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
[*]*jeff.goldsmith@columbia.edu*
[2]Department of Neurology, Columbia University Medical Center

November 5, 2015

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). Because of this, there is a great need for development of neurorehabilitative therapies to improve arm movements after stroke. One of the challenges in developing and testing therapeutic interventions to promote motor recovery after stroke is that it remains unclear to what extent these motor deficits are idiosyncratic (or subject-specific) rather than common across affected patients, and how these deficits vary according to the severity of motor impairment. A better understanding of these factors would allow for therapies that target the specific motor deficits that are shared by stroke patients, but perhaps to also tailor therapies based on individual characteristics, such as stroke severity.

In this current study we focus on the effects of stroke on motor control, which we define as the ability to make accurate, goal-directed movements. We use a planar reaching task designed to test a fundamental level of motor control (Kitago et al., 2013), and explore the relationship between patients' performance on this task and the Fugl-Meyer Upper Extremity Motor Assessment (FM-UE, Fugl-Meyer et al. (1974)), a clinical measure of the severity of arm motor impairment, in a population of chronic stroke patients with residual arm paresis. Elderly healthy controls are included as a reference group.

In the reaching task, observations at the subject level are repeated two-dimensional motion trajectories to eight target directions parameterized by time. Our analytical approach for these multilevel 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 motor impairment 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 are 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 (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.

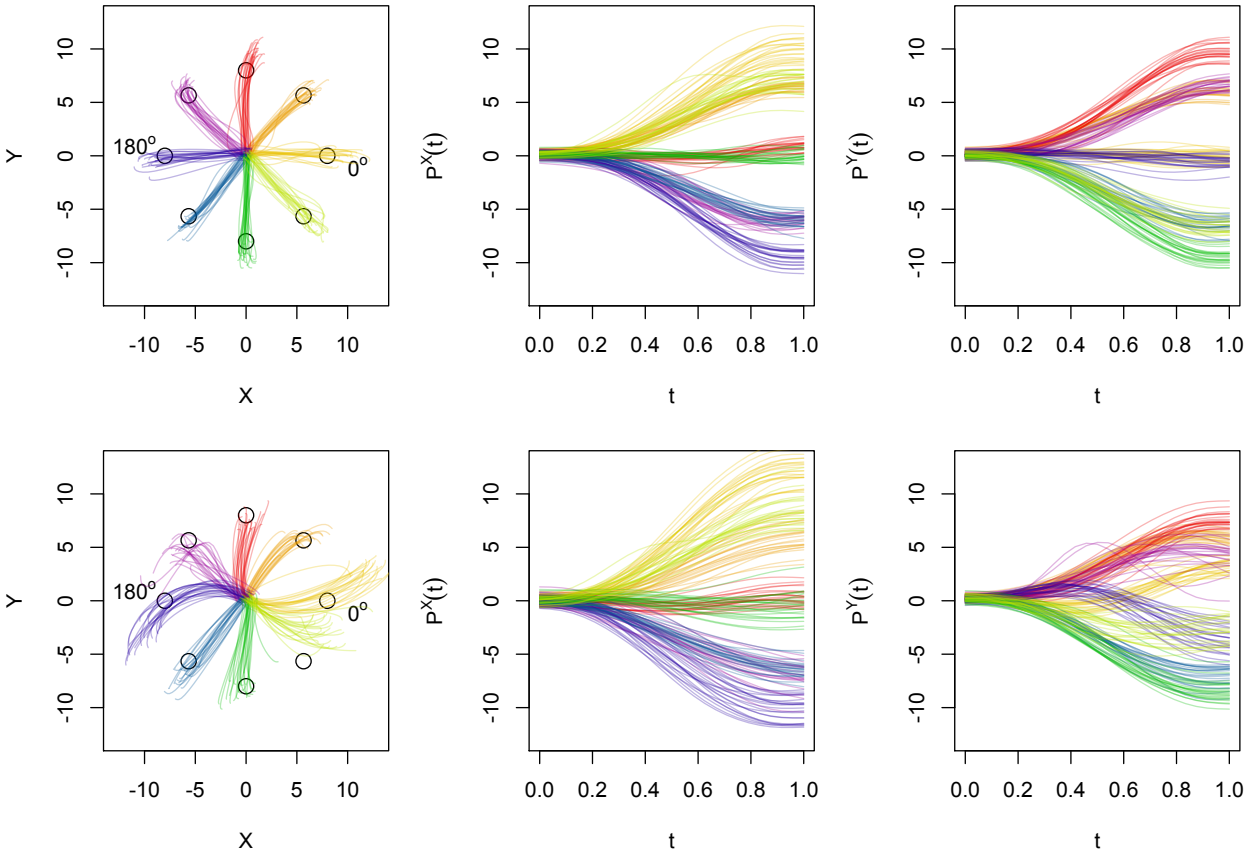

Figure 1: Observed data for two subjects; the top row shows the dominant hand of a healthy control, and the bottom row shows the affected dominant hand of a severe stroke patient. The left column shows observed kinematic data for all reaches observed in the dominant hand. The middle and right columns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

Our goal in this analysis is to explore the extent to which the effects of stroke on motor control are shared across subjects or are subject-specific through a regression analysis using a combination of subject-level scalar covariates, such as the Fugl-Meyer measure of impairment severity, as predictors of interest. Evidence for 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

Conceptually, we observe functional 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 $\boldsymbol{w}_i = [w_{i1}, \ldots, w_{ip}]$ is a length p vector of scalar covariates. We propose the model

$$\begin{cases} P_{ij}^X(t) = \beta_0^X(t) + \sum_{k=1}^p w_{ik}\beta_k^X(t) + b_i^X(t) + \epsilon_{ij}^X(t) \\ P_{ij}^Y(t) = \beta_0^Y(t) + \sum_{k=1}^p w_{ik}\beta_k^Y(t) + b_i^Y(t) + \epsilon_{ij}^Y(t) \end{cases} \tag{1}$$

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 a Bayesian framework that allows the joint modeling of the mean structure (through fixed and random effects) and residual correlation (through the error covariance matrix) in a single Gibbs sampler. Importantly, a variational Bayes algorithm provides a computationally efficient and accurate approximation to the full sampler. Model 1 is analogous to a standard mixed model with subject-level random intercepts; the identifiability of fixed and random effects will depend on the prior specification, hyperparameter selection and sampling framework, which we discuss in Sections 2.1 and 2.3.

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 the domain $[0, 1]$ with D elements, and that this grid is common to all subjects. In our application, trajectories are observed at 120Hz. Although efforts were made to ensure uniform motion times of approximately .5 seconds, motions take different times to complete and we use a linear registration to obtain a common, evenly-space observation grid of length $D = 25$ prior to analysis. After registration, motions are observed from $t = 0$ to

$t = 1$ with 0 and 1 indicating the beginning and end of the motion, respectively. 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 datasets. 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), have modeled curve-specific deviations from a population mean using low-dimensional basis functions estimated from the empirical covariance matrix. These methods did not focus on the flexible estimation of the population mean surface; moreover in assessing uncertainty these methods implicitly conditioned 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 et al. (2010) built 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) proposed a very flexible class of functional outcome models, allowing cross sectional or multilevel 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 uncorrelated measurement error despite clear correlation in the functional domain. One alternative, proposed by Reiss et al. (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.

Several Bayesian methods for function-on-scalar regression exist in the literature. Morris et al. (2003) developed wavelet-based functional mixed models (FMMs) assuming residual curves consist of independent measurement errors; Morris and Carroll (2006) extended this to allow correlated residual curves. Both approaches used a discrete wavelet transform (DWT) of the observed data and model coefficients in the wavelet domain using spike-and-slab priors. It was assumed that errors in the wavelet domain are independent, justified heuristically by the whitening property of the DWT. After modeling using MCMC, the inverse DWT provided estimates in the observed space. A penalized spline approach for FMMs was taken in Baladandayuthapani et al. (2007), while Baladandayuthapani et al. (2010) used a piecewise constant basis; in both cases, correlated residual curves are explicitly modeled rather than treated as errors around the mean. For cross-sectional functional data observed sparsely at the subject level, Montagna et al. (2012) developed a Bayesian latent factor model in which predictors are incorporated at the latent factor level; a potentially large number of factors are allowed, and sparsity is induced through a shrinkage prior on the basis coefficients. The computation burden of the Bayesian procedures can be prohibitive for data exploration and model building even for moderate datasets, which has contributed to the slow adoption of Bayesian methods in functional data analysis. As an example, a comparison of the Bayesian penalized spline method in Baladandayuthapani et al. (2007) to an FPCA-based method on simulated data found computation times of five hours versus five seconds (Staicu et al., 2010).

This paper 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 and, surprisingly, to a 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 multilevel function-on-scalar regression, we consider 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. The web-based supporting materials present 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 Sections 2.1, 2.2 and 2.3. These Sections develop our methodology for estimating fixed effects, random effects, and residual covariance when a single functional response is observed. Once this is established, we consider the bivariate model in Section 2.4.

2.1 Full Model

For now, assume data are $[Y_{ij}(t), \boldsymbol{w}_i]$ for subjects $i = 1, \ldots, I$ and visits $j = 1, \ldots J_i$, giving a total of $n = \sum_i J_i$ functional 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

$$\boldsymbol{y}_{ij} = \boldsymbol{w}_i\boldsymbol{\beta} + \boldsymbol{z}_{ij}\boldsymbol{b} + \boldsymbol{\epsilon}_{ij} \quad \text{with} \quad \boldsymbol{\epsilon}_{ij} \sim \mathrm{N}\left[0, \Sigma\right] \tag{2}$$

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 $\boldsymbol{\epsilon}_{ij}$ is a $1 \times D$ vector of residual curves distributed $\mathrm{N}\left[0, \Sigma\right]$ where Σ is a $D \times D$ covariance matrix. Errors $\boldsymbol{\epsilon}_{ij}$ are conditionally independent given fixed effects and subject-specific random effects, and are *iid* across subjects and visits.

We express the functional effects in the rows of $\boldsymbol{\beta}$ and $\boldsymbol{b}$ using a spline expansion. Let Θ denote a $D \times K_\theta$ cubic B-spline evaluation matrix with K_θ basis functions. Further let B_W and B_Z denote the matrices

whose columns are basis coefficients for $\boldsymbol{\beta}$ and $\boldsymbol{b}$ respectively, so that $\boldsymbol{\beta} = [\Theta B_W]^T$ and $\boldsymbol{b} = [\Theta B_Z]^T$. (This notation is drawn from Ramsay and Silverman (2005, Section 13.4.3) and Reiss et al. (2010); readers should note potential conflicts in notation, particularly with Montagna et al. (2012) and references therein.)

Penalization is a commonly-used technique to avoid overfitting and induce smoothness in functional effects. For spline coefficients in the kth column $B_{W,k}$ of B_W and ith column $B_{Z,i}$ of B_Z, we assume that $B_{W,k} \sim \mathrm{N}\left[0, \sigma^2_{W,k} P^{-1}\right]$ and $B_{Z,i} \sim \mathrm{N}\left[0, \sigma^2_Z P^{-1}\right]$. P is a known penalty matrix shared across fixed and random effects to enforce a common penalty structure. Variances $\left\{\sigma^2_{W,k}\right\}$ are unique to each coefficient function to allow unique levels of smoothness, and σ^2_Z is shared across random effects so that they are draws from a common population. Notationally, the "zeroth" column $B_{W,0}$ of B_W and the variance $\sigma^2_{W,0}$ correspond to the intercept $\beta_0(t)$. The connection between this prior specification and a quadratic roughness penalty well known; see Ruppert et al. (2003, Ch. 4.9) for a detailed treatment. The choice of penalty matrix P is discussed in Section 2.3. In this penalized spline framework, the number and position of knots is typically unimportant provided that the number is sufficient to model the complexity of the coefficient functions (Ruppert, 2002). Using this specification, model (2) can be expressed as

$$\begin{aligned}
\boldsymbol{y}_{ij} &= \boldsymbol{w}_i B_W^T \Theta^T + \boldsymbol{z}_{ij} B_Z^T \Theta^T + \boldsymbol{\epsilon}_{ij} \\
\boldsymbol{\epsilon}_{ij} &\sim \mathrm{N}\left[0, \Sigma\right] \\
B_{Z,i} &\sim \mathrm{N}\left[0, \sigma^2_Z P^{-1}\right] \text{ for } i = 1 \ldots I \\
B_{W,k} &\sim \mathrm{N}\left[0, \sigma^2_{W,k} P^{-1}\right] \text{ for } k = 0 \ldots p
\end{aligned} \tag{3}$$

which has a form similar to a traditional mixed model.

The variance components $\left\{\sigma^2_{W,k}\right\}, \sigma^2_Z$ are assigned inverse-gamma priors, and our model specification is completed using an inverse Wishart prior for the residual covariance matrix Σ. Although these priors are convenient for the development of a straightforward Gibbs sampler and for the derivation of the variational approximation in Section 2.2, they have been criticized by several researchers (Gelman, 2006; Yang and Berger, 1994). Inverse-gamma priors can be sensitive to the choice of hyperparameters a and b, especially for "uninformative" values like $a = b = .001$ that place a large prior mass near zero, and inverse-Wishart priors may not shrink eigenvalues as expected. In Section 2.3 we suggest using the data to help choose

hyperparameters in a reasonable way, but emphasize the need for sensitivity analyses to these choices.

Following Gelfand et al. (1995), we pursue an alternative parameterization of model (3) to improve sampling performance through hierarchical recentering. For a simple example of this idea, consider the (non-functional) model $y_{ij} = \mu + b_i + \epsilon_{ij}$ with priors for μ and b_i having mean zero; alternatively one could let $Y_{ij} = \eta_i + \epsilon_{ij}$ with the prior for η_i having mean α and the prior for α having mean zero. The latter parameterization often results in better behavior of the posterior chains and increased identifiability of fixed and random effects. For our function-on-scalar regression model, let Y be an $n \times D$ matrix of row-stacked functional outcomes, Z be the random effects design matrix, W be the fixed effects matrix constructed by row-stacking the $\boldsymbol{w}_i$, and $\otimes$ represent the Kronecker product operator. Using hierarchical recentering, we reparameterize model (3) using

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

Full conditionals for all model parameters are straightforward to obtain using vector notation and Kronecker products. For matrix M and vector c, let $\mathrm{vec}\,(M)$ be the vector formed by concatenating the columns of M and $\mathrm{diag}\,(c)$ 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

$$p\left[\mathrm{vec}\left(B_W\right)|\mathrm{rest}\right] \propto \mathrm{N}\left[\mu_{B_W}, \Sigma_{B_W}\right]$$

where

$$\Sigma_{B_W} = \left(\frac{1}{\sigma_Z^2}(W \otimes I_{K_\theta})^T(I_I \otimes P)(W \otimes I_{K_\theta}) + \mathrm{diag}\left(\frac{1}{\sigma_{W,k}^2}\right) \otimes P\right)^{-1}$$

and

$$\mu_{B_W} = \Sigma_{B_W}\left(\frac{1}{\sigma_Z^2}(W \otimes I_{K_\theta})^T(I_I \otimes P)\mathrm{vec}\left(B_Z\right)\right).$$

Additionally, we have that

$$p\left[\sigma^2_{W,k}|\text{rest}\right] \propto \text{IG}\left[a_{W,k} + \frac{K_\theta}{2}, b_{W,k} + B^T_{W,k} P B_{W,k}\right].$$

Complete derivations of this and all other full conditional distributions are provided in the web-based supplementary materials.

Model (4) contains a large number of parameters, particularly in the spline coefficient matrices B_W and B_Z. Typically the available data for estimation in the $n \times D$ outcome matrix Y will dwarf the number of parameters in B_W and B_Z, meaning these can be well-estimated. However, in some cases it may be necessary to use a low-dimensional spline basis or other parametric approach for the estimation of fixed and random effects. This is possible using a simplification of model (4), but it should be noted the choice of parametric form will be an implicit tuning parameter that replaces the explicit penalization implemented above. When the number of observed points per curve D is large the covariance matrix Σ will also be large, and the number of parameters in this matrix could be substantial; again, a modification of model (4) to use a parametric form for Σ is possible and perhaps advisable in this case, although the caveats regarding the introduction of a parametric structure still apply. In our motivating dataset n is large (≈ 20000) while D is modest (50 for bivariate curves), and model (4) is reasonable.

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 assume

$$q\left(B_Z, B_W, \sigma^2_{W,0}, \ldots, \sigma^2_{W,p}, \sigma^2_Z, \Sigma\right) \quad = \quad q(B_Z) q(B_W) q(\sigma^2_{W,0}, \ldots, \sigma^2_{W,p}, \sigma^2_Z, \Sigma)$$

where the functions q are distinguished by their argument rather than by subscript l. The additional factorization

$$q\left(B_Z, B_W, \sigma^2_{W,0}, \ldots, \sigma^2_{W,p}, \sigma^2_Z, \Sigma\right) \quad = \quad \left(\prod_{i=1}^{I} q(B_{Z,i})\right) q(B_W) \left(\prod_{k=0}^{p} q(\sigma^2_{W,k})\right) q(\sigma^2_Z) q(\Sigma)$$

is induced by the conditional independence properties of the joint distribution (see Bishop (2006, Sec. 10.2.5); a directed acyclic graph of our model appears in Figure A.1). Using this factorization, it can be shown that the optimal density $q^*(\text{vec}\,(B_W))$ for is $\text{N}\left[\mu_{q(B_W)}, \Sigma_{q(B_W)}\right]$, where

$$\Sigma_{q(B_W)} = \left(\mu_{q(1/\sigma^2_Z)} (W \otimes I_{K_\theta})^T (I_I \otimes P)(W \otimes I_{K_\theta}) + \text{diag}\left(\mu_{q(1/\sigma^2_{W,k})}\right) \otimes P\right)^{-1}$$

and

$$\mu_{q(B_W)} = \Sigma_{q(B_W)} \left(\mu_{q(1/\sigma_Z^2)} (W \otimes I_{K_\theta})^T (I_I \otimes P) \text{vec} \left(\mu_{q(B_Z)} \right) \right).$$

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 $q^*(\text{vec}\,(B_W))$ 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 the web-based supplementary materials.

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 is commonly used in functional data analysis and enforces smoothness in the estimated function, but is non-invertible. The P_0 penalty matrix is the identity matrix, induces general shrinkage and is full rank. Details on the construction of P_0 and P_2 are available in Eilers and Marx (1996). Selecting $0 < \alpha \leq 1$ balances smoothness and shrinkage, and results in a full rank penalty matrix. There is can be some sensitivity to the choice of α, with large values shrinking estimates toward 0; we recommend a relatively small value ($\alpha = 0.01$ or smaller) in keeping with the tendency to enforce smoothness rather than shrinkage.

We use the following procedure based on model (4) to choose hyperparameters. First, we estimate B_Z using ordinary least squares from the regression $\text{E}\,[Y] = ZB_Z^T\Theta^T$ to obtain B_Z^{OLS}. We estimate the error covariance Σ using a functional principal components decomposition of the residuals from this regression to obtain $\hat{\Sigma}$ (Yao et al., 2005), and use $\nu = \sum_i J_i$, $\Psi = \sum_i J_i \hat{\Sigma}$ as hyperparameters in the prior for Σ. Next, we estimate B_W using weighted least squares from the regression $\text{E}\left[\left(B_Z^{OLS} \right)^T \right] = WB_W^T$ with weight

matrix P^{-1} to obtain B_W^{WLS}. We choose

$$a_z = \frac{I * K_\theta}{2}, b_z = \frac{1}{2}\text{tr}\left[\left(\left(B_Z^{OLS}\right)^T - W\left(B_W^{WLS}\right)^T\right)^T P\left(\left(B_Z^{OLS}\right)^T - W\left(B_W^{WLS}\right)^T\right)\right]$$

and

$$a_{W,k} = \frac{K_\theta}{2}, b_{W,k} = \frac{1}{2}\left(B_{W,k}^{WLS}\right)^T P\left(B_{W,k}^{WLS}\right)$$

motivated by the form of the full conditionals for these variance components.

This procedure avoids the default "uninformative" choice $a_z = b_z = a_{W,k} = b_{W,k} = .001$, which places a large prior mass near zero for all variance components. Such a choice favors overshrinkage of fixed and random effects toward zero. In our application, we found that using this default tended to result in the incorporation of subject-level random effects into the error variance Σ. Sensitivity to the choice of tuning parameters should be assessed in each application, and weakly informative priors used where possible; see Section 4.2 and the web-based supplementary materials for details of a sensitivity analysis for our application.

Initial values are sampled from a $\text{N}[0, 100]$ for spline coefficients and from a Uniform$[.1, 10]$ for variance components. The starting value for Σ is $\sigma^2 I_D$ where σ^2 is Uniform$[.1, 10]$. For other applications, different starting values may be needed.

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. Similar extensions to three or more curves observed over a common domain proceed similarly.

Let $Y = [Y_1 Y_2]$ be the concatenation of two outcome matrices Y_1 and Y_2 (in our example, we concatenate the X and Y position curves so that $Y = [P^X P^Y]$). Next, let $B_W^T = [B_{1,W}^T, B_{2,W}^T]$ so that the columns of B_W concatenate the fixed effect spline coefficients for Y_1 and Y_2. Similarly, let $B_Z^T = [B_{1,Z}^T, B_{2,Z}^T]$ so that the columns of B_Z concatenate the random effect spline coefficients for Y_1 and Y_2. Then $B_Z^T\left(I_2 \otimes \Theta^T\right)$ contains subject-specific random effects for the overall outcome matrix Y. For spline coefficients in columns

$B_{W,k}$ we assume that $B_{W,k} \sim \mathrm{N}\left[0, \mathrm{diag}\left(\sigma^2_{1,W_k}, \sigma^2_{2,W_k}\right) \otimes P^{-1}\right]$, so that fixed effects for Y_1 and Y_2 are penalized separately and assumed independent *a priori*. A similar specification is used for the columns $B_{Z,i}$, and variance components are assigned inverse-gamma priors. Again using a hierarchical formulation, for bivariate data our complete model is

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

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).

This bivariate model is motivated by the data structure, in which X and Y position curves are observed concurrently for all motions – in that sense, the bivariate model is more faithful to the observed data than fitting separate univariate models. A bivariate model also explicitly models correlation in X and Y position error curves. In our application, this correlation may provide insight into sensory or visual feedback in reaching motions, or into the biomechanical processes involved. Nonetheless, it is possible to fit a bivariate model using two separate univariate models, especially if X and Y errors are uncorrelated or if this correlation is not scientifically meaningful.

3 Simulations

We demonstrate the performance of our method using a simulation in which generated data mimic the motivating application; all code needed to reproduce these simulations is available on the first author's website. Our simulations consider three groups: control subjects' dominant hand; the affected dominant

hand of moderately affected stroke patients; and the affected dominant hand of severely affected stroke patients. We therefore created a three-level categorical predictor with Groups 1, 2, and 3 referring to controls, moderately affected subjects, and severely affected subjects.

Curves are observed on a common grid of length $D = 25$. Data are generated from the univariate outcome model $\boldsymbol{y}_{ij} = \boldsymbol{w}_i\boldsymbol{\beta} + \boldsymbol{b}_i + \boldsymbol{\epsilon}_{ij}$ where $\boldsymbol{w}_i$ is a length-3 binary vector indicating group membership for subject i, $\boldsymbol{\beta}$ is a $3 \times D$ matrix whose rows are group average curves, $\boldsymbol{b}_i$ is a length-D random effect, and $\boldsymbol{\epsilon}_{ij}$ is a residual vector. For each simulated dataset, the predictors $\boldsymbol{w}_i$ are sampled from a multinomial distribution with probabilities set to the proportions of each group in the motivating data, random effects $\boldsymbol{b}_i$ are drawn from a $\mathrm{N}\left[0, \Sigma^{\boldsymbol{b}}\right]$ and residuals $\boldsymbol{\epsilon}_{ij}$ are drawn from a $\mathrm{N}\left[0, \Sigma^{\epsilon}\right]$.

The quantities $\boldsymbol{\beta}$, $\Sigma^{\boldsymbol{b}}$ and Σ^{ϵ} are chosen to resemble our motivating data. First, we focus on a subset of the full dataset that consists of the observed y-position curves from reaches to the target at 180°. In this subset, we find group-level average curves for each of the groups of interest; these become the coefficient functions in the rows of $\boldsymbol{\beta}$. For each subject in our subset, we find the subject-level average curve and subtract the corresponding group-level mean, then calculate the covariance $\Sigma^{\boldsymbol{b}}$ of these curves. Finally, for each curve in our subset we subtract the subject-level mean, then calculate the covariance Σ^{ϵ} of these curves. 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 $J_i = 5$. 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. 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 `pffrGLS()` function in the `refund R` package. This extends the the penalized function-on-function regression model assuming independent errors implemented in `pffr()` (Scheipl et al., 2013); generalized least squares is used to account for residual correlation in `pffrGLS()`, and a mixed model framework is used for parameter estimation. In a process similar to the GLS method described by Reiss et al. (2010) for cross-sectional function-on-scalar regression, we first fit the model assuming independence and use the residual curves to

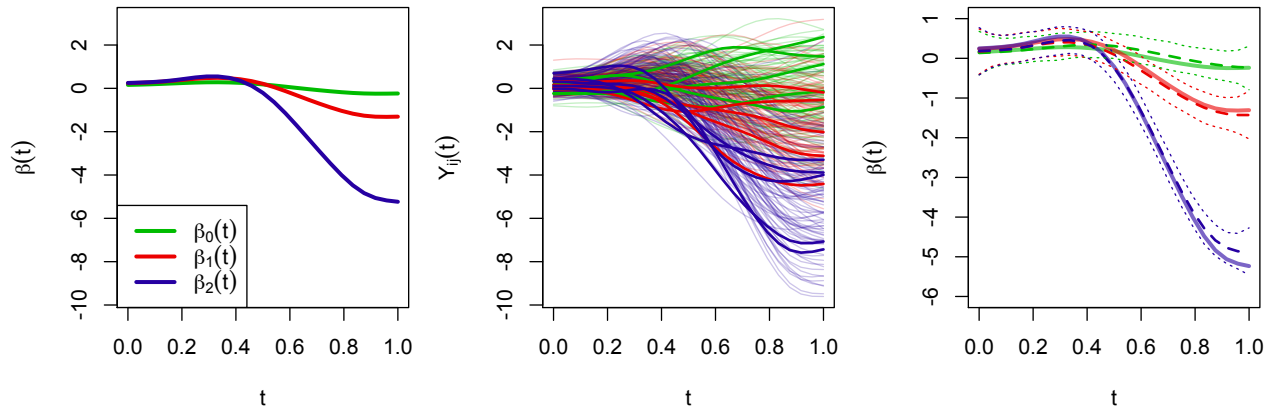

Figure 2: The left panel shows the three coefficient functions used to simulate data. The middle panel shows a complete simulated data set, with three subjects (one from each group) highlighted. The right panel shows the true coefficient functions, as well as their estimates and credible intervals derived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

estimate the covariance matrix for use in `pffrGLS()`. Although it is not described in a manuscript or preprint at the time of writing, to the best of our knowledge `pffrGLS()` represents the current state-of-the-art in function-on-scalar regression with subject-level random effects.

The left panels of Figure 3 show the integrated mean squared error IMSE $= \int \left(\hat{\beta}(t) - \beta(t)\right)^2 dt$ 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 outperform the mixed model approach, sometime substantially. The right panel of Figure 3 shows the computation time for each sample size and approach (simulations were executed in parallel on a compute cluster with Intel Xeon CPUs running at 2.30GHz; memory usage was 2, 4, and 6 GB for $I = 60$, 120 and 180, respectively). 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 nearly two hours. This discrepancy in computation time is in part due to the need to fit both two models (one using `pffr()` and one using `pffrGLS()`) for the mixed model; also, the code for the variational Bayes algorithm is tailored to the model at hand while `pffrGLS()` uses the more general `mgcv` package for estimating parameters. However, we also note that our implementation of `pffrGLS()` used

5 (rather than 10) basis functions. For $I = 180$, using 10 basis functions in `pffrGLS()` required 20,000 seconds – roughly 3.5 times longer than the Gibbs sampler.

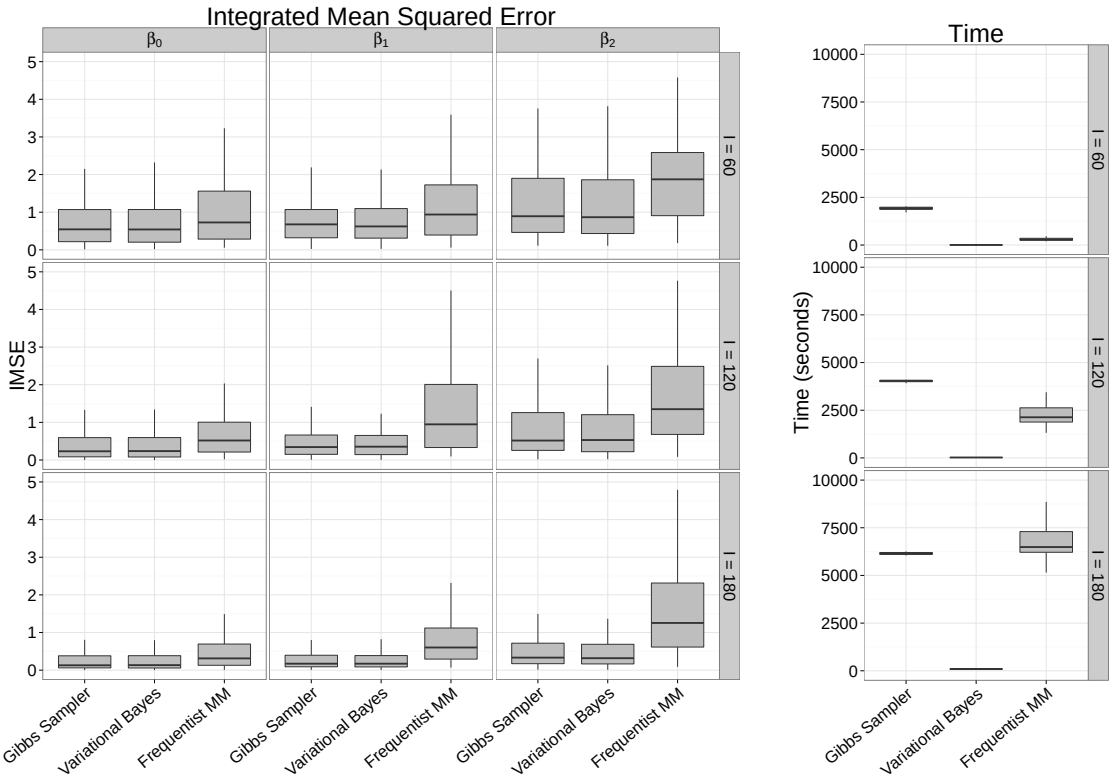

Figure 3: Simualtion results. The left panels show IMSE, defined as IMSE $= \int \left(\hat{\beta}(t) - \beta(t)\right) dt$, for each coefficient function, sample size, and estimation technique. The right panel show computation time for each sample size and estimation method.

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 slightly exceeds the nominal level and is often between .96 and .98. As expected, coverage improves as sample size increases. For the mixed model coverage is well below nominal levels for all coefficients and sample sizes; it is possible that the code is still under development and that future iterations may provide better inference. 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		
	$\beta_0(t)$	$\beta_1(t)$	$\beta_2(t)$	$\beta_0(t)$	$\beta_1(t)$	$\beta_2(t)$	$\beta_0(t)$	$\beta_1(t)$	$\beta_2(t)$
$I = 60$	0.98	0.98	0.98	0.97	0.97	0.96	0.43	0.47	0.43
$I = 120$	0.98	0.99	0.98	0.97	0.99	0.96	0.37	0.34	0.34
$I = 180$	0.97	0.99	0.97	0.96	0.98	0.96	0.40	0.32	0.27

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 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. The assumptions that underly the variational algorithm make in unsuitable for inference in our real-data application, and comparisons are made on the basis of percent variance explained. Therefore, after identifying a plausible final model, we estimate all model parameters using the complete Gibbs sampler described in Section 2.1 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 select hyperparameters as described in Section 2.3 and estimate 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. 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 target-only model using

$$\text{Relative PVE } = 100 \times \left[1 - \frac{\text{Var}\left[\text{vec}\left(Y - \hat{Y}_m\right)\right]}{\text{Var}\left[\text{vec}\left(Y - \hat{Y}_0\right)\right]}\right] \tag{6}$$

for models $m \in 1, \ldots, 7$, where Y is the matrix of observed trajectories and $\hat{Y}_m$ is the matrix of estimated trajectories based on fixed effects in Model m.

The poor performance of Model 1 compared to Model 2 indicates the importance of interaction between

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	4.6
3	FM^2×Tar	32	5.2
4	FM×Tar×Hand	32	8.4
5	FM×Tar×Aff	32	8.8
6	FM^2×Tar×Aff	64	10.1
7	FM×Tar×Hand×Aff	64	11.9

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; "FM" is the continuous Fugl-Meyer score; "+" 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 (6)).

target and impairment severity, due to the target-specific direction of the effect of stroke and to differing levels of biomechanical difficulty. Models 3, 4 and 5 build on Model 2 by adding a quadratic effect of the Fugl-Meyer score and interactions with hand used and affectedness, respectively. The relatively small improvement of the quadratic model may support an assumption of linearity, and the comparability of Models 4 and 5 may be due to the fact that, in our dataset, the affected hand was more likely to be the dominant hand. Models 6 and 7 build on Model 5 by allowing a quadratic effect of the Fugl-Meyer score and an interaction with the hand used. Again, the improvement from using a quadratic term is models compared the effect of the added interaction. 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 deviations, and trajectory-level variation. In Section 4.2 we discuss inference for Model 7.

Figure 4 compares the estimated fixed effects from Models 2, 5, and 7. For each model (in rows), we show the estimated mean trajectories in an affected dominant hand for Fugl-Meyer scores 66, 51, 36, and 21 (in columns). In Model 2, the effect of increasing stroke severity is assumed to be the same in both the affected and unaffected hand. This is unlikely given that our data set consists of patients with unilateral stroke. Model 5 estimates separate effects of stroke severity for the affected and unaffected arm. Comparing Models 2 and 5 for an affected arm in Figure 4, Model 5 indicates large effects of increasing stroke severity; for an unaffected arm (not shown) Model 5 indicates small or no effect. Model 7 additionally separates

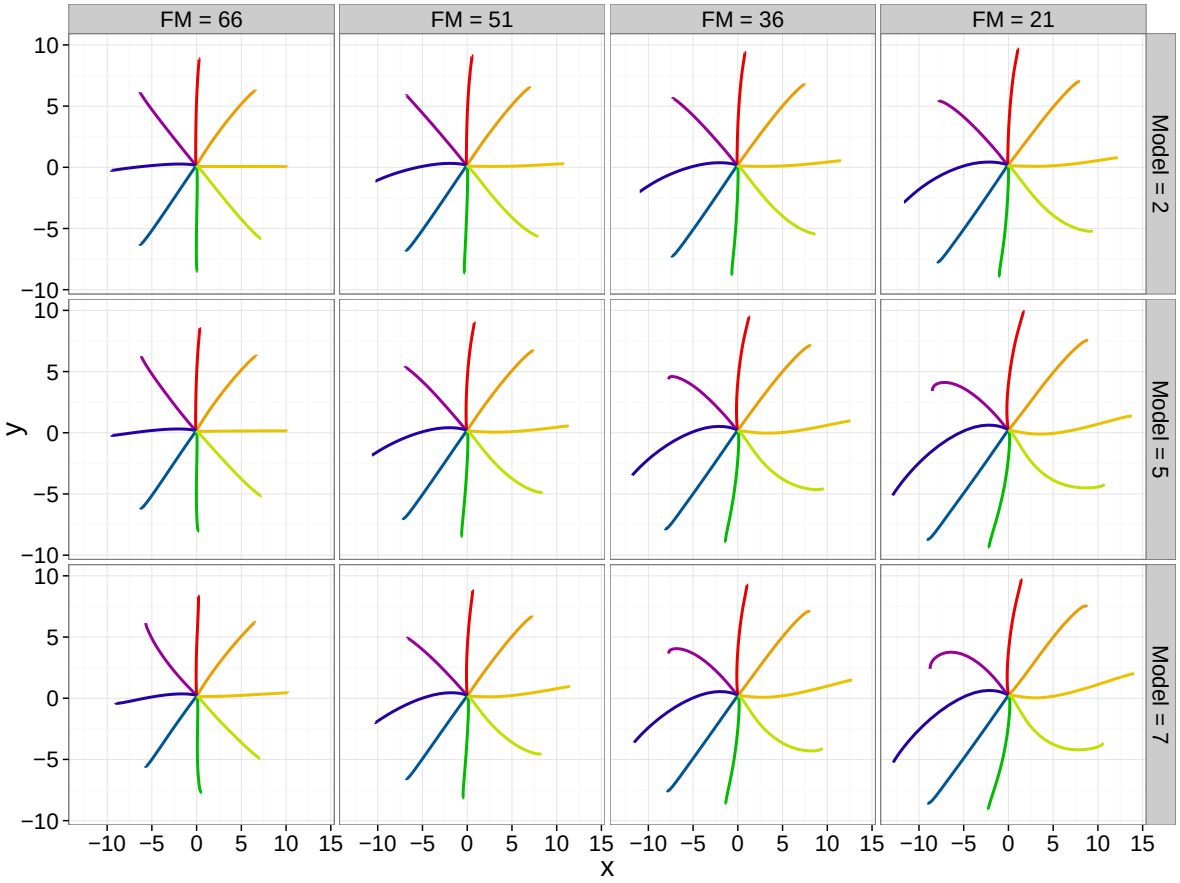

Figure 4: Estimated motions for an affected dominant arm with Fugl-Meyer scores 66, 51, 36, and 21 (in columns) under Models 2, 5, and 7 (in rows). An interactive version of this Figure is available on the first author's website.

the affected dominant from affected non-dominant hands, with the scientific interpretation that a stroke of the same severity could affect these limbs differently. The differences between Models 5 and 7 are subtle for affected dominant arms, but more noticeable for unaffected and non-dominant arms. An interactive version of Figure 4 is available on the first author's website.

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.

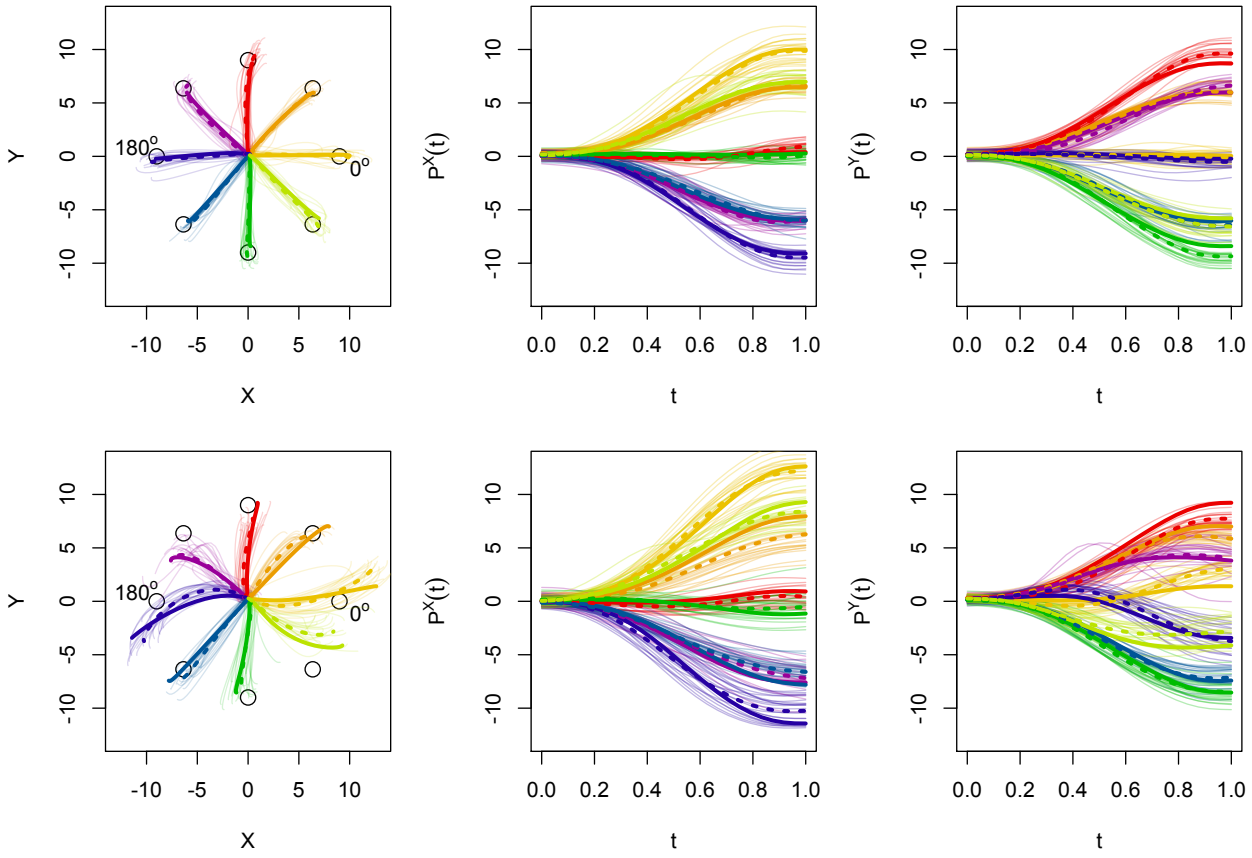

Figure 5: Observed data (faint solid lines) overlayed with estimated fixed (bold solid lines) and random (bold dashed lines) effects. Data for two subjects are shown: in the top row, the dominant hand of a control subject, and in the bottom row the affected dominant hand of a severe stroke patient. Data for these subjects appear 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 five chains with random starting values, setting chain length to 2,000 iterations and discarding the first 500 as burn-in. Hyperparameters and initial values were chosen as described in Section 2.3. Analyses assessing the sensitivity to hyperparameter values appear in the web-based supplementary materials. Computations took 2.5 days per chain on a Intel Xeon CPUs running at 2.30GHz with 8 Gb

memory, which emphasizes the importance of a fast approximation for data exploration and model building.

Figure 6 shows the estimated effect of a ten unit decrease in Fugl-Meyer score in a dominant hand affected by stroke for all target directions. The top and bottom rows show the marginal effect on the X and Y position curves, respectively. Panels show the posterior mean as a bold curve, and a sample from the posterior as translucent curves. These results show clear, significant effects to all targets, verifying that increasing stroke severity has a systematic effect on motor control. For motions to the target at 0°, increasing stroke severity leads to over-reach (increases in the X direction) as well as a vertical shift (increase in the Y position). Other targets can be interpreted similarly. The largest effects are generally in directions that require multi-joint coordination, and are thus more biomechanically difficult.

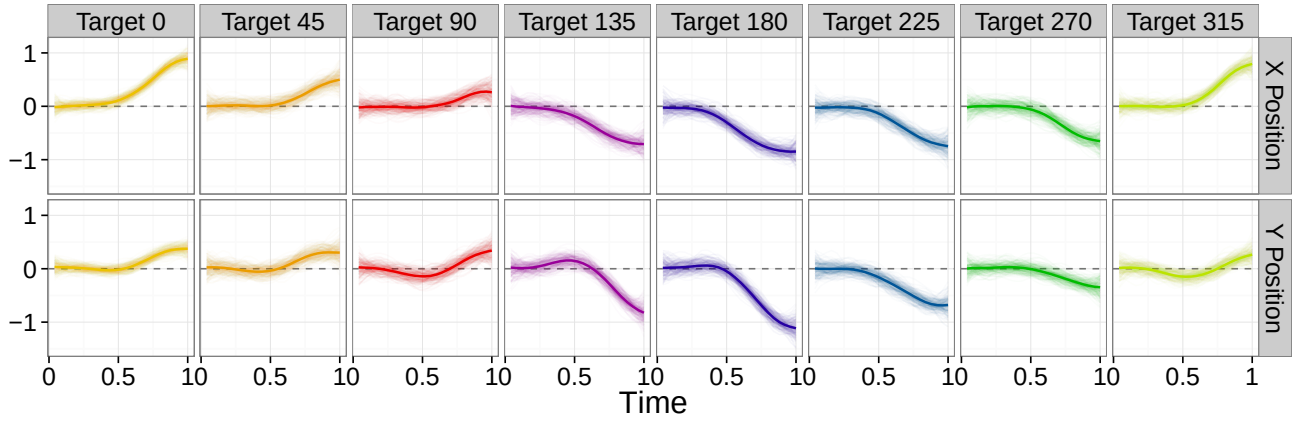

Figure 6: Estimated effect of a 10-unit decrease in Fugl-Meyer score in an affected dominant hand. Effects to eight targets appear in columns; the effect in the X and Y positions appear in rows. Posterior means are shown as bold curves; a posterior sample is shown as translucent curves.

Figure 7 illustrates the quality and convergence of our Markov chains. The top row shows plots for a representative spline coefficient in a fixed effect function $\beta_k(t)$; the bottom row shows the variance component $\sigma_{W,k}$ that controls the degree of penalization in this function. In each row, the left panel shows the five posterior chains for the parameter of interest, started from randomly chosen values with the burn-in period shaded; this shows that chains converge quickly and there is little sensitivity to the choice of starting value. An autocorrelation function is shown in the second panel and indicates low autocorrelation. The third panel of each row shows the convergence criterion of Gelman and Rubin (1992) as a function of iteration number. Values near 1 indicate convergence, which is typically attained after only a few hundred iterations. Finally, at right we show the posterior mean residual covariance surface to illustrate

the correlation within X and Y position residuals and the correlation between them.

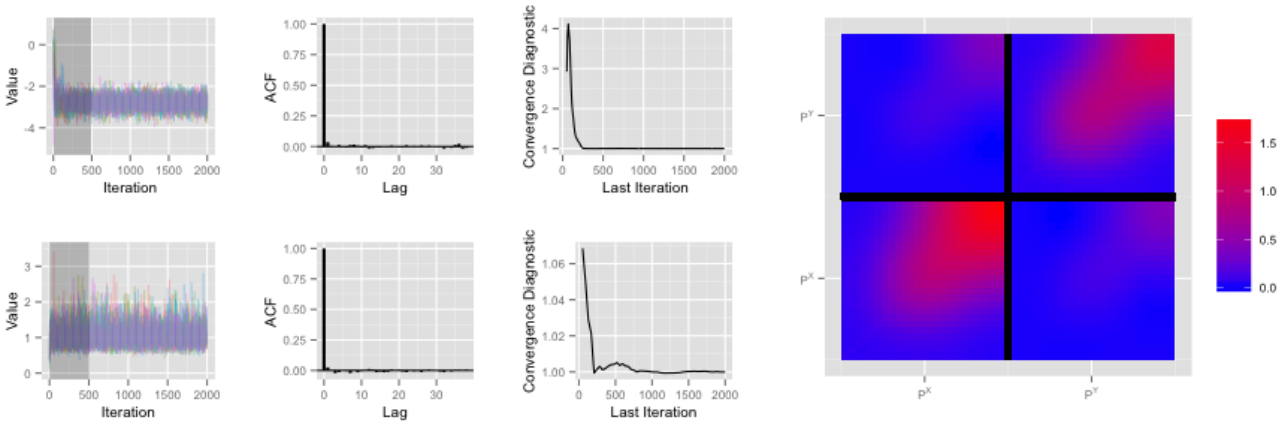

Figure 7: Diagnostic and convergence criteria for representative parameters are show in the grid on the left. The top row shows a spline coefficient in a fixed effect function and the bottom row shows the variance component associated with that coefficient function. The columns in this grid show, from left to right, the five chains with random start values, the autocorrelation function, and Gelman and Rubin's diagnostic criterion. The plot at right show the residual covariance surface.

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.

Variational approximations are often quite inaccurate for the construction of credible intervals due to the factorization assumptions and use of parametric densities to approximate the posterior distribution. In that sense, the good inferential performance found in simulations is perhaps surprising. In part this performance is due to the hierarchical recentering in model (4), in which observed curves are centered around subject effects, subject effects are centered around fixed effects, and fixed effects are centered around zero. This formulation helps to decrease the posterior correlation between fixed and random effects and improves the quality of the approximation (and, importantly, the mixing for the Gibbs sampler). A direct implementation of model (3), which centers observed curves around the combination of fixed and random

effects and centers both fixed and random effects around zero, had similar estimation accuracy but much poorer inference. Despite these results, we recommend caution when using the variational approximation for inference, and prefer this method for model building, cross-validation of results, bootstrapping, or other computationally intensive procedures. Finally, we note that the good performance of the variational approximation suggests that other approximate methods might be suitable for this model and should be explored carefully.

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. Additional work will examine the replicability of these findings in larger studies and quantify possible overfitting.

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. Implementing and testing our methods for the case that curves are observed on a sparse grid would broaden the class of problems to which these methods can be applied. New general-purpose programming languages for Bayesian analysis (including sampling and optimization) will facilitate the implementation of more classes of prior distributions, potentially leading to better models and reducing the reliance on convenient priors (Stan Development Team, 2013). 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 thank John Krakauer for his scientific insight and guidance, and Johnny Liang, Sophia Ryan, and Sylvia Huang for their assistance in data collection. The first author's research was supported in part by Award R01HL123407 from the National Heart, Lung, and Blood Institute and by Award R21EB018917 from the National Institute of Biomedical Imaging and Bioengineering.

References

Baladandayuthapani, V., Ji, Y., Talluri, R., Nieto-Barajas, L. E., and Morris, J. S. "Bayesian Random Segmentation Models to Identify Shared Copy Number Aberrations for Array CGH Data." Journal of the American Statistical Association, 105:1358–1375 (2010).

Baladandayuthapani, V., Mallick, B., Young Hong, M., Lupton, J., Turner, N., and Carroll, R. J. "Bayesian Hierarchical Spatially Correlated Functional Data Analysis with Application to Colon Carcinogenesis." Biometrics, 64:64–73 (2007).

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).

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

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).

Eilers, P. H. C. and Marx, B. D. "Flexible smoothing with B-splines and penalties." Statistical Science, 11:89–121 (1996).

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).

Gelfand, A., Sahu, S. K., and Carlin, B. "Efficient Parameterizations for Generalized Linear Mixed Models,"." Biometrika, 82:479–488 (1995).

Gelman, A. "Prior distributions for variance parameters in hierarchical models." Bayesian Analysis, 1:515–533 (2006).

Gelman, A. and Rubin, D. B. "Inference from iterative simulation using multiple sequences." Statistical science, 7:457–472 (1992).

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).

Montagna, S., Tokdar, S. T., Neelson, B., and Dunson, D. B. "Bayesian Latent Factor Regression for Functional and Longitudinal Data." Biometrics, 69:10641073 (2012).

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

Morris, J. S., Vannucci, M., Brown, P. J., and Carroll, R. J. "Wavelet-Based Nonparametric Modeling of Hierarchical Functions in Colon Carcinogenesis." Journal of the American Statistical Association, 98:573–583 (2003).

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

Ruppert, D. "Selecting the Number of Knots for Penalized Splines." Journal of Computational and Graphical Statistics, 11:735–757 (2002).

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).

Stan Development Team. Stan Modeling Language User's Guide and Reference Manual, Version 1.3 (2013). URL http://mc-stan.org/

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).

Yang, R. and Berger, J. O. "Estimation of a covariance matrix using the reference prior." The Annals of Statistics, 22:1195–1211 (1994).

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 supplementary material consists of the following appendices: A, containing a graphical representation of our model and a brief discussion of induced factorings; B, containing sensitivity analyses for the choice of hyperparameters in our full Bayesian analysis; C, containing derivations of full conditional distributions for the Gibbs sampler; and D, containing derivations optimal densities for the variational Bayes algorithm. Throughout we consider 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 $[Y_{ij}(t), \boldsymbol{w}_i]$ 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 are interested in estimating the parameters in

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

In this model, B_W is the matrix of coefficients for fixed effects and B_Z is the matrix of coefficients for random subject effects; Σ is the residual covariance matrix, and the variance components $\left\{\sigma_{W,k}^2\right\}, \sigma_Z^2$ control the amount of smoothness in the fixed and random effects. Additionally, Y, W, Z, Θ, and P are the observed outcomes, the fixed and random effect design matrices, the b-spline basis matrix, and the known penalty matrix, respectively.

A Graphical model

The joint distribution corresponding to (A.1) is

$$\left[\prod_{i=1}^{I}\left[\prod_{j=1}^{J_i} p(Y_{ij}|B_{Z,i},\Sigma)\right] p(B_{Z,i}|B_W,\sigma_Z^2)\right]\cdot\left[\prod_{k=0}^{p} p(B_{W,k}|\sigma_{W,k}^2)p(\sigma_{W,k}^2)\right]\cdot p(\Sigma)\cdot p(\sigma_Z^2).$$

Here we omit hyperparameters for the densities for the covariance matrix Σ and variances $\left\{\sigma_{W,k}^2\right\}, \sigma_Z^2$. Also note that the "zeroth" column of B_W and the variance $\sigma_{W,0}^2$ correspond to the intercept $\beta_0(t)$. A graphical representation of this model is shown in Figure A.1.

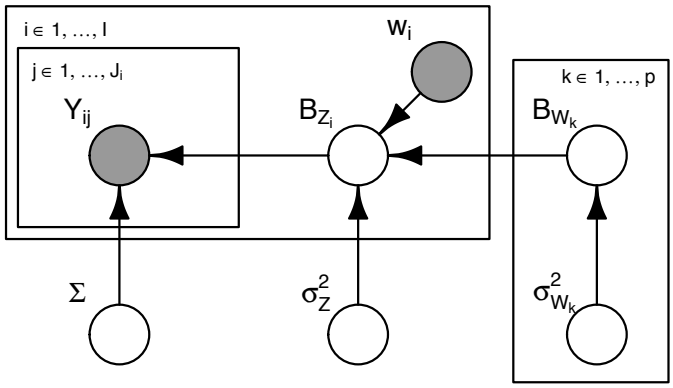

Figure A.1: Graphical illustration of model (A.1). Shaded nodes represent observed data; blank nodes denote inferred parameters; rectangles or plates denote indexing over a set of variables. For example, the plate surrounding only Y_{ij} indicates that for subject i, observations are indexed by j and are independent given parameters outside this plate.

Our variational approach assumes that the posterior distribution $p\left(B_Z, B_W, \sigma_{W,0}^2, \ldots, \sigma_{W,p}^2, \sigma_Z^2, \Sigma|Y\right)$ can be approximated using

$$q\left(B_Z, B_W, \sigma_{W,0}^2, \ldots, \sigma_{W,p}^2, \sigma_Z^2, \Sigma\right) \quad = \quad q(B_Z)q(B_W)q(\sigma_{W,0}^2, \ldots, \sigma_{W,p}^2, \sigma_Z^2, \Sigma)$$

where the functions q are distinguished by their argument rather than by subscript l. The additional factorizations $q(B_Z) = \prod_{i=1}^{I} q(B_{Z,i})$ and $q(\sigma_{W,0}^2, \ldots, \sigma_{W,p}^2, \sigma_Z^2, \Sigma) = \left(\prod_{k=0}^{p} q(\sigma_{W,k}^2)\right) q(\sigma_Z^2)q(\Sigma)$ follow from the result

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

and Figure A.1. In particular, the columns of B_Z are conditionally independent given B_W and the variance components are all conditionally independent given the remaining model parameters. These observations lead to our final factorization:

$$q\left(B_Z, B_W, \sigma^2_{W,0}, \ldots, \sigma^2_{W,p}, \sigma^2_Z, \Sigma\right) \quad = \quad \left(\prod_{i=1}^{I} q(B_{Z,i})\right) q(B_W) \left(\prod_{k=0}^{p} q(\sigma^2_{W,k})\right) q(\sigma^2_Z) q(\Sigma).$$

B Sensitivity Analyses

In Section 2.3 of the manuscript, we describe a procedure for choosing hyperparameters. This procedure is based on an analysis of the data, and is developed to avoid default "uninformative" choices like $a_z = b_z = a_{W,k} = b_{W,k} = .001$, which places a large prior mass near zero for all variance components and can lead to overshrinkage of fixed and random effects toward zero. However there is the potential for sensitivity to these choices, and in this section we repeat our full data analysis using other choices for hyperparameters.

In this analysis, we set $a_z = 10, b_z = 2.5$, and $a_{W,k} = b_{W,k} = 1$ for all k as the hyperparameters for inverse-gamma densities. We set $\nu = 10$ and $\Psi = 10I$, where I is an identity matrix. For reference, in the analysis presented in the manuscript, we set $a_z = 4250$ and $b_z = 2600$; $a_{W,k}$ was set to 5 and $b_{W,k}$ ranged between .001 and 100, with typical values near .05; we set $\nu = 19053$ and Ψ was constructed using an FPCA decomposition. The choices of hyperparameters in this appendix result in more diffuse priors and encode relatively little information. Using these hyperparameters, we repeated our analysis by using five chains with random starting values, with each chain consisting of 2,000 iterations and discarding the first 500 as burn-in.

Figure A.2 shows the estimated effect of a 10-unit change in Fugl-Meyer score in a dominant hand affected by stroke for all target directions; this recreates Figure 6 from the main manuscript. Estimates of effects are very similar in both analysis, indicated that choices of hyperparameters can relax the degree of

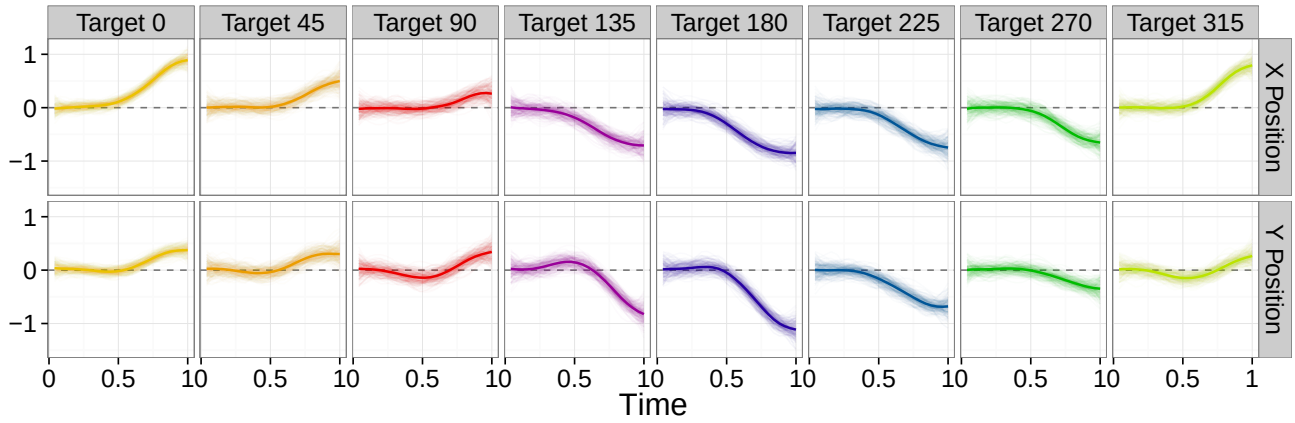

Figure A.2: Estimated effect of a 10-unit decrease in Fugl-Meyer score in an affected dominant hand. Effects to eight targets appear in columns; the effect in the X and Y positions appear in rows. Posterior means are shown as bold curves; a posterior sample is shown as translucent curves. Figure is based on hyperparameters chosen for a sensitivity analysis.

prior information with relatively little effect on estimation. The posterior sample in all panels of Figure A.2 tends to be more variable than the corresponding panel of Figure 6 in the manuscript, but not dramatically so.

Figure A.3 recreates Figure 7 in the main manuscript using the hyperparameter choices given above. The top row shows plots for the same representative spline coefficient in a fixed effect function $\beta_k(t)$ as in Figure 7; the bottom row shows the variance component $\sigma_{W,k}$ that controls the degree of penalization in this function. In each row, the left panel shows the five posterior chains; the second panel shows an autocorrelation function; and the third panel of each row shows the convergence criterion of Gelman and Rubin (1992) as a function of iteration number. Convergence is somewhat slower using the current hyperparameters, but is still reasonable. As seen in Figure A.2, distributions are a bit more variable, likely due to the spreading of prior mass for variance components.

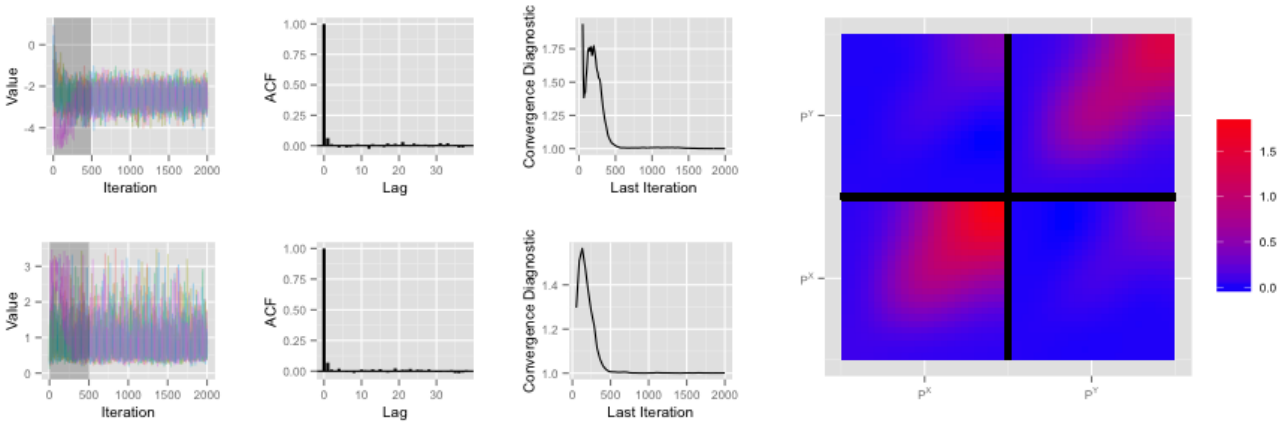

Figure A.3: Diagnostic and convergence criteria for representative parameters are show in the grid on the left. The top row shows a spline coefficient in a fixed effect function and the bottom row shows the variance component associated with that coefficient function. The columns in this grid show, from left to right, the five chains with random start values, the autocorrelation function, and Gelman and Rubin's diagnostic criterion. The plot at right show the residual covariance surface. Figure is based on hyperparameters chosen for a sensitivity analysis.

C Gibbs Sampler

In this section we provide derive full conditional distributions for the parameters in model (A.1). Let $\mathbf{1}_m$ be a length m vector of 1's and Y_i denote the rows of Y for subject i. Also recall that $n = \sum_{i=1}^{I} J_i$ is the total number of observations.

- The full conditional distribution $p(B_{Z,i}|\text{rest})$ is

$$\begin{aligned}
p(B_{Z,i} \mid \text{rest}) &\propto p(Y_i \mid B_{Z,i}, \Sigma) p(B_{Z,i} \mid \sigma_Z^2, B_W) \\
&\propto \exp\left[-\frac{1}{2}\left\{(\text{vec}\left(Y_i^T\right) - (\mathbf{1}_{J_i} \otimes \Theta) B_{Z,i})^T (I_{J_i} \otimes \Sigma^{-1})(\text{vec}\left(Y_i^T\right) - (\mathbf{1}_{J_i} \otimes \Theta) B_{Z,i})\right.\right. \\
&\qquad \left.\left. + (B_{Z_i} - B_W \boldsymbol{w}_i^T)^T (\frac{1}{\sigma_Z^2} P)(B_{Z_i} - B_W \boldsymbol{w}_i^T)\right\}\right] \\
&\propto \exp\left[-\frac{1}{2}\left\{B_{Z,i}^T \left((\mathbf{1}_{J_i} \otimes \Theta)^T (I_{J_i} \otimes \Sigma^{-1})(\mathbf{1}_{J_i} \otimes \Theta) + (\frac{1}{\sigma_Z^2} P)\right) B_{Z,i}\right.\right. \\
&\qquad \left.\left. -2\left(\text{vec}\left(Y_i^T\right)^T (I_{J_i} \otimes \Sigma^{-1})(\mathbf{1}_{J_i} \otimes \Theta) + (B_W \boldsymbol{w}_i^T)^T (\frac{1}{\sigma_Z^2} P)\right) B_{Z,i}\right\}\right] \\
&\propto \exp\left\{-\frac{1}{2}\left(B_{Z,i} - \mu_{B_{Z,i}}\right)^T \Sigma_{B_{Z,i}}^{-1} \left(B_{Z,i} - \mu_{B_{Z,i}}\right)\right\}
\end{aligned}$$

  where

$$\Sigma_{B_{Z,i}} = \left((\mathbf{1}_{J_i} \otimes \Theta)^T (I_{J_i} \otimes \Sigma^{-1})(\mathbf{1}_{J_i} \otimes \Theta) + \frac{1}{\sigma_Z^2} P\right)^{-1}$$

  and

$$\mu_{B_{Z,i}} = \Sigma_{B_{Z,i}} \left((\mathbf{1}_{J_i} \otimes \Theta)^T (I_{J_i} \otimes \Sigma^{-1}) \text{vec}\left(Y_i^T\right) + \frac{1}{\sigma_Z^2} P(B_W \boldsymbol{w}_i^T)\right).$$

  Thus $[B_{Z,i} \mid \text{rest}]$ follows a multivariate normal distribution with mean and variance as described above.

- Derivation of the full conditional distribution $p(\text{vec}\,(B_W)\,|\text{rest})$ follows very similarly to the above.

In particular,

$$\begin{aligned}
p(\text{vec}\,(B_W) \mid \text{rest}) &\propto p(B_Z \mid B_W, \sigma_Z^2) p(\text{vec}\,(B_W) \mid \{\sigma_{W,k}^2\}) \\
&\propto \exp\left[-\frac{1}{2}\left\{(\text{vec}\,(B_Z) - (W \otimes I_{K_\theta})\text{vec}\,(B_W))^T (\frac{1}{\sigma_Z^2} I_I \otimes P)(\text{vec}\,(B_Z) - (W \otimes I_{K_\theta})\text{vec}\,(B_W))\right.\right. \\
&\qquad \left.\left. + \text{vec}\,(B_W)^T \left(\text{diag}\left(1/\sigma_{W,k}^2\right) \otimes P\right)\text{vec}\,(B_W)\right\}\right] \\
&\propto \exp\left\{-\frac{1}{2}\left(\text{vec}\,(B_W) - \mu_{B_W}\right)^T \Sigma_{B_W}^{-1} \left(\text{vec}\,(B_W) - \mu_{B_W}\right)\right\}
\end{aligned}$$

where

$$\Sigma_{B_W} = \left(\frac{1}{\sigma_Z^2}(W \otimes I_{K_\theta})^T (I_I \otimes P)(W \otimes I_{K_\theta}) + \text{diag}\left(\frac{1}{\sigma_{W,k}^2}\right) \otimes P\right)^{-1}$$

and

$$\mu_{B_W} = \Sigma_{B_W}\left(\frac{1}{\sigma_Z^2}(W \otimes I_{K_\theta})^T (I_I \otimes P)\text{vec}\,(B_Z)\right).$$

Thus $[\text{vec}\,(B_W) \mid \text{rest}]$ follows a multivariate normal distribution with mean and variance given above; the conditional distribution of B_W is given by inverting the $\text{vec}\,(\cdot)$ operation.

- The full conditional distribution $p(\sigma_Z^2|\text{rest})$ is

$$\begin{aligned}
p(\sigma_Z^2 \mid \text{rest}) &\propto p(\sigma_Z^2) \prod_{i=1}^{I} p(B_{Z,i} \mid \sigma_Z^2) \\
&\propto (\sigma_Z^2)^{-a_Z - 1} \exp\left\{-\frac{b_Z}{\sigma_Z^2}\right\} \prod_{i=1}^{I} (\sigma_Z^2)^{-K_\theta/2} \exp\left\{-\frac{1}{2}(B_{Z,i} - B_W \boldsymbol{w}_i^T)^T (\frac{1}{\sigma_Z^2} P)(B_{Z,i} - B_W \boldsymbol{w}_i^T)\right\} \\
&= (\sigma_Z^2)^{-a_Z - \frac{I * K_\theta}{2} - 1} \exp\left\{-\frac{1}{\sigma_Z^2}\left(b_z + \frac{1}{2}\sum_{i=1}^{I}(B_{Z,i} - B_W \boldsymbol{w}_i^T)^T (\frac{1}{\sigma_Z^2} P)(B_{Z,i} - B_W \boldsymbol{w}_i^T)\right)\right\}.
\end{aligned}$$

Thus $[\sigma_Z^2 \mid \text{rest}]$ is distributed IG $\left[a_z + \frac{I * K_\theta}{2}, b_z + \frac{1}{2}\sum_{i=1}^{I}\left(B_{Z,i} - B_W \boldsymbol{w}_i^T\right)^T P \left(B_{Z,i} - B_W \boldsymbol{w}_i^T\right)\right]$.

- The derivation of the full conditional distribution for each of the $\sigma_{W,k}^2$ is the same, and is similar to that of σ_Z^2. For any k, the full conditional $p(\sigma_{W,k}^2|\text{rest})$ is

$$p(\sigma_{W,k}^2 \mid \text{rest}) \propto p(\sigma_{W,k}^2) p(B_{W,k} \mid \sigma_{W,k}^2).$$

Straightforward computation shows that $[\sigma_{W,k}^2 \mid \text{rest}]$ is distributed IG $\left[a_{W,k} + \frac{K_\theta}{2}, b_{W,k} + \frac{1}{2} B_{W,k}^T P B_{W,k}\right]$.

- For a given value of B_Z, let $R = Y - ZB_Z^T\Theta^T$ be the matrix containing residual curves evaluated over the discrete grid of observation points such that $\boldsymbol{R}_{ij}$ contains the jth residual vector for subject i. The full conditional distribution $p(\Sigma|\text{rest})$ is

$$\begin{aligned} p(\Sigma \mid \text{rest}) &\propto p(Y \mid B_Z, \Sigma)p(\Sigma) \\ &\propto |\Sigma|^{-\frac{n}{2}} \exp\left\{-\frac{1}{2}\sum_{i=1}^{I}\sum_{j=1}^{J_i}\boldsymbol{R}_{ij}\Sigma^{-1}\boldsymbol{R}_{ij}^T\right\} \cdot |\Sigma|^{\frac{\nu+D+1}{2}} \exp\left\{-\frac{1}{2}\text{tr}\left[\Psi\Sigma^{-1}\right]\right\}. \end{aligned}$$

Note that $\sum_{i=1}^{I}\sum_{j=1}^{J_i}\boldsymbol{R}_{ij}\Sigma^{-1}\boldsymbol{R}_{ij}^T = \text{tr}\left[\boldsymbol{R}\Sigma^{-1}\boldsymbol{R}^T\right] = \text{tr}\left[\boldsymbol{R}^T\boldsymbol{R}\Sigma^{-1}\right]$. Combining terms, the full conditional is

$$\begin{aligned} p(\Sigma \mid \text{rest}) &\propto p(Y \mid B_Z, \Sigma)p(\Sigma) \\ &\propto |\Sigma|^{-\frac{\nu+n+D+1}{2}} \exp\left\{-\frac{1}{2}\text{tr}\left[(\Psi + \boldsymbol{R}^T\boldsymbol{R})\Sigma^{-1}\right]\right\} \end{aligned}$$

so that $[\Sigma \mid \text{rest}]$ is distributed IW $\left[\nu + \sum_i J_i, \Psi + \left(Y - ZB_Z^T\Theta^T\right)^T\left(Y - ZB_Z^T\Theta^T\right)\right]$.

D Variational Bayes Algorithm

In this section we derive the optimal densities and the iterative algorithm 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.

First, we recall that given a partition $\{\boldsymbol{\phi}_1, \ldots, \boldsymbol{\phi}_L\}$ of the parameter space $\boldsymbol{\phi}$, the explicit solution for $q(\boldsymbol{\phi}_l)$, $1 \leq l \leq L$ has the form

$$q_l^*(\boldsymbol{\phi}_l) \quad \propto \quad \exp\left[E_{\boldsymbol{\phi}_{-l}} \log p(\boldsymbol{\phi}_l|\text{rest})\right]; \, 1 \leq l \leq L \tag{A.2}$$

where rest $\equiv \{Y, \boldsymbol{\phi}_1, \ldots, \boldsymbol{\phi}_{l-1}, \boldsymbol{\phi}_{l+1}, \ldots, \boldsymbol{\phi}_L\}$. Notationally, for a scalar random variable ϕ, let

$$\begin{aligned} \mu_{q(\phi)} &\equiv \mathrm{E}_q[\phi] = \int \phi_0 q(\phi_0)\, d\phi_0 \\ \sigma_{q(\phi)} &\equiv \mathrm{Var}_q[\phi] = \int (\phi_0 - \mathrm{E}[\phi])^2 q(\phi_0)\, d\phi_0 \end{aligned}$$

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

- For $B_{Z,i}$, we use the full conditional derivation in Section C to find that

$$\begin{aligned} q^*(B_{Z,i}) &\propto \exp\left[E_{-B_{Z,i}} \log p(B_{Z,i}|\text{rest})\right] \\ &\propto \exp\left[E_{-B_{Z,i}}\left[-\frac{1}{2}\left\{\mathrm{vec}\,(B_{Z,i})^T\left((\mathbf{1}_{J_i} \otimes \Theta)^T(I_{J_i} \otimes \Sigma^{-1})(\mathbf{1}_{J_i} \otimes \Theta) + (\frac{1}{\sigma_Z^2}P)\right)\mathrm{vec}\,(B_{Z,i})\right.\right.\right. \\ &\qquad \left.\left.\left. -2\left(\mathrm{vec}\,(Y_i)^T(I_{J_i} \otimes \Sigma^{-1})(\mathbf{1}_{J_i} \otimes \Theta) + (B_W \boldsymbol{w}_i^T)^T(\frac{1}{\sigma_Z^2}P)\right)B_{Z,i}\right\}\right]\right] \\ &\propto \exp\left\{-\frac{1}{2}\left(B_{Z,i} - \mu_{q(B_{Z,i})}\right)^T \Sigma_{q(B_{Z,i})}^{-1}\left(B_{Z,i} - \mu_{q(B_{Z,i})}\right)\right\} \end{aligned}$$

where

$$\Sigma_{q(B_{Z,i})} = \left((\mathbf{1}_{J_i} \otimes \Theta)^T(I_{J_i} \otimes \mu_{q(\Sigma^{-1})})(\mathbf{1}_{J_i} \otimes \Theta) + \mu_{q(1/\sigma_Z^2)}P\right)^{-1}$$

and

$$\mu_{q(B_{Z,i})} = \Sigma_{q(B_{Z,i})}\left((\mathbf{1}_{J_i} \otimes \Theta)^T(I_{J_i} \otimes \mu_{q(\Sigma^{-1})})\mathrm{vec}\,(Y_i) + \mu_{q(1/\sigma_Z^2)}P(\mu_{q(B_W)}\boldsymbol{w}_i^T)\right).$$

Thus the optimal density $q^*(B_{Z,i})$ is multivariate normal with mean and variance as specified above.

- The derivation for B_W is similar to that for $B_{Z,i}$; specifically, we use the full conditional derivation in Section C to find that

$$\begin{aligned} q^*(B_W) &\propto \exp\left[E_{-B_W} \log p(B_W|\text{rest})\right] \\ &\propto \exp\left\{-\frac{1}{2}\left(\text{vec}\left(B_W\right) - \mu_{q(B_W)}\right)^T \Sigma^{-1}_{q(B_W)} \left(\text{vec}\left(q(B_W)\right) - \mu_{B_W}\right)\right\} \end{aligned}$$

where

$$\Sigma_{q(B_W)} = \left(\mu_{q(1/\sigma^2_Z)}(W \otimes I_{K_\theta})^T(I_I \otimes P)(W \otimes I_{K_\theta}) + \text{diag}\left(\mu_{q(1/\sigma^2_{W,k})}\right) \otimes P\right)^{-1}$$

and

$$\mu_{q(B_W)} = \Sigma_{q(B_W)}\left(\mu_{q(1/\sigma^2_Z)}(W \otimes I_{K_\theta})^T(I_I \otimes P)\text{vec}\left(\mu_{q(B_Z)}\right)\right).$$

Thus the optimal density $q^*(B_W)$ is multivariate normal with mean and variance as specified above.

- For σ^2_Z, we find that

$$\begin{aligned} q^*(\sigma^2_Z) &\propto \exp\left[E_{-\sigma^2_Z} \log p(\sigma^2_Z|\text{rest})\right] \\ &\propto \exp\left\{-\left(a_Z + \frac{I * K_\theta}{2} + 1\right)\log(\sigma^2_Z)\right. \\ &\qquad \left. -\frac{1}{\sigma^2_Z}\text{E}_{-\sigma^2_Z}\left((b_z + \frac{1}{2}\sum_{i=1}^{I}(B_{Z,i} - B_W \boldsymbol{w}_i^T)^T(P)(B_{Z,i} - B_W \boldsymbol{w}_i^T)\right)\right\} \\ &= (\sigma^2_Z)^{-a_Z - \frac{I*K_\theta}{2} - 1} \exp\left\{-\frac{1}{\sigma^2_Z} b_{q(\sigma^2_Z)}\right\} \end{aligned}$$

where

$$\begin{aligned} b_{q(\sigma^2_Z)} = b_Z + \frac{1}{2}\sum_i \Big(&\mu^T_{q(B_{Z,i})} P \mu_{q(B_{Z,i})} + \text{tr}\left[P\Sigma_{q(B_{Z,i})}\right] - 2\boldsymbol{w}_i \mu^T_{q(B_W)} P \mu_{q(B_{Z,i})} \\ &+ \boldsymbol{w}_i \mu^T_{q(B_W)} P \mu_{q(B_W)} \boldsymbol{w}_i^T + \boldsymbol{w}_i \text{diag}\left(\text{tr}\left[P\Sigma_{q(B_{W,k})}\right]\right)\boldsymbol{w}_i^T\Big). \end{aligned}$$

Thus the optimal density $q^*(\sigma^2_Z)$ is IG $\left[a_Z + \frac{I*K_\theta}{2}, b_{q(\sigma^2_Z)}\right]$. Note that the term $\mu_{q(1/\sigma^2_Z)}$ appearing in the optimal densities $q^*(B_{Z,i})$ and $q^*(B_W)$ is equal to $\frac{a_Z + \frac{I*K_\theta}{2}}{b_{q(\sigma^2_Z)}}$.

- The derivation for the $\sigma^2_{W,k}$ is similar to that of σ^2_Z. Specifically,

$$\begin{aligned} q^*(\sigma^2_{W,k}) &\propto \exp\left[E_{-\sigma^2_{W,k}} \log p(\sigma^2_{W,k}|\text{rest})\right] \\ &\propto \exp\left\{-\left(a_{W,k}+\frac{K_\theta}{2}+1\right)\log(\sigma^2_{W,k}) - \frac{1}{\sigma^2_{W,k}}\mathrm{E}_{-\sigma^2_{W,k}}\left((b_{W,k}+\frac{1}{2}B^T_{W,k}PB_{W,k}\right)\right\} \\ &= (\sigma^2_{W,k})^{-a_{W,k}-\frac{K_\theta}{2}-1}\exp\left\{-\frac{1}{\sigma^2_{W,k}}b_{q(\sigma^2_{W,k})}\right\} \end{aligned}$$

where

$$b_{q(\sigma^2_{W,k})} = b_W + \frac{1}{2}\left(\mu^T_{q(B_{W,k})}P\mu_{q(B_{W,k})} + \mathrm{tr}\left[P\Sigma_{q(B_{W,k})}\right]\right).$$

Thus the optimal density $q^*(\sigma^2_{W,k})$ is $\mathrm{IG}\left[a_{W,k}+\frac{K_\theta}{2}, b_{q(\sigma^2_{W,k})}\right]$. Note that the term $\mu_{q(1/\sigma^2_{W,k})}$ appearing in the optimal density $q^*(B_W)$ is equal to $\frac{a_{W,k}+\frac{K_\theta}{2}}{b_{q(\sigma^2_{W,k})}}$.

- For Σ, we have

$$\begin{aligned} q^*(\Sigma) &\propto \exp\left[E_{-\Sigma}\log p(\Sigma|\text{rest})\right] \\ &\propto |\Sigma|^{\frac{\nu+n+D+1}{2}}\exp\left\{-\frac{1}{2}\left(\mathrm{tr}\left[\Psi\Sigma^{-1}\right] + \sum_{m=1}^{n} E_{-\Sigma}[(Y_m - Z_m B^T_Z\Theta^T)\Sigma^{-1}(Y_m - Z_m B^T_Z\Theta^T)^T\right)]\right\} \end{aligned}$$

where m indexes rows of Y and Z and has values between 1 and n. For any m, the expectation above is

$$\begin{aligned} &Y_m\Sigma^{-1}Y^T_m - Z_m B^T_Z\Theta^T\Sigma^{-1}Y^T_m - Y_m\Sigma^{-1}\Theta B_Z Z_m + Z_m\mu^T_{q(B_Z)}\Theta^T\Sigma^{-1}\Theta\mu_{q(B_Z)}Z^T_m \\ &\quad + \mathrm{tr}\left[\Theta^T\Sigma^{-1}\Theta\Sigma_{q(B_{Z,i^*})}\right] \end{aligned}$$

where $\Sigma_{q(B_{Z,i^*})}$ is the covariance matrix defined in the update for $B_{Z,i}$ and subject i^* is the subject to which row m corresponds. Taking the sum over m of this expectation we have that

$$q^*(\Sigma) \propto |\Sigma|^{\frac{\nu+n+D+1}{2}}\exp\left\{-\frac{1}{2}\mathrm{tr}\left[\Psi_{q(\Sigma)}\Sigma^{-1}\right]\right\}$$

with

$$\Psi_{q(\Sigma)} = \Psi + Y^T Y - Y^T Z \mu_{q(B_Z)}^T \Theta^T + \Theta \mu_{q(B_Z)} Z^T Y + \Theta \mu_{q(B_Z)} Z^T Z \mu_{q(B_Z)}^T \Theta^T + \sum_i J_i \Theta \Sigma_{q(B_{Z,i})} \Theta^T.$$

Thus the optimal density $q^*(\Sigma)$ is IW $\left[\nu + n, \Psi_{q(\Sigma)}\right]$. Note that the term $\mu_{q(\Sigma^{-1})}$ appearing in the optimal density $q^*(B_Z)$ is equal to $\left(\frac{\Psi_{q(\Sigma)}}{\nu + \sum_i J_i}\right)^{-1}$.

From these individual updates, we arrive at the following iterative algorithm:

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

Initialize: $\mu_{q(B_{Z,i})}$, $\Sigma_{q(B_{Z,i})}$ for all i; $\mu_{q(B_W)}$; $\Sigma_{q(B_W)}$; $b_{q(\sigma_Z^2)}$, $b_{q(\sigma_{W,k}^2)}$ for all k; and $\Psi_{q(\Sigma)}$. Initial values can be chosen at random, or based on the approach used to select hyperparameters described in Section 2.3 to speed convergence.

Cycle:

- For all i:
  - $\Sigma_{q(B_{Z,i})} \leftarrow \left((\mathbf{1}_{J_i} \otimes \Theta)^T (I_{J_i} \otimes \mu_{q(\Sigma^{-1})})(\mathbf{1}_{J_i} \otimes \Theta) + \mu_{q(1/\sigma_Z^2)} P\right)^{-1}$
  - $\mu_{q(B_{Z,i})} \leftarrow \Sigma_{q(B_{Z,i})} \left((\mathbf{1}_{J_i} \otimes \Theta)^T (I_{J_i} \otimes \mu_{q(\Sigma^{-1})}) \text{vec}\,(Y_i) + \mu_{q(1/\sigma_Z^2)} P(\mu_{q(B_W)} \boldsymbol{w}_i^T)\right)$
- $\Sigma_{q(B_W)} \leftarrow \left(\mu_{q(1/\sigma_Z^2)} (W \otimes I_{K_\theta})^T (I_I \otimes P)(W \otimes I_{K_\theta}) + \text{diag}\left(\mu_{q(1/\sigma_{W,k}^2)}\right) \otimes P\right)^{-1}$
- $\mu_{q(B_W)} \leftarrow \Sigma_{q(B_W)} \left(\mu_{q(1/\sigma_Z^2)} (W \otimes I_{K_\theta})^T (I_I \otimes P) \text{vec}\left(\mu_{q(B_Z)}\right)\right)$
- $b_{q(\sigma_Z^2)} \leftarrow b_Z + \frac{1}{2} \sum_i \Big(\mu_{q(B_{Z,i})}^T P \mu_{q(B_{Z,i})} + \text{tr}\left[P \Sigma_{q(B_{Z,i})}\right] - 2\boldsymbol{w}_i \mu_{q(B_W)}^T P \mu_{q(B_{Z,i})} + \boldsymbol{w}_i \mu_{q(B_W)}^T P \mu_{q(B_W)} \boldsymbol{w}_i^T + \boldsymbol{w}_i \text{diag}\left(\text{tr}\left[P \Sigma_{q(B_{W,k})}\right]\right) \boldsymbol{w}_i^T\Big)$
- For all k:
  - $b_{q(\sigma_{W,k}^2)} \leftarrow b_W + \frac{1}{2}\left(\mu_{q(B_{W,k})}^T P \mu_{q(B_{W,k})} + \text{tr}\left[P \Sigma_{q(B_{W,k})}\right]\right)$
- $\Psi_{q(\Sigma)} \leftarrow \Psi + Y^T Y - Y^T Z \mu_{q(B_Z)}^T \Theta^T + \Theta \mu_{q(B_Z)} Z^T Y + \Theta \mu_{q(B_Z)} Z^T Z \mu_{q(B_Z)}^T \Theta^T + \sum_i J_i \Theta \Sigma_{q(B_{Z,i})} \Theta^T$

until the increase in L_q is negligible.

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

$$
\begin{aligned}
L_{q^*} &= \int q(\boldsymbol{\phi}) \log \frac{p(\boldsymbol{y}, \boldsymbol{\phi})}{q^*(\boldsymbol{\phi})} d\boldsymbol{\phi} = \\
&= \frac{1}{2} \sum_i \log(|\Sigma_{q(B_{Z,i})}|) + \frac{1}{2} \sum_k \log(|\Sigma_{q(B_{W,k})}|) - \\
&\quad \left(a_z + \frac{I * K_\theta}{2} \right) \log(b_{q(\sigma_Z^2)}) - \sum_k \left(a_w + \frac{K_\theta}{2} \right) \log(b_{q(\sigma_{W,k}^2)}) - \\
&\quad \left(\frac{\nu + \sum_i J_i}{2} \right) \log(|\Psi_{q(\Sigma)}|) + \text{ const.}
\end{aligned}
$$

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.