Assessing Systematic Effects of Stroke on Motor Control using Hierarchical Function-on-Scalar Regression

Jeff Goldsmith[1,*] and Tomoko Kitago[2]

[1]Department of Biostatistics, Mailman School of Public Health, Columbia University
[2]Department of Neurology, Columbia University Medical Center
[*]*jeff.goldsmith@columbia.edu*

September 12, 2013

Abstract

This work is concerned with understanding common population-level effects of stroke on motor control while accounting for possible subject-level idiosyncratic effects. Upper extremity motor control for each subject is assessed through repeated planar reaching motions from a central point to eight pre-specified targets arranged on a circle. We observe the kinematic data for hand position as a bivariate function of time for each reach. Our goal is to estimate the bivariate function-on-scalar regression with subject-level random functional effects while accounting for potential correlation in residual curves; covariates of interest are severity of motor impairment and target number. We express fixed effects and random effects using penalized splines, and allow for residual correlation using a Wishart prior distribution. Parameters are jointly estimated in a Bayesian framework, and we implement a computationally efficient approximation algorithm using variational Bayes. Simulations indicate that the proposed method yields accurate estimation and inference, and application results suggest that the effect of stroke on motor control has a systematic component observed across subjects.

Key Words: Penalized Splines, Bivariate Data, Bayesian Regression, Gibbs Sampler, Variational Bayes.

1 Introduction

Stroke is the leading cause of long-term disability in the United States, with an incidence of over 795,000 events each year (Go et al., 2013) – a rate that is expected to grow to over one million by 2025 (Broderick, 2004). Disability induced by stroke is manifested in many activities including motor control, speech, and

cognitive performance. Between 30-66% of stroke patients have clinically apparent motor deficits involving the upper extremity at 6 months (Kwakkel et al., 2003), but it remains unclear to what extent these deficits are idiosyncratic (or subject-specific) rather than common across affected individuals. In this paper we explore the relationship between a clinical measure of arm motor impairment, the Fugl-Meyer Upper Extremity Motor Assessment (FM-UE, Fugl-Meyer et al. (1974)), and performance on a reaching task designed to test a fundamental level of motor control (Kitago et al., 2013) in a population of patients suffering from with persistent arm weakness following stroke. Elderly healthy controls are included as a reference group. In the reaching task, observations at the subject level are repeated two-dimensional motion trajectories parameterized by time. Our analytical approach for these mulitlevel bivariate functional data is to jointly model main effects for motor impairment and target direction, subject-level random effects, and residual correlation in a Bayesian function-on-scalar regression.

1.1 Two-dimensional Planar Reaching Data

We now describe the scientific setting and data structure in more detail. Our study population consists of patients who had a first time ischemic or hemorhagic stroke six or more months in the past, and have residual paresis of the affected arm (FM-UE less than the maximum score of 66). Exclusion criteria include multiple stroke events, hemorrhagic stroke, traumatic brain injury, major non-stroke medical illness that alters brain function, orthopedic or neurological condition that interferes with arm function, or inability to give informed consent. Selected patients exhibit moderate to severe hemiparesis, or weakness and motor control deficit, in the affected arm. Healthy controls with an age distribution similar to that in stroke patients are included as a reference group.

As a measurement of upper extremity motor control, subjects make repeated center-out arm reaching movements to 8 targets in the following experimental design. After subjects are seated to align the shoulder, elbow, and hand in the horizontal plane, the trunk is comfortably secured and the wrist and hand are immobilized with a splint. The forearm is supported on an air-sled system to reduce the effects of friction and gravity, diminishing the impact of strength deficits on motions and isolating motor control. Subjects make reaching movements from a central starting point to eight targets arranged equidistantly on a circle of radius 8cm around the starting point. The center-out reaching movements required can be

performed by all but the most severely impaired subjects. Before data acquisition, a short introductory period familiarizes subjects with the experiment configuration. These data were collected as part of baseline assessments for two longitudinal stroke intervention studies (Kitago et al., 2013; Huang et al., 2012) and a study of cerebral blood flow after stroke (unpublished data), which were approved by the Columbia University Medical Center Institutional Review Board.

Kinematic data is recorded for each motion made by each subject. That is, we observe the X and Y coordinate of the hand position for as a function of time giving bivariate functional observations $(P_{ij}^X(t), P_{ij}^Y(t))$ for subjects i and motions j. Our dataset consists of 24 healthy controls, 25 mildly affected stroke patients (FM-UE 44 and above), and 8 severely affected stroke patients (FM-UE <44); all participants make 22 reaching motions with both their dominant and nondominant hands to each of the eight targets, giving 352 motions for each subject and roughly 20,000 overall bivariate functional observations (note that due to technical errors in recording, some motions are removed from the dataset). Although the data are inherently functional in nature, existing analyses have primarily focused on scalar summaries of observed trajectories including the deviation of endpoint from target, peak velocity, and curvature (Levin, 1996; Lang et al., 2006; Coderre et al., 2010).

Figure 1 shows the observed data for one healthy control in the top row and one severely affected stroke patient in the bottom row. In the left column are complete trajectories, illustrating the full path of each reaching motion colored according to target. There are clear differences comparing the healthy control and stroke patient, particularly in the average motion made to each target: for instance, for the target at 0° the stroke patient exhibits both overextension and increased curvature with respect to the control subject. The middle and right columns show the constituent functions $P_{ij}^X(t)$ and $P_{ij}^Y(t)$ that make up the kinematic data for each trajectory – we will model these using a combination of population-level fixed effects, subject-level random effects, and curve-level FPCA effects. The stroke patient has unilateral tissue damage due to blockage of the right middle cerebral artery, which results in disrupted motor skill in the dominant arm.

Our goal in this analysis is to explore the extent to which the effects of stroke are consistent across subjects through a regression analysis using a combination of subject-level scalar covariates, such as a categorical ranking or a continuous measure of impairment severity, as predictors of interest. Evidence for

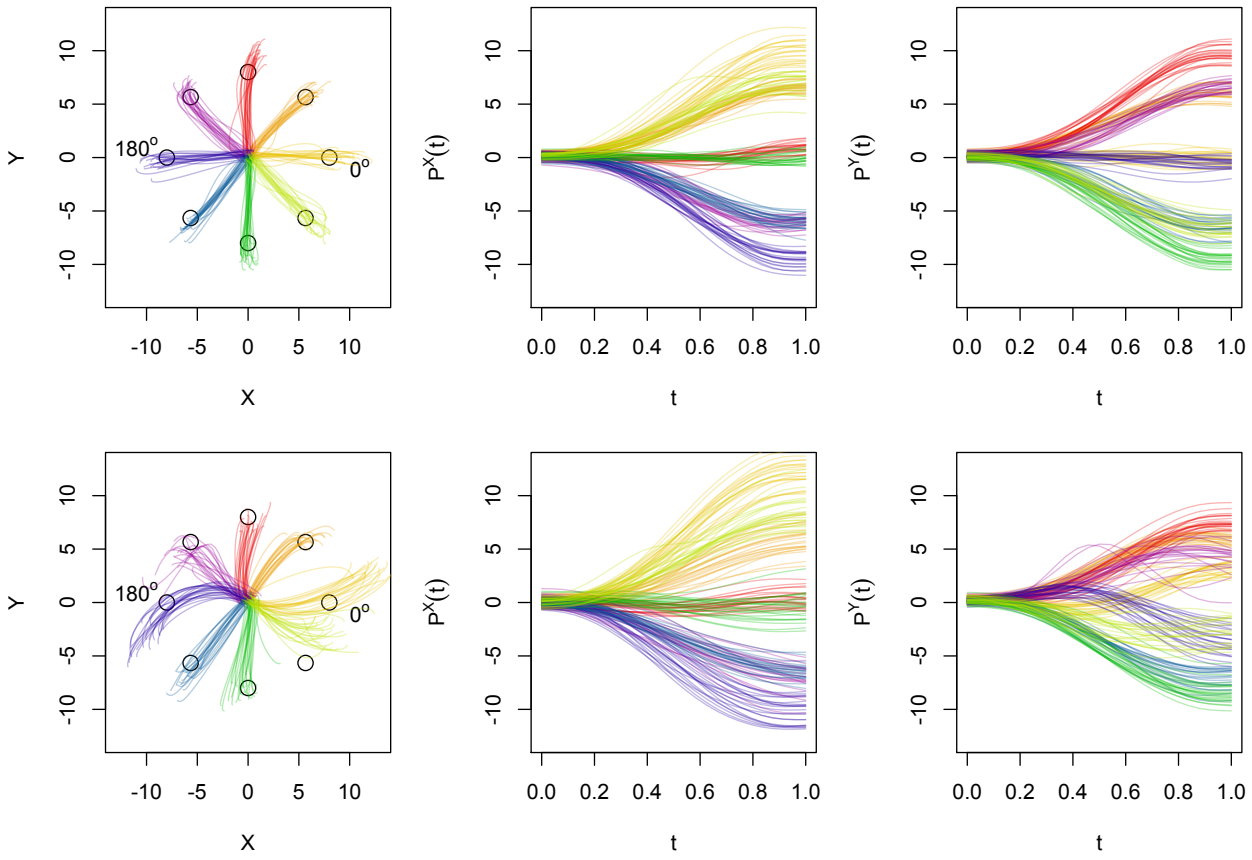

Figure 1: Observed data for two subjects; the top row shows the dominant hand of a healthy control, and the bottom row shows the affected dominant hand of a severe stroke patient. The left column shows observed kinematic data for all reaches observed in the dominant hand. The middle and right columns show the X- and Y- position separately for all reaches.

systematic effects of stroke on motor control would indicate that the induced control abnormality is not entirely subject specific, but rather that disrupting the motor cortex or its descending pathways leads to predictable deficits in upper extremity motor control. The data structure necessitates correctly accounting for subject-level effects through the inclusion of random functional intercepts, and accurate inference depends on incorporating residual correlation. Throughout, our outcome is the bivariate kinematic function for hand position over time.

1.2 Statistical Methods

We observe data $[P_{ij}^X(t), P_{ij}^Y(t), \boldsymbol{w}_i]$ for subjects $i = 1, \ldots, I$ and visits $j = 1, \ldots J_i$ for a total number of observations $n = \sum_i J_i$. In our application $P_{ij}^X(t), P_{ij}^Y(t)$ are the X and Y position curves indexed by time

$t \in [0,1]$ and $\boldsymbol{w}_{ij}$ is a length p vector of scalar covariates. We propose the model

$$\begin{cases} P_{ij}^X(t) = \beta_0^X(t) + \sum_{k=1}^p \boldsymbol{w}_{i,k}\beta_k^X(t) + b_i^X(t) + \epsilon_{ij}^X(t) \\ P_{ij}^Y(t) = \beta_0^Y(t) + \sum_{k=1}^p \boldsymbol{w}_{i,k}\beta_k^Y(t) + b_i^Y(t) + \epsilon_{ij}^Y(t) \end{cases} \tag{1}$$

where $\beta_k^X(t), \beta_k^Y(t)$ are fixed effects associated with scalar covariates, $b_i^X(t), b_i^Y(t)$ are subject-specific random effects, and $\epsilon_{ij}^X(t), \epsilon_{ij}^Y(t)$ are potentially correlated residual curves. Penalized splines are used to estimate fixed and random effects; although many options are possible, we will use a cubic B-splines basis with a combined zeroth and second derivative penalty throughout. All parameters are modeled in Bayesian framework that, in particular, allows the joint modeling of the mean structure (through fixed and random effects) and residual correlation (through the errorl covariance matrix) in a single Gibbs sampler. Importantly, a variational Bayes algorithm provides a computationally efficient and accurate approximation to the full sampler.

In practice, observations are not truly functional but are observed as structured discrete vectors. For notational simplicity, we assume functional observations lie on a dense grid of length D common to all subjects, although the methods described are suitable for different observation grids that can be sparse at the subject or visit level. In our application, trajectories are observed at 120Hz but have been interpolated so that $D = 25$ for all motions. Despite being observed as vectors, objects that are functional in nature will be denoted as $f(t)$ where appropriate to emphasize the structure underlying the data.

There is a large body of existing work for the analysis of functional outcome models. We broadly consider two methodological categories, the first of which consists of approaches that seek to estimate each curve in the dataset. Brumback and Rice (1998) posed a function-on-scalar regression in which population-level coefficients and curve-level deviations are modeled using penalized splines; for computational convenience, intercepts and slopes for curve-level effects were treated as fixed effects. Guo (2002) extended this approach by formulating curve-specific deviations as random effects. Due to the difficulty in estimating all curves using penalized splines, these approaches can be computationally intensive for large dataset. Functional principal component methods for cross sectional data (Yao et al., 2005), as well as recent extensions for multilevel (Di et al., 2009), longitudinal (Greven et al., 2010), and spatially correlated

data (Staicu et al., 2010), model curve-specific deviations from a population mean using low-dimensional basis functions estimated from the empirical covariance matrix. These methods do not focus on the flexible estimation of the population mean surface; moreover in assessing uncertainty these methods implicitly condition on estimated decomposition objects, which can lead to the understatement of total variability (Goldsmith et al., 2013)

Alternatively, one can view individual curves as errors around the population- or subject-level mean of interest. This approach is described in (Ramsay and Silverman, 2005, §13.4), in which fixed effects at the population level are estimated using penalized splines but individual curves are not directly modeled. Reiss and Huang (2010) builds on this approach by taking advantage of the inherent connection between penalized splines and ridge regression to develop fast methods for leave-one-out cross validation to select tuning parameters. Scheipl et al. (2013) propose a very flexible class of functional outcome models, allowing cross sectional or longitudinal data as well as scalar or functional predictors and estimating effects in a mixed model framework; a robust software implementation of this method is provided in the `refund R` package (Crainiceanu et al., 2012). A drawback of these approaches is the assumption that error curves consist only of measurement error despite clear correlation in the functional domain. One alternative, proposed by Reiss and Huang (2010), is an iterative procedure to estimate the mean structure and then, using this mean, the residual covariance matrix followed by a re-estimation of the mean using generalized least squares. Doing so necessarily increases the computational burden and does not allow joint estimation of the mean and covariance; additionally, coverage properties of this approach have not been presented. Morris and Carroll (2006) develop a Bayesian wavelet-based estimation method that allows for correlated errors, but the computation burden of the MCMC procedure is prohibitive for data exploration and model building, especially for large datasets. Moreover, wavelet bases are not be suitable for many applications.

This manuscript presents several methodological advancements. We develop a Bayesian framework for penalized spline function-on-scalar regression, allowing the joint modeling of population-level fixed effects, subject-level random effects and residual covariance. Dramatic computational improvements compared to the fully Bayesian analysis and, surprisingly, frequentist mixed model approaches are obtained through a variational Bayes approximation that is fast and accurate. This algorithm enables model selection and comparison, which for large datasets is infeasible with competing approaches. Novelly in function-on-scalar

regression, we consider multilevel bivariate functional data as the outcome of interest. Finally, the size and structure of the motivating dataset – which consists of nearly 20,000 trajectories, nested within subjects and depending on target, impairment severity and affected hand as covariates – is unique in the functional data analysis literature.

The remainder of the paper is organized as follows. We discuss the model formulation and the variational Bayes approximation in Section 2. In Section 3 we conduct simulations designed to mimic the motivating data. Section 4 presents the analysis of the complete dataset. We close with a discussion in Section 5. An appendix presents the complete Gibbs sampler and variational Bayes algorithm. `R` implementations of all proposed methods and complete simulation code are publicly available.

2 Methods

We begin by focusing on a simplification of model (1) for univariate functional data in order to develop the requisite techniques for fixed effects, random effects, and residual covariance in a Bayesian setting. Once these are established, we consider the bivariate model in Section 2.4.

2.1 Full Model

For now, assume data is $[Y_{ij}(t), \boldsymbol{w}_i]$ for subjects $i = 1, \ldots, I$ and visits $j = 1, \ldots J_i$, giving a total of $n = \sum_i J_i$ observations. Univariate functional outcomes $Y_{ij}(t)$ are observed on a regular grid of length D for all subjects and visits. We pose the outcome model

$$Y_{ij} = \boldsymbol{w}_i\boldsymbol{\beta} + \boldsymbol{z}_{ij}\boldsymbol{b} + \epsilon_{ij} \tag{2}$$

$$\epsilon_{ij} \sim \mathrm{N}[0, \Sigma] \tag{3}$$

where $\boldsymbol{Y}_{ij}$ is the $1 \times D$ observed functional outcome; $\boldsymbol{w}_i$ and $\boldsymbol{z}_{ij}$ are fixed and random effect design vectors of size $1 \times p$ and $1 \times I$ respectively; $\boldsymbol{\beta}$ and $\boldsymbol{b}$ are fixed and random effect coefficient matrices of size $p \times D$ and $I \times D$, respectively; and $\boldsymbol{\epsilon}$ is a $1 \times D$ matrix of residual curves distributed $\mathrm{N}[0, \Sigma]$, assumed *iid* across subjects and visits.

We express the functional effects in the rows of $\boldsymbol{\beta}$ and $\boldsymbol{b}$ using a spline expansion. Let $\boldsymbol{\Theta}$ denote

a $D \times K_\theta$ cubic B-spline evaluation matrix with K_θ basis functions. Further let $\boldsymbol{B}_W$ and $\boldsymbol{B}_Z$ denote the matrices whose columns are basis coefficients for $\boldsymbol{\beta}$ and $\boldsymbol{b}$ respectively, so that $\boldsymbol{\beta} = [\boldsymbol{\Theta}\boldsymbol{B}_W]^T$ and $\boldsymbol{b} = [\boldsymbol{\Theta}\boldsymbol{B}_Z]^T$. Penalization is a commonly-used technique to avoid overfitting and induce smoothness in functional effects. In a frequentist approach quadratic penalization is accomplished through the addition of terms $\lambda_{W_k}\boldsymbol{B}_{W_k}^T P \boldsymbol{B}_{W_k}$ and $\lambda_Z \boldsymbol{B}_{Z_i}^T P \boldsymbol{B}_{Z_i}$ for the columns of $\boldsymbol{B}_W$ and $\boldsymbol{B}_Z$ to the least squares criterion, where P is a known penalty matrix and the $\{\lambda_{W_k}\}_{k=1}^p$ and λ_Z are tuning parameters that controls the degree of smoothness. Distinct tuning parameters $\{\lambda_{W_k}\}_{k=1}^p$ are estimated for the columns of $\boldsymbol{B}_W$, while a single tuning parameter λ_Z controls smoothness of subject random effects in the columns of $\boldsymbol{Z}$. The choice of penalty matrix P is discussed in Section 2.3.

The connection between penalized spline regression and mixed models is well known; see Ruppert et al. (2003, Ch. 4.9) for a detailed treatment. Very briefly, assuming that $\boldsymbol{B}_{W_1} \sim \mathrm{N}\left[0, \sigma_{W_1}^2 P^{-1}\right]$ and $\boldsymbol{\epsilon} \sim \mathrm{N}\left[0, \sigma_\epsilon^2 \boldsymbol{I}_D\right]$, estimating model parameters in a mixed model as best linear unbiased predictors is equivalent to penalized least squares with $\lambda_{W_1} = \frac{\sigma_\epsilon^2}{\sigma_{W_1}^2}$; variance components are typically estimated using restricted maximum likelihood. Frequentist mixed models and Bayesian approaches are closely related, and it is natural to view penalized spline smoothing from a Bayesian perspective (Crainiceanu et al., 2005). To maintain the connection to penalized spline regression, we use mean-zero Normal priors for the B-spline regression coefficients in the columns of $\boldsymbol{B}_W$ and $\boldsymbol{B}_Z$. Meanwhile, inverse-gamma priors are used for the variance components $\left\{\sigma_{W_k}^2\right\}_{k=1}^p, \sigma_Z^2$. Our model specification is completed using an Inverse Wishart prior for the residual covariance matrix Σ.

For estimation, we cast model (2) in a hierarchical framework. Let $\boldsymbol{Y}$ be an $n \times D$ matrix of row-stacked functional outcomes, $\boldsymbol{Z}$ be the random effects design matrix, $\boldsymbol{W}$ be the fixed effects matrix constructed by row-stacking the $\boldsymbol{w}_i$, and $\otimes$ represent the Kronecker product operator. Our full model is

$$
\begin{aligned}
\boldsymbol{Y} &= \boldsymbol{Z}\boldsymbol{B}_Z^T\boldsymbol{\Theta}^T + \boldsymbol{\epsilon} \\
\boldsymbol{\epsilon} &\sim \mathrm{N}\left[0, \Sigma \otimes I_n\right]; \Sigma \sim \mathrm{IW}\left[\nu, \boldsymbol{\Psi}\right] \\
\boldsymbol{B}_{Z_i} &\sim \mathrm{N}\left[\boldsymbol{w}_i\boldsymbol{B}_W, \sigma_Z^2 P^{-1}\right] \text{ for } i = 1 \ldots I; \sigma_Z^2 \sim \mathrm{IG}\left[a_Z, b_Z\right] \\
\boldsymbol{B}_{W_k} &\sim \mathrm{N}\left[0, \sigma_{W_k}^2 P^{-1}\right], \sigma_{W_k}^2 \sim \mathrm{IG}\left[a_{W_k}, b_{W_k}\right] \text{ for } k = 1 \ldots p.
\end{aligned} \tag{4}
$$

Full conditionals for all model parameters are straightforward to obtain using vector notation and Kronecker products. For matrix M and vector c, let $\text{vec}(M)$ be the vector formed by concatenating the columns of M and $\text{diag}(c)$ be the matrix with elements of c on the main diagonal and zero elsewhere. Further let "rest" include both the observed data and all parameters not currently under consideration. As an example of the full conditional distributions resulting from model (4), it can be shown that

$$p\left[\text{vec}\left(\boldsymbol{B}_W\right)|\text{rest}\right] \propto \text{N}\left[\boldsymbol{\mu}_{\boldsymbol{B}_W}, \Sigma_{\boldsymbol{B}_W}\right]$$

where

$$\Sigma_{\boldsymbol{B}_W} = \left(\frac{1}{\sigma_Z^2}(\boldsymbol{W} \otimes I_{K_t})^T(I_I \otimes P)(\boldsymbol{W} \otimes I_{K_t}) + \text{diag}\left(\frac{1}{\sigma_{W_k}^2}\right) \otimes P\right)^{-1}$$

and

$$\boldsymbol{\mu}_{\boldsymbol{B}_W} = \Sigma_{\boldsymbol{B}_W}\left(\frac{1}{\sigma_Z^2}(\boldsymbol{W} \otimes I_{K_t})^T(I_I \otimes P)\text{vec}\left(\boldsymbol{B}_Z\right)\right).$$

Additionally, we have that

$$p\left[\sigma_{W_k}^2|\text{rest}\right] \propto \text{IG}\left[a_{W_k} + \frac{K_t}{2}, b_{W_k} + \boldsymbol{B}_{W_k}^T P \boldsymbol{B}_{W_k}\right].$$

Complete derivations of this and all other full conditional distributions are provided in Appendix A.

2.2 Variational Bayes

Variational Bayes methods are regularly used in the computer science literature, and to a more limited extent in the statistics literature, to provide approximate solutions to intractable inference problems (Jordan, 2004; Jordan et al., 1999; Titterington, 2004; Ormerod and Wand, 2012). These tools have also been used somewhat rarely in functional data analysis (Goldsmith et al., 2011; McLean et al., 2013; van der Linde, 2008). Here we review variational Bayes only as much as needed to develop an iterative algorithm for approximate Bayesian inference in penalized function-on-scalar regression; for a more detailed overview see Ormerod and Wand (2010) and Bishop (2006, Chapter 10). We emphasize that the variational Bayes approach is not intended to supplant a more complete MCMC sampler, but rather is an appealing computationally efficient approximation that is useful for model building and data exploration.

Let $\boldsymbol{y}$ and $\boldsymbol{\phi}$ represent respectively the full data and parameter collection. The goal of variational Bayes methods is to approximate the full posterior $p(\boldsymbol{\phi}|\boldsymbol{y})$ using $q(\boldsymbol{\phi})$, where q is restricted to a class of functions that are more tractable than the full posterior distribution. From the restricted class of functions, we wish to choose the element q^* that minimizes the Kullback-Leibler distance from $p(\boldsymbol{\phi}|\boldsymbol{y})$. Divergence between $p(\boldsymbol{\phi}|\boldsymbol{y})$ and $q(\boldsymbol{\phi})$ is measured using $L_q = \int q(\boldsymbol{\phi}) \log \frac{p(\boldsymbol{y},\boldsymbol{\phi})}{q(\boldsymbol{\phi})} d\boldsymbol{\phi}$, the q-specific lower bound on the marginal log-likelihood $\log p(\boldsymbol{y})$; maximizing L_q across the class of candidate functions gives the best possible approximation to the full posterior distribution. To make the approximation tractable, the candidate functions $q(\boldsymbol{\phi})$ are products over a partition of $\boldsymbol{\phi}$, so that $q(\boldsymbol{\phi}) = \prod_{l=1}^{L} q_l(\phi_l)$, and each q_l is a parametric density function. It can be shown that the optimal q_l^* densities are given by

$$q_l^*(\phi_l) \quad \propto \quad \exp\left[E_{\phi_{-l}} \log p(\boldsymbol{y}, \phi)\right] \propto \exp\left[E_{\phi_{-l}} \log p(\phi_l|\text{rest})\right]$$

where, again, rest $\equiv \{\boldsymbol{y}, \phi_1, \ldots, \phi_{l-1}, \phi_{l+1}, \ldots, \phi_L\}$ is the collection of all remaining parameters and the observed data. In practice, one sets initial values for each of the ϕ_l and updates the respective optimal densities iteratively, similarly to a Gibbs sampler, while monitoring the q-specific lower bound L_q for convergence.

For the function-on-scalar regression model shown in Equation (4), we use the partitioning

$$\begin{aligned} & q\left(\boldsymbol{B}_Z, \boldsymbol{B}_W, \sigma^2_{W_1}, \ldots, \sigma^2_{W_p}, \sigma^2_Z, \Sigma\right) \\ &= \quad q(\boldsymbol{B}_Z) q(\boldsymbol{B}_W) \left(\prod_{k=1}^{p} q(\sigma^2_{W_k})\right) q(\sigma^2_Z) q(\Sigma) \end{aligned}$$

where the functions q are distinguished by their argument rather than by subscript l. Using this factorization, it can be shown that the optimal density $q^*(\text{vec}\,(\boldsymbol{B}_W))$ for is $\text{N}\left[\mu_{q(\boldsymbol{B}_W)}, \Sigma_{q(\boldsymbol{B}_W)}\right]$, where

$$\Sigma_{q(\boldsymbol{B}_W)} = \left(\mu_{q(1/\sigma^2_Z)} (\boldsymbol{W} \otimes I_{K_t})^T (I_I \otimes P)(\boldsymbol{W} \otimes I_{K_t}) + \text{diag}\left(\mu_{q(1/\sigma^2_{W_k})}\right) \otimes P\right)^{-1}$$

and

$$\boldsymbol{\mu}_{q(\boldsymbol{B}_W)} = \Sigma_{q(\boldsymbol{B}_W)} \left(\mu_{q(1/\sigma^2_Z)} (\boldsymbol{W} \otimes I_{K_t})^T (I_I \otimes P) \text{vec}\left(\mu_{q(\boldsymbol{B}_Z)}\right)\right).$$

In the above, the notation $\mu_{q(\phi)}$ and $\Sigma_{q(\phi)}$ indicate the mean and variance of the density $q(\phi)$. Thus,

the optimal density $q^*(\text{vec}(\boldsymbol{B}_W))$ is Normal with mean and variance completely determined by the data and the parameters of the remaining densities. Similar expressions are obtained for all model parameters. Together, these forms suggest an iterative algorithm in which each density is updated in turn using the parameters from the remaining densities; convergence of this algorithm is monitored through L_q, the q-specific lower bound of the marginal log-likelihood.The iterative algorithm and the form for L_q are provided in Appendix B.

2.3 Choice of penalty matrix, hyperparameters, and initial values

In both the Gibbs sampler presented in Section 2.1 and in density updates for the variational Bayes algorithm described in 2.2, it is not necessary that the penalty matrix P be of full rank: although this introduces improper priors for the functional effects, the posteriors are proper. However the lower bound L_q, used to monitor convergence of the variational Bayes algorithm, contains a term of the form $log(|P^{-1}|)$, thus requiring P to be full rank. For this reason, we propose to use $P = \alpha P_0 + (1-\alpha)P_2$, where P_0 and P_2 are the matrices corresponding to zeroth and second derivative penalties. The P_2 penalty matrix enforces smoothness in the estimated function, but is non-invertable. The P_0 penalty matrix induces general shrinkage and is full rank. Selecting $0 < \alpha \leq 1$ balances smoothness and shrinkage, and results in a full rank penalty matrix. Unreported simulations and data analyses indicate low sensitivity to the choice of α, but we recommend a relatively small value ($\alpha = 0.1$) in keeping with the tendency to enforce smoothness rather than shrinkage.

We use the following procedure to choose hyperparameters and initial values. First, we estimate $\boldsymbol{B}_Z$ using ordinary least squares from the regression $\text{E}[\boldsymbol{Y}] = \boldsymbol{Z}\boldsymbol{B}_Z^T\boldsymbol{\Theta}^T$ to obtain $\boldsymbol{B}_Z^{OLS}$. Second, we estimate $\boldsymbol{B}_W$ using weighted least squares from the regression $\text{E}\left[\left(\boldsymbol{B}_Z^{OLS}\right)^T\right] = \boldsymbol{W}\boldsymbol{B}_W^T$ with weight matrix P^{-1} to obtain $\boldsymbol{B}_W^{WLS}$. Initial values for $\boldsymbol{B}_Z$ and $\boldsymbol{B}_W$ are given by $\boldsymbol{B}_Z^{OLS}$ and $\boldsymbol{B}_W^{WLS}$, respectively. Our first regression, an ordinary least squares, assumes the residual covariance is $\Sigma = \sigma_\epsilon^2 I_D$. We estimate σ_ϵ^2 in this framework, and use $\nu = \sum_i J_i$, $\boldsymbol{\Psi} = \sum_i J_i \sigma_\epsilon^2 I_D$ as hyperparameters in the prior for Σ. Effectively, our prior assumes uncorrelated homoscedastic errors and the posterior balances this with the observed residual correlation. Because the observed residual correlation matrix is low rank in our application, adding a diagonal matrix is needed for a full-rank posterior. Similarly, we

choose $a_z = \frac{I*K_t}{2}$, $b_z = \text{tr}\left[\left(\left(\boldsymbol{B}_Z^{OLS}\right)^T - \boldsymbol{W}\left(\boldsymbol{B}_W^{WLS}\right)^T\right)^T P\left(\left(\boldsymbol{B}_Z^{OLS}\right)^T - \boldsymbol{W}\left(\boldsymbol{B}_W^{WLS}\right)^T\right)\right]$, $a_{W_k} = \frac{K_t}{2}$, and $b_{W_k} = \left(\boldsymbol{B}_{W_k}^{WLS}\right)^T P\left(\boldsymbol{B}_{W_k}^{WLS}\right)$ according to the assumed form for the second regression and the prior for $\boldsymbol{B}_W$. These hyperparameters establish maximum levels of penalization and prevent oversmoothing.

2.4 Bivariate data

In the preceding we have focused on a univariate outcome for clarity of exposition while introducing methods. In this section we describe the bivariate outcome model. Only straightforward modifications to the Gibbs sampler and variational Bayes updates given in Sections 2.1 and 2.2, respectively, are needed for this setting.

Let $\boldsymbol{Y} = [\boldsymbol{Y}_1 \boldsymbol{Y}_2]$ be the concatenation of two outcome matrices $\boldsymbol{Y}_1$ and $\boldsymbol{Y}_2$ (in our example, we concatenate the X and Y position curves so that $\boldsymbol{Y} = [P^X P^Y]$). The full model for bivariate outcome data uses Kronecker products of the B-spline basis to expand $\boldsymbol{Y}_1$ and $\boldsymbol{Y}_2$, induces smoothness for fixed and random effects in $\boldsymbol{Y}_1$ and $\boldsymbol{Y}_2$ through separate variance components, and allows residual correlation within and across $\boldsymbol{Y}_1$ and $\boldsymbol{Y}_2$:

$$
\begin{aligned}
\boldsymbol{Y} &= \boldsymbol{Z}\boldsymbol{B}_Z^T\left(I_2 \otimes \boldsymbol{\Theta}^T\right) + \boldsymbol{\epsilon} \\
\boldsymbol{\epsilon} &\sim \text{N}\left[0, \Sigma \otimes I_n\right]; \Sigma \sim \text{IW}\left[\nu, \boldsymbol{\Psi}\right] \\
\boldsymbol{B}_{Z_i} &\sim \text{N}\left[\boldsymbol{w}_i \boldsymbol{B}_W, \text{diag}\left(\sigma^2_{1,Z,}, \sigma^2_{2,Z}\right) \otimes P^{-1}\right] \text{ for } i = 1 \ldots I \\
&\quad \sigma^2_{m,Z} \sim \text{IG}\left[a_{m,Z}, b_{m,Z}\right] \text{ for } m = 1, 2 \\
\boldsymbol{B}_{W_k} &\sim \text{N}\left[0, \text{diag}\left(\sigma^2_{1,W_k}, \sigma^2_{2,W_k}\right) \otimes P^{-1}\right] \\
&\quad \sigma^2_{m,W_k} \sim \text{IG}\left[a_{m,W_k}, b_{,mW_k}\right] \text{ for } m = 1, 2 \text{ and } k = 1 \ldots p.
\end{aligned} \tag{5}
$$

Note that the residual covariance matrix Σ is now of size $(2D) \times (2D)$, and that the columns of $\boldsymbol{B}_Z$, $\boldsymbol{B}_W$ include basis coefficients for both $\boldsymbol{Y}_1$ and $\boldsymbol{Y}_2$. This model extends the univariate outcome model (4), but the Gibbs sampler and variational Bayes approximations can be directly modified for bivariate data. Similarly, the method for setting initial values and choosing hyperparameters given in Section 2.3 can be adapted to model (5).

3 Simulations

We demonstrate the performance of our method using a simulation in which generated data mimic the motivating application. Data are generated from a univariate outcome model in which the categorical predictor has three levels:

$$Y_{ij}(t) = \sum_{p=0}^{2} I(\boldsymbol{w}_i = p)\beta_p^{obs}(t) + b_i(t) + \epsilon_{ij}(t) \tag{6}$$

The number of subjects I is set to (a) 60, (b) 120, or (c) 180. In all cases, we fix the number of observations per subject to be $J_i = 5$, set the size of the grid on which curves are observed to be $D = 25$, and use an equal number of subjects for each of the three levels of $\boldsymbol{w}$. Subject-level random effects $b_i(t)$ are distributed $\mathrm{N}\left[0, \Sigma^{b,obs}\right]$ and curve-level residuals $\epsilon_{ij}(t)$ are distributed $\mathrm{N}\left[0, \Sigma^{\epsilon,obs}\right]$. Coefficient functions $\beta_p^{obs}(t)$, random effect covariance $\Sigma^{b,obs}$, and residual covariance $\Sigma^{\epsilon,obs}$ are empirical estimates derived from the observed $P_{ij}^{Y}(t)$ curves using trajectories to the target at 180°. We use three levels of disease status (control, mild stroke, severe stroke) as the categorical predictor $\boldsymbol{w}$, and use only the dominant arm. From the observed data, we estimate fixed and random effects using the approach taken to select initial values described in Section 2.3; covariance matrices $\Sigma^{b,obs}$ and $\Sigma^{\epsilon,obs}$ are empirical estimates of observed random effects and residuals. To illustrate the simulation design, Figure 2 shows the coefficient functions in the left panel. The middle panel of Figure 2 shows a complete simulated dataset with $I = 60$, and highlights data for three subjects.

For each sample size, we generate 100 datasets according to model (6). Parameters are estimated using the Gibbs sampler and variational Bayes algorithm described in Sections 2.1 and 2.2, respectively, with hyperparameters and initial values chosen as in Section 2.3. For the Gibbs sampler, we used chains of length 5000 and discarded the first 1000 as burn-in. To provide a frame of reference for our methods, we compare to the `pffr()` function in the `refund R` package. This implements the penalized function-on-function regression model (of which ours is a special case), and estimates parameters in a frequentist mixed model framework (Scheipl et al., 2013); to the best of our knowledge, `pffr()` represents the current state-of-the-art in function-on-scalar regression with subject-level random effects.

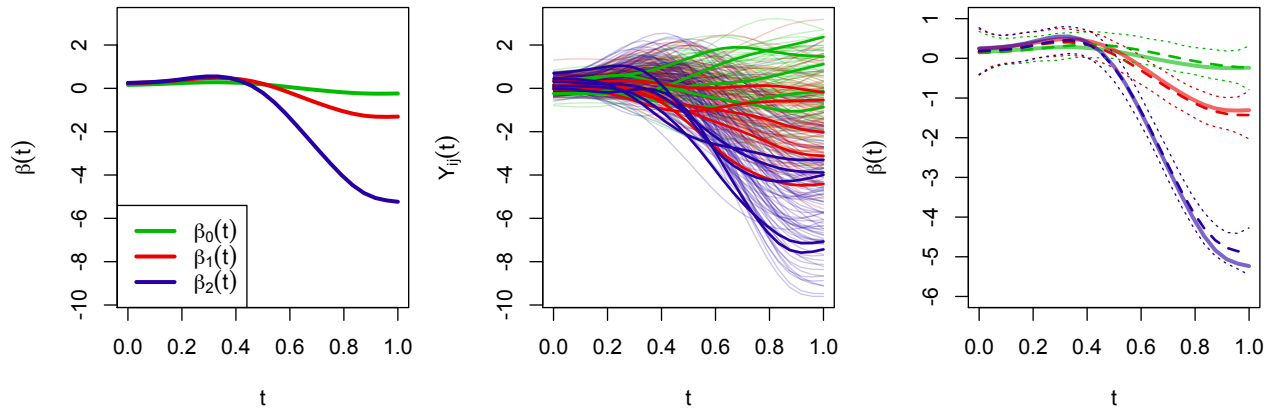

Figure 2: The left panel shows the three coefficient functions used to simulate data. The middle panel shows a complete simulated data set, with three subjects (one from each group) highlighted. The right panel shows the true coefficient functions, as well as their estimates and credible intervals derived from the dataset shown in the middle panel.

The left panels of Figure 3 show the integrated mean squared error IMSE $= \int \left(\hat{\beta}(t) - \beta(t)\right) dt$ for each coefficient function, estimation method, and sample size. IMSEs are indistinguishable for the Gibbs sampler and variational Bayes approaches, indicating that for posterior means the variational Bayes approximation is reasonable. As expected, IMSEs decrease as sample size increases. Both approaches are comparable to or somewhat outperform the frequentist mixed model approach. An abberration in the mixed model's performance is seen for $I = 120$, most likely due to incorporating population-level fixed effects into subject-level random effects. The right panel of Figure 3 shows the computation time for each sample size and approach. Not surprisingly, the variational Bayes algorithm is substantially faster than the complete Gibbs sampler. However, there are also meaningful improvements in computation time comparing the variational Bayes algorithm to the mixed model: for $i = 180$, the median computation time for the variational Bayes approach was roughly 15 seconds, while the median computation time for the mixed model was roughly 45 minutes. This discrepancy in computation time may be due to the fact that both Bayesian approaches are tailored to the model at hand, while `pffr()` can be used in more general settings.

Table 1 presents the average coverage probability of 95% pointwise confidence intervals constructed using the Gibbs sampler, the variational Bayes algorithm, and the frequentist mixed moedl for each coefficient function and sample size. For both of the proposed Bayesian approaches, coverage approaches the nominal level and is often between .92 and .95. Coverage is somewhat lower for $\beta_1(t)$; this may stem from oversmoothing the coefficient function, which is relatively (but not perfectly) flat. As expected, coverage

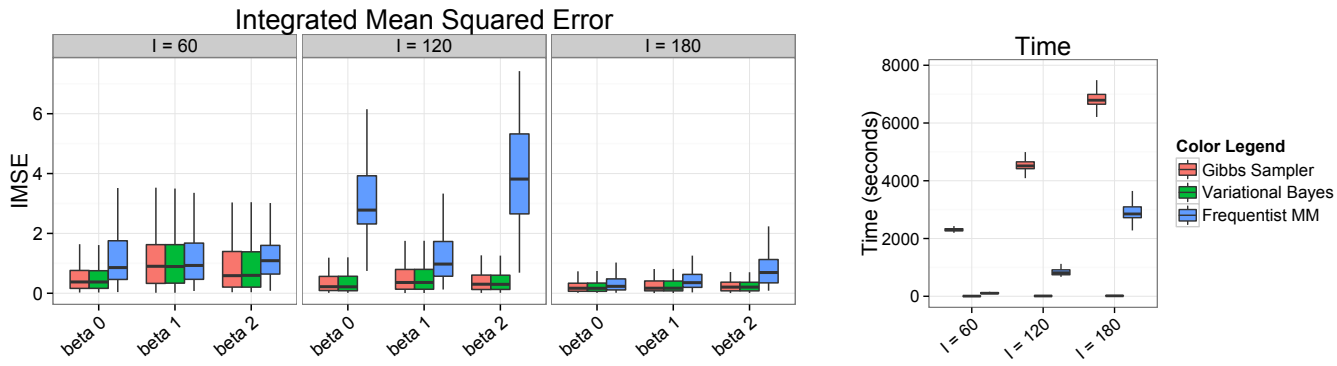

Figure 3: Simualtion results. The left panels show IMSE, defined as IMSE $= \int \left(\hat{\beta}(t) - \beta(t)\right) dt$, for each coefficient function, sample size, and estimation technique. The right panel show computation time for each sample size and estimation method.

improves as sample size increases. For the mixed model coverage is well below nominal levels for all coefficients and sample sizes; possible causes for this are the assumption that errors are independent white noise and the lack of a hierarchical framework for model estimation. Although coverage for the variational Bayes approximation is reasonable in our simulations, we do not necessarily recommend basing inference in practice on this approach due to the difficulty in verifying the assumptions regarding the factorization of the posterior distribution. Rather, we favor the variational algorithm as a fast method for model building and base inference on a full Gibbs sampler.

	Gibbs Sampler			Variational Bayes			Frequentist MM		
	$\beta_0(t)$	$\beta_1(t)$	$\beta_2(t)$	$\beta_0(t)$	$\beta_1(t)$	$\beta_2(t)$	$\beta_0(t)$	$\beta_1(t)$	$\beta_2(t)$
$I = 60$	0.96	0.89	0.92	0.94	0.86	0.90	0.40	0.55	0.47
$I = 120$	0.93	0.91	0.93	0.92	0.89	0.92	0.09	0.33	0.04
$I = 180$	0.94	0.93	0.94	0.93	0.92	0.92	0.46	0.51	0.37

Table 1: Average coverage of 95% credible intervals constructed using the Gibbs sampler, variational Bayes algorithm, and `pffr()`. Coverages are expressed as percents.

4 Application

We now apply the developed methods to the motivating data described in Section 1.1. In our dataset affected patients exhibit arm paresis, a weakness or motor control deficit affecting either the dominant or non-dominant arm, due to a unilateral stroke. Patients experienced stroke more than 6 months prior to data collection, meaning that observed motor control deficits are not due to short-term effects but rather

are chronic in nature. To quantify the severity of arm impairment we use the Upper Extremity portion of the Fugl-Meyer motor assessment, a well known and widely used clinical assessment of motor impairmen. Fugl-Meyer scores were assessed for the affected arm only, and for upper extremity testing scores range from a 0 to 66 with 66 indicating healthy function. Controls were not scored and are were assigned a Fugl-Meyer score of 66. Kinematic data collected for the left hand were reflected through the Y axis, and thus are in the same intrinsic joint space as data for the right hand (i.e., motions to the target at 180° reach across the body and involve both the shoulder and elbow).

Our focus is the effect of the severity of arm impairment on control of visually-guided reaching, where impairment is quantified using the Fugl-Meyer score. In addition to impairment severity, we control for important covariates in our regression modeling. We adjust for target direction (with 8 possible targets, treated as a categorical predictor); hand used (dominant and non-dominant); whether the arm is affected by stroke (affected and unaffected); and, potentially, interactions between these variables. Interactions of impairment severity and other covariates are possible, and are likely for target direction: the effect of stroke may be greatest to the more biomechanically difficult targets that involve coordination of multiple joints.

Our data analysis proceeds in two parts and focuses on estimation of the bivariate model (5). First, we use the variational Bayes algorithm developed in Section 2.2 to explore several possible models that include different combinations of target, hand used, affectedness, and impairment severity as well as potential interactions. In all models, subject-level random effects are estimated for each target and hand; these effects are a priori assumed to be independent. The computational efficiency of the variational Bayes algorithm is crucial at this stage, allowing the fast evaluation and comparison of models. After identifying a plausible final model, we estimate all model parameters using the complete Gibbs sampler described in Section2.1 and base inference for the effect of stroke on this analysis.

4.1 Exploratory analyses using variational Bayes

In the following, we are interested in estimated fixed effects using a variety of structures for the population mean. We estimate all models using the variational approximation described in Section 2.2. Computation time was under 20 minutes for each model we consider; the importance of fast computation in the model

building stage cannot be understated, since it allows the consideration and refinement of many candidate models.

As a reference for the more complex structures that follow, we began with a model that uses only target direction as a predictor. This model addresses directional variation only, but the eight fixed effects account for roughly 90% of observed variance in the outcome. Following this, several models that included the Fugl-Meyer score, hand used, and affectedness as predictors were considered. Finally, we compared to a model in which impairment severity was treated as a categorical variable based on a Fugl-Meyer threshold of 44 (Chae et al., 1998). This model was used to check the assumption that the effect of Fugl-Meyer is roughly linear: separate effects are estimated for each impairment severity category and no restriction is placed on their shape.

Table 2 provides the fixed effects used in each of the models we consider, the number of fixed effects for each model, and the percent of outcome variance explained by fixed effects. Percent variance explained is given relative to the model with only target direction as a predictor using

$$\text{Relative PVE } = 100 \times \left[1 - \frac{\text{Var}\left[\text{vec}\left(\boldsymbol{Y} - \hat{\boldsymbol{Y}}_{\text{Model } m}\right)\right]}{\text{Var}\left[\text{vec}\left(\boldsymbol{Y} - \hat{\boldsymbol{Y}}_{\text{Model } 0}\right)\right]}\right] \tag{7}$$

for models $m \in 1, \ldots, 8$, where $\boldsymbol{Y}$ is the matrix of observed trajectories and $\hat{\boldsymbol{Y}}_{\text{Model } m}$ is the matrix of estimated trajectories based on fixed effects in Model m.

The poor performance of Models 1 and 3 indicate the importance of interaction between target and impairment severity, due to the target-specific effect of stroke due to differing levels of biomechanical difficulty. Models 5 and 6 significantly improve over model 4 by adding interactions with hand used and affectedness, respectively. The comparability of these two models may be due to the fact that in our dataset, the affected hand was more likely to be the dominant hand. Model 7 explains the highest percent of outcome variance, and is the saturated model using Fugl-Meyer as a continuous predictor. Interestingly, Model 7 outperforms Model 8, the saturated model using a categorical impairment severity measure, using fewer fixed effects. For all models, the fixed and random effects together explain roughly 50% of outcome variance; the remaining 50% is residual variance around subject-level means. This partitioning of variance usefully quantifies the extent to which motor control is explainable by covariates, subject-specific

Model	Fixed Effects	Number of fixed effects	Relative PVE for fixed effects
0	Tar	8	Reference
1	FM+Tar	9	0.1
2	FM×Tar	16	4.7
3	FM+Tar+Hand	10	0.3
4	FM×Tar+Hand	17	4.9
5	FM×Tar×Hand	32	8.4
6	FM×Tar×Aff	32	8.9
7	FM×Tar×Hand×Aff	64	11.9
8	Cond×Tar×Hand×Aff	80	9.7

Table 2: Description and comparison of models considered. Fixed effects structure is described in the second column, where "Tar" represents the target direction (as a categorical variable); "Hand" represents hand used (dominant, non-dominant); "Aff" indicates an affected hand; "Cond" represents possible disease conditions (control, moderate stroke, severe stroke,); and "FM" is the continuous Fugl-Meyer score. In the second column, "+" indicates additive effects and "×" indicates interactions. The number of fixed effects induced by the model structure is given in the third column. The fourth column provides the percent of outcome variance explained by the model relative to a model with only target as a covariate (defined in equation (7)).

deviations, and trajectory-level variation. In Section 4.2 we discuss inference for Model 7.

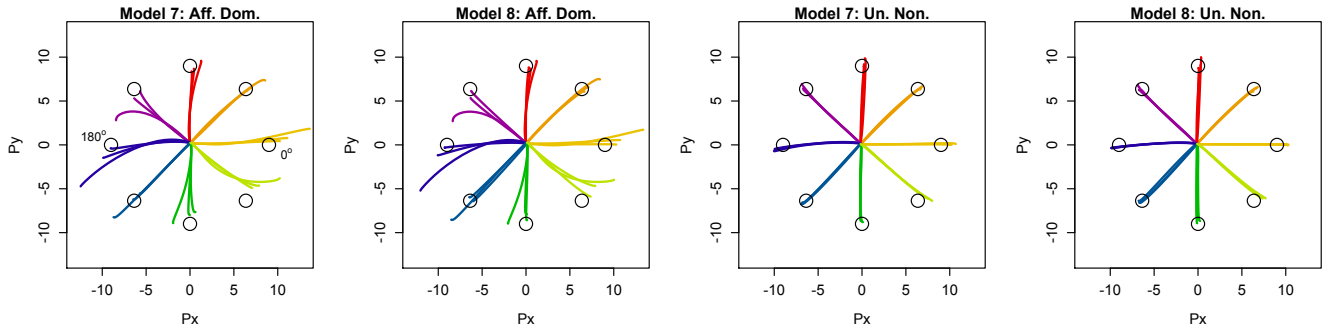

Figure 4: Comparison of fixed effects for Models 7 and 8. Estimated effects for the affected dominant and unaffected non-dominant hand for a moderately and severely affected stroke patient are shown, as well as effects for a healthy control dominant and non-dominant hand.

Figure 4 compares the estimated fixed effects from Models 7 and 8. In particular, we show the estimated effect for a healthy control, for a moderately affected stroke patient (Fugl-Meyer score 56) with stroke affecting the dominant hand, and for a severely affected stroke patient (Fugl-Meyer score 26) with stroke affecting the dominant hand. Fugl-Meyer scores were chosen as the average of observed scores in the moderate and severe groups. Estimated effects for both the dominant (affected) and non-dominant (unaffected) are given. For the dominant hand, the close agreement of estimated fixed effects indicates that the assumption of a linear effect for Fugl-Meyer is reasonable. Neither model indicates a systematic

effect of stroke in the unaffected hand, a reasonable finding given that our data set consists of patients with unilateral stroke.

Estimates of subject-level effects are shown in Figure 5 for two subjects (separately by row) overlayed on the observed trajectories. Fixed effects estimates based on Model 7 are shown in bold solid lines and subject-level estimates including random effects are shown in bold dashed lines. In the top row is a control subject's dominant hand; fixed effects and random effects estimates differ only slightly, indicating relatively little subject deviation from the population mean. In the bottom row is a severely affected (Fugl-Meyer 28) subject's affected dominant hand. Here, fixed effects are noticeably curved for several targets indicating a systematic effect of stroke. Subject-level estimates differ from the fixed effects in some cases (particularly for targets at 0° and 180°), illustrating the idiosyncratic effects of stroke in this patient. Note that data for these patients is shown in Figure 1.

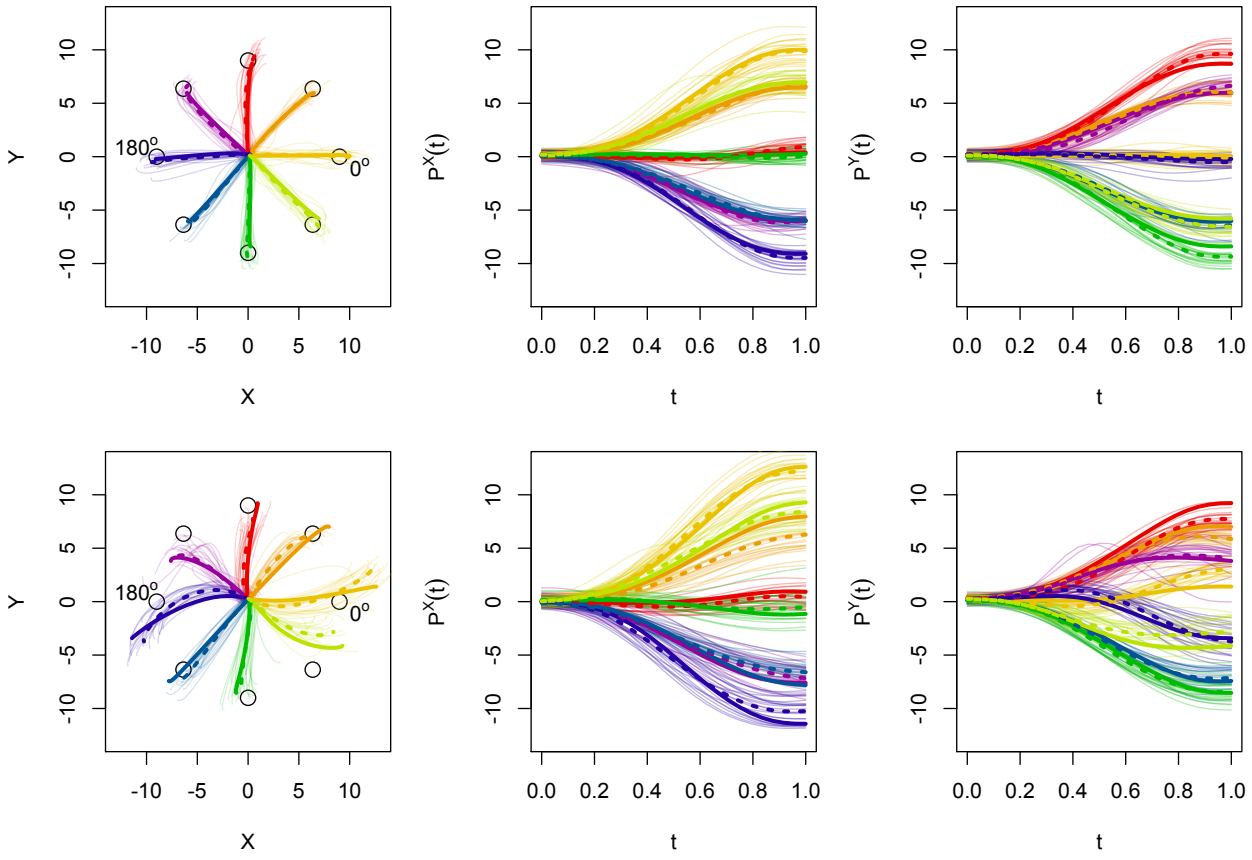

Figure 5: Observed data (faint solid lines) overlayed with estimated fixed (bold solid lines) and random (bold dashed lines) effects. Data for two subjects are shown: in the top row, the dominant hand of a control subject, and in the bottom row the affected dominant hand of a severe stroke patient. Data for these subjects appears in Figure 1.

4.2 Full Bayesian analysis

After exploring several candidate fixed effects structures, we fit Model 7 using a fully Bayesian analysis to explore inferential properties of estimated coefficients. In particular, we are interested in the target- and hand-specific systematic effects of the Fugl-Meyer score as a continuous covariate. For our final model, we used 5,000 iterations of the Gibbs sampler described in Section 2.1 and discard the first 1,000 as burn-in; visual inspection of chains (not shown) indicate good mixing. Hyperparameters and initial values were chosen as described in Section 2.3. Computations took 7.5 days, which emphasizes the importance of a fast approximation for data exploration and model building.

Figure 6 shows the estimated effect of a ten unit difference in Fugl-Meyer score in a dominant hand affected by stroke separately for the targets at 0° and 180°. For both targets, we show the marginal effect on the X and Y position curves in the left panels with pointwise posterior credible intervals. These intervals indicate a significant effect of Fugl-Meyer score on hand position for $t > .5$ and correspond to the over-reach and increased curvature observed in Figure 4. The right panels in Figure 6 show a sample from the posterior distribution of the effect of Fugl-Meyer on X and Y position jointly as a function of t; a red line shows the null of no difference. For the target at 180° this joint distribution suggests an effect for $.25 < t < .5$ not observed in either marginal plot. Though we focus on only the two most biomechanically challenging targets here, similar systematic effects are found to each of the other targets.

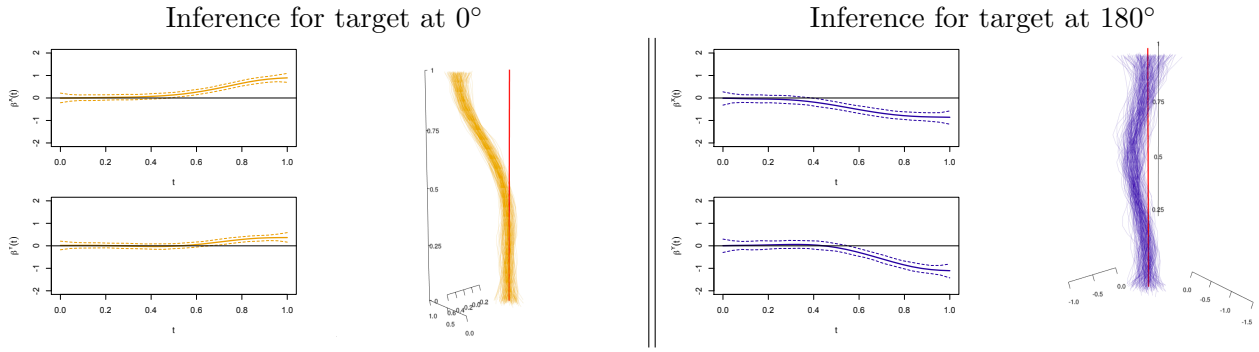

Figure 6: Inference for the linear effect of Fugl-Meyer score for targets at 0° and 180° based on the full Gibbs sampler estimation of Model 9. For both targets, marginal plots of $\beta^X(t)$ and $\beta^Y(t)$ with 95% credible intervals are shown at the left. The joint distribution of $(\beta^X(t), \beta^Y(t))$ as a function of t is shown in the right panel.

5 Concluding remarks

This manuscript has focused on the development of a regression framework for the analysis of kinematic data used to assess motor control in stroke patients. Our model allows flexible mean structures, subject-level random effects, bivariate outcomes and correlated errors. We develop a hierarchical Bayesian estimation framework; crucially, a fast and accurate variational Bayes approximation to the full Gibbs sampler allows extensive data exploration and model building before estimation with the full Bayesian approach. Implementations of both approaches and complete simulation code is publicly available.

The application of our developed methodology to the motivating data yields novel insights into the effect of arm impairment on control of visually-guided reaching. We demonstrate consistent, systematic effects of stroke on reaching trajectories using the Fugl-Meyer score as a continuous covariate that are direction-dependent. Our final model indicates that roughly 10% of variability in observed trajectories is due to systematic effects of impairment severity; subject-specific idiosyncrasies account for an additional 40%. Although not of primary concern here, our application also allows comparisons of dominant and non-dominant hand among controls, as well as consideration of systematic effects in the unaffected hand following stroke.

Future work may take several directions. In statistical methodology, additional flexibility in the mean structure, for instance by allowing non-linear effects of covariates, could broaden the applicability of the model. Parameterizing the residual correlation structure as a function of impairment severity would more accurately reflect the disease process. Adaptive smoothing and penalization for fixed effects could yield more accurate inference in some cases. Reversing the function-on-scalar problem, using kinematic data as a predictor for disease type may illuminate differences between disease types, but classification based on collections of trajectories is an open problem. In the applied setting, extension to three-dimensional kinematics will be necessary as experiments allow more complex reaching motions. Longitudinal experiments to explore treatment effects and describe the natural history of recovery are underway; accompanying methods will be needed to account for within-subject correlations over time.

6 Acknowledgments

We would like to thank John Krakauer for his scientific insight and guidance, and Johnny Liang, Sophia Ryan, and Sylvia Huang for their assistance in data collection.

References

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

Broderick, J. "William M. Feinberg Lecture: stroke therapy in the year 2025: burden, breakthroughs, and barriers to progress." Stroke, 35:205–211 (2004).

Brumback, B. and Rice, J. "Smoothing spline models for the analysis of nested and crossed samples of curves." Journal of the American Statistical Association, 93:961–976 (1998).

Chae, J., Bethoux, F., Bohinc, T., Dobos, L., Davis, T., and Friedl, A. "Neuromuscular stimulation for upper extremity motor and functional recovery in acute hemiplegia." Stroke, 29:975–979 (1998).

Coderre, A. M., Zeid, A. A., Dukelow, S. P., Demmer, M. J., Moore, K. D., Demers, M. J., Bretzke, H., Herter, T. M., Glasgow, J. I., Norman, K. E., et al. "Assessment of upper-limb sensorimotor function of subacute stroke patients using visually guided reaching." Neurorehabilitation and Neural Repair, 24:528–541 (2010).

Crainiceanu, C., Reiss, P., Goldsmith, J., Huang, L., Huo, L., and Scheipl, F. refund: Regression with Functional Data (2012). R package version 0.1-6.
URL http://CRAN.R-project.org/package=refund

Crainiceanu, C. M., Ruppert, D., and Wand, M. P. "Bayesian analysis for penalized spline regression using WinBUGS." Journal of statistical software, 14:165–185 (2005).

Di, C.-Z., Crainiceanu, C. M., Caffo, B. S., and Punjabi, N. M. "Multilevel Functional Principal Component Analysis." Annals of Applied Statistics, 4:458–488 (2009).

Fugl-Meyer, A., Jääskö, L., Leyman, I., Olsson, S., and Steglind, S. "The post-stroke hemiplegic patient. 1. A method for evaluation of physical performance." Scandinavian journal of rehabilitation medicine, 7:13–31 (1974).

Go, A. S., Mozaffarian, D., Roger, V. L., Benjamin, E. J., Berry, J. D., Borden, W. B., Bravata, D. M., Dai, S., Ford, E. S., Fox, C. S., et al. "Heart disease and stroke statistics2013 update a report from the American Heart Association." Circulation, 127:e6–e245 (2013).

Goldsmith, J., Greven, S., and Crainiceanu, C. M. "Corrected Confidence Bands for Functional Data using Principal Components." Biometrics, 69:41–51 (2013).

Goldsmith, J., Wand, M. P., and Crainiceanu, C. M. "Functional Regression via Variational Bayes." Electronic Journal of Statistics, 5:572–602 (2011).

Greven, S., Crainiceanu, C. M., Caffo, B., and Reich, D. "Longitudinal Functional Principal Component Analysis." Electronic Journal of Statistics, 4:1022–1054 (2010).

Guo, W. "Functional mixed effects models." Biometrics, 58:121–128 (2002).

Huang, V., Ryan, S., Kane, L., Huang, S., Berard, J., Kitago, T., Mazzoni, P., and Krakauer, J. "3D Robotic training in chronic stroke improves motor control but not motor function." Society for Neuroscience. October 2012. New Orleans, USA (2012).

Jordan, M. I. "Graphical models." Statistical Science, 19:140–155 (2004).

Jordan, M. I., Ghahramani, Z., Jaakkola, T. S., and Saul, L. K. "An Introduction to Variational Methods for Graphical Models." Machine Learning, 37:183–233 (1999).

Kitago, T., Liang, J., Huang, V. S., Hayes, S., Simon, P., Tenteromano, L., Lazar, R. M., Marshall, R. S., Mazzoni, P., Lennihan, L., and Krakauer, J. W. "Improvement After Constraint-Induced Movement Therapy Recovery of Normal Motor Control or Task-Specific Compensation?" Neurorehabilitation and Neural Repair, 27:99–109 (2013).

Kwakkel, B., Gand Kollen, van der Grond, J., and Prevo, A. J. "Probability of regaining dexterity in the flaccid upper Limb Impact of severity of paresis and time since onset in acute stroke." Stroke, 34:2181–2186 (2003).

Lang, C. E., Wagner, J. M., Edwards, D. F., Sahrmann, S. A., and Dromerick, A. W. "Recovery of grasp versus reach in people with hemiparesis poststroke." Neurorehabilitation and neural repair, 20:444–454 (2006).

Levin, M. F. "Interjoint coordination during pointing movements is disrupted in spastic hemiparesis." Brain, 119:281–293 (1996).

McLean, M. W., Scheipl, F., Hooker, G., Greven, S., and Ruppert, D. "Bayesian Functional Generalized Additive Models for Sparsely Observed Covariates." Under Review (2013).

Morris, J. S. and Carroll, R. J. "Wavelet-based functional mixed models." Journal of the Royal Statistical Society: Series B, 68:179–199 (2006).

Ormerod, J. and Wand, M. P. "Explaining Variational Approximations." The American Statistician, 64:140–153 (2010).

—. "Gaussian Variational Approximation Inference for Generalized Linear Mixed Models." The American Statistician, 21:2–17 (2012).

Ramsay, J. O. and Silverman, B. W. Functional Data Analysis. New York: Springer (2005).

Reiss, P. and Ogden, T. "Functional generalized linear models with images as predictors." Biometrics, 66:61–69 (2010).

Reiss, P. T. and Huang, L. "Fast Function-on-Scalar Regression with Penalized Basis Expansions." International Journal of Biostatistics, 6:Article 28 (2010).

Ruppert, D., Wand, M. P., and Carroll, R. J. Semiparametric Regression. Cambridge: Cambridge University Press (2003).

Scheipl, F., Staicu, A.-M., and Greven, S. "Additive Mixed Models for Correlated Functional Data." Under Review (2013).

Staicu, A.-M., Crainiceanu, C., and Carroll, R. "Fast methods for spatially correlated multilevel functional data." Biostatistics, 11:177–194 (2010).

Titterington, D. M. "Bayesian Methods for Neural Networks and Related Models." Statistical Science, 19:128–139 (2004).

van der Linde, A. "Variational Bayesian functional PCA." Computational Statistics and Data Analysis, 53:517–533 (2008).

Yao, F., Müller, H., and Wang, J. "Functional data analysis for sparse longitudinal data." Journal of the American Statistical Association, 100(470):577–590 (2005).

Appendices to: Assessing Systematic Effects of Stroke on Motor Control using Hierarchical Function-on-Scalar Regression

Jeff Goldsmith and Tomoko Kitago

This appendix contains derivations of full conditional distributions for the full Gibbs sampler and optimal densities for the variational Bayes algorithm. Derivations are for the univariate model; for the bivariate outcome model presented in Section 2.4 slight augmentations are necessary but straightforward. For completeness, we briefly describe the data and model of interest. The data are $[Y_{ij}(t), \boldsymbol{w}_{ij}]$ for subjects $i = 1, \ldots, I$ and visits $j = 1, \ldots J_i$, giving a total of $n = \sum_i J_i$ observations. Univariate functional outcomes $Y_{ij}(t)$ are observed on a regular grid of length D for all subjects and visits. We are interested in estimating the parameters in

$$\begin{aligned}
\boldsymbol{Y} &= \boldsymbol{Z}\boldsymbol{B}_Z^T\boldsymbol{\Theta}^T + \boldsymbol{\epsilon} \\
\boldsymbol{\epsilon} &\sim \text{N}\left[0, \Sigma \otimes I_n\right]; \Sigma \sim \text{IW}\left[\nu, \boldsymbol{\Psi}\right] \\
\boldsymbol{B}_{Z_i} &\sim \text{N}\left[\boldsymbol{w}_i\boldsymbol{B}_W, \sigma_Z^2 P^{-1}\right] \text{ for } i = 1 \ldots I; \sigma_Z^2 \sim \text{IG}\left[a_Z, b_Z\right] \\
\boldsymbol{B}_{W_k} &\sim \text{N}\left[0, \sigma_{W_k}^2 P^{-1}\right], \sigma_{W_k}^2 \sim \text{IG}\left[a_{W_k}, b_{W_k}\right] \text{ for } k = 1 \ldots p.
\end{aligned} \tag{A.1}$$

In this model, $\boldsymbol{B}_W^T$ is the matrix of coefficients for fixed effects and $\boldsymbol{B}_Z^T$ is the matrix of coefficients for random subject effects. Additionally, $\boldsymbol{Y}$, $\boldsymbol{W}$, $\boldsymbol{Z}$ and $\boldsymbol{\Theta}$ are the observed outcomes, the fixed and random effect design matrices, and the b-spline basis matrix, respectively.

A Gibbs Sampler

In this section we provide full conditional distributions for the parameters in model (A.1). Let $\mathbf{1}_m$ be a length m vector of 1's and $\boldsymbol{Y}_i$ denote the rows of $\boldsymbol{Y}$ for subject i.

- For $\boldsymbol{B}_{Z_i}$, we have that

$$p\left[\text{vec}\left(\boldsymbol{B}_{Z_i}\right)|\text{rest}\right] \propto \text{N}\left[\boldsymbol{\mu}_{\boldsymbol{B}_{Z_i}}, \Sigma_{\boldsymbol{B}_{Z_i}}\right]$$

  where

$$\Sigma_{\boldsymbol{B}_{Z_i}} = \left((\mathbf{1}_{J_i} \otimes \boldsymbol{\Theta})^T(I_{J_i} \otimes \Sigma^{-1})(\mathbf{1}_{J_i} \otimes \boldsymbol{\Theta}) + \frac{1}{\sigma_Z^2}P\right)^{-1}$$

and

$$\boldsymbol{\mu}_{\boldsymbol{B}_{Z_i}} = \Sigma_{\boldsymbol{B}_{Z_i}} \left((\mathbf{1}_{J_i} \otimes \boldsymbol{\Theta})^T (I_{J_i} \otimes \Sigma^{-1}) \text{vec}\left((\boldsymbol{Y}_i)^T \right) + \frac{1}{\sigma_Z^2} P(\boldsymbol{B}_W \boldsymbol{w}_i^T) \right).$$

- For $\boldsymbol{B}_W$, we have that

$$p\left[\text{vec}\left(\boldsymbol{B}_W\right)|\text{rest}\right] \propto \text{N}\left[\boldsymbol{\mu}_{\boldsymbol{B}_W}, \Sigma_{\boldsymbol{B}_W}\right]$$

where

$$\Sigma_{\boldsymbol{B}_W} = \left(\frac{1}{\sigma_Z^2} (\boldsymbol{W} \otimes I_{K_t})^T (I_I \otimes P)(\boldsymbol{W} \otimes I_{K_t}) + \text{diag}\left(\frac{1}{\sigma_{W_k}^2} \right) \otimes P \right)^{-1}$$

and

$$\boldsymbol{\mu}_{\boldsymbol{B}_W} = \Sigma_{\boldsymbol{B}_W} \left(\frac{1}{\sigma_Z^2} (\boldsymbol{W} \otimes I_{K_t})^T (I_I \otimes P) \text{vec}\left(\boldsymbol{B}_Z\right) \right).$$

- For σ_Z^2, we have that

$$p\left[\sigma_Z^2|\text{rest}\right] \propto \text{IG}\left[a_z + \frac{I * K_t}{2}, b_z + \frac{1}{2} \sum_i \left(\boldsymbol{B}_{Z_i} - \boldsymbol{B}_W \boldsymbol{w}_i^T\right)^T P \left(\boldsymbol{B}_{Z_i} - \boldsymbol{B}_W \boldsymbol{w}_i^T\right) \right].$$

- For $\sigma_{W_k}^2$, we have that

$$p\left[\sigma_{W_k}^2|\text{rest}\right] \propto \text{IG}\left[a_{W_k} + \frac{K_t}{2}, b_{W_k} + \frac{1}{2} \boldsymbol{B}_{W_k}^T P \boldsymbol{B}_{W_k} \right].$$

- For Σ, we have that

$$p\left[\sigma_{W_k}^2|\text{rest}\right] \propto \text{IW}\left[\nu + \sum_i J_i, \boldsymbol{\Psi} + \left(\boldsymbol{Y} - \boldsymbol{Z}\boldsymbol{B}_Z^T \boldsymbol{\Theta}^T\right)^T \left(\boldsymbol{Y} - \boldsymbol{Z}\boldsymbol{B}_Z^T \boldsymbol{\Theta}^T\right) \right].$$

B Variational Bayes Algorithm

In this section we present the iterative algorithm for updating optimal densities q^* in the variational Bayes approximation to the full posterior of model (A.1). Additionally, we provide the q-specific lower bound L_q used to monitor convergence of the algorithm.

In the following, we approximate the full posterior $p(\boldsymbol{B}_Z, \boldsymbol{B}_W, \sigma^2_{W_1}, \ldots, \sigma^2_{W_p}, \sigma^2_Z, \Sigma|\boldsymbol{Y})$ using $q(\cdot)$, where we assume the partitioning

$$\begin{aligned} & q\left(\boldsymbol{B}_Z, \boldsymbol{B}_W, \sigma^2_{W_1}, \ldots, \sigma^2_{W_p}, \sigma^2_Z, \Sigma\right) \\ = \;& q(\boldsymbol{B}_Z) q(\boldsymbol{B}_W) \left(\prod_{k=1}^{p} q(\sigma^2_{W_k})\right) q(\sigma^2_Z) q(\Sigma). \end{aligned}$$

Based on this factorization, it can be shown that:

- the optimal density $q(\boldsymbol{B}_{Z_i})$ is $\mathrm{N}\left[\mu_{q(\boldsymbol{B}_{Z_i})}, \Sigma_{q(\boldsymbol{B}_{Z_i})}\right]$ for all subjects i;
- the optimal density $q(\boldsymbol{B}_W)$ is $\mathrm{N}\left[\mu_{q(\boldsymbol{B}_W)}, \Sigma_{q(\boldsymbol{B}_W)}\right]$;
- the optimal density $q(\sigma^2_Z)$ is $\mathrm{IG}\left[a_Z + \frac{I*K_t}{2}, b_{q(\sigma^2_Z)}\right]$;
- the optimal density $q(\sigma^2_{W_k})$ is $\mathrm{IG}\left[a_W + \frac{I*K_t}{2}, b_{q(\sigma^2_{W_k})}\right]$ for all coefficients k;
- the optimal density $q(\Sigma)$ is $\mathrm{IW}\left[\nu + \sum_i J_i, \boldsymbol{\Psi}_{q(\Sigma)}\right]$

where the notation $\mu_{q(\phi)}$ and $\Sigma_{q(\phi)}$ indicate the mean and variance for Normal densities $q(\phi)$; similar notation is used for the parameters in inverse Gamma and inverse Wishart densities. Parameters in these densities is updated iteratively according to the following algorithm.

Algorithm 1 *Iterative scheme for obtaining the optimal density parameters in the function-on-scalar regression model (A.1).*

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

Cycle:

For all i:

$$\Sigma_{q(\boldsymbol{B}_{Z_i})} \leftarrow \left((\boldsymbol{1}_{J_i} \otimes \boldsymbol{\Theta})^T (I_{J_i} \otimes \left(\frac{\boldsymbol{\Psi}_{q(\Sigma)}}{\nu + \sum_i J_i} \right)^{-1}) (\boldsymbol{1}_{J_i} \otimes \boldsymbol{\Theta}) + \frac{a_Z + \frac{I * K_t}{2}}{b_{q(\sigma^2_Z)}} P \right)^{-1}$$

$$\mu_{q(\boldsymbol{B}_{Z_i})} \leftarrow \Sigma_{q(\boldsymbol{B}_{Z_i})} \left((\boldsymbol{1}_{J_i} \otimes \boldsymbol{\Theta})^T (I_{J_i} \otimes \left(\frac{\boldsymbol{\Psi} + \boldsymbol{\Psi}_{q(\Sigma)}}{\nu + \sum_i J_i} \right)^{-1}) \text{vec}\left((\boldsymbol{Y}_i)^T \right) + \frac{a_Z + \frac{I * K_t}{2}}{b_{q(\sigma^2_Z)}} P (\mu_{q(\boldsymbol{B}_W)} \boldsymbol{w}_i^T) \right)$$

For all k:

$$\Sigma_{q(\boldsymbol{B}_{W_k})} \leftarrow \left(\frac{a_Z + \frac{I * K_t}{2}}{b_{q(\sigma^2_Z)}} (\boldsymbol{W} \otimes I_{K_t})^T (I_I \otimes P) (\boldsymbol{W} \otimes I_{K_t}) + \text{diag}\left(\frac{a_W + \frac{K_t}{2}}{b_{q(\sigma^2_{W_k})}} \right) \otimes P \right)^{-1}$$

$$\mu_{q(\boldsymbol{B}_{W_k})} \leftarrow \Sigma_{q(\boldsymbol{B}_W)} \left(\frac{a_Z + \frac{I * K_t}{2}}{b_{q(\sigma^2_Z)}} (\boldsymbol{W} \otimes I_{K_t})^T (I_I \otimes P) \text{vec}\left(\mu_{q(\boldsymbol{B}_Z)} \right) \right)$$

$$b_{q(\sigma^2_Z)} \leftarrow b_Z + \tfrac{1}{2} \sum_i \left(\mu^T_{q(\boldsymbol{B}_{Z_i})} P \mu_{q(\boldsymbol{B}_{Z_i})} + \text{tr}\left[P \Sigma_{q(\boldsymbol{B}_{Z_i})} \right] - 2 \boldsymbol{w}_i \mu^T_{q(\boldsymbol{B}_W)} P \mu_{q(\boldsymbol{B}_{Z_i})} + \boldsymbol{w}_i \mu^T_{q(\boldsymbol{B}_W)} P \mu_{q(\boldsymbol{B}_W)} \boldsymbol{w}_i^T + \boldsymbol{w}_i \text{diag}\left(\text{tr}\left[P \Sigma_{q(\boldsymbol{B}_{W_k})} \right] \right) \boldsymbol{w}_i^T \right)$$

For all k:

$$b_{q(\sigma^2_{W_k})} \leftarrow b_W + \tfrac{1}{2} \left(\mu^T_{q(\boldsymbol{B}_{W_k})} P \mu_{q(\boldsymbol{B}_{W_k})} + \text{tr}\left[P \Sigma_{q(\boldsymbol{B}_{W_k})} \right] \right)$$

$$\boldsymbol{\Psi}_{q(\Sigma)} \leftarrow \boldsymbol{\Psi} + \boldsymbol{Y}^T \boldsymbol{Y} - \boldsymbol{Y}^T \boldsymbol{Z} \mu^T_{q(\boldsymbol{B}_Z)} \boldsymbol{\Theta}^T + \boldsymbol{\Theta} \mu_{q(\boldsymbol{B}_Z)} \boldsymbol{Z}^T \boldsymbol{Y} + \boldsymbol{\Theta} \mu_{q(\boldsymbol{B}_Z)} \boldsymbol{Z}^T \boldsymbol{Z} \mu^T_{q(\boldsymbol{B}_Z)} \boldsymbol{\Theta}^T + \sum_i J_i \boldsymbol{\Theta} \Sigma_{q(\boldsymbol{B}_{Z_i})} \boldsymbol{\Theta}^T$$

until the increase in L_q is negligible.

Finally, the expression for the q-specific lower bound of the marginal log-likelihood is

$$
\begin{aligned}
L_{q^*} &= \int q(\boldsymbol{\phi}) \log \frac{p(\boldsymbol{y}, \boldsymbol{\phi})}{q^*(\boldsymbol{\phi})} d\boldsymbol{\phi} = \\
&= \frac{1}{2} \sum_i \log(|\Sigma_{q(\boldsymbol{B}_{Z_i})}|) + \frac{1}{2} \sum_k \log(|\Sigma_{q(\boldsymbol{B}_{W_k})}|) - \\
&\quad \left(a_z + \frac{I * K_t}{2} \right) \log(b_{q(\sigma_Z^2)}) - \sum_k \left(a_w + \frac{K_t}{2} \right) \log(b_{q(\sigma_{W_k}^2)}) - \\
&\quad \left(\frac{\nu + \sum_i J_i}{2} \right) \log(|\boldsymbol{\Psi}_{q(\Sigma)}|) + \text{ const.}
\end{aligned}
$$

where "const." represents an additive constant not affected by updates to the q density parameters. It should be noted that this constant contains the term $\log(|P^1|)$, thus necessitating a full rank penalty matrix if L_q is used to monitor convergence of the algorithm.