BIOMETRICS 000, 1–20 DOI: 000
000 XXXX

FPCA-based Method to Select Optimal Sampling Schedules that Capture Between-Subject Variability in Longitudinal Studies

Meihua Wu[1], Ana Diez-Roux[2], Trivellore E. Ragunathan[3], Brisa N. Sánchez*[,3]

[1] Gilead Sciences, Inc., Foster City, CA USA 94404

[2] Department of Epidemiology and Biostatistics, Drexel University, Philadelphia, PA USA 19104

[3] Department of Biostatistics, University of Michigan, Ann Arbor, MI USA 48109

**email:* brisa@umich.edu

SUMMARY: A critical component of longitudinal study design involves determining the sampling schedule. Criteria for optimal design often focus on accurate estimation of the mean profile, although capturing the between-subject variance of the longitudinal process is also important since variance patterns may be associated with covariates of interest or predict future outcomes. Existing design approaches have limited applicability when one wishes to optimize sampling schedules to capture between-individual variability. We propose an approach to derive optimal sampling schedules based on functional principal component analysis (FPCA), which separately characterizes the mean and the variability of longitudinal profiles and leads to a parsimonious representation of the temporal pattern of the variability. Simulation studies show that the new design approach performs equally well compared to an existing approach based on parametric mixed model (PMM) when a PMM is adequate for the data, and outperforms the PMM-based approach otherwise. We use the methods to design studies aiming to characterize daily salivary cortisol profiles and identify the optimal days within the menstrual cycle when urinary progesterone should be measured.

KEY WORDS: longitudinal design, nonlinear model design, optimal design, temporal pattern

1. Introduction

Carefully designed longitudinal studies with repeated measures can deepen our understanding of how biological processes evolve and enhance our ability to identify predictors of change. Longitudinal study design, however, is complex since it involves: 1) the number of subjects; 2) the number of samples per subject; and, in particular, 3) the spacing between samples (i.e., sampling schedule), while meeting budgetary and logistical constraints. In a motivating example, investigators want to identify times during the day at which to collect salivary cortisol, a stress biomarker that follows a nonlinear profile (Figure 1a). In another example, it is of interest to identify a small number of days during the menstrual cycle at which to measure urinary progesterone (Figure 4a).

Methods to determine the sampling schedule of repeated measures studies have received less attention than those for sample size and power calculations (e.g., Raudenbush and Liu, 2000; Retout et al., 2002; Stroud et al., 2001; Basagaña and Spiegelman, 2010). Available approaches include selecting optimal sampling schedules based on parametric nonlinear mixed models (PMM) (Fedorov and Hackl, 1997; Stroud et al., 2001), which are advantageous when a PMM adequately describes the longitudinal process, e.g. pharmacokinetic models for drug clearance rates. Ji and Müller (2017) develop methods to select the sampling schedule to optimize prediction of either individual trajectories or scalar responses that depend on functional predictors.

We propose approaches for selecting sampling schedules that help capture the between-individual variability of a longitudinal process, a novel area. Capturing such variability is important for at least three reasons: (a) correct variance models improve the quality of inference (Carroll, 2003); (b) the variability of a longitudinal predictor can be directly associated with a subsequent outcome (Elliott, 2007); (c) sampling the process at the times when between-invidual variability is larger relative to random measurement error can help identify correlates of the process.

Our methods can accomodate cases when estimating a parametric model for the population process is needed (i.e., both mean and variance are of interest), or cases when only identifiying sampling times to capture between-individual variance is desirable. We review existing methods based

on PMMs in Section 2, and show that parsimonious approaches to accommodate a broader range of between-subject variability patterns at the design stage is an area that needs improvement. In Section 3, we utilize functional principal component analysis (FPCA) to characterize the variability structure of the longitudinal process, and develop a method for identifying the optimal sampling schedules. Section 4 contains a simulation study to evaluate our approach. We demonstrate the methodology with two design problems in Section 5, and conclude with a discussion in Section 6.

2. Selecting Optimal Schedules Based on Parametric Mixed Models (PMM)

In this existing design approach, a parametric model is assumed for the individual profiles. Specifically, observations $y_i(t_{ij})$ of subject $i = 1, \ldots, N$ collected at time points $\boldsymbol{T} = (t_{i1}, ..., t_{in_i})$ follow

$$y_i(t_{ij}) = f(t_{ij}; \boldsymbol{\eta}_i) + \epsilon_{ij}, \quad \boldsymbol{\eta}_i \overset{iid}{\sim} MVN(\boldsymbol{\eta}, \boldsymbol{\Sigma}^*), \tag{1}$$

where $\epsilon_{ij} \sim N(0, \sigma^2)$ are independent and identically distributed (iid) measurement errors; and $\boldsymbol{\eta}_i$ is a $p \times 1$ subject-specific vector, and f is a known parametric function. These models can be estimated via maximum likelihood (see Web Appendix E and Pinheiro et al., 2015).

Although not needed for estimation, at the design stage it is often assumed that $n_i = m$ and $t_{ij} = t_j$ for $j = 1, ..., m$. The design goal is to find an optimal schedule $\boldsymbol{T}^* = (t_1, ..., t_m)$ that minimizes the estimation variance of $(\boldsymbol{\eta}, \boldsymbol{\Sigma}^*)$. Towards this goal, let $l(\boldsymbol{\eta}, \boldsymbol{\Sigma}, \sigma^2)$ be the log-likelihood, where $\boldsymbol{\Sigma}$ is the vector of unique parameters in $\boldsymbol{\Sigma}^*$ in (1); let $\widehat{\boldsymbol{I}}_{\eta,\Sigma,\sigma^2}(\boldsymbol{T}) = I(\boldsymbol{T}; \widehat{\boldsymbol{\eta}}, \widehat{\boldsymbol{\Sigma}}, \widehat{\sigma}^2)$ denote the information matrix evaluated at schedule $\boldsymbol{T}$ using available estimates $\widehat{\boldsymbol{\eta}}, \widehat{\boldsymbol{\Sigma}}, \widehat{\sigma}^2$; and let $\widehat{\boldsymbol{I}}_{\theta}(\boldsymbol{T})$ denote the submatrix of $\widehat{\boldsymbol{I}}_{\eta,\Sigma,\sigma^2}(\boldsymbol{T})$ corresponding to the subset $\boldsymbol{\theta} \in \{\boldsymbol{\eta}, \boldsymbol{\Sigma}, \sigma^2\}$. Various optimization criteria based on $\widehat{\boldsymbol{I}}_{\eta,\Sigma}(\boldsymbol{T})$ have been developed to select the optimal sampling schedules for $(\boldsymbol{\eta}, \boldsymbol{\Sigma}^*)$ (Atkinson et al., 2007). D-optimality, i.e. maximizing $|\widehat{\boldsymbol{I}}_{\eta,\Sigma}(\boldsymbol{T})|$, where $|\cdot|$ denotes the determinant, has desirable properties: 1) it is the reciprocal of the size of the confidence region for the MLE of $\boldsymbol{\eta}, \boldsymbol{\Sigma}$ for a fixed σ^2; 2) it is invariant under reparameterization of $\boldsymbol{\eta}, \boldsymbol{\Sigma}$; 3) it is convex, allowing the use of special optimization algorithms (Retout et al., 2002; Ogungbenro et al., 2005).

This PMM approach is useful in may applications, although three aspects can be improved. First, the PMM approach might not always be flexible enough to characterize the temporal pattern of the variability, given by $Var(f(t, \boldsymbol{\eta}_i))$ at time t, because both the variance of $\boldsymbol{\eta}_i$ and the functional form of $f(t, \boldsymbol{\eta})$ affect $Var(f_i(t))$. Web Figure 1 shows that PMM cannot always capture the variance patterns even when data generating models share the same population mean. Second, the PMM approach may not adapt to the situation where a parametric model that describes the mechanistic process (e.g., clearance rates for a drug) is unavailable. Last, evaluating the information matrix in PMM is often difficult since typically no closed form solution exists for the integral with respect to $\boldsymbol{\eta}_i$ when $f(t, \boldsymbol{\eta})$ is a general nonlinear function of $\boldsymbol{\eta}$ (Bazzoli et al., 2009).

3. Selecting Optimal Schedules Based on Functional Principal Component Analysis

3.1 *Modeling Strategy*

In constrast to PMM, we model the mean and variability structures separately, using FPCA: For subject i at time t_{ij}, we have $y_i(t_{ij}) = f(t_{ij}; \boldsymbol{\eta}) + g_i(t_{ij}) + \epsilon_{ij}$, where $\epsilon_{ij} \overset{iid}{\sim} N(0, \sigma^2)$ are random errors, and functional subject-specific deviations g_i are iid Gaussian processes with $E(g_i) = 0$, with standard conditions for the covariance structure of g_i (Ramsay and Silverman, 2005).

Mean profile. Our approach to modeling the mean profile includes two cases:

case (a): a known parametric function for the population mean profile is available and it is of interest to estimate the population mean parameters

case (b): a parametric function for the population mean is not available, or only capturing the between-individual variance is of interest.

When there is a parametric model for $f(t, \boldsymbol{\eta})$ that approximates the underlying biological process, it will often be advantageous to use it, since the population-level (mean) parameters would have meaningful interpretations. When a parametric model for $f(t, \boldsymbol{\eta})$ is unavailable, we de-trend the data by removing the population-level average curve, use the residuals to derive the principal

components, and focus the design on capturing the variance only. To de-trend the data, we use smoothers (e.g., regression splines, local regression) that the ensure the information matrix for the model is block diagonal, which yields optimization criteria that are easier to compute (below).

Variance model. To characterize g_i, we use FPCA, which consists of finding smooth functional principal components (FPC) $\beta_k(t)$ $k = 1, 2, 3, \ldots$ that maximize $Var\left(\int f_i(t)\beta_k(t)dt\right)$ with the orthonormal restrictions $\int \beta_k(t) \cdot \beta_{k'}(t)dt = 0$ for $k \neq k'$ and $\int \beta_k(t) \cdot \beta_k(t)dt = 1$. Each FPC accounts for variance $d_k = Var(\int f_i(t)\beta_k(t)dt)$. Under this framework, the subject-specific mean is $f_i(t) = f(t, \boldsymbol{\eta}) + \sum_{k=1}^{\infty} \alpha_{ik}\beta_k(t)$, where $\alpha_{ik} = \int f_i(t)\beta_k(t)dt$ is the loading on the kth FPC $\beta_k(t)$ for subject i. We assume the d_k are in descending order, thus α_{ik} is typically negligible for large k. Hence, the first few principal components $\beta_k(t), k = 1, \ldots, r$ capture the majority of the variability of the process, leading to a reduced rank model (James et al., 2000):

$$y_i(t_{ij}) = f_i(t_{ij}) + \epsilon_{ij} = f(t_{ij}, \boldsymbol{\eta}) + \sum_{k=1}^{r} \alpha_{ik}\beta_k(t_j) + \epsilon_{ij}. \quad (2)$$

With FPCA, the variability of $f_i(t)$ can be summarized by the variance of the component scores $\boldsymbol{D} = diag\{d_1, \ldots, d_r\}$, which is easier to handle than the $p \times p$ (unstructured) $\boldsymbol{\Sigma}^*$ in the PMM (1).

Estimation: Because preliminary data may be sparsely sampled, we use regularized FPCA (Ramsay and Silverman, 2005) to estimate the $\beta_k(t)$. However, ensuring orthogonality of the estimated FPCs is critically important for design purposes, in order to maintain independence of the estimated variance components $\widehat{d}_k$ and thus simplify the optimality criterion. Hence, we use a smoothness penalty, $\lambda \int \beta_k''(t)^2 dt$, on the FPCs $\beta_k(t)$ suggested by Zhou et al. (2008) that maintains orthogonality of the FPCs. Specifically, for fixed values of λ and r, we use the smoothing penalty from Zhou et al. (2008) and adapt the EM algorithm by James et al. (2000) to estimate $\beta_k(t)$ and $\boldsymbol{D}$. At each iteration of the EM algorithm, we employ a singular value decomposition (SVD) to reparameterize the FPCs. SVD improves convergence speed and enforces orthonormality of components, ensuring $\boldsymbol{D}$ is diagonal at every iteration. See Web Appendix A for more details.

We use k-fold cross validation (CV) to select r and λ. We split the preliminary dataset into

training and testing datasets, and define the CV score $s(r, \lambda)$ as the average of marginal log likelihoods of testing data based on the model estimated from training data (Equation (1) in Web Appendix A). Since higher $s(r, \lambda)$ suggests a better model, we choose $\lambda_r^* = argmax_\lambda s(r, \lambda)$ for a FPCA model with r components. To select r, we employ a rule based approach inspired by the "scree plot" (Johnson and Wichern, 2007). Specifically, for a pre-specified threshold of negligible improvement, $b\%$ (e.g., 1% to 5%), we select the smallest r such that the improvement in the CV score due to adding one more component is less than $b\%$. We use this approach because $s_{\lambda_r^*}(r) = s(r, \lambda_r^*)$ tends to increase with r, since a model with $r + 1$ components is more flexible than one with r components; thus choosing r to maximize $s_{\lambda_r^*}(r)$ does not necessarily lead to a parsimonious model that is more useful in the design stage. See Web Appendix B for more details about computing $s(r, \lambda)$ and example scree plots.

3.2 *Sampling times and optimality criteria*

Let S be the set of the admissible sampling times. Theoretically, our method does not place restrictions on S. However, it may be preferable to limit S to sampling times that can be implemented in practice (e.g., time points separated by at least half an hour in the salivary cortisol study). Let $\boldsymbol{T} = (t_1, ..., t_m)$ denote a candidate schedule where t_j, $j = 1, ..., m$ is chosen from S and $t_j \neq t_{j'}$ if $j \neq j'$. Let S_c denote the set of candidate schedules.

We focus on finding schedules for a fixed number m of samples per subject to be collected in the full/future study, all of whom follow the same schedule. Naturally, more samples per person will increase estimation precision (see the Simulation Section), but the number of samples will depend on budgetary constraints. The model has either $p + r$ (case (a)) or r (case (b)) parameters of interest; thus, we assume $m \geqslant p+r$ or $m \geqslant r$, respectively. Unless m is large or the components $\beta_k(t)$ are very smooth, it will in general be difficult to estimate individual profiles and/or $\beta_k(t)$ in the full study since we assume all subjects follow the same schedule. Designs were all subjects follow the same schedule are easier to implement in epidemiological studies that have many

other questionnaire-based and lab components, such the cortisol study. Nevertheless, the proposed methods identify sampling times that capture the between-individual variation as shown in the simulations and examples.

The design goal is to select a sampling schedule $\boldsymbol{T}^* = (t_1, ..., t_m) \in S_c$ that optimizes the D-optimal or D_s-optimal criterion (Atkinson et al., 2007), based on the information matrix for model (2). Web Appendix C shows the derivation of the matrix, which has a block diagonal structure with blocks for mean and variance parameters. We use subscripts to denote blocks of the information matrix done in Section 2, e.g., $\widehat{\boldsymbol{I}}_{\eta,D,\sigma^2}(\boldsymbol{T})$ is the full matrix while $\widehat{\boldsymbol{I}}_D(\boldsymbol{T})$ is the subblock of $\widehat{\boldsymbol{I}}_{\eta,D,\sigma^2}(\boldsymbol{T})$ corresponding only to parameters in $\boldsymbol{D}$. In case (a) we are interested in both $\boldsymbol{\eta}$ and $\boldsymbol{D}$, thus we obtain $\boldsymbol{T}^* = argmax_{T\in S_c} |\widehat{\boldsymbol{I}}_{\eta,D}(\boldsymbol{T})|$. For case (b) we employ the D_s-optimal criterion: $\boldsymbol{T}^* = argmax_{T\in S_c} |\widehat{\boldsymbol{I}}_D(\boldsymbol{T})|$, since the mean is no longer of interest. While in (2) $Var(f_i(t))$ depends on the FPCs $\beta_k(t)$'s and $\boldsymbol{D} = diag(d_1, ..., d_r)$, we focus on the estimation of $\boldsymbol{D}$ because between-individual heterogeneity is driven by subject-specific loadings α_{ik} (hence captured by $\boldsymbol{D}$) whereas the functional principal components $\beta_k(t)$ are shared by all subjects.

3.3 *Implementation*

If there are only a few time points to choose from, enumeration or a grid search works well. However, if there are many possible choices for the time points, more sophisticated optimization methods are needed. We implement a Metropolis-Hastings algorithm (Metropolis et al., 1953; Hastings, 1970) that is guaranteed to reach the global maximum if the Markov chain has converged. Since $|\widehat{\boldsymbol{I}}_{\eta,D}(\boldsymbol{T})|$ and $|\widehat{\boldsymbol{I}}_D(\boldsymbol{T})|$ are positive, we treat them as the probability function (less a constant) of a multivariate distribution of $t_1, ..., t_m$. The maximum of the criterion function corresponds to the mode of the probability distribution, which the Markov chain will visit with probability one. Details about the algorithm are in Web Appendix D.

Preliminary data is assumed to exist with which estimates $\widehat{\boldsymbol{\eta}}, \widehat{\boldsymbol{D}}, \widehat{\sigma}^2$ can be obtained and used to evaluate the information matrix $\widehat{\boldsymbol{I}}_{\eta,D,\sigma^2}$ needed to form the optimization criterion. Sampling over

the entire time interval of interest is needed in the preliminary data to reconstruct the variability pattern. This can be achieved by densely sampling profiles for few individuals or combining data from a larger number of individuals if they have different (sparse) sampling points.

4. Simulation study

Set up. We evaluate design methods for studies falling into case (a), where both FPCA and PMM apply, and case (b) where PMM is no longer applicable. We consider the following scenarios:

(A) The functional form of $f(t;\boldsymbol{\eta})$ is specified based on prior knowledge and designs that capture the mean and variance are of interest. The mean is generated from $f(t;\boldsymbol{\eta}) = \eta_0 + \eta_1 t + \eta_2 t\exp(-\eta_3 t)$ (Figure 1b). Between-subject variability is induced through

(A.1): random parameters in the mean profile, namely, $(\eta_{0i}, \eta_{1i}, \eta_{2i})^\intercal \sim MVN[(\eta_0, \eta_1, \eta_2)^\intercal, \boldsymbol{\Sigma}]$, with $\boldsymbol{\Sigma} = \text{diag}(0.347^2, 0.036^2, 0.3^2)$ ($\eta_{i3} = \eta_3$ is fixed so that $I_{\eta,\Sigma,\sigma^2}(\boldsymbol{T})$ has a closed form).

(A.2): the FPCs in Figure 1c are used with $\boldsymbol{D} = diag(0.8^2, 0.7^2)$.

(B) The functional form of the mean is not known or not of interest; designs that capture the between-subject variance are desirable. Data are generated as:

(B.1): same data as scenario (A.2) (but the focus of the design is different)

(B.2): variability induced using the FPCs in Figure 1d with $\boldsymbol{D} = diag(0.7^2, 0.7^2)$ or $\boldsymbol{D} = diag(1.4^2, 1.4^2)$.

In all scenarios $\sigma^2 = 0.5^2$. This simulation set up allows us to compare FPCA vs. PMM designs when data are from a PMM (A.1) or a reduced rank model (A.2); to see how the focus of the design (mean and variance, or variance only) impacts the $\boldsymbol{T}^*$ by comparing the results of (A.2) and (B.1); and to examine the performance of FPCA when FPCs are smoother (B.1) vs. more complex (B.2).

In all scenarios we assume $S = \{0, 0.5, ..., 15.5, 16\}$. We initially let S_c consist of schedules with $m = 7$ times chosen from S. The ideal sampling schedules, $\boldsymbol{T}_{ideal}$ which optimize the appropriate

design criteria computed using the true model are denoted with triangles along the time axes in Figure 2 for $m = 7$. We also compute $\boldsymbol{T}_{ideal}$ for a range of m, shown in Figure 3 for scenario B.2.

To evaluate the methods in a more realistic setting, we also obtain schedules using parameters estimated from (simulated) preliminary data instead of the true model. For each of the scenarios, we simulate $\ell = 1, ..., 1000$ preliminary data sets with $N = 200$ subjects, and each subject was assumed to take $n = 9$ samples chosen randomly from $\{0, 0.5, \ldots, 16\}$. Random noise $\tau_{ij} \sim N(0, 0.1^2)$ is added to t_{ij} to simulate noncompliance of the subjects. Given the times for each subject, outcome data are generated according to the true models above. We also examined cases when there were fewer subjects ($N = 90$) or fewer samples ($m = 5$).

In scenarios A.1 and A.2, we fit the PMM and FPCA assuming $f(t; \boldsymbol{\eta}) = \eta_0 + \eta_1 t + \eta_2 t \exp(-\eta_3 t)$ to each preliminary dataset. We use $b\% = 1\%$ as the threshold for negligible improvement to select r. In B.1 and B.2 we de-trend the data using local linear regression and only use the FPCA design method. Since in B.1 and B.2 the PMM is not applicable, use Ji and Müller (2017)'s idea to compare the FPCA to a 'random design'. Here, a random design consists of m times chosen from S at random without replacement. For all approaches, we use the Metropolis-Hastings algorithm described in Section 3.3 to obtain $\boldsymbol{T}_\ell^*$ for each preliminary dataset.

We evaluate the performance of the design methods in three ways. First, we investigate whether the $\boldsymbol{T}_\ell^*$ are in agreement with $\boldsymbol{T}_{ideal}$ by tabulating the relative frequency with which each sample time is selected. Second, we calculate the efficiency of $\boldsymbol{T}_\ell^*$ relative to $\boldsymbol{T}_{ideal}$ as a numerical benchmark. Relative efficiency (Atkinson et al., 2007) is calculated as $RE_\ell = \{|\boldsymbol{I}_\theta(\boldsymbol{T}_\ell^*)|/|\boldsymbol{I}_\theta(\boldsymbol{T}_{ideal})|\}^{1/q}$, where q is the number of parameters in θ, the subset of parameters for which the schedule is optimized, and $\boldsymbol{I}_\theta(\cdot)$ is information matrix evaluated at the true parameter values. Because $\boldsymbol{T}_{ideal}$ maximizes $|\boldsymbol{I}_\theta(\boldsymbol{T}_{ideal})|$, $RE_\ell \leqslant 1$, with higher values indicating better schedules. As a summary, we calculate average relative efficiency, $AvRE = \sum_{\ell=1}^{1000} RE_\ell/1000$.

As stated in the introduction, optimizing schedules that better capture between-individual vari-

ance can help identify correlates of the longitudinal process. Hence, in the third evaluation of the designs we examine whether sampling data at schedules $\boldsymbol{T}_{ideal}$, or $\boldsymbol{T}_\ell^*$ helps to detect associations between the profile and other covariates not specified at the design stage. We generate 1,000 testing datasets with the same parameters as, but otherwise independent from, the preliminary datasets. In each testing dataset we also simulate a univariate variable w_i that is correlated with the profiles via the functional principal component scores: $w_i = \gamma_1\alpha_{i1} + \gamma_2\alpha_{i2} + \delta_i$. Then, for each of the schedules $\boldsymbol{T}_\ell^* = (t_{\ell,1}, \ldots, t_{\ell,m})$ found using the $j = 1, \ldots, 1000$ preliminary datasets in a given simulation scenario, we fit the model $E(w_i) = \kappa_0 + \kappa_1 y_i(t_{\ell,1}) + \cdots + \kappa_m y_i(t_{\ell,m})$ and test the significance of the regression ($H_0 : \kappa_1 = \cdots = \kappa_m = 0$). We calculate power as the proportion of testing datasets where the p-value< 0.05 for this H_0. There are several other ways to test the association between the profiles and w_i, with this being a straightforward approach. We apply this evaluation to the ideal schedules as well as the random designs.

Results. The FPCA algorithm converged for all preliminary data sets in all scenarios. However, the nlme() function for fitting the PMM only converged for 858 preliminary data sets in A.1 and 528 in A.2. For A.1 and A.2, we restrict our comparison to the simulations where both methods converged because lack of convergence for the PMM would have triggered additional model diagnostics and model re-specification that are not straightforward to implement in a simulation.

For scenario A.1 (Figure 2a), both FPCA and PMM provide schedules that are similar to $\boldsymbol{T}_{ideal}$. There is perfect agreement between the two methods at the end points, $t = 0$ and $t = 16$, as well as the peak time $t = 1.5$. Furthermore, all the schedules include at least one of $t = 3.5$ or 4, which are near the inflection point of the profile (Figure 1b). The PMM approach has a slight advantage with $AvRE = 0.998$ vs. 0.968 for FPCA, and are both superior to the random design ($AvRE = 0.376$).

For scenario A.2 (Figure 2b), $\boldsymbol{T}_{ideal}$ includes $t = 7$, which captures the peak in $\beta_2(t)$ at $t = 7$ (Figure 1c). The $\boldsymbol{T}_\ell^*$ obtained with the FPCA method have at least one sample in the interval [6, 8] with probability 0.70, whereas the PMM approach never includes sampling times in this interval.

For the rest of the sampling times, FPCA and PMM have good agreement. The $AvRE$ is better for FPCA (0.972) vs. PMM (0.834), primarily due to FPCA's ability to include the sampling times near $t = 7$. Again, both FPCA and PMM are better than the random design (0.489). Scenario A.2 demonstrates the FPCA approach is more applicable than the PMM in scenarios where the temporal variability pattern is not induced by the random effects on the parametric mean profile.

Scenarios B.1 and B.2 shift the attention of the design to capture the variance only. $\boldsymbol{T}_{ideal}$ for scenario B.1 includes points in the [6,8] interval, as that in A.2, and times near $t = 16$ but not $t = 0$ (Figure 2c). This is due to the fact that $\beta_2(t)$ is zero near $t = 0$, but not $t = 16$ (Figure 1c). Again, the $\boldsymbol{T}_\ell^*$ tend to coincide with $\boldsymbol{T}_{ideal}$, although the degree of agreement, and consequently the $AvRE$, depends on the sample size in the preliminary data since the sample size affects the precision with which the FPCs are estimated. When the preliminary data had $N = 200$, $n = 9$, the $AvRE$ was 0.867; this dropped to $AvRE = 0.709$ when $N = 50$, $n = 9$, and to $AvRE = 0.610$ when $N = 90$, $n = 5$. Although the number of samples is $n \times N = 450$ in the latter two cases, $N = 50$, $n = 9$ has better $AvRE$ since having more samples per person improves the estimation of the FPCs. The random design (selecting 7 samples at random) had $AvRE = 0.620$.

Scenario B.2 (Figure 2d) illustrates the impact of the magnitude of the between-individual variance to that of the within-individual or random error variance, and demonstrates that optimal designs do not always include the end points of the sampling space. The agreement and $AvRE$ of the $\boldsymbol{T}_\ell^*$ obtained via the FPCA method depends on the magnitude of the variance of the FPCs (d_k's) relative to the variance of the random error ($\sigma^2 = 0.5^2$). Larger d_k implies the between-subject variability is larger (relative to the noise), and the principal components are more readily identified. With smaller d_k, the number of FPCs (below) and the shape of the FPCs not not well estimated from the preliminary datasets, resulting in a poorer designs (less agreement with ideal design, and lower $AvRE$). Nevertheless, in both cases, the FPCA yieled $AvRE$ that was better than the random design. The $\boldsymbol{T}_{ideal}$ (triangles in Figure 2d) does not include the end points since

both FPCs in this scenario (Figure 1d) are zero at the endpoints (i.e., except for random noise, $\epsilon's$, between-person variability is concentrated in the middle of the sampling space).

Scenario B.2 also illustrates a scenario when FPCs are more complex, and thus we focus on this scenario to show schedules with varying m. Figure 3a shows $\boldsymbol{T}_{ideal}$ for various m per subject, and the associated criterion D_s when $d_1 = d_2 = 0.7$. Clearly, D_s increases for higher m, indicating higher precision to estimate between-subject variance with larger m; the increase plateaus when we normalize by m (i.e., D_s/m). For the lowest m, the samples are placed at the first peak of component 1, and at the peak of component 2. For $m = 3$, the third sample is placed to capture the second peak of component 1, i.e., the more complex component. The placement of additional samples alternates between the timing of the peaks of the FPCs. As expected, when m increases, so does the $AvRE$, for both $\boldsymbol{T}^*_\ell$ obtained by FPCA and for random designs.

From Figure 3a we also see that the power to detect an association between a covariate (unknown at the design stage) and the profile increases as m increases, although it quickly reaches a plateau– around $m = 4$ in this scenario. It is also clear that $\boldsymbol{T}_{ideal}$ designs obtained by the FPCA method have higher power than those obtained from random sampling. Average power for the schedules $\boldsymbol{T}^*_\ell$ obtained when FPCA method relies on preliminary data is also higher than random designs, although of course lower than for $\boldsymbol{T}_{ideal}$. Although both $AvRE$ and power increase with m, there is not a one-to-one correspondence between relative efficiency and power (i.e., designs with the same RE can have different power), since the optimization criterion does not include information on the covariate (Figure 3b). However, the key point is that optimizing the sampling schedule to maximize $|\boldsymbol{I}_D(\boldsymbol{T})|$ tends to yield designs with higher power to detect associations between the profile and covariates, even when these were not included at design stage.

Finally, Figure 3c shows the sensitivity of the relative efficiency of individual designs to the number of components r, which is also selected empirically when using preliminary data. Even though in this simulation scenario the selected r is 1 in 83% of the data sets, more than 75% of the

random designs have relative efficiency below the 50^{th} percentile of the relative efficiency of the designs selected by FPCA with $r = 1$.

5. Examples

Design for Salivary Cortisol Studies. Given its objective nature compared to self-reported questionnaires, salivary cortisol is increasingly common in epidemiological studies seeking to study stress (Adam and Kumari, 2009). Salivary cortisol exhibits a nonlinear diurnal pattern through the length of the day (Figure 1a). We apply the PMM and FPCA methods in the design for salivary cortisol studies. As preliminary data for the design, we use data on 850 individuals from wave I of the Stress Ancillary Study of the Multi-Ethnic Study of Atherosclerosis (MESA Stress) (Hajat et al., 2010) in order to suggest possible designs for wave II of MESA Stress. In wave I, individuals collected 6 samples for 3 days (Figure 1a).

For wave II of MESA Stress, the design goal is to identify a schedule, with 6 samples, that maximizes the precision to estimate the mean profile and also better capture the between-subject variability. We consider sampling times between wake up ($t = 0$) and 16 hours after wake up, which are spaced by either half an hour, or 10 minutes. Spacing by half hour is much more practical, while the 10 minute spacing allows us to examine the sensitivity of the design to the spacing assumed. Existing studies suggest $f(t; \boldsymbol{\eta}) = \eta_0 + \eta_1 t + \eta_2 t \cdot exp(\eta_3 t)$ is a suitable mean profile for salivary cortisol (Stroud et al., 2004). We use this mean profile in the PMM and FPCA (case (a)) approaches. The values of $r = 3$ and $\lambda = 2000$ were determined by 10-fold cross validation for the reduced rank model (2). For the half-hour spacing, we enumerate the $\binom{33}{6}$ candidate schedules and identify the best five schedules. For the 10 minute spacing we use the Metropolis-Hastings algorithm described in Section 3 and Web Appendix D.

Five schedules with the best criterion values obtained from each method are given in Table 1(a) and (b). For both methods, all schedules include four common sampling times: $0, 0.5, 1$ and 16 hour

after wake up. The difference between the two methods is revealed in the time period between 4 and 16. PMM places the remaining sample close to the ends of this time period and none in between since in the 2-16 hrs time period the mean profile is dominated by the linear term $\eta_0 + \eta_1 t$. The linear term induces higher variability at the end points of the time period, thus sampling times are placed there. However, FPCA detects a different temporal pattern for the variability and places remaining sampling time at around 11 hours after wake up. This new sampling time was discovered mainly because the FPCA approach places less restriction on the variance structure. Using the 10 minute spacing yields slightly different optimal schedules.

Urinary Progesterone Study. Urinary progesterone is an important biomarker of reproductive health (De Souza et al., 2010). Studies with a small number of subjects, N, often collect samples everyday during the menstrual cycle (Waller et al., 1998). Since such an intense schedule would be difficult and costly to implement in studies with large N, it is important to find simplified sampling schedules that adequately capture the between-subject variation of the progesterone levels. Since no parametric form fits the mean adequately (Brumback and Rice, 2009), we only use FPCA.

As preliminary data, we use the urinary progesterone data from Brumback and Rice (2009), which were collected as part of early pregnancy loss studies. The data set contained progesterone profiles of 91 menstrual cycles from 51 women (Figure 3a). We randomly selected only one cycle from each of the women who contributed data on multiple cycles to ensure independence. As is standard practice in endocrinological research, progesterone profiles were aligned by the day of ovulation (day=0) and then truncated at each end to present curves of equal length (24 days).

In the absence of a parametric model for the mean, we first center the data at each time point with the mean estimated by a local polynomial smoother, and perform FPCA on the residuals. The number of components $r = 3$ and the smoothing parameter $\lambda = 1000$ were determined by 10-fold cross validation and $b\% = 1\%$ as the threshold. The first component accounts for 53.2% of the variance and can be interpreted as a constant cycle-level deviation from the mean (Figure 4b). The

second and third components account for 36.6% and 1.4% of overall variance, and capture local deviations at various times within the cycle.

Since $r = 3$ components were selected to characterize the progesterone data, we consider sampling schedules with $m = 3$ sampling times. We consider each day of the menstrual cycle as a potential time, thus $S = \{-8, -7, ..., 15\}$, and S_c consists of schedules with 3 sampling times from S. Table 1(c) lists the 5 best schedules with 3 samples each. These schedules exhibit a clear pattern. The earliest time point is in the interval $[-7, -5]$; the second sample is in the interval and $[7, 8]$, and the third sample is always $t = 15$. The choice of the sampling times can be intuitively understood if we refer to the principal components (Figure 4b). The deviation from the mean of the second and third components are relatively higher at days prior to day -4, around day $t = 7$, and towards the end of the cycle near day $t = 15$.

As a 'proof of concept' that selecting sampling times to better capture the variability of the profiles is important for assessing between-individual differences, we examined the performance of the optimal schedule derived from FPCA in predicting conceptive status (whether the woman had conceived during the cycle) the only covariate available in the data set. This complements the simulation study where power to detect associations was illustrated. We consider the leave-one-out prediction rate of all the schedules for predicting the conceptive status (since a replication dataset is not available), and emphasize that conceptive status was not used to inform the design. Since only three samples are selected in the design, prediction rates were computed by a logistic regression model with conceptive status as the outcome and progesterone levels measured at the days indicated by each of the derived schedules as the predictors. The leave-one-out prediction rate of the five best sampling schedules obtained by FPCA are between $93\% - 94\%$ (the fourth column of Table 1c), which suggests that just three progesterone values are highly predictive of conceptive status, provided they are collected with the suggested schedules. When we rank all candidate schedules in S_c by the leave-one-out prediction rate (as if prediction rate was the optimality

criterion), we see that the sampling schedules we obtained perform better than $94\% - 98\%$ of all candidate schedules (the last column of Table 1c). Given the excellent performance of just three samples, in practice it would be difficult to argue in favor denser schedules.

6. Discussion

We propose a semiparametric approach to longitudinal study design that optimizes estimation of the mean profile and between-subject variability. We use FPCA to model the mean and variability of the profiles separately, and in turn obtain a parsimonious and flexible representation of the temporal pattern of the variability. In simulations and data examples, we show that the FPCA approach is comparable or better than existing design methods based on PMM. One key advantage of the FPCA over the PMM approach is that the former does not assume that the variance model follows the pattern specified by the mean model. Another advantage is that computing the optimality criteria for the FPCA approach can be much faster since the information matrix can be computed in closed form whereas PMM often requires numerical integration.

We used FPCA to identify sampling schedules that capture between-individual variability. To our knowledge, our proposed method is the first to employ FPCA to derive optimal sampling schedules for the estimation of the between-subject variability. Fedorov and Hackl (1997) uses an FPCA to model correlated data and consider the design that minimizes the errors in predicting the profile. Ji and Müller (2017) use FPCA to derive schedules to optimize prediction accuracy, and are thus not directly comparable to our approach. Mean and variance models (Davidian et al., 1988) could potentially be used to identify samples to capture the variance, but since the models focus on the marginal distribution of the data they provide less insight into the between-individual variability.

There are some limitations to our approach and several possibilities for extending it. It is of interest to extend the proposed methods to select schedules that optimize power to detect associations between the longitudinal profiles and its correlates. While Ji and Müller (2017) use predictive

criteria for continuous outcomes to advance this area, we demonstrated that our proposed methods have power to detect associations between the profile and correlates even though the correlate was not specified in the design. Retout et al. (2007) evaluate the influence of designs on the power the Wald test has to detect a treatment effect in PMM. Similar to the normality assumptions on the random effects and residuals in the PMM-based approach, our FPCA approach relies on normality of the component scores and normally distributed residuals. These assumptions are needed in deriving the information matrix and thus form the optimality criteria. Defining optimality criteria that don't rely on the information matrix, e.g., prediction of a subsequent outcome, including non-normal outcomes, is an interesting extension in its own right, and could also circumvent normality assumptions. Ji and Müller (2017) have advanced this line of research by using prediction as the optimality criteria, although components of their work rely on normality assumptions.

Our method considers only one optimal schedule in the design due to practical constraints of large scale epidemiology studies. Future studies should consider methods where multiple schedules, randomly assigned to subjects, are combined to optimally capture population-level parameters (e.g., Mentré and Baccar, 1997). The FPCA approach requires preliminary data with dense sampling, and thus there are situations where it cannot be applied but the PMM approach can. In the cortisol example, each individual collects very few samples in the preliminary data, but dense sampling is achieved since many individuals are included. Clearly, neither approach will select sampling points where no preliminary data were collected (e.g., functional principal components cannot be estimated in those regions). However, those areas may be particularly interesting to sample, and a statistical model is not needed to make such determination. Other directions include approaches to obtain the sampling schedules for estimating $\beta_k(t)$, or for the sampling schedules to incorporate multilevel sampling (e.g., repeated cycles from the same woman). Incorporating cost into the optimality criteria may also help determine the number of samples m.

Supplementary Materials

Web Appendices and Figures referenced in Sections 2, 3, 4, and 5 as well as R code to implement the methods are available with this paper at the Biometrics website on Wiley Online Library.

Acknowledgements

This work was funded by National Institutes of Health Grant R21 DA024273 and R01 HL101161. MESA data collection (Section 5) was supported by contracts N01-HC-95159 through N01-HC-95169 from the National Heart, Lung, and Blood Institute. A full list of participating MESA investigators and institutions can be found at http://www.mesa-nhlbi.org. The authors thank Dr. Bill Lasley for generously providing the urinary progesterone dataset.

References

Adam, E. K. and Kumari, M. (2009). Assessing salivary cortisol in large-scale, epidemiological research. *Psychoneuroendocrinology* **34,** 1423–1436.

Atkinson, A. C., Donev, A. N., and Tobias, R. (2007). *Optimum experimental designs, with SAS*. Oxford University Press, Oxford.

Basagaña, X. and Spiegelman, D. (2010). Power and sample size calculations for longitudinal studies comparing rates of change with a time-varying exposure. *Statistics in Medicine* **29,** 181–192.

Bazzoli, C., Retout, S., and Mentré, F. (2009). Fisher information matrix for nonlinear mixed effects multiple response models: Evaluation of the appropriateness of the first order linearization using a pharmacokinetic/pharmacodynamic model. *Statistics in Medicine* **28,** 1940–1956.

Brumback, B.A. and Rice, J.A. (1998). Smoothing spline models for the analysis of nested and crossed samples of curves. *Journal of the American Statistical Association* **93,** 961–976.

Carroll, R. J. (2003). Variances are not always nuisance parameters. *Biometrics* **59,** 211–220.

Davidian, M., Carroll, R. J., and Smith, W. (1988). Variance functions and the minimum detectable concentration in assays. *Biometrika* **75,** 549 –556.

De Souza, M., Toombs, R., Scheid, J., O'Donnell, E., West, S., and Williams, N. (2010). High prevalence of subtle and severe menstrual disturbances in exercising women: confirmation using daily hormone measures. *Human Reproduction* **25,** 491 –503.

Elliott, M. R. (2007). Identifying latent clusters of variability in longitudinal data. *Biostatistics* **8,** 756 –771.

Fedorov, V. V. and Hackl, P. (1997). *Model-oriented design of experiments*. Springer, New York.

Hajat, A., Diez-Roux, A., Franklin, T. G., Seeman, T., Shrager, S., Ranjit, N., et al. (2010). Socioeconomic and race/ethnic differences in daily salivary cortisol profiles: the Multi-Ethnic Study of Atherosclerosis. *Psychoneuroendocrinology* **35,** 932–943.

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.

Ji, H. and Müller, H.G. (2017). Optimal designs for longitudinal and functional data. *Journal of the Royal Statistical Society-Series B* **xx,** xx–xx (in press).

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

Mentré, F., M. and Baccar, D. (1997). Optimal design in random-effects regression models. *Biometrika* **84,** 429–442.

Metropolis, N., Rosenbluth, A. W., Rosenbluth, M. N., Teller, A. H., and Teller, E. (1953). Equation of state calculations by fast computing machines. *The Journal of Chemical Physics* **21,** 1087–1092.

Ogungbenro, K., Graham, G., Gueorguieva, I., and Aarons, L. (2005). The use of a modified

Fedorov exchange algorithm to optimise sampling times for population pharmacokinetic experiments. *Computer Methods and Programs in Biomedicine* **80,** 115–125.

Pinheiro, J., Bates, D., DebRoy, S., Sarkar, D., and R Core Team (2015). *nlme: Linear and Nonlinear Mixed Effects Models*. R package version 3.1.

Ramsay, J. and Silverman, B. (2005). *Functional Data Analysis*. Springer, New York, 2^{nd} edition.

Raudenbush, S. W. and Liu, X. (2000). Statistical power and optimal design for multisite randomized trials. *Psychological Methods* **5,** 199–213.

Retout, S., Mentré, F., and Bruno, R. (2002). Fisher information matrix for non-linear mixed-effects models: evaluation and application for optimal design of enoxaparin population pharmacokinetics. *Statistics in Medicine* **21,** 2623–2639.

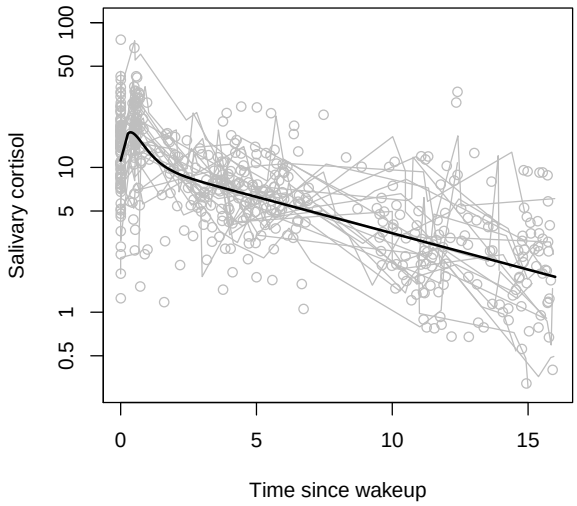
Retout, S., Comets, E., Samson, A., and Mentré, F. (2007). Design in nonlinear mixed effects models: Optimization using the Fedorov-Wynn algorithm and power of the Wald test for binary covariates. *Statistics in Medicine* **26,** 5162–5179.

Stroud, J. R., Müller, P., and Rosner, G. L. (2001). Optimal sampling times in population pharmacokinetic studies. *Journal of the Royal Statistical Society: Series C* **50,** 345–359.

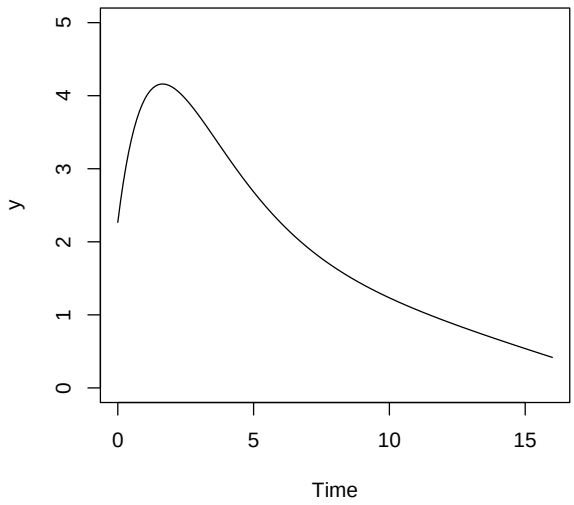
Stroud, L. R., Papandonatos, G. D., Williamson, D. E., and Dahl, R. E. (2004). Applying a nonlinear regression model to characterize cortisol responses to corticotropin-releasing hormone challenge. *Annals of the New York Academy of Sciences* **1032,** 264–266.

Waller, K., Swan, S. H., Windham, G. C., Fenster, L., Elkin, E. P., and Lasley, B. L. (1998). Use of urine biomarkers to evaluate menstrual function in healthy premenopausal women. *American Journal of Epidemiology* **147,** 1071 –1080.

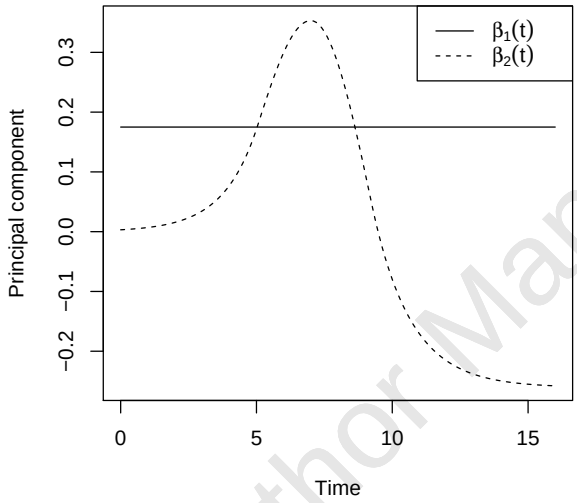
Zhou, L., Huang, J. Z., and Carroll, R. J. (2008). Joint modelling of paired sparse functional data using principal components. *Biometrika* **95,** 601 –619.



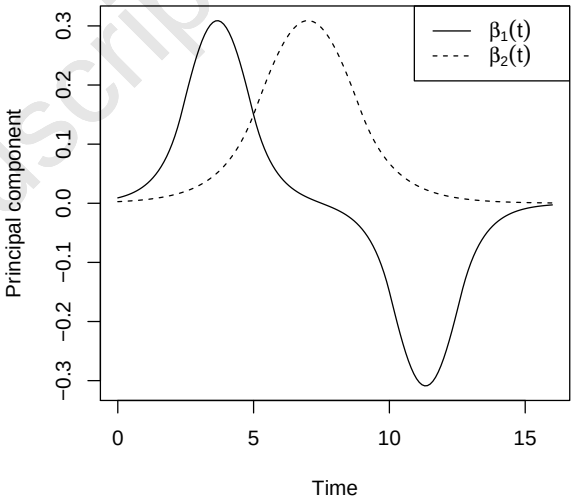
(a) Salivary cortisol data



(b) Mean profile for the simulations

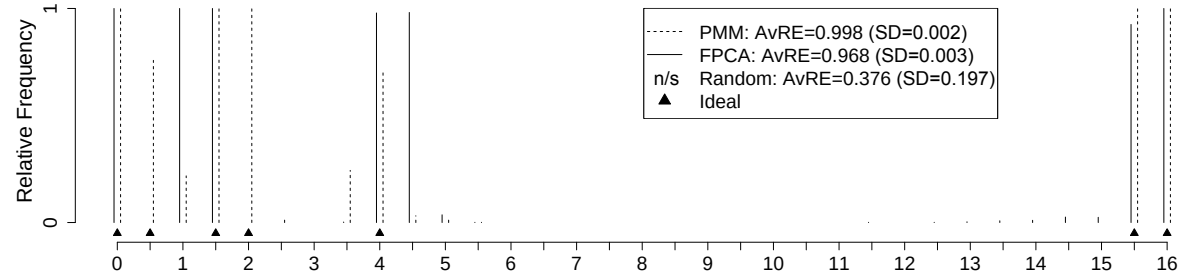


(c) FPCs in Scenarios A.2 and B.1

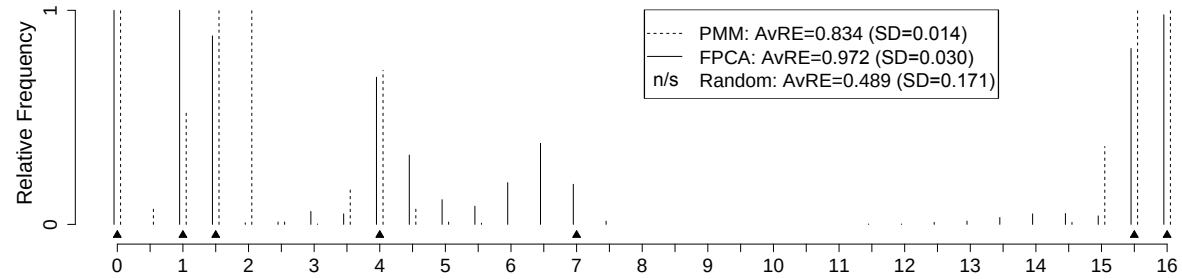


(d) FPCs in Scenario B.2

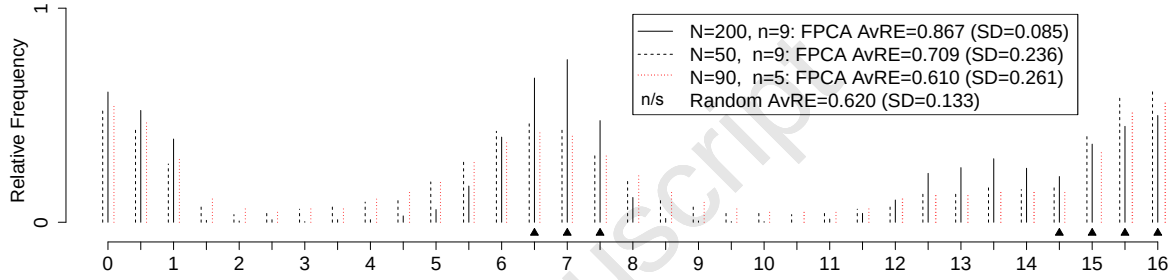
Figure 1: (a) Scatterplot of salivary cortisol data described in Section 5; (b) mean profile and (c)-(d) functional principal components for the variability structure in the simulations.



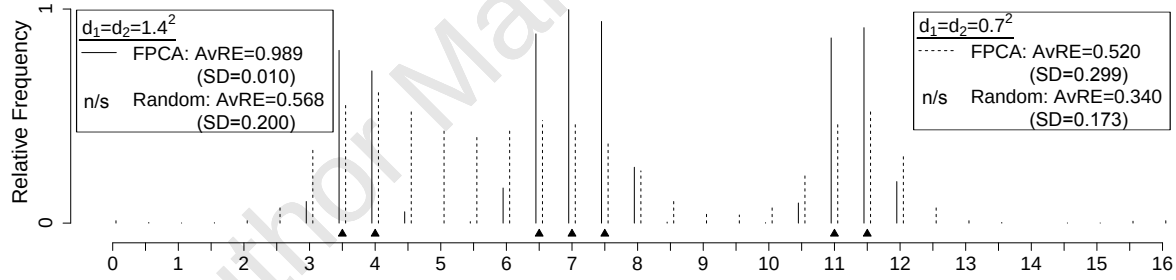
(a) Scenario A.1



(b) Scenario A.2

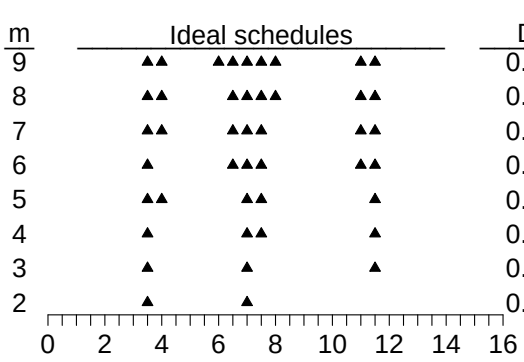


(c) Scenario B.1



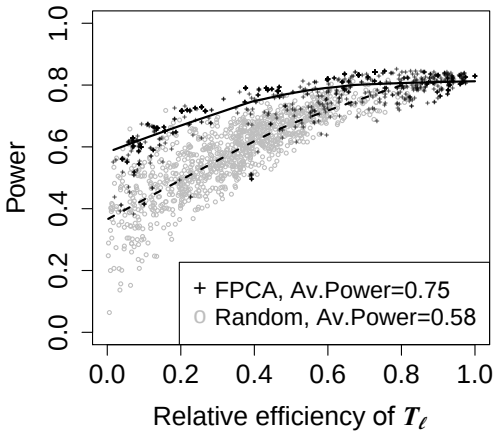
(d) Scenario B.2

Figure 2: Frequency with which each time point is included in the optimal schedules (random design is not shown, n/s). Higher frequencies concentrated near the ideal design $\boldsymbol{T}_{ideal}$ demonstrate higher level of agreement between the designs obtained after fitting the model to preliminary data ($\boldsymbol{T}_{\ell}^{*}$) and $\boldsymbol{T}_{ideal}$. Legends give the average (AvRE) and standard deviation (SD) of the relative efficiency of the designs $\boldsymbol{T}_{\ell}^{*}$ relative to $\boldsymbol{T}_{ideal}$.

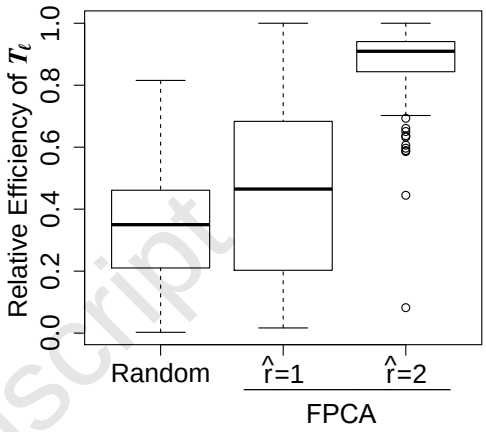


m	Ideal D_s	Ideal Power	Est. FPCA AvRE	Est. FPCA Power	Random AvRE	Random Power
9	0.59	0.83	0.63	0.75	0.42	0.61
8	0.51	0.83	0.58	0.76	0.38	0.6
7	0.42	0.83	0.52	0.75	0.34	0.58
6	0.33	0.81	0.46	0.72	0.3	0.55
5	0.22	0.82	0.41	0.72	0.27	0.52
4	0.14	0.8	0.32	0.69	0.22	0.48
3	0.06	0.75	0.26	0.65	0.2	0.44
2	0.02	0.73	0.2	0.6	0.13	0.37

(a) Precision (D_s) and power for $\boldsymbol{T}_{ideal}$, and average relative efficiency ($AvRE$) and average power for $\boldsymbol{T}^*_\ell$ and random designs for various m

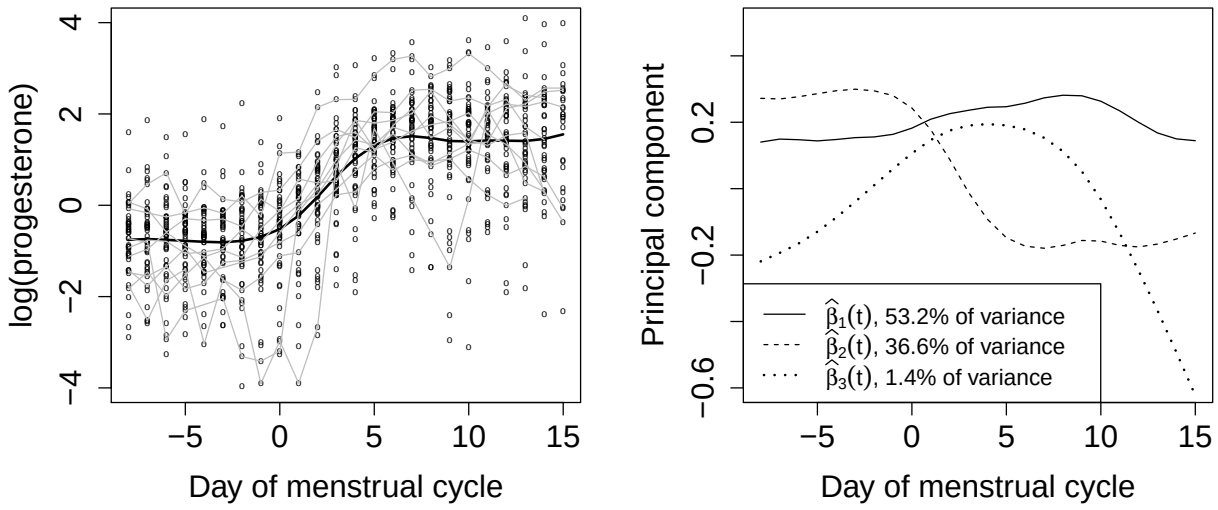


(b) Relationship between power and RE of $\boldsymbol{T}^*_\ell$



(c) Impact of estimating r on RE of individual $\boldsymbol{T}^*_\ell$

Figure 3: Additional simulation results for scenario B.2, d_1=d_2=0.7. Power calculations assumed $\gamma_1 = \gamma_2 = 1$, and $Var(\delta_i) = 6$.



(a) (b)

Figure 4: (a) Scatter plot of log(urinary progesterone) vs. day in the menstrual cycle, including the sample mean (thick black line) and a sample of individual's trajectories (gray), and (b) functional principal components of the urinary progesterone data.

Spacing	t_1	t_2	t_3	t_4	t_5	t_6
30min	0:00	0:30	1:00	3:00	15:30	16:00
	0:00	0:30	1:00	2:30	15:30	16:00
	0:00	0:30	1:00	3:30	15:30	16:00
	0:00	0:30	1:00	3:00	15:00	16:00
	0:00	0:30	1:00	2:30	15:00	16:00
10min	0:00	0:20	0:40	2:20	15:50	16:00

(a) Cortisol study, Schedules selected with PMM

Spacing	t_1	t_2	t_3	t_4	t_5	t_6
30min	0:00	0:30	1:00	4:00	11:00	16:00
	0:00	0:30	1:00	4:00	10:30	16:00
	0:00	0:30	1:00	4:30	11:00	16:00
	0:00	0:30	1:00	4:00	11:30	16:00
	0:00	0:30	1:00	4:00	10:00	16:00
10min	0:00	0:20	1:00	4:00	10:50	16:00

(b) Cortisol study, Schedules selected with FPCA

			Prediction	
t_1	t_2	t_3	Rate	Rank
-6	8	15	93%	94%
-5	8	15	93%	94%
-7	8	15	93%	94%
-6	7	15	93%	94%
-5	7	15	94%	98%

(c) Progesterone study, schedules selected with FPCA

Table 1: Sampling schedules chosen by the (a) PMM and (b) FPCA based approaches in the Cortisol study when samples are spaced by half an hour or 10 minutes. (c) The top five sampling schedules for urinary progesterone chosen by the FPCA approach. In (c) the leave-one-out prediction rate is for the prediction of conceptive status based on the progesterone measured at the stated days. The ranking in the last column is based on the leave-one-out prediction rate.