Web-Based Supplementary Materials for: FPCA-based Method to Select Optimal Sampling Schedules that Capture Between-Subject Variability in Longitudinal Studies

by: Meihua Wu, Ana V. Diez-Roux, Trivelore E. Ragunathan, and Brisa N. Sánchez*

**email: brisa@umich.edu*

Web Appendix A: Estimation of FPCA

There is a rich literature on methods for estimating FPCA (Rice and Silverman, 1991; Yao et al., 2005; Hall et al., 2006; Peng, 2009; James et al., 2000; Silverman, 1996; Yao and Lee, 2006; Zhou et al., 2008; Hastings, 1970). We implement an approach that includes smoothing of the principal component functions and is applicable for sparse data, and that preserves orthogonality of the smoothed principal component functions so that the optimization criterion for the design stage is straightforward to compute. We use an EM approach because it can handle the sparsity of the longitudinal data well, and automatically computes the loadings on each principal component for every subject. In the following paragraphs, we describe our modification of the EM algorithm used by James et al. (2000) that incorporates the smoothing penalty proposed by Zhou et al. (2008).

For subject i, let the sampling times be $\boldsymbol{T}_i = (t_{i1}, , t_{in_i})$ and the observations be $\boldsymbol{Y}_i = (y_i(t_{i1}), \ldots, y_i(t_{in_i}))'$. We model the principal components by linear combinations of spline basis $\beta_k(t) = \sum_{l=1}^{q} b_l(t)\theta_{lk}$. Let $\boldsymbol{B}_i$ be a basis matrix such that $(\boldsymbol{B}_i)_{jl} = b_l(t_{ij}); \boldsymbol{\theta}_k = (\theta_{1k}, ..., \theta_{qk})'; \boldsymbol{\Theta}$ be the coefficient matrix such that $\boldsymbol{\Theta} = (\boldsymbol{\theta}_1, ..., \boldsymbol{\theta}_r)$ and finally $\boldsymbol{\alpha}_i = (\alpha_{i1}, ..., \alpha_{ir})'$. Then the reduced rank model can be written in the matrix form:

$$\boldsymbol{Y}_i = f(\boldsymbol{T}_i; \boldsymbol{\eta}) + \boldsymbol{\beta}(\boldsymbol{T}_i)\boldsymbol{\alpha}_i + \boldsymbol{\epsilon}_i$$

where $\boldsymbol{\beta}(\boldsymbol{T}_i) = [\beta_1(\boldsymbol{T}_i), ..., \beta_r(\boldsymbol{T}_i)] = \boldsymbol{B}_i\boldsymbol{\Theta}, \beta_k(\boldsymbol{T}_i) = (\beta_k(t_{i1}), ..., \beta_k(t_{in_i})), \boldsymbol{\alpha}_i \sim N(0, \boldsymbol{D})$, and $\boldsymbol{\epsilon}_i \sim N(0, \sigma^2 \boldsymbol{I}_{n_i})$. For the purpose of identification, we impose the orthonormal constraints: $\boldsymbol{\Theta}'\boldsymbol{\Theta} = \boldsymbol{I}, \int b_l(t)b_{l'}(t)dt = \delta_{ll'}$. Thus the marginal distribution of $\boldsymbol{Y}_i$ is:

$$\boldsymbol{Y}_i \sim N(f(\boldsymbol{T}_i; \boldsymbol{\eta}), \boldsymbol{A}_i) \qquad \boldsymbol{A}_i = \boldsymbol{\beta}(\boldsymbol{T}_i)\boldsymbol{D}\boldsymbol{\beta}'(\boldsymbol{T}_i) + \sigma^2 \boldsymbol{I}_{n_i}$$

and the observed log likelihood given the data is the sum of the individual's contributions to the log-likelihood, $l = \sum_{i=1}^{N} l_i$

$$l = \sum_{i=1}^{N} \frac{-n_i}{2} log(2\pi) - \frac{1}{2} log \mid \boldsymbol{A}_i \mid - \frac{1}{2}(\boldsymbol{Y}_i - f(\boldsymbol{T}_i, \boldsymbol{\eta}))' \boldsymbol{A}_i^{-1}(\boldsymbol{Y}_i - f(\boldsymbol{T}_i, \boldsymbol{\eta})). \qquad (1)$$

It is difficult to find the MLE for the observed likelihood given its complexity. Instead, we employ the EM algorithm and work with the complete data likelihood with $\boldsymbol{\alpha}_i$ assumed to be known:

$$\sum_{i=1}^{n} \frac{-n_i}{2} log(2\pi) - \frac{n_i}{2} log(\sigma^2) - \frac{1}{2} log \mid \boldsymbol{D} \mid - \frac{1}{2\sigma^2}(\boldsymbol{Y}_i - f(\boldsymbol{T}_i, \eta) - \boldsymbol{B}_i \boldsymbol{\Theta} \boldsymbol{\alpha}_i)'(\boldsymbol{Y}_i - f(\boldsymbol{T}_i, \boldsymbol{\eta}) - \boldsymbol{B}_i \boldsymbol{\Theta} \boldsymbol{\alpha}_i) - \frac{1}{2} \boldsymbol{\alpha}_i' \boldsymbol{D}^{-1} \boldsymbol{\alpha}_i.$$

We encourage the smoothness of $\beta_k(t)$ by introducing a second derivative penalty $\lambda \int \beta_k''(t)^2 dt$ with λ being a positive smoothing parameter. To write the penalty in terms of θ_k, we let $\boldsymbol{H}$ be a matrix such that the ll' element is $\boldsymbol{H}_{ll'} = \int b_l(t) b_{l'}(t) dt$. Thus, we have $\lambda \int \beta_k''(t)^2 dt = \lambda \boldsymbol{\theta}_k' \boldsymbol{H} \boldsymbol{\theta}_k$. We then obtain a smoothed version of FPCA by maximizing the penalized log likelihood

$$Q = \sum_{i=1}^{n} \frac{-n_i}{2} log(2\pi) - \frac{n_i}{2} log(\sigma^2) - \frac{1}{2} log \mid \boldsymbol{D} \mid - \frac{1}{2\sigma^2}(\boldsymbol{Y}_i - f(\boldsymbol{T}_i, \boldsymbol{\eta}) - \boldsymbol{B}_i \boldsymbol{\Theta} \boldsymbol{\alpha}_i)'(Y_i - f(\boldsymbol{T}_i, \boldsymbol{\eta}) -$$

$$\boldsymbol{B}_i \boldsymbol{\Theta} \alpha_i) - \frac{1}{2} \boldsymbol{\alpha}_i' \boldsymbol{D}^{-1} \boldsymbol{\alpha}_i - \sum_{k=1}^{r} \lambda \boldsymbol{\theta}_k' \boldsymbol{H} \boldsymbol{\theta}_k.$$

Green (1990) shows that the EM algorithm is also applicable to the penalized log likelihood when the penalty term does not involve latent variables. Hence we employ the EM algorithm for maximizing Q. The E-step and M-step of the EM algorithm are as follows. For the E-step, we denote $E(\boldsymbol{\alpha}_i | \boldsymbol{Y}_i, \boldsymbol{\eta}, \boldsymbol{\Theta}, \boldsymbol{D}, \sigma^2)$ by $\widehat{\boldsymbol{\alpha}}_i$ and $E(\boldsymbol{\alpha}_i \boldsymbol{\alpha}_i' | \boldsymbol{Y}_i, \boldsymbol{\eta}, \boldsymbol{\Theta}, \boldsymbol{D}, \sigma^2)$ by $\widehat{\boldsymbol{\alpha}_i \boldsymbol{\alpha}_i'}$. We have

$$\widehat{\boldsymbol{\alpha}}_i = (\sigma^2 \boldsymbol{D}^{-1} + \boldsymbol{\Theta}' \boldsymbol{B}_i' \boldsymbol{B}_i \boldsymbol{\Theta})^{-1} \boldsymbol{\Theta}' \boldsymbol{B}_i' (\boldsymbol{Y}_i - f(\boldsymbol{T}, \boldsymbol{\eta}))$$

$$\widehat{\boldsymbol{\alpha}_i \boldsymbol{\alpha}_i'} = \boldsymbol{\alpha}_i \boldsymbol{\alpha}_i' + (\boldsymbol{D}^{-1} + \boldsymbol{\Theta}' \boldsymbol{B}_i' \boldsymbol{B}_i \boldsymbol{\Theta} / \sigma^2)^{-1}.$$

For the M-step, we maximize Q iteratively over $\boldsymbol{\eta}, \boldsymbol{\Theta}, \boldsymbol{D}, \sigma^2$. The formula for $\boldsymbol{\eta}$, $\boldsymbol{D}$, and σ^2 are identical to those in (James et al., 2000) and they are omitted here. The interesting question is how to maximize Q while preserving the orthonormality of the columns of $\boldsymbol{\Theta}$. Our solution to this problem is a reparameterization based on singular value decomposition (SVD). In this setting, the optimization is carried out column-by-column of Θ. Suppose $\boldsymbol{\theta}_k$ is being considered and the rest of the columns, denoted by $\Theta_{(k)}$ are kept fixed. Then $\boldsymbol{\theta}_k$ should be orthogonal to the column space of $\Theta_{(k)}$. We perform a SVD on $\Theta_{(k)}$ and we have

$$\Theta_{(k)} = \boldsymbol{U} \cdot \boldsymbol{S} \cdot \boldsymbol{V}'$$

where $\boldsymbol{U}$ is $q \times q$ orthogonal matrix, $\boldsymbol{S}$ is diagonal matrix and $\boldsymbol{V}$ is $(r-1) \times (r-1)$ orthogonal matrix. Let $(\boldsymbol{u}_1, ..., \boldsymbol{u}_q)$ be the columns of $\boldsymbol{U}$. Then $(\boldsymbol{u}_r, ..., \boldsymbol{u}_q)$ spans the space orthogonal to the column space of $\Theta_{(k)}$. Then we parameterize $\boldsymbol{\theta}_k$ as

$$\boldsymbol{\theta}_k = \sum_{l=r}^{q} \boldsymbol{u}_l \boldsymbol{p}_l = \widetilde{\boldsymbol{U}}\boldsymbol{P}$$

where $\widetilde{\boldsymbol{U}} = (\boldsymbol{u}_r, ..., \boldsymbol{u}_q)$ and $\boldsymbol{P} = (\boldsymbol{p}_r, ..., \boldsymbol{p}_q)'$. Under this parameterization, $\boldsymbol{\theta}_k$ is always orthogonal to the column space of $\Theta_{(k)}$ and there is no restriction on $\boldsymbol{P}$. Now by focusing on the terms in $\boldsymbol{Q}$ that involve $\boldsymbol{\theta}_k$, we only need to minimize:

$$\sum_{i=1}^{N} (\boldsymbol{Y}_i - f(\boldsymbol{T}, \boldsymbol{\eta}) - \boldsymbol{B}_i \Theta \boldsymbol{\alpha}_i)'(\boldsymbol{Y}_i - f(\boldsymbol{T}, \boldsymbol{\eta}) - \boldsymbol{B}_i \Theta \boldsymbol{\alpha}_i) + \widetilde{\lambda} \boldsymbol{\theta}_k' \boldsymbol{H} \boldsymbol{\theta}_k$$

where $\widetilde{\lambda} = 2\sigma^2 \lambda$. We rewrite it as

$$\sum_{i=1}^{N} ((\boldsymbol{Y}_i - f(\boldsymbol{T}, \boldsymbol{\eta}) - \boldsymbol{B}_i \Theta_{(k)} \boldsymbol{\alpha}_{i(k)} - \boldsymbol{B}_i \boldsymbol{\theta}_k)'(\boldsymbol{Y}_i - f(\boldsymbol{T}, \boldsymbol{\eta}) - \boldsymbol{B}_i \Theta_{(k)} \boldsymbol{\alpha}_{i(k)} - \boldsymbol{B}_i \boldsymbol{\theta}_k) + \widetilde{\lambda} \boldsymbol{\theta}_k' \boldsymbol{H} \boldsymbol{\theta}_k$$

$$= \sum_{i=1}^{N} ((\boldsymbol{Y}_i - f(\boldsymbol{T}, \boldsymbol{\eta}) - \boldsymbol{B}_i \Theta_{(k)} \boldsymbol{\alpha}_{i(k)} - \boldsymbol{B}_i \tilde{\boldsymbol{U}} \boldsymbol{P})'(\boldsymbol{Y}_i - f(\boldsymbol{T}, \boldsymbol{\eta}) - \boldsymbol{B}_i \Theta_{(k)} \boldsymbol{\alpha}_{i(k)} - \boldsymbol{B}_i \tilde{\boldsymbol{U}} \boldsymbol{P}) + \widetilde{\lambda} \boldsymbol{P}' \tilde{\boldsymbol{U}}' \boldsymbol{H} \tilde{\boldsymbol{U}} \boldsymbol{P}$$

The closed form solution for $\boldsymbol{P}$ that minimizes the above equation is

$$\widehat{\boldsymbol{P}} = (\sum_{i=1}^{N} \widehat{\boldsymbol{\alpha}_{ik}^2} \widetilde{\boldsymbol{U}}' \boldsymbol{B}_i' \boldsymbol{B}_i \widetilde{\boldsymbol{U}} + \widetilde{\lambda} \widetilde{\boldsymbol{U}'} \boldsymbol{H} \widetilde{\boldsymbol{U}})^{-1} \sum_{i=1}^{N} \widetilde{\boldsymbol{U}'} \boldsymbol{B}_i' (\widehat{\boldsymbol{\alpha}_{ik}} (\boldsymbol{Y}_i - f(\boldsymbol{T}, \boldsymbol{\eta})) - \sum_{l \neq k} \widehat{\boldsymbol{\alpha}_{ik} \boldsymbol{\alpha}_{ik}'} \boldsymbol{B}_i \boldsymbol{\theta}_l).$$

Then we have $\widehat{\boldsymbol{\theta}}_k = \widetilde{\boldsymbol{U}} \widehat{\boldsymbol{P}}$ and we normalize $\widehat{\boldsymbol{\theta}}_k$ by dividing it by its norm. We repeat the same procedures for $k = 1, ..., r$ iteratively until convergence is reached. In summary, the reparameterization based on SVD allows us to optimize Q under the constraint that the columns of Θ are orthonormal. This procedure works when the number of columns, i.e., the number of principal components is strictly larger than the number of spline bases, which is generally true.

Web Appendix B: Selecting the Number of Components and the Smoothing Parameter

The cross validation score $s(r, \lambda)$ is computed by k-fold cross validation. We randomly and evenly split the preliminary data in the simulation into k groups. We treat one group as testing data set and the rest $k-1$ groups as the training data set. We use training data to estimate an FPCA model for the given r and λ. Then we compute the log likelihood of the testing data given the estimated model. We repeat the process k times so that every group becomes the testing data exactly once. Then $s(r, \lambda)$ is computed as the average of log likelihoods across the k testing sets. Because of the normal assumption, the $s(r, \lambda)$ is equivalent to negative of the mean square error, less some constant. In general, the higher the $s(r, \lambda)$, the better the model fit. In the simulations and other two real data applications, we specify $k = 10$.

In the simulation, we consider λ ranges from 10 to 2000. The appropriate λ for a FPCA model with r components is chosen as $\lambda_r = argmax_s(r, \lambda)$. Figure 1 of this supplementary material plots $s(r, \lambda)$ vs. λ for $r = 1, 2, 3$ for a sample of the simulated dataset from simulation scenarios A.1 and A.2. The highest $s(r, \lambda)$ in each case is marked by a triangle and the corresponding λ is chosen as λ_r^*.

Let $s(r) = s(r, \lambda_r)$, i.e. the CV score for a model with r principal components and the appropriate smoothing parameter λ_r . Because the model with $r+1$ components is more flexible than the model with r components, $s(r) = s(r, \lambda_r)$ almost always increases as r increases. Therefore maximizing the CV score $s(r)$ does not always lead to a parsimonious model in practice. In this case, we use a “scree plot” as a tool to visually select the appropriate number of components (Johnson and Wichern, 2007). In a scree plot, the CV score $s(r)$ is plotted against r and we are interested in the elbow point r^* where the improvement of $s(r)$ after r^* is relatively much smaller than those before r^*. In other words, the model fit will not significantly improve if we already have at least r^* principal components in the model. Figure 2a and 2b of this supplement present scree plots for data sets simulated for simulation scenarios A.1 and A.2. In general, $s(r)$ increases substantially from $r = 1$ to $r = 2$ but there is virtually no improvement in $s(r)$ from $r = 2$ to $r = 3$. So $r = 2$ principal components are sufficient for the data in the simulation replicates shown in Figure 2a and 2b.

While the scree plot is simple to understand, it bears some subjective influence from the analyst and cannot be implemented in a simulation scenario where analyst intervention is absent. Inspired by the scree plot, we use an objective rule based approach. We set a threshold b for negligible improvement and select the appropriate r for the data as r^* as the smallest r such that the improvement in CV score by adding one more component is less than the threshold $b\%$ for negligible improvement. The rule based approach can be clearly defined and carried out in the simulation without outside intervention.

To specify the threshold of negligible improvement, we could consult with the investigators or refer to the scree plot. For example, in Figure 2a and 2b, we notice from the scree plots that improvement in CV score is generally larger than 1% from $r = 1$ to $r = 2$ and less than 1% from $r = 2$ to $r = 3$. Therefore we set the threshold to be $b\% = 1\%$ for negligible improvement in the simulation.

Web Appendix C: Derivation of Information Matrix

We rely on formulas for matrix-based derivatives published by Harville (2001) to obtain the information matrix, and use the form of the likelihood given in (1) in Web Appendix A. Let η_l

and η_k be elements of the mean parameter vector $\boldsymbol{\eta}$. For the i^{th} likelihood contribution, we have:

$$\begin{aligned}\frac{\partial l_i}{\partial \eta_k} &= -\frac{1}{2}\frac{\partial}{\partial \eta_k}\{(\boldsymbol{Y}_i - \boldsymbol{\mu}_i)'\boldsymbol{A}_i^{-1}(\boldsymbol{Y}_i - \boldsymbol{\mu}_i)\} \\ &= -\frac{1}{2}\{(\boldsymbol{Y}_i - \boldsymbol{\mu}_i)'\boldsymbol{A}_i^{-1}\frac{\partial(\boldsymbol{Y}_i - \boldsymbol{\mu}_i)'}{\partial \eta_k} + \frac{\partial(\boldsymbol{Y}_i - \boldsymbol{\mu}_i)'}{\partial \eta_k}\boldsymbol{A}_i^{-1}(\boldsymbol{Y}_i - \boldsymbol{\mu}_i)\} \\ &= (\boldsymbol{Y}_i - \boldsymbol{\mu}_i)'\boldsymbol{A}_i^{-1}\frac{\partial \boldsymbol{\mu}_i}{\partial \eta_k}.\end{aligned}$$

The first step holds by the chain rule; the last holds since $\boldsymbol{A}_i^{-1}$ is symmetric, and

$$\frac{\partial(\boldsymbol{Y}_i - \boldsymbol{\mu}_i)}{\partial \eta_k} = -\frac{\partial \boldsymbol{\mu}_i}{\partial \eta_k}.$$

Next we have

$$\frac{\partial l_i}{\partial \eta_l \partial \eta_k} = (\boldsymbol{Y}_i - \boldsymbol{\mu}_i)'\boldsymbol{A}_i^{-1}\frac{\partial \boldsymbol{\mu}_i}{\partial \eta_l \partial \eta_l} - \frac{\partial \boldsymbol{\mu}_i'}{\partial \eta_l}\boldsymbol{A}_i^{-1}\frac{\partial \boldsymbol{\mu}_i}{\partial \eta_k}$$

by applying the chain rule. Since $E(\boldsymbol{Y}_i - \boldsymbol{\mu}_i) = 0$ when the mean model is correct, then

$$E(\frac{\partial l_i}{\partial \eta_l \partial \eta_k}) = -\frac{\partial \boldsymbol{\mu}_i'}{\partial \eta_l}\boldsymbol{A}_i^{-1}\frac{\partial \boldsymbol{\mu}_i}{\partial \eta_k}.$$

When $\boldsymbol{T}_i = \boldsymbol{T}$, then $\boldsymbol{\mu}_i = \boldsymbol{\mu}$, $\boldsymbol{A}_i = \boldsymbol{A}$, and thus

$$E(\frac{\partial l}{\partial \eta_l \partial \eta_k}) = \sum_{i=1}^{N} E(\frac{\partial l_i}{\partial \eta_l \partial \eta_k}) = -N[\frac{\partial \boldsymbol{\mu}'}{\partial \eta_l}\boldsymbol{A}^{-1}\frac{\partial \boldsymbol{\mu}}{\partial \eta_k}]. \tag{2}$$

We now take derivatives with respect to the variance components. Let d_k and d_l denote the k^{th} and l^{th} diagonal element of the diagonal matrix $\boldsymbol{D}$. We have

$$\begin{aligned}\frac{\partial l_i}{\partial d_k} &= -\frac{1}{2}\frac{\partial}{\partial d_k}log|\boldsymbol{A}_i| - \frac{1}{2}\frac{\partial}{\partial d_k}(\boldsymbol{Y}_i - \boldsymbol{\mu}_i)'\boldsymbol{A}_i^{-1}(\boldsymbol{Y}_i - \boldsymbol{\mu}_i) \\ &= -\frac{1}{2}tr\{\boldsymbol{A}_i^{-1}\frac{\partial \boldsymbol{A}_i}{\partial d_k} - \frac{1}{2}tr\{(\boldsymbol{Y}_i - \boldsymbol{\mu}_i)(\boldsymbol{Y}_i - \boldsymbol{\mu}_i)'\frac{\partial}{\partial d_k}\boldsymbol{A}_i^{-1}\} \\ &= -\frac{1}{2}tr\{\boldsymbol{A}_i^{-1}\beta_k(\boldsymbol{T}_i)\beta_k'(\boldsymbol{T}_i)\} + \frac{1}{2}tr\{(\boldsymbol{Y}_i - \boldsymbol{\mu}_i)(\boldsymbol{Y}_i - \boldsymbol{\mu}_i)'\boldsymbol{A}_i^{-1}\beta_k(\boldsymbol{T}_i)\beta_k(\boldsymbol{T}_i)\boldsymbol{A}_i^{-1}\}.\end{aligned}$$

The second equality holds given the formulas for the derivative of the log of the determinant of a matrix (Harville 2001), and because a quadratic term equals its trace and the derivative of a trace is equal to the trace of the derivative. The last one holds because

$$\frac{\partial \boldsymbol{A}_i}{\partial d_k} = \beta_k(\boldsymbol{T}_i)\beta_k'(\boldsymbol{T}_i) \quad \text{and because} \quad \frac{\partial \boldsymbol{A}_i^{-1}}{\partial d_k} = -\boldsymbol{A}_i^{-1}\frac{\partial \boldsymbol{A}_i}{\partial d_k}\boldsymbol{A}_i^{-1}.$$

Next we have

$$\begin{aligned}\frac{\partial l_i}{\partial d_l \partial d_k} &= -\frac{1}{2}tr\{\frac{\partial}{\partial d_l}\boldsymbol{A}_i^{-1}\beta_k(\boldsymbol{T}_i)\beta_k'(\boldsymbol{T}_i)\}\\ &+ \frac{1}{2}tr\{(\boldsymbol{Y}_i - \boldsymbol{\mu}_i)(\boldsymbol{Y}_i - \boldsymbol{\mu}_i)'\frac{\partial}{\partial d_l}\boldsymbol{A}_i^{-1}\beta_k(\boldsymbol{T}_i)\beta_k(\boldsymbol{T}_i)\boldsymbol{A}_i^{-1}\}\\ &+ \frac{1}{2}tr\{(\boldsymbol{Y}_i - \boldsymbol{\mu}_i)(\boldsymbol{Y}_i - \boldsymbol{\mu}_i)'\boldsymbol{A}_i^{-1}\beta_k(\boldsymbol{T}_i)\beta_k(\boldsymbol{T}_i)\frac{\partial}{\partial d_l}\boldsymbol{A}_i^{-1}\}.\end{aligned}$$

Since $E[(\boldsymbol{Y}_i - \boldsymbol{\mu}_i)(\boldsymbol{Y}_i - \boldsymbol{\mu}_i)'] = \boldsymbol{A}_i$, we have

$$\begin{aligned}E[\frac{\partial l_i}{\partial d_l \partial d_k}] &= -\frac{1}{2}tr\{\frac{\partial}{\partial d_l}\boldsymbol{A}_i^{-1}\beta_k(\boldsymbol{T}_i)\beta_k'(\boldsymbol{T}_i)\}\\ &+ \frac{1}{2}tr\{\boldsymbol{A}_i\frac{\partial}{\partial d_l}\boldsymbol{A}_i^{-1}\beta_k(\boldsymbol{T}_i)\beta_k(\boldsymbol{T}_i)\boldsymbol{A}_i^{-1}\}\\ &+ \frac{1}{2}tr\{\beta_k(\boldsymbol{T}_i)\beta_k(\boldsymbol{T}_i)\frac{\partial}{\partial d_l}\boldsymbol{A}_i^{-1}\}.\end{aligned}$$

The first and third term cancel out, and simplifying the second term using rules for the trace of a product, we have

$$\begin{aligned}E[\frac{\partial l_i}{\partial d_l \partial d_k}] &= \frac{1}{2}tr\{\beta_k'(\boldsymbol{T}_i)\frac{\partial}{\partial d_l}\boldsymbol{A}_i^{-1}\beta_k(\boldsymbol{T}_i)\}\\ &= \frac{1}{2}\{\beta_k'(\boldsymbol{T}_i)\boldsymbol{A}_i^{-1}\frac{\partial}{\partial d_l}\boldsymbol{A}_i\boldsymbol{A}_i^{-1}\beta_k(\boldsymbol{T}_i)\}\\ &= \frac{1}{2}\{\beta_k'(\boldsymbol{T}_i)\boldsymbol{A}_i^{-1}\beta_l(\boldsymbol{T}_i)\beta_l'(\boldsymbol{T}_i)\boldsymbol{A}_i^{-1}\beta_k(\boldsymbol{T}_i)\}\\ &= -\frac{1}{2}\{\beta_k'(\boldsymbol{T}_i)\boldsymbol{A}_i^{-1}\beta_l(\boldsymbol{T}_i)\}^2.\end{aligned}$$

When $\boldsymbol{T}_i = \boldsymbol{T}$, then we have $\boldsymbol{A}_i = \boldsymbol{A}$ and therefore

$$E[\frac{\partial l}{\partial d_l \partial d_k}] = -\frac{N}{2}\{\beta_k'(\boldsymbol{T})\boldsymbol{A}^{-1}\beta_l(\boldsymbol{T})\}^2. \tag{3}$$

For σ^2 we have:

$$\frac{\partial l_i}{\partial \sigma^2} = -\frac{1}{2}tr\{\boldsymbol{A}_i^{-1}\frac{\partial \boldsymbol{A}_i}{\partial \sigma^2}\} - \frac{1}{2}tr\{(\boldsymbol{Y}_i - \boldsymbol{\mu}_i)(\boldsymbol{Y}_i - \boldsymbol{\mu}_i)'\frac{\partial \boldsymbol{A}_i^{-1}}{\partial \sigma^2}\}.$$

Since $\frac{\partial \boldsymbol{A}_i}{\partial \sigma^2} = I_{n_i}$ (an $n_i \times n_i$ identity matrix) and $\frac{\partial \boldsymbol{A}_i^{-1}}{\partial \sigma^2} = \boldsymbol{A}_i^{-1} \frac{\partial \boldsymbol{A}_i}{\partial \sigma^2} \boldsymbol{A}_i^{-1} = -\boldsymbol{A}_i^{-1} \boldsymbol{A}_i^{-1}$

$$\frac{\partial l_i}{\partial \sigma^2} = -\frac{1}{2} tr\{\boldsymbol{A}_i^{-1}\} + \frac{1}{2} tr\{(\boldsymbol{Y}_i - \boldsymbol{\mu}_i)(\boldsymbol{Y}_i - \boldsymbol{\mu}_i)' \boldsymbol{A}_i^{-1} \boldsymbol{A}_i^{-1}\}.$$

Then

$$\begin{aligned} \frac{\partial l_i}{\partial \sigma^2 \partial \sigma^2} &= \frac{1}{2} tr\{\boldsymbol{A}_i^{-1} \frac{\partial \boldsymbol{A}_i}{\partial \sigma^2} \boldsymbol{A}_i^{-1}\} \\ &- \frac{1}{2} tr\{(\boldsymbol{Y}_i - \boldsymbol{\mu}_i)(\boldsymbol{Y}_i - \boldsymbol{\mu}_i)' \boldsymbol{A}_i^{-1} \frac{\partial \boldsymbol{A}_i^{-1}}{\partial \sigma^2} \boldsymbol{A}_i^{-1} \boldsymbol{A}_i^{-1}\} \\ &- \frac{1}{2} tr\{(\boldsymbol{Y}_i - \boldsymbol{\mu}_i)(\boldsymbol{Y}_i - \boldsymbol{\mu}_i)' \boldsymbol{A}_i^{-1} \boldsymbol{A}_i^{-1} \frac{\partial \boldsymbol{A}_i^{-1}}{\partial \sigma^2} \boldsymbol{A}_i^{-1}\}. \end{aligned}$$

Since $E[(\boldsymbol{Y}_i - \boldsymbol{\mu}_i)(\boldsymbol{Y}_i - \boldsymbol{\mu}_i)'] = \boldsymbol{A}_i$ and $\frac{\partial \boldsymbol{A}_i}{\partial \sigma^2} = I_{ni}$ we have $E[\frac{\partial l_i}{\partial \sigma^2 \partial \sigma^2}] = -tr\{\boldsymbol{A}_i^{-1} \boldsymbol{A}_i^{-1}\}$. Thus, when $\boldsymbol{T}_i = \boldsymbol{T}$

$$E[\frac{\partial l}{\partial \sigma^2 \partial \sigma^2}] = -N \cdot tr\{\boldsymbol{A}^{-1} \boldsymbol{A}^{-1}\}. \tag{4}$$

Finally,

$$\begin{aligned} \frac{\partial l_i}{\partial d_k \partial \sigma^2} &= \frac{1}{2} tr\{\boldsymbol{A}_i^{-1} \beta_k(\boldsymbol{T}_i) \beta_k'(\boldsymbol{T}_i) \boldsymbol{A}_i^{-1}\} \\ &- \frac{1}{2} tr\{(\boldsymbol{Y}_i - \boldsymbol{\mu}_i)(\boldsymbol{Y}_i - \boldsymbol{\mu}_i)' \boldsymbol{A}_i^{-1} \beta_k(\boldsymbol{T}_i) \beta_k'(\boldsymbol{T}_i) \boldsymbol{A}_i^{-1} \boldsymbol{A}_i^{-1}\} \\ &- \frac{1}{2} tr\{(\boldsymbol{Y}_i - \boldsymbol{\mu}_i)(\boldsymbol{Y}_i - \boldsymbol{\mu}_i)' \boldsymbol{A}_i^{-1} \boldsymbol{A}_i^{-1} \beta_k(\boldsymbol{T}_i) \beta_k'(\boldsymbol{T}_i) \boldsymbol{A}_i^{-1}\}. \end{aligned}$$

Hence, $E[\frac{\partial l_i}{\partial d_k \partial \sigma^2}] = \frac{1}{2} \beta_k'(\boldsymbol{T}_i) \boldsymbol{A}_i^{-1} \boldsymbol{A}_i^{-1} \beta_k(\boldsymbol{T}_i)$. And, when $\boldsymbol{T}_i = \boldsymbol{T}$,

$$E[\frac{\partial l}{\partial d_k \partial \sigma^2}] = \frac{N}{2} \beta_k'(\boldsymbol{T}) \boldsymbol{A}^{-1} \boldsymbol{A}^{-1} \beta_k(\boldsymbol{T}). \tag{5}$$

As far as the interdependence of the mean and variance parameters, we have

$$\frac{\partial l_i}{\partial d_l \partial \eta_k} = (\boldsymbol{Y}_i - \boldsymbol{\mu}_i) \frac{\partial \boldsymbol{A}_i^{-1}}{\partial d_l} \frac{\partial \boldsymbol{\mu}_i}{\partial \eta_k},$$

which has expected value

$$E(\frac{\partial l_i}{\partial d_l \partial \eta_k}) = 0, \tag{6}$$

since $E(\boldsymbol{Y}_i - \boldsymbol{\mu}_i) = 0$ and the last two factors are not random variables. The same argument holds for $E(\frac{\partial l_i}{\partial \sigma^2 \partial \eta_k})$. Hence, the information matrix is block diagonal. The first block is the information matrix for η, i.e. $I(\boldsymbol{T};\boldsymbol{\eta})$, with elements in (2), and the second deals with variance components, with elements given by (3), (4) and (5).

Web Appendix D: Algorithm for Identifying the Optimal Schedule

Let $g(\boldsymbol{T}) = g(t_1, , t_n) > 0$ be the objective function with $\boldsymbol{T} = (t_1, ..., t_n)$ being the sampling schedule. The maximization is with respect to the sampling schedule T and the values of $t_i, i = 1, , n$ are taken (without replacement) from the set of feasible sampling times S. If S contains n_S elements, there are $\binom{n_S}{n}$ sampling schedules in the pool of all candidate schedules S_c, so this number can be extremely large even if n_S and n are only of moderate size, which makes almost impossible to enumerate all candidate sampling schedules for the purpose of maximization, particularly in a simulation setting. Therefore, in the following we describe a more efficient maximization algorithm based on the Metropolis-Hastings algorithm (Metropolis et al. 1953; Hastings, 1970)

Algorithm: Select an initial sampling schedule $\boldsymbol{T}_0$ from the pool of candidate schedules S_c. Also set up two storage arrays x_f and x_T of length M where M is the total number iterations to be run. Let $\boldsymbol{T}_{k-1}$ denote the sampling schedules selected at the $k-1$ iteration.

During the kth iteration:

(a) Let $\boldsymbol{G}_{k-1}$, denote the set of feasible sampling times not included in the sampling schedule $\boldsymbol{T}_{k-1}$ (i.e. the complement of $\boldsymbol{T}_{k-1}$ with respect to S).

(b) Randomly select one sampling time from $\boldsymbol{T}_{k-1}$ and replace it with a sampling time randomly selected from $\boldsymbol{G}_{k-1}$ to create a new sampling schedule $\boldsymbol{T}_{temp}$.

(c) Compute $g(\boldsymbol{T}_{k-1})$ and $g(\boldsymbol{T}_{temp})$.

(d) If $g(\boldsymbol{T}_{temp}) > g(\boldsymbol{T}_{k-1})$ then we let $\boldsymbol{T}_k = \boldsymbol{T}_{temp}$.

(e) Otherwise, we generate a random number $q \sim Binomial(\alpha)$. If $q = 1$, then $\boldsymbol{T}_k = \boldsymbol{T}_{temp}$, otherwise $\boldsymbol{T}_k = \boldsymbol{T}_{k-1}$.

(f) Store the sampling schedule $\boldsymbol{T}_k$ in $x_T[k]$ and objective value $g(\boldsymbol{T}_k)$ in $x_f[k]$. Repeat the iteration for M times. Then we identify the maximum objective values from x_f and corresponding sampling schedule from x_T

In this algorithm, we are treating the $g(\boldsymbol{T})$ as the probability function (less a constant) of a multivariate distribution of $t_1, ..., t_n$. The distribution can be simulated with the Metropolis-Hastings algorithm and a uniform proposal distribution. Since the mode of the distribution is identical to the maximum of the objective function $g(\cdot)$ less a constant, we are guaranteed to reach the maximum if the Metropolis-Hastings algorithm has run long enough to converge. So the optimal objective function and the corresponding optimal schedule will appear in x_f and x_T with probability 1.

Web Appendix E: Estimating Parametric Mixed Model with the R package nlme.

The nlme package is used to estimate linear and nonlinear mixed effect models. The code to estimate the model: $y_{ij} = \eta_{0i} + \eta_{1i}t_j + \eta_2 t_j \cdot exp(\eta_3 t) + \epsilon_{ij}$

where $\epsilon \sim N(0; \sigma^2)$; the random effects are $(\eta_{0i}, \eta_{1i}, \eta_{2i}) \sim N(\eta, \Sigma)$:

```
fit=nlme(y ~ eta0+eta1*time+eta2*time*exp(-eta3*time), data=data,
  fixed=eta0+eta1+eta2+eta3 ~ 1,
  random=eta0+eta1+eta2 ~ 1|ID,
  start=c(2.264, -0.1152, 1.1464, 0.6682) )
```

References

Green, P. J. (1990). On use of the EM for penalized likelihood estimation. *Journal of the Royal Statistical Society. Series B (Methodological)* **52,** 443–452.

Hall, P., Mller, H., and Wang, J. (2006). Properties of principal component methods for functional and longitudinal data analysis. *The Annals of Statistics* **34,** 1493–1517.

Hastings, W. K. (1970). Monte carlo sampling methods using markov chains and their applications. *Biometrika* **57,** 97–109.

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

Johnson, R. A. and Wichern, D. W. (2007). *Applied Multivariate Statistical Analysis*. Pearson Prentice Hall.

Peng, J. (2009). 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.

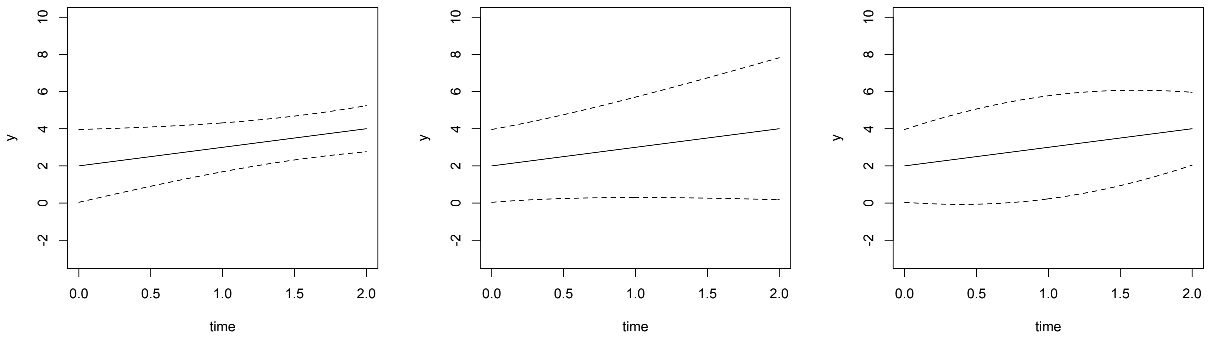
Rice, J. A. and Silverman, B. W. (1991). Estimating the mean and covariance structure nonparametrically when the data are curves. *Journal of the Royal Statistical Society. Series B (Methodological)* **53,** 233–243. ArticleType: research-article / Full publication date: 1991 / Copyright 1991 Royal Statistical Society.

Silverman, B. W. (1996). Smoothed functional principal components analysis by choice of norm. *The Annals of Statistics* **24,** 1–24.

Yao, F. and Lee, T. C. M. (2006). Penalized spline models for functional principal component analysis. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **68,** 3–25.

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

Zhou, L., Huang, J. Z., and Carroll, R. J. (2008). Joint modelling of paired sparse functional data using principal components. *Biometrika* **95,** 601 –619.

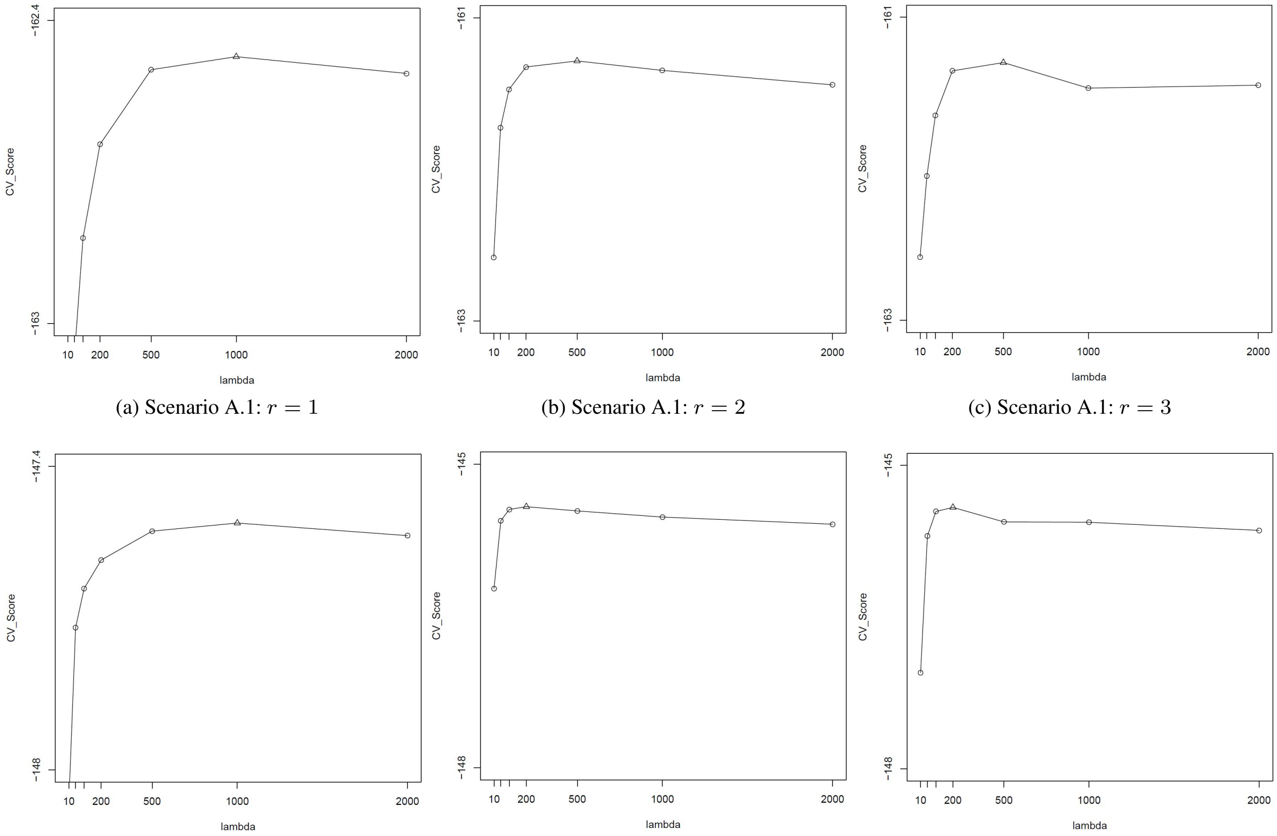


(a)

(b)

(c)

Figure 1: Potential temporal pattern of the mean and variance of a longitudinal process. The solid line represent the mean profile, which is assumed to be $f(t;\eta_0,\eta_1) = \eta_0 + \eta_1 t$ and (η_0,η_1) is a fixed parameter vector. The dashed lines represent the variability pattern. In Figure 1a and 1b, the between subject variability is derived from the mean profile using random effect of the parameters, i.e. the profile for subject i is $f(t;\eta_{i0},\eta_{i1})$with$(\eta_{i0},\eta_{i1}) \sim N((\eta_0,\eta_1);\mathbf{\Sigma}), \mathbf{\Sigma}$ is a 2×2 covariance matrix with elements $\tau_{kl}, k = 1,2; l = 1,2$. The variability $Var(f(t;\eta_{i0},\eta_{i1})) = \eta_0^2\tau_{11} + 2\eta_0\eta_1 t + \eta_0^2\tau_{22}t^2$is a quadratic function of t. In these cases, the variability in the middle of the time interval is always lower than the variability at least one end point. In Figure 1c, we consider a more flexible structure for the between subject variability, i.e. the profile for subject i is $f(t;\eta_0,\eta_1) + g_i(t)$ where $g_i(t)$ is general random process unrelated to the mean profile $f(t;\eta)$. The variability $Var(f(t;\eta_0,\eta_1) + g_i(t)) = Var(g_i(t))$ could demonstrate any pattern depending on the property of $g_i(t)$. In particular, if the variability is higher in the middle of the time interval, the variability structure cannot be characterized by the random effect model employed in Figure 1b and 1a. In the particular case of Figure 1c, $g_i(t) = \eta_{2i}(t-1)^2$ with $\eta_{2i} \sim Normal(0,\tau_{22})$. In the simplified example of Figure 1c, a random effects model with a random quadratic term, but not a quadratic term with fixed coefficient, could be used.



(a) Scenario A.1: $r = 1$

(b) Scenario A.1: $r = 2$

(c) Scenario A.1: $r = 3$

(d) Scenario A.2: $r = 1$

(e) Scenario A.2: $r = 2$

(f) Scenario A.2: $r = 3$

Figure 2: CV score vs. λ for various choices of r. The highest CV score is marked by a triangle in each graph.

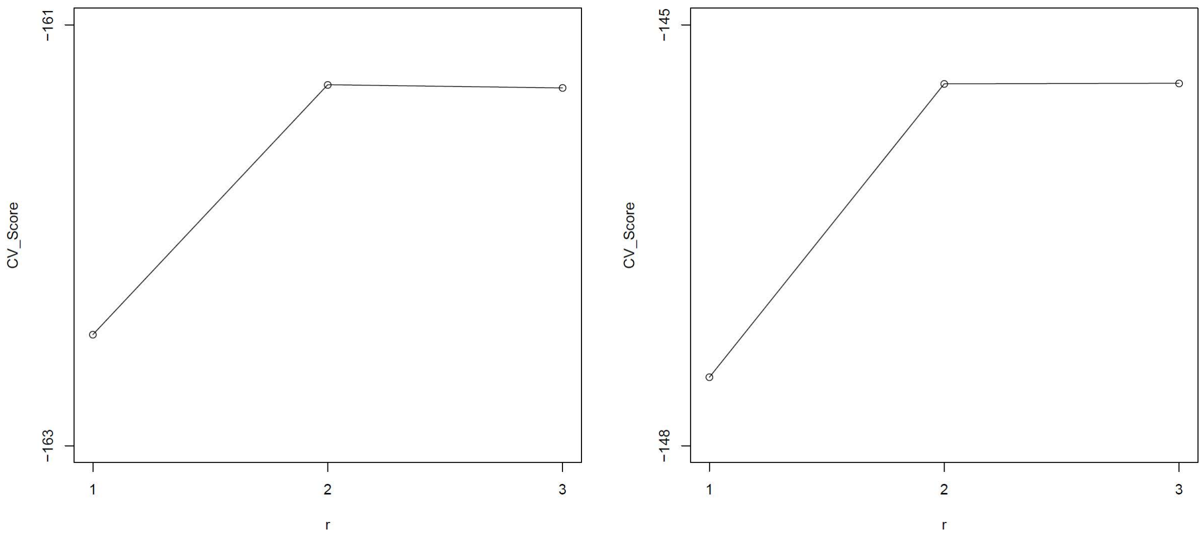


Figure 3: Scree plots for selected replicates in (left) Simulation A.1 and (right) Simulation A.2.