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

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

A Additional results from analysis of kinematic data

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

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

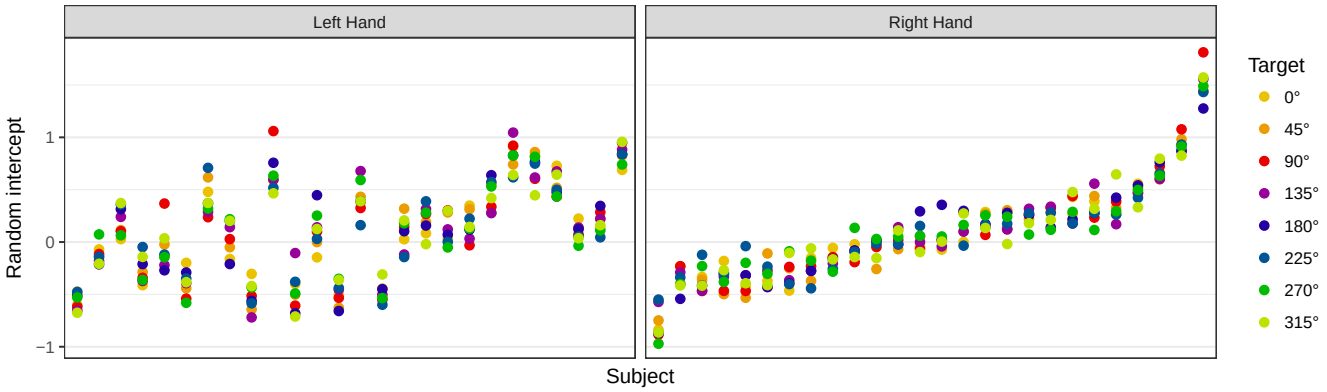

Figure A.1: Estimates of random intercepts. Each panel shows, for the left or the right hand, the estimated first principal component score variance random intercept parameters $g_{il1,int}$ in model (10) for each subject i and target l, for the X coordinate of motion. Targets are colored as in Figure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

B HMC and SE methods applied to kinematic data

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

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

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

C Bivariate model

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

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

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

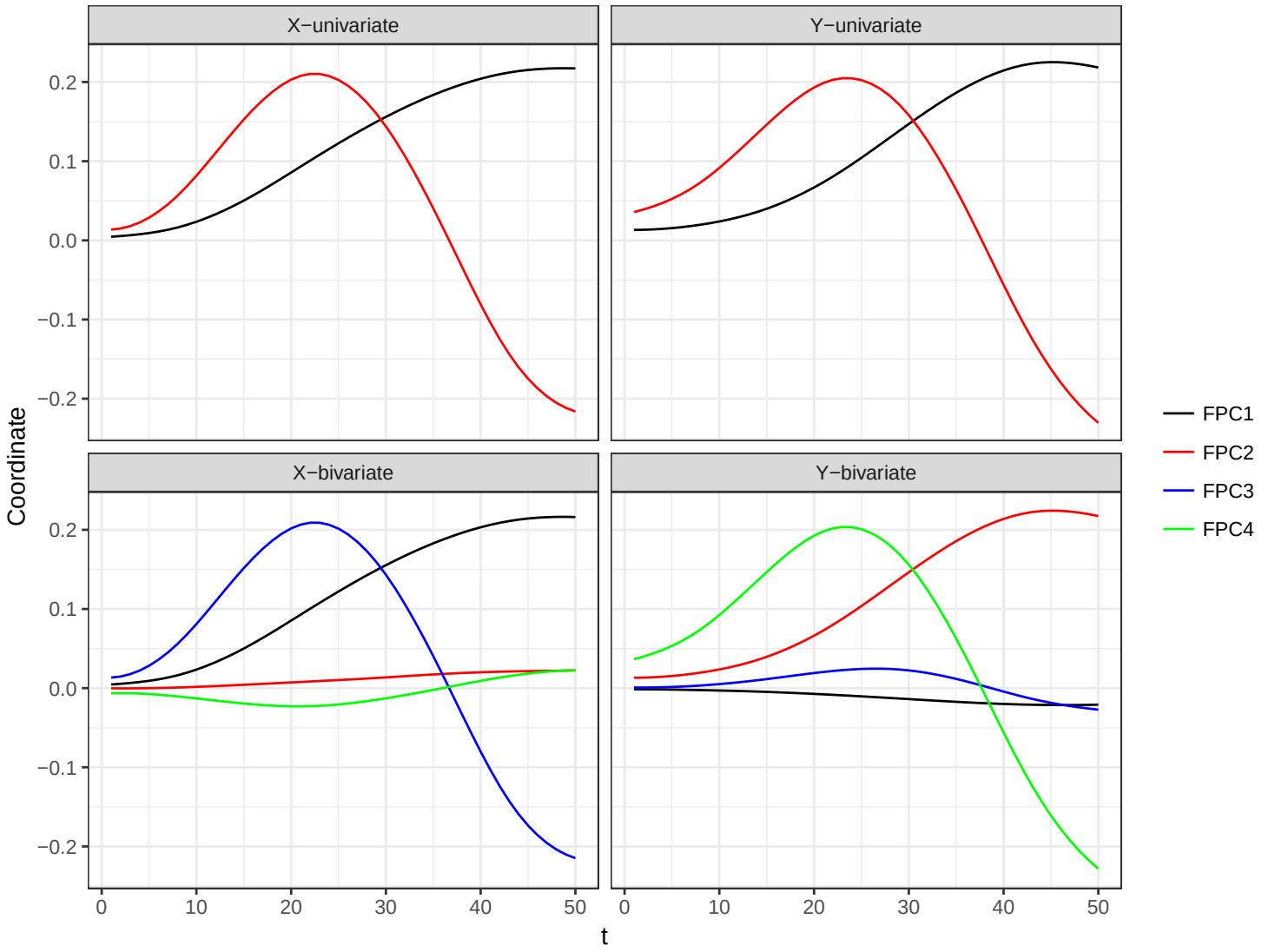

Figure A.2: FPCs from model (9) fit to the univariate and bivariate data. The FPCs on the left are for the X coordinates of motions, those on the right are for the Y-coordinate. The FPCs in the top row were estimated using univariate models, and the FPCs in the bottom row were estimated using bivariate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

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

greater involvement of both dimensions in FPCs 3 and 4.

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

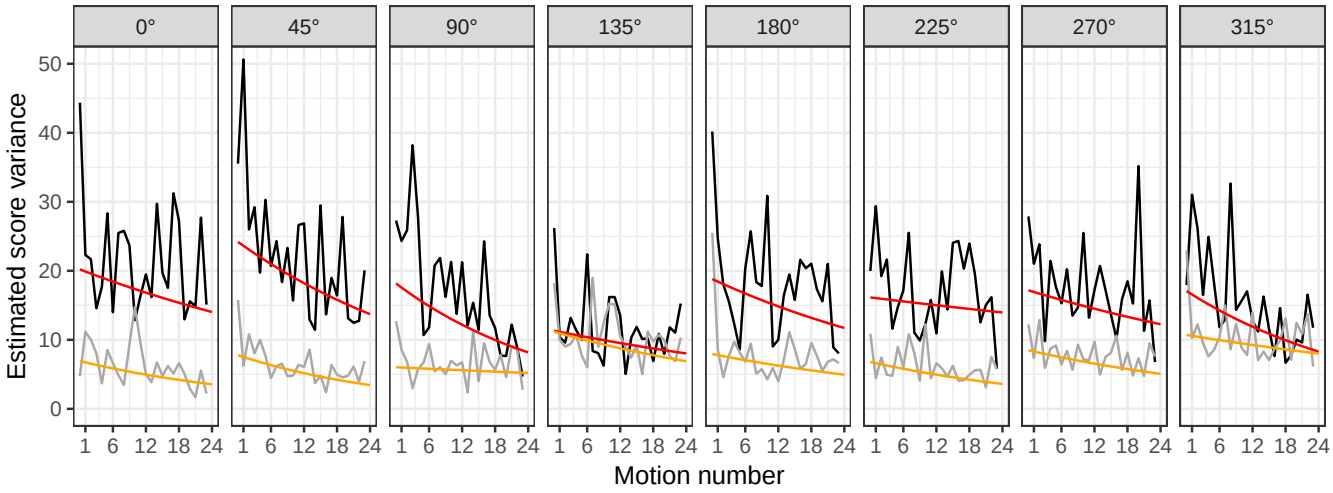

Figure A.3: Estimates of bivariate FPC score variances in the right hand for each target. Panels show the estimates of the score variance as a function of repetition number using the slope-intercept model (10) in red and orange (first and second FPC, respectively), and using the saturated one-parameter-per-repetition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

D Sensitivity Analyses

D.1 Hyperparameters

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

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

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

D.2 Mean Structure

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

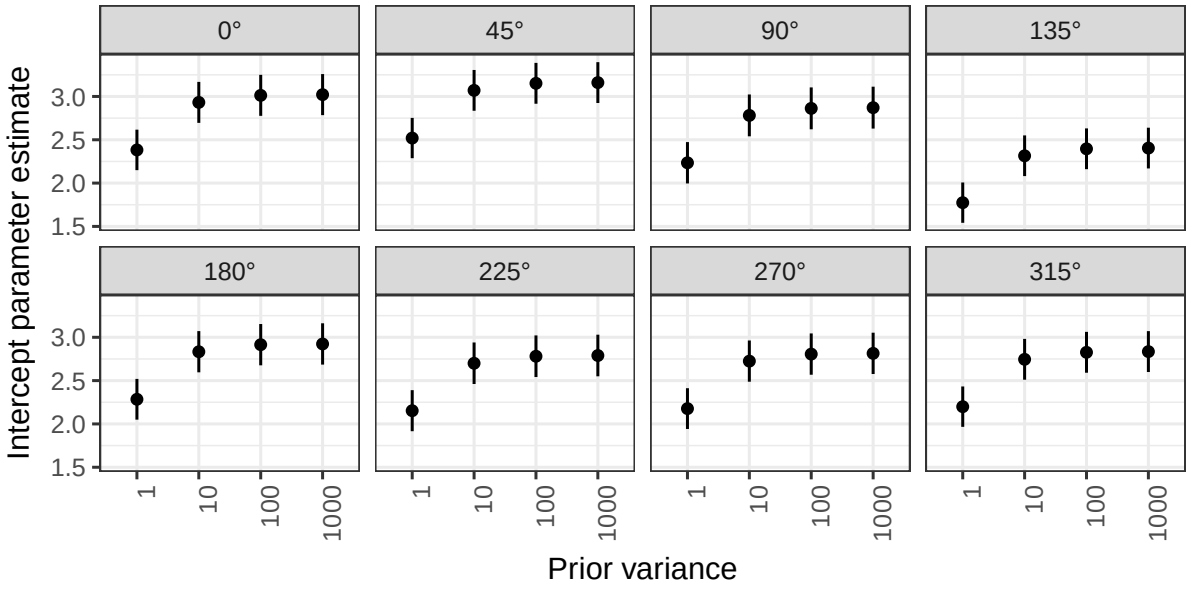

Figure A.4: Estimates and 95% credible intervals for $\gamma_{l1,int}$ as a function of the variance of its normal prior.

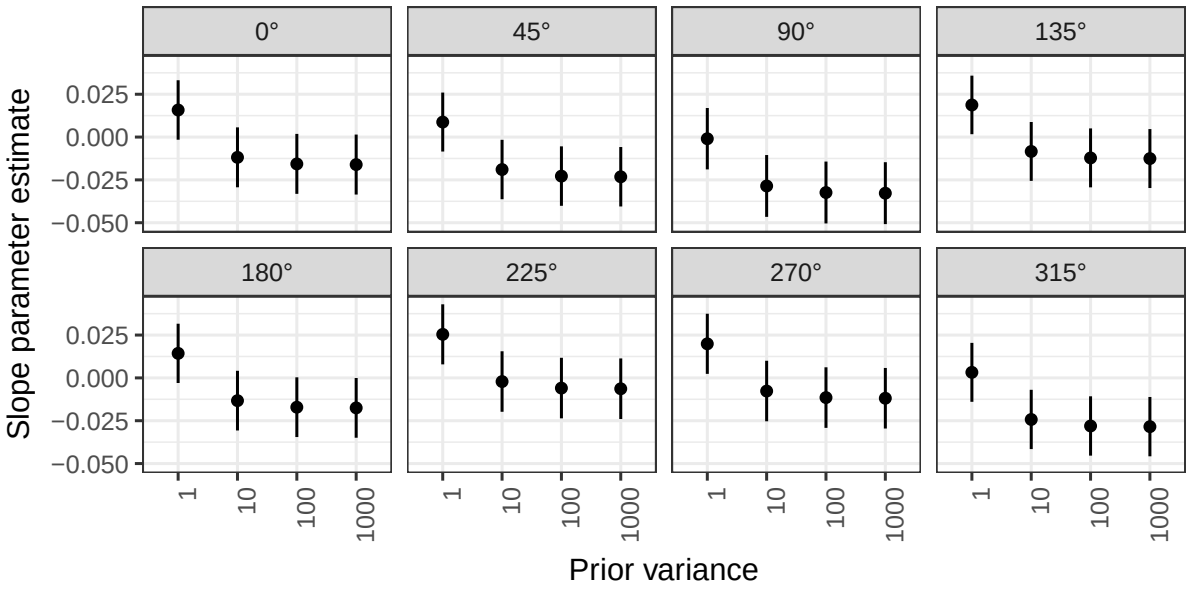

Figure A.5: Estimates and 95% credible intervals for $\gamma_{l1,slope}$ as a function of the variance of its normal prior.

curves with a functional intercept for each target and random functional effects for each subject-target combination. If the mean motion to a target systematically changed as a function of repetition number, then scores at the beginning or end of the training session might be inflated, which could lead to over- or under-estimation of our parameter of principal interest, the motion number score variance slopes $\gamma_{l1,slope}$. To examine this possibility, we conducted an analysis, restricted to data

for right hand motions to target 0°, in which we fit 4 separate functional random effects for each subject, for 4 groups of consecutive motions (motions 1 through 6, motions 7 through 12, et cetera). We found that inference for the slope parameter $\gamma_{11,slope}$ was unchanged, suggesting that model (9) is adequate.

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

E Derivations

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

E.1 Derivation of conditional distributions

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

We rewrite our model using matrix notation as follows:

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

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

elements of $\boldsymbol{c}$ on its main diagonal and 0 elsewhere and $\otimes$ is the Kronecker product operator. The posterior distribution of $\text{vec}\,(\boldsymbol{\beta})$ is then

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

Using the identity

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

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

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

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

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

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

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

that is, multivariate normal with covariance matrix

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

and mean

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

Letting $\boldsymbol{\sigma}_{\boldsymbol{\Phi}}^2$ be the length K vector of prior variances $\sigma_{\boldsymbol{\phi}_k}^2$ (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 $\boldsymbol{\Sigma}_{\boldsymbol{\Phi}} = \text{diag}\,(\boldsymbol{\sigma}_{\boldsymbol{\Phi}}^2) \otimes \boldsymbol{Q}^{-1}$. The posterior distribution of $\text{vec}\,(\boldsymbol{\Phi})$ is then

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

that is, multivariate normal with covariance matrix

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

and mean

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

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

distribution of $\boldsymbol{\xi}_{ij}$ is

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

that is, multivariate normal with covariance matrix

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

and mean

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

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

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

which has the form of the posterior of a gamma generalized linear model with log link, responses given by ξ^2_{ijk}, shape parameter equal to $1/2$ and a mean-zero multivariate normal prior on the

coefficients $\boldsymbol{\gamma}_k$ and $\boldsymbol{g}_{ik}, i = 1, \ldots, I$, with covariance matrix determined by $\boldsymbol{\Sigma}_{\gamma_k}$ and $\boldsymbol{\Sigma}_{\boldsymbol{g}_k}$.

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

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

For this variance parameter and also for the variance parameters $\sigma^2_{\boldsymbol{b}}$ and $\sigma^2_{\boldsymbol{\phi}_k}$, the conditional posterior distributions are the same in the model with bivariate observations, except that, for example, in the conditional posterior distribution of $\sigma^2_{\boldsymbol{\beta}^x_l}$, 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 $\sigma^2_{\boldsymbol{\beta}^y_l}$, the remaining components of $\boldsymbol{\beta}_l$ are used. The conditional distribution of $\sigma^2_{\boldsymbol{b}}$ is similar:

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

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

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

of σ^2:

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

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

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

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

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

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

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

E.2 Overview of variational Bayes

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

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

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

for all the other parameters.

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

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

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

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

E.3 Derivation of variational Bayes algorithm

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

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

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

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

and mean

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

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

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

and mean

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

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

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

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

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

and mean

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

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 $\mu^T_{q(\Phi)}\boldsymbol{\Theta}^T\boldsymbol{\Theta}\mu_{q(\Phi)} + \{M_{ij}\}$ 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 $\mu^T_{q(\Xi)}\mu_{q(\Xi)} + M$, 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 $\mu_{q(\xi^2_{ijk})}$ in place of ξ^2_{ijk}, 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 $\mu_{q\left(\Sigma^{-1}_{\xi_{ij}}\right)}$ in the expression for $\boldsymbol{\Sigma}_{q(\xi_{ij})}$ above is $\exp\left(-\boldsymbol{\mu}^T_{mode}\boldsymbol{x} - \frac{1}{2}\boldsymbol{x}^T\boldsymbol{\Sigma}_{mode}\boldsymbol{x}\right)$.

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

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

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

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

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

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

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

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

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

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

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

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

E.4 Details of implementation of HMC sampler

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

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

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

F Additional simulation results

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

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

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

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

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

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

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

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

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

fixed. As expected, more noise results in larger errors in estimation, of both the FPCs and the score variance parameters.

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

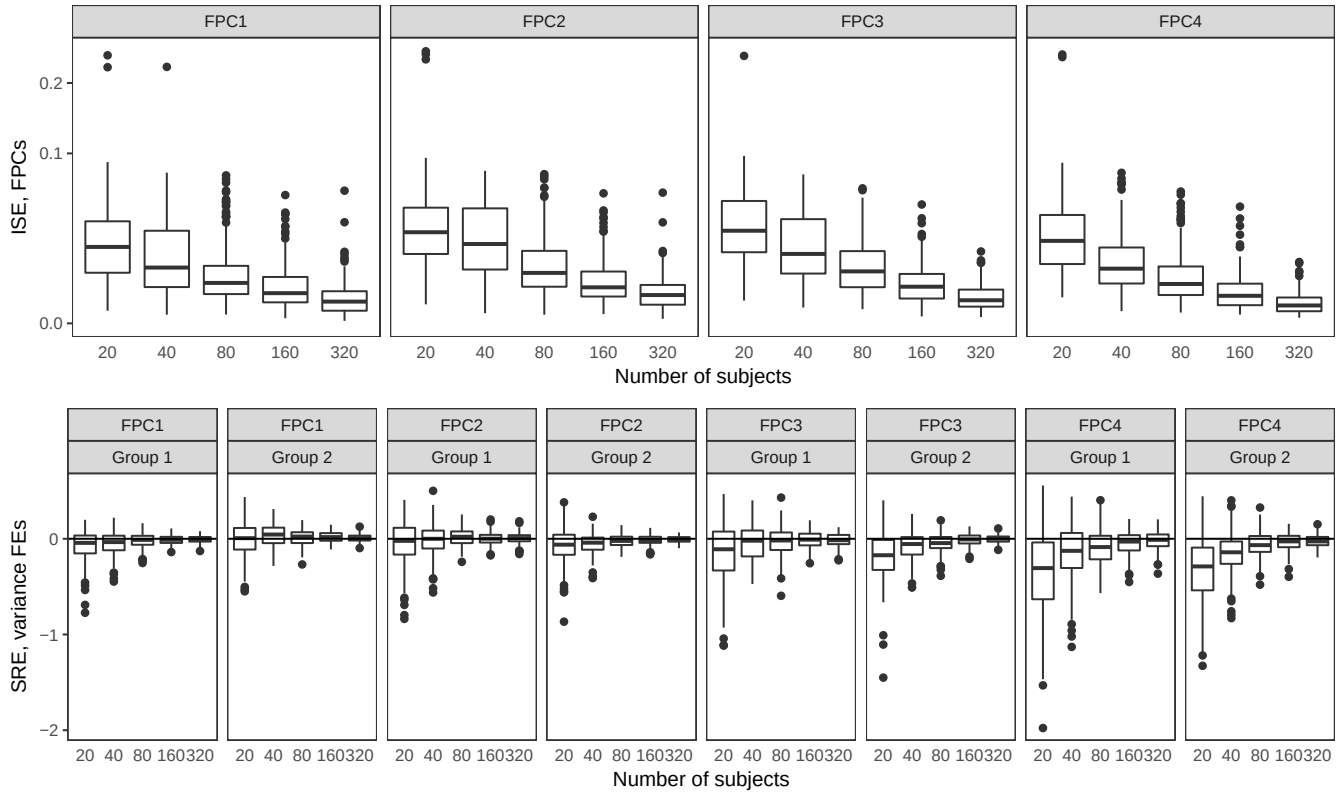

Figure A.6: Varying the number of curves. Integrated squared errors in estimation of FPCs (first row) and signed relative error in estimation of variance parameters (second row) decreases with more curves.

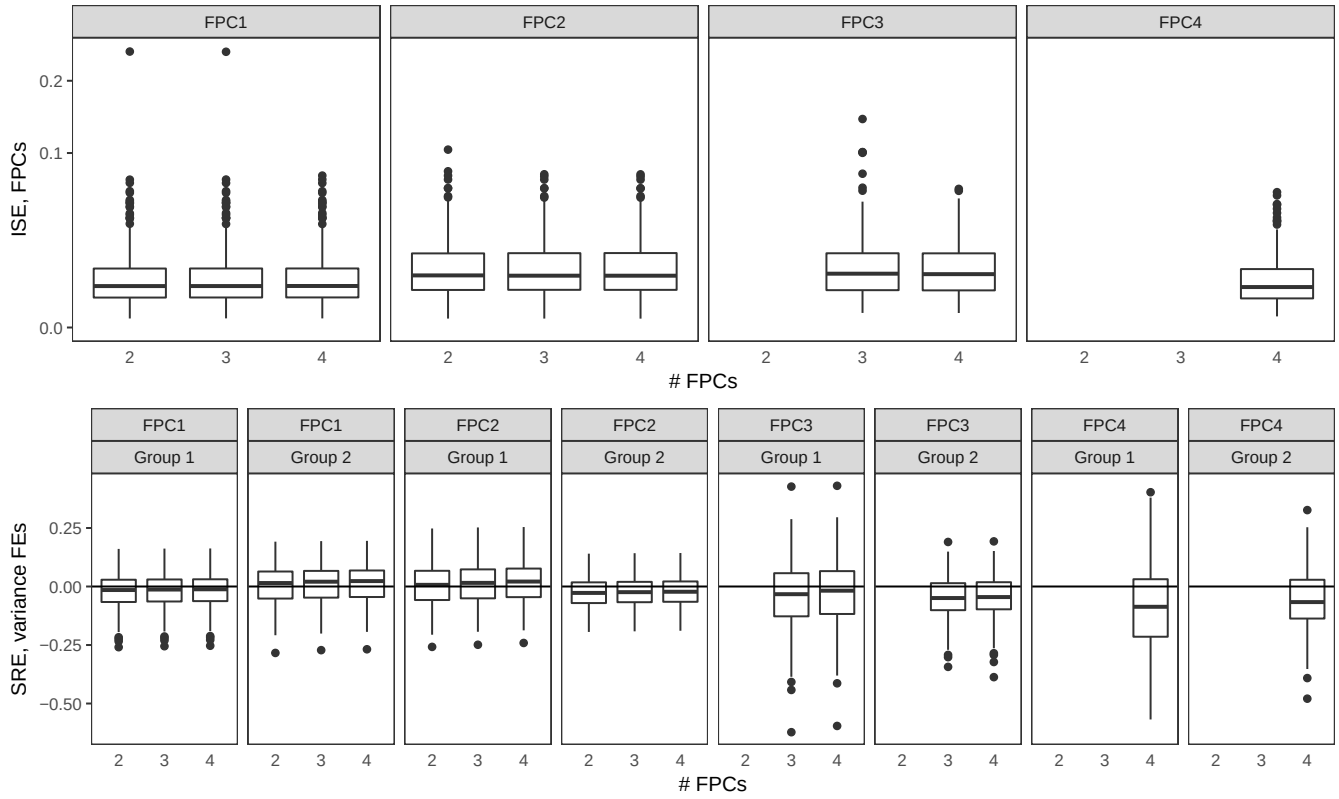

Figure A.7: Varying the number of estimated FPCs. Integrated squared errors in estimation of FPCs (first row) and signed relative error in estimation of variance parameters (second row) for FPCs 1 and 2 is mostly invariant to whether additional FPCs and associated score variance parameters are also estimated.

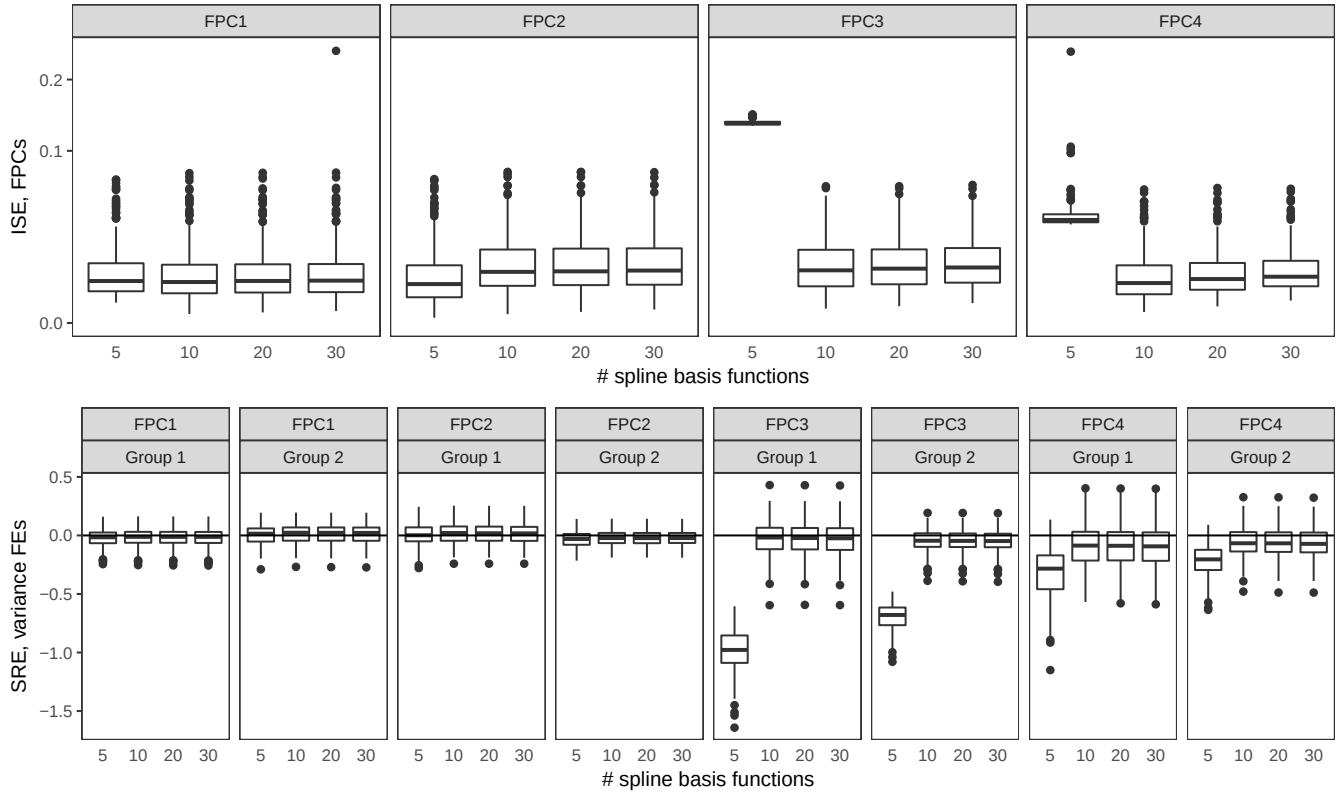

Figure A.8: Varying the number of spline basis functions. 5 spline basis functions are not sufficient to adequately capture the relatively fast variation in FPCs 3 and 4. Otherwise integrated squared errors in estimation of FPCs (first row) and signed relative error in estimation of variance parameters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

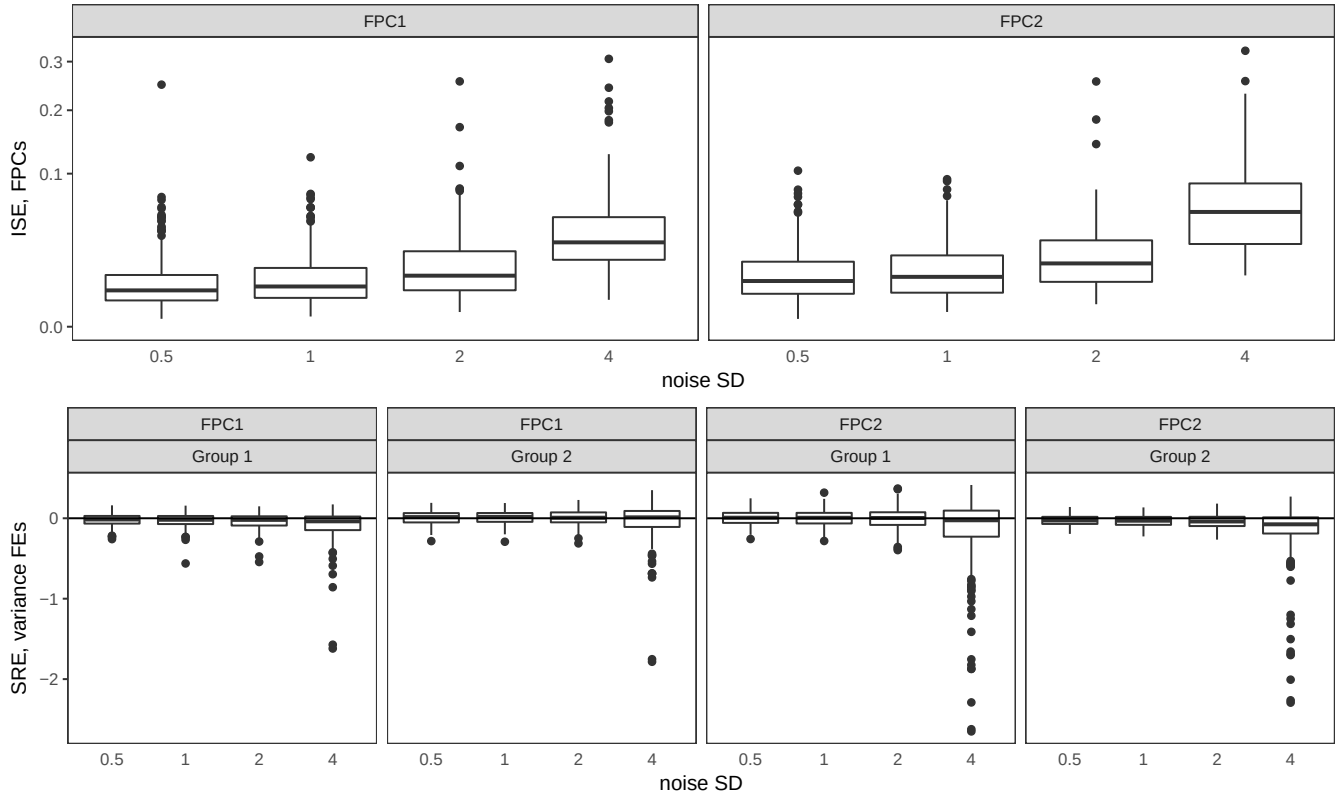

Figure A.9: Varying the measurement error. We varied the measurement error standard deviation to 0.5, 1, 2 and 4. FPC integrated squared errors (first row) and signed relative errors in estimation of the variance parameters (second row) illustrate that results are robust to a significant amount of noise,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

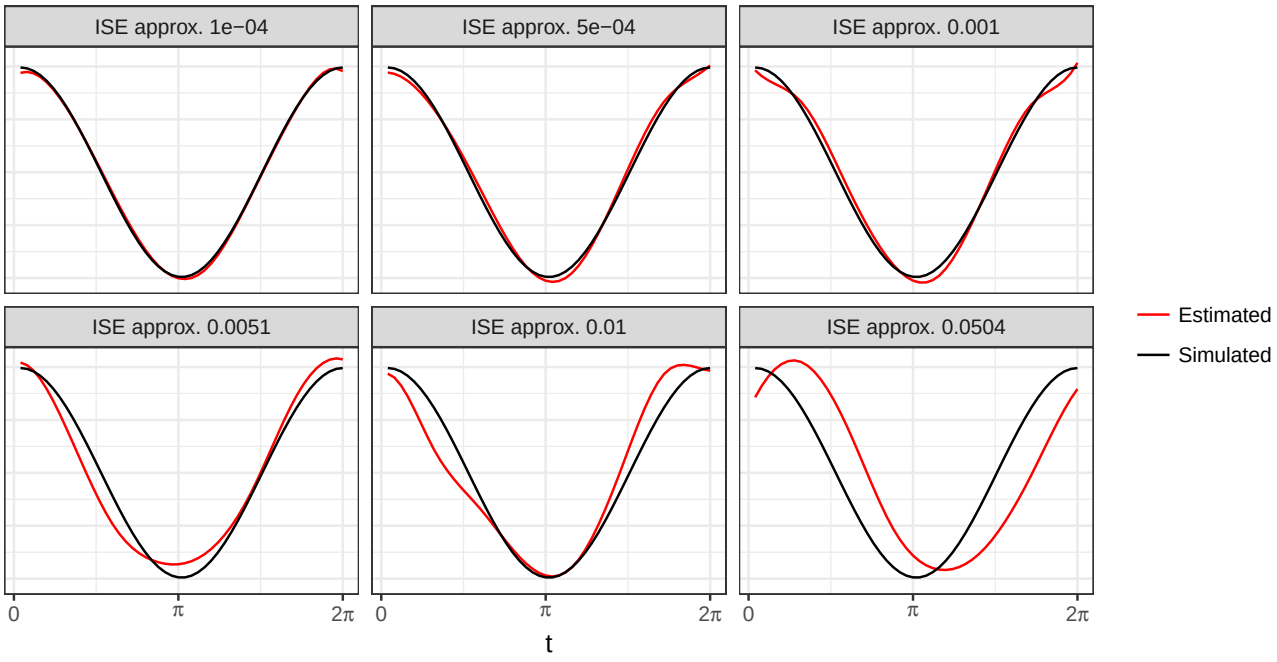

Figure A.10: Examples of estimates of FPC 2 with varying levels of integrated squared error. These estimates come from the longitudinal simulation scenario with $J_i = 4$.