Modeling motor learning using heteroskedastic functional principal components analysis

Daniel Backenroth[1,*], Jeff Goldsmith[1], Michelle D. Harran[2], Juan C. Cortes[2], John W. Krakauer[3], and Tomoko Kitago[2]

[1]Department of Biostatistics, Mailman School of Public Health, Columbia University

[*]*db2175@cumc.columbia.edu*

[2]Department of Neurology, Columbia University Medical Center

[3]Departments of Neurology and Neuroscience, Johns Hopkins University

September 4, 2017

Abstract

We propose a novel method for estimating population-level and subject-specific effects of covariates on the variability of functional data. We extend the functional principal components analysis framework by modeling the variance of principal component scores as a function of covariates and subject-specific random effects. In a setting where principal components are largely invariant across subjects and covariate values, modeling the variance of these scores provides a flexible and interpretable way to explore factors that affect the variability of functional data. Our work is motivated by a novel dataset from an experiment assessing upper extremity motor control, and quantifies the reduction in motion variance associated with skill learning.

Key Words: Variational Bayes, Kinematic Data, Motor Control, Probabilistic PCA, Variance Modeling, Functional Data.

1 Scientific motivation

1.1 Motor learning

Recent work in motor learning has suggested that change in motion variability is an important component of improvement in motor skill. It has been suggested that when a motor task is learned, variance is reduced along dimensions relevant to the successful accomplishment of the task, although it may increase in other dimensions (Scholz and Schöner, 1999; Yarrow et al., 2009). Experimental work, moreover, has shown that learning-induced improvement of motion execution, measured through the trade-off between speed and accuracy, is accompanied by significant reductions in motion variability. In fact, these reductions in motion variability may be a more important feature of learning than changes in the average motion (Shmuelof et al., 2012). These results have typically been based on assessments of variability at a few time points, e.g., at the end of the motion, although high-frequency laboratory recordings of complete motions are often available.

In this paper we develop a modeling framework that can be used to quantify motion variability based on dense recordings of fingertip position throughout motion execution. This framework can be used to explore many aspects of motor skill and learning: differences in baseline skill among healthy subjects, effects of repetition and training to modulate variability over time, or the effect of baseline stroke severity on motion variance and recovery (Krakauer, 2006). By taking full advantage of high-frequency laboratory recordings, we shift focus from particular time points to complete curves. Our approach allows us to model the variability of these curves as they depend on covariates, like the hand used or the repetition number, as well as the estimation of random effects reflecting differences in baseline variability and learning rates among subjects.

Section 1.2 describes our motivating data in more detail, and Section 2 introduces our modeling framework. A review of relevant statistical work appears in Section 3. Details of our estimation approach are in Section 4. Simulations and the application to our motivating data appear in Sections 5 and 6, respectively, and we close with a discussion in Section 7.

1.2 Dataset

Our motivating data were gathered as part of a study of motor learning among healthy subjects. Kinematic data were acquired in a standard task used to measure control of reaching motions. In this task, subjects rest their forearm on an air-sled system to reduce effects of friction and gravity. The subjects are presented with a screen showing eight targets arranged in a circle around a starting point, and they reach with their arm to a target and back when it is illuminated on the screen. Subjects' motions are displayed on the screen, and they are rewarded with 10 points if they turn their hand around within the target, and 3 or 1 otherwise, depending on how far their hand is from the target at the point of return. Subjects are not rewarded for motions outside pre-specified velocity thresholds.

Our dataset consists of 9,481 motions by 26 right-handed subjects. After becoming familiarized with the experimental apparatus, each subject made 24 or 25 reaching motions to each of the 8 targets, in a semi-random order, with both the left and right hand. Motions that did not reach at least 30% of the distance to the target and motions with a direction more than 90° away from the target direction at the point of peak velocity were excluded from the dataset, because of the likelihood that they were made to the wrong target or not attempted due to distraction. Motions made at speeds outside the range of interest, with peak velocity less than 0.04 or greater than 2.0 m/s, were also excluded. These exclusion rules and other similar rules have been used previously in similar kinematic experiments, and are designed to increase the specificity of these experiments for probing motor control mechanisms (Huang et al., 2012; Tanaka et al., 2009; Kitago et al., 2015). A small number of additional motions were removed from the dataset due to instrumentation and recording errors. The data we consider have not been previously reported.

For each motion, the X and Y position of the hand motion is recorded as a function of time from motion onset to the initiation of return to the starting point, resulting in bivariate functional observations denoted $[P_{ij}^X(t), P_{ij}^Y(t)]$ for subject i and motion j. In practice, observations are recorded not as functions but as discrete vectors. There is some variability in motion duration, which we

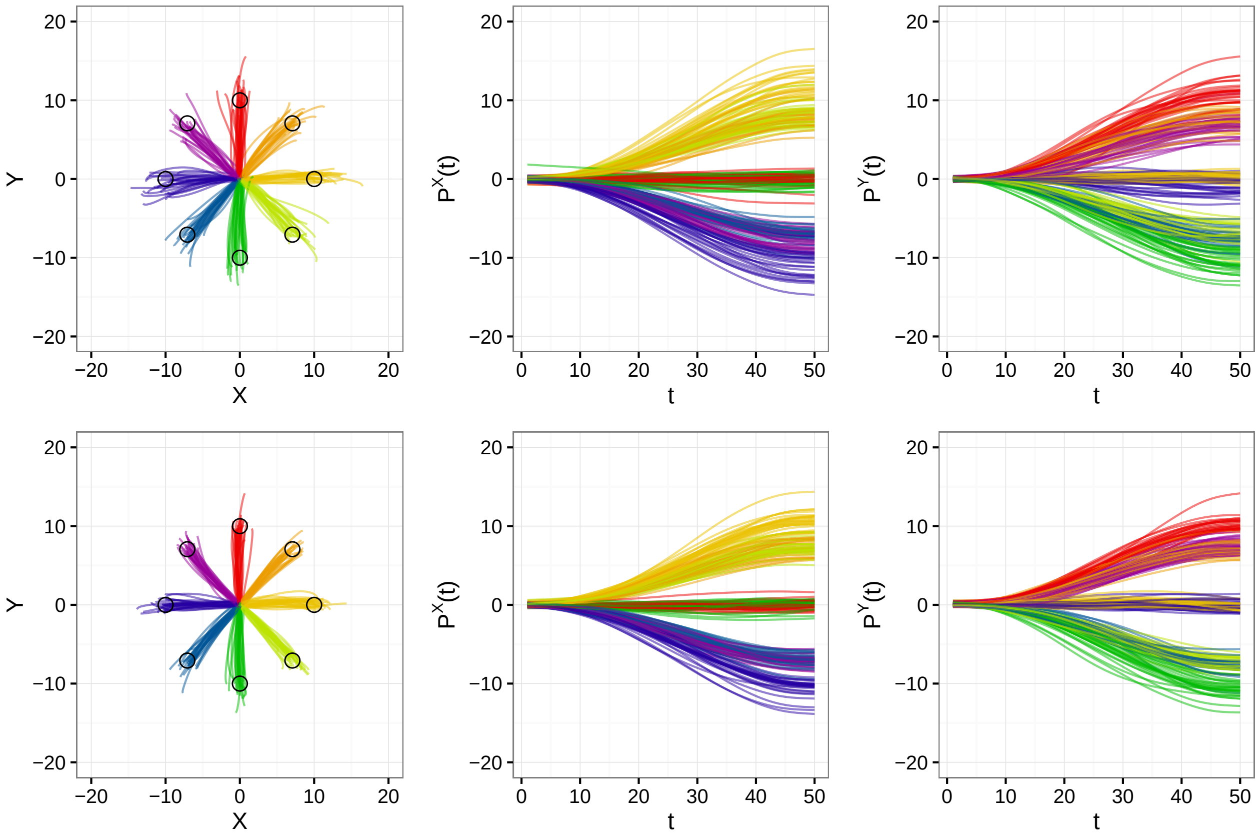

Figure 1: Observed kinematic data. The top row shows the first right-hand motion to each target for each subject; the bottom row shows the last motion. The left panel of each row shows observed reaching data in the X and Y plane. Targets are indicated with circles. The middle and right panels of each row show the $P_{ij}^X(t)$ and $P_{ij}^Y(t)$ curves, respectively.

remove for computational convenience by linearly registering each motion onto a common grid of length $D = 50$. The structure of the registered kinematic data is illustrated in Figure 1. The top and bottom rows show, respectively, the first and last right-hand motion made to each target by each subject. The reduction in motion variance after practice is clear.

Prior to our analyses, we rotate curves so that all motions extend to the target at 0°. This rotation preserves shape and scale, but improves interpretation. After rotation, motion along the X coordinate represents motion parallel to the line between origin and target, and motion along the Y coordinate represents motion perpendicular to this line. We build models for X and Y coordinate curves separately in our primary analysis. An alternative bivariate analysis appears in Appendix C.

2 Model for curve variance

We adopt a functional data approach to model position curves $P_{ij}(t)$. Here we omit the X and Y superscripts for notational simplicity. Our starting point is the functional principal component analysis (FPCA) model of Yao et al. (2005) with subject-specific means. In this model, it is assumed that each curve $P_{ij}(t)$ can be modeled as

$$\begin{aligned} P_{ij}(t) &= \mu_{ij}(t) + \delta_{ij}(t) \\ &= \mu_{ij}(t) + \sum_{k=1}^{\infty} \xi_{ijk}\phi_k(t) + \epsilon_{ij}(t). \end{aligned} \tag{1}$$

Here $\mu_{ij}(t)$ is the mean function for curve $P_{ij}(t)$, the deviation $\delta_{ij}(t)$ is modeled as a linear combination of eigenfunctions $\phi_k(t)$, the ξ_{ijk} are uncorrelated random variables with mean 0 and variances λ_k, where $\sum_k \lambda_k < \infty$ and $\lambda_1 \geq \lambda_2 \geq \cdots$, and $\epsilon_{ij}(t)$ is white noise. Here all the deviations $\delta_{ij}(t)$ are assumed to have the same distribution, that of a single underlying random process $\delta(t)$.

Model (1) is based on a truncation of the Karhunen-Loève representation of the random process $\delta(t)$. The Karhunen-Loève representation, in turn, arises from the spectral decomposition of the covariance of the random process $\delta(t)$ from Mercer's Theorem, from which one can obtain eigenfunctions $\phi_k(t)$ and eigenvalues λ_k.

The assumption of constant score variances λ_k in model (1) is inconsistent with our motivating data because it implies that the variability of the position curves $P_{ij}(t)$ is not covariate- or subject-dependent. However, motion variance can depend on the subject's baseline motor control and may change in response to training. Indeed, these changes in motion variance are precisely our interest.

In contrast to the preceding, we therefore assume that each random process $\delta_{ij}(t)$ has a potentially unique distribution, with a covariance operator that can be decomposed as

$$\text{Cov}[\delta_{ij}(s), \delta_{ij}(t)] = \sum_{k=1}^{\infty} \lambda_{ijk}\phi_k(s)\phi_k(t),$$

so that the eigenvalues λ_{ijk}, but not the eigenfunctions, vary among the curves. We assume that deviations $\delta_{ij}(t)$ are uncorrelated across both i and j.

The model we pose for the $P_{ij}(t)$ is therefore

$$P_{ij}(t) = \mu_{ij}(t) + \sum_{k=1}^{K} \xi_{ijk}\phi_k(t) + \epsilon_{ij}(t), \tag{2}$$

where we have truncated the expansion in model (1) to K eigenfunctions, and into which we incorporate covariate and subject-dependent heteroskedasticity with the score variance model

$$\lambda_{ijk} = \lambda_{k|\boldsymbol{x}^*_{ijk},\boldsymbol{z}^*_{ijk},\boldsymbol{g}_{ik}} = \mathrm{Var}(\xi_{ijk}|\boldsymbol{x}^*_{ijk}, \boldsymbol{z}^*_{ijk}, \boldsymbol{g}_{ik}) = \exp\left(\gamma_{0k} + \sum_{l=1}^{L^*} \gamma_{lk}x^*_{ijlk} + \sum_{m=1}^{M^*} g_{imk}z^*_{ijmk}\right) \tag{3}$$

where, as before, ξ_{ijk} is the kth score for the jth curve of the ith subject. In model (3), γ_{0k} is an intercept for the variance of the scores, γ_{lk} are fixed effects coefficients for covariates x^*_{ijlk}, $l = 1, \dots, L^*$, and g_{imk} are random effects coefficients for covariates z^*_{ijmk}, $m = 1, \dots, M^*$. The vector $\boldsymbol{g}_{ik}$ consists of the concatenation of the coefficients g_{imk}, and likewise for the vectors $\boldsymbol{x}^*_{ijk}$ and $\boldsymbol{z}^*_{ijk}$. Throughout, the subscript k indicates that models are used to describe the variance of scores associated with each basis function $\phi_k(t)$ separately. The covariates x^*_{ijlk} and z^*_{ijmk} in model (3) need not be the same across principal components. This model allows exploration of the dependence of motion variability on covariates, like progress through a training regimen, as well as of idiosyncratic subject-specific effects on variance through the incorporation of random intercepts and slopes.

Together, models (2) and (3) induce a subject- and covariate-dependent covariance structure for $\delta_{ij}(t)$:

$$\mathrm{Cov}[\delta_{ij}(s), \delta_{ij}(t)|\boldsymbol{x}^*_{ijk}, \boldsymbol{z}^*_{ijk}, \phi_k, \boldsymbol{g}_{ik}] = \sum_{k=1}^{K} \lambda_{k|\boldsymbol{x}^*_{ijk},\boldsymbol{z}^*_{ijk},\boldsymbol{g}_{ik}}\phi_k(s)\phi_k(t).$$

In particular, the $\phi_k(t)$ are assumed to be eigenfunctions of a conditional covariance operator. Our proposal can be related to standard FPCA by considering covariate values random and marginalizing across the distribution of random effects and covariates using the law of total covariance:

$$\begin{aligned}\mathrm{Cov}[\delta_{ij}(s),\delta_{ij}(t)] &= E\left\{\mathrm{Cov}[\delta_{ij}(s),\delta_{ij}(t)|\boldsymbol{x}^*,\boldsymbol{z}^*,\boldsymbol{g}]\right\}+\mathrm{Cov}\left\{E[\delta_{ij}(s)|\boldsymbol{x}^*,\boldsymbol{z}^*,\boldsymbol{g}]E[\delta_{ij}(t)|\boldsymbol{x}^*,\boldsymbol{z}^*,\boldsymbol{g}]\right\}\\ &= \sum_{k=1}^{K} E\left[\lambda_{k|\boldsymbol{x}^*_{ijk},\boldsymbol{z}^*_{ijk},\boldsymbol{g}_{ik}}\right]\phi_k(s)\phi_k(t).\end{aligned}$$

We assume that the basis functions $\phi_k(t)$ do not depend on covariate or subject effects, and are therefore unchanged by this marginalization. Scores ξ_{ijk} are marginally uncorrelated over k; this follows from the assumption that scores are uncorrelated in our conditional specification, and holds even if random effects $\boldsymbol{g}_{ik}$ are correlated over k. Lastly, the order of marginal variances $E\left[\lambda_{k|\boldsymbol{x}^*_{ijk},\boldsymbol{z}^*_{ijk},\boldsymbol{g}_{ik}}\right]$ may not correspond to the order of conditional variances $\lambda_{k|\boldsymbol{x}^*_{ijk},\boldsymbol{z}^*_{ijk},\boldsymbol{g}_{ik}}$ for some or even all values of the covariates and random effects coefficients.

In our approach, we assume that the scores ξ_{ijk} have mean zero. For this assumption to be valid, the mean $\mu_{ij}(t)$ in model (2) should be carefully modeled. To this end we use the well-studied multilevel function-on-scalar regression model (Guo, 2002; Di et al., 2009; Morris and Carroll, 2006; Scheipl et al., 2015),

$$\mu_{ij}(t)=\beta_0(t)+\sum_{l=1}^{L}x_{ijl}\beta_l(t)+\sum_{m=1}^{M}z_{ijm}b_{im}(t). \tag{4}$$

Here $\beta_0(t)$ is the functional intercept; x_{ijl} for $l\in 1,\ldots,L$ are scalar covariates associated with functional fixed effects with respect to the curve $P_{ij}(t)$; $\beta_l(t)$ is the functional fixed effect associated with the lth such covariate; z_{ijm} for $m\in 1,\ldots,M$ are scalar covariates associated with functional random effects with respect to the curve $P_{ij}(t)$; and $b_{im}(t)$ for $m\in 1,\ldots,M$ are functional random effects associated with the ith subject.

Keeping the basis functions constant across all subjects and motions, as in conventional FPCA, maintains the interpretability of the basis functions as the major patterns of variation across curves. Moreover, the covariate and subject-dependent score variances reflect the proportion of variation attributable to those patterns. To examine the appropriateness of this assumption for our data, we estimated basis functions for various subsets of motions using a traditional FPCA approach, after

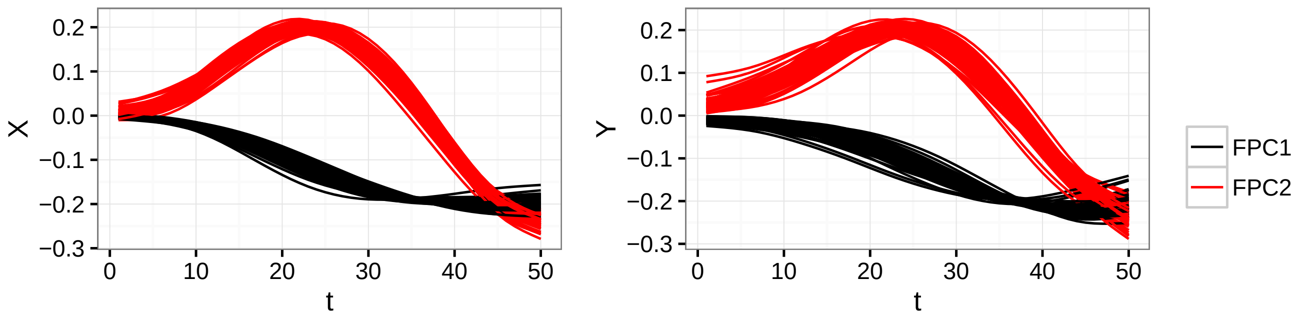

Figure 2: FPC basis functions estimated for various data subsets after rotating curves onto the positive X axis. The left panel shows the first and second FPC basis functions estimated for the X coordinate of motions to each target, for the left and right hand separately, and separately for motion numbers 1-6, 7-12, 13-18 and 19-24. The right panel shows the same for the Y coordinate.

rotating observed data so that all motions extend to the target at 0°. As illustrated in Figure 2, the basis functions for motions made by both hands and at different stages of training are similar.

3 Prior work

FPCA has a long history in functional data analysis. It is commonly performed using a spectral decomposition of the sample covariance matrix of the observed functional data (Ramsay and Silverman, 2005; Yao et al., 2005). Most relevant to our current work are probabilistic and Bayesian approaches based on non-functional PCA methods (Tipping and Bishop, 1999; Bishop, 1999; Peng and Paul, 2009). Rather than proceeding in stages, first by estimating basis functions and then, given these, estimating scores, such approaches estimate all parameters in model (1) jointly. James et al. (2000) focused on sparsely observed functional data and estimated parameters using an EM algorithm; van der Linde (2008) took a variational Bayes approach to estimation of a similar model. Goldsmith et al. (2015) considered both exponential-family functional data and multilevel curves, and estimated parameters using Hamiltonian Monte Carlo.

Some previous work has allowed for heteroskedasticity in FPCA. Chiou et al. (2003) developed a model which uses covariate-dependent scores to capture the covariate dependence of the mean of curves. In a manner that is constrained by the conditional mean structure of the curves, some covariate dependence of the variance of curves is also induced; the development of models for score variance was, however, not pursued. Here, by contrast, our interest is to use FPCA to model the effects of covariates on curve variance, independently of the mean structure. We are not using FPCA to model the mean; rather, the mean is modeled by the function-on-scalar regression model (4). Jiang and Wang (2010) introduce heteroskedasticity by allowing both the basis functions and the scores in an FPCA decomposition to depend on covariates. Briefly, rather than considering a bivariate covariance as the object to be decomposed, the authors pose a covariance surface that depends smoothly on a covariate. Aside from the challenge of incorporating more than a few covariates or subject-specific effects, it is difficult to use this model to explore the effects of covariates on heteroskedasticity: covariates affect both the basis functions and the scores, making the interpretation of scores and score variances at different covariate levels unclear. Although it does not allow for covariate-dependent heteroskedasticity, the model of Huang et al. (2014) allows curves to belong to one of a few different clusters, each with its own FPCs and score variances.

In contrast to the existing literature, our model provides a general framework for understanding covariate and subject-dependent heteroskedasticity in FPCA. This allows the estimation of rich models with multiple covariates and random effects, while maintaining the familiar interpretation of basis functions, scores, and score variances.

Variational Bayes methods, which we use here to approximate Bayesian estimates of the parameters in models (2) and (3), are computationally efficient and typically yield accurate point estimates for model parameters, although they provide only an approximation to the complete posterior distribution and inference may suffer as a result (Ormerod and Wand, 2012; Jordan, 2004; Jordan et al., 1999; Titterington, 2004). These tools have previously been used in functional data analysis (van der Linde, 2008; Goldsmith et al., 2011; McLean et al., 2013); in particular, Goldsmith and Kitago (2016) used variational Bayes methods in the estimation of model (4).

4 Methods

The main contribution of this manuscript is the introduction of subject and covariate effects on score variances in model (3). Several estimation strategies can be used within this framework. Here we describe three possible approaches. Later, these will be compared in simulations.

4.1 Sequential estimation

Models (2) and (3) can be fit sequentially in the following way. First, the mean $\mu_{ij}(t)$ in model (2) is estimated through function-on-scalar regression under a working independence assumption of the errors; we use the function `pffr` in the `refund` package (Goldsmith et al., 2016) in `R`. Next, the residuals from the function-on-scalar regression are modeled using standard FPCA approaches to obtain estimates of principal components and marginal score variances; given these quantities, scores themselves can be estimated (Yao et al., 2005). For this step we use the function `fpca.sc`, also in the `refund` package, which is among the available implementations. Next, we reestimate the mean $\mu_{ij}(t)$ in model (2) with function-on-scalar regression using `pffr`, although now, instead of assuming independence, we decompose the residuals using the principal components and score variances estimated in the previous step. We then reestimate principal components and scores using `fpca.sc`. The final step is to model the score variances given these score estimates. Assuming that the scores are normally distributed conditional on random effects and covariates, model (3) induces a generalized gamma linear mixed model for ξ_{ijk}^2, the square of the scores, with log link, coefficients γ_{lk} and g_{imk}, and shape parameter equal to 1/2. We fit this model with the `lme4` package, separately with respect to the scores for each principal component, in order to obtain estimates of our parameters of interest in the score variance model (Bates et al., 2015).

The first two steps of this approach are consistent with the common strategy for FPCA, and we account for non-constant score variance through an additional modeling step. We anticipate that this sequential approach will work reasonably well in many cases, but note that it arises as a sequence of models that treat estimated quantities as fixed. First, one estimates the mean; then one treats the mean as fixed to estimate the principal components and the scores; finally, one treats the scores as fixed to estimate the score variance model. Overall performance may deteriorate by failing to incorporate uncertainty in estimates in each step, particularly in cases of sparsely observed curves or high measurement error variances (Goldsmith et al., 2013). Additionally, because scores are typically estimated in a mixed model framework, the use of *marginal* score variances in the FPCA step can negatively impact score estimation and the subsequent modeling of *conditional* score variances.

4.2 Bayesian approach

4.2.1 Bayesian model

Jointly estimating all parameters in models (2) and (3) in a Bayesian framework is an appealing alternative to the sequential estimation approach. We expect this to be less familiar to readers than the sequential approach, and therefore provide a more detailed description.

Our Bayesian specification of these models is formulated in matrix form to reflect the discrete nature of the observed data. In the following $\mathbf{\Theta}$ is a known $D \times K_\theta$ matrix of K_θ spline basis functions evaluated on the shared grid of length D on which the curves are observed. We assume a normal distribution of the scores ξ_{ijk} conditional on random effects and covariates:

$$
\begin{aligned}
\boldsymbol{p}_{ij} &= \sum_{l=0}^{L} x_{ijl} \boldsymbol{\Theta} \boldsymbol{\beta}_l + \sum_{m=1}^{M} z_{ijm} \boldsymbol{\Theta} \boldsymbol{b}_{im} + \sum_{k=1}^{K} \xi_{ijk} \boldsymbol{\Theta} \boldsymbol{\phi}_k + \boldsymbol{\epsilon}_{ij} \qquad (5) \\
\boldsymbol{\beta}_l &\sim \text{MVN}\left[0, \sigma^2_{\boldsymbol{\beta}_l} \boldsymbol{Q}^{-1}\right]; \sigma^2_{\boldsymbol{\beta}_l} \sim \text{IG}\left[\alpha, \beta\right] \\
\boldsymbol{b}_i &\sim \text{MVN}\left[0, \sigma^2_{\boldsymbol{b}}((1-\pi)\boldsymbol{Q} + \pi \boldsymbol{I})^{-1}\right]; \sigma^2_{\boldsymbol{b}} \sim \text{IG}\left[\alpha, \beta\right] \\
\boldsymbol{\phi}_k &\sim \text{MVN}\left[0, \sigma^2_{\boldsymbol{\phi}_k} \boldsymbol{Q}^{-1}\right]; \sigma^2_{\boldsymbol{\phi}_k} \sim \text{IG}\left[\alpha, \beta\right] \\
\xi_{ijk} &\sim \text{N}\left[0, \exp\left(\sum_{l=0}^{L^*} \gamma_{lk} x^*_{ijlk} + \sum_{m=1}^{M^*} g_{imk} z^*_{ijmk}\right)\right] \\
\gamma_{lk} &\sim \text{N}\left[0, \sigma^2_{\gamma_{lk}}\right] \\
\boldsymbol{g}_{ik} &\sim \text{MVN}\left[0, \boldsymbol{\Sigma}_{\boldsymbol{g}_k}\right]; \boldsymbol{\Sigma}_{\boldsymbol{g}_k} \sim \text{IW}\left[\boldsymbol{\Psi}_k, \nu\right] \\
\boldsymbol{\epsilon}_{ij} &\sim \text{MVN}\left[0, \sigma^2 \boldsymbol{I}\right]; \sigma^2 \sim \text{IG}\left[\alpha, \beta\right]
\end{aligned}
$$

In model (5), $i = 1, \ldots, I$ refers to subjects, $j = 1, \ldots, J_i$ refers to motions within subjects, and $k = 1, \ldots, K$ refers to principal components. We define the total number of functional observations $n = \sum_{i=1}^{I} J_i$. The column vectors $\boldsymbol{p}_{ij}$ and $\boldsymbol{\epsilon}_{ij}$ are the $D \times 1$ observed functional outcome and independent error term, respectively, on the finite grid shared across subjects for the jth curve of the ith subject. The vectors $\boldsymbol{\beta}_l$, for $l = 0, \ldots, L$, are functional effect spline coefficient vectors, $\boldsymbol{b}_{im}$, for $i = 1, \ldots, I$ and $m = 1, \ldots, M$, are random effect spline coefficient vectors, and $\boldsymbol{\phi}_k$, for $k = 1 \ldots, K$, are principal component spline coefficient vectors, all of length K_θ. $\boldsymbol{Q}$ is a penalty matrix of the form $\boldsymbol{\Theta}^T \boldsymbol{M}^T \boldsymbol{M} \boldsymbol{\Theta}$, where $\boldsymbol{M}$ is a matrix that penalizes the second derivative of the estimated functions. $\boldsymbol{I}$ is the identity matrix. MVN refers to the multivariate normal distribution, N to the normal distribution, IG to the inverse-gamma distribution, and IW to the inverse-Wishart distribution. Models (3) and (4) can be written in the form of model (5) above by introducing into those models covariates x^*_{ij0k} (in model (3), multiplying γ_{0k}) and x_{ij0} (in model (4), multiplying $\beta_0(t)$), identically equal to 1. Some of the models used here, like in our real data analysis, do not have a global functional intercept $\boldsymbol{\beta}_0$ or global score variance intercepts γ_{0k}; in these models there are no such covariates identically equal to 1.

As discussed further in Section 4.2.3, for purposes of identifiability and to obtain FPCs that represent non-overlapping directions of variation, when fitting this model we introduce the additional constraint that the FPCs should be orthonormal and that each FPC should explain the largest possible amount of variance in the data, conditionally on the previously estimated FPCs, if any.

In keeping with standard practice, we set the prior variances $\sigma^2_{\gamma_{lk}}$ for the fixed-effect coefficients in the score variance model to a large constant, so that their prior is close to uniform. We set ν, the degrees of freedom parameter for the inverse-Wishart prior for the covariance matrices $\Sigma_{\boldsymbol{g}_k}$, to the dimension of $\boldsymbol{g}_{ik}$. We use an empirical Bayes approach, discussed further in Section 4.2.4, to specify $\boldsymbol{\Psi}_k$, the scale matrix parameters of these inverse-Wishart priors. When the random effects $\boldsymbol{g}_{ik}$ are one-dimensional, this prior reduces to an inverse-Gamma prior. Sensitivity to prior specifications of this model should be explored, and we do so with respect to our real data analysis in Appendix D.

Variance components $\{\sigma^2_{\boldsymbol{\beta}_l}\}_{l=0}^{L}$ and $\{\sigma^2_{\boldsymbol{\phi}_k}\}_{k=1}^{K}$ act as tuning parameters controlling the smoothness of coefficient functions $\beta_l(t)$ and FPC functions $\phi_k(t)$, and our prior specification for them is related to standard techniques in semiparametric regression. $\sigma^2_{\boldsymbol{b}}$, meanwhile, is a tuning parameter that controls the amount of penalization of the random effects, and is shared across the $b_{im}(t)$, so that all random effects for all subjects share a common distribution. Whereas fixed effects and functional principal components are penalized only through their squared second derivative, the magnitude of the random effects is also penalized through the full-rank penalty matrix $\boldsymbol{I}$ to ensure identifiability (Scheipl et al., 2015; Djeundje and Currie, 2010). The parameter π, $0 < \pi < 1$, determines the balance of smoothness and shrinkage penalties in the estimation of the random effects $b_{im}(t)$. We discuss how to set the value of this parameter in Section 4.2.4. We set α and β, the parameters of the inverse-gamma prior distributions for the variance components, to 1.

Our framework can accommodate more complicated random effect structures. In our application in Section 6, for example, each subject has 8 random effect vectors $\boldsymbol{g}_{ilk}$, one for each target, indexed by $l = 1, \ldots, 8$; the index l is used here since in Section 6 l is used to index targets. We model the correlations between these random effect vectors through a nested random effect structure:

$$\boldsymbol{g}_{ilk} \sim \text{MVN}\left[\boldsymbol{g}_{ik}, \boldsymbol{\Sigma}_{\boldsymbol{g}_{ik}}\right]; \qquad \boldsymbol{g}_{ik} \sim \text{MVN}\left[0, \boldsymbol{\Sigma}_{\boldsymbol{g}_k}\right] \tag{6}$$

Here the random effect vectors $\boldsymbol{g}_{ilk}$ for subject i and FPC k, $l = 1, \ldots 8$, are centered around a subject-specific random effect vector $\boldsymbol{g}_{ik}$. We estimate two separate random effect covariance matrices, $\boldsymbol{\Sigma}_{\boldsymbol{g}_{ik}}$ and $\boldsymbol{\Sigma}_{\boldsymbol{g}_k}$, for each FPC k, one at the subject-target level and one at the subject level. These matrices are given inverse-Wishart priors, and are discussed further in Section 4.2.4.

4.2.2 Estimation strategies

Sampling-based approaches to Bayesian inference of model (5) are challenging due to the constraints we impose on the $\phi_k(t)$ for purposes of interpretability of the score variance models, which are our primary interest. We present two methods for Bayesian estimation and inference for model (5): first, an iterative variational Bayes method, and second, a Hamiltonian Monte Carlo (HMC) sampler, implemented with the `STAN` Bayesian programming language (Stan Development Team, 2013). Our iterative variational Bayes method, which estimates each parameter in turn conditional on currently estimated values of the other parameters, is described in detail in Appendix E. This appendix also includes a brief overview of variational Bayes methods. Our HMC sampler, also described in Appendix E, conditions on estimates of the FPCs and fixed and random functional effects from the variational Bayes method, and estimates the other quantities in model (5).

4.2.3 Orthonormalization

A well-known challenge for Bayesian and probabilistic approaches to FPCA is that the basis functions $\phi_k(t)$ are not constrained to be orthogonal. In addition, when the scores ξ_{ijk} do not have unit variance, the basis functions will also be indeterminate up to magnitude, since any increase in their norm can be accommodated by decreased variance of the scores. Where interest lies in the variance of scores with respect to particular basis functions, it is important for the basis functions to be well-identified and orthogonal, so that they represent distinct and non-overlapping modes of variation. We therefore constrain estimated FPCs to be orthonormal and require each FPC to explain the largest possible amount of variance in the data, conditionally on the previously estimated FPCs, if any.

Let $\boldsymbol{\Xi}$ be the $n \times K$ matrix of principal component scores and $\boldsymbol{\Phi}$ the K by K_θ matrix of principal component spline coefficient vectors. In each step of our iterative variational Bayes algorithm, we apply the singular value decomposition to the matrix product $\boldsymbol{\Xi}\boldsymbol{\Phi}^T\boldsymbol{\Theta}^T$; the orthonormalized principal component basis vectors which satisfy these constraints are then the right singular vectors of this decomposition. A similar approach was used to induce orthogonality of the principal components in the Monte Carlo Expectation Maximization algorithm of (Huang et al., 2014) and as a post-processing step in (Goldsmith et al., 2015). Although explicit orthonormality constraints may be possible in this setting (Šmídl and Quinn, 2007), our simple approach, while not exact, provides for accurate estimation. Our HMC sampler conditions on the variational Bayes estimates of the FPCs, and therefore also satisfies the desired constraints.

4.2.4 Hyperparameter selection

The parameter π in model (5) controls the balance of smoothness and shrinkage penalization in the estimation of the random effects $\boldsymbol{b}_{im}$. In our variational Bayes approach we choose π to minimize the Bayesian information criterion (Schwarz, 1978), following the approach of Djeundje and Currie (2010).

To set the hyperparameter $\boldsymbol{\Psi}_k$ in model (5) (or the hyperparameters in the inverse-Wishart priors for the variance parameters in model (6)), we use an empirical Bayes method. First, we estimate scores ξ_{ijk} using our variational Bayes method, with a constant score variance for each FPC. We then estimate the random effects $\boldsymbol{g}_{ik}$ (or $\boldsymbol{g}_{ilk}$) using a generalized gamma linear mixed model, as described in Section 4.1. Finally, we compute the empirical covariance matrix corresponding to $\boldsymbol{\Sigma}_{\boldsymbol{g}_k}$

(or $\boldsymbol{\Sigma}_{\boldsymbol{g}_{ik}}$ and $\boldsymbol{\Sigma}_{\boldsymbol{g}_k}$), and set the hyperparameter so that the mode of the prior distribution matches this empirical covariance matrix.

5 Simulations

We demonstrate the performance of our method using simulated data. Here we present a simulation that includes functional random effects as well as scalar score variance random effects. Appendix F includes additional simulations in a cross-sectional context which demonstrate the effect of varying the number of estimated FPCs, the number of spline basis functions, and the measurement error.

In our simulation design, the jth curve for the ith subject is generated from the model

$$P_{ij}(t) = 0 + b_i(t) + \sum_{k=1}^{4} \xi_{ijk}\phi_k(t) + \epsilon_{ij}(t) \tag{7}$$

We observe the curves at $D = 50$ equally spaced points on the domain $[0, 2\pi]$. FPCs ϕ_1 and ϕ_2 correspond to the functions $\sin(x)$ and $\cos(x)$ and FPCs ϕ_3 and ϕ_4 correspond to the functions $\sin(2x)$ and $\cos(2x)$. We divide the curves equally into two groups $m = 1, 2$. We define x^*_{ij1} to be equal to 1 if the ith subject is assigned to group 1, and 0 otherwise, and we define x^*_{ij2} to be equal to 1 if the ith subject is assigned to group 2, and 0 otherwise. We generate scores ξ_{ijk} from zero-mean normal distributions with variances equal to

$$\mathrm{Var}(\xi_{ijk}|\boldsymbol{x}^*_{ij}, g_{ik}) = \exp\left(\sum_{l=1}^{2} \gamma_{lk}x^*_{ijl} + g_{ik}\right) \tag{8}$$

We set γ_{1k} for $k = 1, \ldots, 4$ to the natural logarithms of $36, 12, 6$ and 4, respectively, and γ_{2k} for $k = 1, \ldots, 4$ to the natural logarithms of $18, 24, 12$ and 6, respectively. The order of γ_{1k} and γ_{2k} for FPCs (represented by k) 1 and 2 black is purposely reversed between groups 1 and 2 so that the dominant mode of variation is not the same in the two groups. We generate the random effects g_{ik} in the score variance model from a normal distribution with mean zero and variance

$\sigma^2_{g_k}$, setting $\sigma^2_{g_k}$ to 3.0, 1.0, 0.3, and 0.1 across FPCs. We simulate functional random effects $b_i(t)$ for each subject by generating 10 elements of a random effect spline coefficient vector from the distribution MVN $[0, \sigma^2_{\boldsymbol{b}}((1-\pi)\boldsymbol{Q} + \pi\boldsymbol{I})^{-1}]$, and then multiplying this vector by a B-spline basis function evaluation matrix. We set $\pi = \sigma^2_{\boldsymbol{b}} = 1/2000$, resulting in smooth random effects approximately one-third the magnitude of the FPC deviations. The $\epsilon_{ij}(t)$ are independent errors generated at all t from a normal distribution with mean zero and variance $\sigma^2 = 0.25$.

We fix the sample size I at 24 and set the number of curves per subject J_i to 4, 12, 24 and 48. Two hundred replicate datasets were generated for each of the four scenarios. The simulation scenario with $I = J_i = 24$ is closest to the sample size in our real data application, where for each of 8 targets we have $I = 26$ and $J_i \approx 24$.

We fit the following model to each simulated dataset using each of the three approaches described in Section 4:

$$\begin{aligned}
\boldsymbol{p}_{ij} &= \boldsymbol{\Theta}\boldsymbol{\beta}_0 + \boldsymbol{\Theta}\boldsymbol{b}_i + \sum_{k=1}^{4} \xi_{ijk}\boldsymbol{\Theta}\boldsymbol{\phi}_k + \boldsymbol{\epsilon}_{ij} \\
\xi_{ijk} &\sim \mathrm{N}\left[0, \exp\left(\sum_{l=1}^{2} \gamma_{lk} x^*_{ijl} + g_{ik}\right)\right].
\end{aligned}$$

Here $\boldsymbol{p}_{ij}$ is the vectorized observation of $P_{ij}(t)$ from model (7). We use 10 spline basis functions for estimation, so that $\boldsymbol{\Theta}$ is a 50×10 B-spline basis function evaluation matrix. For the Bayesian approaches, we use the priors specified in model (5), including N $[0, 100]$ priors for variance parameters $\sigma^2_{\gamma_{lk}}$. We use the empirical Bayes approach discussed in Section 4.2.4 to set the scale parameters for the inverse-gamma priors for the variances $\sigma^2_{g_k}$ of the random effects g_{ik}.

Figures 3, 4 and 5 illustrate the quality of variational Bayes (VB) estimation of functional random effects, FPCs, and fixed and random effect score variance parameters. The top row of Figure 3 shows the collection of simulated curves for two subjects and includes the true and estimated subject-specific mean. The bottom row of this figure shows the true and estimated score variances across FPCs for a single simulated dataset, and suggests that fixed and random effects in the score

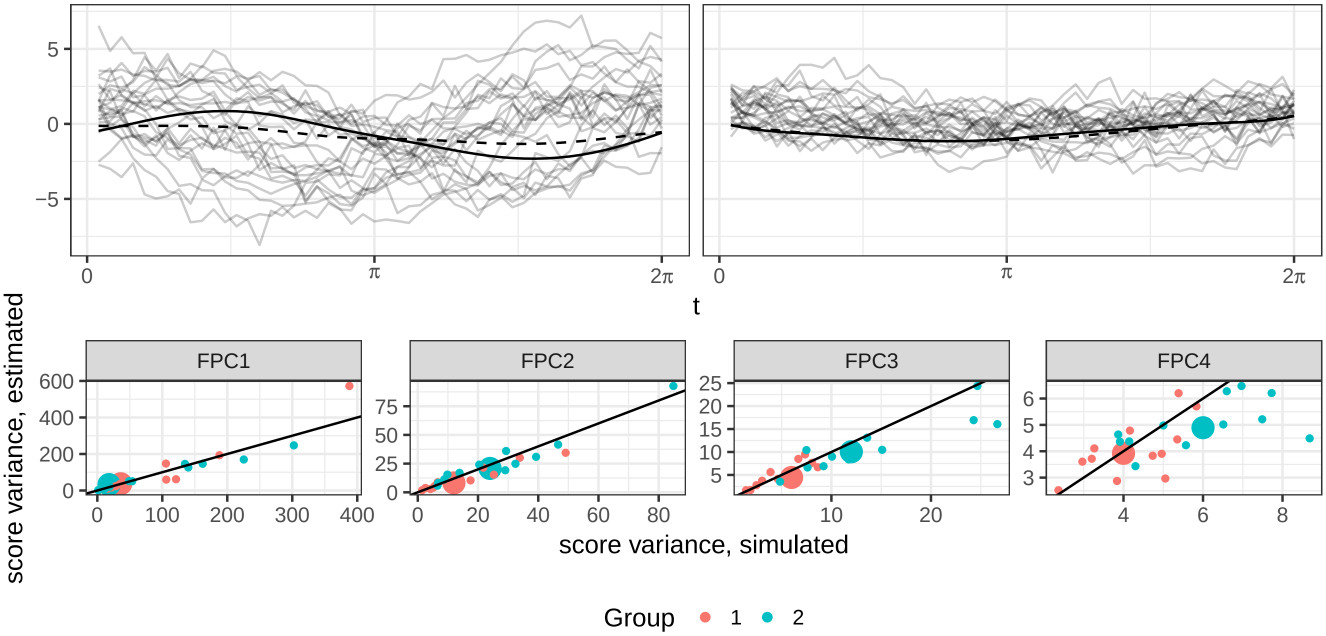

Figure 3: Selected results for the VB method for one simulation replicate with $I = J_i = 24$. This simulation replicate was selected because the estimation quality of the group-level score variances, shown in the bottom row, is close to median with respect to all simulations. Panels in the top row show simulated curves for two subjects in light black, the simulated functional random effect for that subject as a dashed line, and the estimated functional random effect for that subject as a dark solid line. The subjects were selected to show one subject with a poorly estimated functional random effect (left) and one with a well estimated functional random effect (right). Panels in the bottom row show, for each FPC, estimates and simulated values of the group-level and subject-specific score variances. Large colored dots are the group-level score variances, and small colored dots are the estimated score variances for each subject, i.e., they combine the fixed effect and the random effect.

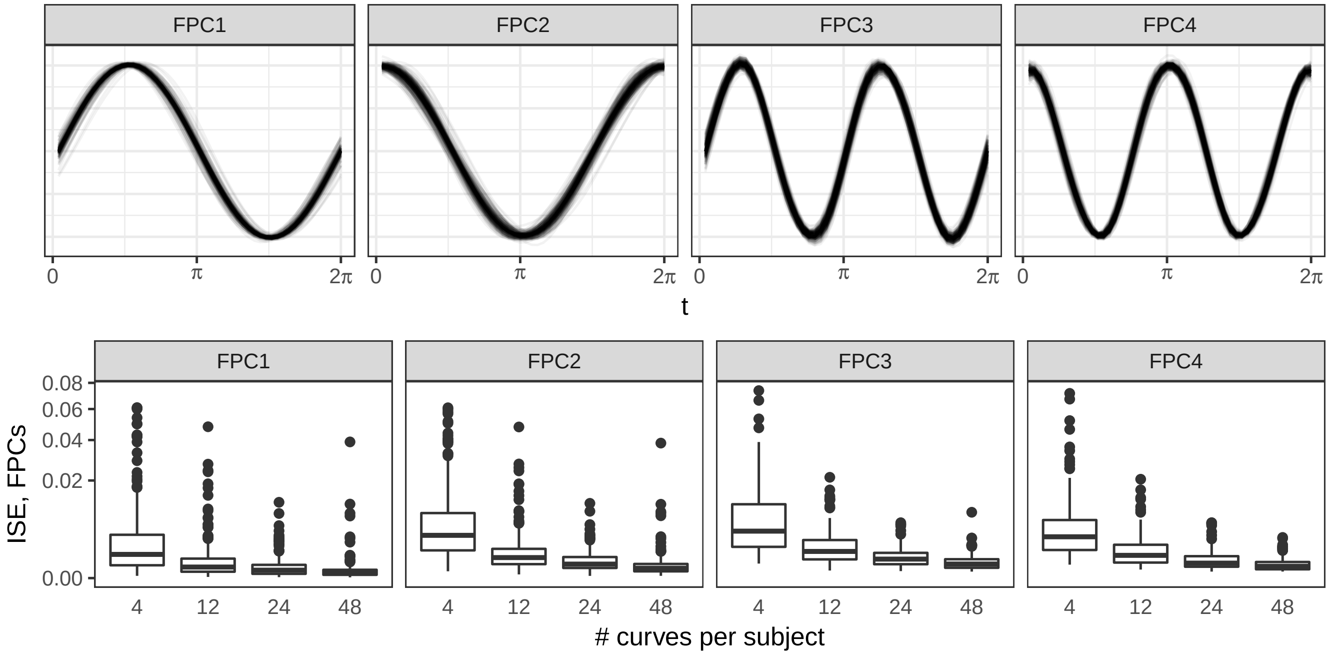

Figure 4: Estimation of FPCs using the VB method. Panels in the top row show a true FPC in dark black, and the VB estimates of that FPC for all simulation replicates with $J_i = 24$ in light black. Panels in the bottom row show, for each FPC and J_i, boxplots of integrated square errors (ISEs) for VB estimates $\widehat{\phi_k}(t)$ of each FPC $\phi_k(t)$, defined as ISE $= \int_0^{2\pi}[\phi_k(t) - \widehat{\phi_k}(t)]^2 dt$. The estimates in the top row therefore correspond to the ISEs for $J_i = 24$ shown in the bottom row. Figure A.10 in Appendix F shows examples of estimates of FPCs with a range of different ISEs.

variance model can be well-estimated.

The top row of Figure 4 shows estimated FPCs across all simulated datasets with $J_i = 24$; the FPCs are well-estimated and have no obvious systematic biases. The bottom row shows integrated squared errors (ISEs) for the FPCs across each possible J_i. As expected, the ISEs are smaller for the FPCs with larger score variances, and decrease as J_i increases. For 12 and especially for 4 curves per subject, estimates of the FPCs correspond to linear combinations of the simulated FPCs, leading to high ISEs and to inaccurate estimates of parameters in our score variance model (examples of poorly estimated FPCs can be seen in Appendix F).

Panels in the top row of Figure 5 show that estimates of fixed effect score variance parameters are shrunk towards zero, especially for lower numbers of curves per subject and FPCs 3 and 4. We attribute this to overfitting of the random effects in the mean model, which incorporates some of the variability attributable to the FPCs into the estimated random effects and reduces estimated

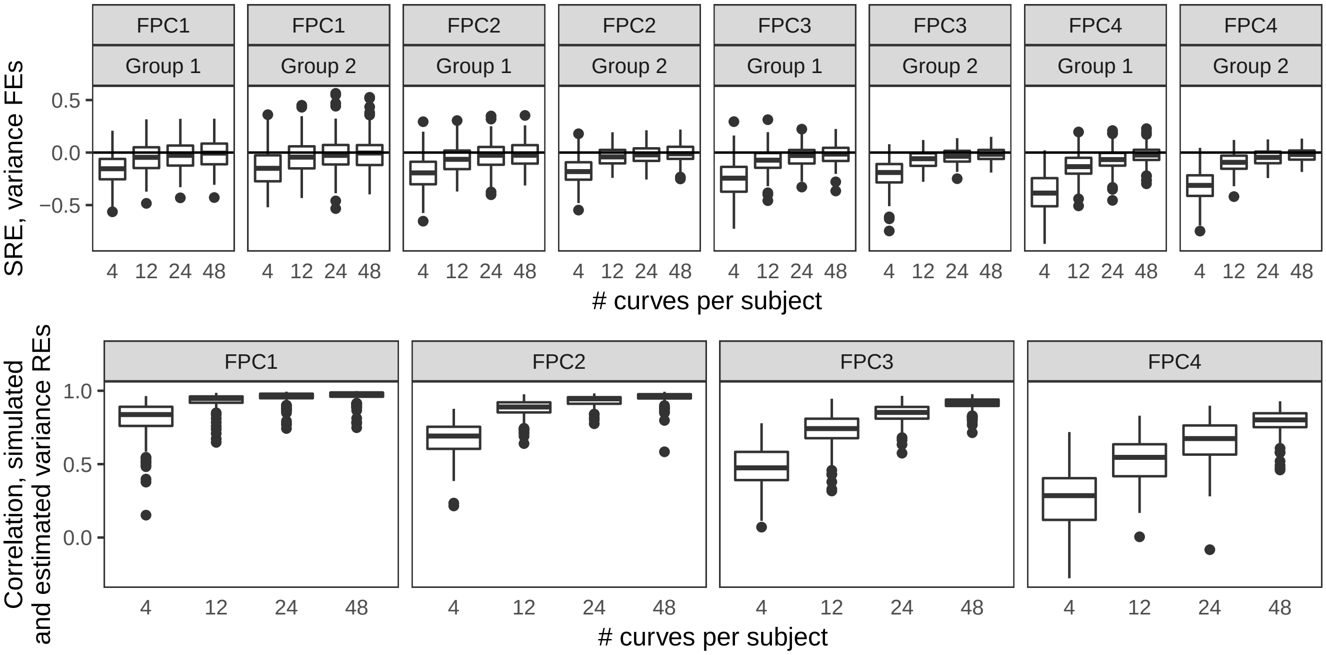

Figure 5: Estimation of score variance fixed and random effects using VB. Panels in the top row show, for each FPC, group, and J_i, boxplots of signed relative errors (SREs) for VB estimates $\widehat{\gamma_{lk}}$ of the fixed effect score variance parameters γ_{lk}, defined as SRE $= \frac{\widehat{\gamma_{lk}} - \gamma_{lk}}{\gamma_{lk}}$. Panels in the bottom row show, for each FPC and J_i, the correlation between random effect score variance parameters g_{ik} and their VB estimates. Intercepts and slopes for linear regressions of estimated on simulated random effect score variances are centered around 0 and 1, respectively (not shown).

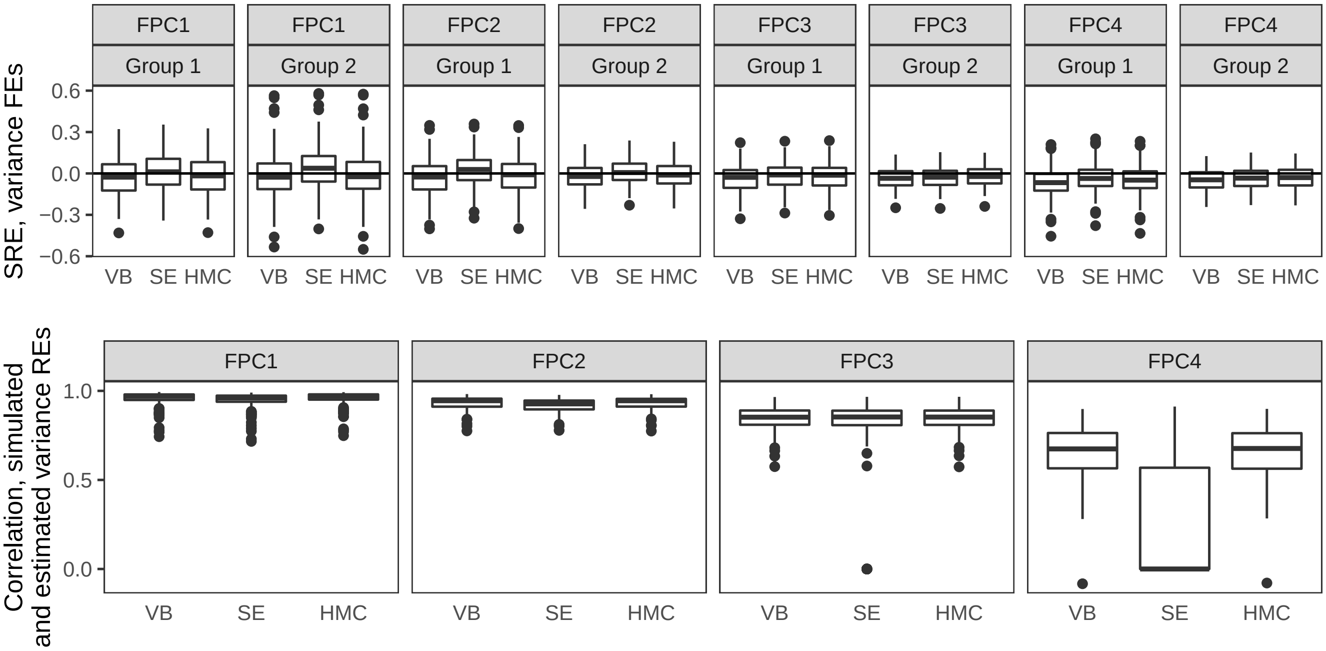

Figure 6: Comparison of estimation of score variance fixed and random effects using three methods. Panels in the top row show, for each FPC, group, and estimation method, boxplots of signed relative errors (SREs) for estimates of the fixed effect score variance parameters γ_{lk} for $J_i = 24$. Panels in the bottom row show, for each FPC and estimation method, the correlation between random effect score variance parameters g_{ik} and their estimates for $J_i = 24$. Intercepts and slopes for linear regressions of estimated on simulated random effect score variances are centered around 0 and 1, respectively (not shown).

score variances. Score variance random effects, shown in the bottom row of Figure 5, are more accurately estimated with more curves per subject.

Figure 6 and Table 1 show results from a comparison of the VB estimation procedure to the sequential estimation (SE) and Hamiltonian Monte Carlo (HMC) methods described in Section 4. We ran 4 HMC chains for 800 iterations each, and discarded the first 400 iterations from each chain. We assessed convergence of the chains by examining the convergence criterion of Gelman and Rubin (1992). Values of this criterion near 1 indicate convergence. For each of our simulation runs the criterion for every sampled variable was less than 1.1, and usually much closer to 1, suggesting convergence of the chains. In general, performance for the VB and HMC methods is comparable, and both methods are in some respects superior to the performance of the SE method. Figure 6 compares the three methods' estimation of the score variance parameters. Especially for FPC 4, the SE method occasionally estimates random effect variances at 0; these are represented in the lower-right panel of Figure 6 as points where the correlation between simulated and estimated score variance random effects is 0. Table 1 shows, based on the simulation scenario with $J_i = 24$, the frequentist coverage of 95% credible intervals for the VB and HMC methods, and of 95% confidence intervals for the SE method, in each case, for the fixed effect score variance parameters γ_{lk}. For FPCs 3 and 4 especially, the SE procedure confidence intervals are too narrow. The median ISE for the functional random effects is about 30% higher with the VB method than with the SE method. This results from the relative tendency of the VB method to shrink FPC score estimates to zero; when the mean of the scores is in fact non-zero, this shifts estimated functional random effects away from zero. Other comparisons of these methods are broadly similar.

The HMC method is more computationally expensive than the other two methods. Running 4 chains for 800 iterations in parallel took approximately 90 minutes for $J_i = 24$. On one processor, by comparison, the SE method took about 20 minutes, almost entirely to run function-on-scalar regression using `pffr`. The VB method took approximately six minutes, including the grid search to set the value of the parameter π, which controls the balance between zeroth and second-derivative penalties in the estimation of functional random effects.

FPC	Group	VB	SE	HMC
1	1	0.955	0.915	0.960
1	2	0.945	0.905	0.945
2	1	0.940	0.935	0.940
2	2	0.980	0.935	0.975
3	1	0.965	0.930	0.975
3	2	0.955	0.885	0.980
4	1	0.930	0.775	0.970
4	2	0.940	0.705	0.965

Table 1: Coverage of 95% credible/confidence intervals for the score variance parameters γ_{lk} using the VB, SE and HMC procedures, for $J_i = 24$.

6 Analysis of kinematic data

We now apply the methods described above to our motivating dataset. To reiterate, our goal is to quantify the process of motor learning in healthy subjects, with a focus on the reduction of motor variance through repetition. Our dataset consists of 26 healthy, right-handed subjects making repeated motions to each of 8 targets. We focus on estimation, interpretation and inference for the parameters in a covariate and subject-dependent heteroskedastic FPCA model, with primary interest in the effect of repetition number in the model for score variance. We hypothesize that variance will be lower for later repetitions due to skill learning.

Prior to fitting the model, we rotate all motions to be in the direction of the target at 0° so that the X axis is the major axis of motion. For this reason, variation along the X axis is interpretable as variation in motion extent and variation along the Y axis is interpretable as variation in motion direction. We present results for univariate analyses of the $P_{ij}^X(t)$ and $P_{ij}^Y(t)$ position curves in the right hand and describe a bivariate approach to modeling the same data.

We present models with 2 FPCs, since 2 FPCs are sufficient to explain roughly 95% of the motion variability (and usually more) of motions remaining after accounting for fixed and random effects in the mean structure. Most of the variability of motions around the mean is explained by the first FPC, so we emphasize score variance of the first FPC as a convenient summary for the motion variance, and briefly present some results for the second FPC.

6.1 Model

We examine the effect of practice on the variance of motions while accounting for target and individual-specific idiosyncrasies. To do this, we use a model for score variance that includes a fixed intercept and slope parameter for each target and one random intercept and slope parameter for each subject-target combination. Correlation between score variance random effects for different targets for the same subject is induced via a nested random effects structure. The mean structure for observed curves consists of functional intercepts $\boldsymbol{\beta}_l$ for each target $l \in \{1, \ldots, 8\}$ and random effects $\boldsymbol{b}_{il}$ for each subject-target combination, to account for heterogeneity in the average motion across subjects and targets. Our heteroskedastic FPCA model is therefore:

$$\boldsymbol{p}_{ij} = \sum_{l=1}^{8} \mathbb{I}(tar_{ij} = l)\left(\boldsymbol{\Theta}\boldsymbol{\beta}_l + \boldsymbol{\Theta}\boldsymbol{b}_{il}\right) + \sum_{k=1}^{K} \xi_{ijk}\boldsymbol{\Theta}\boldsymbol{\phi}_k + \boldsymbol{\epsilon}_{ij} \tag{9}$$

$$\xi_{ijk} \sim \mathrm{N}\left[0, \sigma^2_{\xi_{ijk}} = \exp\left(\sum_{l=1}^{8} \mathbb{I}(tar_{ij} = l)\left(\gamma_{lk,int} + g_{ilk,int} + (rep_{ij} - 1)(\gamma_{lk,slope} + g_{ilk,slope})\right)\right)\right] \tag{10}$$

$$\boldsymbol{g}_{ilk} \sim \mathrm{MVN}\left[\boldsymbol{g}_{ik}, \boldsymbol{\Sigma}_{\boldsymbol{g}_{ik}}\right]; \; \boldsymbol{g}_{ik} \sim \mathrm{MVN}\left[0, \boldsymbol{\Sigma}_{\boldsymbol{g}_k}\right]$$

The covariate tar_{ij} indicates the target to which motion j by subject i is directed. The covariate rep_{ij} indicates the repetition number of motion j, starting at 1, among all motions by subject i to the target to which motion j is directed, and $\mathbb{I}(\cdot)$ is the indicator function. To accommodate differences in baseline variance across targets, this model includes separate population-level intercepts $\gamma_{lk,int}$ for each target l. The slopes $\gamma_{lk,slope}$ on repetition number indicate the change in variance due to practice for target l; negative values indicate a reduction in motion variance. To accommodate subject and target-specific effects, each subject-target combination has a random intercept $g_{ilk,int}$ and a random slope $g_{ilk,slope}$, and each subject has an overall random intercept $g_{ik,int}$ and overall random slope $g_{ik,slope}$, in the score variance model for each functional principal component. This model parameterization allows different baseline variances and changes in variance for each target and subject, but shares FPC basis functions across targets. The model also assumes independence of functional random effects $\boldsymbol{b}_{il}$, $l = 1, \ldots, 8$ by the same subject to different targets, as well as independence of functional random effects $\boldsymbol{b}_{il}$ and score variance random effects $\boldsymbol{g}_{ilk}$ for the same subject. The validity of these assumptions for our data are discussed in Appendix D.

Throughout, fixed effects $\gamma_{lk,int}$ and $\gamma_{lk,slope}$ are given $\mathrm{N}[0, 100]$ priors. Random effects $g_{ilk,int}$ and $g_{ilk,slope}$ are modeled using a bivariate normal distribution to allow for correlation between the random intercept and slope parameters in each FPC score variance model, and with nesting to allow for correlations between the random effects for the same subject and different targets. We use the empirical Bayes method described in Section 4.2.4 to set the scale matrix parameters of the inverse-Wishart priors for $\boldsymbol{g}_{ilk}$ and $\boldsymbol{g}_{ik}$. Appendix D includes an analysis which examines the sensitivity of our results to various choices of prior hyperparameters.

We fit (9) and (10) using our VB method, with $K = 2$ principal components and a cubic B-spline evaluation matrix $\boldsymbol{\Theta}$ with $K_\theta = 10$ basis functions.

6.2 Results

Figure 7 shows estimated score variances as a function of repetition number for X and Y coordinate right hand motions to all targets. There is a decreasing trend in score variance for the first principal component scores for all targets and for both the X and Y coordinates, which agrees with our hypotheses regarding learning. Figure 7 also shows that nearly all of the variance of motion is attributable to the first FPC. Baseline variance is generally higher in the X direction than the Y direction, indicating that motion extent is generally more variable than motion direction.

To examine the adequacy of modeling score variance as a function of repetition number with a linear model, we compared the results of model (10) with a model for the score variances saturated in repetition number, i.e., where each repetition number m has its own set of parameters γ_{lkm} in the model for the score variances:

$$\xi_{ijk} \sim \mathrm{N}\left[0, \sigma^2_{\xi_{ijk}} = \exp\left(\sum_{l=1}^{8}\sum_{m=1}^{24} \mathbb{I}(tar_{ij} = l, rep_{ij} = m)\gamma_{lkm}\right)\right]. \tag{11}$$

The results for these two models are included in Figure 7. The general agreement between the linear and saturated models suggests that the slope-intercept model is reasonable. For some targets score variance is especially high for the first motion, which may reflect a familiarization with the experimental apparatus.

We now consider inference for the decreasing trend in variance for the first principal component scores. We are interested in the parameters $\gamma_{l1,slope}$, which estimate the population-level target-specific changes in score variance for the first principal component with each additional motion.

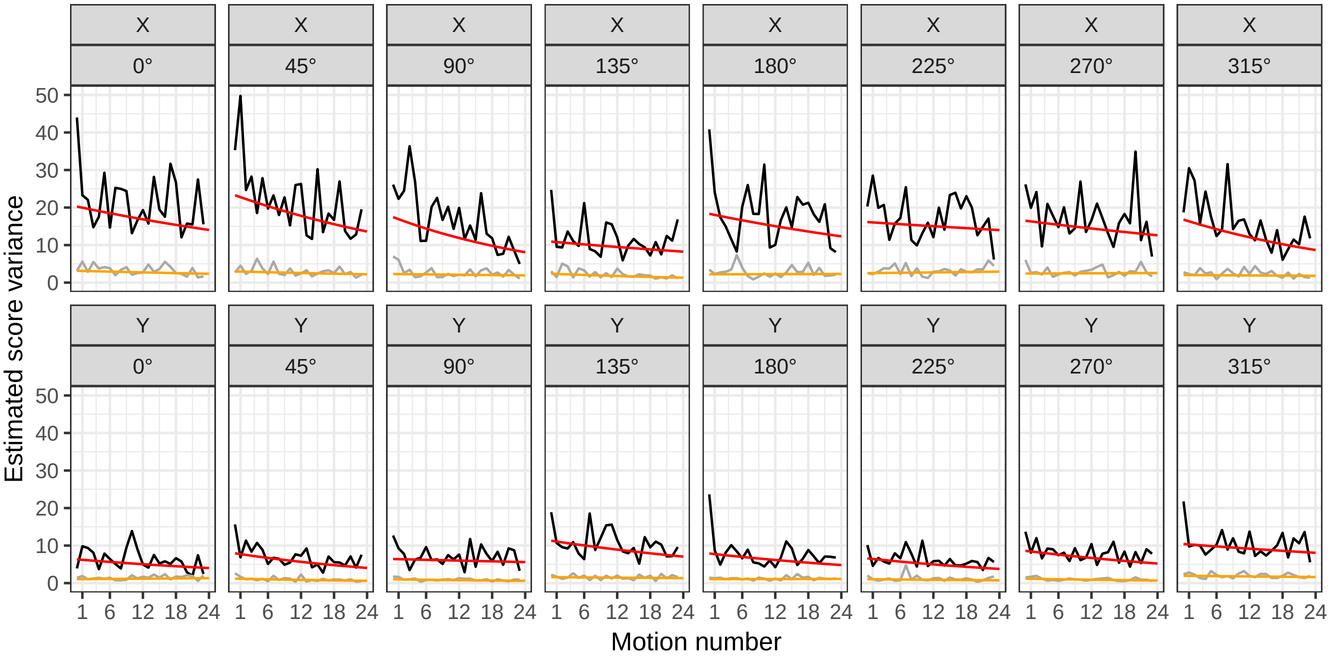

Figure 7: VB estimates of score variances for right hand motions to each target (in columns), separately for each direction (X or Y, in rows). Panels show the VB estimates of the score variance as a function of repetition number using the slope-intercept model (10) in red and orange (first and second FPC, respectively), and using the saturated one-parameter-per-repetition number model (11), in black and grey (first and second FPC, respectively).

Figure 8 shows VB estimates and 95% credible intervals for the $\gamma_{l1,slope}$ parameters for motions by the right hand to each target. All the point estimates $\gamma_{l1,slope}$ are lower than 0, indicating decreasing first principal component score variance with additional repetition. For some targets and coordinates there is substantial evidence that $\gamma_{l1,slope} < 0$; these results are consistent with our understanding of motor learning, although they do not adjust for multiple comparisons.

Appendix C includes results of a bivariate approach to modeling motion kinematics, which accounts for the 2-dimensional nature of the motions. In this approach, the X and Y coordinates of curves are concatenated, and each principal component reflects variation in both X and Y coordinates. For curves rotated to extend in the same direction, the results of this approach suggests that variation in motion extent (represented by the X coordinate) and motion direction (represented by the Y coordinate) are largely uncorrelated: the estimate of the first bivariate FPC represents variation primarily in the X coordinate, and is similar to the estimate of the first FPC in the X coordinate model, and vice versa for the second bivariate FPC. Analyses of score variance, then,

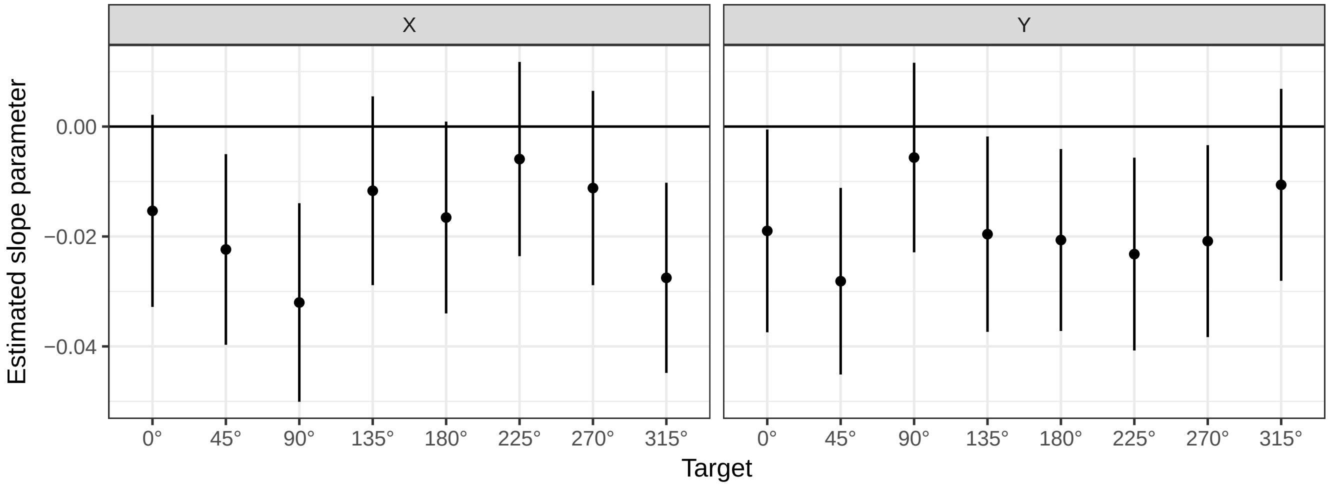

Figure 8: VB estimates of $\gamma_{l1,slope}$. This figure shows VB estimates and 95% credible intervals for target-specific score variance slope parameters $\gamma_{l1,slope}$ for motions by the right hand to each target, for the X and Y coordinates.

closely follow the preceding univariate analyses.

Appendix B includes an analysis of data for one target using the VB, HMC and SE methods. The three methods yield similar results.

7 Discussion

This manuscript develops a framework for the analysis of covariate and subject-dependent patterns of motion variance in kinematic data. Our methods allow for flexible modeling of the covariate-dependence of variance of functional data with easily interpretable results. Our approach allows for the estimation of subject-specific effects on variance, as well as the consideration of multiple covariates.

By applying these methods to our motivating dataset, we have demonstrated that motion variance is reduced with repetition. Results in Appendix A additionally show that the baseline level of skill of subjects is correlated across targets and hands, and that baseline variance is considerably greater in the non-dominant than the dominant hand. Further applications of these methods in scientifically important contexts could focus, for example, on whether motion variance is reduced with training faster in the dominant hand, or on whether training with one hand transfers skill to the other hand. Further research could also investigate target-specific differences in improvement of variance with training. Movements to some of the targets require coordination between the shoulder and elbow, whereas others are primarily single-joint motions; the effectiveness of training may depend on the complexity of the motion.

We have provided three different estimation approaches for fitting heteroskedastic functional principal components models. Given its computational efficiency and comparable accuracy to the HMC and SE methods, we recommend use of the VB approach for exploratory analyses and model building. However, because of its approximate nature, we advise that any conclusions derived from the VB approach be confirmed with one of the other two methods, perhaps with a subset of the data if required for computational feasibility.

An alternative approach to the analysis of this dataset could treat the target effects $\gamma_{lk,int}$ and $\gamma_{lk,slope}$ in model (10) for the score variances as random effects centered around parameters $\mu_{k,int}$ and $\mu_{k,slope}$, representing the average across-target baseline score variance and change in score variance with repetition. Some advantages of this approach would be the estimation of parameters that summarize the global effect of repetition on motion variance and shrinkage of the target-specific score variance parameters. However, with only 8 random target effects, the model would be sensitive to the specification of priors. Moreover, as discussed above, motions to different targets impose different demands on coordination and skill, which may reduce the interpretability of the parameters $\mu_{k,int}$ and $\mu_{k,slope}$.

Our analysis here is of curves linearly registered onto a common time domain, although our method could be applied to curves with different time domains, or to sparsely observed functional data. Our research group is currently working on developing an improved approach to registration in kinematic experiments which will take account of the repeated observations at the subject level by seeking to estimate subject- and curve-specific warping functions. This approach, combined with the methods we present in the current manuscript, will eventually allow a more complete model for motion variability that takes into account both variability in motion duration and variability in motion trajectories.

There are several directions for further development. A full Bayesian treatment could estimate all quantities in model (5) jointly, or could condition on only the FPCs and jointly estimate all other quantities; given the very flexible nature of this model, additional constraints might be required in such a Bayesian treatment to improve identifiability. More complex models could allow for correlations between functional random effects and score variance random effects. Considering our data from the perspective of shape analysis may provide better understanding of interpretable motion features like location, scale and orientation (Kurtek et al., 2012; Gu et al., 2012). Lastly, an alternative approach to that presented here would be to model covariate-dependent score distributions through quantile regression. This may produce valuable insights into the complete distribution of motions, especially when this is not symmetric, but some work is needed to understand the connection of this technique to traditional FPCA.

8 Supplementary Materials

The Supplementary Materials contains the following appendices: A, containing additional results from our real data analysis; B, containing results comparing the VB, HMC and SE methods as applied to the kinematic data; C, containing a description of and results from a bivariate approach to the analysis of our kinematic data; D, containing a sensitivity analysis for our choice of hyperparameters and examining alternative mean specifications; E, containing a derivation of our variational Bayes algorithm and more detail on our HMC implementation; and F, containing additional simulation results.

The Supplementary Materials also include code implementing our methods and all the simulation scenarios presented here. This code can fit models in the form of model (5), provided that, as in our real data application and simulations, only one functional random effect is associated with each functional observation.

9 Acknowledgements

The first and second author's research was supported in part by Award R21EB018917 from the National Institute of Biomedical Imaging and Bioengineering. The second author's research was also supported by Award R01NS097423-01 from the National Institute of Neurological Disorders and Stroke and Award R01HL123407 from the National Heart, Lung, and Blood Institute.

References

Bates, D., Mächler, M., Bolker, B., and Walker, S. "Fitting Linear Mixed-Effects Models Using lme4." Journal of Statistical Software, 67(1):1–48 (2015).

Bishop, C. M. "Bayesian PCA." Advances in Neural Information Processing Systems, 382–388 (1999).

Chiou, J.-M., Müller, H.-G., and Wang, J.-L. "Functional quasi-likelihood regression models with smooth random effects." Journal of the Royal Statistical Society: Series B (Statistical Methodology), 65(2):405–423 (2003).

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).

Djeundje, V. A. B. and Currie, I. D. "Appropriate covariance-specification via penalties for penalized splines in mixed models for longitudinal data." Electronic Journal of Statistics, 4:1202–1224 (2010).

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

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

Goldsmith, J. and Kitago, T. "Assessing Systematic Effects of Stroke on Motor Control using Hierarchical Function-on-Scalar Regression." Journal of the Royal Statistical Society: Series C, 65:215–236 (2016).

Goldsmith, J., Scheipl, F., Huang, L., Wrobel, J., Gellar, J., Harezlak, J., McLean, M. W., Swihart, B., Xiao, L., Crainiceanu, C., and Reiss, P. T. refund: Regression with Functional Data (2016). R package version 0.1-17.
URL http://CRAN.R-project.org/package=refund

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

Goldsmith, J., Zipunnikov, V., and Schrack, J. "Generalized multilevel function-on-scalar regression and principal component analysis." Biometrics, 71(2):344–353 (2015).

Gu, K., Pati, D., and Dunson, D. B. "Bayesian hierarchical modeling of simply connected 2D shapes." arXiv preprint arXiv:1201.1658 (2012).

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

Huang, H., Li, Y., and Guan, Y. "Joint Modeling and Clustering Paired Generalized Longitudinal Trajectories with Application to Cocaine Abuse Treatment Data." Journal of the American Statistical Association, 83:210–223 (2014).

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).

James, G. M., Hastie, T. J., and Sugar, C. A. "Principal component models for sparse functional data." Biometrika, 87:587–602 (2000).

Jiang, C.-R. and Wang, J.-L. "Covariate adjusted functional principal components analysis for longitudinal data." The Annals of Statistics, 38:1194–1226 (2010).

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., Goldsmith, J., Harran, M., Kane, L., Berard, J., Huang, S., Ryan, S. L., Mazzoni, P., Krakauer, J. W., and Huang, V. S. "Robotic therapy for chronic stroke: general recovery of impairment or improved task-specific skill?" Journal of Neurophysiology, 114(3):1885–1894 (2015).

Krakauer, J. W. "Motor learning: its relevance to stroke recovery and neurorehabilitation." Current Opinion in Neurology, 19:84–90 (2006).

Kurtek, S., Srivastava, A., Klassen, E., and Ding, Z. "Statistical modeling of curves using shapes and related features." Journal of the American Statistical Association, 107:1152–1165 (2012).

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

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

Neal, R. "MCMC Using Hamiltonian Dynamics." Handbook of Markov Chain Monte Carlo, Chapter 5, 113–162 (2011).

Nott, D. J., Tran, M.-N., and Leng, C. "Variational approximation for heteroscedastic linear models and matching pursuit algorithms." Statistics and Computing, 22(2):497–512 (2012).

Ormerod, J. and Wand, M. P. "Gaussian Variational Approximation Inference for Generalized Linear Mixed Models." The American Statistician, 21:2–17 (2012).

Peng, J. and Paul, D. "A geometric approach to maximum likelihood estimation of the functional principal components from sparse longitudinal data." Journal of Computational and Graphical Statistics, 18:995–1015 (2009).

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

Scheipl, F., Staicu, A.-M., and Greven, S. "Functional additive mixed models." Journal of Computational and Graphical Statistics, 24:477–501 (2015).

Scholz, J.-P. and Schöner, G. "The uncontrolled manifold concept: identifying control variables for a functional task." Experimental Brain Research, 126:289–306 (1999).

Schwarz, G. "Estimating the Dimension of a Model." The Annals of Statistics, 6:461–464 (1978).

Shmuelof, L., Krakauer, J. W., and Mazzoni, P. "How is a motor skill learned? Change and invariance at the levels of task success and trajectory control." Journal of Neurophysiology, 108(2):578–594 (2012).

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

Tanaka, H., Sejnowski, T. J., and Krakauer, J. W. "Adaptation to visuomotor rotation through interaction between posterior parietal and motor cortical areas." Journal of Neurophysiology, 102:2921–2932 (2009).

Tipping, M. E. and Bishop, C. "Probabilistic Principal Component Analysis." Journal of the Royal Statistical Society: Series B, 61:611–622 (1999).

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).

Šmídl, V. and Quinn, A. "On Bayesian principal component analysis." Computational Statistics & Data Analysis, 51:4101–4123 (2007).

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).

Yarrow, L., Brown, P., and Krakauer, J.-W. "Inside the brain of an elite athlete: the neural processes that support high achievement in sports." Nature Reviews Neuroscience, 10:585–596 (2009).

Appendices to: Modeling motor learning using heteroskedastic functional principal components analysis

Daniel Backenroth, Jeff Goldsmith, Michelle D. Harran. Juan C. Cortes, John W. Krakauer and Tomoko Kitago

A Additional results from analysis of kinematic data

One scientifically interesting question about individual motion characteristics that is addressable in our modeling framework is whether subjects with high baseline motion variance to one target tend to have high baseline motion variance to other targets. Figure A.1 shows the estimated first principal component score variance random intercept parameters $g_{il1,int}$ for each subject and each target for both the left and right hands for the X coordinate of motion, ordered by the average random intercept for each subject across targets for the right hand. There are clear subject-specific patterns of variability shared across and within hands, and clearer subject-specific patterns of variability within each hand across 8 targets. The correlation of average random intercepts for each subject across the 8 targets, one for the left and one for the right hand, was 0.56, indicating a positive correlation between baseline motor skill across hands within an individual.

Our model's point estimate of the correlation between the subject-specific cross-target score variance random intercept and the subject-specific cross-target score variance random slope is -0.80, suggesting a relationship between high baseline motion variance and faster decrease in variance with practice.

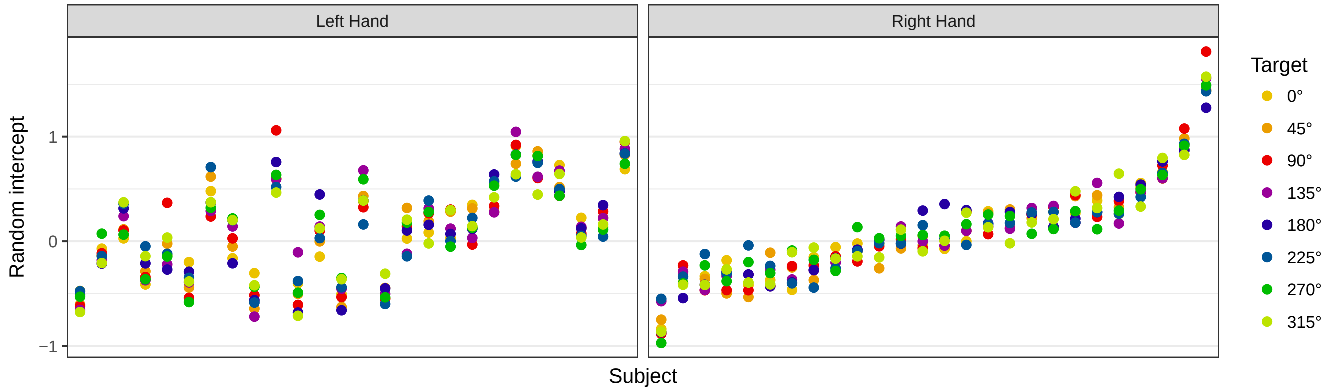

Figure A.1: Estimates of random intercepts. Each panel shows, for the left or the right hand, the estimated first principal component score variance random intercept parameters $g_{il1,int}$ in model (10) for each subject i and target l, for the X coordinate of motion. Targets are colored as in Figure 1, and subjects are ordered by their average random intercept across targets for the right hand.

B HMC and SE methods applied to kinematic data

We applied the VB, HMC and SE methods to the X coordinate of motions by the right hand to the target at 0°, and obtained very similar results. While the estimate and 95% posterior credible interval for the first FPC slope variance parameter using VB was -0.020 $(-0.043, 0.003)$, the corresponding estimate and interval for HMC was -0.020 $(-0.040, -0.001)$ and the SE confidence interval was -0.023 $(-0.041, -0.005)$. The estimates and posterior credible/confidence intervals for the first FPC intercept variance parameter were also similar: 3.12 $(2.81, 3.43)$ for VB versus 3.18 $(2.9, 3.45)$ for HMC and 3.23 $(2.97, 3.49)$ for SE.

Estimates of random effects were also similar using the three methods, with all pairwise correlations between random intercepts and random slopes estimated using the three methods exceeding 0.85.

To generate these HMC results we ran 4 HMC chains for 2000 iterations each, and discarded the first 1000 iterations from each chain. The convergence criterion of Gelman and Rubin (1992) was less than 1.011 for each sampled variable, suggesting convergence of the chains.

C Bivariate model

To fit our model to bivariate data, we make the following modifications to our model. First, $\boldsymbol{p}_{ij}$ is now a $2D \times 1$ observed functional outcome, formed by concatenating the X and Y coordinates of rotated motions. Second, our basis function matrix $\boldsymbol{\Theta}'$ is now the $2D \times 2K_\theta$ matrix $\left(\begin{smallmatrix} \boldsymbol{\Theta} & 0 \\ 0 & \boldsymbol{\Theta} \end{smallmatrix}\right)$, where $\boldsymbol{\Theta}$ is the $D \times K_\theta$ basis function matrix from model (5). Third, the covariance matrices in the multivariate normal distributions for $\boldsymbol{\beta}_l$, $\boldsymbol{b}_i$ and $\boldsymbol{\phi}_k$ are now the matrices (where p* represents the appropriate parameter) $\left(\begin{smallmatrix} \sigma^2_{p*,x} & 0 \\ 0 & \sigma^2_{p*,y} \end{smallmatrix}\right) \otimes \boldsymbol{P}^{-1}_{K_\theta}$, where $\otimes$ is the Kronecker product operator, $\sigma^2_{p*,x}$ and $\sigma^2_{p*,y}$ are independent with IG $[\alpha, \beta]$ priors and $\boldsymbol{P}_{K_\theta}$ is the corresponding penalty matrix from model (5). Finally, $\boldsymbol{\epsilon}_{ij}$ is now a $2D \times 1$ vector of independent error terms with a MVN $[0, \sigma^2 \boldsymbol{I}_{2D}]$ distribution. Since the FPCs are bi-dimensional in this model, each FPC represents a deviation from the mean motion in two dimensions, and each score represents the amount of that bi-dimensional mode of variation reflected in each motion. We assume independence of the first and last D coordinates of the functional random effects (each corresponding to a different coordinate of motion); further work could introduce correlations between them.

Figure A.2 illustrates the FPCs estimated using model (9) fitted to the X and Y coordinates of right hand rotated motions separately (top panels) and together using bivariate curves (bottom panels). The FPCs estimated using X and Y coordinates separately are very similar to one another. The first FPC in the bivariate model is similar to the first FPC from the model fit only to X coordinate data, and shows little variation in the Y coordinate. The second FPC in the bivariate model is similar to the first FPC from the model fit only to Y coordinate data, and shows little variation in the X coordinate. These FPCs therefore show similar patterns of variation but in different dimensions. The same pattern repeats, to a lesser extent, for the third and fourth PCs estimated using the bivariate model.

This pattern indicates that deviations from the mean motion profile in each of the dimensions represented by the X and Y coordinates are for the most part independent. The first FPC, for

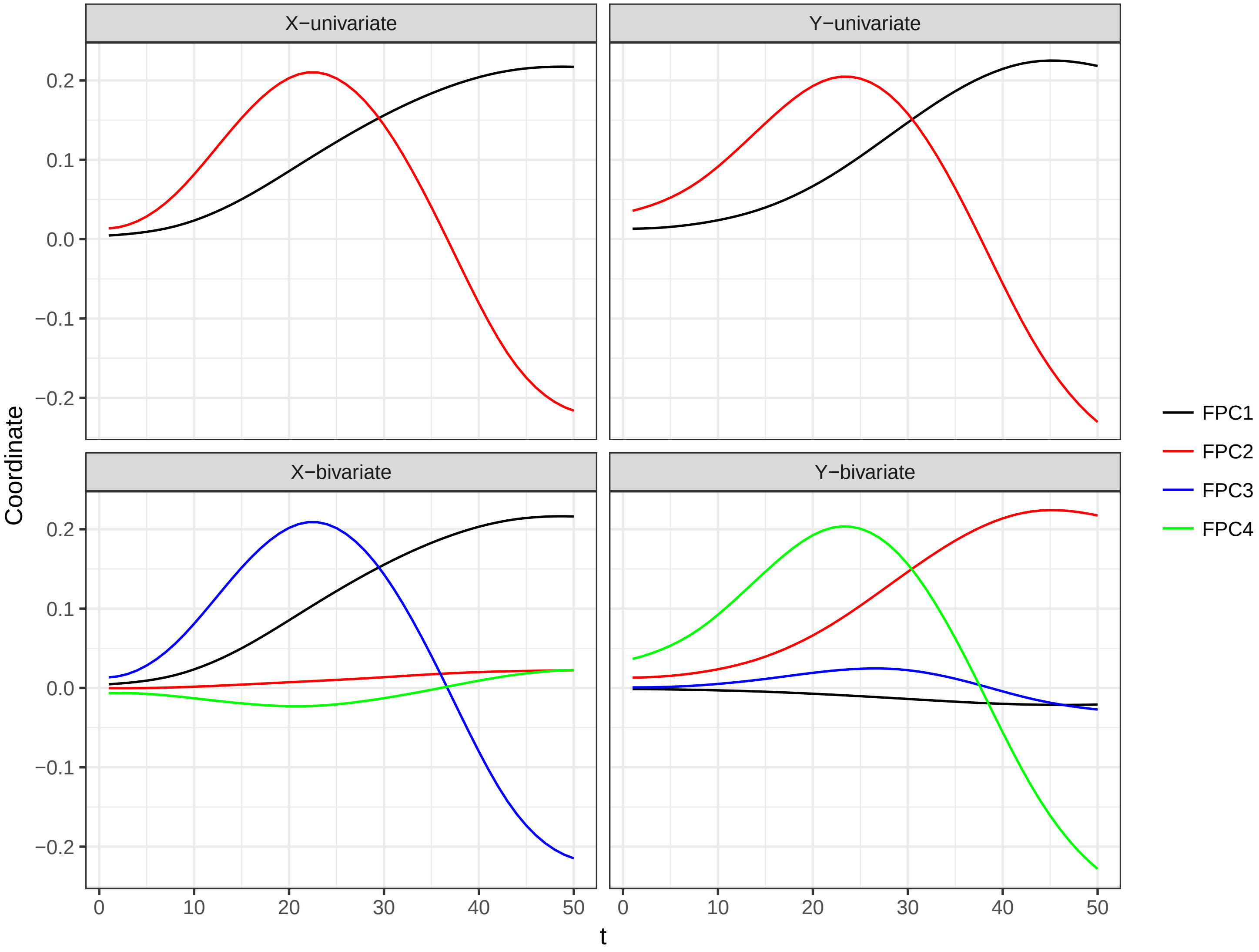

Figure A.2: FPCs from model (9) fit to the univariate and bivariate data. The FPCs on the left are for the X coordinates of motions, those on the right are for the Y-coordinate. The FPCs in the top row were estimated using univariate models, and the FPCs in the bottom row were estimated using bivariate models.

example, which represents a mode of variation in which motions overshoot or undershoot the target with respect to the line connecting the origin and target, is associated only with a slight systematic deviation upwards or downwards from this line. Likewise, the second FPC, which represents a mode of variation in which motions deviate upwards or downwards from the line connecting the origin and the target, is associated with only a slight systematic deviation in length of motion along this line. The third and fourth FPCs represent patterns in which motions are slower than average at the beginning of the motion and then faster than average later (or vice versa). There is slightly greater involvement of both dimensions in FPCs 3 and 4.

Figure A.3 shows the change in variability of first and second bivariate FPC scores as a function of practice at the motion task. For both FPCs and all targets, score variance is estimated to decrease with motion number.

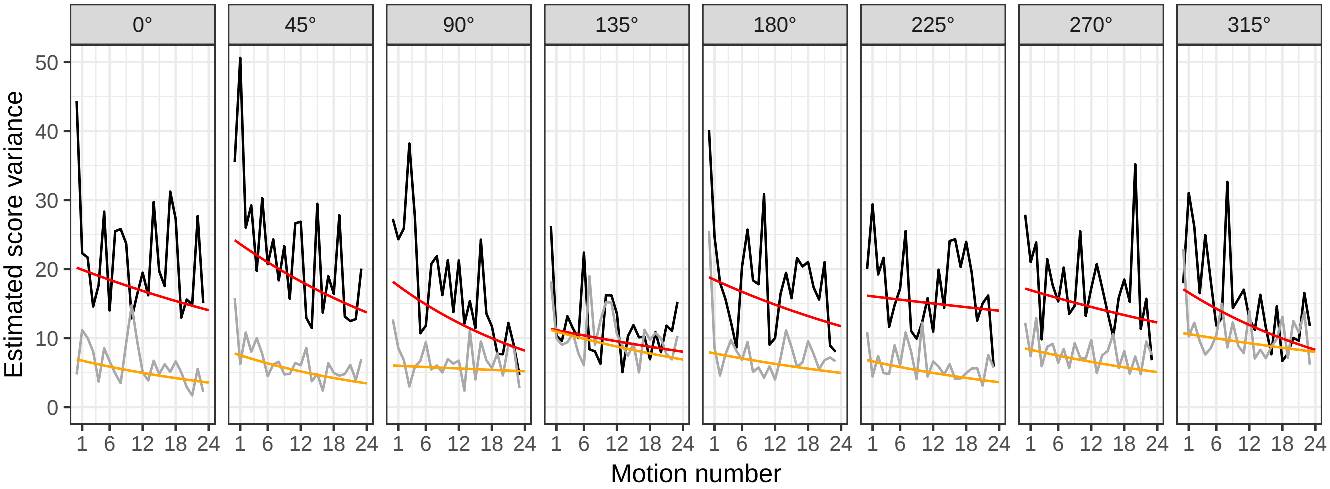

Figure A.3: Estimates of bivariate FPC score variances in the right hand for each target. Panels show the estimates of the score variance as a function of repetition number using the slope-intercept model (10) in red and orange (first and second FPC, respectively), and using the saturated one-parameter-per-repetition number model (11), in black and grey (first and second FPC, respectively).

D Sensitivity Analyses

D.1 Hyperparameters

In our sensitivity analysis we focus on the parameters of principal interest to us in the analysis in Section 6, the fixed effect parameters $\gamma_{l1,slope}$, which measure how much the variability of the first FPC scores decreases with each additional motion. We found that inference for these parameters in our VB model is not sensitive to the choice of the hyperparameters α and β in the inverse-gamma priors for the smoothing parameters $\sigma^2_{\boldsymbol{\beta}_l}$, $\sigma^2_{\boldsymbol{b}}$ and $\sigma^2_{\boldsymbol{\phi}_k}$ (we tried various combinations of values of α and β in the set $\{0.001, 0.01, 0.1, 1\}$), or to the number of spline basis functions used (we tried values in the set $\{5, 10, 15, 20\}$).

When the prior for the parameters $\gamma_{l1,int}$, which measure the baseline variance of scores for the first FPC, becomes too concentrated around zero, for example, when the variance of the mean-zero normal prior for this parameter is decreased to 1, then to compensate for the resulting severely shrunk estimates of these parameters, the estimates of $\gamma_{l1,slope}$ reverse sign. However, inference for $\gamma_{l1,slope}$ was relatively insensitive to values of the variance of this prior in the set $\{10, 100, 100\}$ (see Figures A.4 and A.5).

When using standard prior specifications for the scale matrix parameters of the inverse-Wishart priors for the random effects $\boldsymbol{g}_{ik}$ (like a diagonal identity matrix), we observed that the variance of the random effects, and credible intervals for the fixed effect parameters γ, showed dependence on the scale matrix parameters $\boldsymbol{\Psi}_k$. For this reason we use the empirical Bayes method described in Section 4.2.4 to set the value of these priors.

D.2 Mean Structure

We conducted various analyses to critically examine various modeling assumptions inherent in models (9) and (10). First, model (9) assumes that it is adequate to model the mean of the observed

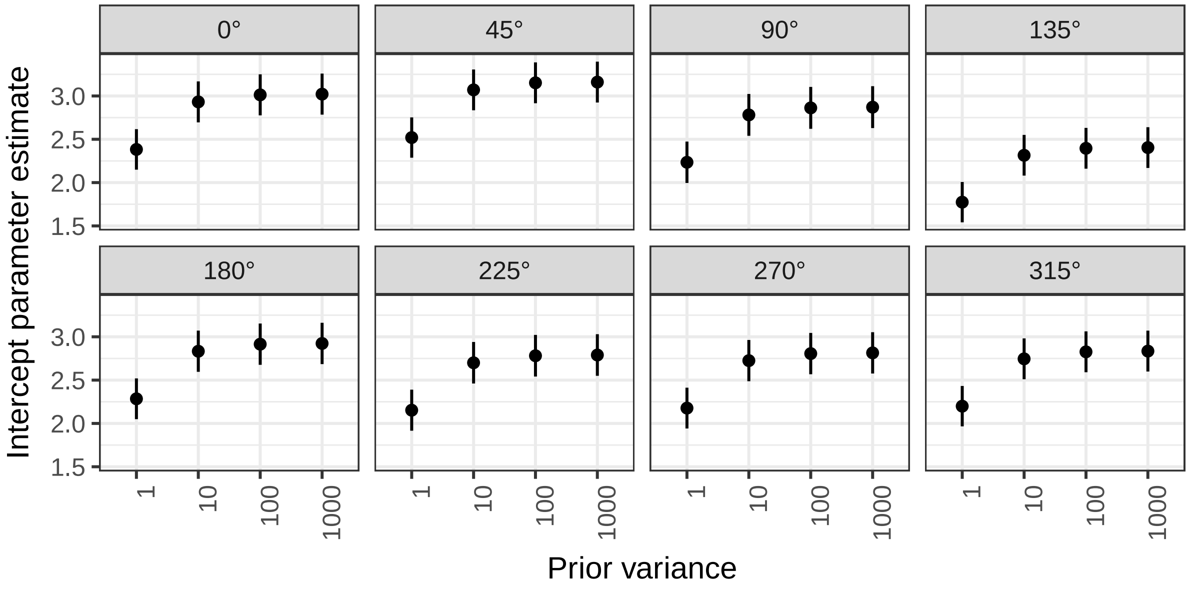

Figure A.4: Estimates and 95% credible intervals for $\gamma_{l1,int}$ as a function of the variance of its normal prior.

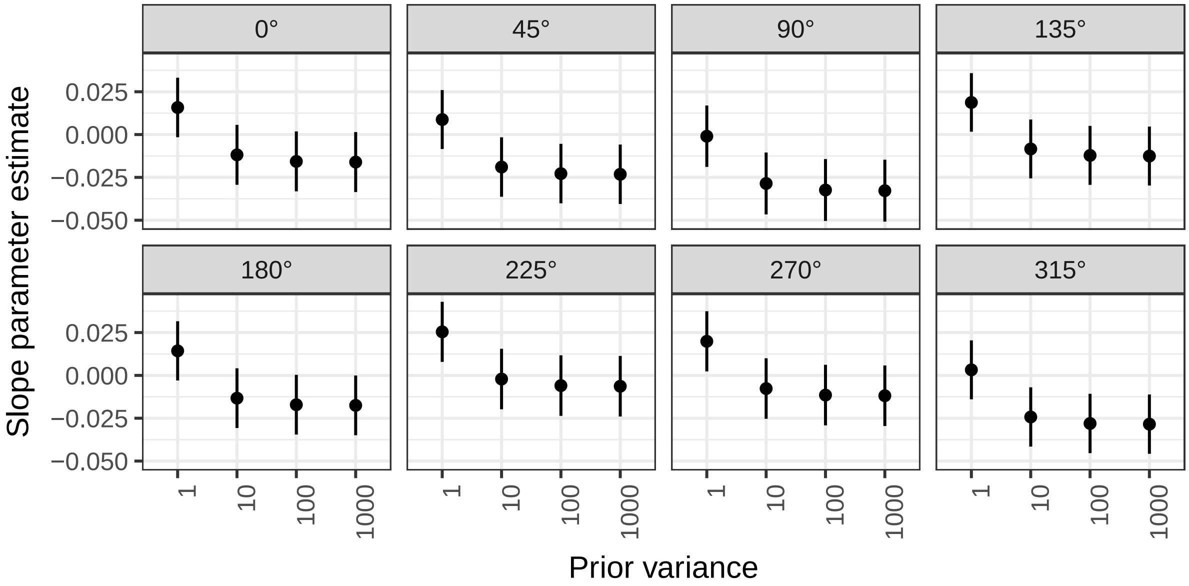

Figure A.5: Estimates and 95% credible intervals for $\gamma_{l1,slope}$ as a function of the variance of its normal prior.

curves with a functional intercept for each target and random functional effects for each subject-target combination. If the mean motion to a target systematically changed as a function of repetition number, then scores at the beginning or end of the training session might be inflated, which could lead to over- or under-estimation of our parameter of principal interest, the motion number score variance slopes $\gamma_{l1,slope}$. To examine this possibility, we conducted an analysis, restricted to data for right hand motions to target 0°, in which we fit 4 separate functional random effects for each subject, for 4 groups of consecutive motions (motions 1 through 6, motions 7 through 12, et cetera). We found that inference for the slope parameter $\gamma_{11,slope}$ was unchanged, suggesting that model (9) is adequate.

Models (9) and (10) also make several simplifying independence assumptions. First, we assume independence of functional random effects for motions made by the same subject to different targets. Analysis of more complex models that modeled correlation between these functional random effects showed that although taking into account these correlations did shrink together functional random effects for the same subject, it did not change inference for our parameters of interest in the model above, the score variance repetition number slope parameters $\gamma_{l1,slope}$. Second, we assume independence of functional random effects and score variance random effects. In an ad hoc analysis to check the effects of this simplifying assumption, we included the endpoint of the estimated functional random effects as a predictor in our score variance model for data for right hand motions to target 0°. Although the 95% credible interval for this endpoint parameter did not include 0, its inclusion in the score variance model did not alter the credible interval for the repetition number slope parameter. In other contexts, for example, motions by stroke patients, correlations between functional and score variance model random effects might be stronger, and might need to be taken into account in order for inference to be correct.

E Derivations

This section includes derivations of conditional distributions of all quantities in model (5), an overview of variational Bayes, a derivation of our variational Bayes algorithm, and additional details on the implementation of our HMC sampler. The derivations of conditional distributions are included because they are used in the derivation of our variational Bayes algorithm. Throughout this section we consider a model where each subject has one functional random effect $\boldsymbol{b}_i$. It is straightforward to extend the derivations below to the case where there are different functional random effects $\boldsymbol{b}_{im}$ for different sets of curves for each subject.

E.1 Derivation of conditional distributions

Let $n = \sum_{i=1}^{I} J_i$ be the total number of motions by all subjects. Let $\boldsymbol{P}$ be the $D \times n$ matrix of functional outcomes, $\boldsymbol{\beta}$ the $K_\theta \times (L+1)$ matrix of fixed effect coefficient vectors and $\boldsymbol{X}$ the corresponding $n \times (l+1)$ fixed effects design matrix, $\boldsymbol{B}$ the $K_\theta \times I$ matrix of random effect coefficient vectors and $\boldsymbol{V}$ the corresponding $n \times I$ random effects design matrix, $\boldsymbol{\Phi}$ the $K_\theta \times K$ matrix of principal component coefficient vectors and $\boldsymbol{\Xi}$ the corresponding $n \times K$ matrix of principal component scores and $\boldsymbol{E}$ the $D \times n$ error matrix of error vectors $\boldsymbol{\epsilon}_i$.

We rewrite our model using matrix notation as follows:

$$\boldsymbol{P} = \boldsymbol{\Theta}\boldsymbol{\beta}\boldsymbol{X}^T + \boldsymbol{\Theta}\boldsymbol{B}\boldsymbol{V}^T + \boldsymbol{\Theta}\boldsymbol{\Phi}\boldsymbol{\Xi}^T + \boldsymbol{E}$$

We will first derive the posterior distribution of $\boldsymbol{\beta}$ conditional on the values of the other parameters in the model. Let $\boldsymbol{\sigma}^2_{\boldsymbol{\beta}}$ be the length $L+1$ vector of prior variances $\sigma^2_{\boldsymbol{\beta}_l}$ or, in the model with bivariate observations, the length $2L+2$ vector of prior variances $(\sigma^2_{\boldsymbol{\beta}^x_0}\sigma^2_{\boldsymbol{\beta}^y_0}, \ldots, \sigma^2_{\boldsymbol{\beta}^x_L}, \sigma^2_{\boldsymbol{\beta}^y_L})$. Let $\text{vec}\,(\boldsymbol{M})$ be the vector formed by concatenating the columns of the matrix $\boldsymbol{M}$. Then the covariance matrix of the normal prior distribution of $\text{vec}\,(\boldsymbol{\beta})$ is $\boldsymbol{\Sigma}_{\boldsymbol{\beta}} = \text{diag}\left(\boldsymbol{\sigma}^2_{\boldsymbol{\beta}}\right) \otimes \boldsymbol{Q}^{-1}$, where $\text{diag}\,(\boldsymbol{c})$ is the matrix with the elements of $\boldsymbol{c}$ on its main diagonal and 0 elsewhere and $\otimes$ is the Kronecker product operator. The posterior distribution of $\text{vec}\,(\boldsymbol{\beta})$ is then

$$\begin{aligned} p(\text{vec}\,(\boldsymbol{\beta})\,|\text{rest}) &\propto p(\text{vec}\,(\boldsymbol{P})\,|\boldsymbol{\beta}, \boldsymbol{B}, \boldsymbol{\Phi}, \Xi, \sigma^2) p(\text{vec}\,(\boldsymbol{\beta})\,|\boldsymbol{\Sigma}_{\boldsymbol{\beta}}) \\ &\propto \exp\left\{-\frac{1}{2}\left[\frac{1}{\sigma^2}\|\text{vec}\left(\boldsymbol{P} - \boldsymbol{\Theta}\boldsymbol{\beta}\boldsymbol{X}^T - \boldsymbol{\Theta}\boldsymbol{B}\boldsymbol{V}^T - \boldsymbol{\Theta}\boldsymbol{\Phi}\Xi^T\right)\|^2 + \text{vec}\,(\boldsymbol{\beta})^T\,\boldsymbol{\Sigma}_{\boldsymbol{\beta}}^{-1}\text{vec}\,(\boldsymbol{\beta})\right]\right\} \end{aligned}$$

Using the identity

$$\text{vec}\,(\boldsymbol{A}\boldsymbol{B}\boldsymbol{C}) = (\boldsymbol{C}^T \otimes \boldsymbol{A})\text{vec}\,(\boldsymbol{B}) \tag{A.1}$$

we see that the exponent in this posterior distribution is a quadratic in $\text{vec}\,(\boldsymbol{\beta})$, and so the posterior distribution is multivariate normal. The inverse of the coefficient of the quadratic term is the covariance matrix of this posterior distribution:

$$\boldsymbol{\Sigma}'_{\boldsymbol{\beta}} = \left[(\boldsymbol{X} \otimes \boldsymbol{\Theta})^T \frac{1}{\sigma^2}(\boldsymbol{X} \otimes \boldsymbol{\Theta}) + \boldsymbol{\Sigma}_{\boldsymbol{\beta}}^{-1}\right]^{-1}.$$

This covariance matrix multiplied by the linear term of this exponent gives the mean of this posterior distribution:

$$\boldsymbol{\mu}'_{\boldsymbol{\beta}} = \boldsymbol{\Sigma}'_{\boldsymbol{\beta}}(\boldsymbol{X} \otimes \boldsymbol{\Theta})^T \frac{1}{\sigma^2}\left[\text{vec}\left(\boldsymbol{P} - \boldsymbol{\Theta}\boldsymbol{B}\boldsymbol{V}^T - \boldsymbol{\Theta}\boldsymbol{\Phi}\Xi^T\right)\right].$$

The derivations of the conditional posterior distributions of $\boldsymbol{B}$ and $\boldsymbol{\Phi}$ are similar. Let $\boldsymbol{b}_i$ be the random effect for the ith subject. The covariance matrix of the normal prior distribution of $\boldsymbol{b}_i$ is $\boldsymbol{\Sigma}_{\boldsymbol{b}} = \text{diag}\,(\boldsymbol{\sigma}_{\boldsymbol{b}}^2) \otimes ((1-\pi)\boldsymbol{Q} + \pi\boldsymbol{I})^{-1}$, where, in the model with bivariate observations, $\boldsymbol{\sigma}_{\boldsymbol{b}}^2 = (\sigma_{\boldsymbol{b}^x}^2, \sigma_{\boldsymbol{b}^y}^2)$. Let $\boldsymbol{P}_i, \boldsymbol{X}_i$ and Ξ_i be the submatrices of the matrices $\boldsymbol{P}, \boldsymbol{X}$ and Ξ corresponding to the observations for the ith subject. The posterior distribution of $\boldsymbol{b}_i$ is then

$$\begin{aligned} p(\boldsymbol{b}_i|\text{rest}) &\propto p(\text{vec}\left(\boldsymbol{P}_i\right)|\boldsymbol{\beta}, \boldsymbol{b}_i, \boldsymbol{\Phi}, \boldsymbol{\Xi}_i, \sigma^2) p(\text{vec}\left(\boldsymbol{b}_i\right)|\boldsymbol{\Sigma}_{\boldsymbol{b}}) \\ &\propto \exp\left\{-\frac{1}{2}\left[\frac{1}{\sigma^2}\|\text{vec}\left(\boldsymbol{P}_i - \boldsymbol{\Theta}\boldsymbol{\beta}\boldsymbol{X}_i^T - \boldsymbol{\Theta}\boldsymbol{b}_i\boldsymbol{1}_{J_i}^T - \boldsymbol{\Theta}\boldsymbol{\Phi}\boldsymbol{\Xi}_i^T\right)\|^2 + \boldsymbol{b}_i^T\boldsymbol{\Sigma}_{\boldsymbol{b}}^{-1}\boldsymbol{b}_i\right]\right\}, \end{aligned}$$

that is, multivariate normal with covariance matrix

$$\boldsymbol{\Sigma}'_{\boldsymbol{b}} = \left[(\boldsymbol{1}_{J_i} \otimes \boldsymbol{\Theta})^T \frac{1}{\sigma^2} (\boldsymbol{1}_{J_i} \otimes \boldsymbol{\Theta}) + \boldsymbol{\Sigma}_{\boldsymbol{b}}^{-1}\right]^{-1}$$

and mean

$$\boldsymbol{\mu}'_{\boldsymbol{b}_i} = \boldsymbol{\Sigma}'_{\boldsymbol{b}} (\boldsymbol{1}_{J_i} \otimes \boldsymbol{\Theta})^T \frac{1}{\sigma^2}\left[\text{vec}\left(\boldsymbol{P}_i - \boldsymbol{\Theta}\boldsymbol{\beta}\boldsymbol{X}_i^T - \boldsymbol{\Theta}\boldsymbol{\Phi}\boldsymbol{\Xi}_i^T\right)\right].$$

Letting $\boldsymbol{\sigma}_{\boldsymbol{\Phi}}^2$ be the length K vector of prior variances $\sigma_{\phi_k}^2$ (or, in the model with bivariate observations, the length $2K$ vector $(\sigma_{\phi_1^x}^2, \sigma_{\phi_1^y}^2, \ldots, \sigma_{\phi_K^x}^2, \sigma_{\phi_K^y}^2)$), the covariance matrix of the normal prior distribution of $\text{vec}\left(\boldsymbol{\Phi}\right)$ is $\boldsymbol{\Sigma}_{\boldsymbol{\Phi}} = \text{diag}\left(\boldsymbol{\sigma}_{\boldsymbol{\Phi}}^2\right) \otimes \boldsymbol{Q}^{-1}$. The posterior distribution of $\text{vec}\left(\boldsymbol{\Phi}\right)$ is then

$$\begin{aligned} p(\text{vec}\left(\boldsymbol{\Phi}\right)|\text{rest}) &\propto p(\text{vec}\left(\boldsymbol{P}\right)|\boldsymbol{\beta}, \boldsymbol{B}, \boldsymbol{\Phi}, \Xi, \sigma^2) p(\text{vec}\left(\boldsymbol{\Phi}\right)|\boldsymbol{\Sigma}_{\boldsymbol{\Phi}}) \\ &\propto \exp\left\{-\frac{1}{2}\left[\frac{1}{\sigma^2}\|\text{vec}\left(\boldsymbol{P} - \boldsymbol{\Theta}\boldsymbol{\beta}\boldsymbol{X}^T - \boldsymbol{\Theta}\boldsymbol{B}\boldsymbol{V}^T - \boldsymbol{\Theta}\boldsymbol{\Phi}\boldsymbol{\Xi}^T\right)\|^2 + \text{vec}\left(\boldsymbol{\Phi}\right)^T \boldsymbol{\Sigma}_{\boldsymbol{\Phi}}^{-1} \text{vec}\left(\boldsymbol{\Phi}\right)\right]\right\}, \end{aligned}$$

that is, multivariate normal with covariance matrix

$$\boldsymbol{\Sigma}'_{\boldsymbol{\Phi}} = \left[(\boldsymbol{\Xi} \otimes \boldsymbol{\Theta})^T \frac{1}{\sigma^2} (\boldsymbol{\Xi} \otimes \boldsymbol{\Theta}) + \boldsymbol{\Sigma}_{\boldsymbol{\Phi}}^{-1}\right]^{-1}$$

and mean

$$\boldsymbol{\mu}'_{\boldsymbol{\Phi}} = \boldsymbol{\Sigma}'_{\boldsymbol{\Phi}} (\boldsymbol{\Xi} \otimes \boldsymbol{\Theta})^T \frac{1}{\sigma^2}\left[\text{vec}\left(\boldsymbol{P} - \boldsymbol{\Theta}\boldsymbol{\beta}\boldsymbol{X}^T - \boldsymbol{\Theta}\boldsymbol{B}\boldsymbol{V}^T\right)\right].$$

To compute the conditional posterior distribution of $\boldsymbol{\xi}_{ij}$, the vector of scores for the jth motion for the ith subject, we let the covariance matrix of the normal prior distribution of $\boldsymbol{\xi}_{ij}$ be $\boldsymbol{\Sigma}_{\boldsymbol{\xi}_{ij}} = \text{diag}\left(\boldsymbol{\sigma}_{\boldsymbol{\xi}ij}^2\right)$, where $\boldsymbol{\sigma}_{\boldsymbol{\xi}_{ij}}^2$ is the length K vector of prior variances for $\boldsymbol{\xi}_{ij}$. Then the posterior distribution of $\boldsymbol{\xi}_{ij}$ is

$$\begin{aligned}
&p(\boldsymbol{\xi}_{ij}|\text{rest})\\
&\quad\propto p(\boldsymbol{p}_{ij}|\boldsymbol{\beta},\boldsymbol{b}_i,\boldsymbol{\Phi},\boldsymbol{\xi}_{ij},\sigma^2)p(\boldsymbol{\xi}_{ij}|\boldsymbol{\Sigma}_{\boldsymbol{\xi}_{ij}})\\
&\quad\propto \exp\left(-\frac{1}{2}\left\{\frac{1}{\sigma^2}\|\boldsymbol{p}_{ij}-\boldsymbol{\Theta}\boldsymbol{\beta}\boldsymbol{x}_{ij}-\boldsymbol{\Theta}\boldsymbol{b}_i-\boldsymbol{\Theta}\boldsymbol{\Phi}\boldsymbol{\xi}_{ij}\|^2+\boldsymbol{\xi}_{ij}^T\boldsymbol{\Sigma}_{\boldsymbol{\xi}_{ij}}^{-1}\boldsymbol{\xi}_{ij}\right\}\right),
\end{aligned}$$

that is, multivariate normal with covariance matrix

$$\boldsymbol{\Sigma}'_{\boldsymbol{\xi}_{ij}}=\left\{\frac{1}{\sigma^2}\boldsymbol{\Phi}^T\boldsymbol{\Theta}^T\boldsymbol{\Theta}\boldsymbol{\Phi}+\boldsymbol{\Sigma}_{\boldsymbol{\xi}_{ij}}^{-1}\right\}^{-1}$$

and mean

$$\boldsymbol{\mu}'_{\boldsymbol{\xi}_{ij}}=\boldsymbol{\Sigma}'_{\boldsymbol{\xi}_{ij}}\boldsymbol{\Phi}^T\boldsymbol{\Theta}^T\frac{1}{\sigma^2}\left(\boldsymbol{p}_{ij}-\boldsymbol{\Theta}\boldsymbol{\beta}\boldsymbol{x}_{ij}-\boldsymbol{\Theta}\boldsymbol{b}_i\right).$$

In the model for the variance of the kth principal component scores, let $\boldsymbol{x}^*_{ijk}$ be the length L^*+1 vector of fixed effect coefficients for the jth motion by the ith subject and $\boldsymbol{\gamma}_k$ the corresponding vector of fixed effects, shared across all subjects and motions, and let $\boldsymbol{z}^*_{ijk}$ be the length M^* vector of random effect coefficients for the jth motion by the ith subject and $\boldsymbol{g}_{ik}$ the corresponding vector of random effects for the ith subject. If we let $\boldsymbol{\sigma}^2_{\gamma_k}$ be the vector of the $\sigma^2_{\gamma_{lk}}$, the prior variances of the components of γ_k, then the covariance matrix of the prior distribution of $\boldsymbol{\gamma}_k$ is $\boldsymbol{\Sigma}_{\gamma_k}=\text{diag}\left(\boldsymbol{\sigma}^2_{\gamma_k}\right)$. Let the covariance matrix of the prior distribution of $\boldsymbol{g}_{ik}$ be $\boldsymbol{\Sigma}_{g_k}$. The conditional posterior distribution of $\boldsymbol{\gamma}_k$ and the vectors $\boldsymbol{g}_{ik}, i=1,\dots,I$ is then

$$\begin{aligned}
p(\boldsymbol{\gamma}_k,\boldsymbol{g}_{1k},\boldsymbol{g}_{2k},\dots,\boldsymbol{g}_{Ik}|\text{rest})&\propto\left(\prod_{i=1}^{I}\prod_{j=1}^{J_i}p(\xi_{ijk}|\boldsymbol{\gamma}_k,\boldsymbol{g}_{ik})\right)p(\boldsymbol{\gamma}_k)\left(\prod_{i=1}^{I}p(\boldsymbol{g}_{ik})\right)\\
&\propto\left(\prod_{i=1}^{I}\prod_{j=1}^{J_i}\frac{e^{-\xi^2_{ijk}/2e^{(\boldsymbol{\gamma}_k\boldsymbol{x}^*_{ijk}+\boldsymbol{g}_{ik}\boldsymbol{z}^*_{ijk})}}}{e^{(\boldsymbol{\gamma}_k\boldsymbol{x}^*_{ijk}+\boldsymbol{g}_{ik}\boldsymbol{z}^*_{ijk})/2}}\right)\exp\left[-\frac{1}{2}\left(\boldsymbol{\gamma}_k^T\boldsymbol{\Sigma}_{\gamma_k}\gamma_k+\sum_{i=1}^{I}\boldsymbol{g}_{ik}^T\boldsymbol{\Sigma}_{\boldsymbol{g}_k}\boldsymbol{g}_{ik}\right)\right],
\end{aligned}$$

which has the form of the posterior of a gamma generalized linear model with log link, responses given by ξ^2_{ijk}, shape parameter equal to $1/2$ and a mean-zero multivariate normal prior on the coefficients $\boldsymbol{\gamma}_k$ and $\boldsymbol{g}_{ik}, i = 1, \ldots, I$, with covariance matrix determined by $\boldsymbol{\Sigma}_{\gamma_k}$ and $\boldsymbol{\Sigma}_{\boldsymbol{g}_k}$.

Now we derive the conditional distributions of the variance parameters, starting with $\sigma^2_{\boldsymbol{\beta}_l}$. The inverse gamma density is $p(x|\alpha,\beta) = \frac{\beta^\alpha}{\Gamma(\alpha)} x^{-\alpha-1} \exp\left(-\frac{\beta}{x}\right)$. Therefore the posterior distribution of $\sigma^2_{\boldsymbol{\beta}_l}$ is

$$\begin{aligned} p(\sigma^2_{\boldsymbol{\beta}_l}|\text{rest}) &\propto p(\sigma^2_{\boldsymbol{\beta}_l}|\alpha,\beta) p(\boldsymbol{\beta}_l|\sigma^2_{\boldsymbol{\beta}_l}) \\ &\propto \left(\sigma^2_{\boldsymbol{\beta}_l}\right)^{-\alpha-1} \exp\left(-\frac{\beta}{\sigma^2_{\boldsymbol{\beta}_l}}\right) \frac{1}{\left(\sigma^2_{\boldsymbol{\beta}_l}\right)^{K_\theta/2}} \exp\left(-\frac{1}{2\sigma^2_{\boldsymbol{\beta}_l}} \boldsymbol{\beta}_l^T \boldsymbol{Q} \boldsymbol{\beta}_l\right) \\ &\propto \text{IG}\left[\alpha + \frac{K_\theta}{2}, \beta + \frac{1}{2}\boldsymbol{\beta}_l^T \boldsymbol{Q} \boldsymbol{\beta}_l\right]. \end{aligned}$$

For this variance parameter and also for the variance parameters $\sigma^2_{\boldsymbol{b}}$ and $\sigma^2_{\boldsymbol{\phi}_k}$, the conditional posterior distributions are the same in the model with bivariate observations, except that, for example, in the conditional posterior distribution of $\sigma^2_{\boldsymbol{\beta}_l^x}$, the quadratic form in the expression for the second parameter of the inverse gamma posterior distribution is computed with respect to only the first K_θ components of the vector $\boldsymbol{\beta}_l$. In the conditional distribution of $\sigma^2_{\boldsymbol{\beta}_l^y}$, the remaining components of $\boldsymbol{\beta}_l$ are used. The conditional distribution of $\sigma^2_{\boldsymbol{b}}$ is similar:

$$\begin{aligned} p(\sigma^2_{\boldsymbol{b}}|\text{rest}) &\propto p(\sigma^2_{\boldsymbol{b}}|\alpha,\beta) \prod_{i=1}^{I} p(\boldsymbol{b}_i|\sigma^2_{\boldsymbol{b}}) \\ &\propto \left(\sigma^2_{\boldsymbol{b}}\right)^{-\alpha-1} \exp\left(-\frac{\beta}{\sigma^2_{\boldsymbol{b}}}\right) \frac{1}{\left(\sigma^2_{\boldsymbol{b}}\right)^{IK_\theta/2}} \exp\left(-\frac{1}{2\sigma^2_{\boldsymbol{b}}} \sum_{i=1}^{I} \boldsymbol{b}_i^T((1-\pi)\boldsymbol{Q} + \pi\boldsymbol{I})\boldsymbol{b}_i\right) \\ &\propto \text{IG}\left[\alpha + \frac{IK_\theta}{2}, \beta + \frac{1}{2}\sum_{i=1}^{I} \boldsymbol{b}_i^T((1-\pi)\boldsymbol{Q} + \pi\boldsymbol{I})\boldsymbol{b}_i\right], \end{aligned}$$

as is the conditional distribution of $\sigma^2_{\boldsymbol{\phi}_k}$:

$$\begin{aligned}
p(\sigma^2_{\boldsymbol{\phi}_k}|\text{rest}) &\propto p(\sigma^2_{\boldsymbol{\phi}_k}|\alpha,\beta)p(\boldsymbol{\phi}_k|\sigma^2_{\boldsymbol{\phi}_k}) \\
&\propto \left(\sigma^2_{\boldsymbol{\phi}_k}\right)^{-\alpha-1}\exp\left(-\frac{\beta}{\sigma^2_{\boldsymbol{\phi}_k}}\right)\frac{1}{\left(\sigma^2_{\boldsymbol{\phi}_k}\right)^{K_\theta/2}}\exp\left(-\frac{1}{2\sigma^2_{\boldsymbol{\phi}_k}}\boldsymbol{\phi}_k^T\boldsymbol{Q}\boldsymbol{\phi}_k\right) \\
&\propto \text{IG}\left[\alpha+\frac{K_\theta}{2},\beta+\frac{1}{2}\boldsymbol{\phi}_k^T\boldsymbol{Q}\boldsymbol{\phi}_k\right],
\end{aligned}$$

of σ^2:

$$\begin{aligned}
p(\sigma^2|\text{rest}) &\propto p(\sigma^2|\alpha,\beta)p(\text{vec}\,(\boldsymbol{P})\,|\boldsymbol{\beta},\boldsymbol{B},\boldsymbol{\Phi},\boldsymbol{\Xi},\sigma^2) \\
&\propto \left(\sigma^2\right)^{-\alpha-1}\exp\left(-\frac{\beta}{\sigma^2}\right)\frac{1}{\left(\sigma^2\right)^{nD/2}}\exp\left[-\frac{1}{2\sigma^2}\|\text{vec}\left(\boldsymbol{P}-\boldsymbol{\Theta\beta X}^T-\boldsymbol{\Theta B V}^T-\boldsymbol{\Theta\Phi\Xi}^T\right)\|^2\right] \\
&\propto \text{IG}\left[\alpha+\frac{nD}{2},\beta+\frac{1}{2}\|\text{vec}\left(\boldsymbol{P}-\boldsymbol{\Theta\beta X}^T-\boldsymbol{\Theta B V}^T-\boldsymbol{\Theta\Phi\Xi}^T\right)\|^2\right],
\end{aligned}$$

and of $\sigma^2_{g_k}$ (this is the case where there is just one scalar random effect):

$$\begin{aligned}
p(\sigma^2_{g_k}|\text{rest}) &\propto p(\sigma^2_{g_k}|\alpha,\beta)\prod_{i=1}^{I}p(g_{ik}|\sigma^2_{g_k}) \\
&\propto \left(\sigma^2_{g_k}\right)^{-\alpha-1}\exp\left(-\frac{\beta}{\sigma^2_{g_k}}\right)\frac{1}{\left(\sigma^2_{g_k}\right)^{I/2}}\exp\left(-\frac{1}{2\sigma^2_{g_k}}\sum_{i=1}^{I}g_{ik}^2\right) \\
&\propto \text{IG}\left[\alpha+\frac{I}{2},\beta+\frac{1}{2}\sum_{i=1}^{I}g_{ik}^2\right].
\end{aligned}$$

In our real data application, we consider a model where two random effects $g_{ik,int}$ and $g_{ik,slope}$ have a bivariate, mean-zero normal prior distribution with covariance matrix $\boldsymbol{\Sigma}_{g_k}$. This covariance matrix has an inverse-Wishart prior distribution. The inverse-Wishart density is $p(\boldsymbol{\Sigma}|\boldsymbol{\Psi},\nu) = |\boldsymbol{\Sigma}|^{-\frac{\nu+p+1}{2}}\exp\left(-\frac{1}{2}\text{tr}\left[\boldsymbol{\Psi\Sigma}^{-1}\right]\right)$, where p is the number of rows and columns of the covariance matrix

$\boldsymbol{\Sigma}$. The conditional posterior distribution of $\boldsymbol{\Sigma}_{g_k}$ is therefore

$$\begin{aligned}
p(\boldsymbol{\Sigma}_{g_k}|\text{rest}) &\propto p(\boldsymbol{\Sigma}_{g_k}) \prod_{i=1}^{I} p(\boldsymbol{g}_{ik}|\boldsymbol{\Sigma}_g) \\
&\propto |\boldsymbol{\Sigma}_{g_k}|^{-\frac{\nu+p+1}{2}} \exp\left(-\frac{1}{2}\text{tr}\left[\boldsymbol{\Psi}\boldsymbol{\Sigma}_{g_k}^{-1}\right]\right) |\boldsymbol{\Sigma}_{g_k}|^{-I/2} \exp\left(-\frac{1}{2}\sum_{i=1}^{I} \boldsymbol{g}_{ik}^T \boldsymbol{\Sigma}_{g_k}^{-1} \boldsymbol{g}_{ik}\right) \\
&\propto |\boldsymbol{\Sigma}_{g_k}|^{-\frac{\nu+p+I+1}{2}} \exp\left[-\frac{1}{2}\left(\sum_{i=1}^{I} \text{tr}\left[\boldsymbol{g}_{ik}\boldsymbol{g}_{ik}^T \boldsymbol{\Sigma}_{g_k}^{-1}\right] + \text{tr}\left[\boldsymbol{\Psi}\boldsymbol{\Sigma}_{g_k}^{-1}\right]\right)\right] \\
&\propto \text{IW}\left[\boldsymbol{\Psi} + \sum_{i=1}^{I} \boldsymbol{g}_{ik}\boldsymbol{g}_{ik}^T, \nu + I\right].
\end{aligned}$$

Straightforward extensions of these derivations apply in the case of nested random effects, as in model extension (6).

E.2 Overview of variational Bayes

Let $\boldsymbol{y}$ and $\boldsymbol{\zeta}$ represent the data and parameters, respectively, in a Bayesian model. Using variational Bayes, we approximate the posterior $p(\boldsymbol{\zeta}|\boldsymbol{y})$ using $q(\boldsymbol{\zeta})$, where q is a member of a restricted class of functions Q more easily estimated than the posterior $p(\boldsymbol{\zeta}|\boldsymbol{y})$. To find the best q in this restricted class, we choose the element $q^* \in Q$ that minimizes the Kullback-Leibler distance from $p(\boldsymbol{\zeta}|\boldsymbol{y})$. The class Q is often the class of posterior distributions satisfying some factorization property, so that $q(\boldsymbol{\zeta}) = \prod_{h=1}^{H} q_h(\zeta_h)$, with each $q_h(\zeta_h)$ a parametric density function. It can then be shown that the optimal q_h^* densities are given by

$$q_h^*(\zeta_h) \propto \exp\left[E_{-\zeta_h} \log p(\zeta_h|\text{rest})\right] \tag{A.2}$$

where $E_{-\zeta_h}$ represents the expectation with respect to the currently estimated values of all parameters except ζ_h, and "rest" represents the observed data plus all parameters other than ζ_h. This suggests the use of an iterative algorithm, setting initial values for all parameters and then updating the optimal density for each parameter ζ_h in turn, conditionally on the currently estimated values for all the other parameters.

Let $\{\sigma_s^2\}_{s\in S}$ represent the collection of all variance parameters in model (5). Let $\boldsymbol{\xi}_{ij}$ represent the vector of scores for the jth motion of the ith subject. The factorization we use to approximate the posterior distribution $q(\boldsymbol{\zeta})$ for model (5) is

$$q(\boldsymbol{\beta}_0,\ldots,\boldsymbol{\beta}_L)\left\{\prod_{i=1}^{I}\prod_{m=1}^{M}q(\boldsymbol{b}_{im})\right\}q(\boldsymbol{\phi}_1,\ldots,\boldsymbol{\phi}_K)\left\{\prod_{i=1}^{I}\prod_{j=1}^{J_i}q(\boldsymbol{\xi}_{ij})\right\}\left\{\prod_{k=1}^{K}q(\gamma_{0k},\ldots,g_{11k},\ldots,)\right\}\left\{\prod_{s\in S}q(\sigma_s^2)\right\} \tag{A.3}$$

In the case of the model extension (6), each term $\boldsymbol{g}_{ik}$ would have its own factor $q(\boldsymbol{g}_{ik})$ in the factorization above.

The quality of this approximation depends on the extent to which the true posterior distribution factors as above. It is expected that the parameters in the curve mean $\mu_{ij}(t)$ and the deviation $\delta_{ij}(t)$ will be correlated, which suggests that assumptions underlying the variational approximation will be violated for these components of the posterior. Nonetheless, the assumptions related to the score variance model, which is our main interest, may be sufficiently accurate to provide a reasonable approximation.

E.3 Derivation of variational Bayes algorithm

To find the optimal $q^*(\cdot)$ distributions for $\boldsymbol{\beta}, \boldsymbol{B}, \boldsymbol{\Phi}$ and $\boldsymbol{\Xi}$, we use the following result: if the conditional distribution of a parameter ζ is multivariate normal with mean $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$, then the distribution $q^*(\zeta)$ is multivariate normal with covariance matrix $\Sigma_{q(\zeta)} = \left(E_{-\zeta}\left[\boldsymbol{\Sigma}^{-1}\right]\right)^{-1}$ and mean $\mu_{q(\zeta)} = \left(E_{-\zeta}\left[\boldsymbol{\Sigma}^{-1}\right]\right)^{-1} E_{-\zeta}\left[\boldsymbol{\Sigma}^{-1}\mu\right]$, where we use the notation $\mu_{q(\zeta)}$ and $\Sigma_{q(\zeta)}$, respectively, to denote the mean and variance of a parameter ζ under its optimal q^* distribution.

Throughout this section we make extensive use of the conditional distributions derived in Appendix E.1.

For vec ($\boldsymbol{\beta}$), the optimal density $q^*(\mathrm{vec}\,(\beta))$ is thus multivariate normal with covariance matrix

$$\boldsymbol{\Sigma}_{q(\mathrm{vec}(\beta))} = \left[\mu_{q\left(\frac{1}{\sigma^2}\right)}((\boldsymbol{X} \otimes \boldsymbol{\Theta})^T(\boldsymbol{X} \otimes \boldsymbol{\Theta})) + \mathrm{diag}\left(\mu_{q\left(1/\sigma^2_{\beta_l}\right)}\right) \otimes \boldsymbol{Q}\right]^{-1}$$

and mean

$$\mu_{q(\mathrm{vec}(\beta))} = \boldsymbol{\Sigma}_{q(\mathrm{vec}(\beta))}(\boldsymbol{X} \otimes \boldsymbol{\Theta})^T \mu_{q\left(\frac{1}{\sigma^2}\right)} \left[\mathrm{vec}\left(\boldsymbol{P} - \boldsymbol{\Theta}\mu_{q(B)}\boldsymbol{V}^T - \boldsymbol{\Theta}\mu_{q(\Phi)}\mu^T_{q(\Xi)}\right)\right].$$

For $\boldsymbol{b}_i$, the optimal density $q^*(\boldsymbol{b}_i)$ is multivariate normal with covariance matrix

$$\boldsymbol{\Sigma}_{q(b_i)} = \left[\mu_{q\left(\frac{1}{\sigma^2}\right)}(\mathbf{1}_{J_i} \otimes \boldsymbol{\Theta})^T(\mathbf{1}_{J_i} \otimes \boldsymbol{\Theta}) + \mathrm{diag}\left(\mu_{q\left(1/\sigma^2_b\right)}\right) \otimes ((1-\pi)\boldsymbol{Q} + \pi\boldsymbol{I})\right]^{-1}$$

and mean

$$\mu_{q(b_i)} = \boldsymbol{\Sigma}_{q(b_i)}(\mathbf{1}_{J_i} \otimes \boldsymbol{\Theta})^T \mu_{q\left(\frac{1}{\sigma^2}\right)} \left[\mathrm{vec}\left(\boldsymbol{P}_i - \boldsymbol{\Theta}\mu_{q(\beta)}\boldsymbol{X}_i^T - \boldsymbol{\Theta}\mu_{q(\Phi)}\mu_{q\left(\Xi_i^T\right)}\right)\right].$$

For vec ($\boldsymbol{\Phi}$), the optimal density $q^*(\mathrm{vec}\,(\boldsymbol{\Phi}))$ is multivariate normal with covariance matrix

$$\boldsymbol{\Sigma}_{q(\mathrm{vec}(\Phi))} = \left[\mu_{q(\boldsymbol{\Xi}^T\boldsymbol{\Xi})} \otimes (\boldsymbol{\Theta}^T\boldsymbol{\Theta}) + \mathrm{diag}\left(\mu_{q\left(1/\sigma^2_{\Phi}\right)}\right) \otimes \boldsymbol{Q}\right]^{-1}$$

and mean

$$\mu_{q(\mathrm{vec}(\Phi))} = \boldsymbol{\Sigma}_{q(\mathrm{vec}(\Phi))}(\mu_{q(\Xi)} \otimes \boldsymbol{\Theta})^T \mu_{q\left(\frac{1}{\sigma^2}\right)} \left[\mathrm{vec}\left(\boldsymbol{P} - \boldsymbol{\Theta}\mu_{q(\beta)}\boldsymbol{X}^T - \boldsymbol{\Theta}\mu_{q(B)}\boldsymbol{V}^T\right)\right].$$

For $\boldsymbol{\xi}_{ij}$, letting $\mu_{q\left(\Sigma^{-1}_{\xi_{ij}}\right)}$ represent the expectation under the current distributions of the parameters γ_{lk} and g_{imk} of the precision matrix of the $\boldsymbol{\xi}_{ij}$, the optimal density $q^*(\boldsymbol{\xi}_{ij})$ is multivariate normal with covariance matrix

$$\boldsymbol{\Sigma}_{q(\xi_{ij})} = \left\{\mu_{q\left(\frac{1}{\sigma^2}\right)}\mu_{q(\boldsymbol{\Phi}^T\boldsymbol{\Theta}^T\boldsymbol{\Theta}\boldsymbol{\Phi})} + \mu_{q\left(\Sigma^{-1}_{\xi_{ij}}\right)}\right\}^{-1}$$

and mean

$$\mu_{q(\xi_{ij})} = \boldsymbol{\Sigma}_{q(\xi_{ij})} \mu_{q(\Phi)}^T \boldsymbol{\Theta}^T \mu_{q\left(\frac{1}{\sigma^2}\right)} \left(\boldsymbol{p}_{ij} - \boldsymbol{\Theta}\mu_{q(\beta)}\boldsymbol{x}_{ij} - \boldsymbol{\Theta}\mu_{q(b_i)}\right).$$

The expectation $\mu_{q(\Phi^T\Theta^T\Theta\Phi)}$ appearing in the above expression for $\boldsymbol{\Sigma}_{q(\xi_{ij})}$ is the $K \times K$ matrix given by $\mu_{q(\Phi)}^T \boldsymbol{\Theta}^T \boldsymbol{\Theta} \mu_{q(\Phi)} + \{M_{ij}\}$ where $M_{ij} = \mathrm{tr}\left[\boldsymbol{\Theta}^T\boldsymbol{\Theta}\mathrm{cov}(\boldsymbol{\phi}_i, \boldsymbol{\phi}_j)\right]$ and $\mathrm{cov}(\boldsymbol{\phi}_i, \boldsymbol{\phi}_j)$ is a submatrix of $\boldsymbol{\Sigma}_{q(\mathrm{vec}(\Phi))}$. The expectation $\mu_{q(\Xi^T\Xi)}$ appearing in the above expression for $\boldsymbol{\Sigma}_{q(\mathrm{vec}(\Phi))}$ is the $K \times K$ matrix given by $\mu_{q(\Xi)}^T \mu_{q(\Xi)} + M$, where $M = \sum_{i,j} \boldsymbol{\Sigma}_{q(\xi_{ij})}$.

Let $(\boldsymbol{\gamma}, \boldsymbol{g})_k$ represent the vector $(\boldsymbol{\gamma}_k, \boldsymbol{g}_{1k}, \boldsymbol{g}_{2k}, \ldots, \boldsymbol{g}_{Ik})$. As in Nott et al. (2012), we use a multivariate normal approximation to the density $q((\boldsymbol{\gamma}, \boldsymbol{g})_k)$. Using a routine from Nott et al. (2012), we approximate the mean $\mu_{q((\boldsymbol{\gamma},\boldsymbol{g})_k)}$ of the density $q((\boldsymbol{\gamma}, \boldsymbol{g})_k)$ with the posterior mode of the Bayesian gamma generalized linear model corresponding to the conditional posterior distribution of $(\boldsymbol{\gamma}, \boldsymbol{g})_k$, using as responses the expectations $\mu_{q(\xi_{ijk}^2)}$ in place of ξ_{ijk}^2, and we approximate the variance $\Sigma_{q((\boldsymbol{\gamma},\boldsymbol{g})_k)}$ with the negative inverse Hessian of the log posterior at the mode. Let these approximations be $\boldsymbol{\mu}_{mode}$ and $\boldsymbol{\Sigma}_{mode}$. Then, if ξ_{ijk} has the distribution $N[0, \exp\left(\boldsymbol{x}^T(\boldsymbol{\gamma}, \boldsymbol{g})_k\right)]$ for some coefficient vector $\boldsymbol{x}$, then by completing the square, we find that the expectation $\mu_{q\left(\Sigma_{\xi_{ij}}^{-1}\right)}$ in the expression for $\boldsymbol{\Sigma}_{q(\xi_{ij})}$ above is $\exp\left(-\boldsymbol{\mu}_{mode}^T\boldsymbol{x} - \frac{1}{2}\boldsymbol{x}^T\boldsymbol{\Sigma}_{mode}\boldsymbol{x}\right)$.

To find the optimal $q^*(\cdot)$ distributions for $\sigma_{\boldsymbol{\beta}_l}^2$, $\sigma_{\boldsymbol{b}}^2$, $\sigma_{\boldsymbol{\phi}_k}^2$ and σ^2, we use the following result: if the conditional distribution of a parameter ζ is inverse gamma with parameters α and β, then the distribution $q^*(\zeta)$ is inverse gamma with parameters $E_{-\zeta}[\alpha]$ and $E_{-\zeta}[\beta]$, and the expectation $\mu_{q(1/\zeta)}$ is $E_{-\zeta}[\alpha]/E_{-\zeta}[\beta]$.

For $\sigma_{\boldsymbol{\beta}_l}^2$, the optimal density $q^*(\sigma_{\boldsymbol{\beta}_l}^2)$ is inverse gamma with parameters $\alpha + \frac{K_\theta}{2}$ and $\beta + \frac{1}{2}\mu_{q\left(\beta_l^T Q \beta_l\right)}$. For $\sigma_{\boldsymbol{b}}^2$, the optimal density $q^*(\sigma_{\boldsymbol{b}}^2)$ is inverse gamma with parameters $\alpha + \frac{IK_\theta}{2}$ and $\beta + \frac{1}{2}\mu_{q\left(\sum_{i=1}^I b_i^T((1-\pi)Q+\pi I)b_i\right)}$. For $\sigma_{\boldsymbol{\phi}_k}^2$, the optimal density $q^*(\sigma_{\boldsymbol{\phi}_k}^2)$ is inverse gamma with parameters $\alpha + \frac{K_\theta}{2}$ and $\beta + \frac{1}{2}\mu_{q\left(\phi_k^T Q \phi_k\right)}$. All of these expectations can be found using the optimal $q^*()$ distributions for $\boldsymbol{\beta}_l$, $\boldsymbol{b}_i$ and $\boldsymbol{\phi}_k$ and the formula for the expectation of a quadratic form.

For σ^2, let $\boldsymbol{x}_{ij}$ be the row of the matrix $\boldsymbol{X}$ corresponding to the jth motion of the ith subject.

Then the optimal density $q^*(\sigma^2)$ is inverse gamma with parameters $\alpha + \frac{nD}{2}$ and

$$\beta + \frac{1}{2}\sum_{i=1}^{I}\sum_{j=1}^{J_i}\big[\|\boldsymbol{p}_{ij} - \boldsymbol{\Theta}\mu_{q(\beta)}\boldsymbol{x}_{ij} - \boldsymbol{\Theta}\mu_{q(b_i)} - \boldsymbol{\Theta}\mu_{q(\Phi)}\mu_{q(\xi_{ij})}\|^2 \\ + \boldsymbol{x}_{ij}\boldsymbol{L}\boldsymbol{x}_{ij}^T + m_i + n_{ij}\big]$$

where the matrix $\boldsymbol{L}$ is the $(l+1) \times (l+1)$ matrix whose i, j entry is the trace of $\boldsymbol{\Theta}^T\boldsymbol{\Theta}$ times the covariance between the ith and jth column of $\boldsymbol{\beta}$ under the current distribution of $\boldsymbol{\beta}$, $m_i = \text{tr}\left[\boldsymbol{\Theta}^T\boldsymbol{\Theta}\boldsymbol{\Sigma}_{q(b_i)}\right]$, and

$$n_{ij} = \mu_{q(\xi_{ij})}^T\mu_{q(\Phi^T\Theta^T\Theta\Phi)}\mu_{q(\xi_{ij})} + \text{tr}\left[\mu_{q(\Phi^T\Theta^T\Theta\Phi)}\boldsymbol{\Sigma}_{q(\xi_{ij})}\right] - \mu_{q(\xi_{ij})}^T\mu_{q(\Phi)}^T\boldsymbol{\Theta}^T\boldsymbol{\Theta}\mu_{q(\Phi)}\mu_{q(\xi_{ij})}.$$

The optimal $q^*(\Sigma_{g_k})$ density is given by

$$\begin{aligned} q^*(\Sigma_{g_k}) &\sim \exp[E_{-\Sigma_{g_k}}\log p(\Sigma_{g_k}|\text{rest})] \\ &\sim \exp\left[E_{-\Sigma_{g_k}}\left\{-\frac{\nu + I + p + 1}{2}\log|\Sigma| - \frac{1}{2}\left(\text{tr}\left[(\Psi + \sum_{i=1}^{I}\boldsymbol{g}_{ik}\boldsymbol{g}_{ik}^T)\Sigma^{-1}\right]\right)\right\}\right] \end{aligned}$$

Therefore the optimal density is inverse-Wishart with parameters $\nu + I$ and $\boldsymbol{\Psi} + \sum_{i=1}^{I}\mu_{q(\boldsymbol{g}_{ik}\boldsymbol{g}_{ik}^T)}$. The expectation $\mu_{q(\boldsymbol{g}_{ik}\boldsymbol{g}_{ik}^T)}$ in this expression is $\mu_{q(\boldsymbol{g}_{ik})}\mu_{q(\boldsymbol{g}_{ik})}^T + M$, where M is the covariance of $\boldsymbol{g}_{ik}$ under the posterior distribution of $(\boldsymbol{\gamma}, \boldsymbol{g})_k$. The mean of this density is

$$\mu_{q(\Sigma_{gk})} = \frac{\boldsymbol{\Psi} + \sum_{i=1}^{I}\mu_{q(\boldsymbol{g}_{ik}\boldsymbol{g}_{ik}^T)}}{\nu + I - p - 1}.$$

Straightforward extensions of these derivations apply in the case of nested random effects, as in model extension (6).

E.4 Details of implementation of HMC sampler

Our HMC samplers in Sections 5 and 6 fit the same models as fit by our VB model, while conditioning on VB estimates of the parameters $\boldsymbol{\beta}_l$, $\boldsymbol{b}_{im}$ and $\boldsymbol{\phi}_k$ in model (5), and therefore implicitly also conditioning on the associated variance parameters and on the VB estimate of π. The HMC samplers estimate all other parameters in these models: the scores ξ_{ijk}, the fixed effect variance parameters γ_{lk}, the random effect variance parameters $\boldsymbol{g}_{ik}$ (and $\boldsymbol{g}_{ilk}$, in model extension (6)), the random effect variance parameter covariance matrices, and the error variance σ^2. The samplers were implemented in the `STAN` Bayesian programming language (Stan Development Team, 2013). `STAN` implements Hamiltonian Monte Carlo, an MCMC algorithm that uses the gradient of the log-posterior to avoid random walk behavior and therefore more quickly generate samples from the posterior (Neal, 2011).

We ran all HMC samplers here using 4 chains and checked for convergence using the convergence criterion of Gelman and Rubin (1992). We ran the HMC sampler used in Section 5 for 800 iterations per chain, and discarded the first 400 iterations from each chain, which took about 90 minutes per chain. We ran the HMC sampler used in Appendix B for 2000 iterations per chain, and discarded the first 1000 iterations from each chain.

Code implementing the `STAN` model used in Section 5 is included in the Supplementary Materials.

F Additional simulation results

Here we present cross-sectional simulations to illustrate the effect of varying the number of curves, the number of estimated FPCs, the number of spline basis functions and the measurement error on the quality of estimation using the VB method. In this cross-sectional design, curves are generated from the model

$$P_i(t) = 0 + \sum_{k=1}^{4} \xi_{ik}\phi_k(t) + \epsilon_i(t).$$

FPCs and group and FPC-specific score variances are as in the simulations in Section 5.

All results are for 200 replicates per simulation scenario. We present one simulation where we fix the number of estimated FPCs at 4, the number of spline basis functions at 10, and the measurement error standard deviation at 0.25, and vary the number of curves in the set $\{20, 40, 80, 160, 320\}$. In the other simulations we fix the sample size at 80 and vary one of the other parameters.

For each simulated dataset, we use the methods described in Section 4 to fit the model

$$\boldsymbol{p}_i = \boldsymbol{\Theta}\boldsymbol{\beta}_0 + \sum_{k=1}^{K} \xi_{ik}\boldsymbol{\Theta}\boldsymbol{\phi}_k + \boldsymbol{\epsilon}_i \tag{A.4}$$

$$\xi_{ik} \sim \mathrm{N}\left[0, \exp\left(\sum_{m=1}^{2} \gamma_{lk} x^*_{il}\right)\right]. \tag{A.5}$$

The covariates x^*_{il} are defined like the analogous covariates in Section 5.

Figure A.6 shows that accuracy in estimation of FPCs and bias in estimation of variance model parameters decreases with more curves. Figure A.7 shows that when 2 or 3 FPCs are estimated instead of the 4 that actually exist, estimates of the quantities that are estimated are not negatively affected. Figure A.8 shows the result of changing the number of spline basis functions used for estimation. 5 spline basis functions are not sufficient to adequately capture the relatively fast variation in FPCs 3 and 4; otherwise, because we induce smoothness in the estimated FPCs using the penalty matrix $\boldsymbol{Q}$, using richer spline bases does not negatively affect estimation accuracy. Figure A.9 shows the result of adding more noise to the simulated curves, keeping the sample size fixed. As expected, more noise results in larger errors in estimation, of both the FPCs and the score variance parameters.

Figure A.10 shows examples of estimates of FPC 2 with varying levels of integrated squared error. These estimates are from the longitudinal simulation scenario with $J_i = 4$.

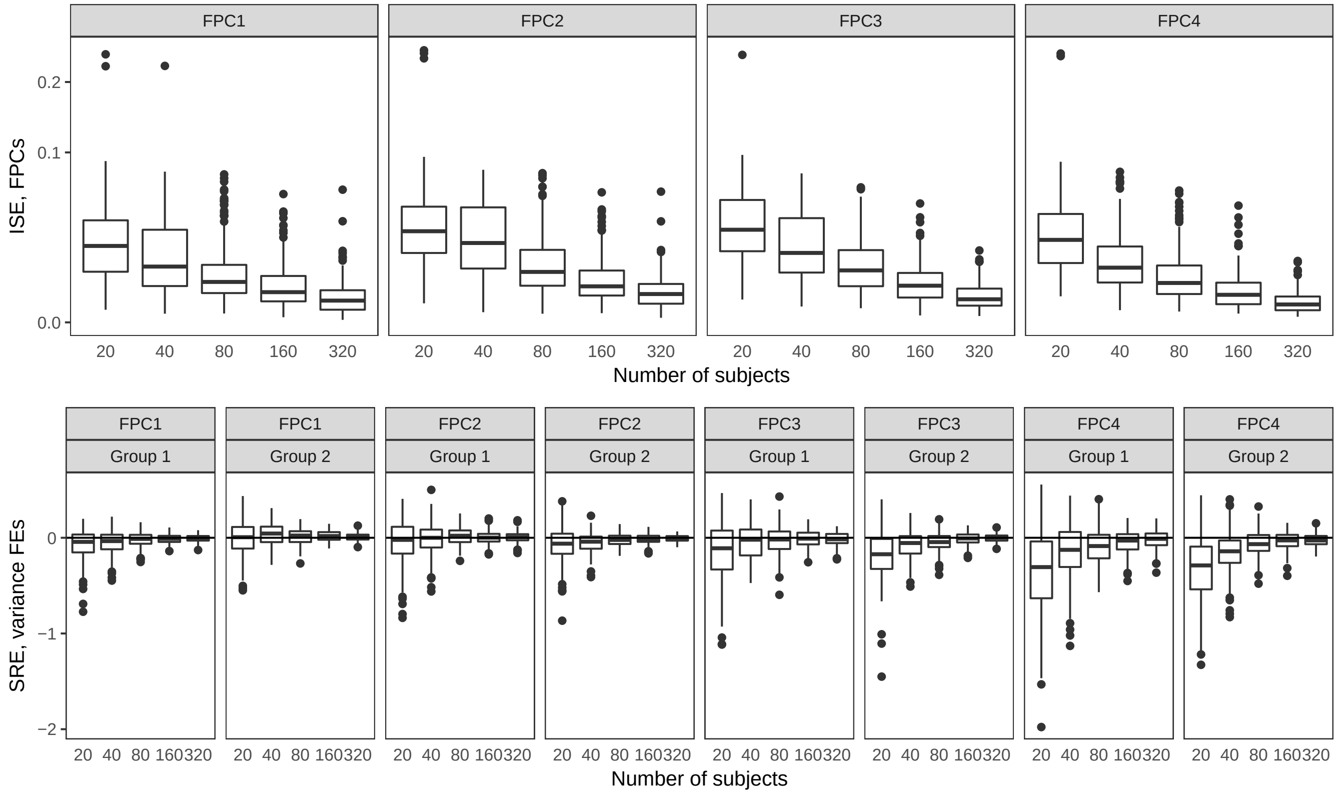

Figure A.6: Varying the number of curves. Integrated squared errors in estimation of FPCs (first row) and signed relative error in estimation of variance parameters (second row) decreases with more curves.

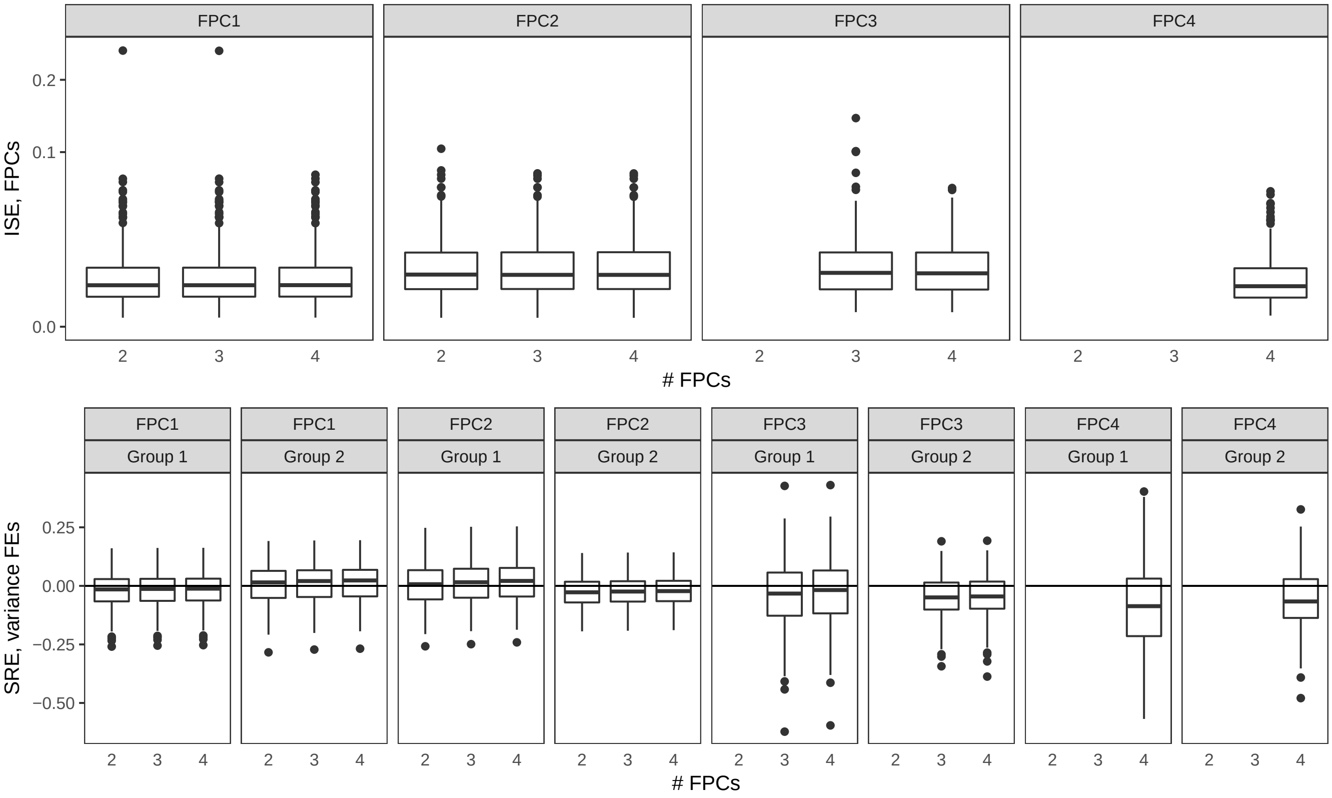

Figure A.7: Varying the number of estimated FPCs. Integrated squared errors in estimation of FPCs (first row) and signed relative error in estimation of variance parameters (second row) for FPCs 1 and 2 is mostly invariant to whether additional FPCs and associated score variance parameters are also estimated.

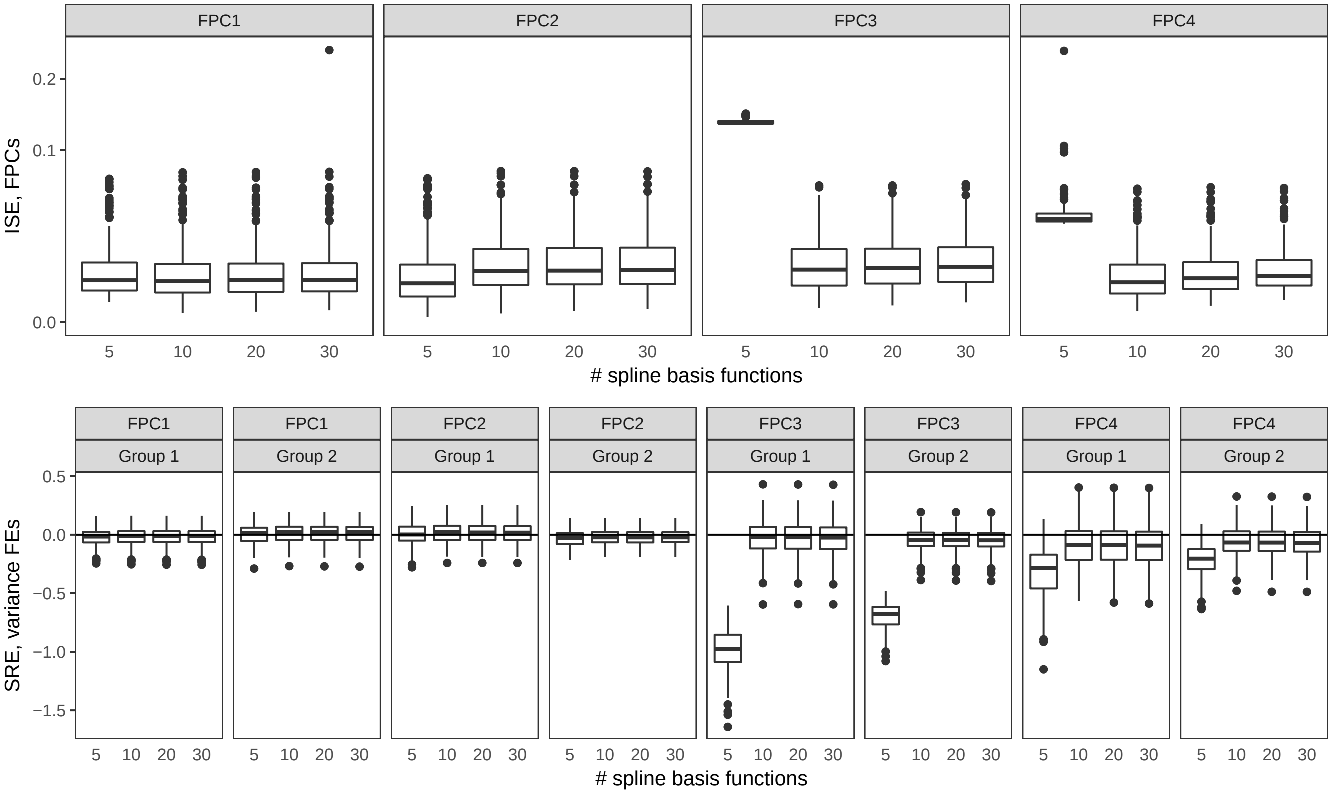

Figure A.8: Varying the number of spline basis functions. 5 spline basis functions are not sufficient to adequately capture the relatively fast variation in FPCs 3 and 4. Otherwise integrated squared errors in estimation of FPCs (first row) and signed relative error in estimation of variance parameters (second row) are mostly invariant to the number of spline basis functions used in simulation.

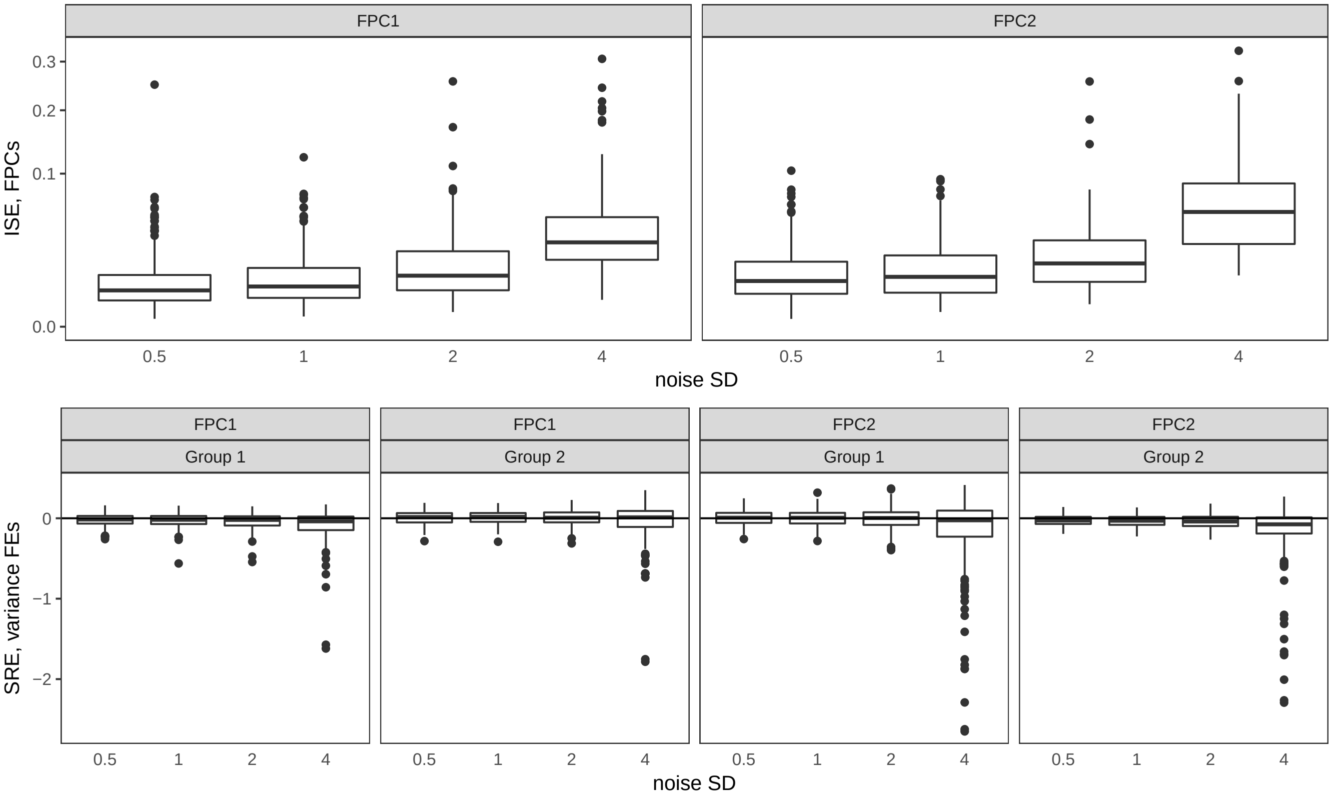

Figure A.9: Varying the measurement error. We varied the measurement error standard deviation to 0.5, 1, 2 and 4. FPC integrated squared errors (first row) and signed relative errors in estimation of the variance parameters (second row) illustrate that results are robust to a significant amount of noise, but estimation of parameters becomes poorer as the amount of noise increases. Four FPCs were simulated but only 2 were estimated.

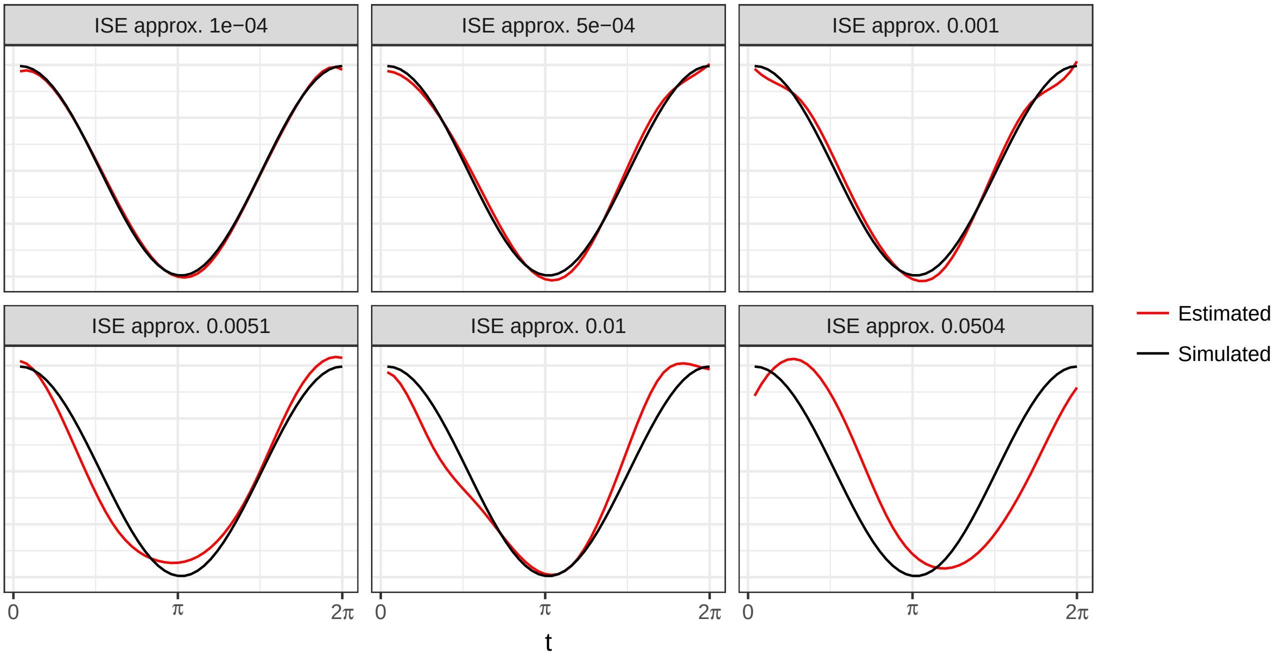

Figure A.10: Examples of estimates of FPC 2 with varying levels of integrated squared error. These estimates come from the longitudinal simulation scenario with $J_i = 4$.