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

December 2, 2016

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

Recent work in motor learning has suggested that change in motion variability is an important component of improvement in motor skill. For example, 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 of 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.

2 Dataset and Model

2.1 Dataset

Our motivating data were gathered as part of a study of motor learning among healthy subjects. Kinematic data were acquired in a standard planar reaching task used to measure upper extremity reaching control. In this task, subjects first rest their forearm on an air-sled system to reduce friction and gravity effects. They then perform repeated reaching motions with their arm to 8 equally-spaced targets arranged in a circle around the starting point, with targets appearing in a semi-random order. Subjects 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 to the starting point. Subjects are not rewarded for motions outside pre-specified velocity thresholds.

Our dataset consists of 9,481 motions by 26 right-handed subjects. After being familiarized with the experimental apparatus, each subject made 24 reaching motions to each of the 8 targets with both the left and right hand. In order to exclude motions mistakenly made to the wrong target or not attempted by the subject due to distraction, 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. In addition, to exclude motions made at speeds outside the range of interest, motions 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 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 have been removed from the dataset due to instrumentation and recording errors. In addition, a few subjects made 25 motions to a target instead of 24. 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

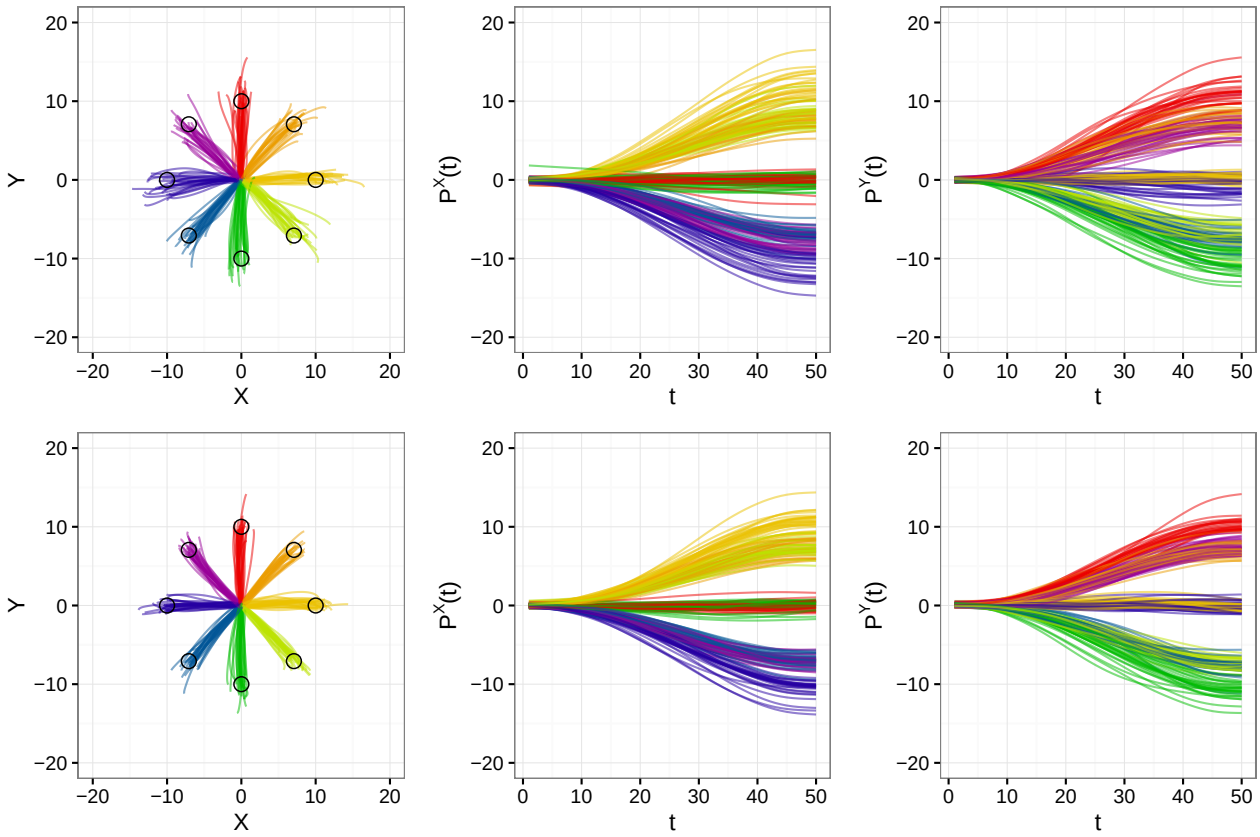

Figure 1: Observed kinematic data. The top row shows the first motion to each target for each of the 26 subjects using the right hand; the bottom row shows the last motion to each target for each of the 26 subjects using the right hand. In the left panel of each row, observed reaching data are shown in the X and Y plane, and targets are indicated using circles. In the middle and right panels of each row, the $P_{ij}^X(t)$ and $P_{ij}^Y(t)$ curves are shown separately as a function of time.

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 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 motion made to each target by each subject using the right hand. Within each row, the left panel shows reaching data in the X and Y plane, and the middle and right panels show the $P_{ij}^X(t)$ and $P_{ij}^Y(t)$ curves separately as functions of time. 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: 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. Comments on this univariate and an alternative bivariate approach are included in Section 6, and analyses implementing the bivariate approach appear in the Supplementary Materials.

2.2 Model for Curve Variance

We adopt a functional data approach to model position curves $P_{ij}(t)$, omitting the X and Y superscripts for notational simplicity. Our starting point is to model deviations from curve-specific expected values using functional principal components analysis (FPCA):

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

where $\mu_{ij}(t)$ is the expected value of $P_{ij}(t)$ and the deviation $\delta_{ij}(t)$ is modeled as a linear combination of K functional principal components (FPC) basis functions $\phi_k(t)$ with coefficients ξ_{ijk}. FPC basis functions $\phi_k(t)$ are constant across all individuals and curves, while the scores ξ_{ijk} are individual and curve-specific. Model (1) is based on a truncation of the covariance decomposition

$$\text{Cov}[\delta_{ij}(s), \delta_{ij}(t)] = \sum_{k=1}^{\infty} \lambda_k \phi_k(s)\phi_k(t) \tag{2}$$

using eigenfunctions $\phi_k(t)$ and corresponding non-increasing eigenvalues λ_k. It is traditionally assumed that scores ξ_{ijk} are uncorrelated random variables with $E(\xi_{ijk}) = 0$ and $\text{Var}(\xi_{ijk}) = \lambda_k$.

The distribution of the scores ξ_{ijk} in (1) determines the variability of a subject's curves around the subject-specific mean. The usual assumption of constant score variance implies, for our data, that the variability of the position curves $P_{ij}(t)$ is not covariate- or subject-dependent. Such a

model is unfounded in this context, where 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.

We therefore model the variance of scores ξ_{ijk}, which continue to have zero mean, by introducing covariate and subject-dependent heteroskedasticity into (1). We pose the score variance model

$$\text{Var}(\xi_{ijk}|\boldsymbol{w}_{ijk}, \boldsymbol{z}_{ijk}, \boldsymbol{g}_{ik}) = \lambda_{k|\boldsymbol{w}_{ijk},\boldsymbol{z}_{ijk},\boldsymbol{g}_{ik}} = \exp\left(\gamma_{k0} + \sum_{m=1}^{q} \gamma_{km} w_{ijkm} + \sum_{h=1}^{r} g_{ikh} z_{ijkh}\right) \tag{3}$$

where, as before, ξ_{ijk} is the kth score for the jth curve of the ith subject. In (3), γ_{k0} is an intercept for the variance of the scores, γ_{km} are fixed effects coefficients for covariates w_{ijkm}, $m = 1, \ldots, q$, and g_{ikh} are random effects coefficients for covariates z_{ijkh}, $h = 1, \ldots, r$. The vector $\boldsymbol{g}_{ik}$ consists of the concatenation of the coefficients g_{ikh}, likewise for the vectors $\boldsymbol{w}_{ijk}$ 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 w_{ijkm} and z_{ijkh} 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. As in standard FPCA approaches, we assume that scores have mean zero and are not correlated across k.

Together, models (1) and (3) induce a subject- and covariate-dependent covariance structure for $\delta_{ij}(t)$:

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

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:

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

We assume that the basis functions $\phi_k(t)$ do not depend on covariate or subject effects; they 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 $E\left[\lambda_{k|\boldsymbol{w}_{ijk}, \boldsymbol{z}_{ijk}, \boldsymbol{g}_{ik}}\right]$ may not correspond to the order of conditional variances $\lambda_{k|\boldsymbol{w}_{ijk}, \boldsymbol{z}_{ijk}, \boldsymbol{g}_{ik}}$ 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 (1) 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)

$$\mu_{ij}(t) = \beta_0(t) + \sum_{l=1}^{p} x_{ijl}\beta_l(t) + b_i(t). \tag{4}$$

Here $\beta_0(t)$ is the functional intercept, x_{ijl} for $l \in 1 \ldots p$ are covariates associated with curve $P_{ij}(t)$, $\beta_l(t)$ is the coefficient function associated with the lth covariate, and $b_i(t)$ is a random functional intercept for the ith subject.

We also assume that the basis functions $\phi_k(t)$ do not themselves depend on covariate or subject effects. 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 rotating observed data so that all motions extend to the target at 0°. In the left

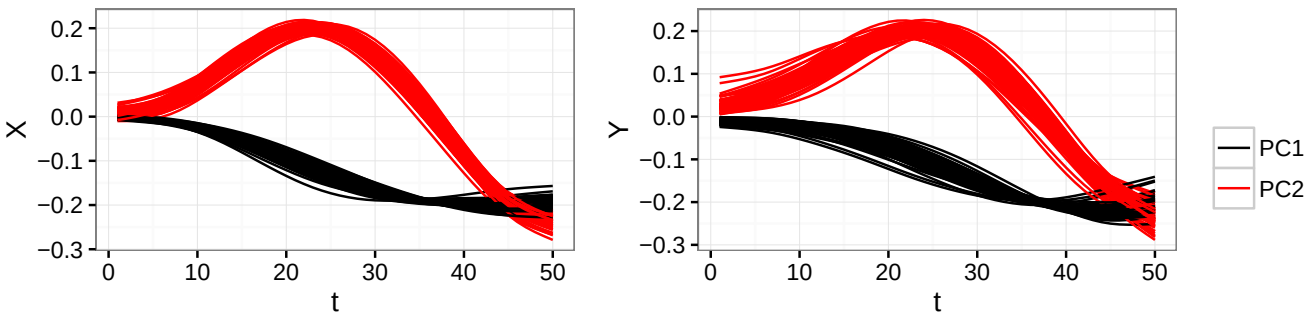

Figure 2: FPC basis functions estimated for various data subsets after rotating curves onto the positive X axis. The left panel shows the first and second FPC basis functions estimated for the X coordinate of motions to each target, for the left and right hand separately, and separately for motion numbers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

panel of Figure 2, we show the first two basis functions for the X coordinate of motions made to different targets by the left and right hands separately, and at different stages of training; the right panel shows the same for the Y coordinate of motion. In all cases, the basis functions are similar.

Our estimation strategy for models (1) and (3) is discussed in Section 4.

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; in a similar model, van der Linde (2008) took a variational Bayes approach to estimation. 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 dependence of the mean of curves on covariates. As a byproduct, 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 multilevel 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 principal components 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.

4 Methods

A main contribution of this manuscript is the introduction of subject and covariate effects on score variances in model (3). Several estimation strategies could be used within this framework. We briefly describe and contrast some possible approaches before developing one in detail; later, these approaches will be compared in simulations.

Perhaps the most common approach to traditional FPCA is that in Yao et al. (2005). In short, one uses observed data to estimate the mean and covariance, the latter of which is then decomposed as in equation (2) to obtain principal components and score variances. Given these quantities, scores are estimated using best linear unbiased predictors in a mixed model representation of model (1). One could, given these score estimates, then fit a model for the score variances. This approach is expected to work reasonably well in many cases, but it arises as a sequence of models that treat estimated quantities as fixed: first one estimates the mean, principal components, and score variances; then one treats these as fixed to estimate the scores; then 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. Iterative or resampling-based procedures can help to address these concerns (Goldsmith et al., 2013), but require some care.

Jointly estimating all parameters is an appealing alternative. After specifying an appropriate model, both maximum likelihood and Bayesian approaches are possible. Sampling-based approaches to Bayesian inference are expected to be computationally intensive; additionally, they may encounter difficulty due to the high correlation between the mean $\mu_{ij}(t)$ and the deviation $\delta_{ij}(t)$. Estimating some model components from the data and treating them as fixed, as in the multi-stage approach, may address the latter concern at the expense of posing a fully Bayesian model.

We adopt a variational approach as an alternative to sampling-based Bayesian estimation and inference. This technique is computationally efficient and typically yields accurate point estimates for model parameters, although it is an approximation to the complete posterior distribution and inference may suffer as a result. Comparisons to multi-stage and sampling-based approaches appear in Section 5.

4.1 Bayesian model specification

We now present the full model specification, in which we reformulate models (1) and (4) in matrix form to reflect the discrete nature of the observed data. In the following $\boldsymbol{\Theta}$ is a known $D \times K_\theta$ matrix of K_θ spline basis functions evaluated on the grid of length D to which the curves are registered. Assuming a normal distribution of the scores ξ_{ijk}, conditional on random effects and covariates, models (1), (3) and (4) may be combined as:

$$
\begin{aligned}
\boldsymbol{p}_{ij} &= \sum_{l=0}^{p} x_{ijl}\boldsymbol{\Theta}\boldsymbol{\beta}_l + \boldsymbol{\Theta}\boldsymbol{b}_i + \sum_{k=1}^{K} \xi_{ijk}\boldsymbol{\Theta}\boldsymbol{\phi}_k + \boldsymbol{\epsilon}_{ij} \qquad (5)\\
\boldsymbol{\beta}_l &\sim \text{MVN}\left[0, \sigma^2_{\boldsymbol{\beta}_l}\boldsymbol{Q}^{-1}_{K_\theta}\right]; \sigma^2_{\boldsymbol{\beta}_l} \sim \text{IG}\left[\alpha, \beta\right]\\
\boldsymbol{b}_i &\sim \text{MVN}\left[0, \sigma^2_{\boldsymbol{b}}\boldsymbol{Q}^{-1}_{K_\theta}\right]; \sigma^2_{\boldsymbol{b}} \sim \text{IG}\left[\alpha, \beta\right]\\
\boldsymbol{\phi}_k &\sim \text{MVN}\left[0, \sigma^2_{\boldsymbol{\phi}_k}\boldsymbol{Q}^{-1}_{K_\theta}\right]; \sigma^2_{\boldsymbol{\phi}_k} \sim \text{IG}\left[\alpha, \beta\right]\\
\xi_{ijk} &\sim \text{N}\left[0, \exp\left(\sum_{m=0}^{q}\gamma_{km}w_{ijkm} + \sum_{h=1}^{r} g_{ikh}z_{ijkh}\right)\right]\\
\gamma_{km} &\sim \text{N}\left[0, \sigma^2_{\gamma_{km}}\right]\\
g_{ikh} &\sim \text{N}\left[0, \sigma^2_{g_{kh}}\right]; \sigma^2_{g_{kh}} \sim \text{IG}\left[\alpha, \beta\right]\\
\boldsymbol{\epsilon}_{ij} &\sim \text{MVN}\left[0, \sigma^2\boldsymbol{I}_D\right]; \sigma^2 \sim \text{IG}\left[\alpha, \beta\right]
\end{aligned}
$$

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. The vectors $\boldsymbol{p}_{ij}$ and $\boldsymbol{\epsilon}_{ij}$ are the $D \times 1$ observed functional outcome and independent error term, respectively, for the jth curve of the ith subject. The vectors $\boldsymbol{\beta}_l$, for $l = 0, \ldots, p$, are functional effect spline coefficient vectors, $\boldsymbol{b}_i$, for $i = 1, \ldots, I$, 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. $\boldsymbol{Q}_{K_\theta}$ is a matrix penalizing the second derivative of the estimated functions, commonly used to encourage smoothness, and $\boldsymbol{I}_D$ is the $D \times D$ identity matrix. MVN refers to the multivariate normal distribution, N to the normal distribution and IG to the

inverse-gamma distribution.

In keeping with standard practice, we set the prior variances $\sigma^2_{\gamma_{km}}$ for the fixed-effect coefficients in the score variance model to a large constant, so that the prior is close to uniform. The variances $\sigma^2_{g_{kh}}$ correspond to random effects in the score variance model, and are assigned inverse gamma priors. Variance components $\{\sigma^2_{\boldsymbol{\beta}_l}\}_{l=0}^{p}$, $\{\sigma^2_{\boldsymbol{\phi}_k}\}_{k=1}^{K}$ and $\sigma^2_{\boldsymbol{b}}$ have a somewhat different interpretation. The first two act as tuning parameters controlling the smoothness of coefficient functions $\beta_l(t)$ and FPC functions $\phi(t)$, and our prior specification for them is related to standard techniques in semi-parametric regression. The latter, meanwhile, is a tuning parameter that controls the smoothness of random effects $b_i(t)$, and is shared across subjects so that random effects share a common distribution. The penalty matrix $\boldsymbol{Q}_{K_\theta}$ is common across fixed and random effects and FPCs to induce smoothness of the same type but with a degree controlled by the preceding tuning parameters. Lastly, we set the values of α and β, the parameters of the inverse-gamma distributions for the variance components, to 1. In practice, sensitivity to prior specifications should be explored.

In our real data analysis in Section 6, we extend this model by assuming that two random effects in the score variance model for each FPC (one a random intercept and one a random slope) have a common bivariate normal distribution.

4.2 Estimation via variational Bayes

Variational Bayes methods are a computationally efficient approach to obtain approximate solutions in Bayesian models (Ormerod and Wand, 2012; Jordan, 2004; Jordan et al., 1999; Titterington, 2004). These tools have been used in functional data analysis (van der Linde, 2008; Goldsmith et al., 2011; McLean et al., 2013; Goldsmith and Kitago, 2016); in particular, Goldsmith and Kitago used variational Bayes in the estimation of model (4). Here we present a brief overview of variational Bayes. A detailed derivation of our variational algorithm is provided in Appendix Section E of the Supplementary Material.

Let $\boldsymbol{y}$ and $\boldsymbol{\lambda}$ represent the data and parameters, respectively. Using variational Bayes, we

approximate the posterior $p(\boldsymbol{\lambda}|\boldsymbol{y})$ using $q(\boldsymbol{\lambda})$, where q is a member of a restricted class of functions Q more easily estimated than the posterior $p(\boldsymbol{\lambda}|\boldsymbol{y})$. To find the best q in this restricted class, we choose the element $q^* \in Q$ that minimizes the Kullback-Leibler distance from $p(\boldsymbol{\lambda}|\boldsymbol{y})$. The class Q is often the class of posterior distributions satisfying some factorization property, so that $q(\boldsymbol{\lambda}) = \prod_{l=1}^{L} q_l(\lambda_l)$, with each $q_l(\lambda_l)$ a parametric density function. It can then be shown that the optimal q_l^* densities are given by

$$q_l^*(\lambda_l) \propto \exp\left[E_{-\lambda_l} \log p(\lambda_l|\text{rest})\right] \tag{6}$$

where $E_{-\lambda_l}$ represents the expectation with respect to the currently estimated values of all parameters except λ_l, and "rest" represents the observed data plus all parameters other than λ_l. This suggests the use of an iterative algorithm, setting initial values for all parameters and then updating the optimal density for each parameter λ_l 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 $q(\boldsymbol{\lambda})$ is

$$q(\boldsymbol{\beta}_0, \ldots, \boldsymbol{\beta}_p) \left\{\prod_{i=1}^{I} q(\boldsymbol{b}_i)\right\} q(\phi_1, \ldots, \phi_K) \left\{\prod_{i=1}^{I}\prod_{j=1}^{J_i} q(\boldsymbol{\xi}_{ij})\right\} \left\{\prod_{k=1}^{K} q(\gamma_{k0}, \ldots, \gamma_{kq}, g_{1k1}, \ldots, g_{Ikr})\right\} \left\{\prod_{s\in S} q(\sigma_s^2)\right\} \tag{7}$$

The quality of this approximation depends on the extent to which the true posterior distribution factors as above. As stated elsewhere, 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.

4.3 Orthonormalization

One of the challenges of a Bayesian approach to FPCA is that the basis functions $\phi_k(t)$ are indeterminate up to a nonsingular transformation, at least up to sign and rotation. 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 interpretable. We therefore impose two constraints on the FPC basis functions $\phi_k(t)$, which we estimate with the vectors $\boldsymbol{\Theta}\boldsymbol{\phi}_k$.

First, we impose the familiar constraint from principal component analysis that the first principal component should explain the largest possible amount of variance in the data, and that each succeeding principal component should both be orthonormal to all preceding principal components and should explain the largest possible amount of the remaining variance in the data. 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. To enforce this ordering constraint, 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. Second, we impose the constraint that the first component of each principal component basis function coefficient vector should be positive. To enforce this constraint, we divide each vector $\boldsymbol{\Theta}\boldsymbol{\phi}_k$ by the sign of its first component. These constraints are enforced after the estimation of $\boldsymbol{\Phi}$ in each step of our iterative algorithm. 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.

5 Simulations

We demonstrate the performance of our method using simulated data. In all simulations, principal components $\boldsymbol{\phi}_1$ and $\boldsymbol{\phi}_2$ are the functions $\sin(x)$ and $\cos(x)$ and principal components $\boldsymbol{\phi}_3$ and $\boldsymbol{\phi}_4$ are the functions $\sin(2x)$ and $\cos(2x)$, observed over the domain $[0, 2\pi]$.

5.1 Cross-sectional simulations

We start with a cross-sectional simulation setting. In this design, curves are generated from the model

$$\boldsymbol{p}_i = 0 + \sum_{k=1}^{4} \xi_{ik}\boldsymbol{\phi}_k + \boldsymbol{\epsilon}_i$$

where $\boldsymbol{\epsilon}_i$ is a vector of independent normal errors with variance σ^2. We observe the curves at $D = 50$ equally spaced points on the domain. We divide the curves into two groups and generate scores for each curve and FPC k from a normal distribution with mean zero and variance $\lambda_{k|group}$. The variances for the scores for each FPC and group are shown in Table 1.

Two hundred replicate datasets were generated according to the above design for sample sizes $I \in \{20, 40, 80, 160, 320\}$. For now, we fix $\sigma^2 = 0.25$. For each simulated dataset, we use the methods described in Section 4 to fit the model

$$\boldsymbol{p}_i = \boldsymbol{\Theta}\boldsymbol{\beta}_0 + \sum_{k=1}^{K} \xi_{ik}\boldsymbol{\Theta}\boldsymbol{\phi}_k + \boldsymbol{\epsilon}_i \tag{8}$$

$$\xi_{ijk} \sim \mathrm{N}\left[0, \exp\left(\sum_{m=1}^{2} \mathbb{I}(w_i = m)\gamma_{km}\right)\right] \tag{9}$$

where w_i is the group (1 or 2) to which curve i is assigned, and $\mathbb{I}$ is the indicator function. Throughout this simulation section, we use the priors specified in model 5, setting the fixed effect hyperparameter $\sigma^2_{\gamma_{km}}$ to 100 and the inverse gamma hyperparameters α and β to 1. Lastly, here we fit the model using 10 basis functions and 4 FPCs.

	$\lambda_{1\|group}$	$\lambda_{2\|group}$	$\lambda_{3\|group}$	$\lambda_{4\|group}$
Group 1	36	12	6	4
Group 2	18	24	12	6

Table 1: Score variances $\lambda_{k|group}$ for each FPC k and group 1 or 2, used in simulation.

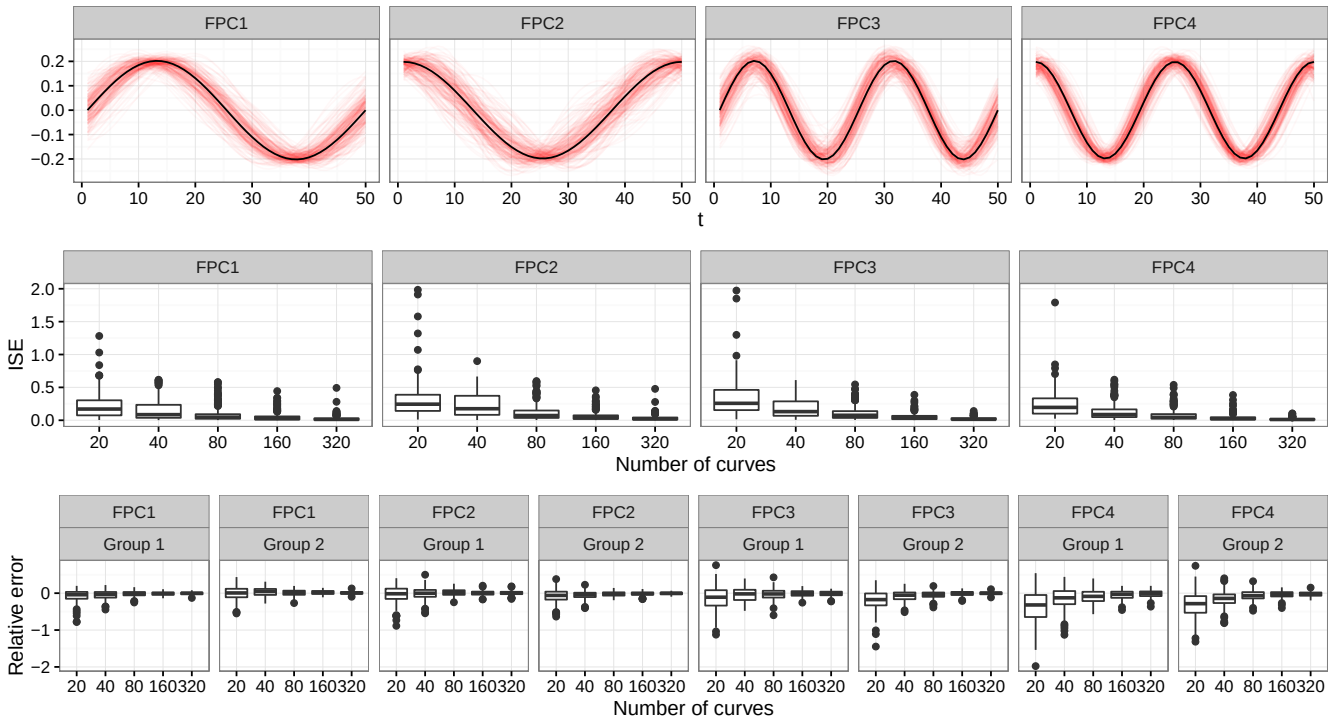

Figure 3: Cross-sectional simulation results. The top row shows estimates of the 4 FPCs using the variational Bayes procedure from each simulation replicate with sample size $I = 80$; true FPCs are plotted in black. The second row shows boxplots of the integrated squared error of estimates of the four FPCs for each sample size. The third row shows boxplots of the relative error of estimates of the two parameters for each FPC score variance, one for each of the two groups.

Figure 3 illustrates the results for this simulation scenario. The top row shows estimates of the four FPCs for a sample size of $I = 80$ per replicate simulation, with true FPCs plotted in black; estimates are generally accurate and have no obvious systematic biases. The second row shows the integrated squared error ISE $= \int_0^1 [\phi_k(t) - \widehat{\phi_k}(t)]^2 \, dt$ of estimates $\widehat{\phi_k}(t)$ of the four FPCs for each of the 5 possible sample sizes. As expected, error in estimation of the FPCs decreases with an increasing number of curves. The third row shows the relative error RE $= \frac{\widehat{\gamma_{km}} - \gamma_{km}}{\gamma_{km}}$ of estimates $\widehat{\gamma_{km}}$ of the two parameters for each FPC score variance, one for each of the two groups. Estimates are somewhat shrunk towards zero for smaller sample sizes, especially for FPCs 3 and 4; errors decrease for larger sample sizes.

Next, we compare our variational Bayes estimation procedure to the alternative procedures described in Section 4. In the smoothed covariance (SC) procedure, we first estimate FPCs using a smoothed covariance decomposition. We then estimate scores as best linear unbiased predictors in a mixed model formulation of model (8). Given these score estimates, we then fit to the square of the scores the gamma generalized linear model induced by (9) (see Appendix Section E in the Supplementary Materials) using standard generalized linear model software.

Our initial implementation of a Gibbs sampler to estimate all parameters in models 8 and 9 included an SVD step to enforce constraints on the eigenfunctions. However, we found that this did not provide correct inference for the FPCs, an issue that may have arisen due to the rotation step. In the sampling procedure we present here (Gibbs), we therefore condition on the mean and FPCs estimated using a smoothed covariance decomposition, and estimate the scores and all parameters in the variance model using a Gibbs sampler. We implement this sampler using `JAGS`, an open-source, general purpose programming language for Bayesian analysis (Plummer, 2003).

Figure 4 compares these three different estimates of the score variance parameters for the simulation scenario with 80 curves per replicate simulation. The estimates of the variance parameters produced by the three procedures are very similar. In terms of computation time, the VB procedure took about 20 seconds, the SC procedure about 1 second, and the Gibbs procedure about

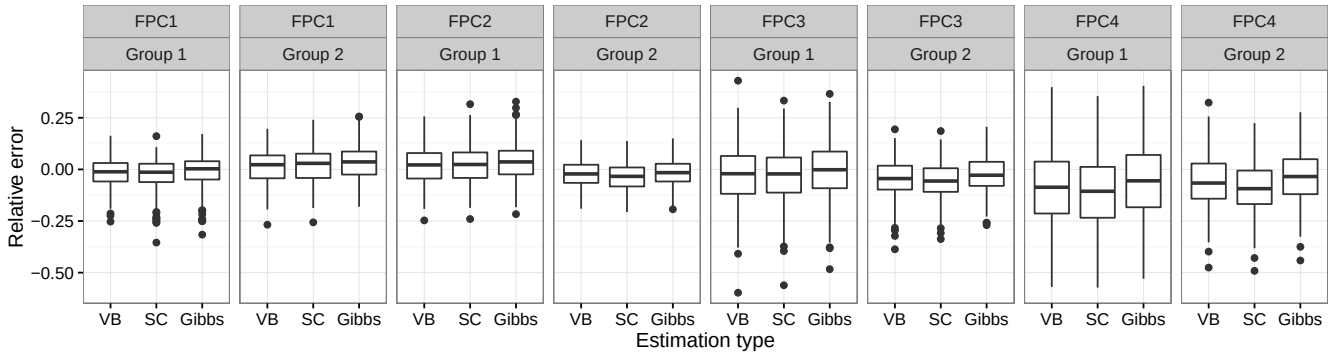

Figure 4: Boxplots show the relative error of estimates of the score variance parameters using the VB, SC and Gibbs methods for all simulation replicates with sample size $I = 80$.

10 minutes, running 3 chains for 2,000 iterations each. The time for running the Gibbs procedure could be shortened by parallelizing the computation; in addition, it is possible that fewer iterations could approximate the posterior to sufficient accuracy.

FPC	Group	VB	SC	Gibbs
1	1	0.890	0.880	0.905
1	2	0.945	0.880	0.885
2	1	0.950	0.920	0.910
2	2	0.935	0.895	0.965
3	1	0.920	0.925	0.930
3	2	0.860	0.855	0.930
4	1	0.875	0.860	0.935
4	2	0.905	0.870	0.950

Table 2: Coverage of 95% confidence intervals for the score variance parameters using the VB, SC and Gibbs procedures, based on all simulation replicates with sample size $I = 80$.

All three of the methods discussed allow for inference regarding the score variance parameters. Table 2 shows, based on 80 curves per simulation replicate and 200 simulation replicates, that coverage of confidence intervals for the VB, SC and Gibbs procedures is close to the nominal level; coverage improves for larger sample sizes. In theory, the VB method can also provide confidence intervals for the FPCs; in practice, however, these have extremely poor coverage, likely due to the failure of the independence assumptions inherent in the factorization (7).

Lastly, we briefly summarize additional cross sectional simulations; full results are given in Appendix Section F in the Supplementary Materials. We vary the measurement error variance

keeping the sample size fixed and find, as expected, that larger variances result in larger ISEs and score variance relative errors in estimation. Estimating fewer FPCs than actually exist (e.g., $K = 2$ in our design) does not detrimentally affect the estimates of the components that are estimated. Finally, because we induce smoothness in the estimated FPCs, using larger spline bases (e.g., $K_\theta = 30$) does not negatively affect performance.

5.2 Multilevel simulations

We next investigate how our model performs when random effects are included. We introduce subject-specific random effects for the scores by generating scores for each subject and FPC from a normal distribution with mean zero and variance $\lambda_{k|group} e^{g_{ik}}$, where g_{ik} is generated from a normal distribution with mean zero and variance $\sigma^2_{g_k}$. We set the variances $\sigma^2_{g_1}, \ldots, \sigma^2_{g_4}$ equal to $3, 1, 0.3$ and 0.1, respectively, and use the group-level baseline variances $\lambda_{k|group}$ shown in Table 1. We introduce random effects $\boldsymbol{b}_i$ for the mean curve for each subject by generating elements of a random effect spline coefficient vector independently from the standard normal distribution and then multiplying this vector by the spline basis functions.

To fit this model, we replace model 9 for the scores with the following model:

$$\xi_{ijk} \sim \mathrm{N}\left[0, \exp\left(\sum_{m=1}^{2} \mathbb{I}(w_i = m)\gamma_{km} + g_{ik}\right)\right], \tag{10}$$

where the g_{ik} are random effects, and we replace model 8 with the following model:

$$\boldsymbol{p}_{ij} = \boldsymbol{\Theta}\boldsymbol{\beta}_0 + \boldsymbol{\Theta}\boldsymbol{b}_i + \sum_{k=1}^{K} \xi_{ijk}\boldsymbol{\Theta}\boldsymbol{\phi}_k + \boldsymbol{\epsilon}_{ij},$$

where $\boldsymbol{b}_i$ are random effects.

We fix the sample size at $I = 40$ and vary the number of curves per subject $J_i \in \{8, 16, 32\}$. Two hundred replicate datasets were generated for each scenario. Figure 5 illustrates the results of this simulation. The top row shows the relative error of estimates of the $\lambda_{k|group}$ for each FPC

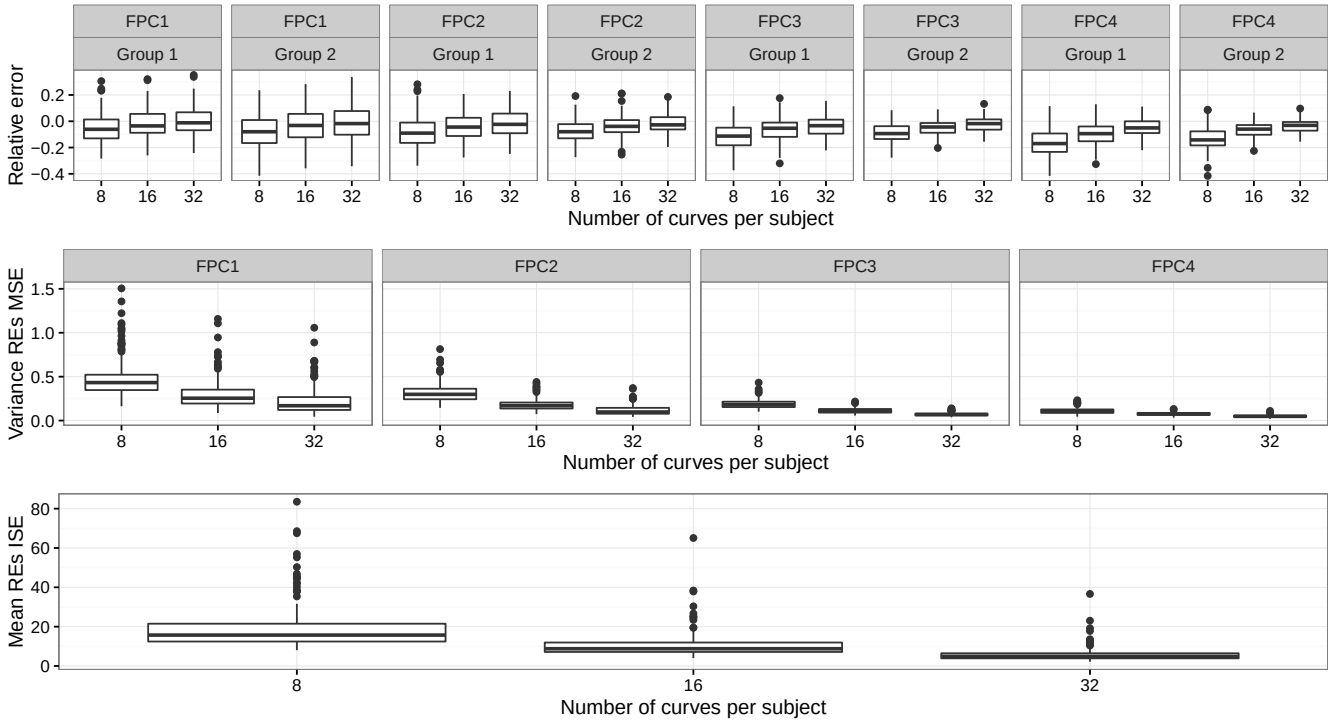

Figure 5: Longitudinal simulation results. The top row shows the relative error of estimates of the two parameters for each FPC score variance, one for each of the two groups, for sample size $I = 40$. The second row shows the mean squared error of estimates of the random effects g_{ik} in the score variance model. The bottom row shows the average ISE across subjects of estimates of the random effects $\boldsymbol{b}_i$ in the mean model. In all plots the number of curves per subject is varied in $\{8, 16, 32\}$.

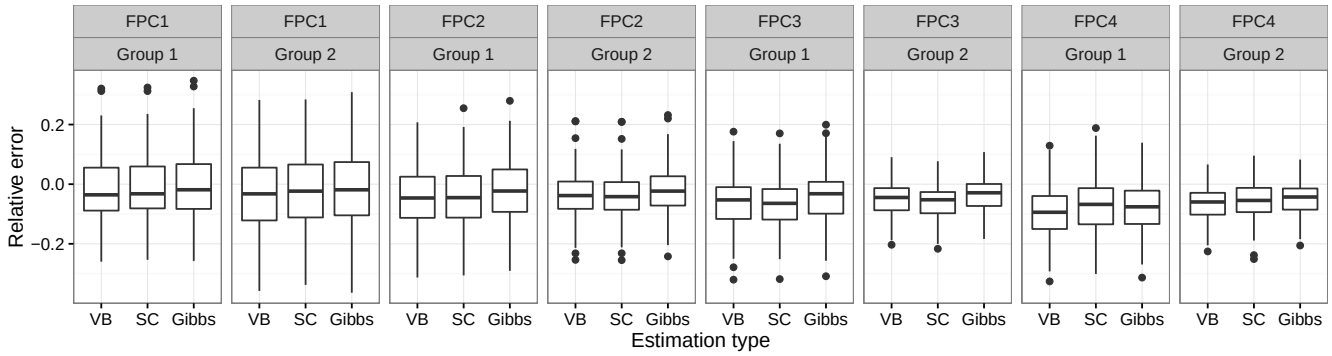

Figure 6: Comparison of three methods for longitudinal score variance model estimation. These boxplots compare the relative error in estimation of the score variance parameters using the VB, SC and Gibbs methods, for the longitudinal simulation scenario scenario with 40 subjects and 16 curves per subject.

and each group. Estimates are shrunk towards zero, especially for lower numbers of curves per subject. We attribute this to overfitting of the random effects in the mean model, which tends to decrease loadings of curves on the FPCs. The second row shows the mean squared error of estimates of the random effects g_{ik} in the score variance model. The bottom row shows the average ISE across subjects of estimates of the random effects $\boldsymbol{b}_i$ in the mean model. With more curves per subject, estimation of the random effects in both the score variance and mean models becomes more accurate.

We compare our variational Bayes estimation procedure to the SC and Gibbs alternatives in the longitudinal context. Prior to application of the SC and Gibbs procedures, random effects were estimated using function-on-scalar regression and subtracted from the simulated curves. For the SC procedure, we fit to the square of the scores the gamma generalized linear mixed-effects model induced by (10) (see Appendix Section E in the Supplementary Materials) using the `lme4` package for fitting linear mixed-effects models (Bates et al., 2015). Figure 6 compares the estimates of the score variance parameters for the scenario with 16 curves per subject. As in the cross sectional case, the results for the three methods are very similar. However, there is a significant increase in the computational burden for the Gibbs sampler: the VB procedure took about 70 seconds, the SC procedure about 40 seconds, and the Gibbs procedure about 80 minutes (running 3 chains for 2,000 iterations each).

Table 3 shows, based on the simulation scenario with 40 subjects and 16 curves per subject, that coverage of confidence intervals for the VB and Gibbs procedures is close to the nominal level. For FPCs 3 and 4 especially, the SC procedure confidence intervals are too narrow. For the SC method, replicates where there were convergence failures fitting the generalized linear mixed model are omitted (27 out of 200 replicates for FPC 3, and 15 out of 200 replicates for FPC 4).

FPC	Group	VB	SC	Gibbs
1	1	0.955	0.950	0.935
1	2	0.945	0.945	0.925
2	1	0.930	0.900	0.935
2	2	0.895	0.885	0.910
3	1	0.905	0.838	0.945
3	2	0.885	0.786	0.925
4	1	0.825	0.692	0.880
4	2	0.895	0.697	0.940

Table 3: Coverage of 95% confidence intervals for the score variance parameters using the VB, SC and Gibbs procedures, based on all simulation replicates with sample size $I = 40$ and number of curves per subject $J_i = 16$.

6 Analysis of kinematic data

We now apply the methods described above to the motivating data. 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. Our primary covariate of interest in the model for score variance is the repetition number. We expect that variance will be lower for later repetitions due to practice and learning. Since most of the variability of motions around the mean is explained by the first FPC, we use the score variance of the first FPC as a convenient summary for the motion variance, although we also present results for the second FPC.

Prior to fitting the model, we rotated 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, for all targets 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 briefly outline a bivariate approach to modeling this kinematic data.

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. 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 model for the mean and variance structure is therefore:

$$\boldsymbol{p}_{ij} = \sum_{l=1}^{8} \mathbb{I}(x_{ij0} = l) \left(\boldsymbol{\Theta}\boldsymbol{\beta}_l + \boldsymbol{\Theta}\boldsymbol{b}_{il}\right) + \sum_{k=1}^{K} \xi_{ijk}\boldsymbol{\Theta}\boldsymbol{\phi}_k \tag{11}$$

$$\xi_{ijk} \sim \mathrm{N}\left[0, \sigma^2_{\xi_{ijk}} = \exp\left(\sum_{l=1}^{8} \mathbb{I}(x_{ij0} = l) \left(\gamma_{lk0} + g_{ilk0} + (x_{ij1} - 1)(\gamma_{lk1} + g_{ilk1})\right)\right)\right]. \tag{12}$$

The covariate x_{ij0} indicates the target to which motion j by subject i is directed. The covariate x_{ij1} 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. $\mathbb{I}(\cdot)$ is the indicator function. To accommodate differences in baseline variance across targets, this model includes separate population-level intercepts γ_{lk0} for each target l. The slopes γ_{lk1} 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_{ilk0} and a random slope g_{ilk1} 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.

Throughout, fixed effects γ_{lk0} and γ_{lk1} are given $N[0, 100]$ priors. Random effects g_{ilk0} and g_{ilk1} are modeled using a bivariate normal distribution to allow for correlation between the random intercept and slope parameters in each FPC score variance model. The mean of this bivariate normal distribution is set to 0; its covariance matrix is given an inverse-Wishart prior with scale matrix Σ_{g_k} and 2 degrees of freedom. We use empirical Bayes to select the hyperparameter Σ_{g_k}: in short, we use the smoothed covariance approach to obtain estimates of random effects g_{il10} and g_{il11}, and set the hyperparameter to be the empirical covariance of these estimates.

We fit models (11) and (12) using $K = 2$ principal components and a cubic B-spline evaluation matrix $\mathbf{\Theta}$ with $K_\theta = 10$ basis functions. In addition to presenting our variational Bayes models, we also present results from using a Gibbs sampler to estimate the scores and the parameters in model (12), conditioning on FPCs estimated using a smoothed covariance decomposition and subject-specific mean curves estimated using function-on-scalar regression. We ran this Gibbs sampler with 4 chains for 15,000 iterations, discarding the first 3,000 iterations; doing so took just over 24 hours using 4 cores running in parallel. The Gelman-Rubin diagnostic for each variable was less than 1.06, suggesting convergence to the posterior distribution. Results for the left hand and results for the bivariate analysis are presented in Appendix Sections B and C in the Supplementary materials.

6.2 Results

Figure 7 shows estimated score variances as a function of repetition number for the first two FPCs, and includes estimates for the X and Y coordinate models for all targets in the right hand. 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

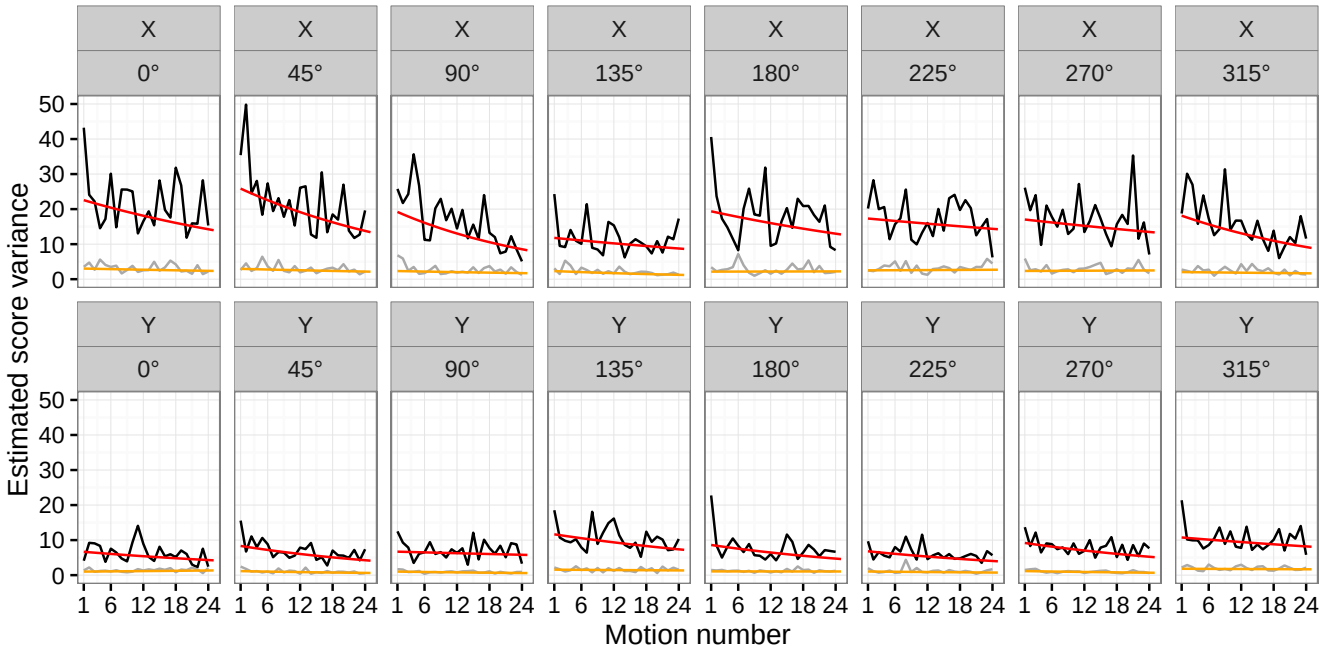

Figure 7: Estimates of score variances in the right hand for each target (in columns) and direction (X or Y, in rows). Panels show the estimates of the score variance as a function of repetition number using the slope-intercept model (12) in red and orange (first and second FPC, respectively), and using the saturated one-parameter-per-repetition number model (13), in black and grey (first and second FPC, respectively).

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 (12) with a model for the score variances saturated in repetition number, i.e., where each repetition number j has its own parameter γ_{jk} in the model for the score variances:

$$\xi_{ijk} \sim \mathrm{N}\left[0, \sigma^2_{\xi_{ijk}} = \exp\left(\sum_{j=1}^{24} \gamma_{jk} \mathbb{I}(x_{ij1} = j)\right)\right]. \tag{13}$$

The results for these two models are included 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

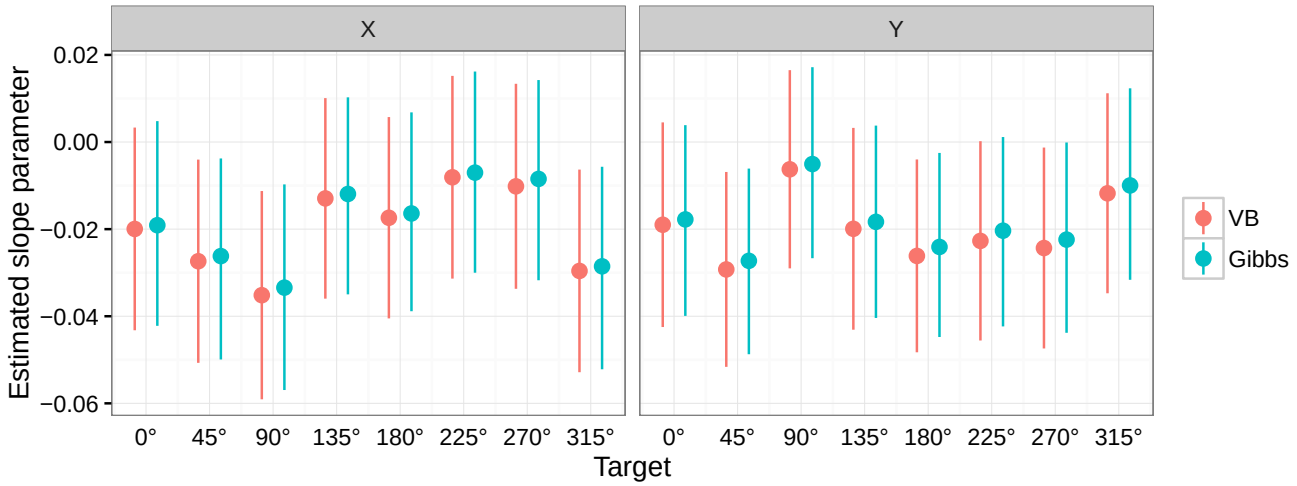

Figure 8: Estimates of γ_{l11} using a variational Bayes approximation and Gibbs sampling, conditioning on FPCs and estimates of subject-specific mean curves. This figure shows estimates and upper and lower 95% confidence bounds for target-specific score variance slope parameters γ_{l11} for motions by the right hand to each target for the VB and Gibbs methods, for the X and Y coordinates.

scores. We are interested in the parameters γ_{l11}, which estimate the population-level target-specific changes in score variance for the first principal component with each additional motion. Figure 8 shows both estimates and 95% confidence bounds for the γ_{l11} parameters for motions by the right hand to each target, for both the VB and Gibbs methods and for both coordinates. All the point estimates γ_{l11} are lower than 0, indicating decreasing first principal component score variance with additional repetition. For some targets and coordinates there is substantial evidence against the null hypothesis $\gamma_{l11} = 0$; these results are consistent with our understanding of motor learning, although they do not adjust for multiple comparisons. The estimates and confidence bands from the VB and Gibbs are nearly identical, although the Gibbs estimates are slightly closer to zero than the VB estimates.

Appendix Section B of the Supplementary Materials contains results for the left hand, which are broadly similar to those for the right hand: clear training effects can be observed in both the X and Y directions, and the first principal component explains most of the overall variability. Appendix Section C includes results of a bivariate approach to modeling motion kinematics, which acknowledges that motions exist in two dimensions. In short, the X and Y coordinates of curves are concatenated for each motion, 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, closely follow the preceding univariate analyses.

7 Discussion

We have focused here on developing 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. New in this context, our approach allows for the estimation of subject-specific effects on variance, as well as the consideration of multiple covariates.

Applying our methods to our motivating dataset, we have demonstrated that motion variance is reduced with repetition. Results in Appendix Section B in the Supplementary Material 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.

An alternative hierarchical approach to the analysis of this dataset could treat the target effects γ_{lk0} and γ_{lk1} in model (12) for the score variances as random effects centered around parameters μ_{k0} and μ_{k1}, 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 μ_{k0} and μ_{k1}.

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, as it could also be applied to sparsely observed functional data. Our research group is currently working on developing an improved approach to registration in kinematic experiments. This approach 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. Allowing for outlying scores, for example by posing a prior distribution with wider tails or a mixture of normal distributions, would extend the applicability of the model. Extensions of our approach to exponential-family response curves (e.g., binary-valued "curves") will be useful in several scientific contexts, including studies of physical activity. 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 online Supplementary Materials includes an appendix giving

- a graphical representation of our real data analysis model;
- results of the application of our model to the left hand;
- results for the bivariate approach;
- a sensitivity analysis for our choice of priors in our Bayesian model;
- derivations of full conditional distributions and our variational Bayes algorithm;
- and additional simulation results.

The Supplementary Materials also include code implementing our methods, including code implemented all the simulation scenarios presented here.

9 Acknowledgments

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).

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., 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).

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).

Plummer, M. "JAGS: A program for analysis of Bayesian graphical models using Gibbs sampling." In Proceedings of the 3rd International Workshop on Distributed Statistical Computing (DSC 2003). March, 20–22 (2003).

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).

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).

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

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

This supplementary material contains the following appendices: A, containing a graphical representation of our model; B, containing results from applying our univariate model to the X and Y coordinates of curves for the left hand; 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 priors in our Bayesian model; E, containing a derivation of our variational Bayes algorithm and F, containing some additional simulation results.

A Graphical model

Figure A.1 contains a graphical illustration of models (11) and (12). Shaded nodes represent observed data. Blank nodes denote inferred parameters. Rectangles or plates denote indexing over a set of variables. For example, the plate surrounding $\boldsymbol{p}_{ij}$ indicates that for subject i, observations are indexed by j and are independent given parameters outside this plate, like, for example, the subject and target-specific random effects $\boldsymbol{b}_{il}$, g_{ilk0} and g_{ilk1}.

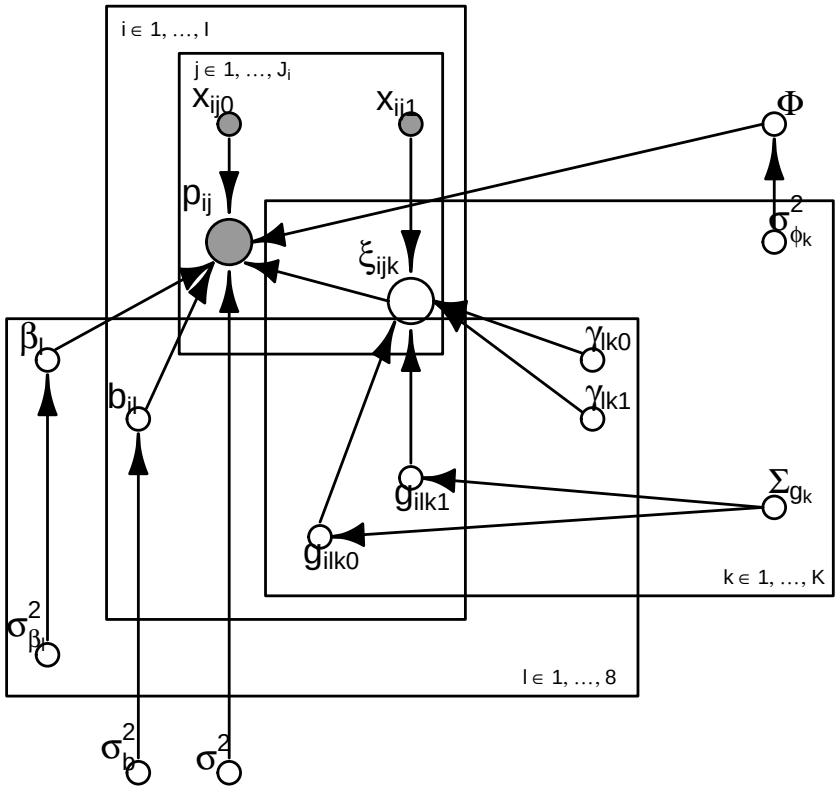

Figure A.1: Graphical illustration of models (11) and (12).

B Results for the left hand

Figure A.2 shows the change in variability of the first and second FPC scores as a function of practice for the left hand. As for the right hand (Figure 7), the slopes for the first principal component score variance as a function of motion number are all negative, indicating decreased motion variance with training. Except for one target for the y direction, baseline variability for the first FPC scores is higher in both directions and for all targets in the left hand than in the right hand, as expected for the non-dominant hand. For the targets at 135° and 315°, as in the right hand, there is more variability perpendicular to the line connecting the target and the origin than for the other targets.

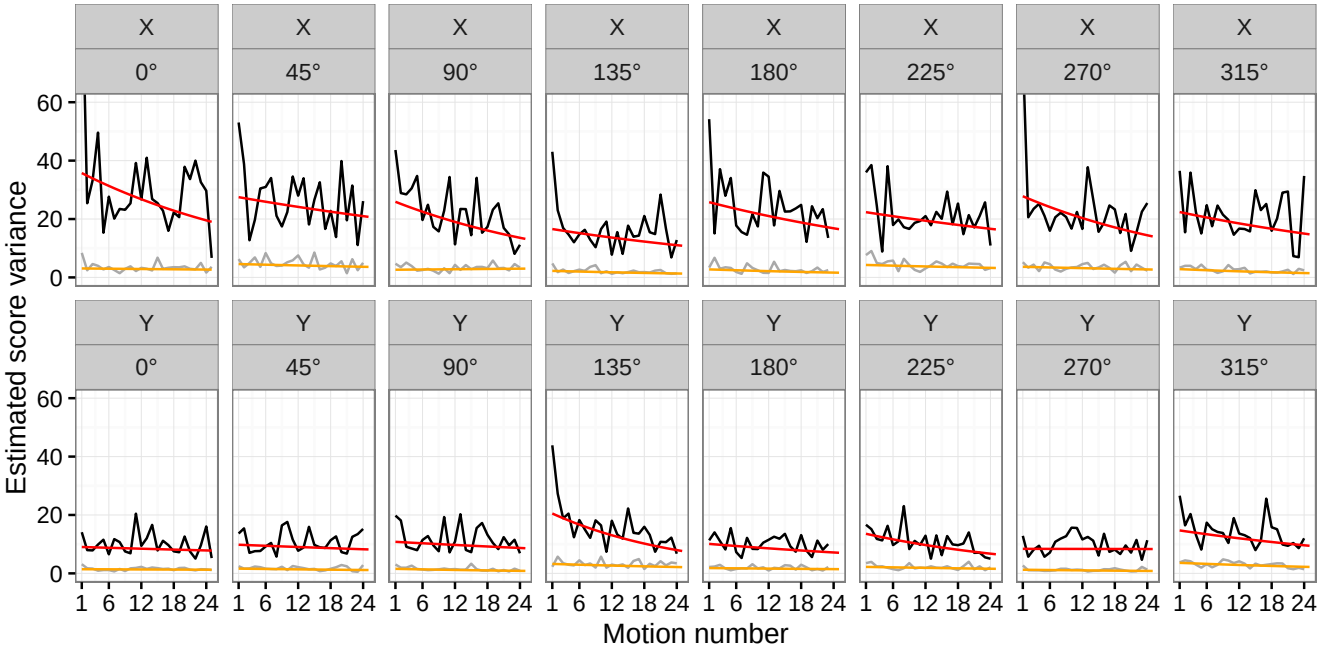

Figure A.2: Estimates of score variance in the left hand for each target (in columns) and direction (X or Y, in rows). Panels show the estimates of the score variance as a function of repetition number using the slope-intercept model (12) in red and orange (first and second FPC, respectively), and using the saturated one-parameter-per-repetition number model (13), in black and grey (first and second FPC, respectively).

Figure A.3 shows the estimates and confidence bounds for the target-specific score variance slope parameters γ_{l11} for each coordinate of motions by the left hand to each target. All but one of the point estimates are less than 0, indicating decreasing motion variance with training, and for some targets there is substantial evidence against the null hypothesis that $\gamma_{l11} = 0$.

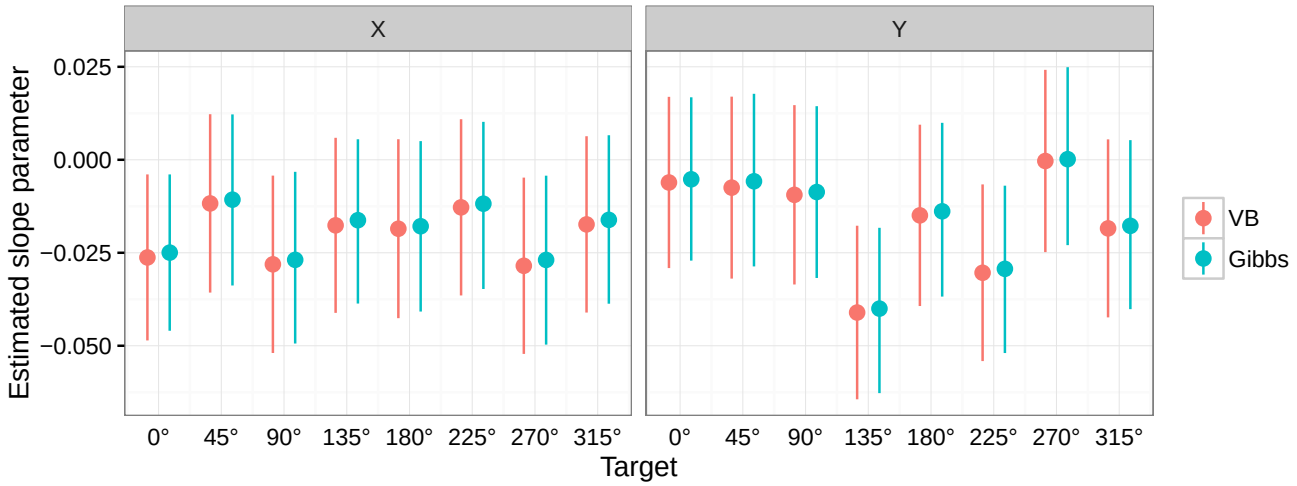

Figure A.3: Estimates of γ_{l11} for the left hand using a variational Bayes approximation and Gibbs sampling, conditioning on FPCs and estimates of subject-specific mean curves. This figure shows estimates and upper and lower 95% confidence bounds for target-specific score variance slope parameters γ_{l11} for motions by the left hand to each target, for the VB and Gibbs methods, for the X and Y coordinates.

There are several scientifically relevant questions about individual motion characteristics that are addressable in our modeling framework. As an example, we examine whether subjects with high baseline motion variance to one target tend to have high baseline motion variance to other targets. Figure A.4 shows the estimated random intercept parameters g_{il10} for each subject and each target for both the left and right hands for the x direction 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. Treating the random intercepts as observed data, we calculated the intraclass correlation coefficient to be 0.34 for the left-hand random intercepts and 0.45 for the right-hand random intercepts, indicating that subjects tend to have consistent patterns of variability across different targets with the same hand. To determine if the random intercepts in the left and right hand were correlated, we calculated the average random intercept across all targets for each subject, in the left and right hands separately. The correlation of these two vectors, one for the left and one for the right hand, was 0.68, indicating a positive correlation between baseline motor skill across hands within an individual.

When we fit models for the left hand, we use the same empirical Bayes hyperparameter Σ_{g_k} as we

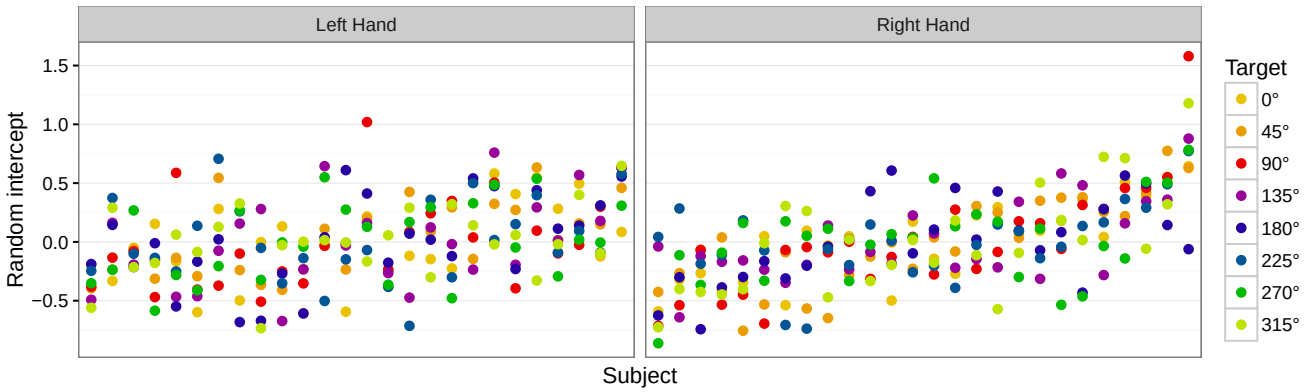

Figure A.4: Estimates of random intercepts. Each panel shows, for the left or the right hand, the estimated score variance random intercept parameters g_{il10} in model (12) for each subject i and target l, for the first principal component for the X coordinate of motion. Targets are colored as in Figure 1, and subjects are ordered by their average random intercept across targets for the right hand.

use for the model for the right hand, for the sake of the consistency, and since the hyperparameter estimated using data for the left hand is similar.

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) $\begin{pmatrix} \sigma^2_{p*,x} & 0 \\ 0 & \sigma^2_{p*,y} \end{pmatrix} \otimes \boldsymbol{Q}^{-1}_{K_\theta}$, where $\otimes$ is the Kronecker product operator, $\sigma^2_{p*,x}$ and $\sigma^2_{p*,y}$ are independent with IG $[\alpha, \beta]$ priors and $\boldsymbol{Q}_{K_\theta}$ is the 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.

Figure A.5 illustrates the FPCs estimated using model 11 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 showing similar patterns of variation but in different dimensions. The same pattern repeats, to a lesser extent, for the third and fourth PCs estimated using 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 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

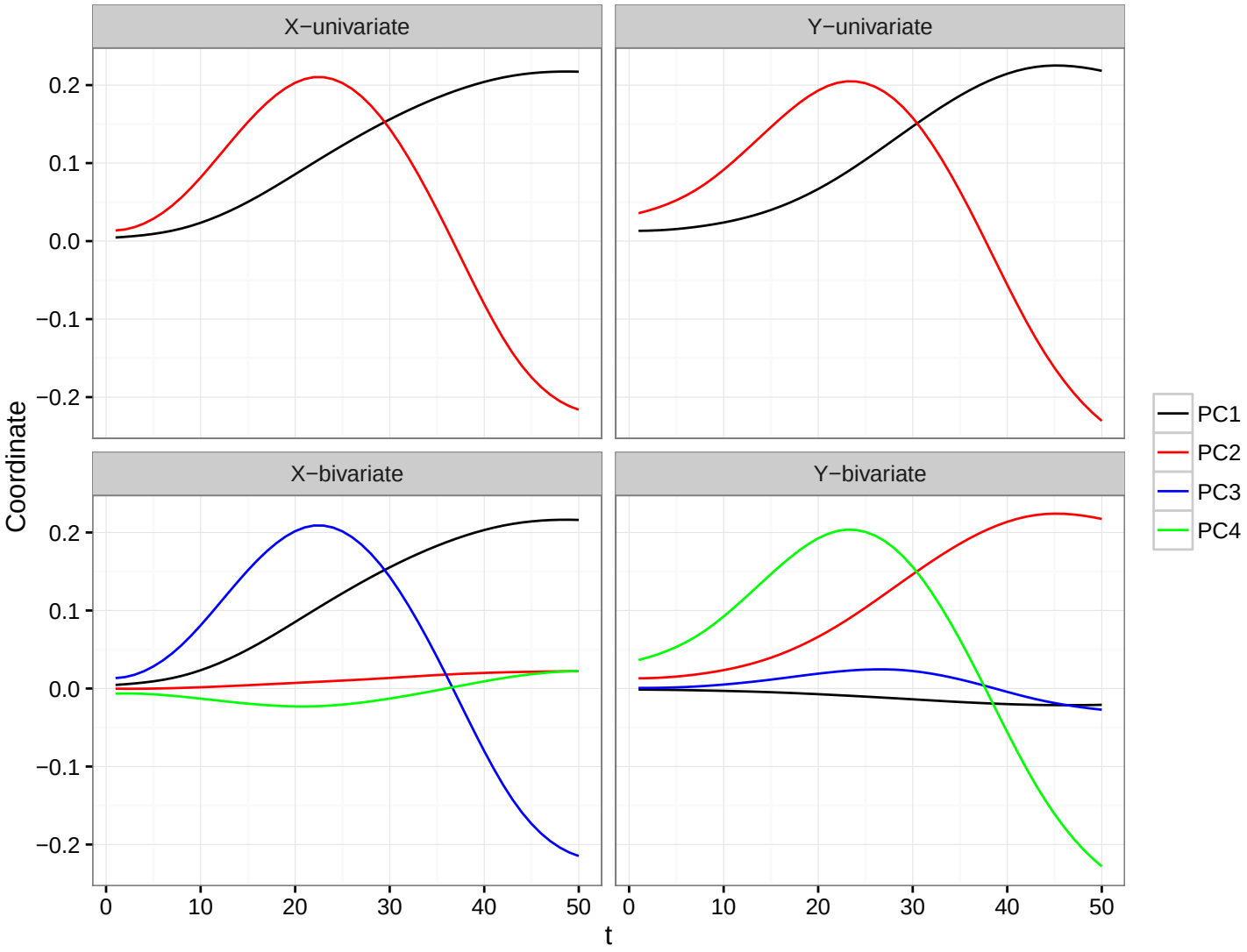

Figure A.5: FPCs from model 11 fitted to the univariate and bivariate data. The FPCs on the left are for the X coordinates of motions, those on the right are for the Y-coordinate. The FPCs in the top row were estimated using univariate models, and the FPCs in the bottom row were estimated using bivariate models.

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.6 show the estimated overall mean functions and estimated individual-specific mean

functions from the bivariate model. The X coordinate profiles of estimated mean functions are much more consistent across targets than the Y coordinate profiles, which, for example, have different curvature depending on the target to which the motion is directed.

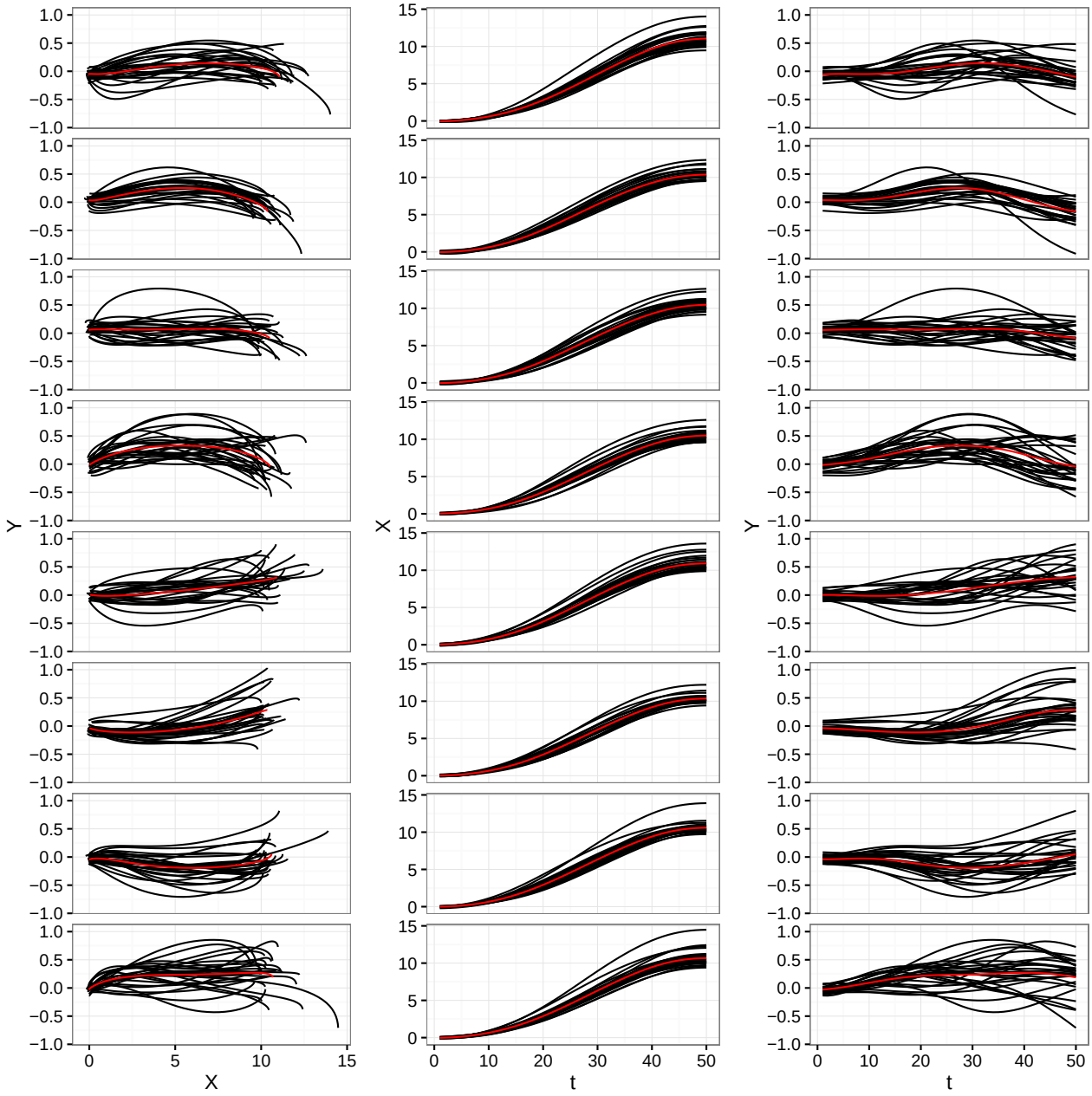

Figure A.6: Overall mean (red) and subject-specific (black) mean curves from bivariate model.

Figure A.7 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.

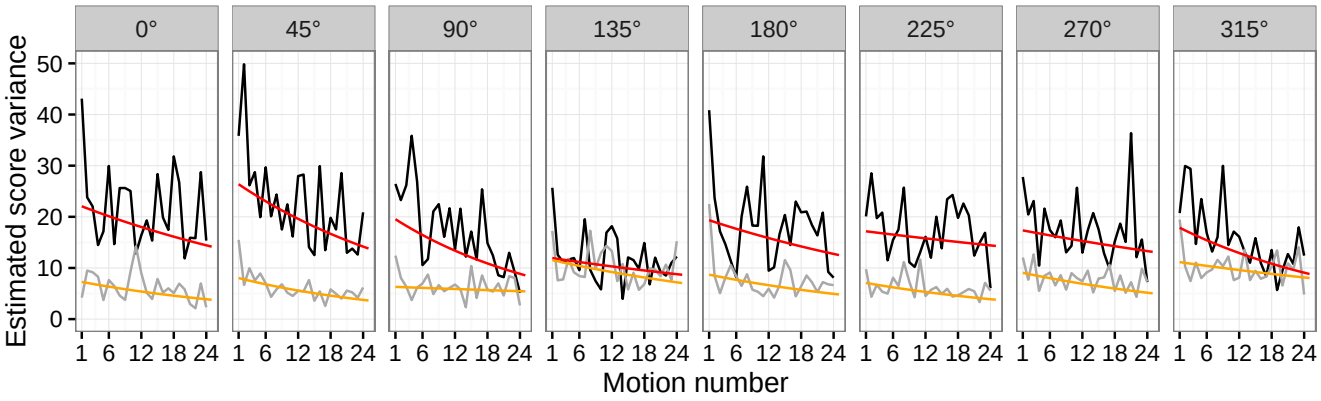

Figure A.7: Estimates of bivariate FPC score variances in the right hand for each target. Panels show the estimates of the score variance as a function of repetition number using the slope-intercept model (12) in red and orange (first and second FPC, respectively), and using the saturated one-parameter-per-repetition number model (13), in black and grey (first and second FPC, respectively).

When we fit bivariate models, we use the same empirical Bayes hyperparameter Σ_{g_k} as we use for the model for the right hand, for the sake of the consistency, and since the hyperparameter estimated using data for the bivariate data is similar.

D Sensitivity Analysis

In our sensitivity analysis we focus on the parameter of principal interest to us, the fixed effect parameters γ_{l11}, which measure how much variability of the first FPC scores decreases with each additional motion. We plot in Figure A.8 the estimates of γ_{l11} as a result of changing the variance of the prior for γ_{l11} from 10 to 100 to 1000. As can be seen in the figure, the variance of the prior for γ_{l11} has a negligible impact on estimation and inference.

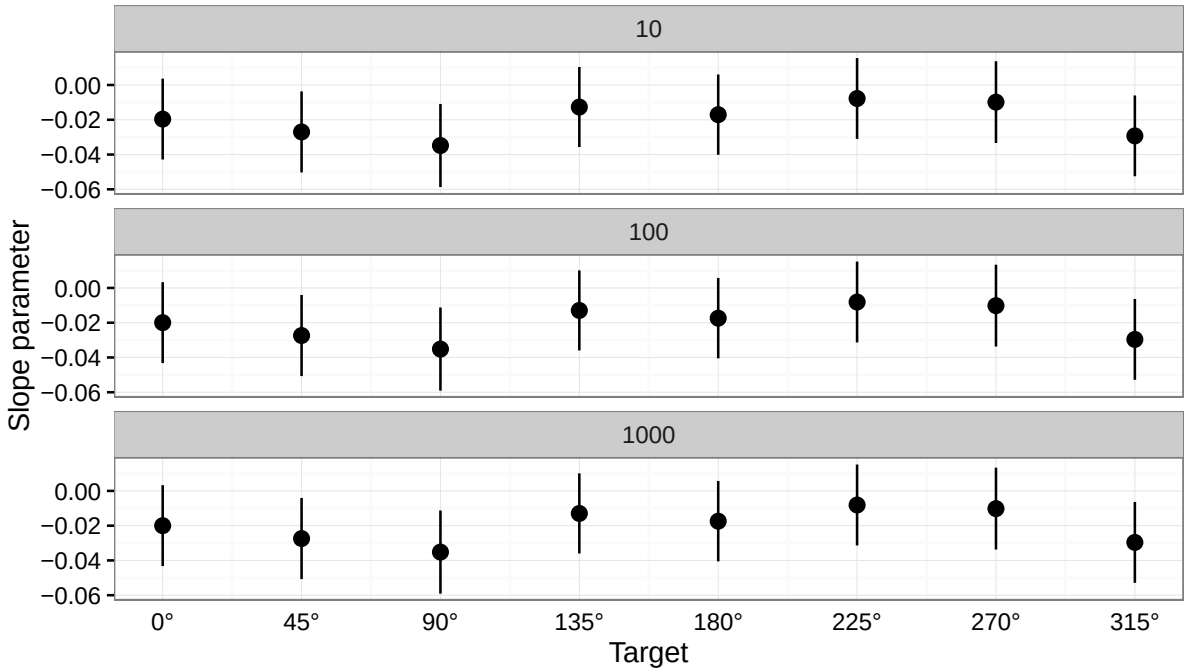

Figure A.8: Estimates and 95% confidence intervals for γ_{l11} as a function of the variance of its normal prior.

E Derivations

$$
\begin{aligned}
\boldsymbol{p}_{ij} &= \sum_{l=0}^{p} x_{ijl}\boldsymbol{\Theta}\boldsymbol{\beta}_l + \boldsymbol{\Theta}\boldsymbol{b}_i + \sum_{k=1}^{K} \xi_{ijk}\boldsymbol{\Theta}\boldsymbol{\phi}_k + \boldsymbol{\epsilon}_{ij} \qquad \text{(A.1)} \\
\boldsymbol{\beta}_l &\sim \text{MVN}\left[0, \sigma^2_{\boldsymbol{\beta}_l}\boldsymbol{Q}^{-1}_{K_\theta}\right]; \sigma^2_{\boldsymbol{\beta}_l} \sim \text{IG}\left[\alpha, \beta\right] \\
\boldsymbol{b}_i &\sim \text{MVN}\left[0, \sigma^2_{\boldsymbol{b}}\boldsymbol{Q}^{-1}_{K_\theta}\right]; \sigma^2_{\boldsymbol{b}} \sim \text{IG}\left[\alpha, \beta\right] \\
\boldsymbol{\phi}_k &\sim \text{MVN}\left[0, \sigma^2_{\boldsymbol{\phi}_k}\boldsymbol{Q}^{-1}_{K_\theta}\right]; \sigma^2_{\boldsymbol{\phi}_k} \sim \text{IG}\left[\alpha, \beta\right] \\
\xi_{ijk} &\sim \text{N}\left[0, \exp\left(\sum_{m=0}^{q} \gamma_{km} w_{ijkm} + \sum_{h=1}^{r} g_{ikh} z_{ijkh}\right)\right] \\
\gamma_{km} &\sim \text{N}\left[0, \sigma^2_{\gamma_{km}}\right] \\
g_{ikh} &\sim \text{N}\left[0, \sigma^2_{g_{kh}}\right]; \sigma^2_{g_{kh}} \sim \text{IG}\left[\alpha, \beta\right] \\
\boldsymbol{\epsilon}_{ij} &\sim \text{MVN}\left[0, \sigma^2\boldsymbol{I}_D\right]; \sigma^2 \sim \text{IG}\left[\alpha, \beta\right]
\end{aligned}
$$

Above, $i = 1, \ldots, I$ refers to subjects, $j = 1, \ldots, J_i$ refers to motions within subjects, and $k = 1, \ldots, K$ refers to principal components. $\boldsymbol{p}_{ij}$ is the $D \times 1$ observed functional outcome for the jth curve of the ith subject, $\boldsymbol{\beta}_l$ for $l = 0, \ldots, p$ are functional effect coefficient vectors, $\boldsymbol{b}_i$ for $i = 1, \ldots, I$ are random effect coefficient vectors, and $\boldsymbol{\phi}_k$ for $k = 1 \ldots, K$ are principal component coefficient vectors, all of length K_θ. $\boldsymbol{\epsilon}_{ij}$ is a $D \times 1$ vector of independent error terms. $\boldsymbol{\Theta}$ is a $D \times K_\theta$ matrix of orthonormal basis functions, $\boldsymbol{I}_D$ is the $D \times D$ identity matrix, and $\boldsymbol{Q}_{K_\theta}$ is a penalty matrix.

In our real data application, we consider a model where two random effects g_{ik1} and g_{ik2} for the ith subject have a bivariate, mean-zero normal prior distribution with covariance matrix $\boldsymbol{\Sigma}_{g_k}$. Let the vector of these two random effects for the ith subject and kth principal component be $\boldsymbol{g}_{ik}$. We give the covariance matrix $\boldsymbol{\Sigma}_{g_k}$ an inverse-Wishart prior distribution, yielding the model

specification

$$\boldsymbol{g}_{ik} \sim \text{MVN}\left[0, \boldsymbol{\Sigma}_{g_k}\right]$$

$$\boldsymbol{\Sigma}_{gk} \sim \text{IW}\left[\boldsymbol{\Psi}, \nu\right].$$

In our real data application, we also consider a model where each functional outcome vector $\boldsymbol{p}_{ij}$ is a $2D \times 1$ bivariate observation instead of a $D \times 1$ univariate observation. The first D components of each vector $\boldsymbol{p}_{ij}$ are the x coordinates of the observation, and the last D components are the y coordinates, so the new matrix of basis functions is $\left(\begin{smallmatrix} \boldsymbol{\Theta} & 0 \\ 0 & \boldsymbol{\Theta} \end{smallmatrix}\right)$. This does not entail any changes to the model for the variance of the scores, but it does result in the following modified model for the mean of the functional observations, where now $\boldsymbol{\beta}_l, \boldsymbol{b}_i$ and $\boldsymbol{\phi}_k$ are vectors of length $2K_\theta$:

$$\begin{aligned}
\boldsymbol{\beta}_l &\sim \text{N}\left[0, \text{diag}\left(\sigma^2_{\boldsymbol{\beta}_l^x}, \sigma^2_{\boldsymbol{\beta}_l^y}\right) \otimes \boldsymbol{Q}^{-1}_{K_\theta}\right]; \sigma^2_{\boldsymbol{\beta}_l^x}, \sigma^2_{\boldsymbol{\beta}_l^y} \sim \text{IG}\left[\alpha, \beta\right] \\
\boldsymbol{b}_i &\sim \text{N}\left[0, \text{diag}\left(\sigma^2_{\boldsymbol{b}^x}, \sigma^2_{\boldsymbol{b}^y}\right) \otimes \boldsymbol{Q}^{-1}_{K_\theta}\right]; \sigma^2_{\boldsymbol{b}^x}, \sigma^2_{\boldsymbol{b}^y} \sim \text{IG}\left[\alpha, \beta\right] \\
\boldsymbol{\phi}_k &\sim \text{N}\left[0, \text{diag}\left(\sigma^2_{\boldsymbol{\phi}_k^x}, \sigma^2_{\boldsymbol{\phi}_k^y}\right) \otimes \boldsymbol{Q}^{-1}_{K_\theta}\right]; \sigma^2_{\boldsymbol{\phi}_k^x}, \sigma^2_{\boldsymbol{\phi}_k^y} \sim \text{IG}\left[\alpha, \beta\right].
\end{aligned}$$

In the above expressions $\text{diag}\left(\boldsymbol{c}\right)$ is the matrix with the elements of $\boldsymbol{c}$ on its main diagonal and 0 elsewhere and $\otimes$ is the Kronecker product operator.

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:

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

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 $\sigma^2_{\beta_l}$ or, in the model with bivariate observations, the length $2l+2$ vector of prior variances $(\sigma^2_{\beta^x_0}\sigma^2_{\beta^y_0}, \ldots, \sigma^2_{\beta^x_l}, \sigma^2_{\beta^y_l})$. 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 $\boldsymbol{\Sigma_\beta} = \text{diag}\left(\boldsymbol{\sigma}^2_{\boldsymbol{\beta}}\right) \otimes \boldsymbol{Q}^{-1}_{K_\theta}$. The posterior distribution of $\text{vec}\,(\boldsymbol{\beta})$ is then

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

Using the identity

$$\text{vec}\,(\boldsymbol{ABC}) = (\boldsymbol{C}^T \otimes \boldsymbol{A})\text{vec}\,(\boldsymbol{B}) \tag{A.2}$$

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:

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

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

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

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 $\boldsymbol{\Sigma_b} = \text{diag}\,(\boldsymbol{\sigma_b^2}) \otimes \boldsymbol{Q}_{K_\theta}^{-1}$, where, in the model with bivariate observations, $\boldsymbol{\sigma_b^2} = (\sigma_{\boldsymbol{b}^x}^2, \sigma_{\boldsymbol{b}^y}^2)$. Let $\boldsymbol{P}_i, \boldsymbol{X}_i$ and $\boldsymbol{\Xi}_i$ be the submatrices of the matrices $\boldsymbol{P}, \boldsymbol{X}$ and $\boldsymbol{\Xi}$ corresponding to the observations for the ith subject. The posterior distribution of $\boldsymbol{b}_i$ is then

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

that is, multivariate normal with covariance matrix

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

and mean

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

Letting $\boldsymbol{\sigma_\Phi^2}$ be the length K vector of prior variances $\sigma_{\boldsymbol{\phi}_k}^2$ (or, in the model with bivariate observations, the length $2K$ vector $(\sigma_{\boldsymbol{\phi}_1^x}^2, \sigma_{\boldsymbol{\phi}_1^y}^2, \ldots, \sigma_{\boldsymbol{\phi}_K^x}^2, \sigma_{\boldsymbol{\phi}_K^y}^2)$), the covariance matrix of the normal prior distribution of $\text{vec}\,(\boldsymbol{\Phi})$ is $\boldsymbol{\Sigma_\Phi} = \text{diag}\,(\boldsymbol{\sigma_\Phi^2}) \otimes \boldsymbol{Q}_{K_\theta}^{-1}$. The posterior distribution of $\text{vec}\,(\boldsymbol{\Phi})$ is then

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

that is, multivariate normal with covariance matrix

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

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 $\boldsymbol{\Sigma}_{\boldsymbol{\xi}_{ij}} = \text{diag}\left(\boldsymbol{\sigma}^2_{\boldsymbol{\xi}ij}\right)$, where $\boldsymbol{\sigma}^2_{\boldsymbol{\xi}_{ij}}$ is the length K vector of prior variances for $\boldsymbol{\xi}_{ij}$. Then the posterior distribution of $\boldsymbol{\xi}_{ij}$ is

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

that is, multivariate normal with covariance matrix

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

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 $\boldsymbol{w}_{ijk}$ be the length $q+1$ 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 r 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 $\sigma^2_{\gamma_{km}}$, 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 $\boldsymbol{g}_{ik}, i = 1, \ldots, I$ is then

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

which has the form of the posterior of a gamma generalized linear model with log link, responses given by ξ_{ijk}^2, 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

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

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:

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

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

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

of σ^2:

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

and of $\sigma^2_{g_{kh}}$:

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

In our real data application, we consider a model where two random effects g_{ik1} and g_{ik2} have a bivariate, mean-zero normal prior distribution with covariance matrix $\mathbf{\Sigma}_{g_k}$. This covariance matrix has an inverse-Wishart prior distribution. The inverse-Wishart density is $p(\mathbf{\Sigma}|\mathbf{\Psi},\nu) = |\mathbf{\Sigma}|^{-\frac{\nu+p+1}{2}} \exp\left(-\frac{1}{2}\text{tr}\left[\mathbf{\Psi}\mathbf{\Sigma}^{-1}\right]\right)$, where p is the number of rows and columns of the covariance matrix $\mathbf{\Sigma}$. The conditional posterior distribution of $\mathbf{\Sigma}_{g_k}$ is therefore

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

E.2 Derivation of variational Bayes algorithm

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

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

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

and mean

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

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

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

and mean

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

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

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

and mean

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

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

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

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 $M_{ij} = \text{tr}\left[\boldsymbol{\Theta}^T\boldsymbol{\Theta}\text{cov}(\boldsymbol{\phi}_i, \boldsymbol{\phi}_j)\right]$ and $\text{cov}(\boldsymbol{\phi}_i, \boldsymbol{\phi}_j)$ is a submatrix of $\boldsymbol{\Sigma}_{q(\text{vec}(\Phi))}$. The expectation $\mu_{q(\Xi^T\Xi)}$ appearing in the above expression for $\boldsymbol{\Sigma}_{q(\text{vec}(\Phi))}$ 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 $\sigma^2_{\boldsymbol{\beta}_l}$, $\sigma^2_{\boldsymbol{b}}$, $\sigma^2_{\phi_k}$ and σ^2, we use the following result: if the conditional distribution of a parameter ϕ is inverse gamma with parameters α and β, then the distribution $q^*(\phi)$ is inverse gamma with parameters $E_{-\phi}[\alpha]$ and $E_{-\phi}[\beta]$, and the expectation $\mu_{q(1/\phi)}$ is $E_{-\phi}[\alpha]/E_{-\phi}[\beta]$.

For $\sigma^2_{\boldsymbol{\beta}_l}$, the optimal density $q^*(\sigma^2_{\boldsymbol{\beta}_l})$ is inverse gamma with parameters $\alpha + \frac{K_\theta}{2}$ and $\beta + \frac{1}{2}\mu_{q\left(\beta_l^T Q_{K_\theta}\beta_l\right)}$. For $\sigma^2_{\boldsymbol{b}}$, the optimal density $q^*(\sigma^2_{\boldsymbol{b}})$ is inverse gamma with parameters $\alpha + \frac{IK_\theta}{2}$ and $\beta + \frac{1}{2}\mu_{q\left(b_i^T Q_{K_\theta} b_i\right)}$. For $\sigma^2_{\boldsymbol{\phi}_k}$, the optimal density $q^*(\sigma^2_{\boldsymbol{\phi}_k})$ is inverse gamma with parameters $\alpha + \frac{K_\theta}{2}$ and $\beta + \frac{1}{2}\mu_{q\left(\phi_k^T Q_{K_\theta}\phi_k\right)}$. 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 $\alpha + \frac{nD_\theta}{2}$ and

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

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

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

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

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

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 $\mu_{q(\boldsymbol{g}_{ik})}\mu^T_{q(\boldsymbol{g}_{ik})} + M$, 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

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

F Additional simulation results

This section includes some additional simulation results. All of these simulations use the same simulation framework as that used in the cross-sectional simulation described in section 5.1, with 80 curves per replicate simulation and 200 replicates per simulation setting.

Figure A.9 shows the result of estimating fewer FPCs than actually exist. The top row shows integrated squared errors in estimation of the FPCs, and the bottom row shows relative error in estimation of the score variance parameters. 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.10 shows the result of changing the number of spline basis functions, from the 10 used in section 5.1 to 5, 20 and 30. Again, the top row shows integrated squared errors in estimation of the FPCs, and the bottom row shows relative error in estimation of the score variance parameters. 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 Q_{K_θ}, using richer spline bases does not negatively affect estimation accuracy.

Figure A.11 shows the result of adding more noise to the simulated curves, keeping the sample size fixed. The top row shows one simulated curve at each of the different settings of the measurement error standard deviation. The second panel illustrates the estimated FPCs when the standard deviation equals 2. The third panel shows integrated squared errors in estimation of the FPCs and the bottom error shows the relative error in estimation of the score variance parameters. As expected, larger variances results in larger errors in estimation, of both the FPCs and the score variance parameters.

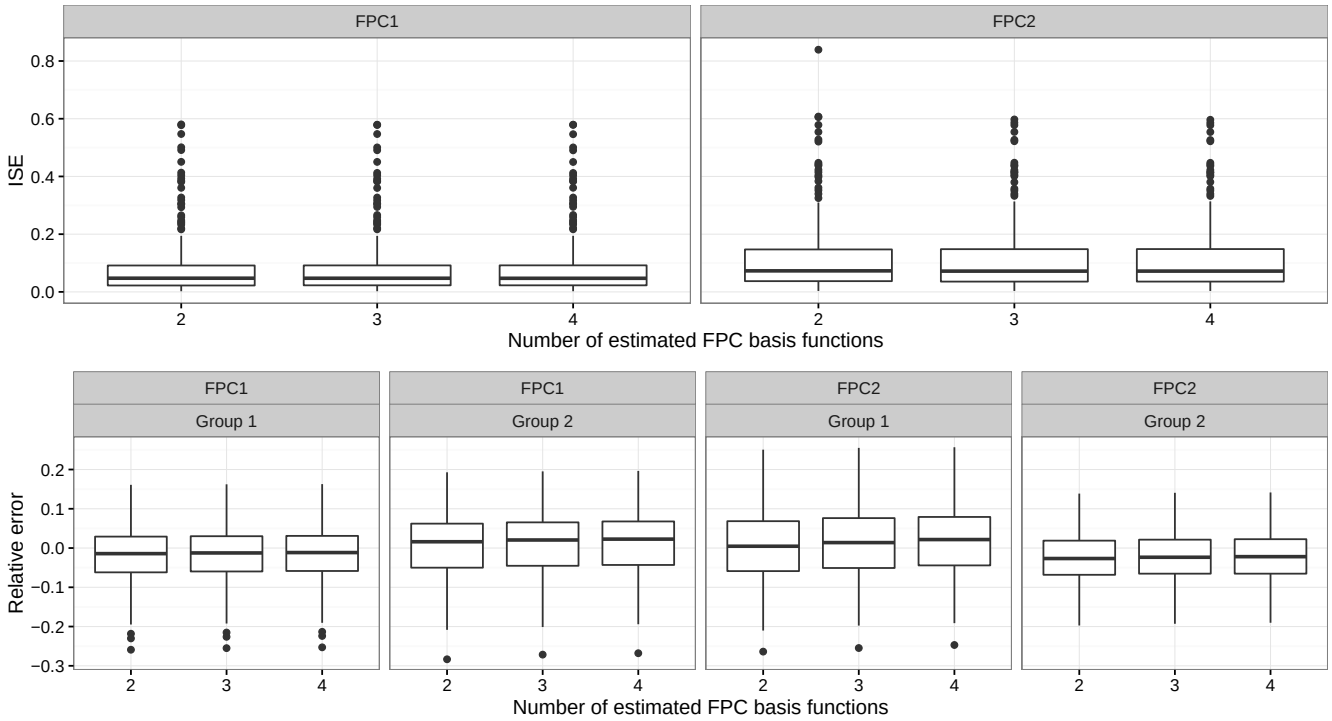

Figure A.9: Varying the number of estimated FPCs. Using the cross-sectional simulation set up described in Section 5 with 80 curves per simulation, we varied the number of FPCs used in estimation to 2 and 3. As shown here, ISE in estimation of FPCs (first row) and relative error in estimation of variance parameters (second row) for FPCs 1 and 2 is mostly invariant to whether additional FPCs and associated score variance parameters are also estimated. Results shown are for 200 replicates per simulation scenario.

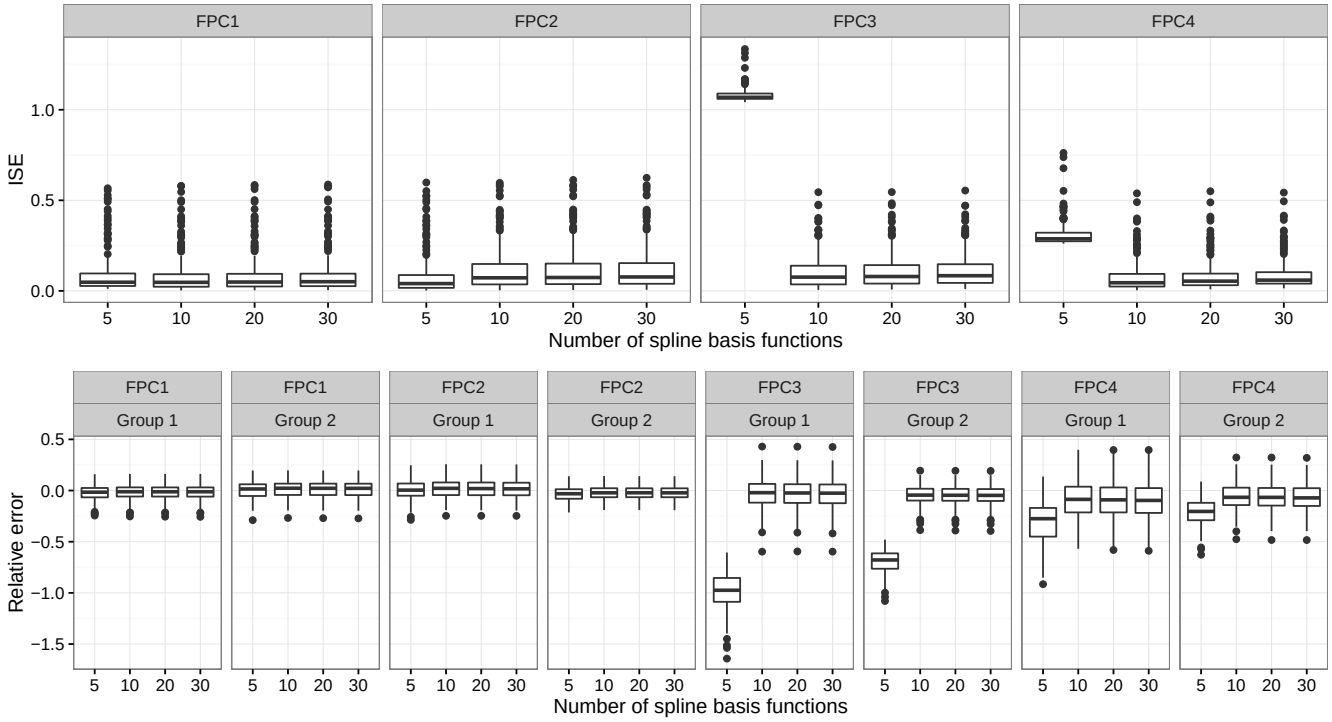

Figure A.10: Varying the number of spline basis functions. Using the cross-sectional simulation set up described in Section 5 with 80 curves per simulation, we varied the number of spline basis functions to 5, 20 and 30. As shown here, 5 spline basis functions are not sufficient to adequately capture the relatively fast variation in FPCs 3 and 4. Otherwise ISE in estimation of FPCs (first row) and relative error in estimation of variance parameters (second row) is mostly invariant to the number of spline basis functions used in simulation. Results shown are for 200 replicates per simulation scenario.

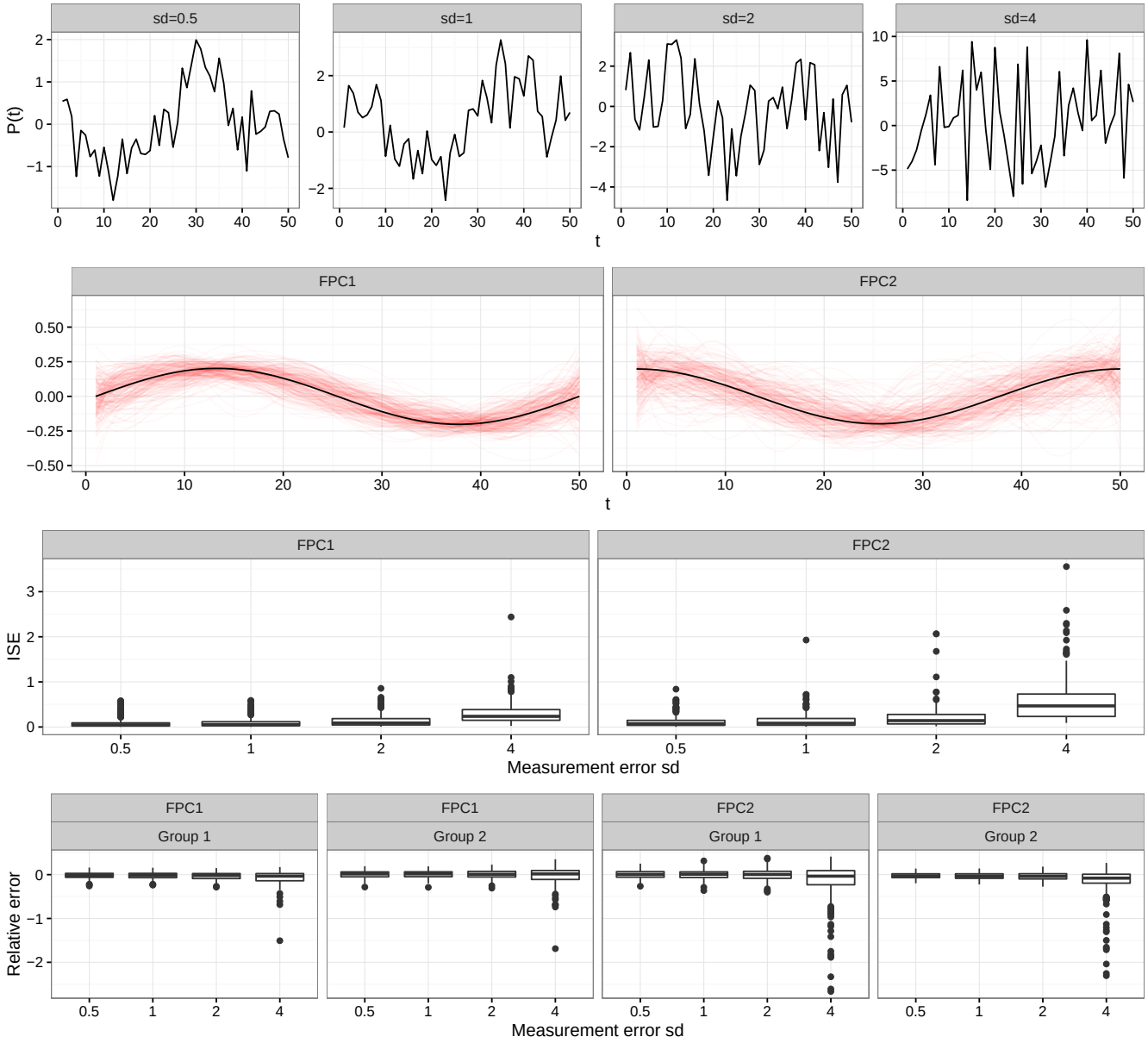

Figure A.11: Varying the measurement error. Using the cross-sectional simulation set up described in Section 5 with 80 curves per simulation, we varied the measurement error standard deviation to 0.5, 1, 2 and 4. The top panel illustrates one simulated curve for each setting of the measurement error standard deviation (note the different Y-axis scales). The second panel illustrates the estimated FPCs for standard deviation equal to 4. The FPC ISEs (panel 3) and relative error in estimation of the variance parameters (panel 4) illustrate that results are robust to a significant amount of noise, but estimation of parameters becomes poorer as the amount of noise increases. Results shown are for 200 replicates per simulation scenario. Four FPCs were simulated but only 2 were estimated.