Published in final edited form as:
*Electron J Stat*. 2011 January 1; 5: 572–602.

# Functional regression via variational Bayes

**Jeff Goldsmith**[*,†],
Johns Hopkins Bloomberg School of Public Health Department of Biostatistics 615 North Wolfe Street Baltimore, Maryland 21205, USA

**Matt P. Wand**[‡], and
School of Mathematical Sciences University of Technology, Sydney P.O. Box 123 Broadway, 2007, Australia

**Ciprian Crainiceanu**[*]
Johns Hopkins Bloomberg School of Public Health Department of Biostatistics 615 North Wolfe Street Baltimore, Maryland 21205, USA

Jeff Goldsmith: jgoldsmi@jhsph.edu; Matt P. Wand: Matt.Wand@uts.edu.au; Ciprian Crainiceanu: ccrainic@jhsph.edu

## Abstract

We introduce variational Bayes methods for fast approximate inference in functional regression analysis. Both the standard cross-sectional and the increasingly common longitudinal settings are treated. The methodology allows Bayesian functional regression analyses to be conducted without the computational overhead of Monte Carlo methods. Confidence intervals of the model parameters are obtained both using the approximate variational approach and nonparametric resampling of clusters. The latter approach is possible because our variational Bayes functional regression approach is computationally efficient. A simulation study indicates that variational Bayes is highly accurate in estimating the parameters of interest and in approximating the Markov chain Monte Carlo-sampled joint posterior distribution of the model parameters. The methods apply generally, but are motivated by a longitudinal neuroimaging study of multiple sclerosis patients. Code used in simulations is made available as a web-supplement.

### Keywords and phrases

Approximate Bayesian inference; Markov chain Monte Carlo; penalized splines

## 1. Introduction

Due to ever-expanding methods for the acquisition and storage of information, functional data is often encountered in scientific applications. A common problem in the field of functional data analysis is determining the relationship between a scalar outcome $Y$ and a densely observed functional predictor $X(t)$ [18, 22, 9]. Increasingly, this problem is longitudinal in that both the functional predictors and scalar outcomes are observed at several visits for each subject. Bayesian approaches to cross-sectional and longitudinal functional regression possess a number of advantages, including the ability to jointly model

[*]The work of Goldsmith and Crainiceanu was supported by Award Number R01NS060910 from the National Institute Of Neurological Disorders And Stroke. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institute Of Neurological DIsorders And Stroke or the National Institutes of Health.
[†]The first author would also like to acknowledge partial support from Training Grant 2T32ES012871 from the US, National Institutes of Health, National Institute of Environmental Health Sciences.
[‡]The work of Wand was partially supported by Australian Research Council Discovery Project DP110100061.

the observed functions and scalar outcomes and easily constructed credible intervals [5, 7]. However, these approaches require computationally expensive Markov chain Monte Carlo (MCMC) simulations of joint posterior distributions. The goal of this paper is to introduce a fast and scalable alternative to accommodate new types of data sets.

Variational approximations, now regularly used in computer science, are a collection of techniques for deriving approximate solutions to inference problems [10, 11, 25]. They have a growing visibility in the statistics literature [16, 19, 24]. In the Bayesian context, these methods are useful in approximating intractable posterior density functions. While this approximation sacrifices some of MCMC's accuracy, it provides large gains in terms of computational feasibility, especially in large-data settings.

In this article, we derive an iterative algorithm for approximate Bayesian inference in functional regression. Using this algorithm, inference on model parameters can be obtained several orders of magnitude faster than MCMC sampling methods. Importantly, the construction of credible intervals for the functional coefficient is straightforward. Moreover, this procedure retains the ability to jointly model the predictor process and the scalar outcome. The computational advantage conveyed by the variational methods also allows resampling techniques, such as the nonparametric bootstrap of subjects, to be used. Unlike MCMC sampling, the variational approach cannot be made arbitrarily accurate. However, simulations indicate that the quality of the approximation is high in our setting. Our variational method is not designed to replace MCMC sampling, but it is a useful *additional* inferential tool in that it provides near-instant and highly accurate approximate posterior distributions. This will become increasingly relevant as functional datasets become larger and more complex.

In particular, we develop variational Bayes methods for two functional regression models: the classic cross-sectional case, in which a single scalar outcome and functional predictor are observed for each subject; and the more recent longitudinal case, in which scalar outcomes and functional predictors are observed repeatedly for each subject. This methodology is based on a penalized approach to functional regression that is flexible and widely applicable [6]. Although variational techniques typically incur initial algebraic and implementation costs, the present article alleviates these considerations.

We apply the methods developed to a longitudinal neuroimaging study, in which multiple sclerosis patients undergo both tests of cognitive ability and a diffusion tensor imaging scan at each of several visits. From the diffusion tensor imaging scans, we construct functional predictors that provide detailed quantitative information about major white matter fiber bundles (see Figure 1). Because multiple sclerosis results in the degradation of cerebral white matter, researchers hope to use the functional predictors and cognitive disability measures to understand the progression of the disease. This study was previously analyzed in [7].

In Section 2 we introduce variational Bayes and a penalized approach to functional regression. Section 3 combines these ideas and develops a scalable iterative algorithm for approximate Bayesian inference in functional regression. The results of a simulation study are described in Section 4 and a real-data analysis is performed in Section 5. We conclude the main text with a discussion in Section 6. Appendices A and B contain algebraic derivations and expressions used in the construction of the iterative algorithm. All code used in the simulation study is available as a web-supplement to this article.

## 2. Background

In the following subsections we introduce variational approximations for Bayesian inference and an approach to functional regression which uses penalized B-splines to estimate the coefficient function.

### 2.1. Variational Bayes

Here we give an overview of variational Bayes; for a more complete treatment see [19] and [3], Chapter 10.

Bayesian inference is based on the posterior density function

$$p(\theta|\boldsymbol{y})=\frac{p(\boldsymbol{y},\theta)}{p(\boldsymbol{y})}$$

where $\boldsymbol{\theta} \in \Theta$ is the parameter vector, $\boldsymbol{y}$ is the observed data, $p(\boldsymbol{y})$ is the marginal likelihood of the observed data, and $p(\boldsymbol{y}, \boldsymbol{\theta})$ is the joint likelihood of the data and model parameters. The goal of the density transform approach is to approximate the posterior density $p(\boldsymbol{\theta}|\boldsymbol{y})$ by a function $q(\boldsymbol{\theta})$ for which the $q$-specific lower bound on the marginal likelihood (defined below) is more tractable than the marginal likelihood itself. The first step is to restrict $q$ to a more manageable class of densities and choose the element of that class with minimum Kullback-Leibler distance from $p(\boldsymbol{\theta}|\boldsymbol{y})$.

More concretely, let $q$ be an arbitrary density function over $\Theta$. Then

$$\log p(\boldsymbol{y}) \geq \int q(\theta)\log\left\{\frac{p(\boldsymbol{y},\theta)}{q(\theta)}\right\}d\theta. \tag{2.1}$$

with equality if and only if $q(\boldsymbol{\theta}) = p(\boldsymbol{\theta}|\boldsymbol{y})$ almost everywhere [12]. It follows that $p(\boldsymbol{y}) \geq \exp\int q(\theta)\log\left\{\frac{p(\boldsymbol{y},\theta)}{q(\theta)}\right\}d\theta$; we define the $q$-specific lower bound on the marginal likelihood as

$$\underline{p}(\boldsymbol{y};q) \equiv \exp\int q(\theta)\log\left\{\frac{p(\boldsymbol{y},\theta)}{q(\theta)}\right\}d\theta. \tag{2.2}$$

It can be shown that minimizing the Kullback-Leibler distance between $q(\boldsymbol{\theta})$ and $p(\boldsymbol{\theta}|\boldsymbol{y})$ is equivalent to maximizing $\underline{p}(\boldsymbol{y}; q)$. Stated generally, the following result holds.

**Result 2.1—**Let $\boldsymbol{u}$ and $\boldsymbol{v}$ be continuous random vectors with joint density $p(\boldsymbol{u}, \boldsymbol{v})$. Then

$$sup_q\left\{\int q(\boldsymbol{u})\log\left[\frac{p(\boldsymbol{u},\boldsymbol{v})}{q(\boldsymbol{u})}d\boldsymbol{u}\right]\right\}$$

is achieved by $q^*(\boldsymbol{u}) = p(\boldsymbol{u}|\boldsymbol{v})$.

*The work of Goldsmith and Crainiceanu was supported by Award Number R01NS060910 from the National Institute Of Neurological Disorders And Stroke. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institute Of Neurological DIsorders And Stroke or the National Institutes of Health.

Next, we restrict $q$ to a class of functions for which $\underline{p}(\boldsymbol{y}; q)$ is more tractable than $p(\boldsymbol{y})$. While several restrictions are possible, here we focus on the product density transform: we assume that for some partition $\{\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_L\}$ of $\boldsymbol{\theta}$ it is possible to write $q(\theta)=\prod_{l=1}^{L} q_l(\theta_l)$. In the functional regression setting, the posterior dependence of some subsets of the model parameters is weak and the assumption that $q$ factorizes provides accurate approximate inference. In other settings where the posterior dependence of parameters is stronger, this assumption may lead to poor approximations and inference due to the failure to account for correlations between model parameters. There are three simple strategies to gain insight into what sets of parameters are a-posteriori weakly correlated: 1) theoretical work on asymptotic posterior correlations; 2) Bayesian inference on smaller or simpler data sets; and 3) prior experience. If posterior correlation is potentially problematic, a more flexible component density $q_l$ that allows for this correlation could be used; however, this must be balanced against the simplification desired in the approximating class of functions. While none of the approaches above is infallible, when combined with powerful variational approximations they can provide a valuable alternative to Bayesian inference. The methods provided in this paper are intended as a reasonable and tractable *complement of* and not *replacement for* Bayesian computations.

Combining the assumption that $q$ factorizes over a partition of $\boldsymbol{\theta}$ with Result 2.1, we can derive explicit solutions for each factor $q_l(\boldsymbol{\theta}_l)$, $1 \le l \le L$, in terms of the remaining factors. Solving for each factor in terms of the others leads to an iterative algorithm for obtaining a solution for $q$. The explicit solution for each $q_l(\boldsymbol{\theta}_l)$ is derived as follows. Assuming that $q$ is subject to the factorization restriction, it follows that

$$\begin{aligned}\log \underline{p}(\boldsymbol{y};q) &= \int \left(\prod_{l=1}^{L} q_i(\theta_i)\right)\left(\log p(\boldsymbol{y},\theta) - \sum_{l=1}^{L} \log q_i(\theta_i)\right) d\theta_1 \ldots d\theta_L \\ &= \int q_1(\theta_1)\left(\int \log p(\boldsymbol{y},\theta) q_2(\theta_2)\ldots q_L(\theta_L) d\theta_2 \ldots d\theta_L\right) d\theta_1 - \int q_1(\theta_1)\log q_1(\theta_1) d\theta_1 + \text{terms not involving } q_1\end{aligned}$$

Define the joint density function $\tilde{p}(\boldsymbol{y}, \boldsymbol{\theta}_1)$ to be

$$\tilde{p}(\boldsymbol{y},\theta_1) \equiv \frac{\exp \int \log p(\boldsymbol{y},\theta) q_2(\theta_2)\cdots q_M(\theta_M) d\theta_2 \cdots d\theta_M}{\int\int \left\{\exp \int \log p(\boldsymbol{y},\theta) q_2(\theta_2)\cdots q_M(\theta_M) d\theta_2 \cdots d\theta_M\right\} d\theta_1 d\boldsymbol{y}}$$

so that

$$\log \underline{p}(\boldsymbol{y};q) = \int q_1(\theta_1) \log\left[\frac{\tilde{p}(\boldsymbol{y},\theta_1)}{q_1(\theta_1)}\right] + \text{terms not involving } q_1.$$

Then, using Result 2.1, the optimal $q_1$ is

$$\begin{aligned} q_1^*(\theta_1) = \tilde{p}(\theta_1|\boldsymbol{y}) &= \frac{\tilde{p}(\boldsymbol{y},\theta_1)}{\int \tilde{p}(\boldsymbol{y},\theta_1) d\theta_1} \\ &\propto \exp\left[\int \log p(\boldsymbol{y},\theta) q_2(\theta_2)\ldots q_L(\theta_L) d\theta_2 \ldots d\theta_L\right] \\ &= \exp[E_{\theta_{-1}} \log p(\boldsymbol{y},\theta)] \end{aligned}$$

where $E_{\theta-1} \log p(\boldsymbol{y}, \boldsymbol{\theta})$ is the expectation with respect to $q_2(\boldsymbol{\theta}_2) \ldots q_L(\boldsymbol{\theta}_L)$. The same argument for $l$ in 1, …, $L$ yields optimal densities satisfying

$$q_l^*(\theta_l) \propto \exp[E_{\theta_{-l}} \log p(\boldsymbol{y}, \theta)] \propto \exp[E_{\theta_{-l}} \log p(\theta_l | \text{rest})] \tag{2.3}$$

where rest ≡ $\{\boldsymbol{y}, \boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_{l-1}, \boldsymbol{\theta}_{l+1}, \ldots, \boldsymbol{\theta}_L\}$ is the collection of all remaining parameters and the observed data. Solving for each factor in terms of the others leads to an iterative algorithm for obtaining a solution for $q$. We update each factor in turn until the change in $p(\boldsymbol{y}; q)$ is negligible.

### 2.2. Penalized functional regression

Next we introduce penalized approaches to cross-sectional and longitudinal functional regression [5, 6, 7].

In the cross-sectional case, we observe data of the form $[Y_i, X_i(t), \boldsymbol{z}_i]$ for subjects $1 \le i \le I$, where $Y_i$ is a continuous outcome, $X_i(t)$ is a functional covariate, and $\boldsymbol{z}_i$ is a $1 \times p$ vector of non-functional covariates. The linear functional regression model is given by [4, 21]

$$\begin{gathered} Y_i = \boldsymbol{z}_i \beta + \int_0^1 X_i(t) \gamma(t) dt + \varepsilon_i^Y \\ \varepsilon_i^Y \sim \mathrm{N}(0, \sigma_\gamma^2). \end{gathered} \tag{2.4}$$

We call the parameter $\gamma(t)$ the coefficient function. In practice, the predictor functions $X_i(t)$ are observed over a discrete grid, and often with error. That is, we observe $\{W_i(t_{ij}): t_{ij} \in [0, 1]\}$ for $1 \le i \le I$ and $1 \le j \le J_i$, where $W_i(t_{ij}) = X_i(t_{ij}) + \varepsilon_i^X(t_{ij})$ and $\varepsilon_i^X(t_{ij}) \sim \mathrm{N}(0, \sigma_x^2)$. The sampling scheme on which the functional predictors are observed may take a variety of forms: points may be equally or unequally spaced, sparse or dense at the subject level, identical or different across subjects. For simplicity, we will assume that all subjects are observed over the same grid $\{t_1, \ldots, t_N\}$ and are observed at an equal number of visits $J$. Extensions to different grids and different number of visits is straightforward, but with considerable increase in notational complexity.

To estimate the parameters in model (2.4), we use the following two-stage procedure. First, the predictor functions $X_i(t)$ are expressed using a principal components (PC) decomposition. Second, the coefficient function $\gamma(t)$ is estimated using penalized B-splines. Smoothness of $\hat{\gamma}(t)$ is explicitly induced via a mixed effects model. Specifically, let $\hat{\Sigma}^X(s, t)$ be an estimator of the covariance operator Cov $(X_i(s), X_i(t))$ based on the available functional observations. Further, let $\sum_{k=1}^{\infty} \lambda_k \psi_k(s) \psi_k(t)$ be the spectral decomposition of $\hat{\Sigma}^X(s, t)$, where $\lambda_1 \ge \lambda_2 \ge \cdots$ are the non-increasing eigenvalues and $\boldsymbol{\psi}(t) = \{\psi_k(t): k \in \boldsymbol{Z}^+\}$ are the corresponding orthonormal eigenfunctions. An approximation for $X_i(t)$, based on a truncated Karhunen-Lóeve expansion, is given by $X_i(t) = \mu(t) + \sum_{k=1}^{K_x} c_{ik} \psi_k(t)$, where $K_x$ is the truncation lag, the PC loadings $c_{ik} = \int_0^1 \{X_i(t) - \mu(t)\} \psi_k(t) dt$ are uncorrelated random variables with variance $\lambda_k$, and $\mu(t)$ is the mean function over all subjects and visits.

Next, we use a large cubic B-spline basis to smoothly estimate the coefficient function $\gamma(t)$ using a mixed effects model. Let $\boldsymbol{\varphi}(t) = \{\varphi_1(t), \ldots, \varphi_{K_g}(t)\}$ be a cubic B-spline basis of dimension $K_g$. Then the integral in model (2.4) can be written as

$\int_0^1 X_i(t)\gamma(t)dt = a + \int_0^1 \boldsymbol{c}_i' \boldsymbol{\psi}^T(t)\boldsymbol{\varphi}(t)\boldsymbol{g}dt = a + \boldsymbol{c}_i' \boldsymbol{M}\boldsymbol{g}$ where $a = \int_0^1 \mu(t)\gamma(t)dt$, $\boldsymbol{c}_i' = [c_{i1}, \ldots, c_{iK_x}]$ is the row vector of subject $i$'s PC loadings, and $\boldsymbol{M}$ is a $K_x \times K_g$ matrix with $(k, l)^{th}$ entry $\int_0^1 \psi_k(t)\varphi_l(t)dt$. Smoothness of $\hat{\gamma}(t)$ is enforced by assuming a modified first order random walk prior on the vector $\boldsymbol{g}$ [13]. That is, we assume $g_l \sim \mathrm{N}(g_{l-1}, \sigma_{\boldsymbol{g}}^2)$ for $2 \le l \le K_g$ and let $g_1 \sim \mathrm{N}(0, 0.01\sigma_{\boldsymbol{g}}^2)$. These are standard assumptions in Bayesian P-splines modeling [23, 13]. Taken together, we jointly model the scalar outcome $Y_i$ and the functional exposure $X_i(t)$ using the following model:

$$\begin{aligned} Y_i &\sim \mathrm{N}(\boldsymbol{z}_i\beta + \boldsymbol{c}_i'\boldsymbol{M}\boldsymbol{g}, \sigma_Y^2);\ \sigma_Y^2 \sim \mathrm{IG}(A_Y, B_Y) \\ W_i(t) &\sim \mathrm{N}(\mu(t) + \boldsymbol{c}_i'\boldsymbol{\psi}(t)^T, \sigma_{\boldsymbol{X}}^2 \boldsymbol{I});\ \sigma_{\boldsymbol{X}}^2 \sim \mathrm{IG}(A_{\boldsymbol{X}}, B_{\boldsymbol{X}}) \\ \boldsymbol{c}_i' &\sim \mathrm{N}(0, \Lambda);\ \lambda_k \sim \mathrm{IG}(A_\lambda, B_\lambda) \text{ for } 1 \le k \le K_x \\ \boldsymbol{g} &\sim \mathrm{N}(0, \sigma_{\boldsymbol{g}}^2 \boldsymbol{D});\ \sigma_{\boldsymbol{g}}^2 \sim \mathrm{IG}(A_{\boldsymbol{g}}, B_{\boldsymbol{g}}) \\ \beta &\sim \mathrm{N}\left(0, \sigma_\beta^2 \boldsymbol{I}\right) \end{aligned} \tag{2.5}$$

where $\boldsymbol{\beta}$ are treated as fixed parameters with diffuse priors, $\boldsymbol{D}$ is the covariance matrix induced by the first order random walk prior, and $\Lambda = \mathrm{diag}[\lambda_1, \ldots, \lambda_{K_X}]$. Inference for the functional regression model is based on the posterior density

$$p(\boldsymbol{g}, \boldsymbol{C}, \beta, \sigma_Y^2, \sigma_{\boldsymbol{X}}^2, \sigma_{\boldsymbol{g}}^2, \lambda_1, \ldots, \lambda_{K_x} | \boldsymbol{Y}, \boldsymbol{W}) \tag{2.6}$$

where $\boldsymbol{C}$ is the matrix of PC loadings constructed by row-stacking the $\boldsymbol{c}_i$, $\boldsymbol{Y} = \{Y_i\}_{i=1}^I$, and $\boldsymbol{W}$ is the matrix of observed predictor functions constructed by row-stacking the $W_i(t)$. Because the functional predictors $X_i(t)$ are observed with error, this model extends Bayesian inference for measurement error regression problems to the functional setting. A directed acyclic graph depicting model (2.5) is presented in Figure 2.

In the longitudinal case, we observe data of the form $[Y_{ij}, X_{ij}(t), \boldsymbol{z}_{ij}]$ for $1 \le i \le I$ and $1 \le j \le J_i$. Thus we observe a distinct functional predictor and scalar outcome for each subject over several visits, and again note that in place of the true functional predictors $X_{ij}(t)$ we often observe a measured-with-error function $W_{ij}(t)$. The longitudinal functional regression model is given by [7]

$$\begin{gathered} Y_{ij} = \boldsymbol{Z}_i\boldsymbol{b} + \boldsymbol{z}_{ij}\beta + \int_0^1 X_{ij}(t)\gamma(t)dt + \varepsilon_{ij}^Y \\ \varepsilon_i^Y \sim \mathrm{N}(0, \sigma_Y^2); \end{gathered} \tag{2.7}$$

this differs from model (2.4) in the use of subject-specific random effects $\boldsymbol{Z}_i$ to account for correlation in the repeated outcomes at the subject level. Moreover, longitudinal data sets tend to be much larger than cross-sectional data sets because of the number of visits.

Given the advent of multiple observational studies collecting dense functional data at multiple visits, the importance of longitudinal functional regression cannot be understated. Unfortunately, with the exception of the work in [7], no other approach can currently deal with the combination of subject-specific random effects and functional predictors necessary to capture the structure of the data. While a wide array of functional regression methods exist, we contend that the specific modeling choices described here made the extension not

only possible, but seamless. Estimation of the parameters in the longitudinal setting extends naturally from the procedure outlined for the cross-sectional setting. Again, we express the functional predictors using a PC basis and use a penalized B-spline expansion for the coefficient function. The joint model for the outcome, $Y_{ij}$, and exposure, $X_{ij}(t)$, becomes

$$\begin{aligned} Y_{ij} &\sim \mathrm{N}(\boldsymbol{Z}_i\boldsymbol{b}+\boldsymbol{z}_{ij}\beta+\boldsymbol{c}_{ij}'\boldsymbol{M}\boldsymbol{g}, \sigma_Y^2);\ \sigma_Y^2 \sim \mathrm{IG}(A_Y, B_Y) \\ W_{ij}(t) &\sim \mathrm{N}(\mu(t)+\boldsymbol{c}_{ij}'\boldsymbol{\psi}(t)^T, \sigma_X^2\boldsymbol{I});\ \sigma_X^2 \sim \mathrm{IG}(A_X, B_X) \\ \boldsymbol{c}_{ij}' &\sim \mathrm{N}(0, \Lambda);\ \lambda_k \sim \mathrm{IG}(A_\lambda, B_\lambda) \text{ for } 1 \le k \le K_x \\ \boldsymbol{g} &\sim \mathrm{N}(0, \sigma_{\boldsymbol{g}}^2\boldsymbol{D});\ \sigma_{\boldsymbol{g}}^2 \sim \mathrm{IG}(A_{\boldsymbol{g}}, B_{\boldsymbol{g}}) \\ \boldsymbol{b} &\sim \mathrm{N}(0, \sigma_{\boldsymbol{b}}^2\boldsymbol{I});\ \sigma_{\boldsymbol{b}}^2 \sim \mathrm{IG}(A_{\boldsymbol{b}}, B_{\boldsymbol{b}}) \\ \beta &\sim \mathrm{N}\left(0, \sigma_\beta^2\boldsymbol{I}\right) \end{aligned} \tag{2.8}$$

Again, inference is based on the posterior density

$$p(\boldsymbol{g}, \boldsymbol{C}, \beta, \boldsymbol{b}, \sigma_Y^2, \sigma_X^2, \sigma_{\boldsymbol{g}}^2, \sigma_{\boldsymbol{b}}^2, \lambda_1, \ldots, \lambda_{K_x} | \boldsymbol{Y}, \boldsymbol{W}) \tag{2.9}$$

A directed acyclic graph of the longitudinal functional regression model appears in Figure 2.

## 3. Variational approximations for penalized functional regression

We now combine the ideas introduced above to develop a scalable iterative algorithm for approximate Bayesian inference in functional regression. We will focus on the longitudinal functional regression model (2.8); the cross-sectional case can be obtained as a special case by omitting the vector of subject-specific random effects $\boldsymbol{b}$. We pause briefly to introduce the following useful notation: for a scalar random variables $\theta$, let

$$\begin{aligned} \mu_{q(\theta)} &\equiv \mathrm{E}_q[\theta] = \int \theta_0 q(\theta_0) d\theta_0 \\ \sigma_{q(\theta)} &\equiv \mathrm{Var}_q[\theta] = \int (\theta_0 - \mathrm{E}[\theta])^2 q(\theta_0) d\theta_0 \end{aligned}$$

be the mean and variance with respect to the $q$ distribution. For a vector parameter $\boldsymbol{\theta}$, we use the analogously defined $\boldsymbol{\mu}_{q(\boldsymbol{\theta})}$ and $\boldsymbol{\Sigma}_{q(\boldsymbol{\theta})}$.

As noted, inference in the longitudinal functional regression model is based on the posterior density (2.9). Using variational Bayes, we approximate this posterior density using

$$q(\boldsymbol{g}, \boldsymbol{C}, \beta, \boldsymbol{b}, \sigma_Y^2, \sigma_X^2, \sigma_{\boldsymbol{g}}^2, \sigma_{\boldsymbol{b}}^2, \lambda_1, \ldots, \lambda_{K_x}) = q(\boldsymbol{g}) q(\boldsymbol{C}) q(\beta) q(\boldsymbol{b}) q(\sigma_Y^2, \sigma_X^2, \sigma_{\boldsymbol{g}}^2, \sigma_{\boldsymbol{b}}^2, \lambda_1, \ldots, \lambda_{K_x})$$

(3.1)

and by solving for each factor $q(\cdot)$ in terms of the remaining factors. The additional factorization

$$q(\sigma_{\boldsymbol{Y}}^2, \sigma_{\boldsymbol{X}}^2, \sigma_{\boldsymbol{g}}^2, \sigma_{\boldsymbol{b}}^2, \lambda_1, \ldots, \lambda_{K_x}) = q(\sigma_{\boldsymbol{Y}}^2) q(\sigma_{\boldsymbol{X}}^2) q(\sigma_{\boldsymbol{g}}^2) q(\sigma_{\boldsymbol{b}}^2) \prod_{j=1}^{K_x} q(\lambda_j)$$

follows as a consequence of (2.3) and the structure of the current model as shown in Figure 2 [3, Sec. 10.2.5]. We take advantage of this *induced factorization* in deriving optimal densities for the variance components in the penalized functional regression model.

To provide an example of how optimal densities are constructed, we derive the optimal densities $q^*(\boldsymbol{g})$ and $q^*(\sigma_{\boldsymbol{g}}^2)$; derivations of these and for the other parameters are provided in Appendix A. Recall that $\boldsymbol{g} \sim \mathrm{N}(0, \sigma_{\boldsymbol{g}}^2 \boldsymbol{D})$ and $\sigma_{\boldsymbol{g}}^2 \sim \mathrm{IG}(A_{\boldsymbol{g}}, B_{\boldsymbol{g}})$. According to (2.3), the optimal densities are given by

$$q^*(\boldsymbol{g}) \propto \exp\{E_{-\boldsymbol{g}} \log p(\boldsymbol{g}|\text{rest})\} \quad \text{and} \quad q^*(\sigma_{\boldsymbol{g}}^2) \propto \exp\left\{E_{-\sigma_{\boldsymbol{g}}^2} \log p(\sigma_{\boldsymbol{g}}^2|\text{rest})\right\}$$

where rest includes both the observed data and all parameters not currently under consideration.

Using the full conditional distribution $p(\boldsymbol{g}|\text{rest}) \propto p(\boldsymbol{Y}|\beta, \boldsymbol{b}, \boldsymbol{C}, \boldsymbol{g}, \sigma_{\boldsymbol{Y}}^2) p(\boldsymbol{g}|\sigma_{\boldsymbol{g}}^2)$, the optimal density $q^*(\boldsymbol{g})$ is

$$\begin{aligned} q^*(\boldsymbol{g}) &\propto \exp\{E_{-\boldsymbol{g}} \log p(\boldsymbol{Y}|\beta, \boldsymbol{b}, \boldsymbol{C}, \boldsymbol{g}, \sigma_{\boldsymbol{Y}}^2) p(\boldsymbol{g}|\sigma_{\boldsymbol{g}}^2)\} \\ &\propto \exp\left[-\tfrac{1}{2}\mathrm{E}_{-\boldsymbol{g}}\left\{\boldsymbol{g}^T\left(\tfrac{1}{\sigma_{\boldsymbol{Y}}^2}\boldsymbol{M}^T\boldsymbol{C}^T\boldsymbol{C}\boldsymbol{M} + \tfrac{1}{\sigma_{\boldsymbol{g}}^2}\boldsymbol{D}^{-1}\right)\boldsymbol{g} - 2\left(\left(\boldsymbol{Y}^T - \beta^T\boldsymbol{z}^T - \boldsymbol{b}^T\boldsymbol{Z}^T\right)\left(\tfrac{1}{\sigma_{\boldsymbol{Y}}^2}\boldsymbol{C}\boldsymbol{M}\right)\right)\boldsymbol{g}\right\}\right] \\ &\propto \exp\left\{-\tfrac{1}{2}\left(\boldsymbol{g} - \mu_{q(\boldsymbol{g})}\right)^T \Sigma_{q(\boldsymbol{g})}^{-1}\left(\boldsymbol{g} - \mu_{q(\boldsymbol{g})}\right)\right\} \end{aligned}$$

where

$$\begin{aligned} \Sigma_{q(\boldsymbol{g})} &= \left\{\mu_{q(1/\sigma_{\boldsymbol{Y}}^2)} \boldsymbol{M}^T(\mu_{q(\boldsymbol{C})}^T \mu_{q(\boldsymbol{C})} + n\Sigma_{q(\boldsymbol{C})})\boldsymbol{M} + \mu_{q(1/\sigma_{\boldsymbol{g}}^2)}\boldsymbol{D}^{-1}\right\}^{-1}, \\ \mu_{q(\boldsymbol{g})} &= \Sigma_{q(\boldsymbol{g})}\left\{\mu_{q(1/\sigma_{\boldsymbol{Y}}^2)}(\boldsymbol{Y}^T - \mu_{q(\beta)}^T\boldsymbol{z}^T - \mu_{q(\boldsymbol{b})}^T\boldsymbol{Z}^T)(\mu_{q(\boldsymbol{C})}\boldsymbol{M})\right\}^T. \end{aligned} \tag{3.2}$$

Thus the optimal density $q^*(\boldsymbol{g})$ is $\mathrm{N}(\boldsymbol{\mu}_{q(\boldsymbol{g})}, \boldsymbol{\Sigma}_{q(\boldsymbol{g})})$. Similarly, the optimal density $q^*(\sigma_{\boldsymbol{g}}^2)$ is

$$\begin{aligned} q^*(\sigma_{\boldsymbol{g}}^2) &\propto \exp\{E_{-\boldsymbol{g}} \log p(\boldsymbol{g}|\sigma_{\boldsymbol{g}}^2) p(\sigma_{\boldsymbol{g}}^2)\} \\ &\propto (\sigma_{\boldsymbol{g}}^2)^{-A_{\boldsymbol{g}} - K_g/2 - 1} \exp\left\{-\tfrac{1}{\sigma_{\boldsymbol{g}}^2}\mathrm{E}_{-\sigma_{\boldsymbol{g}}^2}\left(B_{\boldsymbol{g}} + \tfrac{1}{2}\boldsymbol{g}^T\boldsymbol{D}^{-1}\boldsymbol{g}\right)\right\}. \end{aligned}$$

Thus $q^*(\sigma_{\boldsymbol{g}}^2)$ is $\mathrm{IG}(A_{\boldsymbol{g}} + K_g/2, B_{q(\sigma_{\boldsymbol{g}}^2)})$ where

$$B_{q(\sigma^2_{\boldsymbol{g}})} = B_{\boldsymbol{g}} + \tfrac{1}{2}\left(\mu^T_{q(\boldsymbol{g})}\boldsymbol{D}^{-1}\mu_{q(\boldsymbol{g})} + \mathrm{tr}(\boldsymbol{D}^{-1}\textstyle\sum_{q(\boldsymbol{g})})\right) \tag{3.3}$$

Note that, when $q(\sigma^2_{\boldsymbol{g}}) = q^*(\sigma^2_{\boldsymbol{g}})$, the term $\mu_{q(1/\sigma^2_{\boldsymbol{g}})}$ appearing in (3.2) is equal to $\frac{A_{\boldsymbol{g}} + K_g/2}{B_{q(\sigma^2_{\boldsymbol{g}})}}$.

Thus, the optimal densities $q^*(\boldsymbol{g})$ and $q^*(\sigma^2_{\boldsymbol{g}})$ belong to parametric families with the parameters explicitly determined by the distributions of the remaining model parameters and the observed data. Similar derivations for the parameters of the remaining optimal densities are derived in Appendix A. Taken together, these solutions lead to Algorithm 1 for approximate Bayesian inference in the functional linear regression setting.

Further, as shown in Appendix B, the $q$-specific lower bound on the marginal log-likelihood has the form

$$\begin{aligned}
\log \underline{p}(\boldsymbol{Y}, \boldsymbol{W}; q) = {} & \tfrac{1}{2}\log\left(\frac{|\textstyle\sum_{q(\beta)}|}{\sigma_\beta^{2p}}\right) - \frac{1}{2\sigma^2_\beta}\left\{\mu^T_{q(\beta)}\mu_{q(\beta)} + \mathrm{tr}(\textstyle\sum_{q(\beta)})\right\} \\
& + \tfrac{1}{2}\mathrm{E}_{q^*}[\log(|\textstyle\sum_{q(\boldsymbol{g})}|)] + \sum_{k=1}^{K_x}\frac{nJ}{2}\log((\textstyle\sum_{q(C)})_{kk}) \\
& + \tfrac{1}{2}\mathrm{E}_{q^*}[\log(|\textstyle\sum_{q(\boldsymbol{b})}|)] - \left(A_{\boldsymbol{g}} + \frac{K_g}{2}\right)\log\left(B_{q(\sigma^2_{\boldsymbol{g}})}\right) \\
& - \left(A_{\boldsymbol{b}} + \frac{n}{2}\right)\log\left(B_{q(\sigma^2_{\boldsymbol{b}})}\right) - \left(A_{\boldsymbol{Y}} + \frac{nJ}{2}\right)\log\left(B_{q(\sigma^2_{\boldsymbol{Y}})}\right) \\
& - \left(A_{\boldsymbol{X}} + \frac{nJN}{2}\right)\log\left(B_{q(\sigma^2_{\boldsymbol{X}})}\right) \\
& - \sum_{k=1}^{K_x}\left\{\left(A_\lambda + \frac{nJ}{2}\right)\log\left(B_{q(\lambda_k)}\right)\right\} + \text{const.}
\end{aligned}$$

where const. is an additive constant that remains unchanged in the iterations of Algorithm 1. All parameters denoted $A$ and $B$ and indexed by a subscript are hyperparameters of the inverse gamma prior distributions of the variance components. The quantity $\log \underline{p}(\boldsymbol{Y}, \boldsymbol{W}; q)$ is typically monitored for convergence in place of $\underline{p}(\boldsymbol{Y}, \boldsymbol{W}; q)$. Note that, because several substitutions are made to simplify the expression, this form for $\log \underline{p}(\boldsymbol{Y}, \boldsymbol{W}; q)$ is only valid at the end of each iteration of Algorithm 1, and only if the parameters are updated in the order given.

Finally, posterior credible intervals are readily obtained for all model parameters. However, variational approximations in effect fit a parametric distribution to a mode of the posterior density, which may have consequences when the posterior is multi-modal or more diffuse than the approximating parametric distribution; in such cases one could expect that credible intervals from variational Bayes and MCMC sampling may not agree. This was not a problem in our simulations, where the agreement between the approximate and MCMC-sampled posterior distribution is high, although in our application the variational credible intervals are slightly narrower than those from MCMC sampling.

## 4. Simulations

In this section we undertake simulation exercises with two goals. First, we evaluate our approach's overall ability to accurately estimate all coefficients in a functional regression model. Second, we compare the individual approximate posterior distributions $q^*(\boldsymbol{\theta}_l) \approx p(\boldsymbol{\theta}_l \mid \text{rest})$ to those given by Markovchain Monte Carlo (MCMC) sampling in order to examine

the quality of the variational approximation in the functional regression setting. We conduct separate simulations for the cross-sectional and longitudinal situations. The MCMC sampling was executed in WinBUGS and the variational Bayes approach was implemented in R.

### 4.1. Cross-sectional functional regression

We generate samples from the model

$$Y_i = \beta_1 + z_i\beta_2 + \int_0^1 X_i^S(t)\gamma(t)dt + \varepsilon_i^Y$$
$$\varepsilon_i^Y \sim \mathrm{N}(0, \sigma_Y^2). \tag{4.1}$$

Here we assume $I = \{100, 500\}$ subjects and generate $z_i \sim$ Unif $[-5, 5]$. We take $\sigma_Y^2 = 5, \beta_1 = \int_0^1 \mu(t)\gamma(t)dt = 3.47$, $\beta_2 = 3$, and $\gamma(t) = \cos(2\pi t)$.

To generate our simulated functional predictors $X_i^S(t)$, we use the functional predictors $X_i^A(t)$ from our scientific application in the following way. First, we compute a functional principal components decomposition of the $X_i^A(t)$ with eigenfunctions $\psi_1(t)$, $\psi_2(t)$, … and corresponding eigenvalues $\lambda_1$, $\lambda_2$, …. Recall that the application predictors can be approximated using $X_i^A(s) = \mu(t) + \sum_{k=1}^{K_x} c_{ik}\psi_k(t)$ where $\mu(t)$ is a population mean function, $K_x$ is the truncation lag and the $c_{ik}$ are uncorrelated random variables with variance $\lambda_k$. Using this, we construct simulated regressors

$$X_i^S(t) = \mu(t) + \sum_{k=1}^{K_x} c_{ik}\psi_k(t) + \varepsilon_i^X(t)$$
$$\boldsymbol{c}_i \sim \mathrm{N}(\mathbf{0}, \mathrm{diag}(\lambda_1, \ldots, \lambda_{K_x})); \ \varepsilon_i^X(t) \sim \mathrm{N}(0, \sigma_X^2).$$

This parametric construction of the simulated functional predictors is related to the application predictors through the mean function $\mu(t)$, the eigenfunctions $\psi_k(t)$, and the variance components $\lambda_k$. As in our application, the simulated predictors are observed on a grid of length 93.

We generate 100 such datasets for $I = 100$ and $I = 500$ and fit model (4.1) using both MCMC simulation and the variational approximation approach. For the MCMC simulation, we use chains of length 2500 with the first 1000 as burn-in. Representative examples of the MCMC model fits were inspected using trace and autocorrelation plots to ensure that the posterior samples were reasonable and that the comparison with variational Bayes was fair. To evaluate the ability of the proposed approach to estimate the functional coefficient $\gamma(t)$ we use the mean squared error (MSE) $\int_0^1 (\widehat{\gamma}(t) - \gamma(t))^2 dt$. A comparison of MSEs for the variational approach with the more computationally intensive MCMC sampling is given in table 1. To provide context for this table, in the left panel of Figure 3 we plot the estimated coefficient function resulting in the median MSE for $I = 100$.

Interestingly, when $I = 500$ the MSEs for MCMC sampling contain several large outliers, which raises the average MSE for the coefficient function in Table 1. Upon inspection, it was found that these large values corresponded to model fits in which the chains for $\boldsymbol{g}$ were bimodal. A large primary mode surrounded the true parameter value but a smaller, more

diffuse mode corresponded to a model overfit. We refit these models using as initial parameter values the estimates provided by the variational approach, which caused the bimodal behavior to disappear and brought the MSEs (and average MSE) down to levels similar to the remaining model fits.

We also quantify the quality of the variational approximation to the MCMC-sampled posterior by computing the accuracy for each parameter in the model using

$\text{Accuracy}=1-\frac{|p(\theta_l|\text{rest})-q^*(\theta_l)|}{2}\in[0,1]$; scores near 1 indicate a high 2 level of agreement between the two densities. Due to the large number of parameters in the model, we present only a subset of the average accuracies in Table 2. As with the MSE, context for this table in given in Figure 3. The accuracy of $\boldsymbol{g}$ is affected by the presence of outliers, attributable to the same bimodal MCMC samples that caused the very large MSEs appearing in Table 1.

Due to the substantial decrease in computation time using variational Bayes over MCMC methods, we are able to construct 95% bootstrap confidence intervals by sampling subjects with replacement and refitting model (4.1) using the variational approach. While credible intervals provided by MCMC or by a single variational fit are overly conservative, the bootstrap intervals are on average .46 times narrower and, averaged over the domain, have coverage probability 93.4% for $I = 100$ and 93.6% for $I = 500$. The far right panel of Figure 3 displays the coverage probabilities of the various credible and confidence intervals for $I = 500$.

As demonstrated in Table 2 and Figure 3, the variational approximation performs well in this functional regression setting, both in terms of low MSEs and of agreement the with MCMC-sample posterior density. This stems from the low posterior dependence between the parameters, which is assumed in the use of the density transform approach. Additionally, the use of the bootstrap allows the construction of confidence intervals that are not overly conservative.

Importantly, even in this simulation the computational burden is greatly reduced through the use of variational approximations. For $I = 100$, the MCMC sampling took on average 315 seconds, while the approximation was computed in on average 0.04 seconds (Dual Core 3.06GHz Processor; 4 GB RAM; OS X 10.6.4). For $I = 500$, the respective times were 1614 and 0.2 seconds. Constructing the bootstrap confidence intervals, based on 400 bootstrap samples, took on average an additional 20 and 76 seconds for $I = 100$ and $I = 500$, respectively.

### 4.2. Longitudinal functional regression

Next, we generate samples from the model

$$Y_{ij}=\beta_1+z_{ij}\beta_2+\boldsymbol{b}_i+\int_0^1 X_{ij}^S(t)\gamma(t)dt+\varepsilon_{ij}^Y \qquad \varepsilon_{ij}^Y\sim \mathrm{N}(0,\sigma_\gamma^2). \tag{4.2}$$

We take $I = 100$ subjects with $J = 3$ visits per subject; random effects $\boldsymbol{b}$ are $\mathrm{N}(0,\sigma_{\boldsymbol{b}}^2)$ with $\sigma_{\boldsymbol{b}}^2=5$. Again, we generate $z_i \sim \text{Unif}[-5, 5]$, take $\sigma_\gamma^2=5, \beta_1=\int_0^1\mu(t)\gamma(t)dt=12.68$, $\beta_2 = 3$, and select $\gamma(t) = \cos(2\pi t)$. The functional predictors $X_{ij}^S(t)$ are constructed as above; we take $\boldsymbol{c}_{ij} \sim \mathrm{N}(\boldsymbol{0}, \text{diag}(\lambda_1, \ldots, \lambda_{ijK_X}))$ so that the PC loadings are not correlated within subjects.

We fit model (2.8) for 100 simulated datasets. As in the cross-sectional case, we use chain lengths of 2500, with 1000 as burn-in, for the MCMC sampling. In Table 3 we display the average MSE of the estimated functional and scalar parameters and the subject-specific random effects. Again, the variational approximation performs as well as the MCMC sampling with a substantial difference in computation time: the MCMC sampling took on average 973 seconds, while the approximation was calculated in on average .2 seconds.

## 5. Application

In our scientific application, we analyze the association between measures of intracranial white matter and cognitive decline in multiple sclerosis patients. White matter is made up of myelinated axons, the long fibers used to transmit electrical signals in the brain, and is organized into bundles, or tracts. Major examples of white matter tracts are the corpus callosum, the corticospinal tracts, and the optic radiations. Here we focus on the corpus callosum, a collection of white-matter fibers which connects the two hemispheres of the brain.

Myelin, the fatty insulation surrounding white matter fibers, allows electrical signals to be propagated at high speeds along white matter tracts. Multiple sclerosis is a demyelinating autoimmune disease that causes in lesions in the white matter. These lesions disrupt electrical signals, and, over time, result in severe disability in affected patients. To measure cognitive disability, we use the Paced Auditory Serial Addition Test (PASAT), which assesses auditory processing speed and calculation ability. In this test, a proctor reads aloud a sequence of 60 numbers at three-second intervals, while the subject provides the sum of the previous two numbers spoken. This test has scores between 0 and 60 indicating the number of correct sums provided by the subject; the score 60 indicates the highest level of cognitive ability [8].

To quantify white matter, we use diffusion tensor imaging, a magnetic resonance imaging that measures the diffusivity of water in the brain. Because white matter is organized in bundles, water tends to diffuse anisotropically along the tract, which makes their reconstruction from MRI possible. By measuring diffusivity along several gradients, diffusion tensor imaging is able to produce detailed images of intracranial white matter [1, 2, 14, 17]. Moreover, continuous summaries of individual white matter tracts, parameterized by distance along the tract and called tract profiles, can be constructed from diffusion tensor images. Here we study the fractional anisotropy tract profile of the right corticospinal tract; this gives a measure of how anisotropic diffusion is along the tract.

Our study consists of 100 multiple sclerosis patients with between two and eight visits each; a total of 334 visits were observed. Study participants had ages between 21 and 71 years, and 63% were women. We fit model (2.8), using age and gender as non-functional covariates and the mean diffusivity tract profile of the corpus callosum as a functional predictor. We include subject-specific random intercepts to account for the repeated observations at the subject level. The model was fit using both the variational approximation and MCMC sampling; the results are shown in Figure 4.

Previous studies have linked damage in the corpus callosum to cognitive decline as measured by PASAT and other tests [15, 20]. However, these studies lacked the spatial information present in the functional treatment here, which proves to be important. From the estimated coefficient function and bootstrapped confidence intervals, we see that the region from roughly 0 to .2 is negatively associated with the PASAT outcome – that is, subjects with above-average mean diffusivity in these regions tend to have lower PASAT scores. A second region, from .65 to .8, is positively associated with the outcome. We base inference on the bootstrapped intervals due to the overly conservative coverage of the MCMC and

variational Bayes credible intervals; however, there is broad agreement between all intervals regarding the location of regions of interest. Note the interpretation of the coefficient function is marginal, rather than conditional on a subject's random intercept. The random intercepts are an important component of the model: a model including only random intercepts explains roughly 80% of the outcome variability, while adding functional and nonfunctional covariates raises this to 89%. Finally, age and gender were not found to be statistically significant, but were retained as scientifically important covariates. Their inclusion did not meaningfully affect the shape or significance of the functional predictor.

There is broad agreement between the variational Bayes and MCMC model fits: the point estimates of the coefficient and the random intercepts are very similar, and the credible intervals indicate the same regions of significance. On the other hand, the credible interval using MCMC is wider than that using variational Bayes. As noted above, the variational method can result in narrower confidence intervals if the approximating density is less diffuse than the MCMC-sampled posterior which appears to be the case here. In this application, we posit that the lesser importance of the functional predictors in comparison to the random intercepts leads to increased posterior variability in the estimated functional coefficient. Indeed, when we fit a model without the random subject-specific intercept the confidence intervals for Bayesian and variational Bayesian approximations became indistinguishable.

Also shown in Figure 4 as a grey band is the 95% bootstrap confidence interval, constructed by nonparametrically resampling subjects and fitting the longitudinal functional regression model using variational Bayes. Inference for the coefficient function is largely unchanged based on the bootstrap interval except in the region from .2 to .4, which does not appear to be significantly associated with the outcome. Although the credible intervals using variational Bayes are likely too narrow, the computational gain and accurate point estimates provided by this method allow for the construction of bootstrap confidence intervals, which performed much better in our simulations.

## 6. Discussion

The variational Bayes approach to functional regression was motivated by a pressing need for computationally feasible Bayesian inference in a large-data setting. We have developed iterative algorithms for approximate inference in both the cross-sectional and longitudinal regression settings, and analyzed a longitudinal neuroimaging study. The methods developed: 1) flexibly estimate all parameters in the cross-sectional and longitudinal functional regression models; 2) accurately approximate the posterior distributions of all model parameters; 3) retain the advantages of Bayesian inference, including the ability to jointly model the functional predictors and scalar outcomes and easily constructed credible bands; 4) require orders of magnitude less computational effort than MCMC techniques; and 5) allow the construction of nonparametric bootstrap confidence intervals, which seem to have good coverage probabilities.

A few limitations of the variational Bayes method are apparent. While our simulations indicate that variational techniques can be used with confidence in the functional regression setting, the approximation cannot be made more accurate by increasing computation time. Additionally, the iterative algorithms are based on involved algebraic derivations; those needed for functional regression have been carried out here, but additional work may be needed to adapt these algorithms to specific scientific settings. Lastly, the performance of credible intervals approximated using variational Bayes may not be satisfactory if the posterior distribution is multimodal or more diffuse than the approximating distribution, although the use of the nonparametric bootstrap can alleviate this issue.

Future work may proceed in several directions. The adaptation of the approach to non-Gaussian outcomes will expand the class of applications in which variational Bayes may be used for functional regression. Very large gains in computation time may be found in functional magnetic resonance imaging or other studies where the predictors are sampled at thousands or tens of thousands of points. More generally, variational Bayes has potential applications in several functional data analysis topics, including function-on-function regression and the decomposition of populations of functions.

## References


1. Basser P, Mattiello J, LeBihan D. MR Diffusion Tensor Spectroscopy and Imaging. Biophysical Journal. 1994; 66:259–267. [PubMed: 8130344]
2. Basser P, Pajevic S, Pierpaoli C, Duda J. In vivo fiber tractography using DT-MRI data. Magnetic Resonance in Medicine. 2000; 44:625–632. [PubMed: 11025519]
3. Bishop, CM. Pattern Recognition and Machine Learning. New York: Springer; 2006.
4. Cardot H, Ferraty F, Sarda P. Functional Linear Model. Statistics and Probability Letters. 1999; 45:11–22.
5. Crainiceanu CM, Goldsmith J. Bayesian Functional Data Analysis using WinBUGS. Journal of Statistical Software. 2010; 32:1–33.
6. Goldsmith J, Bobb J, Crainiceanu CM, Caffo B, Reich D. Penalized Functional Regression. Journal of Computational and Graphical Statistics. To Appear.
7. Goldsmith J, Crainiceanu CM, Caffo B, Reich D. A Case Study of Longitudinal Association Between Disability and Neuronal Tract Measurements. Under Review. 2011
8. Gronwall DMA. Paced auditory serial-addition task: A measure of recovery from concussion. Perceptual and Motor Skills. 1977; 44:367–373. [PubMed: 866038]
9. James GM, JW, JZ. Functional Linear Regression That's Interpretable. Annals of Statistics. 2009; 37:2083–2108.
10. Jordan MI. Graphical models. Statistical Science. 2004; 19:140–155.
11. Jordan MI, Ghahramani Z, Jaakkola TS, Saul LK. An Introduction to Variational Methods for Graphical Models. Machine Learning. 1999; 37:183–233.
12. Kullback S, Leibler D. On Information and Sufficiency. The Annals of Mathematical Statistics. 1951; 22:79–86.
13. Lang S, Brezger A. Bayesian P-splines. Journal of Computational and Graphical Statistics. 2004; 13:183–212.
14. LeBihan D, Mangin J, Poupon C, Clark C. Diffusion Tensor Imaging: Concepts and Applications. Journal of Magnetic Resonance Imaging. 2001; 13:534–546. [PubMed: 11276097]
15. Lin X, Tench CR, Morgan PS, Constantinescu CS. Use of combined conventional and quantitative MRI to quantify pathology related to cognitive impairment in multiple sclerosis. Journal of Neurology, Neurosurgery, and Psychiatry. 2008; 237:437–441.
16. McGrory CA, Titterington DM, Reeves R, Pettitt AN. Variational Bayes for Estimating the Parameters of a Hidden Potts Model. Statistics and Computing. 2009; 19:329–340.
17. Mori S, Barker P. Diffusion magnetic resonance imaging: its principle and applications. The Anatomical Record. 1999; 257:102–109. [PubMed: 10397783]
18. Müller H-G, Stadtmüller U. Generalized functional linear models. Annals of Statistics. 2005; 33:774–805.
19. Ormerod J, Wand MP. Explaining Variational Approximations. The American Statistician. 2010; 64:140–153.
20. Ozturk A, Smith S, Gordon-Lipkin E, Harrison D, Shiee N, Pham D, Caffo B, Calabresi P, Reich D. MRI of the corpus callosum in multiple sclerosis: association with disability. Multiple Sclerosis. 2010; 16:166–177. [PubMed: 20142309]
21. Ramsay, JO.; Silverman, BW. Functional Data Analysis. New York: Springer; 2005.
22. Reiss P, Ogden R. Functional Principal Component Regression and Functional Partial Least Squares. Journal of the American Statistical Association. 2007; 102:984–996.

23. Ruppert, D.; Wand, MP.; Carroll, RJ. Semiparametric Regression. Cambridge: Cambridge University Press; 2003.
24. Teschendorff AE, Wang Y, Barbosa-Morais NL, Brenton JD, Caldas C. A Variational Bayesian Mixture Modeling Framework for Cluster Analysis of Gene-Expression Data. Bioinformatics. 2005; 21:3025–3033. [PubMed: 15860564]
25. Titterington DM. Bayesian Methods for Neural Networks and Related Models. Statistical Science. 2004; 19:128–139.

## Appendix A: Derivations

In this appendix we derive the optimal densities $q^*$ for approximate Bayesian inference in the longitudinal functional regression model. For the cross-sectional case, one may omit the random effects $\boldsymbol{b}$. We recall that, given a partition $\{\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_L\}$ of the parameter space $\boldsymbol{\theta}$, the explicit solution for $q(\boldsymbol{\theta}_l)$, $1 \le l \le L$, has the form

$$q_l^*(\theta_l) \propto \exp\{E_{\theta_{-l}} \log p(\theta_l|\text{rest})\};\ 1 \le l \le L \tag{A.1}$$

where rest $\equiv \{\boldsymbol{y}, \boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_{l-1}, \boldsymbol{\theta}_{l+1}, \ldots, \boldsymbol{\theta}_L\}$

## A.1. Optimal densities for g and $\sigma_g^2$

Recall that $\boldsymbol{g} \sim \mathrm{N}(0, \sigma_{\boldsymbol{g}}^2 \boldsymbol{D})$ and $\sigma_{\boldsymbol{g}}^2 \sim \mathrm{IG}(A_{\boldsymbol{g}}, B_{\boldsymbol{g}})$. According to (A.1), the optimal densities are given by

$$q^*(\boldsymbol{g}) \propto \exp\{E_{-\boldsymbol{g}} \log p(\boldsymbol{g}|\text{rest})\} \quad \text{and} \quad q^*(\sigma_{\boldsymbol{g}}^2) \propto \exp\left\{E_{-\sigma_{\boldsymbol{g}}^2} \log p(\sigma_{\boldsymbol{g}}^2|\text{rest})\right\}.$$

The full conditional distribution $p(\boldsymbol{g}|\text{rest})$ appearing above is given by

$$\begin{aligned} p(\boldsymbol{g}|\text{rest}) &\propto p(\boldsymbol{Y}|\beta, \boldsymbol{b}, \boldsymbol{C}, \boldsymbol{g}, \sigma_Y^2) p(\boldsymbol{g}|\sigma_{\boldsymbol{g}}^2) \\ &\propto \exp\left[-\tfrac{1}{2}\left\{\tfrac{1}{\sigma_Y^2}(\boldsymbol{Y} - \boldsymbol{z}\beta - \boldsymbol{Z}\boldsymbol{b} - \boldsymbol{C}\boldsymbol{M}\boldsymbol{g})^T (\boldsymbol{Y} - \boldsymbol{z}\beta - \boldsymbol{Z}\boldsymbol{b} - \boldsymbol{C}\boldsymbol{M}\boldsymbol{g}) + \tfrac{1}{\sigma_{\boldsymbol{g}}^2}\boldsymbol{g}^T \boldsymbol{D}^{-1}\boldsymbol{g}\right\}\right] \\ &\propto \exp\left[-\tfrac{1}{2}\left\{\boldsymbol{g}^T\left(\tfrac{1}{\sigma_Y^2}\boldsymbol{M}^T\boldsymbol{C}^T\boldsymbol{C}\boldsymbol{M} + \tfrac{1}{\sigma_{\boldsymbol{g}}^2}\boldsymbol{D}^{-1}\right)\boldsymbol{g}\right.\right. \\ &\quad \left.\left. -2\left(\left(\boldsymbol{Y}^T - \beta^T\boldsymbol{z}^T - \boldsymbol{Z}\boldsymbol{b}\right)\left(\tfrac{1}{\sigma_Y^2}\boldsymbol{C}\boldsymbol{M}\right)\right)\boldsymbol{g}\right\}\right]. \end{aligned}$$

Therefore the optimal density $q^*(\boldsymbol{g})$ is

$$\begin{aligned} q^*(\boldsymbol{g}) &\propto \exp\left[-\tfrac{1}{2}\mathrm{E}_{-\boldsymbol{g}}\left\{\boldsymbol{g}^T\left(\tfrac{1}{\sigma_Y^2}\boldsymbol{M}^T\boldsymbol{C}^T\boldsymbol{C}\boldsymbol{M} + \tfrac{1}{\sigma_{\boldsymbol{g}}^2}\boldsymbol{D}^{-1}\right)\boldsymbol{g} - 2\left(\left(\boldsymbol{Y}^T - \beta^T\boldsymbol{z}^T - \boldsymbol{b}^T\boldsymbol{Z}^T\right)\left(\tfrac{1}{\sigma_Y^2}\boldsymbol{C}\boldsymbol{M}\right)\right)\boldsymbol{g}\right\}\right] \\ &\propto \exp\left\{-\tfrac{1}{2}\left(\boldsymbol{g} - \mu_{q(\boldsymbol{g})}\right)^T \Sigma_{q(\boldsymbol{g})}^{-1}\left(\boldsymbol{g} - \mu_{q(\boldsymbol{g})}\right)\right\} \end{aligned}$$

where

$$\begin{aligned}\sum_{q(\boldsymbol{g})} &= \left\{\mu_{q(1/\sigma_{\boldsymbol{Y}}^2)}\boldsymbol{M}^T(\mu_{q(\boldsymbol{C})}^T\mu_{q(\boldsymbol{C})}+n\sum_{q(\boldsymbol{C})})\boldsymbol{M}+\mu_{q(1/\sigma_{\boldsymbol{g}}^2)}\boldsymbol{D}^{-1}\right\}^{-1}\\ \mu_{q(\boldsymbol{g})} &= \sum_{q(\boldsymbol{g})}\left\{\mu_{q(1/\sigma_{\boldsymbol{Y}}^2)}\left(\boldsymbol{Y}^T-\mu_{q(\beta)}^T\boldsymbol{z}^T-\mu_{q(\boldsymbol{b})}^T\boldsymbol{Z}^T\right)\left(\mu_{q(\boldsymbol{C})}\boldsymbol{M}\right)\right\}^T.\end{aligned} \tag{A.2}$$

Thus the optimal density $q^*(\boldsymbol{g})$ is $N(\boldsymbol{\mu}_{q(\boldsymbol{g})}, \boldsymbol{\Sigma}_{q(\boldsymbol{g})})$.

Further, the full conditional $p(\sigma_{\boldsymbol{g}}^2|\text{rest})$ is given by

$$\begin{aligned}p(\sigma_{\boldsymbol{g}}^2|\text{rest}) &\propto p(\boldsymbol{g}|\sigma_{\boldsymbol{g}}^2)p(\sigma_{\boldsymbol{g}}^2)\\ &\propto (\sigma_{\boldsymbol{g}}^2)^{-K_g/2}\exp\left\{-\frac{1}{2\sigma_{\boldsymbol{g}}^2}\boldsymbol{g}^T\boldsymbol{D}^{-1}\boldsymbol{g}\right\}\times(\sigma_{\boldsymbol{g}}^2)^{-A_{\boldsymbol{g}}-1}\exp\left\{-\frac{1}{\sigma_{\boldsymbol{g}}^2}B_{\boldsymbol{g}}\right\}\\ &=(\sigma_{\boldsymbol{g}}^2)^{-A_{\boldsymbol{g}}-K_g/2-1}\exp\left\{-\frac{1}{\sigma_{\boldsymbol{g}}^2}\left(B_{\boldsymbol{g}}+\frac{1}{2}\boldsymbol{g}^T\boldsymbol{D}^{-1}\boldsymbol{g}\right)\right\}\end{aligned}$$

so that the optimal density $q^*(\sigma_{\boldsymbol{g}}^2)$ is

$$q^*(\sigma_{\boldsymbol{g}}^2)\propto\exp\left\{-(A_{\boldsymbol{g}}+K_g/2+1)\log(\sigma_{\boldsymbol{g}}^2)-\frac{1}{\sigma_{\boldsymbol{g}}^2}\mathrm{E}_{-\sigma_{\boldsymbol{g}}^2}\left(B_{\boldsymbol{g}}+\frac{1}{2}\boldsymbol{g}^T\boldsymbol{D}^{-1}\boldsymbol{g}\right)\right\}.$$

Thus $q^*(\sigma_{\boldsymbol{g}}^2)$is $\mathrm{IG}\left(A_{\boldsymbol{g}}+K_g/2, B_{q(\sigma_{\boldsymbol{g}}^2)}\right)$ where

$$B_{q(\sigma_{\boldsymbol{g}}^2)}=B_{\boldsymbol{g}}+\tfrac{1}{2}\left(\mu_{q(\boldsymbol{g})}^T\boldsymbol{D}^{-1}\mu_{q(\boldsymbol{g})}+\mathrm{tr}(\boldsymbol{D}^{-1}\sum_{q(\boldsymbol{g})})\right) \tag{A.3}$$

Note that, when $q(\sigma_{\boldsymbol{g}}^2)=q^*(\sigma_{\boldsymbol{g}}^2)$, the term $\mu_{q(1/\sigma_{\boldsymbol{g}}^2)}$ appearing in (A.2) is equal to $\frac{A_{\boldsymbol{g}}+K_g/2}{B_{q(\sigma_{\boldsymbol{g}}^2)}}$.

## A.2. Optimal densities for b and $\sigma_{\boldsymbol{b}}^2$

Recall that $\boldsymbol{b}\sim N(0,\sigma_{\boldsymbol{b}}^2\boldsymbol{I})$ and $\sigma_{\boldsymbol{b}}^2\sim \mathrm{IG}(A_{\boldsymbol{b}}, B_{\boldsymbol{b}})$.

The full conditional distribution $p(\boldsymbol{b}|\text{rest})$ is given by

$$\begin{aligned}p(\boldsymbol{b}|\text{rest}) &\propto p(\boldsymbol{Y}|\beta,\boldsymbol{b},\boldsymbol{C},\boldsymbol{g},\sigma_{\boldsymbol{Y}}^2)p(\boldsymbol{b}|\sigma_{\boldsymbol{b}}^2)\\ &\propto\exp\left[-\tfrac{1}{2}\left\{\tfrac{1}{\sigma_{\boldsymbol{Y}}^2}(\boldsymbol{Y}-\boldsymbol{z}\beta-\boldsymbol{Z}\boldsymbol{b}-\boldsymbol{C}\boldsymbol{M}\boldsymbol{g})^T(\boldsymbol{Y}-\boldsymbol{z}\beta-\boldsymbol{Z}\boldsymbol{b}-\boldsymbol{C}\boldsymbol{M}\boldsymbol{g})+\tfrac{1}{\sigma_{\boldsymbol{b}}^2}\boldsymbol{b}^T\boldsymbol{I}^{-1}\boldsymbol{b}\right\}\right]\\ &\propto\exp\left[-\tfrac{1}{2}\left\{\boldsymbol{b}^T\left(\tfrac{1}{\sigma_{\boldsymbol{Y}}^2}\boldsymbol{Z}^T\boldsymbol{Z}+\tfrac{1}{\sigma_{\boldsymbol{b}}^2}\boldsymbol{I}\right)\boldsymbol{b}-2\left(\left(\boldsymbol{Y}^T-\beta^T\boldsymbol{z}^T-\boldsymbol{g}^T\boldsymbol{M}^T\boldsymbol{C}^T\right)\left(\tfrac{1}{\sigma_{\boldsymbol{Y}}^2}\boldsymbol{Z}\right)\right)\boldsymbol{b}\right\}\right]\end{aligned}$$

Therefore, by (A.1), the optimal density $q^*(\boldsymbol{b})$ is

$$q^*(\boldsymbol{b}) \propto \exp\left[-\frac{1}{2}\mathrm{E}_{-\boldsymbol{b}}\left\{\boldsymbol{b}^T\left(\frac{1}{\sigma_Y^2}\boldsymbol{Z}^T\boldsymbol{Z}+\frac{1}{\sigma_{\boldsymbol{b}}^2}\boldsymbol{I}\right)\boldsymbol{b}-2\left(\left(\boldsymbol{Y}^T-\beta^T\boldsymbol{z}^T-\boldsymbol{g}^T\boldsymbol{M}^T\boldsymbol{C}^T\right)\left(\frac{1}{\sigma_Y^2}\boldsymbol{Z}\right)\right)\boldsymbol{b}\right\}\right].$$

After taking the expectation above, the optimal density $q^*(\boldsymbol{b})$ is $\mathrm{N}(\boldsymbol{\mu}_{q(\boldsymbol{b})}, \boldsymbol{\Sigma}_{q(\boldsymbol{b})})$ where

$$\begin{aligned}\Sigma_{q(\boldsymbol{b})} &= \left\{\mu_{q(1/\sigma_Y^2)}\boldsymbol{Z}^T\boldsymbol{Z}+\mu_{q(1/\sigma_{\boldsymbol{b}}^2)}\boldsymbol{I}\right\}^{-1}\\ \mu_{q(\boldsymbol{b})} &= \Sigma_{q(\boldsymbol{b})}\left\{\mu_{q(1/\sigma_Y^2)}\left(\boldsymbol{Y}^T-\mu_{q(\beta)}^T\boldsymbol{z}^T-\mu_{q(\boldsymbol{g})}^T\boldsymbol{M}\mu_{q(\boldsymbol{C})}^T\right)\boldsymbol{Z}\right\}^T.\end{aligned} \tag{A.4}$$

Further, the full conditional $p(\sigma_{\boldsymbol{b}}^2|\text{rest})$ is given by

$$\begin{aligned}p(\sigma_{\boldsymbol{b}}^2|\text{rest}) &\propto p(\boldsymbol{b}|\sigma_{\boldsymbol{b}}^2)p(\sigma_{\boldsymbol{b}}^2)\\ &\propto (\sigma_{\boldsymbol{b}}^2)^{-n/2}\exp\left\{-\tfrac{1}{2\sigma_{\boldsymbol{b}}^2}\boldsymbol{b}^T\boldsymbol{b}\right\}\times(\sigma_{\boldsymbol{b}}^2)^{-A_{\boldsymbol{b}}-1}\exp\left\{-\tfrac{1}{\sigma_g^2}B_{\boldsymbol{b}}\right\}\\ &= (\sigma_{\boldsymbol{b}}^2)^{-A_{\boldsymbol{b}}-n/2-1}\exp\left\{-\tfrac{1}{\sigma_{\boldsymbol{b}}^2}\left(B_{\boldsymbol{b}}+\tfrac{1}{2}\boldsymbol{b}^T\boldsymbol{b}\right)\right\}\end{aligned}$$

so that the optimal density $q^*(\sigma_{\boldsymbol{b}}^2)$ is

$$q^*(\sigma_{\boldsymbol{b}}^2) \propto \exp\left\{-(A_{\boldsymbol{b}}+n/2+1)\log(\sigma_{\boldsymbol{b}}^2)-\frac{1}{\sigma_{\boldsymbol{b}}^2}\mathrm{E}_{-\sigma_{\boldsymbol{b}}^2}\left(B_{\boldsymbol{b}}+\frac{1}{2}\boldsymbol{b}^T\boldsymbol{b}\right)\right\}.$$

Thus $q^*(\sigma_{\boldsymbol{b}}^2)$ is $\mathrm{IG}(A_b+n/2, B_{q(\sigma_{\boldsymbol{b}}^2)})$ where

$$B_{q(\sigma_{\boldsymbol{b}}^2)} = B_{\boldsymbol{b}}+\tfrac{1}{2}\left(\mu_{q(\boldsymbol{b})}^T\mu_{q(\boldsymbol{b})}+\mathrm{tr}\left(\sum\nolimits_{q(\boldsymbol{b})}\right)\right) \tag{A.5}$$

Note that, when $q(\sigma_{\boldsymbol{b}}^2)=q^*(\sigma_{\boldsymbol{b}}^2)$, the term $\mu_{q(1/\sigma_{\boldsymbol{b}}^2)}$ appearing in (A.4) is equal to $\frac{A_{\boldsymbol{b}}+n/2}{B_{q(\sigma_{\boldsymbol{b}}^2)}}$.

## A.3. Optimal densities for C and $\lambda_j$

Recall that $\boldsymbol{c}_{ij} \sim \mathrm{N}(0, \Lambda)$, where $\Lambda = \mathrm{diag}(\lambda_1, \ldots, \lambda_{K_X})$ and $\lambda_k \sim \mathrm{IG}(A_\lambda, B_\lambda)$ for $1 \le k \le K_X$. In the following, we continue to use $\boldsymbol{c}_{ij}$ as the PC loadings for subject $i$ at visit $j$ and $\boldsymbol{C}$ as the matrix constructed by row-stacking the $\boldsymbol{c}_{ij}$. We additionally use $\boldsymbol{\mu}_{q(\boldsymbol{c}),ij}$ as the expected value of $\boldsymbol{c}_{ij}$ with respect to the $q(\boldsymbol{C})$ distribution and $\boldsymbol{\mu}_{q(\boldsymbol{C})}$ as the matrix constructed by row-stacking the $\boldsymbol{\mu}_{q(\boldsymbol{c}),ij}$. Finally, let $\Lambda_q^{-1}=\mathrm{diag}(\mu_{q(1/\lambda_1)}, \ldots, \mu_{q(1/\lambda_k)})$

The full conditional distribution $p(\boldsymbol{C}|\text{rest})$ is given by

$$p(\boldsymbol{C}|\text{rest}) \propto p(\boldsymbol{Y}|\beta, \boldsymbol{b}, \boldsymbol{C}, \boldsymbol{g}, \sigma_{\boldsymbol{Y}}^2) p(\boldsymbol{W}|\boldsymbol{C}, \sigma_{\boldsymbol{X}}^2) p(C|\Lambda)$$

$$\propto \exp\left[-\tfrac{1}{2}\text{tr}\left\{\tfrac{1}{\sigma_{\boldsymbol{Y}}^2}\left(\boldsymbol{Y}^T - \beta^T \boldsymbol{z}^T - \boldsymbol{b}^T \boldsymbol{Z}^T - \boldsymbol{g}^T \boldsymbol{M}^T \boldsymbol{C}^T\right)^T \left(\boldsymbol{Y}^T - \beta^T \boldsymbol{z}^T - \boldsymbol{b}^T \boldsymbol{Z}^T - \boldsymbol{g}^T \boldsymbol{M}^T \boldsymbol{C}^T\right)\right\}\right] \times \exp\left[-\tfrac{1}{2}\text{tr}\left\{\tfrac{1}{\sigma_{\boldsymbol{X}}^2}\left(\boldsymbol{W}^T - \boldsymbol{\psi}\boldsymbol{C}^T\right)^T \left(\boldsymbol{W}^T - \boldsymbol{\psi}\boldsymbol{C}^T\right)\right\}\right]$$

$$\propto \exp\left[-\tfrac{1}{2}\text{tr}\left\{\boldsymbol{C}\left(\tfrac{\boldsymbol{M}\boldsymbol{g}\boldsymbol{g}^T\boldsymbol{M}^T}{\sigma_{\boldsymbol{Y}}^2} + \tfrac{\boldsymbol{\psi}^T\boldsymbol{\psi}}{\sigma_{\boldsymbol{X}}^2} + \Lambda^{-1}\right)\boldsymbol{C}^T - 2\left(\tfrac{\boldsymbol{Y}\boldsymbol{g}^T\boldsymbol{M}^T}{\sigma_{\boldsymbol{Y}}^2} - \tfrac{\boldsymbol{z}\beta\boldsymbol{g}^T\boldsymbol{M}^T}{\sigma_{\boldsymbol{Y}}^2} - \tfrac{\boldsymbol{Z}\boldsymbol{b}\boldsymbol{g}^T\boldsymbol{M}^T}{\sigma_{\boldsymbol{Y}}^2} + \tfrac{\boldsymbol{W}\boldsymbol{\psi}}{\sigma_{\boldsymbol{X}}^2}\right)\boldsymbol{C}^T\right\}\right]$$

Therefore, by (A.1), the optimal density $q^*(\boldsymbol{C})$ is

$$q^*(\boldsymbol{C}) \propto \exp\left[-\tfrac{1}{2}\text{E}_{-\boldsymbol{C}}\left[\text{tr}\left\{\boldsymbol{C}\left(\tfrac{\boldsymbol{M}\boldsymbol{g}\boldsymbol{g}^T\boldsymbol{M}^T}{\sigma_{\boldsymbol{Y}}^2} + \tfrac{\boldsymbol{\psi}^T\boldsymbol{\psi}}{\sigma_{\boldsymbol{X}}^2} + \Lambda^{-1}\right)\boldsymbol{C}^T - 2\left(\tfrac{\boldsymbol{Y}\boldsymbol{g}^T\boldsymbol{M}^T}{\sigma_{\boldsymbol{Y}}^2} - \tfrac{\boldsymbol{z}\beta\boldsymbol{g}^T\boldsymbol{M}^T}{\sigma_{\boldsymbol{Y}}^2} - \tfrac{\boldsymbol{Z}\boldsymbol{b}\boldsymbol{g}^T\boldsymbol{M}^T}{\sigma_{\boldsymbol{Y}}^2} + \tfrac{\boldsymbol{W}\boldsymbol{\psi}}{\sigma_{\boldsymbol{X}}^2}\right)\boldsymbol{C}^T\right\}\right]\right]$$

$$\propto \exp\left\{-\tfrac{1}{2}\left[\sum_{i=1}^{I}\sum_{j=1}^{J}\left\{(\boldsymbol{c}_{ij})^T - (\mu_{q(\boldsymbol{c}),ij})^T\right\}^T \Sigma_{q(\boldsymbol{C})}^T \left\{(\boldsymbol{c}_{ij})^T - (\mu_{q(\boldsymbol{c}),ij})^T\right\}\right]\right\}$$

where

$$\Sigma_{q(\boldsymbol{C})} = \left\{\mu_{q(1/\sigma_{\boldsymbol{Y}}^2)} \boldsymbol{M}(\mu_{q(\boldsymbol{g})}\mu_{q(\boldsymbol{g})}^T + \Sigma_{q(\boldsymbol{g})})\boldsymbol{M}^T + \mu_{q(1/\sigma_{\boldsymbol{X}}^2)} \boldsymbol{\psi}^T\boldsymbol{\psi} + \Lambda_q^{-1}\right\}^{-1}$$

$$\mu_{q(\boldsymbol{C})}^T = \Sigma_{q(\boldsymbol{C})}\left(\mu_{q(1/\sigma_{\boldsymbol{Y}}^2)} \boldsymbol{Y}\mu_{q(\boldsymbol{g})}^T\boldsymbol{M}^T - \mu_{q(1/\sigma_{\boldsymbol{Y}}^2)}\mu_{q(\beta)}\boldsymbol{z}\mu_{q(\boldsymbol{g})}^T\boldsymbol{M}^T - \mu_{q(1/\sigma_{\boldsymbol{Y}}^2)}\boldsymbol{Z}\mu_{q(\boldsymbol{b})}\mu_{q(\boldsymbol{g})}^T\boldsymbol{M}^T + \mu_{q(1/\sigma_{\boldsymbol{X}}^2)}\boldsymbol{W}\boldsymbol{\psi}\right)^T.$$

(A.6)

Thus the optimal density $q^*(\boldsymbol{C})$ is a product of Normally distributed random vectors sharing a common covariance matrix and with means the rows of $\boldsymbol{\mu}_{q(\boldsymbol{C})}$.

In the derivation of the optimal density $q^*(\lambda_k)$, $1 \le k \le K_x$, we let $\boldsymbol{C}^k$ denote the $k^{th}$ column of $\boldsymbol{C}$ and $\mu_{q(\boldsymbol{C})}^k$ denote the $k^{th}$ column of $\boldsymbol{\mu}_{q(\boldsymbol{C})}$. Further, we let $(\boldsymbol{\Sigma}_{q(\boldsymbol{C})})_{kk}$ denote the $(k, k)^{th}$ element of $\boldsymbol{\Sigma}_{q(\boldsymbol{C})}$. The full conditional $p(\lambda_k|\text{rest})$ is given by

$$p(\lambda_k|\text{rest}) \propto p(\boldsymbol{C}^k|\lambda_k)p(\lambda_k)$$

$$\propto (\lambda_k)^{-(nJ)/2}\exp\left\{-\tfrac{1}{2\lambda_k}(\boldsymbol{C}^k)^T(\boldsymbol{C}^k)\right\} \times (\lambda_k)^{-A_\lambda - 1}\exp\left\{-\tfrac{1}{\lambda_k}B_\lambda\right\}$$

$$= (\lambda_k)^{-A_\lambda - (nJ)/2 - 1}\exp\left\{-\tfrac{1}{\lambda_k}\left(B_\lambda + \tfrac{1}{2}(\boldsymbol{C}^k)^T(\boldsymbol{C}^k)\right)\right\}$$

so that the optimal density $q^*(\lambda_k)$ is

$$q^*(\lambda_k) \propto \exp\left\{-(A_\lambda + (nJ)/2 + 1)\log(\lambda_k) - \frac{1}{\lambda_k}\text{E}_{-\lambda_k}\left(B_\lambda + \frac{1}{2}(\boldsymbol{C}^k)^T(\boldsymbol{C}^k)\right)\right\}.$$

Thus $q^*(\lambda_k)$ is IG($A_\lambda + (nJ)/2$, $B_{q(\lambda k)}$ where

$$B_{q(\lambda_k)}=B_\lambda+\tfrac{1}{2}\left((\mu^k_{q(\boldsymbol{C})})^T(\mu^k_{q(\boldsymbol{C})})+n(\textstyle\sum_{q(\boldsymbol{C})})_{kk}\right) \tag{A.7}$$

Note that, when $q(\lambda_k) = q^*(\lambda_k)$, the term $(k, k)^{th}$ entry of $\Lambda_q^{-1}$ appearing in (A.6) is equal to $\frac{A_\lambda+(nJ)/2}{B_{q(\lambda_k)}}$.

## A.4. Optimal density for β

Recall that $\beta \sim \mathrm{N}(0, \sigma_\beta^2 \boldsymbol{I})$.

The full conditional distribution $p(\boldsymbol{\beta}|\text{rest})$ is given by

$$\begin{aligned} p(\beta|\text{rest}) &\propto p(\boldsymbol{Y}|\beta, \boldsymbol{b}, \boldsymbol{C}, \boldsymbol{g}, \sigma_{\boldsymbol{Y}}^2)p(\beta) \\ &\propto \exp\left[-\tfrac{1}{2}\left\{\tfrac{1}{\sigma_{\boldsymbol{Y}}^2}(\boldsymbol{Y}-\boldsymbol{z}\beta-\boldsymbol{Z}\boldsymbol{b}-\boldsymbol{C}\boldsymbol{M}\boldsymbol{g})^T(\boldsymbol{Y}-\boldsymbol{z}\beta-\boldsymbol{Z}\boldsymbol{b}-\boldsymbol{C}\boldsymbol{M}\boldsymbol{g})+\tfrac{1}{\sigma_\beta^2}\beta^T\boldsymbol{I}^{-1}\beta\right\}\right] \\ &\propto \exp\left[-\tfrac{1}{2}\left\{\beta^T\left(\tfrac{1}{\sigma_{\boldsymbol{Y}}^2}\boldsymbol{z}^T\boldsymbol{z}+\tfrac{1}{\sigma_\beta^2}\boldsymbol{I}\right)\beta-2\left(\left(\boldsymbol{Y}^T-\boldsymbol{b}^T\boldsymbol{Z}^T-\boldsymbol{g}^T\boldsymbol{M}^T\boldsymbol{C}^T\right)\left(\tfrac{1}{\sigma_{\boldsymbol{Y}}^2}\boldsymbol{z}\right)\right)\beta\right\}\right] \end{aligned}$$

Therefore, by (A.1), the optimal density $q^*(\boldsymbol{\beta})$ is

$$q^*(\beta) \propto \exp\left[-\frac{1}{2}\mathrm{E}_{-\beta}\left\{\beta^T\left(\frac{1}{\sigma_{\boldsymbol{Y}}^2}\boldsymbol{z}^T\boldsymbol{z}+\frac{1}{\sigma_\beta^2}\boldsymbol{I}\right)\beta-2\left(\left(\boldsymbol{Y}^T-\boldsymbol{b}^T\boldsymbol{Z}^T-\boldsymbol{g}^T\boldsymbol{M}^T\boldsymbol{C}^T\right)\left(\frac{1}{\sigma_{\boldsymbol{Y}}^2}\boldsymbol{z}\right)\right)\beta\right\}\right].$$

After taking the expectation above, the optimal density $q^*(\boldsymbol{\beta})$ is $\mathrm{N}(\boldsymbol{\mu}_{q(\boldsymbol{\beta})}, \boldsymbol{\Sigma}_{q(\boldsymbol{\beta})})$ where

$$\begin{aligned} \textstyle\sum_{q(\beta)} &= \left\{\mu_{q(1/\sigma_{\boldsymbol{Y}}^2)}\boldsymbol{z}^T\boldsymbol{z}+\tfrac{1}{\sigma_\beta^2}\boldsymbol{I}\right\}^{-1} \\ \mu_{q(\beta)} &= \textstyle\sum_{q(\beta)}\left\{\mu_{q(1/\sigma_{\boldsymbol{Y}}^2)}\left(\boldsymbol{Y}^T-\mu_{q(\boldsymbol{b})}^T\boldsymbol{Z}^T-\mu_{q(\boldsymbol{g})}^T\boldsymbol{M}\mu_{q(\boldsymbol{C})}^T\right)\boldsymbol{z}\right\}^T. \end{aligned} \tag{A.8}$$

## A.5. Optimal density for $\sigma_X^2$

Recall that the functional predictors are observed over a grid of length $N$. The full conditional $p(\sigma_{\boldsymbol{X}}^2|\text{rest})$ is given by

$$\begin{aligned} p(\sigma_{\boldsymbol{X}}^2|\text{rest}) &\propto p(\boldsymbol{W}|\sigma_{\boldsymbol{X}}^2)p(\sigma_{\boldsymbol{X}}^2) \\ &\propto \left[\prod_{i=1}^{I}\prod_{j=1}^{J}(\sigma_{\boldsymbol{X}}^2)^{-N/2}\exp\left\{-\tfrac{1}{2\sigma_{\boldsymbol{X}}^2}(W_{ij}-C_{ij}\boldsymbol{\psi}^T)(W_{ij}-C_{ij}\boldsymbol{\psi}^T)^T\right\}\right]\cdot\left[(\sigma_{\boldsymbol{X}}^2)^{-A_{\boldsymbol{X}}-1}\exp\left\{-\tfrac{1}{\sigma_{\boldsymbol{X}}^2}B_{\boldsymbol{X}}\right\}\right] \\ &= (\sigma_{\boldsymbol{X}}^2)^{-A_{\boldsymbol{X}}-(nJN)/2-1}\exp\left\{-\tfrac{1}{\sigma_{\boldsymbol{X}}^2}(B_{\boldsymbol{X}}+\tfrac{1}{2}\sum_{i=1}^{I}\sum_{j=1}^{J}(\boldsymbol{W}_{ij}-\boldsymbol{C}_{ij}\boldsymbol{\psi}^T)(\boldsymbol{W}_{ij}-\boldsymbol{C}_{ij}\boldsymbol{\psi}^T)^T)\right\} \end{aligned}$$

so that the optimal density $q^*\sigma_X^2$ is

$$q^*(\sigma_X^2) \propto \exp\left\{-A_X+(nNJ/2+1)\log(\sigma_X^2) - \frac{1}{\sigma_X^2}\left(B_X+\frac{1}{2}\mathrm{E}_{-\sigma_X^2}\sum_{i=1}^{I}\sum_{j=1}^{J}(\boldsymbol{W}_{ij}-\boldsymbol{C}_{ij}\boldsymbol{\psi}^T)(\boldsymbol{W}_{ij}-\boldsymbol{C}_{ij}\boldsymbol{\psi}^T)^T\right)\right\}.$$

Thus $q^*(\sigma_X^2)$ is $\mathrm{IG}(A_X+(nNJ)/2, B_{q(\sigma_X^2)})$ where

$$B_{q(\sigma_X^2)}=B_X+\frac{1}{2}\left\{\sum_{i=1}^{I}\sum_{j=1}^{J}\left\|(\boldsymbol{W}_{ij}-\mu_{q(\boldsymbol{C}),ij}\boldsymbol{\psi}^T)^T\right\|^2+(nJ)\mathrm{tr}\left(\boldsymbol{\psi}^T\boldsymbol{\psi}\sum\nolimits_{q(C)}\right)\right\} \tag{A.9}$$

Note that, when $q(\sigma_X^2)=q^*(\sigma_X^2)$, the term $\mu_{q(1/\sigma_X^2)}$ appearing in (A.6) is equal to $\frac{A_X+nJ/2}{B_{q(\sigma_X^2)}}$.

## A.6. Optimal density for $\sigma_Y^2$

Finally, the full conditional $p(\sigma_Y^2|\mathrm{rest})$ is given by

$$\begin{aligned} p(\sigma_Y^2|\mathrm{rest}) &\propto p(\boldsymbol{Y}|\beta,\boldsymbol{b},\boldsymbol{C},\boldsymbol{g},\sigma_Y^2)p(\sigma_Y^2) \\ &\propto (\sigma_Y^2)^{-(nJ)/2}\exp\left\{-\tfrac{1}{2\sigma_Y^2}\left\|\boldsymbol{Y}-\boldsymbol{z}\beta-\boldsymbol{Z}\boldsymbol{b}-\boldsymbol{C}\boldsymbol{M}\boldsymbol{g}\right\|^2\right\}\times(\sigma_Y^2)^{-A_Y-1}\exp\left\{-\tfrac{1}{\sigma_Y^2}B_Y\right\} \\ &=(\sigma_Y^2)^{-A_Y/(nJ)/2-1}\exp\left\{-\tfrac{1}{\sigma_Y^2}\left(B_Y+\tfrac{1}{2}\left\|\boldsymbol{Y}-\boldsymbol{z}\beta-\boldsymbol{Z}\boldsymbol{b}-\boldsymbol{C}\boldsymbol{M}\boldsymbol{g}\right\|^2\right)\right\} \end{aligned}$$

so that the optimal density $q^*(\sigma_Y^2)$ is

$$q^*(\sigma_Y^2) \propto \exp\left\{-(A_Y+(nJ)/2+1)\log(\sigma_Y^2) - \frac{1}{\sigma_Y^2}\left(B_Y+\frac{1}{2}\mathrm{E}_{-\sigma_Y^2}\left\|\boldsymbol{Y}-\boldsymbol{z}\beta-\boldsymbol{Z}\boldsymbol{b}-\boldsymbol{C}\boldsymbol{M}\boldsymbol{g}\right\|^2\right)\right\}.$$

Next, we see that

$$\begin{aligned}
&\mathrm{E}_{-\sigma_{\boldsymbol{Y}}^2}\left\{\left\|\boldsymbol{Y}-z\mu_{q(\beta)}-\boldsymbol{Z}\boldsymbol{b}-\mu_{q(\boldsymbol{C})}\boldsymbol{M}\mu_{q(\boldsymbol{g})}\right\|^2\right\}\\
&=\left\|\boldsymbol{Y}-z\mu_{q(\beta)}-\boldsymbol{Z}\boldsymbol{b}-\mu_{q(\boldsymbol{C})}\boldsymbol{M}\mu_{q(\boldsymbol{g})}\right\|^2\\
&+\mathrm{E}_{-\sigma_{\boldsymbol{Y}}^2}\left\{(z\mu_{q(\beta)}-z\beta)^T(z\mu_{q(\beta)}-z\beta)\right\}\\
&+\mathrm{E}_{-\sigma_{\boldsymbol{Y}}^2}\left\{(\mu_{q(\boldsymbol{C})}\boldsymbol{M}\mu_{q(\boldsymbol{g})}-\boldsymbol{C}\boldsymbol{M}\boldsymbol{g})^T(\mu_{q(\boldsymbol{C})}\boldsymbol{M}\mu_{q(\boldsymbol{g})}-\boldsymbol{C}\boldsymbol{M}\boldsymbol{g})\right\}\\
&+\mathrm{E}_{-\sigma_{\boldsymbol{Y}}^2}\left\{(\boldsymbol{Z}\mu_{q(\boldsymbol{b})}-\boldsymbol{Z}\boldsymbol{b})^T(\boldsymbol{Z}\mu_{q(\boldsymbol{b})}-\boldsymbol{Z}\boldsymbol{b})\right\}\\
&=\left\|\boldsymbol{Y}-z\mu_{q(\beta)}-\boldsymbol{Z}\boldsymbol{b}-\mu_{q(\boldsymbol{C})}\boldsymbol{M}\mu_{q(\boldsymbol{g})}\right\|^2+\mathrm{tr}(z^Tz\textstyle\sum_{q(\beta)})\\
&-\left(\mu_{q(\boldsymbol{g})}\boldsymbol{M}\mu_{q(\boldsymbol{C})}\right)^T\left(\mu_{q(\boldsymbol{g})}\boldsymbol{M}\mu_{q(\boldsymbol{C})}\right)\\
&+\mu_{q(\boldsymbol{g})}^T\boldsymbol{M}^T(\mu_{q(\boldsymbol{C})}^T\mu_{q(\boldsymbol{C})}+(nJ)\textstyle\sum_{q(\boldsymbol{C})})\boldsymbol{M}\mu_{q(\boldsymbol{g})}\\
&+\mathrm{tr}\left\{\boldsymbol{M}^T\left(\mu_{q(\boldsymbol{C})}^T\mu_{q(\boldsymbol{C})}+(nJ)\textstyle\sum_{q(\boldsymbol{C})}\right)\boldsymbol{M}\textstyle\sum_{q(\boldsymbol{g})}\right\}\\
&+\mathrm{tr}\left\{\boldsymbol{Z}^T\boldsymbol{Z}\textstyle\sum_{q(\boldsymbol{b})}\right\}.
\end{aligned}\tag{A.10}$$

Thus $q^*(\sigma_{\boldsymbol{Y}}^2)$ is $\mathrm{IG}(A_{\boldsymbol{Y}}+(nJ)/2, B_{q(\sigma_{\boldsymbol{Y}}^2)})$ where

$$\begin{aligned}
B_{q(\sigma_{\boldsymbol{Y}}^2)}=B_{\boldsymbol{Y}}+\tfrac{1}{2}\Big[&\left\|\boldsymbol{Y}-z\mu_{q(\beta)}-\boldsymbol{Z}\boldsymbol{b}-\mu_{q(\boldsymbol{C})}\boldsymbol{M}\mu_{q(\boldsymbol{g})}\right\|^2+\mathrm{tr}\left(z^Tz\textstyle\sum_{q(\beta)}\right)\\
&-\left(\mu_{q(\boldsymbol{g})}\boldsymbol{M}\mu_{q(\boldsymbol{C})}\right)^T\left(\mu_{q(\boldsymbol{g})}\boldsymbol{M}\mu_{q(\boldsymbol{C})}\right)\\
&+\mu_{q(\boldsymbol{g})}^T\boldsymbol{M}^T(\mu_{q(\boldsymbol{C})}^T\mu_{q(\boldsymbol{C})}+(nJ)\textstyle\sum_{q(\boldsymbol{C})})\boldsymbol{M}\mu_{q(\boldsymbol{g})}\\
&+\mathrm{tr}\left\{\boldsymbol{M}^T\left(\mu_{q(\boldsymbol{C})}^T\mu_{q(\boldsymbol{C})}+(nJ)\textstyle\sum_{q(\boldsymbol{C})}\right)\boldsymbol{M}\textstyle\sum_{q(\boldsymbol{g})}\right\}+\mathrm{tr}\left\{\boldsymbol{Z}^T\boldsymbol{Z}\textstyle\sum_{q(\boldsymbol{b})}\right\}\Big]
\end{aligned}\tag{A.11}$$

Note that, when $q(\sigma_{\boldsymbol{Y}}^2)=q_{\sigma_{\boldsymbol{Y}}^2}^*(\sigma_{\boldsymbol{Y}}^2)$, the term $\mu_{q(1/\sigma_{\boldsymbol{Y}}^2)}$ appearing regularly above is equal to $\frac{A_{\boldsymbol{Y}}+(nJ)/2}{B_{q(\sigma_{\boldsymbol{Y}}^2)}}$.

## Appendix B: Expression for p(Y, W; q)

In this appendix we derive an expression for the lower bound of the log likelihood. This quantity is use to monitor convergence in Algorithm 1, and its derivation takes advantage of the order of updates in the algorithm to simplify the expression.

We have that $\log\underline{p}(\boldsymbol{Y},\boldsymbol{W};q^*)=\int q(\theta)\log(\frac{p(\boldsymbol{Y},\boldsymbol{W},\theta)}{q^*(\theta)})d\theta=\mathrm{E}_{q^*}[\log p(\boldsymbol{Y},\boldsymbol{W},\theta)-\log q^*(\theta)]$. Now,

$$\begin{aligned}
&\mathrm{E}_{q^*}[\log p(\boldsymbol{Y},\boldsymbol{W},\theta)-\log q^*(\theta)]\\
&=\mathrm{E}_{q^*}[\log p(\boldsymbol{Y}|,\beta,\boldsymbol{b},\boldsymbol{C},\boldsymbol{g},\sigma_{\boldsymbol{Y}}^2)]+\mathrm{E}_{q^*}[\log p(\boldsymbol{W}|\boldsymbol{C},\sigma_{\boldsymbol{X}}^2)]\\
&+\mathrm{E}_{q^*}[\log p(\beta)-\log\{q^*(\beta)\}]+\mathrm{E}_{q^*}[\log p(\boldsymbol{g}|\sigma_{\boldsymbol{g}}^2)-\log\{q^*(\boldsymbol{g})\}]\\
&+\mathrm{E}_{q^*}[\log p(\boldsymbol{b}|\sigma_{\boldsymbol{b}}^2)-\log\{q^*(\boldsymbol{b})\}]+\mathrm{E}_{q^*}[\log p(\boldsymbol{C}|\lambda)-\log\{q^*(\boldsymbol{C})\}]\\
&+\mathrm{E}_{q^*}[\log p(\sigma_{\boldsymbol{g}}^2)-\log\{q^*(\sigma_{\boldsymbol{g}}^2)\}]+\mathrm{E}_{q^*}[\log p(\sigma_{\boldsymbol{b}}^2)-\log\{q^*(\sigma_{\boldsymbol{b}}^2)\}]\\
&+\mathrm{E}_{q^*}[\log p(\sigma_{\boldsymbol{Y}}^2)-\log\{q^*(\sigma_{\boldsymbol{Y}}^2)\}]+\mathrm{E}_{q^*}[\log p(\sigma_{\boldsymbol{X}}^2)-\log\{q^*(\sigma_{\boldsymbol{X}}^2)\}]\\
&+\sum_{k=1}^{K_x}\mathrm{E}_{q^*}[\log p(\lambda_k)-\log\{q^*(\lambda_k)\}]
\end{aligned}\tag{B.1}$$

The first term appearing in (B.1) is

$$
\begin{aligned}
&\mathrm{E}_{q^*}[\log p(\boldsymbol{Y}|\beta,\boldsymbol{b},\boldsymbol{C},\boldsymbol{g},\sigma_{\boldsymbol{Y}}^2]\\
=&\mathrm{E}_{q^*}\left[-\tfrac{nJ}{2}\log(2\pi)-\tfrac{1}{2}\log(|\sigma_{\boldsymbol{Y}}^2\boldsymbol{I}|)-\tfrac{1}{2}\tfrac{1}{\sigma_{\boldsymbol{Y}}^2}\left\|\boldsymbol{Y}-\boldsymbol{z}\beta-\boldsymbol{Z}\boldsymbol{b}-\boldsymbol{C}\boldsymbol{M}\boldsymbol{g}\right\|^2\right]\\
=&-\tfrac{nJ}{2}\log(2\pi)-\tfrac{nJ}{2}\mathrm{E}_{q^*}\log(\sigma_{\boldsymbol{Y}}^2)\\
&-\tfrac{1}{2}\mu_{q(1/\sigma_{\boldsymbol{Y}}^2)}\left\{\left\|\boldsymbol{Y}-\boldsymbol{z}\mu_{q(\beta)}-\boldsymbol{Z}\boldsymbol{b}-\mu_{q(\boldsymbol{C})}\boldsymbol{M}\mu_{q(\boldsymbol{g})}\right\|^2+\mathrm{tr}(\boldsymbol{z}^T\boldsymbol{z}\textstyle\sum_{q(\beta)})\right.\\
&-\left(\mu_{q(\boldsymbol{g})}\boldsymbol{M}\mu_{q(\boldsymbol{C})}\right)^T\left(\mu_{q(\boldsymbol{g})}\boldsymbol{M}\mu_{q(\boldsymbol{C})}\right)\\
&+\mu_{q(\boldsymbol{g})}^T\boldsymbol{M}^T(\mu_{q(\boldsymbol{C})}^T\mu_{q(\boldsymbol{C})}+(nJ)\textstyle\sum_{q(\boldsymbol{C})})\boldsymbol{M}\mu_{q(\boldsymbol{g})}\\
&\left.+\mathrm{tr}\left\{\boldsymbol{M}^T\left(\mu_{q(\boldsymbol{C})}^T\mu_{q(\boldsymbol{C})}+(nJ)\textstyle\sum_{q(\boldsymbol{C})}\right)\boldsymbol{M}\sum_{q(\boldsymbol{g})}\right\}+\mathrm{tr}\left\{\boldsymbol{Z}^T\boldsymbol{Z}\textstyle\sum_{q(\boldsymbol{b})}\right\}\right\}.
\end{aligned}
$$

The second term is

$$
\begin{aligned}
&\mathrm{E}_{q^*}[\log p(\boldsymbol{W}|\boldsymbol{C},\sigma_{\boldsymbol{X}}^2)]=\mathrm{E}_{q^*}\left[\sum_{i=1}^{n}\sum_{j=1}^{J}\left(\tfrac{-N}{2}\log(2\pi)-\tfrac{N}{2}\log(\sigma_{\boldsymbol{X}}^2)-\tfrac{1}{2}\tfrac{1}{\sigma_{\boldsymbol{X}}^2}\left\|\left(\boldsymbol{W}_{ij}-\boldsymbol{C}_{ij}\boldsymbol{\psi}^T\right)^T\right\|^2\right)\right]\\
=&-\tfrac{nJN}{2}\log(2\pi)-\tfrac{nJN}{2}\mathrm{E}_{q^*}\log(\sigma_{\boldsymbol{X}}^2)-\tfrac{1}{2}\mu_{q(1/\sigma_{\boldsymbol{X}}^2)}\left[\sum_{i=1}^{n}\sum_{j=1}^{J}\left\|\left(\boldsymbol{W}_{ij}-\mu_{q(\boldsymbol{c}),ij}\boldsymbol{\psi}^T\right)^T\right\|^2+(nJ)\mathrm{tr}\left(\boldsymbol{\psi}^T\boldsymbol{\psi}\textstyle\sum_{q(\boldsymbol{C})}\right)\right].
\end{aligned}
$$

Next, we have

$$
\begin{aligned}
&\mathrm{E}_{q^*}[\log p(\beta)-\log\{q^*(\beta)\}]=\mathrm{E}_{q^*}\left[\tfrac{1}{2}\log\left(\tfrac{|\sum_{q(\beta)}|}{\sigma_\beta^{2p}}\right)-\tfrac{1}{2\sigma_\beta^2}\beta^T\beta+\tfrac{1}{2}(\beta-\mu_{q(\beta)})^T\textstyle\sum_{q(\beta)}^{-1}(\beta-\mu_{q(\beta)})\right]\\
=&\tfrac{1}{2}\log\left(\tfrac{|\sum_{q(\beta)}|}{\sigma_\beta^{2p}}\right)-\tfrac{1}{2\sigma_\beta^2}\left\{\mu_{q(\beta)}^T\mu_{q(\beta)}+\mathrm{tr}(\textstyle\sum_{q(\beta)})\right\}+\tfrac{p}{2}.
\end{aligned}
$$

The fourth term is given by

$$
\begin{aligned}
&\mathrm{E}_{q^*}[\log p(\boldsymbol{g}|\sigma_{\boldsymbol{g}}^2)-\log\{q^*(\boldsymbol{g})\}]=\mathrm{E}_{q^*}\left[\tfrac{1}{2}\log\left(\tfrac{|\sum_{q(\boldsymbol{g})}|}{|\sigma_{\boldsymbol{g}}^2D|}\right)-\tfrac{1}{2}\tfrac{1}{\sigma_{\boldsymbol{g}}^2}\boldsymbol{g}^T\boldsymbol{D}^{-1}\boldsymbol{g}+\tfrac{1}{2}(\boldsymbol{g}-\mu_{q(\boldsymbol{g})})^T\textstyle\sum_{q(\boldsymbol{g})}^{-1}(\boldsymbol{g}-\mu_{q(\boldsymbol{g})})\right]\\
=&\tfrac{1}{2}\log(|\textstyle\sum_{q(\boldsymbol{g})}|)-\tfrac{K_g}{2}\mathrm{E}_{q^*}[\log\sigma_{\boldsymbol{g}}^2]-\tfrac{1}{2}\log(|\boldsymbol{D}|)-\mu_{q(1/\sigma_{\boldsymbol{g}}^2)}\tfrac{1}{2}\left\{\mu_{q(\boldsymbol{g})}^T\boldsymbol{D}^{-1}\mu_{q(\boldsymbol{g})}+\mathrm{tr}(\boldsymbol{D}^{-1}\textstyle\sum_{q(\boldsymbol{g})})\right\}\tfrac{K_g}{2}.
\end{aligned}
$$

Further, we have

$$
\begin{aligned}
&\mathrm{E}_{q^*}[\log p(\boldsymbol{C}|\Lambda)-\log\{q^*(\boldsymbol{C})\}]=\mathrm{E}_{q^*}\left[\sum_{k=1}^{K_x}\left\{\tfrac{nJ}{2}\log\left(\tfrac{(\sum_{q(\boldsymbol{C})})_{kk}}{\lambda_k}\right)-\tfrac{1}{2}\tfrac{1}{\lambda_k}\boldsymbol{C}_k^T\boldsymbol{C}_k+\tfrac{1}{2}\left((\boldsymbol{C}^k)-(\mu_{q(\boldsymbol{C})}^k)^T\right)^T(\textstyle\sum_{q(\boldsymbol{C})}^{-1})_{kk}\left((\boldsymbol{C}^k)-(\mu_{q(\boldsymbol{C})}^k)^T\right)\right\}\right]\\
=&\sum_{k=1}^{K_x}\tfrac{nJ}{2}\log\{(\textstyle\sum_{q(\boldsymbol{C})})_{kk}\}-\sum_{k=1}^{K_x}\tfrac{nJ}{2}\mathrm{E}_{q^*}\log(\lambda_k)-\sum_{j=1}^{K_x}\tfrac{1}{2}\mu_{q(1/\lambda_k)}\left((\mu_{q(\boldsymbol{C})}^k)^T(\mu_{q(\boldsymbol{C})}^k)+(nJ)(\textstyle\sum_{q(\boldsymbol{C})})_{kk}\right)+\tfrac{(nJ)K_x}{2}.
\end{aligned}
$$

The sixth term in (B.1) is

$$
\begin{aligned}
&\mathrm{E}_{q^*}[\log p(\boldsymbol{b}|\sigma_{\boldsymbol{b}}^2)-q^*(\boldsymbol{b})]=\mathrm{E}_{q^*}\left[\tfrac{1}{2}\log\left(\tfrac{|\sum_{q(\boldsymbol{b})}|}{|\sigma_{\boldsymbol{b}}^2\boldsymbol{I}|}\right)-\tfrac{1}{2}\tfrac{1}{\sigma_{\boldsymbol{b}}^2}\boldsymbol{b}^T\boldsymbol{b}+\tfrac{1}{2}(\boldsymbol{b}-\mu_{q(\boldsymbol{b})})^T\textstyle\sum_{q(\boldsymbol{b})}^{-1}(\boldsymbol{b}-\mu_{q(\boldsymbol{b})})\right]\\
=&\tfrac{1}{2}\mathrm{E}_{q^*}[\log(|\textstyle\sum_{q(\boldsymbol{b})}|)-\tfrac{n}{2}\mathrm{E}_{q^*}[\log\sigma_{\boldsymbol{b}}^2]-\mu_{q(1/\sigma_{\boldsymbol{b}}^2)}\tfrac{1}{2}\left[\mu_{q(\boldsymbol{b})}^T\mu_{q(\boldsymbol{b})}+\mathrm{tr}(\textstyle\sum_{q(\boldsymbol{b})})\right]+\tfrac{n}{2}.
\end{aligned}
$$

Next,

$$\begin{aligned}&\mathrm{E}_{q^*}[\log p(\sigma_{\boldsymbol{g}}^2) - \log\{q^*(\sigma_{\boldsymbol{g}}^2)\}]\\ =&\tfrac{K_g}{2}\mathrm{E}_{q^*}\log(\sigma_{\boldsymbol{g}}^2)+\mu_{q(1/\sigma_{\boldsymbol{g}}^2)}\left(B_{q(\sigma_{\boldsymbol{g}}^2)} - B_{\boldsymbol{g}}\right)+A_{\boldsymbol{g}}\log(B_{\boldsymbol{g}}) - \log\{\Gamma(A_{\boldsymbol{g}})\}\\ &-\left(A_{\boldsymbol{g}}+\tfrac{K_g}{2}\right)\log\left(B_{q(\sigma_{\boldsymbol{g}}^2)}\right)+\log\left\{\Gamma\left(A_{\boldsymbol{g}}+\tfrac{K_g}{2}\right)\right\}.\end{aligned}$$

Additionally, the eighth term in (B.1) is

$$\begin{aligned}&\mathrm{E}_{q^*}[\log p(\sigma_{\boldsymbol{b}}^2) - \log\{q^*(\sigma_{\boldsymbol{b}}^2)\}]\\ =&\tfrac{n}{2}\mathrm{E}_{q^*}\log(\sigma_{\boldsymbol{b}}^2)+\mu_{q(1/\sigma_{\boldsymbol{b}}^2)}\left(B_{q(\sigma_{\boldsymbol{b}}^2)} - B_{\boldsymbol{b}}\right)+A_{\boldsymbol{b}}\log(B_{\boldsymbol{b}}) - \log\{\Gamma(A_{\boldsymbol{b}})\}\\ &-\left(A_{\boldsymbol{b}}+\tfrac{n}{2}\right)\log\left(B_{q(\sigma_{\boldsymbol{b}}^2)}\right)+\log\left\{\Gamma\left(A_{\boldsymbol{b}}+\tfrac{n}{2}\right)\right\}.\end{aligned}$$

Next, we have

$$\begin{aligned}&\mathrm{E}_{q^*}[\log p(\sigma_{\boldsymbol{Y}}^2) - \log\{q^*(\sigma_{\boldsymbol{Y}}^2)\}]\\ =&\tfrac{nJ}{2}\mathrm{E}_{q^*}\log(\sigma_{\boldsymbol{Y}}^2)+\mu_{q(1/\sigma_{\boldsymbol{Y}}^2)}\left(B_{q(\sigma_{\boldsymbol{Y}}^2)} - B_{\boldsymbol{Y}}\right)+A_{\boldsymbol{Y}}\log(B_{\boldsymbol{Y}}) - \log\{\Gamma(A_{\boldsymbol{Y}})\}\\ &-\left(A_{\boldsymbol{Y}}+\tfrac{n}{2}\right)\log\left(B_{q(\sigma_{\boldsymbol{Y}}^2)}\right)+\log\left\{\Gamma\left(A_{\boldsymbol{Y}}+\tfrac{n}{2}\right)\right\}\end{aligned}$$

The tenth term is

$$\begin{aligned}&\mathrm{E}_{q^*}[\log p(\sigma_{\boldsymbol{X}}^2) - \log\{q^*(\sigma_{\boldsymbol{X}}^2)\}]\\ =&\tfrac{nJN}{2}\mathrm{E}_{q^*}\log(\sigma_{\boldsymbol{X}}^2)+\mu_{q(1/\sigma_{\boldsymbol{X}}^2)}\left(B_{q(\sigma_{\boldsymbol{X}}^2)} - B_{\boldsymbol{X}}\right)+A_{\boldsymbol{X}}\log(B_{\boldsymbol{X}})\\ &-\log\{\Gamma(A_{\boldsymbol{X}})\} - \left(A_{\boldsymbol{X}}+\tfrac{nJN}{2}\right)\log\left(B_{q(\sigma_{\boldsymbol{X}}^2)}\right)+\log\left\{\Gamma\left(A_{\boldsymbol{X}}+\tfrac{nJN}{2}\right)\right\}\end{aligned}$$

Finally, for $1 \le k \le K_X$

$$\begin{aligned}&\mathrm{E}_{q^*}[\log p(\lambda_k) - \log\{q^*(\lambda_k)\}]\\ =&\tfrac{nJ}{2}\mathrm{E}_{q^*}\log(\lambda_k)+\mu_{q(1/\lambda_K)}\left(B_{q(\lambda_k)} - B_{\lambda}\right)+A_{\lambda}\log(B_{\lambda}) - \log\{\Gamma(A_{\lambda})\}\\ &-\left(A_{\lambda}+\tfrac{nJ}{2}\right)\log\left(B_{q(\lambda_k)}\right)+\log\left\{\Gamma\left(A_{\lambda}+\tfrac{nJ}{2}\right)\right\}\end{aligned}$$

We combine the above factors noting that many terms cancel. For example, the terms $-\frac{nJ}{2}\mathrm{E}_{q^*}\log(\sigma_{\boldsymbol{Y}}^2)$ and $\frac{nJ}{2}\mathrm{E}_{q^*}\log(\sigma_{\boldsymbol{Y}}^2)$ appear in $\mathrm{E}_{q^*}[\log p(\boldsymbol{Y}|\beta, \boldsymbol{b}, \boldsymbol{C}, \boldsymbol{g}, \sigma_{\boldsymbol{Y}}^2)]$ and $\mathrm{E}_{q^*}[\log p(\sigma_{\boldsymbol{Y}}^2) - \log\{q^*(\sigma_{\boldsymbol{Y}}^2)]$ respectively. Moreover, we can make substitutions for terms appearing in the updates given in Algorithm 1 and again simplify the expression. An example is to combine $-\mu_{q(1/\sigma_{\boldsymbol{g}}^2)}B_{\boldsymbol{g}}$ and $-\mu_{q(1/\sigma_{\boldsymbol{g}}^2)}\frac{1}{2}[\mu_{q(\boldsymbol{g})}^T\boldsymbol{D}^{-1}\mu_{q(\boldsymbol{g})}+\mathrm{tr}(\boldsymbol{D}^{-1}\sum_{q(\boldsymbol{b})})]$ and substitute for $-\mu_{q(1/\sigma_{\boldsymbol{g}}^2)}B_{q(\sigma_{\boldsymbol{g}}^2)}$; this term cancels with another appearing in $\mathrm{E}_{q^*}[\log p(\sigma_{\boldsymbol{g}}^2) - \log\{q^*(\sigma_{\boldsymbol{g}}^2)\}]$. Thus we have

$$
\begin{aligned}
&\log \underline{p}(\boldsymbol{Y}, \boldsymbol{W}; q) \\
&= \tfrac{-nJ}{2}\log(2\pi) + \tfrac{-nJN}{2}\log(2\pi)\tfrac{1}{2}\log\left(\frac{|\sum_{q(\beta)}|}{\sigma_\beta^{2p}}\right) \\
&\quad -\tfrac{1}{2\sigma_\beta^2}\left\{\mu_{q(\beta)}^T\mu_{q(\beta)} + \mathrm{tr}(\textstyle\sum_{q(\beta)})\right\} + \tfrac{p}{2} \\
&\quad + \tfrac{1}{2}\log(|\textstyle\sum_{q(\boldsymbol{g})}|)] + \tfrac{K_g}{2} + \sum_{k=1}^{K_x}\tfrac{nJ}{2}\log((\textstyle\sum_{q(\boldsymbol{C})})_{kk}) + \tfrac{nJK_x}{2} \\
&\quad + \tfrac{1}{2}\log(|\textstyle\sum_{q(\boldsymbol{b})}|) + \tfrac{n}{2} \\
&\quad + A_{\boldsymbol{g}}\log(B_{\boldsymbol{g}}) - \log(\Gamma(A_{\boldsymbol{g}})) - \left(A_{\boldsymbol{g}} + \tfrac{K_g}{2}\right)\log\left(B_{q(\sigma_g^2)}\right) \\
&\quad + \log\left\{\Gamma\left(A_{\boldsymbol{g}} + \tfrac{K_g}{2}\right)\right\} \\
&\quad + A_{\boldsymbol{b}}\log(B_{\boldsymbol{b}}) - \log(\Gamma(A_{\boldsymbol{b}})) - \left(A_{\boldsymbol{b}} + \tfrac{n}{2}\right)\log\left(B_{q(\sigma_b^2)}\right) + \log\left\{\Gamma\left(A_{\boldsymbol{b}} + \tfrac{n}{2}\right)\right\} \\
&\quad + A_{\boldsymbol{Y}}\log(B_{\boldsymbol{Y}}) - \log(\Gamma(A_{\boldsymbol{Y}})) - \left(A_{\boldsymbol{Y}} + \tfrac{nJ}{2}\right)\log\left(B_{q(\sigma_{\boldsymbol{Y}}^2)}\right) \\
&\quad + \log\left\{\Gamma\left(A_{\boldsymbol{Y}} + \tfrac{nJ}{2}\right)\right\} \\
&\quad + A_{\boldsymbol{X}}\log(B_{\boldsymbol{X}}) - \log(\Gamma(A_{\boldsymbol{X}})) - \left(A_{\boldsymbol{X}} + \tfrac{nJN}{2}\right)\log\left(B_{q(\sigma_{\boldsymbol{X}}^2)}\right) \\
&\quad + \log\left\{\Gamma\left(A_{\boldsymbol{X}} + \tfrac{nJN}{2}\right)\right\} \\
&\quad + \sum_{k=1}^{K_x}\left[A_\lambda\log(B_\lambda) - \log(\Gamma(A_\lambda)) - \left(A_\lambda + \tfrac{nJ}{2}\right)\log\left(B_{q(\lambda_k)}\right)\right. \\
&\quad \left. + \log\left\{\Gamma\left(A_\lambda + \tfrac{nJ}{2}\right)\right\}\right]
\end{aligned}
$$

Using const. to represent an additive constant that is not affected by updates in Algorithm 1, we have

$$
\begin{aligned}
&\log \underline{p}(\boldsymbol{Y}, \boldsymbol{W}; q) \\
&= \tfrac{1}{2}\log\left(\frac{|\sum_{q(\beta)}|}{\sigma_\beta^{2p}}\right) - \tfrac{1}{2\sigma_\beta^2}\left\{\mu_{q(\beta)}^T\mu_{q(\beta)} + \mathrm{tr}(\textstyle\sum_{q(\beta)})\right\} \\
&\quad + \tfrac{1}{2}\mathrm{E}_{q^*}[\log(|\textstyle\sum_{q(\boldsymbol{g})}|)] + \sum_{k=1}^{K_x}\tfrac{nJ}{2}\log((\textstyle\sum_{q(\boldsymbol{C})})_{kk}) + \tfrac{1}{2}\mathrm{E}_{q^*}[\log(|\textstyle\sum_{q(\boldsymbol{b})}|)] \\
&\quad - \left(A_{\boldsymbol{g}} + \tfrac{K_g}{2}\right)\log\left(B_{q(\sigma_g^2)}\right) - \left(A_{\boldsymbol{b}} + \tfrac{n}{2}\right)\log\left(B_{q(\sigma_b^2)}\right) \\
&\quad - \left(A_{\boldsymbol{Y}} + \tfrac{nJ}{2}\right)\log\left(B_{q(\sigma_{\boldsymbol{Y}}^2)}\right) \\
&\quad - \left(A_{\boldsymbol{X}} + \tfrac{nJN}{2}\right)\log\left(B_{q(\sigma_{\boldsymbol{X}}^2)}\right) - \sum_{k=1}^{K_x}\left\{\left(A_\lambda + \tfrac{nJ}{2}\right)\log\left(B_{q(\lambda_k)}\right)\right\} + \mathrm{const.}
\end{aligned}
$$

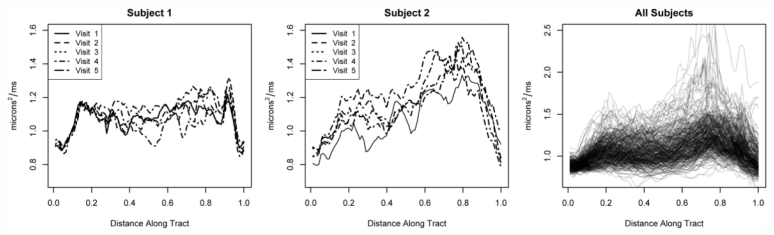


**Fig 1.**

The functional predictor used in our diffusion tensor imaging application. The left and middle panels show the functional predictors observed for individual subjects; the right panel shows the collection of all observed functions.

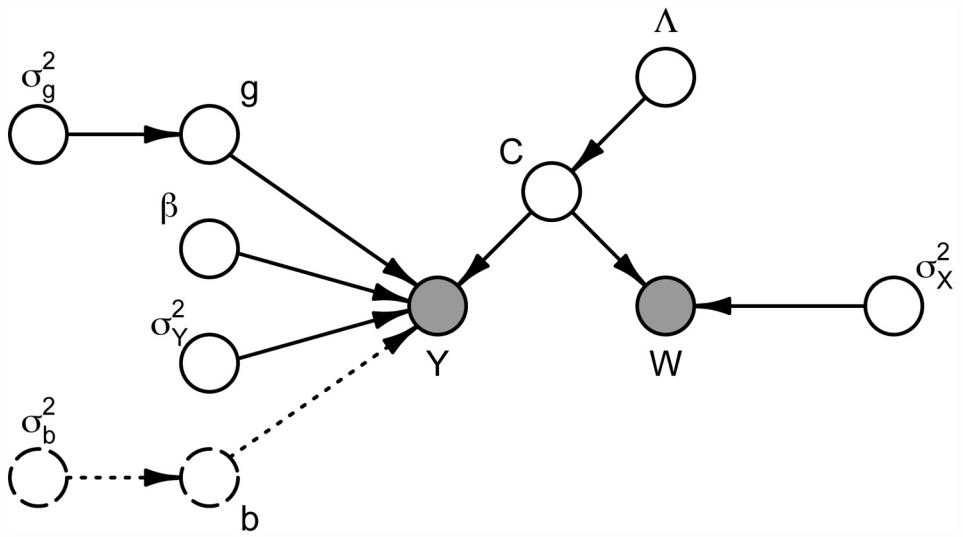


**Fig 2.**

Directed acyclic graph corresponding to the functional regression model (2.5). Shaded nodes correspond to observed data, and unshaded nodes to model parameters. Arrows indicate conditional dependence. The nodes for $\sigma_{\boldsymbol{b}}^2$ and $\boldsymbol{b}$, shown as dashed lines, appear in the longitudinal functional regression model (2.8) but not in the cross-sectional model (2.5).

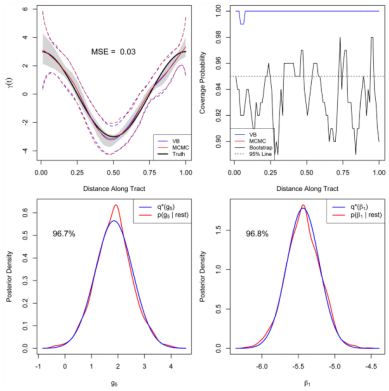


**Fig 3.**

The top left panel shows the estimated coefficient function corresponding to the median MSE = 0.049, as well as variational and MCMC 95% credible intervals (dashed lines) and the 95% bootstrap interval (shaded region). The top-right panel displays the coverage probabilities of the credible intervals over the domain of the predictor (note there is perfect overlap of the VB and MCMC coverage probabilities). The bottom panels show posterior densities estimated by variational approximations and by MCMC sampling from the same simulated dataset (shown in dashed and solid lines, respectively), and provide the accuracy of the approximation expressed as a percent.

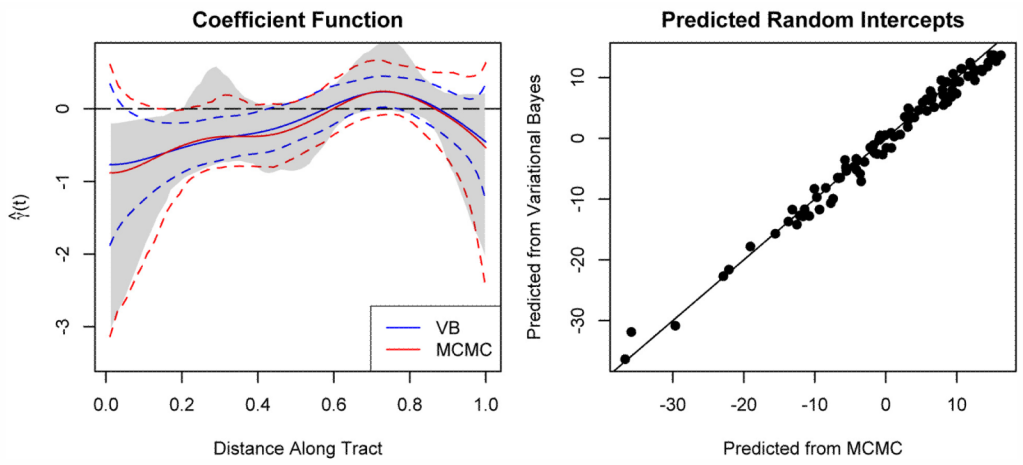


**Fig 4.**

Results of fitting model (2.8) to the diffusion tensor imaging dataset. The left panel shows the estimated coefficient function; credible intervals for both methods are shown in dashed lines, and the nonparametrically bootstrapped interval shown in grey; the right panel shows the random intercepts predicted by both variational Bayes and MCMC sampling.

**Table 1**

Average integrated MSE for $\gamma(t)$ and average MSE for the non-functional covariates $\beta_1$, $\beta_2$ estimated using the variational approximation, taken over 100 simulated datasets. For I = 500, large outlier for the MCMC MSE were removed in the calculation of the average

| | | $\gamma(t)$ | $\beta_1$ | $\beta_2$ |
|---|---|---|---|---|
| $I = 100$ | VB | .050 | .071 | .051 |
| | MCMC | .054 | .071 | .051 |
| $I = 500$ | VB | .046 | .008 | .001 |
| | MCMC | .120 | .008 | .001 |

**Table 2**

The accuracy of the variational approximation to the MCMC-sampled posterior, expressed as a percentage, for a subset of parameters in the cross-sectional functional regression model (2.5)

| **Accuracy** | $g_5$ | $g_{20}$ | $c_{1,1}$ | $c_{1,10}$ | $\lambda_{1,1}$ | $\lambda_{1,10}$ | $\sigma^2_Y$ |
|---|---|---|---|---|---|---|---|
| $I = 100$ | 96.3 | 95.1 | 98.3 | 98.0 | 96.9 | 97.2 | 95.0 |
| $I = 500$ | 86.7 | 82.6 | 97.6 | 97.8 | 97.6 | 88.3 | 96.3 |

**Table 3**

Average integrated MSE for $\gamma(t)$ and average MSE for the non-functional covariates $\beta_1$, $\beta_2$ estimated using the variational approximation, taken over 100 simulated datasets

| | | $\gamma(t)$ | $\beta_1$ | $\beta_2$ |
|---|---|---|---|---|
| $I = 100, J = 3$ | VB | .026 | .0003 | .0002 |
| | MCMC | .030 | .0002 | .0002 |

**Algorithm 1**

Iterative scheme for obtaining the parameters in the optimal densities in the longitudinal functional regression model (2.8).

Initialize: $B_{q(\sigma_g^2)} \ldots B_{q(\sigma_Y^2)} > 0$, , $\boldsymbol{\mu}_{q(C)} = \mathbf{0}$, $\boldsymbol{\mu}_{q(g)} = \mathbf{0}$, $\boldsymbol{\mu}_{q(\beta)} = \mathbf{0}$, $\boldsymbol{\Sigma}_{q(g)} = \boldsymbol{I}$, $\Lambda_q = \boldsymbol{I}$.

Cycle:

$$\boldsymbol{\Sigma}_{q(\beta)} \leftarrow \left\{ \mu_{q\left(1/\sigma_Y^2\right)} z^T z + \frac{1}{\sigma_\beta^2} I \right\}^{-1}$$

$$\boldsymbol{\mu}_{q(\beta)} \leftarrow \boldsymbol{\Sigma}_{q(\beta)} \left\{ \mu_{q\left(1/\sigma_Y^2\right)} \left( Y^T - \boldsymbol{\mu}_{q(b)}^T Z^T - \boldsymbol{\mu}_{q(g)}^T M \boldsymbol{\mu}_{q(C)}^T \right) z \right\}^T$$

$$\boldsymbol{\Sigma}_{q(b)} \leftarrow \left\{ \mu_{q\left(1/\sigma_Y^2\right)} Z^T Z + \mu_{q\left(1/\sigma_b^2\right)} I \right\}^{-1}$$

$$\boldsymbol{\mu}_{q(b)} \leftarrow \boldsymbol{\Sigma}_{q(b)} \left\{ \mu_{q\left(1/\sigma_Y^2\right)} \left( Y^T - \boldsymbol{\mu}_{q(\beta)}^T z^T - \boldsymbol{\mu}_{q(g)}^T M \boldsymbol{\mu}_{q(C)}^T \right) Z \right\}^T$$

$$\boldsymbol{\Sigma}_{q(C)} \leftarrow \left\{ \mu_{q\left(1/\sigma_Y^2\right)} M (\boldsymbol{\mu}_{q(g)} \boldsymbol{\mu}_{q(g)}^T + \boldsymbol{\Sigma}_{q(g)}) M^T + \mu_{q\left(1/\sigma_X^2\right)} \psi^T \psi + \Lambda_q^{-1} \right\}^{-1}$$

$$\boldsymbol{\mu}_{q(C)}^T \leftarrow \boldsymbol{\Sigma}_{q(C)} \left( \mu_{q\left(1/\sigma_Y^2\right)} Y \boldsymbol{\mu}_{q(g)}^T M^T - \mu_{q\left(1/\sigma_Y^2\right)} \boldsymbol{\mu}_{q(\beta)} z \boldsymbol{\mu}_{q(g)}^T M^T \mu_{q\left(1/\sigma_Y^2\right)} Z \boldsymbol{\mu}_{q(b)} \boldsymbol{\mu}_{q(g)}^T M^T + \mu_{q\left(1/\sigma_X^2\right)} W \psi \right)^T$$

$$\boldsymbol{\Sigma}_{q(g)} \leftarrow \left\{ \mu_{q\left(1/\sigma_Y^2\right)} M^T (\boldsymbol{\mu}_{q(C)}^T \boldsymbol{\mu}_{q(C)} + n \boldsymbol{\Sigma}_{q(C)}) M + \mu_{q\left(1/\sigma_g^2\right)} D^{-1} \right\}^{-1}$$

$$\boldsymbol{\mu}_{q(g)} \leftarrow \boldsymbol{\Sigma}_{q(g)} \left\{ \mu_{q\left(1/\sigma_Y^2\right)} \left( Y^T - \boldsymbol{\mu}_{q(\beta)}^T z^T - \boldsymbol{\mu}_{q(b)}^T Z^T \right) (\boldsymbol{\mu}_{q(C)} M) \right\}^T$$

$$B_{q(\lambda_k)} \leftarrow B_\lambda + \frac{1}{2} \left( (\boldsymbol{\mu}_{q(C)}^k)^T (\boldsymbol{\mu}_{q(C)}^k) + n (\boldsymbol{\Sigma}_{q(C)})_{kk} \right), \ 1 \le j \le K_x$$

$$B_{q(\sigma_X^2)} \leftarrow B_X + \frac{1}{2} \left\{ \sum_{i=1}^I \sum_{j=1}^J \left| \left| (W_{ij} - \mu_{q(c),ij} \psi^T)^T \right| \right|^2 + (nJ) \mathrm{tr}(\psi^T \psi \boldsymbol{\Sigma}_{q(C)}) \right\}$$

$$B_{q(\sigma_b^2)} \leftarrow B_b + \frac{1}{2} \left( \boldsymbol{\mu}_{q(b)}^T \boldsymbol{\mu}_{q(b)} + \mathrm{tr}(\boldsymbol{\Sigma}_{q(b)}) \right)$$

$$B_{q(\sigma_g^2)} \leftarrow B_g + \frac{1}{2} \left( \boldsymbol{\mu}_{q(g)}^T D^{-1} \boldsymbol{\mu}_{q(g)} + \mathrm{tr}(D^{-1} \boldsymbol{\Sigma}_{q(g)}) \right)$$

$$B_{q(\sigma_Y^2)} \leftarrow B_Y + \frac{1}{2} \Big[ \left| \left| Y - z \boldsymbol{\mu}_{q(\beta)} - Z b - \boldsymbol{\mu}_{q(C)} M \boldsymbol{\mu}_{q(g)} \right| \right|^2 + \mathrm{tr}(z^T z \boldsymbol{\Sigma}_{q(\beta)})$$

$$- (\boldsymbol{\mu}_{q(g)} M \boldsymbol{\mu}_{q(C)})^T (\boldsymbol{\mu}_{q(g)} M \boldsymbol{\mu}_{q(C)})$$

$$+ \boldsymbol{\mu}_{q(g)}^T M^T (\boldsymbol{\mu}_{q(C)}^T \boldsymbol{\mu}_{q(C)} + (nJ) \boldsymbol{\Sigma}_{q(C)}) M \boldsymbol{\mu}_{q(g)}$$

$$+ \mathrm{tr} \left\{ M^T \left( \boldsymbol{\mu}_{q(C)}^T \boldsymbol{\mu}_{q(C)} + (nJ) \boldsymbol{\Sigma}_{q(C)} \right) M \boldsymbol{\Sigma}_{q(g)} \right\}$$

$$+ \mathrm{tr} \left\{ Z^T Z \boldsymbol{\Sigma}_{q(b)} \right\} \Big]$$

until the increase in $\underline{p}(\boldsymbol{Y}, \boldsymbol{W}; q)$ is negligible.