_\varepsilon^2} + (\sigma_x^{-2} + \sigma_v^{-2})x_i^2 - \frac{2x_i\mu_x}{\sigma_x^2}\right\} + \left\{\frac{x_i w_i}{\sigma_v^2} + \frac{y_i\mathbf{c}(x_i)^T\boldsymbol{\nu}}{\sigma_\varepsilon^2}\right\}\right] \quad (6)$$

which, in general, is not easily sampled from due to the nonlinearity of $\mathbf{c}(x_i)$ in x_i. We now use the fact that the first braced term in (6) does not depend on the data $(\mathbf{y}, \mathbf{w})$ whereas the second braced term in (6) depends linearly on w_i and y_i. This observation allows relatively efficient sampling from (6) via the griddy-Gibbs sampling method (Ritter & Tanner, 1992) by using the same grid $\mathbf{g} = (g_1, \ldots, g_M)$ for each x_i. See Appendix A for details.

The MFVB approximation corresponding to the factorization

$$q(\boldsymbol{\nu}, \mu_x, \sigma_x^2, \sigma_\varepsilon^2, \sigma_u^2, \mathbf{x}) = q(\boldsymbol{\nu}, \mu_x)q(\sigma_x^2, \sigma_\varepsilon^2, \sigma_u^2)q(\mathbf{x})$$

leads to the induced factorization $q(\boldsymbol{\nu})q(\mu_x)q(\sigma_x^2)q(\sigma_\varepsilon^2)q(\sigma_u^2)\prod_{i=1}^n q(x_i)$ and the following q-densities

$$\begin{aligned} &q^*(\boldsymbol{\nu}) \text{ is the } N(\boldsymbol{\mu}_{q(\boldsymbol{\nu})}, \boldsymbol{\Sigma}_{q(\boldsymbol{\nu})}) \text{ density function,} \\ &q^*(\sigma_\varepsilon^2) \text{ is the IG}(A_\varepsilon + \tfrac{n}{2}, B_{q(\sigma_\varepsilon^2)}) \text{ density function,} \\ &q^*(\mu_x) \text{ is the } N(\mu_{q(\mu_x)}, \sigma^2_{q(\mu_x)}) \text{ density function,} \\ &q^*(\sigma_x^2) \text{ is the IG}(A_x + \tfrac{n}{2}, B_{q(\sigma_x^2)}) \text{ density function, and} \\ &q^*(\sigma_u^2) \text{ is the IG}(A_u + \tfrac{n}{2}, B_{q(\sigma_u^2)}) \text{ density function} \end{aligned}$$

where the parameters are updated according to Algorithm 2.

The q-density for $\mathbf{x}$ satisfies $q(\mathbf{x}) = \prod_{i=1}^n q(x_i)$ where

$$\begin{aligned} q(x_i) \quad \propto \exp\Big[&-\tfrac{1}{2}\Big\{\mu_{q(1/\sigma_\varepsilon^2)}\mathbf{c}(x_i)^T\left(\boldsymbol{\Sigma}_{q(\boldsymbol{\nu})} + \boldsymbol{\mu}_{q(\boldsymbol{\nu})}^T\boldsymbol{\mu}_{q(\boldsymbol{\nu})}\right)\mathbf{c}(x_i) + (\mu_{q(1/\sigma_x^2)} + \sigma_v^{-2})x_i^2 \\ &-2\mu_{q(1/\sigma_x^2)}x_i\mu_{q(\mu_x)}\Big\} + \left\{\frac{x_i w_i}{\sigma_v^2} + \mu_{q(1/\sigma_\varepsilon^2)}y_i\mathbf{c}(x_i)^T\boldsymbol{\mu}_{q(\boldsymbol{\nu})}\right\}\Big]. \end{aligned} \quad (7)$$

Efficient calculation of the normalizing constant of the density $q(x_i)$ and the various expectations with respect to $q(x_i)$ is given in Appendix A.

Algorithm 2 summarizes the steps for finding each of the q-densities and will be analyzed empirically in Section 5. The variational lower bound on the marginal log-likelihood at the bottom of the main loop is derived in Appendix B. In Appendix B $\mathbf{e}_i$ is a vector of zeros except for 1 in the ith entry.

Set M, the size of the quadrature grid. Obtain $\mathbf{g} = (g_1, \ldots, g_M)$, and then set $\mathbf{C_g}$ using (8) in Appendix A. Initialize: $\mu_{q(1/\sigma_\varepsilon^2)}$, $\mu_{q(1/\sigma_x^2)} > 0$, $\mu_{q(\mu_x)}$, $\boldsymbol{\mu}_{q(\boldsymbol{\nu})}$ and $\boldsymbol{\Sigma}_{q(\boldsymbol{\nu})}$, where $\boldsymbol{\mu}_{q(\boldsymbol{\nu})}$ is $(K+2)$ vector and $\boldsymbol{\Sigma}_{q(\boldsymbol{\nu})}$ is a $(K+2)\times(K+2)$ matrix.

Cycle:

$$\mathbf{b} \leftarrow \left[g_j \mu_{q(1/\sigma_\varepsilon^2)} \mu_{q(\mu_x)} - \frac{\mu_{q(1/\sigma_\varepsilon^2)} \mathbf{c}(g_j)^T (\boldsymbol{\Sigma}_{q(\boldsymbol{\nu})} + \boldsymbol{\mu}_{q(\boldsymbol{\nu})} \boldsymbol{\mu}_{q(\boldsymbol{\nu})}^T) \mathbf{c}(g_j)}{2} - \frac{(\mu_{q(1/\sigma_x^2)} + \sigma_\nu^{-2}) g_j^2}{2} \right]_{1 \le j \le M}$$

$$\mathbf{Q_g} \leftarrow \exp\left[\mathbf{1}_n \mathbf{b}^T + \mathbf{w}\mathbf{g}^T/\sigma_\nu^2 + \mu_{q(1/\sigma_\varepsilon^2)} \mathbf{y}\boldsymbol{\nu}^T \mathbf{C_g}^T \right]$$

$$E_q(\mathbf{C}) \leftarrow \left[\frac{\mathbf{Q_g C_g}}{\mathbf{1}^T \otimes (\mathbf{Q_g 1})} \right] \quad ; \quad E_q(\mathbf{C}^T\mathbf{C}) \leftarrow \mathbf{C_g}^T \mathrm{diag}\left(\sum_{i=1}^n \frac{(\mathbf{e}_i^T \mathbf{Q_g}) \odot \mathbf{1}}{\mathbf{e}_i^T \mathbf{Q_g 1}} \right) \mathbf{C_g}$$

for all $1 \le i \le n$:

$$\mu_{q(x_i)} \leftarrow \frac{\mathbf{e}_i^T \mathbf{Q_g g}}{\mathbf{e}_i^T \mathbf{Q_g 1}}, \quad \mu_{q(x_i^2)} \leftarrow \frac{\mathbf{e}_i^T \mathbf{Q_g}(\mathbf{g}^2)}{\mathbf{e}_i^T \mathbf{Q_g 1}}, \quad \sigma^2_{q(x_i)} \leftarrow \mu_{q(x_i^2)} - \mu^2_{q(x_i)}$$

$$\boldsymbol{\Sigma}_{q(\boldsymbol{\nu})} \leftarrow \left[\mu_{q(1/\sigma_\varepsilon^2)} E_q(\mathbf{C}^T\mathbf{C}) + \mathrm{blockdiag}(\sigma_\beta^{-2}\mathbf{I}, \mu_{q(1/\sigma_u^2)}\mathbf{I}) \right]^{-1}$$

$$\boldsymbol{\mu}_{q(\boldsymbol{\nu})} \leftarrow \mu_{q(1/\sigma_\varepsilon^2)} \boldsymbol{\Sigma}_{q(\boldsymbol{\nu})} E_q(\mathbf{C})^T \mathbf{y}$$

$$\sigma^2_{q(\mu_x)} \leftarrow 1/\left[n\mu_{q(1/\sigma_x^2)} + 1/\sigma^2_{\mu_x} \right] \quad ; \quad \mu_{q(\mu_x)} \leftarrow \sigma^2_{q(\mu_x)} \mu_{q(1/\sigma_x^2)} \boldsymbol{\mu}^T_{q(\mathbf{x})} \mathbf{1}$$

$$B_{q(\sigma_\varepsilon^2)} \leftarrow B_\varepsilon + \tfrac{1}{2}\left[\|\mathbf{y}\|^2 - 2\mathbf{y}^T E_q(\mathbf{C}) \boldsymbol{\mu}_{q(\boldsymbol{\nu})} + \mathrm{tr}\left(E_q(\mathbf{C}^T\mathbf{C})(\boldsymbol{\Sigma}_{q(\boldsymbol{\nu})} + \boldsymbol{\mu}_{q(\boldsymbol{\nu})}^T \boldsymbol{\mu}_{q(\boldsymbol{\nu})}) \right) \right]$$

$$B_{q(\sigma_u^2)} \leftarrow B_u + \tfrac{1}{2}\left[\|\boldsymbol{\mu}_{q(\mathbf{u})}\|^2 + \mathrm{tr}(\boldsymbol{\Sigma}_{q(\mathbf{u})}) \right]$$

$$B_{q(\sigma_x^2)} \leftarrow B_x + \tfrac{1}{2}\left[\|\boldsymbol{\mu}_{q(\mathbf{x})} - \mu_{q(\mu_x)}\mathbf{1}\|^2 + \mathbf{1}^T \boldsymbol{\sigma}^2_{q(\mathbf{x})} + n\sigma^2_{q(\mu_x)} \right]$$

$$\mu_{q(1/\sigma_\varepsilon^2)} \leftarrow \tfrac{A_\varepsilon + n/2}{B_{q(\sigma_\varepsilon^2)}} \quad ; \quad \mu_{q(1/\sigma_x^2)} \leftarrow \tfrac{A_x + n/2}{B_{q(\sigma_x^2)}} \quad ; \quad \mu_{q(1/\sigma_u^2)} \leftarrow \tfrac{A_u + K/2}{B_{q(\sigma_u^2)}}$$

until the increase in $\underline{p}(\mathbf{y}, \mathbf{w}; q)$ is negligible.

Algorithm 2: Iterative scheme for obtaining the parameters in the optimal densities $q^*(\boldsymbol{\nu})$, $q^*(\sigma_\varepsilon^2)$, $q^*(x_i)$, $q^*(\mu_x)$, $q^*(\sigma_x^2)$ and $q^*(\sigma_u^2)$ for the nonparametric regression with measurement error on the predictor data.

5. Examples

We now focus on an empirical assessment of the accuracy of the MFVB algorithms presented in Sections 3 and 4. As in Wand et al. (2011) and Faes et al. (2011) we will use a measure of accuracy based on the L_1 loss or integrated absolute error. For a particular generic parameter θ we use

$$\mathrm{accuracy}(q^*) \equiv 1 - \tfrac{1}{2} \int |q^*(\theta) - p(\theta|\mathbf{y}, \mathbf{w})| d\theta$$

which can be shown to satisfy $0 \leq \text{accuracy}(q^*) \leq 1$ and is expressed as a percentage in our summaries. An approximation of accuracy(q^*) is made by replacing $p(\theta|\mathbf{y}, \mathbf{w})$ with a kernel density estimate of $p(\theta|\mathbf{y}, \mathbf{w})$ based on a large number of MCMC samples via the `R` function `bkde()` in the package `KernSmooth` (Wand and Ripley, 2010).

For simulated data we also assess the credible interval coverage against the actual coverage formed from the MFVB q-densities. For this assessment we carry out many replications of each simulation setting and we report the true parameter coverage for the 95% credible intervals. For the simple linear regression case we report coverage probabilities for $\{\beta_0, \beta_1, \sigma_\varepsilon^2, \mu_x, \sigma_x^2, x_1, x_2, x_3\}$ and for the nonparametric regression case we report coverage probabilities for $\{f(Q_1), f(Q_2), f(Q_3), \sigma_\varepsilon^2, \mu_x, \sigma_x^2, x_{(75)}, x_{(150)}, x_{(225)}\}$ where Q_1, Q_2 and Q_3 are the empirical first, second and third quartiles of the x_is and $x_{(i)}$ is the ith order statistic. Furthermore, for the nonparametric regression case, figures will be used to illustrate 95% credible interval approximations for the mean function for the MFVB approach against their MCMC counterparts.

A commonly used scale-free measure of measurement error is the reliability ratio (RR) (Ganguli et al., 2005, see for example,) defined by:

$$\text{RR} = \frac{\sigma_x^2}{\sigma_x^2 + \sigma_v^2}$$

In our simulations we used the values RR$\in \{0.9, 0.8, 0.7, 0.6\}$ where RR = 0.9 corresponds to a small amount or measurement error and RR = 0.6 corresponds to a substantial corruption of the predictors. In our examples we use the reliability ratio to determine σ_v^2. For the simulations where σ_x^2 is known we use the known value of σ_x^2, otherwise we set σ_x^2 to be the variance of the predictor.

For the simple linear regression simulations we use the prior hyperparameter values $\sigma_\beta^2 = \sigma_{\mu_x}^2 = 10^8$ and $A_\varepsilon = B_\varepsilon = A_x = B_x = 0.01$. We use the same prior hyperparameter values for the nonparametric regression examples and also use $A_u = B_u = 0.01$. Lastly we use $K = 30$ interior knots when using O-splines with spacing as described in Wand and Ormerod (2008).

5.1. Simple Linear Regression Simulations

We conducted a simulation study for the simple linear regression model with true parameter values

$$\beta_0 = -1, \quad \beta_1 = 1, \quad \sigma_\varepsilon^2 = 0.35, \quad \mu_x = \tfrac{1}{2}, \quad \sigma_x^2 = \tfrac{1}{36} \quad \text{and} \quad n \in \{50, 500\}$$

where σ_v^2 is determined using reliability ratios defined in Section 5 above.

Accuracies based on 300 simulated datasets and coverage probabilities based on 10000 simulated datasets for each simulation setting are summarized in Table 1 and Figure 1 below respectively. Table 2 summarizes the means and standard deviations of MFVB and MCMC for 300 simulated datasets. MCMC summaries are based on 10000 MCMC samples (after a 5000 sample burn-in and without thinning).

From Table 1 we can see that coverage probabilities are a little below the 95% level for RR=0.9. For lower values of RR coverage probabilities drop but are reasonable for most parameters. For larger n coverage estimates seem to improve for most parameters. Similarly in Figure 1 we see that parameter accuracies tend to degrade as RR becomes

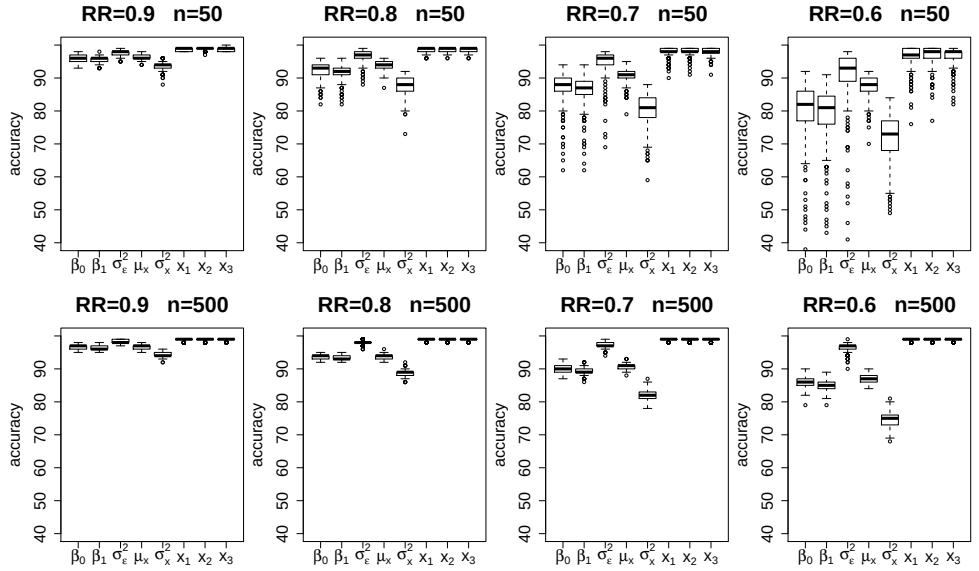

Figure 1: *Boxplots of parameter accuracies for the MFVB approximation applied to the simple linear regression case corresponding to the simulation setting described in Section 5.1.*

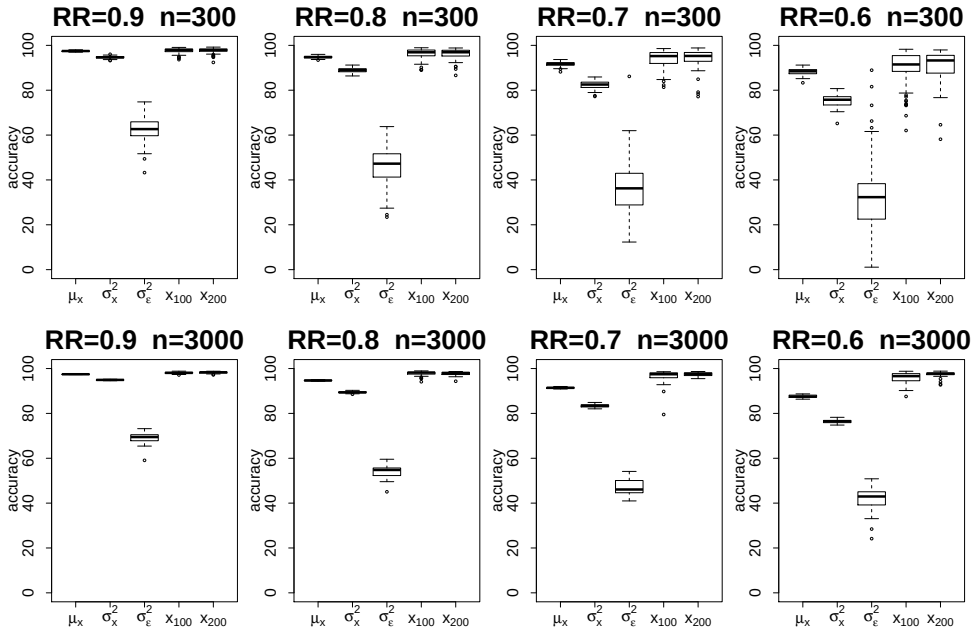

Figure 2: *Boxplots of parameter accuracies for the MFVB approximation applied to the nonparametric regression case corresponding to the simulation setting described in Section 5.2.*

smaller. We also see that the variances of parameter accuracies tends to reduce for larger n. Based on Table 2 we see that the MFVB approach is about 1000 times faster than MCMC.

5.2. Nonparametric Regression Case Simulations

In our simulation for the nonparametric regression case we use

$$f(x_i) = \sin(4\pi x_i), \quad \sigma_\varepsilon^2 = 0.35, \quad \mu_x = \tfrac{1}{2}, \quad \sigma_x^2 = \tfrac{1}{36} \quad \text{and} \quad n \in \{300, 3000\}$$

where σ_v^2 is determined using reliability ratios defined in Section 5 above. Accuracies based on 300 simulated datasets and coverage probabilities based on 1000 simulated datasets are summarized in Table 3 and Figure 2 below respectively. Typical MFVB fits along with 95% credible intervals for the mean function along with corresponding MCMC fits are illustrated in Figure 3. MCMC summaries are based on 5000 MCMC samples (after a 1000 sample burn-in and without thinning).

From Figure 2 we see that accuracies for MFVB seem to be quite good for the x_is for all values of RR, good for μ_x and σ_x^2 for smaller values of RR. However accuracies are poor for σ_ε^2 for all values of RR. These results are consistent with Table 3 where coverage appears reasonable for $f(Q_2)$, μ_x, σ_x^2, $x_{(75)}$, $x_{(75)}$ and $x_{(225)}$ but poor to very poor for $f(Q_1)$, $f(Q_3)$ and σ_ε^2 with performance degrading for smaller values of RR. Furthermore, the variances of accuracies are smaller and coverage is better for larger n.

We see from Figure 3 that the means of the MFVB fits are quite similar to the fits obtained by MCMC. However, 95% credible intervals for the mean of the fitted functions become increasingly underestimated for smaller values of RR. However, for larger values of n the posterior means are nearly indistinguishable (to the eye) when comparing MFVB and MCMC methods and 95% credible intervals for the mean are only slightly underestimated for smaller RR and larger n. Finally, Table 4 shows that the MFVB method appears between 35 and 144 times faster than MCMC depending on the value of RR.

5.3. Illustration for Fossil Data

Consider the *Fossil* data set initially collected and analyzed by Bralower et al. (1997) and subsequently analyzed in Chaudhuri and Marron (1999). For this example the y_is consist of the strontium isotope levels for various fossils whereas the x_is correspond to fossil dates which are obtained using biostratigraphic methods (see Bralower et al., 1997, for details). The biostratigraphic process used to date the fossils can be plausibly envisaged to have some level of measurement error.

Comparative plots of the approximated posterior distributions and the mean fitted functions with corresponding 95% credible sets illustrated in Figure 4 for RR$\in \{0.9, 0.8, 0.7, 0.6\}$. Diagnostic plots from the MCMC analysis are illustrated in Figures 5 and 6 using only the first 1000 MCMC samples.

From Figure 5 and 6 we see strong agreement between MCMC and MFVB for in posterior density estimates of all parameters except σ_ε^2, at least in terms of the posterior means. Posterior variances are slightly underestimated for MFVB estimates. The MFVB-approximate posterior density function possess the same multimodality as their MCMC counterparts. MCMC samples exhibit good convergence.

n	50				500			
RR	0.9	0.8	0.7	0.6	0.9	0.8	0.7	0.6
β_0	93	91	89	85	93	92	90	87
β_1	93	91	88	85	94	92	89	86
σ_ε^2	94	94	94	93	95	94	94	94
μ_x	94	92	89	86	93	92	89	86
σ_x^2	92	88	84	78	92	88	82	76
x_1	95	95	94	94	95	95	95	95
x_2	95	94	94	93	95	95	95	95
x_3	95	94	94	94	95	95	95	95

Table 1: *Coverage probabilities for the MFVB approximation applied to the simple linear regression case corresponding to the simulation setting described in Section 5.1.*

n	50				500			
RR	0.9	0.8	0.7	0.6	0.9	0.8	0.7	0.6
MFVB	0.01	0.01	0.02	0.01	0.01	0.01	0.01	0.01
MCMC	11.89	14.21	14.92	11.14	12.02	13.34	12.37	13.27
Ratio	2050	1319	892	949	1632	1235	959	1037

Table 2: *Times in seconds for the MFVB and MCMC methods applied to the simple linear regression simulation settings described in Section 5.1. The standard errors for all times are less than* 0.001 *for MFVB times and* 0.5 *for MCMC times.*

n	300				3000			
RR	0.9	0.8	0.7	0.6	0.9	0.8	0.7	0.6
$f(Q_1)$	89	65	37	15	92	84	79	74
$f(Q_2)$	91	86	82	76	86	84	79	67
$f(Q_3)$	88	66	39	18	89	85	76	70
σ_ε^2	69	37	20	11	75	53	42	37
μ_x	95	93	91	89	94	93	92	90
σ_x^2	91	88	82	74	89	87	83	75
$x_{(75)}$	97	98	99	99	95	94	92	91
$x_{(150)}$	95	95	96	98	94	90	89	87
$x_{(225)}$	97	99	98	98	95	92	91	92

Table 3: *Coverage probabilities for the MFVB approximation applied to the simple linear regression case corresponding to the simulation setting described in Section 5.2*

n	300				3000			
RR	0.9	0.8	0.7	0.6	0.9	0.8	0.7	0.6
MFVB	6.5	9.1	16.6	49.1	63.9	118.5	167.1	229.2
	(0.2)	(0.4)	(3.9)	(11.9)	(1.3)	(2.7)	(4.1)	(8.2)
MCMC	782.5	782.5	782.3	777.2	9086.8	9101.2	9114.9	9021.6
	(8.5)	(8.7)	(8.7)	(10.2)	(103.2)	(97.0)	(98.5)	(96.0)
Ratio	122.7	89.3	60.0	35.6	143.6	77.6	55.1	40.5

Table 4: *Times (standard errors below) in seconds for the MFVB and MCMC methods applied to the nonparametric regression simulation settings described in Section 5.2.*

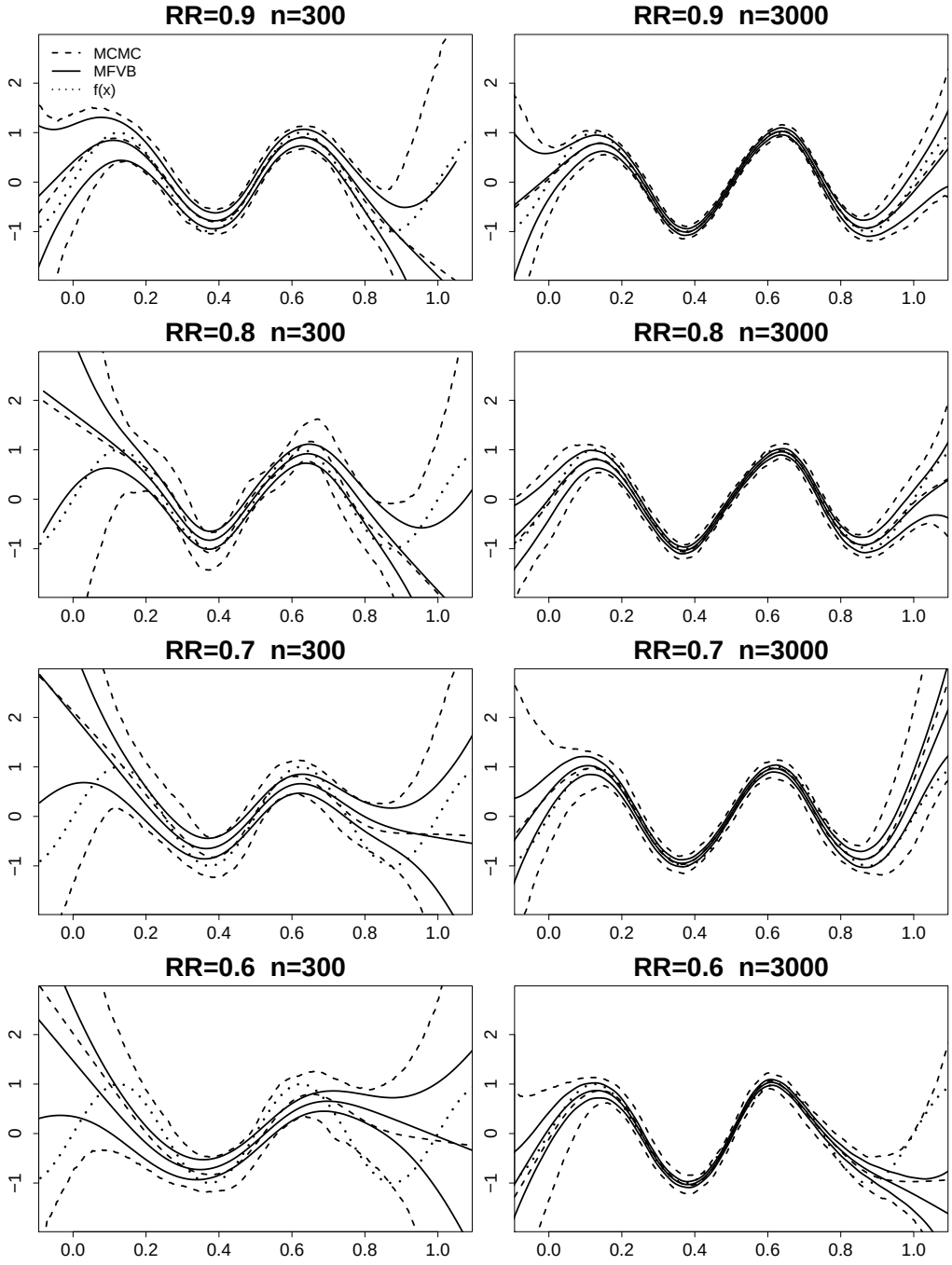

Figure 3: *Comparative plots of the approximated posterior distributions and the mean fitted functions with corresponding 95% credible sets for typical MFVB fits for simulated data described in Section 5.2 and for RR*$\in \{0.9, 0.8, 0.7, 0.6\}$. *MFVB estimates appear as solid lines whereas MCMC estimates appear as dashed lines.*

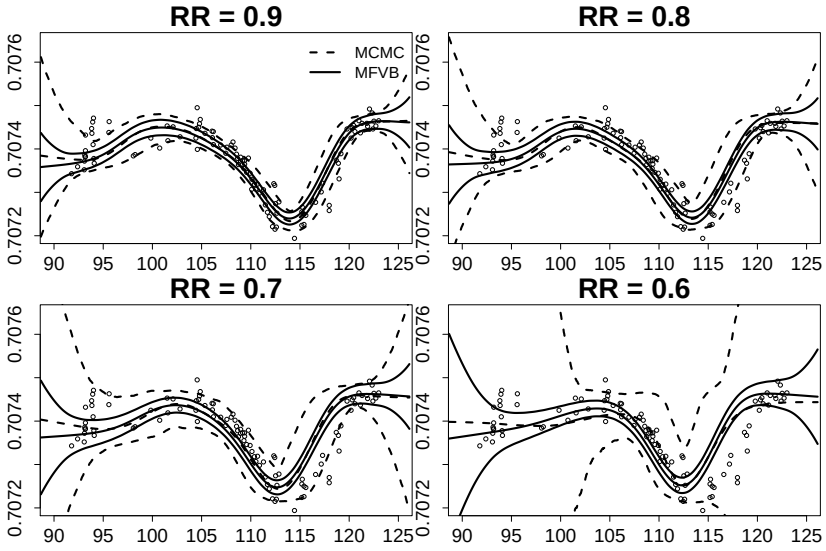

Figure 4: *Comparative plots of the approximated posterior distributions and the mean fitted functions with corresponding 95% credible sets for the Fossil data set for RR*$\in \{0.9, 0.8, 0.7, 0.6\}$. *The x-axes correspond to date (in millions of years) and the y-axes correspond to strontium isotope levels. MFVB estimates appear as solid lines whereas MCMC estimates appear as dashed lines.*

From Figure 4 we observe strong agreement between MCMC and MFVB estimates of $f(\cdot)$, at least for RR$\in \{0.9, 0.8, 0.7\}$. However, MFVB estimates of the 95% credible sets for the posterior mean of $f(\cdot)$ appear increasingly underestimated as RR becomes smaller.

6. Conclusions

In this paper we derived efficient MFVB and MCMC approaches to nonparametric regression problems with classical measurement error. The MFVB approach was shown to have reasonable accuracy with great improvements of speed over the MCMC approach. For the cases involving simple linear regression reasonable coverage probabilities were observed. For the nonparametric regression examples, using the MFVB approach, the estimated mean function approximated the true posterior mean with high accuracy and the 95% credible sets for the mean were only slightly underestimated.

Several simple extensions and a few less simple extensions can be envisaged. Simple extensions include binary response via the probit link, Berkson measurement error, semiparametric regression, non-Gaussian x_i models and models where repeated w_i observations are available to estimate σ_v^2. Extensions for various types of non-Gaussian response are also possible due to Wand et al. (2011), however it is anticipated that these extensions of the techniques presented here would require a reasonable amount of modification.

Finally, the methods presented here are streamlined for one-dimensional spline models. We anticipate that our grid-based MFVB and MCMC approaches would be ex-

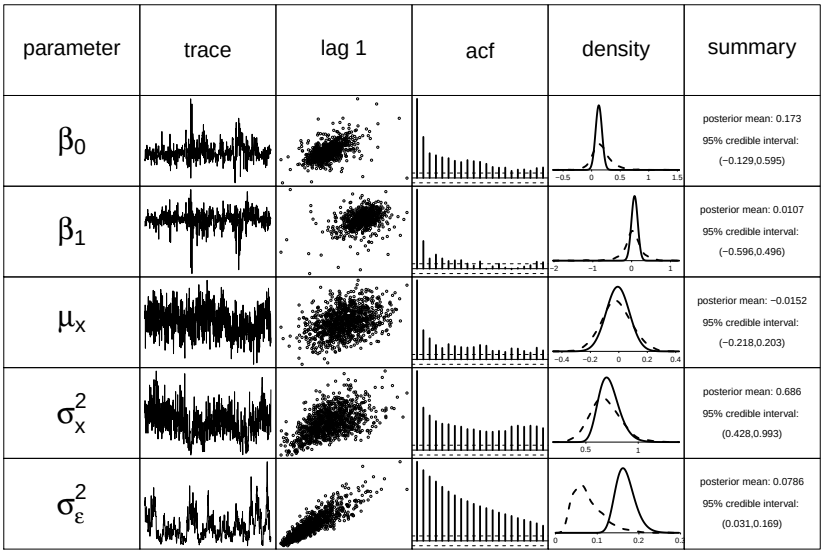

Figure 5: *Diagnostic plots for MCMC analysis along with MFVB posterior density estimates for β_0, β_1, μ_x, σ_x^2 and σ_ε^2. The columns are: missing predictor, trace plot of MCMC sample, plot of sample against 1-lagged sample, sample autocorrelation function, kernel estimates posterior density and basic numerical summaries. For the density column MCMC estimates appear as solid black lines whereas MFVB estimates appear as dashed lines.*

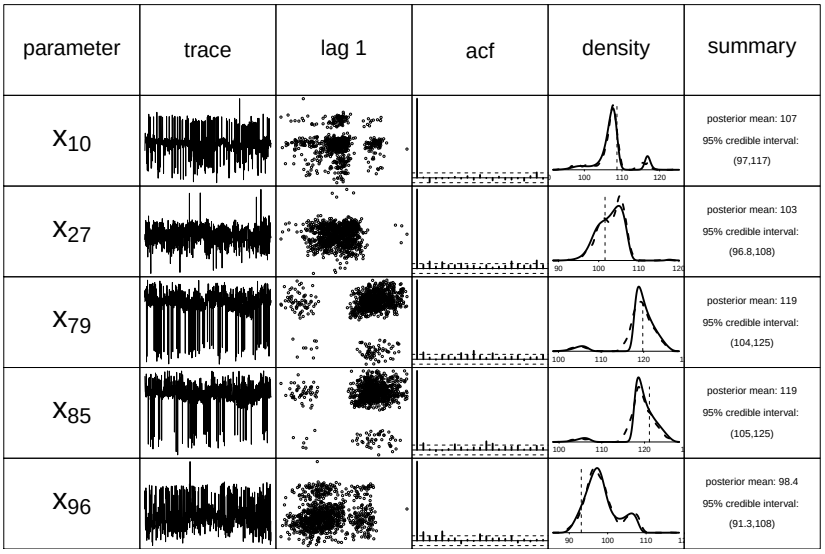

Figure 6: *Diagnostic plots for MCMC analysis along with MFVB posterior density estimates for a random sample of x_is (corresponding to indices: 10, 27, 79, 85 and 96). The columns are: missing predictor, trace plot of MCMC sample, plot of sample against 1-lagged sample, sample autocorrelation function, kernel estimates posterior density and basic numerical summaries. For the density column MCMC estimates appear as solid black lines whereas MFVB estimates appear as dashed lines.*

tendible to higher dimensional measurement error models, for example generalized additive models, provided they do not contain interaction terms. However, such models with measurement error in more than one variable and interaction terms would require more extensive modification or a different approach altogether. Furthermore, factorizing q-densities down to one-dimensional distributions in such high dimensional settings is likely to incur a reduction in accuracy of MFVB approximations.

Acknowledgments

This research was partially supported by Australian Research Council Discovery Project DP110100061.

Appendix A: Approximation of Unobserved Predictor Posterior Densities

We now show how to sample from (6) via griddy-Gibbs sampling. Firstly, let $p(x_i|\text{rest}) = z_i^{-1} P_i(x_i)$ where

$$P_i(x) = \exp\left[-\frac{1}{2}\left\{\frac{(\mathbf{c}(x)^T\boldsymbol{\nu})^2}{\sigma_\varepsilon^2} + (\sigma_x^{-2} + \sigma_v^{-2})x^2 - \frac{2x\mu_x}{\sigma_x^2}\right\} + \left\{\frac{xw_i}{\sigma_v^2} + \frac{y_i\mathbf{c}(x)^T\boldsymbol{\nu}}{\sigma_\varepsilon^2}\right\}\right]$$

and $z_i = \int P_i(x)dx$ is the normalizing constant of $p(x_i|\text{rest})$. To sample from (6) via griddy-Gibbs sampling we approximate $p(x_i|\text{rest})$ by a probability mass function which takes the values over a grid $\mathbf{g} = (g_1, \ldots, g_M)$ with probabilities $\{P_i(g_j)/(\sum_{j=1}^M P_i(g_j))\}_{1\le j\le M}$ which is easily sampled from.

For the examples in Sections 5.2 and 5.3 we choose the grid to be a regular grid of 1000 points between $w_{\min} - (w_{\max} - w_{\min})/10$ and $w_{\max} + (w_{\max} - w_{\min})/10$ where $w_{\min}$ and $w_{\max}$ are the minimum and maximum values of $\mathbf{w}$ respectively.

Using the same grid $\mathbf{g}$ for each predictor allows efficient calculation of $P_i(\cdot)$ over the grid for each i. Let

$$\mathbf{C}_\mathbf{g} = \begin{bmatrix} 1 & g_1 & z_1(g_1) & \ldots & z_K(g_1) \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & g_M & z_1(g_M) & \ldots & z_K(g_M) \end{bmatrix} = \begin{bmatrix} \mathbf{c}(g_1) \\ \vdots \\ \mathbf{c}(g_M) \end{bmatrix} \tag{8}$$

and $\mathbf{a} = (a_1, \ldots, a_M)$ where $a_j = -\frac{1}{2}\left[(\mathbf{C}_\mathbf{g}\boldsymbol{\nu})_j^2/\sigma_\varepsilon^2 + (\sigma_x^{-2} + \sigma_v^{-2})g_j^2 - 2g_j\mu_x/\sigma_x^2\right]$. Then

$$\mathbf{P}_\mathbf{g} = [P_i(g_j)]_{1\le i\le n, 1\le j\le M} = \exp\left[\mathbf{1}_n\mathbf{a}^T + \mathbf{w}\mathbf{g}^T/\sigma_v^2 + \mathbf{y}\boldsymbol{\nu}^T\mathbf{C}_\mathbf{g}^T/\sigma_\varepsilon^2\right]$$

can be calculated in $O(M(n + K))$ operations.

The MFVB calculation of $q(x_i)$ defined by (7) and expectations with respect to $q(x_i)$ can be performed analogously to the griddy-Gibbs procedure for sampling from $p(x_i|\text{rest})$. Firstly, let $q(x_i) = \zeta_i^{-1} Q_i(x_i)$ where

$$\begin{aligned} Q_i(x) \quad \propto \exp\Big[&-\tfrac{1}{2}\Big\{\mu_{q(1/\sigma_\varepsilon^2)}\mathbf{c}(x)^T\left(\boldsymbol{\Sigma}_{q(\boldsymbol{\nu})} + \boldsymbol{\mu}_{q(\boldsymbol{\nu})}^T\boldsymbol{\mu}_{q(\boldsymbol{\nu})}\right)\mathbf{c}(x) + (\mu_{q(1/\sigma_x^2)} + \sigma_v^{-2})x^2 \\ &-2\mu_{q(1/\sigma_x^2)}x\mu_{q(\mu_x)}\Big\} + \left\{\frac{xw_i}{\sigma_v^2} + \mu_{q(1/\sigma_\varepsilon^2)}y_i\mathbf{c}(x)^T\boldsymbol{\mu}_{q(\boldsymbol{\nu})}\right\}\Big] \end{aligned}$$

and $\zeta_i = \int Q_i(x)dx$. Then

$$\mathbf{Q}_\mathbf{g} = [Q_i(g_j)]_{1\le i\le n, 1\le j\le M} = \exp\left[\mathbf{1}_n\mathbf{b}^T + \mathbf{w}\mathbf{g}^T/\sigma_v^2 + \mu_{q(1/\sigma_\varepsilon^2)}\mathbf{y}\boldsymbol{\nu}^T\mathbf{C}_\mathbf{g}^T\right] \tag{9}$$

where $\mathbf{b} = [b_1, \ldots, b_M]^T$ and

$$b_j = -\tfrac{1}{2}\left[\mu_{q(1/\sigma_\varepsilon^2)}((\mathbf{C_g}\boldsymbol{\mu}_{q(\boldsymbol{\nu})})_j^2 + \mathbf{c}(g_j)^T\boldsymbol{\Sigma}_{q(\boldsymbol{\nu})}\mathbf{c}(g_j)) + (\mu_{q(1/\sigma_x^2)} + \sigma_v^{-2})g_j^2 - 2g_j\mu_{q(1/\sigma_x^2)}\mu_{q(\mu_x)}\right].$$

We then approximate $q(x_i)$ by a discrete distribution taking the values $\mathbf{g} = (g_1, \ldots, g_M)$ with probabilities $\{Q_i(g_j)/(\sum_{j=1}^M Q_i(g_j))\}_{1\le j\le M}$. Using this approximation the relevant expectations with respect to $q(x_i)$ are given by

$$\mu_{q(x_i)} \approx \frac{\mathbf{e}_i^T\mathbf{Q_g}\mathbf{g}}{\mathbf{e}_i^T\mathbf{Q_g}\mathbf{1}}, \quad \mu_{q(x_i^2)} \approx \frac{\mathbf{e}_i^T\mathbf{Q_g}(\mathbf{g}^2)}{\mathbf{e}_i^T\mathbf{Q_g}\mathbf{1}}, \quad \sigma^2_{q(x_i)} \approx \mu_{q(x_i^2)} - \mu^2_{q(x_i)}, \quad 1 \le i \le n,$$
$$E_q(\mathbf{C}) \approx \left[\frac{\mathbf{Q_g}\mathbf{C_g}}{\mathbf{1}^T \otimes (\mathbf{Q_g}\mathbf{1})}\right], \quad \text{and} \quad E_q(\mathbf{C}^T\mathbf{C}) \approx \mathbf{C}_\mathbf{g}^T\mathrm{diag}\left(\sum_{i=1}^n \frac{(\mathbf{e}_i^T\mathbf{Q_g}) \odot \mathbf{1}}{\mathbf{e}_i^T\mathbf{Q_g}\mathbf{1}}\right)\mathbf{C_g}.$$

Appendix B: Lower Bound Derivations

For the linear model case $\underline{p}(\mathbf{y}, \mathbf{w}; q) = T_1 + T_2 + T_3 + T_4$ where

$$\begin{aligned} T_1 &= E_q\left[\log\left\{\frac{p(\mathbf{y}|\boldsymbol{\beta}, \sigma_\varepsilon^2)p(\boldsymbol{\beta})p(\sigma_\varepsilon^2)}{q(\boldsymbol{\beta})q(\sigma_\varepsilon^2)}\right\}\right], \quad T_2 = E_q\left[\log\left\{\frac{p(\mathbf{x}|\mu_x, \sigma_x^2)p(\mu_x)p(\sigma_x^2)}{q(\mu_x)q(\sigma_x^2)}\right\}\right], \\ T_3 &= E_q\left[\log\left\{p(\mathbf{w}|\mathbf{x}, \sigma_v^2)\right\}\right] \quad \text{and} \quad T_4 = -E_q\left[\log q(\mathbf{x})\right]. \end{aligned} \tag{10}$$

After minor simplification, T_1 can be written as

$$\begin{aligned} T_1 = {} & \tfrac{p}{2} - \tfrac{n}{2}\log(2\pi) - \tfrac{p}{2}\log(\sigma_\beta^2) + \tfrac{1}{2}\log|\boldsymbol{\Sigma}_{q(\boldsymbol{\beta})}| - \tfrac{1}{2\sigma_\beta^2}\left[\|\boldsymbol{\mu}_{q(\boldsymbol{\beta})}\|^2 + \mathrm{tr}\left(\boldsymbol{\Sigma}_{q(\boldsymbol{\beta})}\right)\right] \\ & + A_\varepsilon\log(B_\varepsilon) - \log\Gamma(A_\varepsilon) - \left(A_\varepsilon + \tfrac{n}{2}\right)\log(B_{q(\sigma_\varepsilon^2)}) + \log\Gamma\left(A_\varepsilon + \tfrac{n}{2}\right) \\ & + \left[B_{q(\sigma_\varepsilon^2)} - B_\varepsilon - \tfrac{1}{2}E_q\left[\|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|^2\right]\right]\mu_{q(1/\sigma_\varepsilon^2)}. \end{aligned}$$

If we perform the updates for σ_ε^2 last then the equality $B_{q(\sigma_\varepsilon^2)} = B_\varepsilon + \frac{1}{2}E_q\left[\|\mathbf{y} - \mathbf{C}\boldsymbol{\nu}\|^2\right]$ holds. Using this expression for $B_{q(\sigma_\varepsilon^2)}$, the term T_1 simplifies to

$$\begin{aligned} T_1 = {} & \tfrac{p}{2} - \tfrac{n}{2}\log(2\pi) - \tfrac{p}{2}\log(\sigma_\beta^2) + \tfrac{1}{2}\log|\boldsymbol{\Sigma}_{q(\boldsymbol{\beta})}| - \tfrac{1}{2\sigma_\beta^2}\left[\|\boldsymbol{\mu}_{q(\boldsymbol{\beta})}\|^2 + \mathrm{tr}\left(\boldsymbol{\Sigma}_{q(\boldsymbol{\beta})}\right)\right] \\ & + A_\varepsilon\log(B_\varepsilon) - \log\Gamma(A_\varepsilon) - \left(A_\varepsilon + \tfrac{n}{2}\right)\log\left(B_{q(\sigma_\varepsilon^2)}\right) + \log\Gamma\left(A_\varepsilon + \tfrac{n}{2}\right). \end{aligned}$$

Similarly for T_2 we have

$$\begin{aligned} T_2 = {} & \tfrac{1}{2} - \tfrac{n}{2}\log(2\pi) + \tfrac{1}{2}\log(\sigma^2_{q(\mu_x)}/\sigma^2_{\mu_x}) - \frac{(\mu_{q(\mu_x)} - \mu_\mu)^2 + \sigma^2_{q(\mu_x)}}{2\sigma^2_{\mu_x}} \\ & + A_x\log(B_x) - \log\Gamma(A_x) - \left(A_x + \tfrac{n}{2}\right)\log(B_{q(\sigma_x^2)}) + \log\Gamma(A_x + n/2) \end{aligned} \tag{11}$$

where we use the update for σ_x^2 last. Note that T_2 is the same expression for the lower bound in Section 2.2.2 of Ormerod and Wand (2010).

For T_3 we obtain

$$T_3 = -\tfrac{n}{2}\log(2\pi\sigma_v^2) - \frac{\|\boldsymbol{\mu}_{q(\mathbf{x})} - \mathbf{w}\|^2 + \mathbf{1}^T\boldsymbol{\sigma}^2_{q(\mathbf{x})}}{2\sigma_v^2} \tag{12}$$

and, since $q(\mathbf{x})$ is a product of univariate Normals, for T_4 we obtain

$$T_4 = \tfrac{n}{2} + \tfrac{n}{2}\log(2\pi) + \tfrac{1}{2}\mathbf{1}^T \log(\sigma^2_{q(\mathbf{x})}).$$

For the nonparametric regression model case $\underline{p}(\mathbf{y}, \mathbf{w}; q) = T_2 + T_3 + T_4 + T_5$ where

$$T_5 = E_q\left[\log\left\{\frac{p(\mathbf{y}|\boldsymbol{\beta}, \mathbf{u}, \sigma^2_\varepsilon)p(\boldsymbol{\beta})p(\mathbf{u}|\sigma^2_u)p(\sigma^2_\varepsilon)p(\sigma^2_u)}{q(\boldsymbol{\beta}, \mathbf{u})q(\sigma^2_\varepsilon)q(\sigma^2_u)}\right\}\right],$$

and the expressions for T_2, T_3 and T_4 are the same as in (10).

The calculation of T_2, T_3 for the nonparametric regression model case are the same for the simple linear model case and are given by equations (11) and (12) respectively. The calculation of T_4 is different for the nonparametric regression model case since $q(\mathbf{x})$ is no longer a product of univariate Normals. For T_4 we use the approximation

$$\begin{aligned} T_4 &= -\sum_{i=1}^{n} E_q\left\{\log q(x_i)\right\} \approx -\sum_{i=1}^{n}\left[\frac{\mathbf{e}_i^T\mathbf{Q}_\mathbf{g}}{\mathbf{e}_i^T\mathbf{Q}_\mathbf{g}\mathbf{1}} \odot \log\left(\frac{\mathbf{e}_i^T\mathbf{Q}_\mathbf{g}}{\mathbf{e}_i^T\mathbf{Q}_\mathbf{g}\mathbf{1}}\right)\right]\mathbf{1} \\ &= \sum_{i=1}^{n}\left[\log(\mathbf{e}_i^T\mathbf{Q}_g\mathbf{1}) - \frac{\left\{\mathbf{e}_i^T\mathbf{Q}_g \odot \log(\mathbf{e}_i^T\mathbf{Q}_g)\right\}^T\mathbf{1}}{\mathbf{e}_i^T\mathbf{Q}_g\mathbf{1}}\right]. \end{aligned}$$

Finally, the expression for T_5 simplifies to

$$\begin{aligned} T_5 &= \tfrac{1}{2}(p+K) - \tfrac{n}{2}\log(2\pi) - \tfrac{p}{2}\log(\sigma^2_\beta) + \tfrac{1}{2}\log|\boldsymbol{\Sigma}_{q(\boldsymbol{\beta},\mathbf{u})}| - \tfrac{1}{2\sigma^2_\beta}\left[\|\boldsymbol{\mu}_{q(\boldsymbol{\beta})}\|^2 + \mathrm{tr}\left(\boldsymbol{\Sigma}_{q(\boldsymbol{\beta})}\right)\right] \\ &\quad + A_\varepsilon\log(B_\varepsilon) - \log\Gamma(A_\varepsilon) - \left(A_\varepsilon + \tfrac{n}{2}\right)\log\left(B_{q(\sigma^2_\varepsilon)}\right) + \log\Gamma\left(A_\varepsilon + \tfrac{n}{2}\right) \\ &\quad + A_u\log(B_u) - \log\Gamma(A_u) - \left(A_u + \tfrac{K}{2}\right)\log\left(B_{q(\sigma^2_u)}\right) + \log\Gamma\left(A_u + \tfrac{m}{2}\right). \end{aligned}$$

References

Berry, S., Carroll, R. J., Ruppert, D., 2002. Bayesian smoothing and regression splines for measurement error problems. Journal of the American Statistical Association 97, 160–169.

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

Bralower, T. J., Fullager, P. D., Paull, C. K., Dwyer, G. S., Leckie, R. M., 1997. Mid-Cretaceous strontium-isotope stratigraphy of deep-sea sections. Geological Society of America Bulletin 109, 1421–1442.

Carroll, R. J., Delaigle, A., Hall, P., 2008. Nonparametric regression estimation from data contaminated by a mixture of Berkson and classical errors. Journal of the Royal Statistical Society, Series B 69, 859–878.

Carroll, R. J., Ruppert, D., Stefanski, L. A., Crainiceanu, C. M., 2006. Measurement Error in Nonlinear Models (Second Edition). Chapman & Hall/CRC, Boca Raton, Florida.

Carroll, R. J., Ruppert, D., Tosteson, T. D., Crainiceanu, C. M., Karagas, M. R., 2004. Nonlinear and nonparametric regression and instrumental variables. Journal of the American Statistical Association 99, 661–671.

Chaudhuri, P., Marron, J. S., 1999. SiZer for exploration of structures in curves. Journal of the American Statistical Association 94, 807–823.

Faes, C., Ormerod, J. T., Wand, M. P., 2011. Variational Bayesian inference for parametric and nonparametric regression with missing data. Journal of the American Statistical Association 106, 959–971.

Ganguli, B., Studenmayer, J., Wand, M. P., 2005. Additive models with predictors subject to measurement error. Australian and New Zealand Journal of Statistics 47, 193–202.

Liang, H., Wu, H., Carroll, R. J., 2003. The relationship between virologic and immunologic responses in AIDS clinical research using mixed-effects varying coefficient semiparametric models with measurement error. Biostatistics 4, 297–312.

Lunn, D. J., Thomas, A., Best, N., Spiegelhalter, D., 2000. WinBUGS – A Bayesian modelling framework: concepts, structure, and extensibility. Statistics and Computing 10, 325–337.

Mallick, B., Hoffman, F. O., Carroll, R. J., 2002. Semiparametric regression modelling with mixtures of Berkson and classical error, with application to fallout from the Nevada Test Site. Biometrics 58, 13–20.

Ormerod, J. T., Wand, M. P., 2010. Explaining variational approximations. The American Statistician 64, 140–153.

Richardson, S., Green, P. J., 2002. Mixture models in measurement error problems, with reference to epidemiological studies. Journal of the Royal Statistical Society, Series A 165, 549–566.

Ritter, C., Tanner, M. A., 1992. Facilitating the Gibbs sampler: the Gibbs stopper and the griddy-Gibbs sampler. Journal of the American Statistical Association 87, 861–868.

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

Wand, M. P., Ormerod, J. T., Padoan, S. A., Fruhwirth, R., 2011. Mean Field Variational Bayes for Elaborate Distributions. Bayesian Analysis 6, 847–900.

Wand, M. P., Ripley, B. D., 2010. KernSmooth 2.23. R package. Functions for kernel smoothing, `http://cran.r-project.org`.