

Variational Bayesian Inference for Parametric and Nonparametric Regression With Missing Data

C. FAES, J. T. ORMEROD, and M. P. WAND

Bayesian hierarchical models are attractive structures for conducting regression analyses when the data are subject to missingness. However, the requisite probability calculus is challenging and Monte Carlo methods typically are employed. We develop an alternative approach based on deterministic variational Bayes approximations. Both parametric and nonparametric regression are considered. Attention is restricted to the more challenging case of missing predictor data. We demonstrate that variational Bayes can achieve good accuracy, but with considerably less computational overhead. The main ramification is fast approximate Bayesian inference in parametric and nonparametric regression models with missing data. Supplemental materials accompany the online version of this article.

KEY WORDS: Directed acyclic graphs; Incomplete data; Mean field approximation; Penalized splines; Variational approximation.

1. INTRODUCTION

Bayesian inference for parametric regression has a long history (e.g., Box and Tiao 1973; Gelman et al. 2004). Mixed model representations of smoothing splines and penalized splines afford Bayesian inference for nonparametric regression (e.g., Ruppert, Wand, and Carroll 2003). Whilst this notion goes back at least to Wahba (1978), recent developments in Bayesian inference methodology, especially Markov chain Monte Carlo (MCMC) algorithms and software, has led to Bayesian approaches to nonparametric regression becoming routine. See, for example, Crainiceanu, Ruppert, and Wand (2005) and Gurin, Scurrah, and Hazelton (2005). There is also a large literature on Bayesian nonparametric regression using regression splines with a variable selection approach (e.g., Denison et al. 2002). The present article deals only with penalized spline nonparametric regression, where hierarchical Bayesian models for nonparametric regression are relatively simple.

When the data are susceptible to missingness a Bayesian approach allows relatively straightforward incorporation of standard missing data models (e.g., Little and Rubin 2004; Daniels and Hogan 2008), resulting in a larger hierarchical Bayesian model. Inference via MCMC is simple in principle, but can be costly in processing time. For example, on the third author's laptop computer (Mac OS X; 2.33 GHz processor, 3 GBytes of random access memory), obtaining 10,000 MCMC samples for a 25-knot penalized spline model, and sample size of 500, takes about 2.6 minutes via the R language (R Development Core Team 2010) package *BRugs* (Ligges et al. 2010). If 30% of the predictor data are reset to be missing completely at random and the appropriate missing data adjustment is made to the model then 10,000 MCMC draws takes about 7.3 minutes; representing an approximate three-fold increase. The situation worsens for more complicated nonparametric and semiparametric regression models. MCMC-based inference, via *BRugs*, for

the missing data/bivariate smoothing example in section 7 of Wand (2009) requires about a week on the aforementioned laptop.

This article is concerned with fast Bayesian parametric and nonparametric regression analysis in situations where some of the data are missing. Speed is achieved by using variational approximate Bayesian inference, often shortened to *variational Bayes*. This is a deterministic approach that yields approximate inference, rather than 'exact' inference produced by an MCMC approach. However, as we shall see, the approximations can be very good. An accuracy assessment, described in Section 3.4, showed that variational Bayes achieves good to excellent accuracy for the main model parameters even with missingness levels as high as 40%.

Variational Bayes is now part of mainstream computer science methodology (e.g., Bishop 2006) and are used in problems such as speech recognition, document retrieval (e.g., Jordan 2004) and functional magnetic resonance imaging (e.g., Flandin and Penny 2007). Recently, they have seen use in statistical problems such as cluster analysis for gene-expression data (Teschendorff et al. 2005) and finite mixture models (McGrory and Titterton 2007). Ormerod and Wand (2010) contains an exposition on variational Bayes from a statistical perspective. A pertinent feature is their heavy algebraic nature. Even relatively simple models require significant notation and algebra for description of variational Bayes.

To the best of our knowledge, the present article is the first to develop and investigate variational Bayes for regression analysis with missing data. In principle, variational Bayes methods can be used in essentially all missing data regression contexts: for example, generalized linear models, mixed models, generalized additive models, geostatistical models and their various combinations. It is prudent, however, to start with simpler regression models where the core tenets can be elucidated without excessive notation and algebra. Hence, the present article treats the simplest parametric and nonparametric regression models: single predictor with homoscedastic Gaussian errors. The full array of missing data scenarios: missing completely at random (MCAR), missing at random (MAR) and missing not at random (MNAR) is treated. For parametric regression

C. Faes is Assistant Professor, Interuniversity Institute for Biostatistics and Statistical Bioinformatics, Hasselt University, BE3590 Diepenbeek, Belgium. J. T. Ormerod is Lecturer, School of Mathematics and Statistics, University of Sydney, Sydney 2006, Australia. M. P. Wand is Distinguished Professor, School of Mathematical Sciences, University of Technology, Sydney, Broadway 2007, Australia (E-mail: matt.wand@uts.edu.au). The authors are grateful to the editor, associate editor and two referees for their feedback on earlier versions of this article. This research was partially supported by the Flemish Fund for Scientific Research, Interuniversity Attraction Poles (Belgian Science Policy) network number P6/03 and Australian Research Council Discovery Project DP0877055.

with probit missing data mechanisms, we show that variational Bayes is purely algebraic, without the need for quadrature or Monte Carlo-based approximate integration. The nonparametric regression extension enjoys many of the assets of parametric regression, but requires some univariate quadrature. Comparisons with MCMC show quite good accuracy, but with computation in the order of seconds rather than minutes. The upshot is fast approximate Bayesian inference in parametric and nonparametric regression models with missing data.

Section 2 summarizes the variational Bayes approach. Inference in the simple linear regression model with missing data is the focus of Section 3. In Section 4 we describe extension to nonparametric regression. Some closing discussion is given in Section 5. An Appendix summarizes notation used throughout this article. Detailed derivations are given in the online supplemental material.

2. ELEMENTS OF VARIATIONAL BAYES

Variational Bayes methods are a family of approximate inference techniques based on the notions of minimum Kullback–Leibler divergence and product assumptions on the posterior densities of the model parameters. They are known as *mean field* approximations in the statistical physics literature (e.g., Parisi 1988). Detailed expositions on variational Bayes may be found in Bishop (2006, sections 10.1–10.4) and Ormerod and Wand (2010). The elements of variational Bayes are described here.

Consider a generic Bayesian model with parameter vector $\theta \in \Theta$ and observed data vector \mathbf{D} . Bayesian inference is based on the posterior density function

$$p(\theta|\mathbf{D}) = \frac{p(\mathbf{D}, \theta)}{p(\mathbf{D})}.$$

We will suppose that \mathbf{D} and θ are continuous random vectors, which conforms with the models in Sections 3 and 4. Let q be an arbitrary density function over Θ . Then the marginal likelihood $p(\mathbf{D})$ satisfies $p(\mathbf{D}) \geq \underline{p}(\mathbf{D}; q)$ where

$$\underline{p}(\mathbf{D}; q) \equiv \exp \int_{\Theta} q(\theta) \log \left\{ \frac{p(\mathbf{D}, \theta)}{q(\theta)} \right\} d\theta.$$

The gap between $\log\{p(\mathbf{D})\}$ and $\log\{\underline{p}(\mathbf{D}; q)\}$ is known as the *Kullback–Leibler divergence* and is minimized by

$$q_{\text{exact}}(\theta) = p(\theta|\mathbf{D}),$$

the exact posterior density function. However, for most models of practical interest, $q_{\text{exact}}(\theta)$ is intractable and restrictions need to be placed on q to achieve tractability. Variational Bayes relies on product density restrictions:

$$q(\theta) = \prod_{i=1}^M q_i(\theta_i) \quad \text{for some partition } \{\theta_1, \dots, \theta_M\} \text{ of } \theta. \quad (1)$$

This restriction is usually governed by tractability considerations. The derivations in Supplement C of the supplemental materials demonstrate the tractability advantages of product forms for the types of models considered in Section 3. As we explain in Section 3.2, the notion of *d-separation* can be used to guide the choice of the partition. A cost of the tractability afforded by (1) is that it imposes posterior independence between the

partition elements $\theta_1, \dots, \theta_M$. Depending on the amounts of actual posterior dependence, the accuracy of variational Bayes can range from excellent to poor. For example, in linear regression models with noninformative independent priors, regression coefficients and the error variance tend to have negligible posterior dependence and the assumption of posterior independence has a mild effect. In Section 3.3 we describe a situation where the posterior dependence is nonnegligible and the variational Bayes approximation suffers.

Under restriction (1), the optimal densities (with respect to minimum Kullback–Leibler divergence) can be shown to satisfy

$$q_i^*(\theta_i) \propto \exp\{E_{-\theta_i} \log p(\mathbf{D}, \theta)\}, \quad 1 \leq i \leq M, \quad (2)$$

where $E_{-\theta_i}$ denotes expectation with respect to the density $\prod_{j \neq i} q_j(\theta_j)$. Equations (2) are a set of consistency conditions for the maximum of $\underline{p}(\mathbf{D}; q)$ subject to constraint (1). Each update uniquely maximizes $\underline{p}(\mathbf{D}, q)$ with respect to the parameters of $q_i^*(\theta_i)$. Convergence of such a scheme, known the *method of alternating variables* or the *coordinate ascent method*, is guaranteed under mild assumptions (Luenberger and Ye 2008, p. 253). Convergence can be assessed by monitoring relative increases in $\log\{\underline{p}(\mathbf{D}; q)\}$. In all of the examples in this article, convergence was achieved within a few hundred iterations with a stringent tolerance level.

An equivalent form for the solutions is

$$q_i^*(\theta_i) \propto \exp\{E_{-\theta_i} \log p(\theta_i|\text{rest})\}, \quad 1 \leq i \leq M, \quad (3)$$

where $\text{rest} \equiv \{\mathbf{D}, \theta_1, \dots, \theta_{i-1}, \theta_{i+1}, \dots, \theta_M\}$ is the set containing the rest of the random vectors in the model, apart from θ_i . The distributions $\theta_i|\text{rest}$, $1 \leq i \leq M$, are known as the *full conditionals* in the MCMC literature. Gibbs sampling (e.g., Robert and Casella 2004) involves successive draws from these full conditionals. We prefer (3) to (2), since it lends itself to considerable simplification via graph theoretic results that we describe next.

2.1 Directed Acyclic Graphs and Markov Blanket Theory

The missing data regression models of Sections 3 and 4 are hierarchical Bayesian models, and hence can be represented as probabilistic directed acyclic graphs (DAGs). DAGs provide a useful ‘road map’ of the model’s structure, and aid the algebra required for variational Bayes. Random variables or vectors correspond to *nodes* while *directed edges* (i.e., *arrows*) convey conditional dependence. The observed data components of the DAG are sometimes called *evidence* nodes, whilst the model parameters correspond to *hidden* nodes. Bishop (2006, chapter 8) and Wasserman (2004, chapter 17) provide very good summaries of DAGs and their probabilistic properties. Figures 1 and 4 (in Sections 3.2 and 4, respectively) contain DAGs for models considered in the present article.

The formulation of variational Bayes algorithms greatly benefit from a DAG-related known concept known as *Markov blanket* theory. First we define the Markov blanket of a node on a DAG:

Definition. The Markov blanket of a node on a DAG is the set of children, parents, and coparents of that node. Two nodes are coparents if they have at least one child node in common.

Markov blankets are important in the formulation of variational Bayes algorithms because of:

Theorem (Pearl 1988). For each node on a probabilistic DAG, the conditional distribution of the node given the rest of the nodes is the same as the conditional distribution of the node given its Markov blanket.

For our generic Bayesian example, this means that

$$p(\theta_i | \text{rest}) = p(\theta_i | \text{Markov blanket of } \theta_i).$$

It immediately follows that

$$q_i^*(\theta_i) \propto \exp\{E_{-\theta_i} \log p(\theta_i | \text{Markov blanket of } \theta_i)\}, \quad 1 \leq i \leq M. \quad (4)$$

For large DAGs, such as those in Figure 4, (4) yields considerable algebraic economy. In particular, it shows that the $q_i^*(\theta_i)$ require only local calculations on the model's DAG.

3. SIMPLE LINEAR REGRESSION WITH MISSING PREDICTOR DATA

In this section we confine attention to the simple linear regression model with homoscedastic Gaussian errors. For complete data on the predictor/response pairs (x_i, y_i) , $1 \leq i \leq n$, this model is

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i, \quad \varepsilon_i \stackrel{\text{ind.}}{\sim} N(0, \sigma_\varepsilon^2).$$

We couch this in a Bayesian framework by taking $\beta_0, \beta_1 \stackrel{\text{ind.}}{\sim} N(0, \sigma_\beta^2)$ and $\sigma_\varepsilon^2 \sim \text{IG}(A_\varepsilon, B_\varepsilon)$ for hyperparameters $\sigma_\beta^2, A_\varepsilon, B_\varepsilon > 0$. Use of these conjugate priors simplifies the variational Bayes algebra. Other priors, such as those described by Gelman (2006), may be used. However, they result in more complicated variational Bayes algorithms.

Now suppose that the predictors are susceptible to missingness. Bayesian inference then requires a probabilistic model for the x_i 's. We will suppose that

$$x_i \stackrel{\text{ind.}}{\sim} N(\mu_x, \sigma_x^2) \quad (5)$$

and take $\mu_x \sim N(0, \sigma_{\mu_x}^2)$ and $\sigma_x^2 \sim \text{IG}(A_x, B_x)$ for hyperparameters $\sigma_{\mu_x}^2, A_x, B_x > 0$. If normality of the x_i 's cannot be reasonably assumed then (5) should be replaced by an appropriate parametric model. The variational Bayes algorithm will need to be changed accordingly. For concreteness and simplicity we will assume that (5) is reasonable for the remainder of the article.

For $1 \leq i \leq n$ let R_i be a binary random variable such that

$$R_i = \begin{cases} 1, & \text{if } x_i \text{ is observed,} \\ 0, & \text{if } x_i \text{ is missing.} \end{cases}$$

Bayesian inference for the regression model parameters differs according to the dependence of the distribution of R_i on the observed data (e.g., Gelman et al. 2004, section 17.2). With Φ denoting the standard normal cumulative distribution function, we will consider the following three missingness mechanisms:

1. $P(R_i = 1) = p$ for some constant $0 < p < 1$. In this case the missing-data mechanism is independent of the data, and the x_i 's are said to be missing completely at random (MCAR). Under MCAR, the observed data are a simple random sample of the complete data.

2. $P(R_i = 1 | \phi_0, \phi_1, y_i) = \Phi(\phi_0 + \phi_1 y_i)$ for parameters $\phi_0, \phi_1 \stackrel{\text{ind.}}{\sim} N(0, \sigma_\phi^2)$ and hyperparameter $\sigma_\phi^2 > 0$. In this case, the missing-data mechanism depends on the observed y_i 's but not on the missing x_i 's. Inference for the regression parameters β_0, β_1 , and σ_ε^2 is unaffected by the ϕ_0 and ϕ_1 or the conditional distribution $R_i | \phi_0, \phi_1, y_i$. The x_i 's are said to be missing at random (MAR). In addition, the independence of the priors for (ϕ_0, ϕ_1) from those of the regression parameters means that the missingness is *ignorable* (Little and Rubin 2004).
3. $P(R_i = 1 | \phi_0, \phi_1) = \Phi(\phi_0 + \phi_1 x_i)$ for parameters $\phi_0, \phi_1 \stackrel{\text{ind.}}{\sim} N(0, \sigma_\phi^2)$ and hyperparameter $\sigma_\phi^2 > 0$. In this case, the missing-data mechanism depends on the unobserved x_i 's and inference for the regression parameters β_0, β_1 , and σ_ε^2 depends on the ϕ_0 and ϕ_1 and $R_i | \phi_0, \phi_1, x_i$. The x_i 's are said to be missing not at random (MNAR).

Define the matrices

$$\mathbf{X} = \begin{bmatrix} 1 & x_1 \\ \vdots & \vdots \\ 1 & x_n \end{bmatrix}, \quad \mathbf{Y} = \begin{bmatrix} 1 & y_1 \\ \vdots & \vdots \\ 1 & y_n \end{bmatrix}, \quad \boldsymbol{\beta} = \begin{bmatrix} \beta_0 \\ \beta_1 \end{bmatrix},$$

and

$$\boldsymbol{\phi} = \begin{bmatrix} \phi_0 \\ \phi_1 \end{bmatrix}.$$

Then the three missing data models can be summarized as follows:

$$\begin{aligned} y_i | x_i, \boldsymbol{\beta}, \sigma_\varepsilon^2 &\stackrel{\text{ind.}}{\sim} N((\mathbf{X}\boldsymbol{\beta})_i, \sigma_\varepsilon^2), & x_i | \mu_x, \sigma_x^2 &\stackrel{\text{ind.}}{\sim} N(\mu_x, \sigma_x^2), \\ \boldsymbol{\beta} &\sim N(\mathbf{0}, \sigma_\beta^2 \mathbf{I}), & \mu_x &\sim N(0, \sigma_{\mu_x}^2), \\ \sigma_\varepsilon^2 &\sim \text{IG}(A_\varepsilon, B_\varepsilon), & \sigma_x^2 &\sim \text{IG}(A_x, B_x), \end{aligned} \quad (6)$$

$$R_i | \boldsymbol{\phi}, x_i, y_i \stackrel{\text{ind.}}{\sim} \begin{cases} \text{Bernoulli}(p), & \\ \text{model with } x_i \text{ MCAR}, & \\ \text{Bernoulli}[\Phi\{(\mathbf{Y}\boldsymbol{\phi})_i\}], & \\ \text{model with } x_i \text{ MAR}, & \\ \text{Bernoulli}[\Phi\{(\mathbf{X}\boldsymbol{\phi})_i\}], & \\ \text{model with } x_i \text{ MNAR}, & \end{cases}$$

$$\boldsymbol{\phi} \sim N(\mathbf{0}, \sigma_\phi^2 \mathbf{I}).$$

Of course, for the model with x_i MCAR, the assumption $R_i | \boldsymbol{\phi} \stackrel{\text{ind.}}{\sim} \text{Bernoulli}(p)$ simplifies to $R_i \stackrel{\text{ind.}}{\sim} \text{Bernoulli}(p)$ and $\boldsymbol{\phi}$ is superfluous.

The following additional notation is useful in the upcoming sections. Let n_{obs} denote the number of observed x_i 's and n_{mis} be the number of missing x_i 's. Let \mathbf{x}_{obs} be the $n_{\text{obs}} \times 1$ vector containing the observed x_i 's and \mathbf{x}_{mis} be $n_{\text{mis}} \times 1$ vector containing the missing x_i 's. We reorder the data so that the observed data is first. Hence, the full vector of predictors is

$$\mathbf{x} \equiv \begin{bmatrix} \mathbf{x}_{\text{obs}} \\ \mathbf{x}_{\text{mis}} \end{bmatrix}.$$

Finally, let $y_{x_{\text{mis},i}}$ be the value of the response variable corresponding to $x_{\text{mis},i}$, the i th entry of \mathbf{x}_{mis} .

3.1 Incorporation of Auxiliary Variables

It is now well established that Bayesian models with probit regression components benefit from the introduction of auxiliary variables. This was demonstrated by Albert and Chib (1993) for inference via Gibbs sampling and by Girolami and Rogers (2006) for variational Bayes inference. Appropriate auxiliary variables are

$$a_i | \phi \sim N((\mathbf{Y}\phi)_i, 1) \quad \text{for the model with } x_i \text{ MAR,} \quad \text{and} \quad (7)$$

$$a_i | \phi \sim N((\mathbf{X}\phi)_i, 1) \quad \text{for the model with } x_i \text{ MNAR.}$$

A consequence of (7) is

$$P(R_i = r | a_i) = I(a_i \geq 0)^r I(a_i < 0)^{1-r}, \quad r = 0, 1.$$

As will become clear in Section 3.3, variational Bayes becomes completely algebraic (i.e., without the need for numerical integration or Monte Carlo methods) if auxiliary variables are incorporated into the model.

3.2 Directed Acyclic Graphs Representations

Figure 1 provides DAG summaries of the three missing data models, after the incorporation of the auxiliary variables $\mathbf{a} = (a_1, \dots, a_n)$ given by (7). To enhance clarity, the hyperparameters are suppressed in the DAGs.

The DAGs in Figure 1 show the interplay between the regression parameters and missing data mechanism parameters. For the MCAR model the observed data indicator vector $\mathbf{R} = (R_1, \dots, R_n)$ is completely separate from the rest of the DAG. Delineation between the MAR and MNAR is more subtle, but can be gleaned from the directed edges in the respective DAGs and graph theoretical results. The Markov blanket theorem of Section 2.1 provides one way to distinguish MAR from MNAR. Table 1 lists the Markov blankets for each of the hidden nodes (i.e., model parameters or missing predictors) under the two missing-data models. Under MAR, there is a separation between the two hidden node sets

$$\{\beta, \sigma_\epsilon^2, \mathbf{x}_{\text{mis}}, \mu_x, \sigma_x^2\} \quad \text{and} \quad \{\mathbf{a}, \phi\}$$

in that their Markov blankets have no overlap. It follows immediately that Bayesian inference for the regression parameters based on Gibbs sampling or variational Bayes is not impacted by the missing-data mechanism. In the MNAR case, this separation does not occur since, for example, the Markov blanket of \mathbf{x}_{mis} includes $\{\mathbf{a}, \phi\}$.

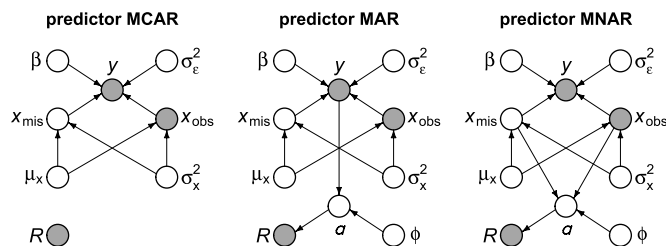


Figure 1. DAGs for the three missing data models for simple linear regression, given by (6). Shaded nodes correspond to the observed data.

Table 1. The Markov blankets for each node in the second and third DAGs of Figure 1

Node	Markov blanket under MAR	Markov blanket under MNAR
β	$\{y, \sigma_\epsilon^2, \mathbf{x}_{\text{mis}}, \mathbf{x}_{\text{obs}}\}$	$\{y, \sigma_\epsilon^2, \mathbf{x}_{\text{mis}}, \mathbf{x}_{\text{obs}}\}$
σ_ϵ^2	$\{y, \beta, \mathbf{x}_{\text{mis}}, \mathbf{x}_{\text{obs}}\}$	$\{y, \beta, \mathbf{x}_{\text{mis}}, \mathbf{x}_{\text{obs}}\}$
\mathbf{x}_{mis}	$\{y, \beta, \sigma_\epsilon^2, \mathbf{x}_{\text{obs}}, \mu_x, \sigma_x^2\}$	$\{y, \beta, \sigma_\epsilon^2, \mathbf{x}_{\text{obs}}, \mu_x, \sigma_x^2, \mathbf{a}, \phi\}$
μ_x	$\{\mathbf{x}_{\text{mis}}, \mathbf{x}_{\text{obs}}, \sigma_x^2\}$	$\{\mathbf{x}_{\text{mis}}, \mathbf{x}_{\text{obs}}, \sigma_x^2\}$
σ_x^2	$\{\mathbf{x}_{\text{mis}}, \mathbf{x}_{\text{obs}}, \mu_x\}$	$\{\mathbf{x}_{\text{mis}}, \mathbf{x}_{\text{obs}}, \mu_x\}$
\mathbf{a}	$\{y, \mathbf{R}, \phi\}$	$\{\mathbf{x}_{\text{mis}}, \mathbf{x}_{\text{obs}}, \mathbf{R}, \phi\}$
ϕ	$\{\mathbf{a}, y\}$	$\{\mathbf{x}_{\text{mis}}, \mathbf{x}_{\text{obs}}, \mathbf{a}\}$

One can also use *d-separation theory* (Pearl 1988; see also section 8.2 of Bishop 2006) to establish that, under MAR,

$$\{\beta, \sigma_\epsilon^2, \mathbf{x}_{\text{mis}}, \mu_x, \sigma_x^2\} \perp\!\!\!\perp \{\mathbf{a}, \phi\} | \{y, \mathbf{x}_{\text{obs}}, \mathbf{R}\},$$

where $\mathbf{u} \perp\!\!\!\perp \mathbf{v} | \mathbf{w}$ denotes conditional independence of \mathbf{u} and \mathbf{v} given \mathbf{w} . The key to this result is the fact that all paths from the nodes in $\{\beta, \sigma_\epsilon^2, \mathbf{x}_{\text{mis}}, \mu_x, \sigma_x^2\}$ to those in $\{\mathbf{a}, \phi\}$ must pass through the y node. In Figure 1 we see that the y node has ‘head-to-tail’ pairs of edges that block the path between \mathbf{a} and the regression parameters.

3.3 Approximate Inference via Variational Bayes

We will now provide details on approximate inference for the simple linear regression missing data models with predictors MNAR. Details for the simpler MCAR and MAR cases are given in Supplement A of the supplemental materials.

As we shall see, variational Bayes boils down to iterative schemes for the parameters of the optimal q densities. The current subsection does little more than listing the algorithm for variational Bayes inference. Section 3.4 addresses accuracy of this algorithm and its MCAR analogue.

For a generic random variable v and density function $q(v)$ let

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

Also, in the special case that $q(v)$ is an Inverse Gamma density function we let

$$(A_{q(v)}, B_{q(v)}) \equiv \text{shape and rate parameters of } q(v).$$

In other words, $v \sim \text{IG}(A_{q(v)}, B_{q(v)})$. Note the relationship $\mu_{q(1/v)} = A_{q(v)}/B_{q(v)}$. For a generic random vector \mathbf{v} and density function $q(\mathbf{v})$ let $\mu_{q(\mathbf{v})} \equiv E_q(\mathbf{v})$ and

$$\Sigma_{q(\mathbf{v})} \equiv \text{Cov}_q(\mathbf{v}) = \text{covariance matrix of } \mathbf{v} \text{ under density } q(\mathbf{v}).$$

To avoid notational clutter we will omit the asterisk when applying these definitions to the optimal q^* densities.

The updates for $E_{q(\mathbf{x}_{\text{mis}})}(\mathbf{X})$ and $E_{q(\mathbf{x}_{\text{mis}})}(\mathbf{X}^T \mathbf{X})$ are the same for each of Algorithms 1, S.1, and S.2 (the latter two algorithms are in the online supplemental materials) so we list them here:

$$E_{q(\mathbf{x}_{\text{mis}})}(\mathbf{X}) \leftarrow \begin{bmatrix} \mathbf{1} & \mathbf{x}_{\text{obs}} \\ \mathbf{1} & \mu_{q(\mathbf{x}_{\text{mis}})} \end{bmatrix},$$

$$E_{q(\mathbf{x}_{\text{mis}})}(\mathbf{X}^T \mathbf{X}) \leftarrow \begin{bmatrix} n & & & \\ \mathbf{1}^T \mathbf{x}_{\text{obs}} + \mathbf{1}^T \mu_{q(\mathbf{x}_{\text{mis}})} & & & \\ & \mathbf{1}^T \mathbf{x}_{\text{obs}} + \mathbf{1}^T \mu_{q(\mathbf{x}_{\text{mis}})} & & \\ & & \|\mathbf{x}_{\text{obs}}\|^2 + \|\mu_{q(\mathbf{x}_{\text{mis}})}\|^2 + n_{\text{mis}} \sigma_{q(\mathbf{x}_{\text{mis}})}^2 & \end{bmatrix}. \quad (8)$$

Algorithm 1 Iterative scheme for obtaining the parameters in the optimal densities $q^*(\boldsymbol{\beta})$, $q^*(\sigma_\varepsilon^2)$, $q^*(\mu_x)$, $q^*(\sigma_x^2)$, $q^*(x_{\text{mis},i})$, and $q^*(\boldsymbol{\phi})$ for the MNAR simple linear regression model.

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

Cycle:

$$\sigma_{q(x_{\text{mis}})}^2 \leftarrow 1 / [\mu_{q(1/\sigma_\varepsilon^2)} + \mu_{q(1/\sigma_x^2)} \{ \mu_{q(\beta_1)}^2 + (\boldsymbol{\Sigma}_{q(\boldsymbol{\beta})})_{22} \} + \mu_{q(\phi_1)}^2 + (\boldsymbol{\Sigma}_{q(\boldsymbol{\phi})})_{22}]$$

for $i = 1, \dots, n_{\text{mis}}$:

$$\begin{aligned} \mu_{q(x_{\text{mis},i})} \leftarrow & \sigma_{q(x_{\text{mis}})}^2 [\mu_{q(1/\sigma_x^2)} \mu_{q(\mu_x)} + \mu_{q(1/\sigma_\varepsilon^2)} \{ y_{x_{\text{mis},i}} \mu_{q(\beta_1)} - (\boldsymbol{\Sigma}_{q(\boldsymbol{\beta})})_{12} - \mu_{q(\beta_0)} \mu_{q(\beta_1)} \} \\ & + \mu_{q(a_{x_{\text{mis},i}})} \mu_{q(\phi_1)} - (\boldsymbol{\Sigma}_{q(\boldsymbol{\phi})})_{12} - \mu_{q(\phi_0)} \mu_{q(\phi_1)}], \end{aligned}$$

update $E_{q(\mathbf{x}_{\text{mis}})}(\mathbf{X})$ and $E_{q(\mathbf{x}_{\text{mis}})}(\mathbf{X}^T \mathbf{X})$ using (8)

$$\boldsymbol{\Sigma}_{q(\boldsymbol{\beta})} \leftarrow \left\{ \mu_{q(1/\sigma_\varepsilon^2)} E_{q(\mathbf{x}_{\text{mis}})}(\mathbf{X}^T \mathbf{X}) + \frac{1}{\sigma_\beta^2} \mathbf{I} \right\}^{-1}; \quad \boldsymbol{\mu}_{q(\boldsymbol{\beta})} \leftarrow \boldsymbol{\Sigma}_{q(\boldsymbol{\beta})} \mu_{q(1/\sigma_\varepsilon^2)} E_{q(\mathbf{x}_{\text{mis}})}(\mathbf{X})^T \mathbf{y}$$

$$\sigma_{q(\mu_x)}^2 \leftarrow 1 / (n \mu_{q(1/\sigma_x^2)} + 1/\sigma_{\mu_x}^2); \quad \mu_{q(\mu_x)} \leftarrow \sigma_{q(\mu_x)}^2 \mu_{q(1/\sigma_x^2)} (\mathbf{1}^T \mathbf{x}_{\text{obs}} + \mathbf{1}^T \boldsymbol{\mu}_{q(\mathbf{x}_{\text{mis}})})$$

$$B_{q(\sigma_\varepsilon^2)} \leftarrow B_\varepsilon + \frac{1}{2} \|\mathbf{y}\|^2 - \mathbf{y}^T E_{q(\mathbf{x}_{\text{mis}})}(\mathbf{X}) \boldsymbol{\mu}_{q(\boldsymbol{\beta})} + \frac{1}{2} \text{tr} \{ E_{q(\mathbf{x}_{\text{mis}})}(\mathbf{X}^T \mathbf{X}) (\boldsymbol{\Sigma}_{q(\boldsymbol{\beta})} + \boldsymbol{\mu}_{q(\boldsymbol{\beta})} \boldsymbol{\mu}_{q(\boldsymbol{\beta})}^T) \}$$

$$B_{q(\sigma_x^2)} \leftarrow B_x + \frac{1}{2} (\|\mathbf{x}_{\text{obs}} - \mu_{q(\mu_x)} \mathbf{1}\|^2 + \|\boldsymbol{\mu}_{q(\mathbf{x}_{\text{mis}})} - \mu_{q(\mu_x)} \mathbf{1}\|^2 + n \sigma_{q(\mu_x)}^2 + n_{\text{mis}} \sigma_{q(x_{\text{mis}})}^2)$$

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

$$\boldsymbol{\Sigma}_{q(\boldsymbol{\phi})} \leftarrow \left\{ E_{q(\mathbf{x}_{\text{mis}})}(\mathbf{X}^T \mathbf{X}) + \frac{1}{\sigma_\phi^2} \mathbf{I} \right\}^{-1}; \quad \boldsymbol{\mu}_{q(\boldsymbol{\phi})} \leftarrow \boldsymbol{\Sigma}_{q(\boldsymbol{\phi})} E_{q(\mathbf{x}_{\text{mis}})}(\mathbf{X})^T \boldsymbol{\mu}_{q(\mathbf{a})}$$

$$\boldsymbol{\mu}_{q(\mathbf{a})} \leftarrow E_{q(\mathbf{x}_{\text{mis}})}(\mathbf{X}) \boldsymbol{\mu}_{q(\boldsymbol{\phi})} + (2\mathbf{R} - \mathbf{1}) \odot \frac{(2\pi)^{-1/2} \exp\{-1/2(E_{q(\mathbf{x}_{\text{mis}})}(\mathbf{X}) \boldsymbol{\mu}_{q(\boldsymbol{\phi})})^2\}}{\Phi((2\mathbf{R} - \mathbf{1}) \odot (E_{q(\mathbf{x}_{\text{mis}})}(\mathbf{X}) \boldsymbol{\mu}_{q(\boldsymbol{\phi})}))}$$

until the increase in $p(\mathbf{y}, \mathbf{x}_{\text{obs}}, \mathbf{R}; q)$ is negligible.

For the simple linear regression model, with predictors MNAR, we impose the product density restriction

$$q(\boldsymbol{\beta}, \sigma_\varepsilon^2, \mathbf{x}_{\text{mis}}, \mu_x, \sigma_x^2, \boldsymbol{\phi}, \mathbf{a}) = q(\boldsymbol{\beta}, \mu_x, \boldsymbol{\phi}) q(\sigma_\varepsilon^2, \sigma_x^2) q(\mathbf{x}_{\text{mis}}) q(\mathbf{a}).$$

The variational Bayes approximate posterior density functions for the model parameters have the forms:

$q^*(\boldsymbol{\beta}) =$ Bivariate Normal density,

$q^*(\sigma_\varepsilon^2) =$ Inverse Gamma density,

$q^*(\mathbf{x}_{\text{mis}}) =$ product of n_{mis} univariate Normal densities,

$q^*(\mu_x) =$ univariate normal density,

$q^*(\sigma_x^2) =$ Inverse Gamma density,

$q^*(\boldsymbol{\phi}) =$ Bivariate Normal density,

and

$q^*(\mathbf{a}) =$ product of n Truncated Normal densities.

Supplement C, in the supplemental materials, provides full expressions for these density functions, as well as their derivation. The parameters for these approximate posterior density functions may be obtained via Algorithm 1. Note that, for $1 \leq i \leq n_{\text{mis}}$, $a_{x_{\text{mis},i}}$ denotes the entry of \mathbf{a} corresponding to $x_{\text{mis},i}$.

The lower bound on the marginal log-likelihood, $\log\{p(\mathbf{y}, \mathbf{x}_{\text{obs}}, \mathbf{R}; q)\}$, has the explicit expression

$$\begin{aligned} \log p(\mathbf{y}, \mathbf{x}_{\text{obs}}, \mathbf{R}; q) &= \frac{1}{2} (n_{\text{mis}} + 5) - \left(n - \frac{1}{2} n_{\text{mis}} \right) \log(2\pi) + \frac{n_{\text{mis}}}{2} \log(\sigma_{q(x_{\text{mis}})}^2) \\ &+ \frac{1}{2} \log \left| \frac{1}{\sigma_\beta^2} \boldsymbol{\Sigma}_{q(\boldsymbol{\beta})} \right| - \frac{1}{2\sigma_\beta^2} \{ \|\boldsymbol{\mu}_{q(\boldsymbol{\beta})}\|^2 + \text{tr}(\boldsymbol{\Sigma}_{q(\boldsymbol{\beta})}) \} \\ &+ \frac{1}{2} \log(\sigma_{q(\mu_x)}^2 / \sigma_{\mu_x}^2) - \frac{1}{2} (\mu_{q(\mu_x)}^2 + \sigma_{q(\mu_x)}^2) / \sigma_{\mu_x}^2 \\ &+ A_\varepsilon \log(B_\varepsilon) - A_{q(\sigma_\varepsilon^2)} \log(B_{q(\sigma_\varepsilon^2)}) \\ &+ \log \Gamma(A_{q(\sigma_\varepsilon^2)}) - \log \Gamma(A_\varepsilon) \\ &+ A_x \log(B_x) - A_{q(\sigma_x^2)} \log(B_{q(\sigma_x^2)}) \\ &+ \log \Gamma(A_{q(\sigma_x^2)}) - \log \Gamma(A_x) \\ &+ \frac{1}{2} \|E_{q(\mathbf{x}_{\text{mis}})}(\mathbf{X}) \boldsymbol{\mu}_{q(\boldsymbol{\phi})}\|^2 \\ &- \frac{1}{2} \text{tr} \{ E_{q(\mathbf{x}_{\text{mis}})}(\mathbf{X}^T \mathbf{X}) (\boldsymbol{\mu}_{q(\boldsymbol{\phi})} \boldsymbol{\mu}_{q(\boldsymbol{\phi})}^T + \boldsymbol{\Sigma}_{q(\boldsymbol{\phi})}) \} \\ &+ \mathbf{R}^T \log \{ \Phi(E_{q(\mathbf{x}_{\text{mis}})}(\mathbf{X}) \boldsymbol{\mu}_{q(\boldsymbol{\phi})}) \} \end{aligned}$$

$$\begin{aligned}
 &+ (\mathbf{1} - \mathbf{R})^T \log \{ \mathbf{1} - \Phi(E_{q(\mathbf{x}_{\text{mis}})}(\mathbf{X})\boldsymbol{\mu}_{q(\boldsymbol{\phi})}) \} \\
 &+ \frac{1}{2} \log \left| \frac{1}{\sigma_\phi^2} \boldsymbol{\Sigma}_{q(\boldsymbol{\phi})} \right| - \frac{1}{2\sigma_\phi^2} \{ \|\boldsymbol{\mu}_{q(\boldsymbol{\phi})}\|^2 + \text{tr}(\boldsymbol{\Sigma}_{q(\boldsymbol{\phi})}) \}.
 \end{aligned}$$

Note that, within each iteration of Algorithm 1, this expression applies only after each of the parameter updates has been made.

3.4 Assessment of Accuracy

We now turn attention to the issue of accuracy of variational Bayes inference for models (6). Algorithms 1, S.1, and S.2 provide speedy approximate inference for the model parameters, but come with no guarantees of achieving an acceptable level of accuracy. Here we provide an accuracy assessment of Algorithm 1 using simulated data. An accuracy assessment of Algorithm S.1 is given in Supplement A of the supplemental materials.

Let θ denote a generic univariate parameter. There are numerous means by which the accuracy of a variational Bayes approximate density $q^*(\theta)$ can be measured with respect to the exact posterior density $p(\theta|\mathbf{y})$. Kullback–Leibler distance is an obvious choice but can be dominated by the tail behavior of the densities involved (e.g., Hall 1987). We recommend working with the L_1 loss, or *integrated absolute error* (IAE) of q^* , given by

$$\text{IAE}(q^*) = \int_{-\infty}^{\infty} |q^*(\theta) - p(\theta|\mathbf{y})| d\theta.$$

This error measure has the attractions of being (a) invariant to monotone transformations on the parameter θ and (b) a scale-independent number between 0 and 2 (e.g., Devroye and Györfi 1985). The second of these motivates the accuracy measure

$$\begin{aligned}
 \text{accuracy}(q^*) &= 1 - \left\{ \text{IAE}(q^*) / \sup_{q \text{ a density}} \text{IAE}(q) \right\} \\
 &= 1 - \frac{1}{2} \text{IAE}(q^*). \tag{9}
 \end{aligned}$$

Note that $0 \leq \text{accuracy}(q^*) \leq 1$ and will be expressed as a percentage in the examples to follow.

Computation of $\text{accuracy}(q^*)$ is a little challenging, since it depends on the posterior $p(\theta|\mathbf{y})$ that we are trying to avoid by using approximate inference methods. However, MCMC with sufficiently large samples can be used to approximate $p(\theta|\mathbf{y})$ arbitrarily well. The accuracy assessments that we present in this section are based on MCMC samples obtained using BRugs (Ligges et al. 2010) with a burn-in of size 10,000. A thinning factor of 5 was applied to postburn-in samples of size 50,000. This resulted in MCMC samples of size 10,000 for density estimation. Density estimates were obtained using the binned kernel density estimate `bkde()` function in the R package `KernSmooth` (Wand and Ripley 2009). The bandwidth was chosen using a direct plug-in rule, corresponding to the default version of `dpik()`. These density estimates act as a proxy for the exact posterior densities. For sample sizes as large as 10,000 and well-behaved posteriors the quality of these proxies should be quite good. Nevertheless, it must be noted that they are subject to errors inherent in density estimation and bandwidth selection.

In our simulation for the MNAR model, the missingness is controlled by the two pairs of probit coefficients:

$$\begin{aligned}
 (\phi_0, \phi_1) &= (2.95, -2.95) \quad \text{and} \\
 (\phi_0, \phi_1) &= (0.85, -1.05). \tag{10}
 \end{aligned}$$

In each case, the probability of missingness increases as a function of the covariate. For the first pair the missingness probability ranges from 0.0 to 0.5 with an average of 0.25. For the second pair the range is 0.2 to 0.58 with an average of 0.39, representing more severe missingness. The hyperparameter is set at $\sigma_\phi^2 = 10^8$ to give a noninformative prior distribution for (ϕ_0, ϕ_1) .

Figure 2 summarizes the accuracy results based on 100 simulated datasets while Figure 3 plots the variational Bayes and MCMC approximate posteriors for a typical realization from the simulation study with $\sigma_\epsilon = 0.2$ and $(\phi_0, \phi_1) = (2.95, -2.95)$.

The parameters corresponding to the regression part of the model $(\beta_0, \beta_1, \sigma_\epsilon^2)$ show high accuracy, with almost all accuracy levels above 80%. The accuracy drops considerably when the amount of missing data is large or when the data are noisy. This might be expected since there is a decrease in the amount of information about the parameters. The accuracy of the missing covariates is high in all situations, even when the missing data percentage is very large.

The variational Bayes approximations generally are poor for the missingness mechanism parameters ϕ_0 and ϕ_1 . Whilst variational Bayes tends to do well in terms of location, it gives posterior density functions with a deflated amount of spread. This is due to strong posterior correlation between $\boldsymbol{\phi}$ and \mathbf{a} in probit auxiliary variable models, as is reported in section 2.1 of Holmes and Held (2006), for example. This deficiency of variational Bayes is restricted to the lower nodes of the right-most DAG in Figure 1 and can only be remedied through a more elaborate variational approximation—for example, one that allows posterior dependence between $\boldsymbol{\phi}$ and \mathbf{a} . Such elaboration will bring computational costs, which need to be traded off against the importance of making inference about the MNAR parameters. In many applied contexts, these parameters are not of primary interest.

3.5 Credible Interval Coverage

Another important type of accuracy assessment involves comparison between the advertised coverage of variational Bayes approximate credible intervals and the actual coverage. We carried out this assessment by simulating 10,000 replications from the MCAR and MNAR simple linear regression models with true parameters as given by (10), and (S.3) and (S.4) in Supplement A. Table 2 shows the percentages of true parameter coverage for the approximate 95% credible intervals formed from the variational Bayes posterior densities with 0.025 probability mass in each tail. The margin of error (i.e., twice the asymptotic standard error) is less than 1% for all entries in Table 2.

For the MCAR models the coverage is generally very good and does not fall below 86% for all model parameters and missing x_i 's. For MNAR models, this claim is true for the regression coefficients and error variance. There is some degradation

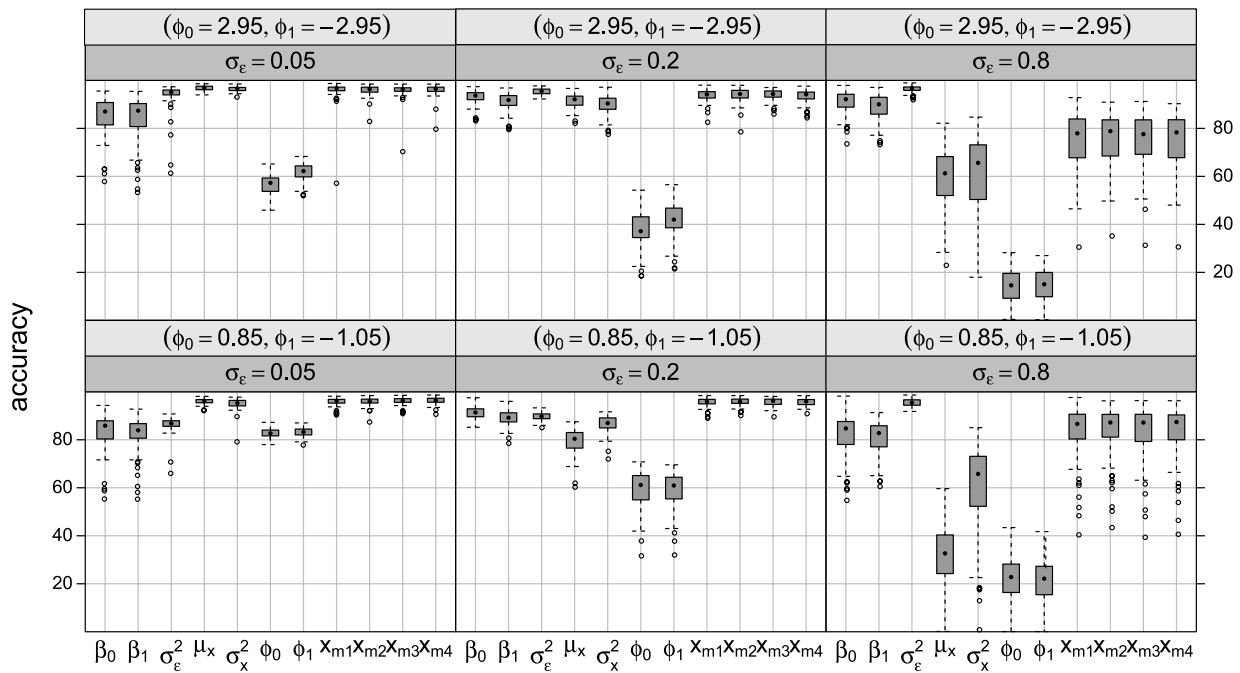


Figure 2. Summary of simulation for simple linear regression with predictor MNAR. For each setting, the accuracy values are summarized as a boxplot.

for μ_x for high noise and missingness. The coverage for the missing data mechanism parameters, ϕ_0 and ϕ_1 , is generally quite poor. Interestingly, the coverage is better in the high missingness case. We do not have an explanation for this counter-intuitive result, except to note that the principles that drive variational Bayes accuracy do not necessarily match those that drive statistical accuracy.

3.6 Prediction Interval Coverage

Prediction intervals for the response are often of interest in missing data problems. For each of the missing data regression models (6) the Bayesian prediction intervals depend on $p(\beta, \sigma_\epsilon^2 | \mathbf{y})$, the joint posterior of the regression coefficients and error variance. In the case of noninformative independent priors there is negligible posterior dependence between β and σ_ϵ^2 . As indicated by Figure 2 and Figure S.1 in Supplement A, the variational Bayes approximation to $p(\beta, \sigma_\epsilon^2 | \mathbf{y})$ is very good and the variational Bayes prediction intervals have good coverage properties, at least for the simulation settings in the current section. We have also visually compared the variational Bayes prediction intervals with their MCMC counterparts for several replications of each simulation setting and found excellent agreement between the two.

3.7 Speed Comparisons

While running the simulation studies described in Section 3.4 we kept track of the time taken for each model to be fitted. The results are summarized in Table 3. The computer involved used the Mac OS X operating system with a 2.33 GHz processor and 3 GBytes of random access memory.

As with most speed comparisons, some caveats need to be taken into account. First, the MCMC and variational Bayes answers 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. 2010) in the R computing environment (R Development Core Team 2010). The variational Bayes model fits were implemented using R. Second, no effort was made to tailor the MCMC scheme to the models at hand. Third, as detailed in Section 3.4, both methods had arbitrarily chosen stopping criteria. Despite these caveats, Table 3 gives an impression of the relative computing times involved if an ‘off-the-shelf’ MCMC implementation is used.

Caveats aside, the results indicate that variational Bayes is at least 60 times faster than MCMC across all models. Hence, a model that takes minutes to run in MCMC takes only seconds with a variational Bayes approach.

4. NONPARAMETRIC REGRESSION WITH MISSING PREDICTOR DATA

We now describe extension to nonparametric regression with missing predictor data. The essence of this extension is replacement of the linear mean function

$$\beta_0 + \beta_1 x \quad \text{by} \quad f(x),$$

where f is a smooth flexible function. There are numerous approaches to modeling and estimating f . The one which is most conducive to inference via variational Bayes is penalized splines with mixed model representation. This involves the model

$$f(x) = \beta_0 + \beta_1 x + \sum_{k=1}^K u_k z_k(x), \quad u_k \stackrel{\text{ind.}}{\sim} N(0, \sigma_u^2), \quad (11)$$

where the $\{z_k(\cdot) : 1 \leq k \leq K\}$ are an appropriate set of spline basis functions. Several options exist for the z_k . Our preference is suitably transformed O’Sullivan penalized splines (Wand

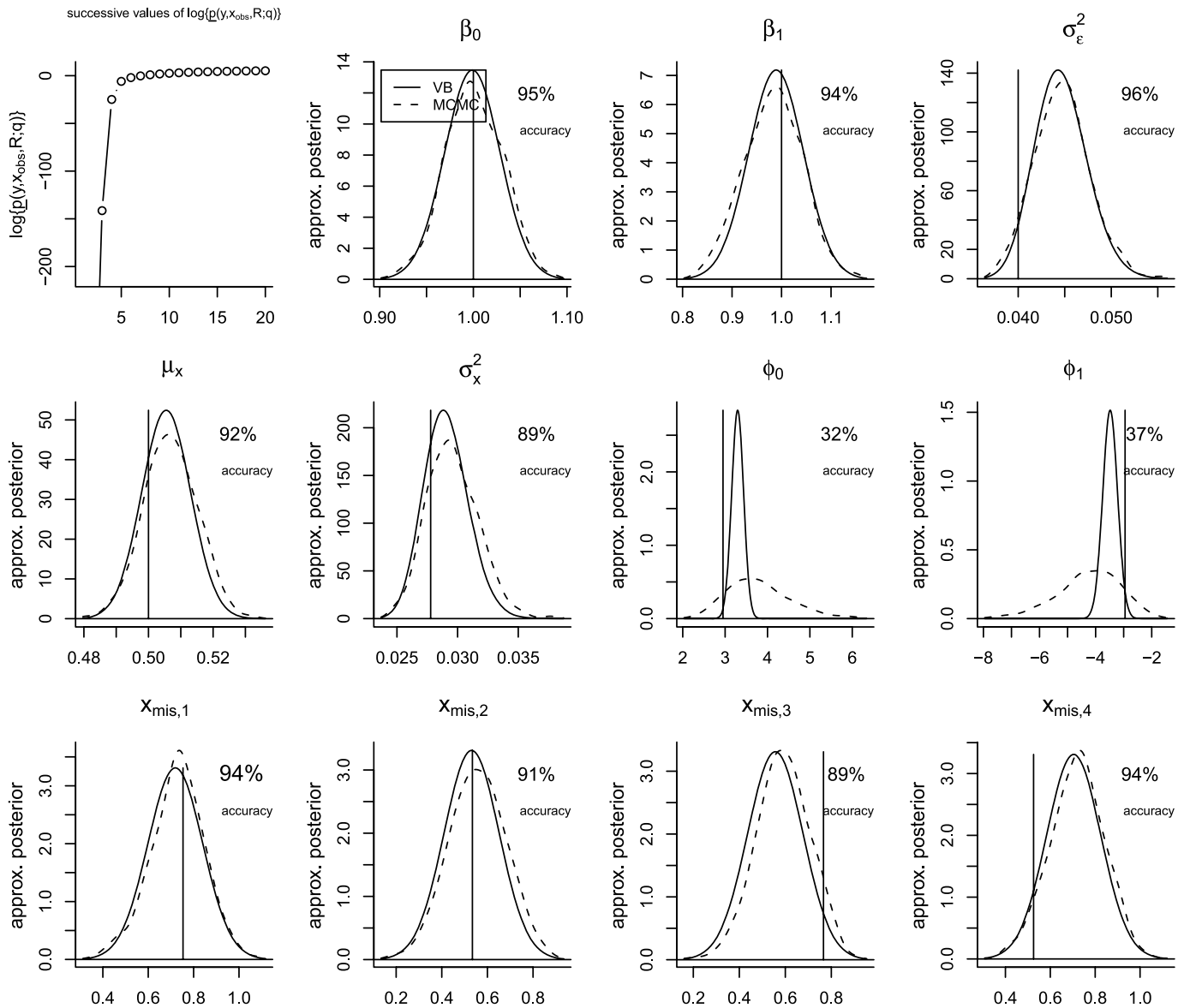


Figure 3. Variational Bayes approximate posteriors for the regression model parameters and three missing x_i 's for simple linear regression with predictors MNAR. The regression parameters are $(\beta_0, \beta_1, \sigma_\varepsilon) = (1, 1, 0.2)$ and the probability of x_i being observed is $\Phi(\phi_0 + \phi_1 x_i)$ where $(\phi_0, \phi_1) = (2.95, -2.95)$. The vertical lines correspond to the true values of the parameters from which the data were simulated (described in the text). The MCMC posteriors are based on samples of size 10,000 and kernel density estimation. The accuracy values correspond to the definition given at (9).

and Ormerod 2008) since this leads to approximate smoothing splines, which have good boundary and extrapolation properties.

From the graphical model standpoint, moving from parametric regression to nonparametric regression using mixed model-based penalized splines simply involves enlarging the DAGs from parametric regression. Figure 4 shows the nonparametric regression DAGs for the three missing data mechanisms treated in Section 3. Comparison with Figure 1 shows the only difference is the addition of the σ_u^2 node, and replacement of β by (β, \mathbf{u}) . Note that (β, \mathbf{u}) could be broken up into separate nodes, but the update expressions are simpler if these two random vectors are kept together.

The variational Bayes algorithms for the DAGs in Figure 4 simply involve modification of Algorithms 1, S.1, and S.2 to ac-

commodate the additional nodes and edges. However, the spline basis functions give rise to nonstandard forms and numerical integration is required. We will give a detailed account of this extension in the MNAR case only. The MCAR and MAR cases require similar arguments, but are simpler.

Define the $1 \times (K + 2)$ vector

$$\mathbf{C}_x \equiv (1, x, z_1(x), \dots, z_K(x))$$

corresponding to evaluation of penalized spline basis functions at an arbitrary location $x \in \mathbb{R}$. Then the optimal densities for the $x_{\text{mis},i}$, $1 \leq i \leq n_{\text{mis}}$, take the form

$$q^*(x_{\text{mis},i}) \propto \exp\left(-\frac{1}{2} \mathbf{C}_{x_{\text{mis},i}} \mathbf{\Lambda}_{\text{mis},i} \mathbf{C}_{x_{\text{mis},i}}^T\right), \quad (12)$$

where the $(K + 2) \times (K + 2)$ matrices $\mathbf{\Lambda}_{\text{mis},i}$, $1 \leq i \leq n_{\text{mis}}$, correspond to each entry of $\mathbf{x}_{\text{mis}} = (x_{\text{mis},1}, \dots, x_{\text{mis},n_{\text{mis}}})$ but does

Table 2. Percentage coverage of true parameter values and the first three missing x_i values by approximate 95% credible intervals based on variational Bayes approximate posterior density functions. Low missingness for the MCAR model corresponds to $p = 0.8$ and high missingness to $p = 0.6$. Low missingness for the MNAR model corresponds to $(\phi_0, \phi_1) = (2.95, -2.95)$ and high missingness to $(\phi_0, \phi_1) = (0.85, -1.05)$.

The percentages are based on 10,000 replications, which guarantees a margin of error (twice the asymptotic standard error) less than 1%

mis'ness	MCAR low miss.			MCAR high miss.			MNAR low miss.			MNAR high miss.			
	σ_ε	0.05	0.2	0.8	0.05	0.2	0.8	0.05	0.2	0.8	0.05	0.2	0.8
β_0		92	89	93	91	92	89	94	94	94	90	91	88
β_1		93	89	93	91	92	88	94	94	94	89	91	88
σ_ε^2		91	86	93	89	95	94	93	94	95	87	90	94
μ_x		94	94	93	90	92	87	95	93	77	94	86	53
σ_x^2		95	94	92	89	92	87	95	93	89	94	89	88
ϕ_0		—	—	—	—	—	—	58	41	11	85	72	18
ϕ_1		—	—	—	—	—	—	64	47	12	85	73	16
$x_{\text{mis},1}$		95	95	95	95	95	94	95	94	92	95	95	94
$x_{\text{mis},2}$		95	95	95	95	95	95	95	94	92	95	95	94
$x_{\text{mis},3}$		95	95	95	95	95	95	95	94	91	95	95	94

not depend on $x_{\text{mis},i}$. A derivation of (12) and expressions for the $\Lambda_{\text{mis},i}$ are given in Supplement D of the supplemental materials.

The right-hand side of (12) does not have a closed-form integral, so numerical integration is required to obtain the normalizing factors and required moments. We will take a basic quadrature approach. In the interests of computational efficiency, we use the same quadrature grid over all $1 \leq i \leq n_{\text{mis}}$. Let

$$\mathbf{g} = (g_1, \dots, g_M)$$

be an equally spaced grid of size M in \mathbb{R} . An example of numerical integration via quadrature is

$$\int_{-\infty}^{\infty} z_1(x) dx \approx \sum_{j=1}^M w_j z_1(g_j) = \mathbf{w}^T z_1(\mathbf{g}),$$

where $\mathbf{w} = (w_1, \dots, w_M)$ is vector of quadrature weights. Examples of \mathbf{w} for common quadrature schemes are

$$\mathbf{w} = \begin{cases} \frac{1}{2}\delta \times (1, 2, 2, 2, 2, 2, \dots, 2, 2, 2, 1) & \text{for the trapezoidal rule,} \\ \frac{1}{3}\delta \times (1, 4, 2, 4, 2, 4, 2, \dots, 4, 2, 4, 1) & \text{for Simpson's rule,} \end{cases}$$

where $\delta = (g_M - g_1)/(M - 1)$ is the distance between successive grid points. Next, define the $M \times (K + 2)$ matrix:

$$\mathbf{C}_{\mathbf{g}} \equiv \begin{bmatrix} 1 & g_1 & z_1(g_1) & \cdots & z_K(g_1) \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & g_M & z_1(g_M) & \cdots & z_K(g_M) \end{bmatrix} = \begin{bmatrix} \mathbf{C}_{g_1} \\ \vdots \\ \mathbf{C}_{g_M} \end{bmatrix}. \quad (13)$$

Table 3. 99% Wilcoxon confidence intervals based on computation times, in seconds, from the simulation study described in Section 3.4

	MAR models	MNAR models
MCMC	(5.89, 5.84)	(33.8, 33.9)
Var. Bayes	(0.0849, 0.0850)	(0.705, 0.790)
Ratio	(76.6, 78.7)	(59.5, 67.8)

For a given quadrature grid \mathbf{g} , $\mathbf{C}_{\mathbf{g}}$ contains the totality of basis function evaluations required for variational Bayes updates.

For succinct statement of quadrature approximations to $E_{q(\mathbf{x}_{\text{mis}})}(\mathbf{C})$ and $E_{q(\mathbf{x}_{\text{mis}})}(\mathbf{C}^T \mathbf{C})$ the following additional matrix notation is useful:

$$\mathbf{Q}_{\mathbf{g}} \equiv \left[\exp\left(-\frac{1}{2} \mathbf{C}_{g_j} \Lambda_{\text{mis},i} \mathbf{C}_{g_j}^T\right) \right]_{\substack{1 \leq j \leq M \\ 1 \leq i \leq n_{\text{mis}}}} \quad \text{and}$$

$$\mathbf{C} \equiv \begin{bmatrix} \mathbf{C}_{\text{obs}} \\ \mathbf{C}_{\text{mis}} \end{bmatrix},$$

where \mathbf{C}_{obs} corresponds to the \mathbf{x}_{obs} component of \mathbf{C} and \mathbf{C}_{mis} corresponds to the \mathbf{x}_{mis} component of \mathbf{C} . Clearly

$$E_{q(\mathbf{x}_{\text{mis}})}(\mathbf{C}) \equiv \begin{bmatrix} \mathbf{C}_{\text{obs}} \\ E_{q(\mathbf{x}_{\text{mis}})}(\mathbf{C}_{\text{mis}}) \end{bmatrix} \quad \text{and}$$

$$E_{q(\mathbf{x}_{\text{mis}})}(\mathbf{C}^T \mathbf{C}) = \mathbf{C}_{\text{obs}}^T \mathbf{C}_{\text{obs}} + E_{q(\mathbf{x}_{\text{mis}})}(\mathbf{C}_{\text{mis}}^T \mathbf{C}_{\text{mis}}).$$

Then we have the following efficient quadrature approximations:

$$E_{q(\mathbf{x}_{\text{mis}})}(\mathbf{C}_{\text{mis}}) \approx \frac{\mathbf{Q}_{\mathbf{g}} \text{diag}(\mathbf{w}) \mathbf{C}_{\mathbf{g}}}{\mathbf{1}^T \otimes (\mathbf{Q}_{\mathbf{g}} \mathbf{w})} \quad \text{and}$$

$$E_{q(\mathbf{x}_{\text{mis}})}(\mathbf{C}_{\text{mis}}^T \mathbf{C}_{\text{mis}}) \approx \mathbf{C}_{\mathbf{g}}^T \text{diag}\left(\sum_{i=1}^{n_{\text{mis}}} \frac{\mathbf{e}_i^T \mathbf{Q}_{\mathbf{g}} \odot \mathbf{w}}{\mathbf{e}_i^T \mathbf{Q}_{\mathbf{g}} \mathbf{w}}\right) \mathbf{C}_{\mathbf{g}}$$

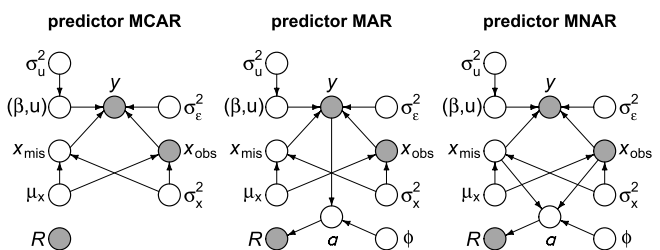


Figure 4. DAGs for the three missing data models for nonparametric regression with mixed model-based penalized spline modeling of the regression function, given by (11). Shaded nodes correspond to the observed data.

with \mathbf{e}_i denoting the $n_{\text{mis}} \times 1$ vector with 1 in the i th position and zeroes elsewhere. Since there are exponentials in entries of \mathbf{Q}_g , some care needs to be taken to avoid overflow and underflow. Working with logarithms is recommended.

Algorithm 2 chronicles the iterative scheme for nonparametric regression with predictors MNAR. The lower bound on the marginal log-likelihood is

$$\begin{aligned} & \log\{p(\mathbf{y}, \mathbf{x}_{\text{obs}}, \mathbf{R}; q)\} \\ &= \left(\frac{1}{2}n_{\text{mis}} - n\right) \log(2\pi) + \frac{1}{2}(K + 5 + n_{\text{mis}}) - \log(\sigma_\beta^2) \\ & \quad - \frac{1}{2\sigma_\beta^2} [\|\boldsymbol{\mu}_{q(\beta)}\|^2 + \text{tr}(\boldsymbol{\Sigma}_{q(\beta)})] + \frac{1}{2} \log|\boldsymbol{\Sigma}_{q(\beta, \mathbf{u})}| \end{aligned}$$

$$\begin{aligned} & + \frac{1}{2} \log\{\sigma_{q(\mu_x)}^2 / \sigma_{\mu_x}^2\} - \frac{1}{2\sigma_{\mu_x}^2} \{\mu_{q(\mu_x)}^2 + \sigma_{q(\mu_x)}^2\} \\ & - \frac{\mathbf{Q}_g \text{diag}(\mathbf{w}) \log(\mathbf{Q}_g)}{\mathbf{1}^T \otimes (\mathbf{Q}_g \mathbf{w})} \\ & + A_\varepsilon \log(B_\varepsilon) - \log \Gamma(A_\varepsilon) \\ & - A_{q(\sigma_\varepsilon^2)} \log(B_{q(\sigma_\varepsilon^2)}) + \log(A_{q(\sigma_\varepsilon^2)}) \\ & + A_u \log(B_u) - \log \Gamma(A_u) \\ & - A_{q(\sigma_u^2)} \log(B_{q(\sigma_u^2)}) + \log(A_{q(\sigma_u^2)}) \\ & + A_x \log(B_x) - \log \Gamma(A_x) \\ & - A_{q(\sigma_x^2)} \log(B_{q(\sigma_x^2)}) + \log(A_{q(\sigma_x^2)}) \end{aligned}$$

Algorithm 2 Iterative scheme for obtaining the parameters in the optimal densities $q^*(\boldsymbol{\beta}, \mathbf{u})$, $q^*(\sigma_\varepsilon^2)$, $q^*(\sigma_u^2)$, $q^*(\mu_x)$, $q^*(\sigma_x^2)$, $q^*(x_{\text{mis}, i})$, and $q^*(\boldsymbol{\phi})$ for the MNAR nonparametric regression model.

Set M , the size of the quadrature grid, and g_1 and g_M , the quadrature grid limits. The interval (g_1, g_M) should contain each of the observed x_i 's. Obtain $\mathbf{g} = (g_1, \dots, g_M)$ where $g_j = g_1 + (j - 1)\delta$, $1 \leq j \leq M$, and $\delta = (g_M - g_1)/(M - 1)$. Obtain the quadrature weights $\mathbf{w} = (w_1, \dots, w_M)$ and set \mathbf{C}_g using (13). Initialize: $\mu_{q(1/\sigma_\varepsilon^2)}, \mu_{q(1/\sigma_x^2)} > 0$, $\mu_{q(\mu_x)}, \boldsymbol{\mu}_{q(\beta, \mathbf{u})} ((K + 2) \times 1)$, $\boldsymbol{\Sigma}_{q(\beta, \mathbf{u})} ((K + 2) \times (K + 2))$, $\boldsymbol{\mu}_{q(\boldsymbol{\phi})} (2 \times 1)$, $\boldsymbol{\Sigma}_{q(\boldsymbol{\phi})} (2 \times 1)$, and $\boldsymbol{\mu}_{q(\mathbf{a})} (n \times 1)$.

Cycle:

update $\boldsymbol{\Lambda}_{\text{mis}, i}$, $1 \leq i \leq n_{\text{mis}}$, using (S.5)–(S.8) in Supplement D.

$$\mathbf{Q}_g \leftarrow \left[\exp\left(-\frac{1}{2} \mathbf{C}_{g_j} \boldsymbol{\Lambda}_{\text{mis}, i} \mathbf{C}_{g_j}^T\right) \right]_{1 \leq j \leq M}; \quad E_{q(\mathbf{x}_{\text{mis}})}(\mathbf{C}) \leftarrow \left[\frac{\mathbf{C}_{\text{obs}}}{\mathbf{Q}_g \text{diag}(\mathbf{w}) \mathbf{C}_g} \right]$$

for $i = 1, \dots, n_{\text{mis}}$:

$$\mu_{q(x_{\text{mis}, i})} \leftarrow \{E_{q(\mathbf{x}_{\text{mis}})}(\mathbf{C}_{\text{mis}})\}_{i2}; \quad \sigma_{q(x_{\text{mis}, i})}^2 \leftarrow \frac{\mathbf{Q}_g \text{diag}(\mathbf{w})(\mathbf{g} - \mu_{q(x_{\text{mis}, i})} \mathbf{1})^2}{\mathbf{1}^T \otimes (\mathbf{Q}_g \mathbf{w})}$$

$$E_{q(\mathbf{x}_{\text{mis}})}(\mathbf{C}^T \mathbf{C}) \leftarrow \mathbf{C}_{\text{obs}}^T \mathbf{C}_{\text{obs}} + \mathbf{C}_g^T \text{diag}\left(\sum_{i=1}^{n_{\text{mis}}} \frac{\mathbf{e}_i^T \mathbf{Q}_g \odot \mathbf{w}}{\mathbf{e}_i^T \mathbf{Q}_g \mathbf{w}}\right) \mathbf{C}_g$$

$$\boldsymbol{\Sigma}_{q(\beta, \mathbf{u})} \leftarrow \left\{ \mu_{q(1/\sigma_\varepsilon^2)} E_{q(\mathbf{x}_{\text{mis}})}(\mathbf{C}^T \mathbf{C}) + \frac{1}{\sigma_\beta^2} \mathbf{I} \right\}^{-1}$$

$$\boldsymbol{\mu}_{q(\beta, \mathbf{u})} \leftarrow \boldsymbol{\Sigma}_{q(\beta, \mathbf{u})} \mu_{q(1/\sigma_\varepsilon^2)} E_{q(\mathbf{x}_{\text{mis}})}(\mathbf{C})^T \mathbf{y}$$

$$\sigma_{q(\mu_x)}^2 \leftarrow 1 / (n \mu_{q(1/\sigma_x^2)} + 1/\sigma_{\mu_x}^2); \quad \mu_{q(\mu_x)} \leftarrow \sigma_{q(\mu_x)}^2 \mu_{q(1/\sigma_x^2)} (\mathbf{1}^T \mathbf{x}_{\text{obs}} + \mathbf{1}^T \boldsymbol{\mu}_{q(\mathbf{x}_{\text{mis}})})$$

$$B_{q(\sigma_\varepsilon^2)} \leftarrow B_\varepsilon + \frac{1}{2} \|\mathbf{y}\|^2 - \mathbf{y}^T E_{q(\mathbf{x}_{\text{mis}})}(\mathbf{C}) \boldsymbol{\mu}_{q(\beta, \mathbf{u})} + \frac{1}{2} \text{tr}\{E_{q(\mathbf{x}_{\text{mis}})}(\mathbf{C}^T \mathbf{C})(\boldsymbol{\Sigma}_{q(\beta, \mathbf{u})} + \boldsymbol{\mu}_{q(\beta, \mathbf{u})} \boldsymbol{\mu}_{q(\beta, \mathbf{u})}^T)\}$$

$$B_{q(\sigma_u^2)} \leftarrow B_u + \frac{1}{2} \{\|\boldsymbol{\mu}_{q(\mathbf{u})}\|^2 + \text{tr}(\boldsymbol{\Sigma}_{q(\mathbf{u})})\}$$

$$B_{q(\sigma_x^2)} \leftarrow B_x + \frac{1}{2} \left(\|\mathbf{x}_{\text{obs}} - \mu_{q(\mu_x)} \mathbf{1}\|^2 + \|\boldsymbol{\mu}_{q(\mathbf{x}_{\text{mis}})} - \mu_{q(\mu_x)} \mathbf{1}\|^2 + n \sigma_{q(\mu_x)}^2 + \sum_{i=1}^{n_{\text{mis}}} \sigma_{q(x_{\text{mis}, i})}^2 \right)$$

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

$$\boldsymbol{\Sigma}_{q(\boldsymbol{\phi})} \leftarrow \left\{ E_{q(\mathbf{x}_{\text{mis}})}(\mathbf{X}^T \mathbf{X}) + \frac{1}{\sigma_\phi^2} \mathbf{I} \right\}^{-1}; \quad \boldsymbol{\mu}_{q(\boldsymbol{\phi})} \leftarrow \boldsymbol{\Sigma}_{q(\boldsymbol{\phi})} E_{q(\mathbf{x}_{\text{mis}})}(\mathbf{X})^T \boldsymbol{\mu}_{q(\mathbf{a})}$$

$$\boldsymbol{\mu}_{q(\mathbf{a})} \leftarrow E_{q(\mathbf{x}_{\text{mis}})}(\mathbf{X}) \boldsymbol{\mu}_{q(\boldsymbol{\phi})} + (2\mathbf{R} - \mathbf{1}) \odot \frac{(2\pi)^{-1/2} \exp\{-1/2(E_{q(\mathbf{x}_{\text{mis}})}(\mathbf{X}) \boldsymbol{\mu}_{q(\boldsymbol{\phi})})^2\}}{\Phi((2\mathbf{R} - \mathbf{1}) \odot (E_{q(\mathbf{x}_{\text{mis}})}(\mathbf{X}) \boldsymbol{\mu}_{q(\boldsymbol{\phi})}))}$$

until the increase in $p(\mathbf{y}, \mathbf{x}_{\text{obs}}, \mathbf{R}; q)$ is negligible.

$$\begin{aligned}
 & + \frac{1}{2} \|E_{q(\mathbf{x}_{\text{mis}})}(\mathbf{X})\boldsymbol{\mu}_{q(\boldsymbol{\phi})}\|^2 \\
 & - \frac{1}{2} \text{tr}\{E_{q(\mathbf{x}_{\text{mis}})}(\mathbf{X}^T\mathbf{X})(\boldsymbol{\mu}_{q(\boldsymbol{\phi})}\boldsymbol{\mu}_{q(\boldsymbol{\phi})}^T + \boldsymbol{\Sigma}_{q(\boldsymbol{\phi})})\} \\
 & + \mathbf{R}^T \log \Phi(E_{q(\mathbf{x}_{\text{mis}})}(\mathbf{X})\boldsymbol{\mu}_{q(\boldsymbol{\phi})}) \\
 & + (\mathbf{1} - \mathbf{R})^T \log\{\mathbf{1} - \Phi(E_{q(\mathbf{x}_{\text{mis}})}(\mathbf{X})\boldsymbol{\mu}_{q(\boldsymbol{\phi})})\} \\
 & + \frac{1}{2} \log \left| \frac{1}{\sigma_\phi^2} \boldsymbol{\Sigma}_{q(\boldsymbol{\phi})} \right| - \frac{1}{2\sigma_\phi^2} \{ \|\boldsymbol{\mu}_{q(\boldsymbol{\phi})}\|^2 + \text{tr}(\boldsymbol{\Sigma}_{q(\boldsymbol{\phi})}) \}.
 \end{aligned}$$

4.1 Illustration

Our first illustration involves data simulated according to

$$y_i \sim N(f(x_i), \sigma_\epsilon^2), \quad f(x) = \sin(4\pi x), \quad x_i \sim N\left(\frac{1}{2}, \frac{1}{36}\right),$$

and

$$\sigma_\epsilon^2 = 0.35, \quad 1 \leq i \leq 300,$$

and with 20% of the x_i 's removed completely at random. This simulation setting, with identical parameters, was also used in Wand (2009).

We applied the MCAR analogue of Algorithm 2 and compared the results with MCMC fitting via BRugs. The penalized splines used the truncated linear spline basis with 30 knots: $z_k(x) = (x - \kappa_k)_+$, $1 \leq k \leq 30$, with the knots equally spaced over the range of the observed x_i 's. Truncated linear splines were used to allow straightforward coding in BUGS. If a comparison with MCMC is not being done then O'Sullivan splines are recommended for variational Bayesian inference in this context. The hyperparameters were set at the values

$$\sigma_\beta^2 = \sigma_{\mu_x}^2 = 10^8 \quad \text{and} \quad A_\epsilon = B_\epsilon = A_x = B_x = \frac{1}{100}. \quad (14)$$

The MCMC sampling involved a burnin of size 20,000, and a thinning factor of 20 applied to postburn-in samples of size 200,000 resulting in samples of size 10,000 being retained for inference. In addition, we used the over-relaxed form of MCMC (Neal 1998). In BRugs this involves setting overRelax=TRUE in the modelUpdate() function. Using these settings, all chains appeared to behave reasonably well.

The resulting posterior densities for the model parameters and three randomly chosen missing x_i values are shown in Figure S.3 in Supplement B of the supplemental materials. The vertical lines correspond to the true values, except σ_u^2 where 'truth' is not readily defined. Good to excellent accuracy of variational Bayes is apparent for all posterior densities. There is some noticeable discordance in the case of σ_u^2 . This is perhaps due to some lack of identifiability for this parameter.

A novel aspect of this example is the multimodality of the posteriors for the $x_{\text{mis},i}$. This arises from the periodic nature of f , since more than one x conforms with a particular y . It is noteworthy that the variational Bayes approximations are able to handle this multimodality quite well.

We then applied Algorithm 2 to data simulated according to

$$y_i \sim N(f(x_i), \sigma_\epsilon^2), \quad f(x) = \sin(4\pi x^2), \quad x_i \sim N\left(\frac{1}{2}, \frac{1}{36}\right),$$

and

$$\sigma_\epsilon^2 = 0.35, \quad 1 \leq i \leq 500,$$

and the observed predictor indicators generated according to

$$R_i \sim \text{Bernoulli}(\Phi(\phi_0 + \phi_1 x_i)) \quad \text{with } \phi_0 = 3 \text{ and } \phi_1 = -3.$$

The hyperparameters were as in (14) and $\sigma_\phi^2 = 10^8$. We also ran an MCMC analysis using BRugs. The spline basis functions and MCMC sample sizes were the same as those used in the MCAR example. Figure S.4 in Supplement B shows the resulting posterior density functions. As with the parametric regression examples, variational Bayes is seen to have good to excellent performance for all parameters except ϕ_0 and ϕ_1 .

Our last example involves two variables from Ozone data-frame (source: Breiman and Friedman 1985) in the R package mlbench (Leisch and Dimitriadou 2009). The response variable is daily maximum one-hour-average ozone level and the predictor variable is daily temperature (degrees Fahrenheit) at El Monte, California, U.S.A. The Ozone data frame is such that five of the response values are missing and 137 of the predictor values are missing. So that we could apply the methodology of the current section directly, we omitted the five records for which the response was missing. This resulted in a sample size of $n = 361$ with $n_{\text{mis}} = 137$ missing predictor values.

Preliminary checks shown the normality assumption for the predictors and errors, along with homoscedasticity, to be quite reasonable. We then assumed MNAR nonparametric regression model and fed the standardized data into Algorithm 2. MCMC fitting of the same model via BRugs was also done for comparison. The results were then transformed to the original scale. Figure 5 shows resulting posterior density functions approximations.

In Figure 6 the fitted function estimates for all three examples are shown. Good agreement is seen between variational Bayes and MCMC.

Finally, it is worth noting that these three penalized spline examples had much bigger speed increases for variational Bayes compared with MCMC in BUGS. The total elapsed time for the variational Bayes analysis was 75 seconds. For BRugs, with the MCMC sample sizes described above, the three examples required 15.5 hours to run. This corresponds to a speed-up in the order of several hundreds.

5. DISCUSSION

We have derived variational Bayes algorithms for fast approximate inference in parametric and nonparametric regression with missing predictor data. The central finding of this article is that, for using regression models with missing predictor data, variational Bayes inference achieves good to excellent accuracy for the main parameters of interest. Poor accuracy is realized for the missing data mechanism parameters. As we note at the end of Section 3.4, better accuracy for these auxiliary parameters maybe achievable with a more elaborate variational scheme—in situations where they are of interest. The nonparametric regression examples illustrate that variational Bayes approximates multimodal posterior densities with a high degree of accuracy.

The article has been confined to single predictor models so that the main ideas could be maximally elucidated. Numerous extensions could be made relatively straightforwardly, based on the methodology developed here. Examples include missing response data, generalized responses, multiple regression, additive models, and additive mixed models.

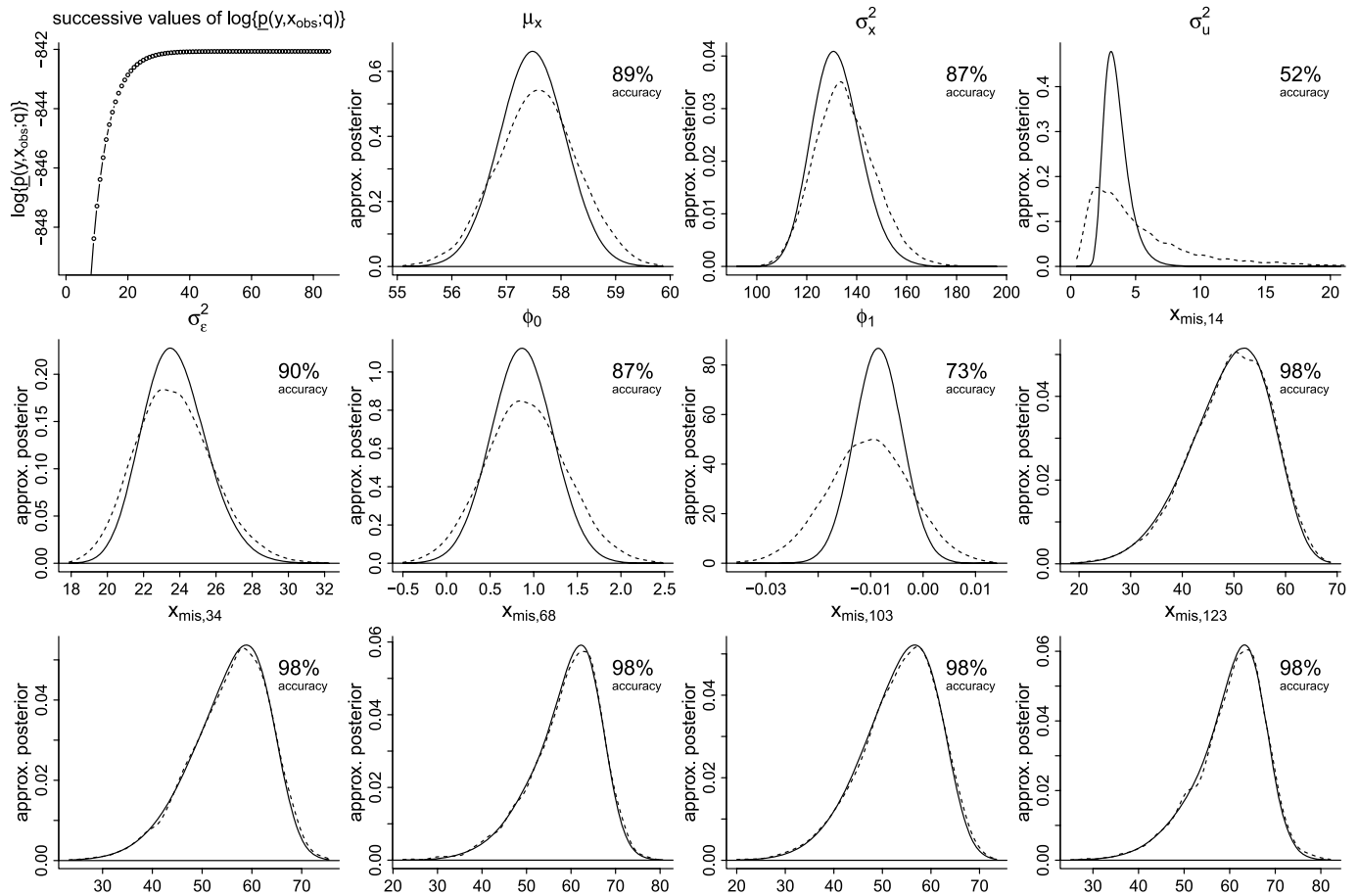


Figure 5. Variational Bayes approximate posteriors for the regression model parameters and four missing x_i 's for nonparametric regression applied to the ozone data with predictors MNAR. The MCMC posteriors are based on samples of size 10,000 and kernel density estimation. The accuracy values correspond to the definition given at (9). Summary of nonparametric regression for ozone data with with predictor MNAR.

APPENDIX: NOTATION

If \mathcal{P} is a logical condition then $I(\mathcal{P}) = 1$ if \mathcal{P} is true and $I(\mathcal{P}) = 0$ if \mathcal{P} is false. We use Φ to denote the standard normal distribution function.

Column vectors with entries consisting of subscripted variables are denoted by a bold-faced version of the letter for that variable. Round

brackets will be used to denote the entries of column vectors. For example $\mathbf{x} = (x_1, \dots, x_n)$ denotes a $n \times 1$ vector with entries x_1, \dots, x_n . The element-wise product of two matrices \mathbf{A} and \mathbf{B} is denoted by $\mathbf{A} \odot \mathbf{B}$. We use $\mathbf{1}_d$ to denote the $d \times 1$ column vector with all entries equal to 1. The norm of a column vector \mathbf{v} , defined to be $\sqrt{\mathbf{v}^T \mathbf{v}}$, is denoted by $\|\mathbf{v}\|$. Scalar functions applied to vectors are evaluated

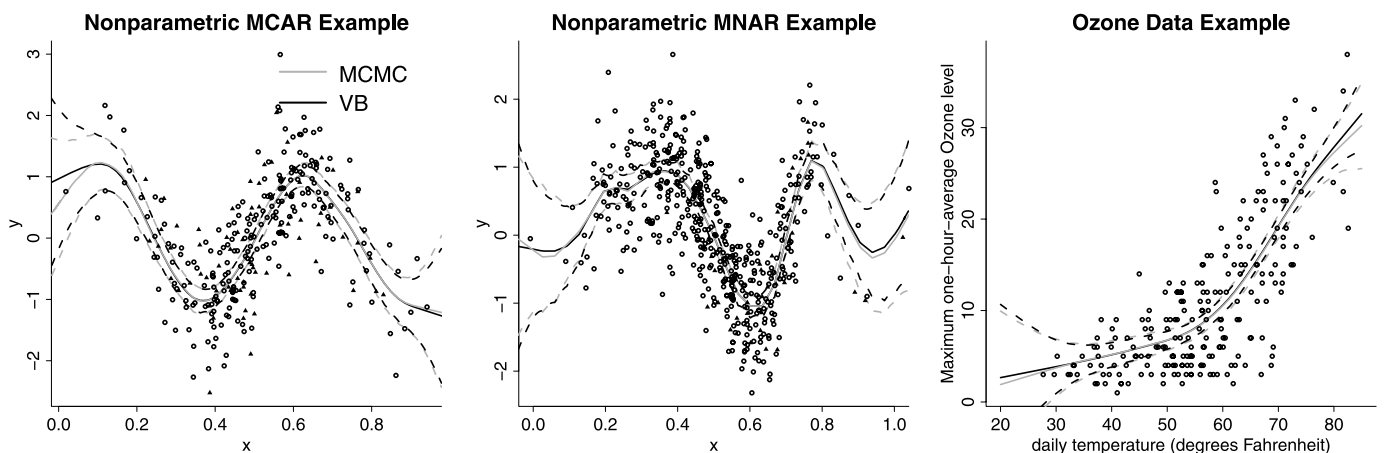


Figure 6. Posterior mean functions and corresponding pointwise 95% credible sets for all three nonparametric regression examples. The grey curves correspond to MCMC-based inference, whilst the black curves correspond to variational Bayesian inference.

element-wise. For example,

$$\Phi(a_1, a_2, a_3) \equiv (\Phi(a_1), \Phi(a_2), \Phi(a_3)).$$

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 $p(\mathbf{u}|\mathbf{v})$. The covariance matrix of \mathbf{u} is denoted by $\text{Cov}(\mathbf{u})$. The notation $\mathbf{x} \sim N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ means that the random vector \mathbf{x} has a Multivariate Normal distribution with mean $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$. A random variable x has an Inverse Gamma distribution with parameters $A, B > 0$, denoted by $x \sim \text{IG}(A, B)$, if its density function is $p(x) = B^A \Gamma(A)^{-1} x^{-A-1} e^{-B/x}$, $x > 0$. If y_i has distribution D_i for each $1 \leq i \leq n$, and the y_i are independent, then we write $y_i \stackrel{\text{ind.}}{\sim} D_i$.

SUPPLEMENTARY MATERIALS

Additional Results and Derivations: The supplemental material is a single document (FaesOrmerodWandSupplement.pdf, PDF file) with the following four components:

Supplement A: Details of variational Bayes for the MCAR and MAR parametric regression models.

Supplement B: Accuracy assessment summaries for non-parametric regression simulations.

Supplement C: Derivation of Algorithm 1.

Supplement D: Derivation of (12) and expression for $\Lambda_{\text{mis},i}$.

[Received May 2010. Revised February 2011.]

REFERENCES

- Albert, J. H., and Chib, S. (1993), "Bayesian Analysis of Binary and Polychotomous Response Data," *Journal of the American Statistical Association*, 88, 669–679. [962]
- Bishop, C. M. (2006), *Pattern Recognition and Machine Learning*, New York: Springer. [959,960,962]
- Box, G. P., and Tiao, G. C. (1973), *Bayesian Inference in Statistical Analysis*, Reading, MA: Addison-Wesley. [959]
- Breiman, L., and Friedman, J. H. (1985), "Estimating Optimal Transformations for Multiple Regression and Correlation," *Journal of the American Statistical Association*, 80, 580–598. [969]
- Crainiceanu, C., Ruppert, D., and Wand, M. P. (2005), "Bayesian Analysis for Penalized Spline Regression Using WinBUGS," *Journal of Statistical Software*, 14 (14), 1–24. [959]
- Daniels, M. J., and Hogan, J. W. (2008), *Missing Data in Longitudinal Studies: Strategies for Bayesian Modeling and Sensitivity Analysis*, Boca Raton, FL: Chapman & Hall/CRC Press. [959]
- Denison, D., Holmes, C., Mallick, B., and Smith, A. (2002), *Bayesian Methods for Nonlinear Classification and Regression*, Chichester, U.K.: Wiley. [959]
- Devroye, L., and Györfi, L. (1985), *Density Estimation: The L_1 View*, New York: Wiley. [964]
- Flandin, G., and Penny, W. D. (2007), "Bayesian fMRI Data Analysis With Sparse Spatial Basis Function Priors," *NeuroImage*, 34, 1108–1125. [959]
- Gelman, A. (2006), "Prior Distributions for Variance Parameters in Hierarchical Models," *Bayesian Analysis*, 1, 515–533. [961]
- Gelman, A., Carlin, J. B., Stern, H. S., and Rubin, D. B. (2004), *Bayesian Data Analysis*, Boca Raton, FL: Chapman & Hall. [959,961]
- Girolami, M., and Rogers, S. (2006), "Variational Bayesian Multinomial Probit Regression," *Neural Computation*, 18, 1790–1817. [962]
- Gurrin, L. C., Scurrah, K. J., and Hazelton, M. L. (2005), "Tutorial in Biostatistics: Spline Smoothing With Linear Mixed Models," *Statistics in Medicine*, 24, 3361–3381. [959]
- Hall, P. (1987), "On Kullback–Leibler Loss and Density Estimation," *The Annals of Statistics*, 15, 1491–1519. [964]
- Holmes, C. C., and Held, L. (2006), "Bayesian Auxiliary Variable Models for Binary and Multinomial Regression," *Bayesian Analysis*, 1, 145–168. [964]
- Jordan, M. I. (2004), "Graphical Models," *Statistical Science*, 19, 140–155. [959]
- Leisch, F., and Dimitriadou, E. (2009), "mlbench 1.1-6: Machine Learning Benchmark Problems," R package. Available at <http://cran.r-project.org>. [969]
- Ligges, U., Thomas, A., Spiegelhalter, D., Best, N., Lunn, D., Rice, K., and Sturtz, S. (2010), "BRugs 0.5: OpenBUGS and Its R/S-PLUS Interface BRugs," R package. Available at <http://www.stats.ox.ac.uk/pub/RWin/bin/windows/contrib/2.14>. [959,964,965]
- Little, R. J., and Rubin, D. B. (2004), *Statistical Analysis With Missing Data* (2nd ed.), New York: Wiley. [959,961]
- Luenberger, D. G., and Ye, Y. (2008), *Linear and Nonlinear Programming* (3rd ed.), New York: Springer. [960]
- Lunn, D. J., Thomas, A., Best, N., and Spiegelhalter, D. (2000), "WinBUGS—A Bayesian Modelling Framework: Concepts, Structure, and Extensibility," *Statistics and Computing*, 10, 325–337. [965]
- McGrory, C. A., and Titterton, D. M. (2007), "Variational Approximations in Bayesian Model Selection for Finite Mixture Distributions," *Computational Statistics and Data Analysis*, 51, 5352–5367. [959]
- Neal, R. (1998), "Suppressing Random Walks in Markov Chain Monte Carlo Using Ordered Over-Relaxation," in *Learning in Graphical Models*, ed. M. I. Jordan, Dordrecht: Kluwer Academic, pp. 205–230. [969]
- Ormerod, J. T., and Wand, M. P. (2010), "Explaining Variational Approximations," *The American Statistician*, 64, 140–153. [959,960]
- Parisi, G. (1988), *Statistical Field Theory*, Redwood City, CA: Addison-Wesley. [960]
- Pearl, J. (1988), *Probabilistic Reasoning in Intelligent Systems*, San Mateo, CA: Morgan Kaufmann. [961,962]
- R Development Core Team (2010), *R: A Language and Environment for Statistical Computing*, Vienna, Austria: R Foundation for Statistical Computing. Available at <http://www.R-project.org>. [959,965]
- Robert, C. P., and Casella, G. (2004), *Monte Carlo Statistical Methods* (2nd ed.), New York: Springer-Verlag. [960]
- Ruppert, D., Wand, M. P., and Carroll, R. J. (2003), *Semiparametric Regression*, New York: Cambridge University Press. [959]
- Teschendorff, A. E., Wang, Y., Barbosa-Morais, N. L., Brenton, J. D., and Caldas, C. (2005), "A Variational Bayesian Mixture Modelling Framework for Cluster Analysis of Gene-Expression Data," *Bioinformatics*, 21, 3025–3033. [959]
- Wahba, G. (1978), "Improper Priors, Spline Smoothing and the Problem of Guarding Against Model Errors in Regression," *Journal of the Royal Statistical Society, Ser. B*, 40, 364–372. [959]
- Wand, M. P. (2009), "Semiparametric Regression and Graphical Models," *Australian and New Zealand Journal of Statistics*, 51, 9–41. [959,969]
- Wand, M. P., and Ormerod, J. T. (2008), "On O'Sullivan Penalised Splines and Semiparametric Regression," *Australian and New Zealand Journal of Statistics*, 50, 179–198. [966]
- Wand, M. P., and Ripley, B. D. (2009), "KernSmooth 2.23: Functions for Kernel Smoothing Corresponding to the Book: Wand, M. P., and Jones, M. C. (1995), *Kernel Smoothing*," R package. Available at <http://cran.r-project.org>. [964]
- Wasserman, L. (2004), *All of Statistics*, New York: Springer. [960]