Modeling and Analysis of Functional Method Comparison Data

Galappaththige S. R. de Silva

Lasitha N. Rathnayake

Pankaj K. Choudhary[1]

Department of Mathematical Sciences, FO 35

University of Texas at Dallas

Richardson, TX 75080-3021, USA

Abstract

We consider modeling and analysis of functional data arising in method comparison studies. The observed data consist of repeated measurements of a continuous variable obtained using multiple methods of measurement on a sample of subjects. The data are treated as multivariate functional data that are observed with noise at a common set of discrete time points which may vary from subject to subject. The proposed methodology uses functional principal components analysis within the framework of a mixed-effects model to represent the observations in terms of a small number of method-specific principal components. Two approaches for estimating the unknowns in the model, both adaptations of general techniques developed for multivariate functional principal components analysis, are presented. Bootstrapping is employed to get estimates of bias and covariance matrix of model parameter estimates. These in turn are used to compute confidence intervals for parameters and functions thereof, such as the measures of similarity and agreement between the measurement methods, that are necessary for data analysis. The estimation approaches are evaluated using simulation. The methodology is illustrated by analyzing two datasets.

Keywords: Agreement; bootstrap; functional principal component analysis; Karhunen-Loeve expansion; mixed model; PACE method.

[1]Corresponding author. Email: pankaj@utdallas.edu, Tel: (972) 883-4436, Fax: (972)-883-6622.

1 Introduction

Multivariate functional data arise when repeated measurements of $J\,(\geq 2)$ variables are taken over time on every subject [1–5]. The measurements of each variable on a subject are assumed to be values of an underlying smooth random function that is observed with noise at discrete time points. For each subject, all the J variables are recorded at every observation time. Thus, these data consist of J curves per subject, observed at a common set of discrete observation times. This set of times, however, may vary from subject to subject. There is dependence in the J curves as they come from the same subject.

We are specifically interested in the special case of multivariate functional data arising in method comparison studies [6]. They involve measuring a continuous variable on every subject using multiple methods of measurement in a common unit. All methods measure the variable with error. The primary goal in these studies is to evaluate whether the methods agree sufficiently well to be used interchangeably. It is evident from more than 25,000 citations of Bland and Altman [7], which proposed the popular limits of agreement approach for agreement evaluation with scalar observations, that such studies are common in biomedical sciences. The measurements from the multiple methods are the dependent functional variables here.

For example, consider two method comparison datasets, both with $J = 2$, that motivated this work—body fat data Chinchilli et al. [8] and body temperature data Li and Chow [9]. In the first, we have measurements of percentage body fat made using skinfold calipers and dual energy x-ray absorptiometry (DEXA) in a cohort of adolescent girls over a period of about 4 years. These longitudinal data are an example of sparse bivariate functional data. In the second, we have core body temperature—the temperature of tissues deep within the body, measured every minute over a period of 90 minutes at two locations in the body—esophagus and rectum. These data are an example of dense bivariate functional data. Our interest is

in evaluating agreement between measurements from caliper and DEXA methods in the first case and between measurements taken at the two body locations in the second case.

There is a growing body of literature on the analysis of method comparison data. See Barnhart et al. [10] and Choudhary and Nagaraja [6] for an introduction. Nevertheless, almost all the literature assumes that the observations are scalar. For scalar data, evaluation of agreement between two methods involves quantifying how far the methods are from having perfect agreement, in which case the joint distribution of the methods is concentrated on the line of equality. In other words, two methods in perfect agreement have equal means, equal variances, and a correlation of one; or equivalently, their differences are zero with probability one. In the statistical literature, agreement is commonly evaluated by performing inference on *measures of agreement* such as concordance correlation coefficient (CCC) of Lin [11] and total deviation index (TDI) of Lin [12]. However, in the biomedical literature, the limits of agreement approach of Bland and Altman [7] is the most popular. These measures are defined in Section 4. A reader interested in their comparison may consult Barnhart et al. [10]. In addition to evaluation of agreement, a secondary goal of a method comparison study is to evaluate similarity of methods by comparing their marginal characteristics such as means and precisions. This is typically done by performing inference on *measures of similarity* such as mean difference and precision ratio [13]. Evaluation of similarity is a necessary supplement to evaluation of agreement as it provides information about the sources of disagreement between the methods [6, Chapter 1].

In the method comparison literature, we are only aware of Li and Chow [9] that deals with functional observations. It extends the ideas of Lin [11] to develop a CCC for functional data from $J = 2$ methods. But this approach has drawbacks that limit its usefulness. First, it produces a single overall index of agreement over the entire time interval. However, given the functional nature of the data, an index that changes smoothly over time may be preferable over the overall scalar index because the former allows insight into how the extent

of agreement changes over time. Second, the approach is specifically designed for CCC—a function of first and second order moments of the measurements. It is unclear how the approach can be adapted for other measures of agreement such as TDI, which is a percentile (see Section 4). This is an issue because CCC is often criticized for being unduly influenced by the between-subject variation in the data as it may lead to misleading conclusions (see, e.g., Barnhart et al. [10]). Third, the approach assumes that all curves are observed at the same time points. This assumption is unnecessarily restrictive. For example, it does not hold for the body fat data although it holds for the body temperature data. Fourth, the approach in its present form cannot deal with $J > 2$ methods. These drawbacks may be overcome by a model-based approach for analyzing functional method comparison data. The model parameters can be used to obtain functional analogs of any measure of similarity and agreement for scalar observations. The model would allow the observation times to differ between the subjects. It can also accommodate more than two methods. This is the approach we take in this article.

Functional data analysis is currently an active area of research, see Ramsay and Silverman [1] for an introduction. A common analytical approach involves performing a functional principal components analysis (FPCA) to obtain a parsimonious representation of the data [1, Chapter 8]. The PACE (principal components analysis through conditional expectation) methodology of Yao et al. [14] is a popular approach for FPCA of data that are observed with measurement error. It involves decomposing the functional observations via a Karhunen-Loève expansion and using the framework of mixed-effects model for estimating coefficients in the expansion as best linear unbiased predictors of random effects, and estimating error variance by smoothing the covariance function. This approach and its refinement due to Goldsmith et al. [15] are implemented in the `refund` [16] and `MFPCA` [17] packages for the statistical software system `R` [18].

Methodologies for FPCA of multivariate functional data have also been developed, see,

e.g., Ramsay and Silverman [1, Chapter 8], Berrendero et al. [2], Jacques and Preda [3], Chiou et al. [4], and Happ and Greven [5]. Among these, the approaches of Chiou et al. [4] and Happ and Greven [5] are of specific interest in this article as they can be used for data observed with measurement error, which is the case for our method comparison data. Although Chiou et al. [4] and Happ and Greven [5] differ in their basic premise regarding univariate components of the multivariate observation—in particular, they may have different units in Chiou et al. [4] and they may be observed on different (dimensional) domains in Happ and Greven [5]—both first obtain a Karhunen-Loève expansion of the multivariate observations. Thereafter, Chiou et al. [4] estimate the unknowns by a generalization of the PACE methodology. They also employ normalization to deal with the different units. On the other hand, Happ and Greven [5] establish a relation between univariate and multivariate FPC decompositions and employ it to obtain estimates of the unknowns in the multivariate model using their estimates from the univariate models. The univariate estimates may be obtained, e.g., using the PACE approach of Yao et al. [14]. This methodology is implemented in an `R` package `MFPCA` [17].

This brings us to our approach for analysis of functional method comparison data. In Section 2, we begin by writing a subject's observed curve from a measurement method as a sum of an unobservable true smooth curve and a random measurement error. Each measurement method has its own mean and covariance functions and error variance. Next, the method-specific true curves are represented via a multivariate Karhunen-Loève expansion. In Section 3, we consider two approaches for estimating the unknowns in the model. The first approach—termed MPACE—directly adapts the PACE methodology to deal with multivariate data along the lines of Chiou et al. [4]. The second approach—termed UPACE—adapts the methodology of Happ and Greven [5]. Bootstrap is used to construct relevant confidence intervals and bands. In Section 4, we discuss evaluation of similarity and agreement under the assumed model. Section 5 presents a simulation study to evaluate properties of the two estimation approaches. The body fat data are analyzed in Section 6. Section 7 concludes

with a discussion. Appendix A contains some technical details. An analysis of the body temperature data and additional simulation results are presented in the online Supplemental Material, which can be accessed from the journal website.

2 Modeling of Data

Let the random function X_j denote the true unobservable curve measured using method $j = 1, \ldots, J\,(\geq 2)$ for a randomly selected subject from the population of interest. The curves are defined on a common domain $\mathcal{T} = [a, b]$, $a < b \in \mathbb{R}$. Let the mean and covariance functions of the random functions be denoted by

$$\mu_j(t) = E(X_j(t)),\ G_{jl}(s,t) = \text{cov}(X_j(s), X_l(t)), \quad j, l = 1, \ldots, J;\ s, t \in \mathcal{T}.$$

Let $\mathbf{X} = (X_1, \ldots, X_J)^T$ denote the $J \times 1$ vector of the curves and $\boldsymbol{\mu}(t) = (\mu_1(t), \ldots, \mu_J(t))^T$ be the $J \times 1$ vector of its mean.

2.1 Model for Population Curves

Under certain conditions [4, 5], the multivariate Karhunen-Loève Theorem provides a stochastic representation of $\mathbf{X}$ as

$$\mathbf{X}(t) = \boldsymbol{\mu}(t) + \sum_{k=1}^{\infty} \xi_k \boldsymbol{\phi}_k(t), \quad t \in \mathcal{T}. \tag{1}$$

Here, $\boldsymbol{\phi}_k(t) = (\phi_{k1}(t), \ldots, \phi_{kJ}(t))^T$ are orthonormal eigenfunctions, satisfying the property that the inner product of $\boldsymbol{\phi}_k$ and $\boldsymbol{\phi}_l$, given as $\sum_{j=1}^{J} \int_{\mathcal{T}} \phi_{kj}(t)\phi_{lj}(t)dt$, equals zero if $k \neq l$ and one if $k = l$; and ξ_k—called "scores"—are uncorrelated random variables with mean zero and variance λ_k. The variances λ_k are eigenvalues associated with the eigenfunctions $\boldsymbol{\phi}_k$ and are non-increasing, i.e., $\lambda_1 \geq \lambda_2 \geq \ldots \geq 0$. We can write (1) as

$$X_j(t) = \mu_j(t) + \sum_{k=1}^{\infty} \xi_k \phi_{kj}(t), \quad j = 1, \ldots, J;\ t \in \mathcal{T}. \tag{2}$$

Thus, the Karhunen-Loève representation provides a basis expansion of the curve X_j in terms of the basis functions $\phi_{1j}, \phi_{2j}, \ldots$ that depend on method j, whereas the random coefficients $\xi_1, \xi_2, \ldots$ are common to all methods. It is these coefficients that induce dependence within and between the curves. In particular, under (2), the covariance functions can be written as

$$G_{jl}(s,t) = \sum_{k=1}^{\infty} \lambda_k \phi_{kj}(s)\phi_{kl}(t), \quad j,l = 1,\ldots,J;\ s,t \in \mathcal{T}. \tag{3}$$

The true curves $X_j(t)$ are observed with error as $Y_j(t) = X_j(t) + \epsilon_j(t)$, where the errors $\epsilon_j(t)$ are independent random variables with mean zero and variance τ_j^2, $j = 1,\ldots,J$, and are independent of the true values. Using (2), we can write this model as

$$Y_j(t) = \mu_j(t) + \sum_{k=1}^{\infty} \xi_k \phi_{kj}(t) + \epsilon_j(t), \quad j = 1,\ldots,J;\ t \in \mathcal{T}. \tag{4}$$

Thus, the mean and autocovariance functions of the observed curves are

$$E(Y_j(t)) = \mu_j(t),\ \operatorname{cov}(Y_j(s), Y_j(t)) = G_{jj}(s,t) + \tau_j^2 I(s=t), \quad j = 1,\ldots J, \tag{5}$$

and their cross covariance function is $\operatorname{cov}(Y_j(s), Y_l(t)) = G_{jl}(s,t)$, $j \neq l = 1,\ldots,J$. Here I is the indicator function. It follows that, for each $t \in \mathcal{T}$, the vector $(Y_1(t),\ldots,Y_J(t))$ has a J-variate distribution with mean $(\mu_1(t),\ldots,\mu_J(t))$, variance $(\sigma_1^2(t),\ldots,\sigma_J^2(t))$, and correlation $\rho_{jl}(t)$, where

$$\sigma_j^2(t) = G_{jj}(t,t) + \tau_j^2, \quad \rho_{jl}(t) = \frac{G_{jl}(t,t)}{\sigma_j(t)\sigma_l(t)}, \quad j \neq l = 1,\ldots,J. \tag{6}$$

Further, for $j \neq l$, the difference $D_{jl}(t) = Y_j(t) - Y_l(t)$ has a distribution with mean $\delta_{jl}(t)$ and variance $\eta_{jl}^2(t)$, where

$$\delta_{jl}(t) = \mu_j(t) - \mu_l(t),\ \eta_{jl}^2(t) = \sigma_j^2(t) + \sigma_l^2(t) - 2G_{jl}(t,t). \tag{7}$$

These distributions are used in Section 4 to get functional analogs of measures of similarity and agreement.

2.2 Model for Observed Data

Suppose there are n subjects in the study, indexed as $i = 1, \ldots, n$. The observed data consist of J curves per subject, one from each method, observed at discrete observation times. Specifically, let $Y_{ij}(t_{im})$ denote the observation from method j on subject i taken at time t_{im}, $m = 1, \ldots, N_i$, $j = 1, \ldots, J$, $i = 1, \ldots, n$. The J curves for a subject are linked in that they are observed at common observation times t_{im}, $m = 1, \ldots, N_i$. Thus, subject i contributes JN_i observations. The number of observations and the observation times need not be the same for each subject. The design is balanced if the observation times are common for all subjects and the linked observations are available at each observation time from every subject. Otherwise, the design is unbalanced. The functional data are usually said to be dense when the design is balanced and the common N_i is large, and they are said to be sparse when the design is unbalanced and N_i is small.

To obtain a model for the observed data, let $X_{ij}(t)$, $\epsilon_{ij}(t)$, $Y_{ij}(t)$, and ξ_{ik} denote the respective counterparts of the population quantities $X_j(t)$, $\epsilon_j(t)$, $Y_j(t)$, and ξ_k, given by (2) and (4), for subject i. The quantities for subject i are assumed to be independent copies of the corresponding population quantities. Thus, the model for the data can be written as

$$Y_{ij}(t_{im}) = \mu_j(t_{im}) + \sum_{k=1}^{\infty} \xi_{ik}\phi_{kj}(t_{im}) + \epsilon_{ij}(t_{im}), \quad m = 1, \ldots, N_i;\ j = 1, \ldots, J;\ i = 1, \ldots, n, \quad (8)$$

where the errors $\epsilon_{ij}(t_{im})$ are independent random variables with mean zero and variance τ_j^2. The model postulates that a subject's true curve from a method is an infinite linear combination of method-specific basis functions that are common to all subjects but with subject-specific coefficients that are common to all methods. The eigenfunctions $\phi_{kj}(t)$ serve as the basis functions and the scores ξ_{ik} serve as the coefficients.

To analyze these data, first we perform a dimension reduction by truncating the infinite

sum in (8) to K terms, where K is the number of FPC to be selected. This leads to

$$Y_{ij}(t_{im}) \approx \mu_j(t_{im}) + \sum_{k=1}^{K} \xi_{ik}\phi_{kj}(t_{im}) + \epsilon_{ij}(t_{im}) \tag{9}$$

as the approximate model. It has the structure of a mixed-effects model. The true model (4) is used to define the parameters and their functions that are the target of inference. But they are estimated by fitting this approximate model to the data. The number of components K is treated as an unknown component in the model. The issue of estimation of unknowns is taken up in Section 3.

To write (9) in the matrix notation, define the $N_i \times 1$ vectors $\mathbf{t}_i = (t_{i1}, \ldots, t_{iN_i})^T$, $\mathbf{Y}_{ij}(\mathbf{t}_i) = (Y_{ij}(t_{i1}), \ldots, Y_{ij}(t_{iN_i}))^T$, $\boldsymbol{\mu}_j(\mathbf{t}_i) = (\mu_j(t_{i1}), \ldots, \mu_j(t_{iN_i}))^T$, and $\boldsymbol{\epsilon}_{ij}(\mathbf{t}_i) = (\epsilon_{ij}(t_{i1}), \ldots, \epsilon_{ij}(t_{iN_i}))^T$. These respectively represent the vectors of the observation times for subject i, the corresponding observations from method j, their means, and the associated random errors. Next, define the $JN_i \times 1$ vectors

$$\mathbf{Y}_i(\mathbf{t}_i) = \begin{pmatrix} \mathbf{Y}_{i1}(\mathbf{t}_i) \\ \vdots \\ \mathbf{Y}_{iJ}(\mathbf{t}_i) \end{pmatrix}, \ \boldsymbol{\mu}(\mathbf{t}_i) = \begin{pmatrix} \boldsymbol{\mu}_1(\mathbf{t}_i) \\ \vdots \\ \boldsymbol{\mu}_J(\mathbf{t}_i) \end{pmatrix}, \ \boldsymbol{\epsilon}_i(\mathbf{t}_i) = \begin{pmatrix} \boldsymbol{\epsilon}_{i1}(\mathbf{t}_i) \\ \vdots \\ \boldsymbol{\epsilon}_{iJ}(\mathbf{t}_i) \end{pmatrix}, \tag{10}$$

and take the $JN_i \times JN_i$ diagonal matrix $\mathbf{R}_i = \text{diag}\{\tau_1^2, \ldots, \tau_1^2, \ldots, \tau_J^2, \ldots, \tau_J^2\}$, where τ_j^2 is repeated N_i times for each j, as the covariance matrix of $\boldsymbol{\epsilon}_i(\mathbf{t}_i)$. Further, define $\boldsymbol{\xi}_i = (\xi_{i1}, \ldots, \xi_{iK})^T$ as the $K \times 1$ vector of scores and $\boldsymbol{\Lambda} = \text{diag}\{\lambda_1, \ldots, \lambda_K\}$ as its $K \times K$ diagonal covariance matrix; $\boldsymbol{\phi}_{kj}(\mathbf{t}_i) = (\phi_{kj}(t_{i1}), \ldots, \phi_{kj}(t_{iN_i}))^T$ as the $N_i \times 1$ vector of values of kth eigenfunction ϕ_{kj} associated with method j; $\boldsymbol{\phi}_k(\mathbf{t}_i) = (\boldsymbol{\phi}_{k1}^T(\mathbf{t}_i), \ldots, \boldsymbol{\phi}_{kJ}^T(\mathbf{t}_i))^T$ as the $JN_i \times 1$ vector by stacking the values for all the methods; and $\boldsymbol{\Phi}(\mathbf{t}_i) = (\boldsymbol{\phi}_1(\mathbf{t}_i), \ldots, \boldsymbol{\phi}_K(\mathbf{t}_i))$ as their $JN_i \times K$ matrix.

With this notation, the model (9) can be written as

$$\mathbf{Y}_i(\mathbf{t}_i) \approx \boldsymbol{\mu}(\mathbf{t}_i) + \boldsymbol{\Phi}(\mathbf{t}_i)\boldsymbol{\xi}_i + \boldsymbol{\epsilon}_i(\mathbf{t}_i), \quad i = 1, \ldots, n. \tag{11}$$

Here the $\boldsymbol{\xi}_i$ follow independent distributions with mean $\mathbf{0}$ and covariance matrix $\boldsymbol{\Lambda}$, $\boldsymbol{\epsilon}_i(\mathbf{t}_i)$ follow independent distributions with mean $\mathbf{0}$ and covariance matrix $\mathbf{R}_i$, and the two vectors are mutually independent. It follows that

$$E(\mathbf{Y}_i(\mathbf{t}_i)) \approx \boldsymbol{\mu}(\mathbf{t}_i),\ \text{var}(\mathbf{Y}_i(\mathbf{t}_i)) \approx \boldsymbol{\Phi}(\mathbf{t}_i)\boldsymbol{\Lambda}\boldsymbol{\Phi}^T(\mathbf{t}_i) + \mathbf{R}_i. \tag{12}$$

The elements of the first term of $\text{var}(\mathbf{Y}_i(\mathbf{t}_i))$ consist of values of covariance functions given by (3) but with the infinite sums therein truncated to K terms. The unknowns in the model are

$$\boldsymbol{\theta} = \{\mu_1, \ldots, \mu_J, K, \lambda_1, \ldots, \lambda_K, \phi_{11}, \ldots, \phi_{J1}, \ldots, \phi_{1K}, \ldots, \phi_{JK}, \tau_1^2, \ldots, \tau_J^2\}.$$

The mean functions and eigenfunctions in $\boldsymbol{\theta}$ depend on $t \in \mathcal{T}$ as well but this dependency is suppressed for convenience. Next, we discuss estimation of $\boldsymbol{\theta}$ to get the plug-in estimator $\hat{\boldsymbol{\theta}}$.

3 Parameter Estimation and FPCA

3.1 Parameter Estimation

Let N_0 be the number of unique observation times in the data and $\mathbf{t}_0 = (t_{01}, \ldots, t_{0N_0})^T$ be the $N_0 \times 1$ vector of these times in increasing order. The elements of $\mathbf{t}_0$ form a grid in the domain $\mathcal{T}$. By definition, there is at least one observation from all measurement methods at each time in $\mathbf{t}_0$. For estimation, we begin by pooling observations from each method on all subjects and smoothing them ignoring the within-subject dependence. Separate smoothing is performed for each method. This results in a smooth estimate $\hat{\mu}_j(t)$ of $\mu_j(t)$, $j = 1, \ldots, J$. Then, each observation in the data is centered by subtracting off the corresponding estimated mean as $\tilde{Y}_{ij}(t_{im}) = Y_{ij}(t_{im}) - \hat{\mu}_j(t_{im})$. These centered observations are used to form $JN_0 \times 1$ vectors $\tilde{\mathbf{Y}}_i(\mathbf{t}_0)$ in the same way as $\mathbf{Y}_i(\mathbf{t}_i)$ are formed in (10). If the subject i does not have an observation for some $t \in \mathbf{t}_0$, that observation is set to be missing in $\tilde{\mathbf{Y}}_i(\mathbf{t}_0)$.

In Appendix A, we describe the two approaches—MPACE and UPACE—for estimating the remaining unknown components of $\boldsymbol{\theta}$ and the multivariate scores $\boldsymbol{\xi}$ in the model (11). Both use the centered data as inputs and involve the PACE methodology of Yao et al. [14] for univariate functional data. MPACE directly adapts the PACE methodology to deal with multivariate data along the lines of Chiou et al. [4], whereas UPACE adapts the approach of Happ and Greven [5]. UPACE is computationally simpler of the two as it involves first applying the univariate PACE methodology separately to each component of the multivariate data and then processing the results. However, this may result in loss of efficiency in estimates, especially of the error variances τ_j^2 because they also come from univariate analyses rather than a multivariate analysis as in MPACE. Although the smoothing needed in these approaches and also for estimating the mean functions is performed here using `gam` function in `R` package `mgcv` [19], any other smoothing technique—e.g., local linear regression as in Yao et al. [14] and Chiou et al. [4]—can also be used without affecting the general methodology. Upon model fitting, the fitted curves are $\hat{\mathbf{Y}}_i(\mathbf{t}_i) = \hat{\boldsymbol{\mu}}_i(\mathbf{t}_i) + \hat{\boldsymbol{\Phi}}(\mathbf{t}_i)\hat{\boldsymbol{\xi}}_i$, $i = 1, \ldots, n$.

3.2 Confidence Intervals and Bands

Suppose $\psi \equiv \psi(\boldsymbol{\theta})$ is a function of model parameters of interest. Examples of ψ include the precision ratio τ_1^2/τ_2^2. Often, the parameter function depends on t, i.e., it has the form $\psi(t) \equiv \psi(t, \boldsymbol{\theta})$, $t \in \mathcal{T}$. Examples of $\psi(t)$ include the mean difference $\delta_{jl}(t)$ and the agreement measures defined in next section. Since ψ can be considered a special case of $\psi(t)$, we focus on constructing one- and two-sided confidence bands for $\psi(t)$. In effect, we construct pointwise and simultaneous intervals on a relatively fine grid $\mathbf{t}$ of L points in $\mathcal{T}$, say, $t_1, \ldots, t_L$. This grid may be the same as the grid $\mathbf{t}_0$ formed by the observed time points, used for estimation in Section 3. Or it may consist of a subset of these time points. In practice, $L \in [25, 50]$ is often adequate.

Let $\hat{\psi}(t) \equiv \psi(t, \hat{\boldsymbol{\theta}})$ be the plug-in estimator of $\psi(t)$. Also, let $\hat{\boldsymbol{\psi}}(\mathbf{t})$ and $\boldsymbol{\psi}(\mathbf{t})$ be $L \times 1$

vectors representing the values of the two functions evaluated at the elements of $\mathbf{t}$. When n is large, the joint distribution of $\hat{\boldsymbol{\psi}}(\mathbf{t}) - \boldsymbol{\psi}(\mathbf{t})$ can be approximated by a $\mathcal{N}_L(\mathbf{b}, \mathbf{S})$ distribution, possibly after a applying normalizing transformation, where the $L \times 1$ vector $\mathbf{b} = (b_1, \ldots, b_L)^T$ and the $L \times L$ matrix $\mathbf{S} = (s_{jk})_{j,k=1,\ldots,L}$ respectively represent the estimated bias vector and covariance matrix of the estimators. Once $\mathbf{b}$ and $\mathbf{S}$ are available, an approximate $100(1-\alpha)\%$ one- or two-sided pointwise confidence band for $\psi(t)$, $t \in \mathcal{T}$ can be computed as

$$\begin{aligned} &\text{lower band: } \hat{\psi}(t_l) - b_l - z_{1-\alpha}\sqrt{s_{ll}}, \quad \text{upper band: } \hat{\psi}(t_l) - b_l + z_{1-\alpha}\sqrt{s_{ll}}, \\ &\text{two-sided band: } \hat{\psi}(t_l) - b_l \pm z_{1-\alpha/2}\sqrt{s_{ll}}, \quad l = 1, \ldots, L, \end{aligned} \tag{13}$$

where z_α is the 100αth percentile of a $\mathcal{N}_1(0,1)$ distribution. A simultaneous band can be constructed by replacing z_α in (13) by an appropriate percentile [6, Chapter 3] that can be computed using the `multcomp` package of Hothorn et al. [20] in `R` or via simulation as we do here. We now present a bootstrap methodology to compute $\mathbf{b}$ and $\mathbf{S}$. It has the following steps:

1. Sample n indices with replacement from the integers $1, \ldots, n$. Take the observed curves associated with the sampled subject indices as a resample of the original data.

2. Apply the estimation and FPCA approach described in Appendix A to estimate $\boldsymbol{\theta}$ from the resampled data to get $\hat{\boldsymbol{\theta}}^*$.

3. Use $\hat{\boldsymbol{\theta}}^*$ to estimate $\boldsymbol{\psi}(\mathbf{t})$ as $\hat{\boldsymbol{\psi}}^*(\mathbf{t})$. This $\hat{\boldsymbol{\psi}}^*(\mathbf{t})$ is a resample of $\hat{\boldsymbol{\psi}}(\mathbf{t})$.

4. Repeat the previous steps Q times to get the resamples $\hat{\boldsymbol{\psi}}_q^*(\mathbf{t})$, $q = 1, \ldots, Q$. Compute the bias vector $\mathbf{b}$ as $\sum_{q=1}^{Q} \hat{\boldsymbol{\psi}}_q^*(\mathbf{t})/Q - \hat{\boldsymbol{\psi}}(\mathbf{t})$, and the covariance matrix $\mathbf{S}$ as the sample covariance matrix of the resamples.

In practice, $Q = 500$ is often enough to estimate $\mathbf{b}$ and $\mathbf{S}$. If there is evidence that a bias correction is not needed, then the term b_l in (13) can be dropped (see Section 5 for an

example). Note that a separate FPCA is performed in each bootstrap repetition. Therefore, the resulting confidence intervals also account for the uncertainty due to FPC decomposition in addition to the usual uncertainty due to sampling [15]. The procedure of this subsection can be easily adapted to construct confidence interval for a parameter function ψ that does not depend on t.

4 Evaluation of Similarity and Agreement

We now focus on how to evaluate similarity and agreement of a pair of measurement methods j and l, $j \neq l = 1, \ldots, J$. This evaluation can be repeated for all such pairs of interest. For similarity evaluation, inference is performed on two measures of similarity—difference in means of the methods and ratio of their precisions [13]. Under the true model (4), $\delta_{jl}(t)$ given by (7) is the mean difference and τ_j^2/τ_l^2 is the precision ratio.

For agreement evaluation, inference is performed on functional analogs of agreement measures originally developed for scalar data. These are obtained by using the definitions of the measures under the bivariate distribution of $(Y_j(t), Y_l(t))$ induced by the true model (4) for each $t \in \mathcal{T}$.

We specifically consider two agreement measures. One is the concordance correlation coefficient (CCC) due to Lin [11]. It is defined in terms of first and second order moments of the paired observations. Using (5) and (6), the functional CCC can be expressed as

$$\mathrm{CCC}_{jl}(t) = \frac{2G_{jl}(t,t)}{\{\mu_j(t) - \mu_l(t)\}^2 + \sigma_j^2(t) + \sigma_l^2(t)} = \rho_{jl}(t) \frac{2}{\frac{\{\mu_j(t)-\mu_l(t)\}^2}{\sigma_j(t)\sigma_l(t)} + \frac{\sigma_j(t)}{\sigma_l(t)} + \frac{\sigma_l(t)}{\sigma_j(t)}}. \tag{14}$$

See Lin [11] for properties of a CCC. Here we just note that $|\mathrm{CCC}_{jl}(t)| \leq |\rho_{jl}(t)| \leq 1$ and $\mathrm{CCC}_{jl}(t) = \rho_{jl}(t)$ if $\mu_j(t) = \mu_l(t)$ and $\sigma_j^2(t) = \sigma_l^2(t)$. A large positive value for CCC implies good agreement. The methods j and l have perfect agreement when $\mathrm{CCC}_{jl}(t) = 1$ for all t.

The other measure is the total deviation index (TDI) due to Lin [12]. For a given large probability p_0, it is defined as the p_0th percentile of absolute difference in the paired

observations. For inference on TDI, we additionally assume that the scores and the errors in the models (4) and (9) follow normal distributions. Under this assumption, for each $t \in \mathcal{T}$, the difference $D_{jl}(t)$ follows a normal distribution with mean $\delta_{jl}(t)$ and variance $\eta^2_{jl}(t)$, given by (7). This implies that the functional TDI can be expressed as

$$\mathrm{TDI}_{jl}(p_0, t) = 100p_0\text{th percentile of } |D_{jl}(t)| = \eta_{jl}(t) \left\{ \chi^2_{1,p_0} \left(\frac{\delta^2_{jl}(t)}{\eta^2_{jl}(t)} \right) \right\}^{1/2}, \qquad (15)$$

where $\chi^2_{1,p_0}(\Delta)$ is the $100p_0$th percentile of a noncentral χ^2 distribution with one degree of freedom and noncentrality parameter Δ. A TDI is non-negative, and its small value implies good agreement. Agreement between methods j and l is perfect when $\mathrm{TDI}_{jl}(t) = 0$ for all t.

The measures of similarity and agreement are estimated by plug-in. Similarity of the methods is evaluated by examining a two-sided confidence band for $\delta_{jl}(t)$ and a two-sided confidence interval for τ^2_j/τ^2_l. Agreement between the methods is evaluated by examining appropriate one-sided confidence bands for agreement measures. Since a large value for CCC and a small value for TDI imply good agreement, an upper confidence band for CCC and a lower confidence band for TDI are appropriate. The construction of confidence intervals and bands was discussed in the previous subsection. To improve accuracy, the intervals for precision ratio and TDI are obtained by first applying a log transformation and those for CCC are obtained by first applying the Fisher's z-transformation. The results are then transformed back to the original scale.

As mentioned in Section 1, the limits of agreement approach of Bland and Altman [7] is quite popular in the biomedical literature for agreement evaluation. This involves, under the normality assumption for the differences, computing estimated mean $\pm$ 1.96 times the estimated standard deviation of the differences, and examining whether the limits contain any unacceptably large differences. Using (7), the functional limits of agreement are $\hat{\delta}_{jl}(t) \pm 1.96\,\hat{\eta}_{jl}(t)$, $t \in \mathcal{T}$.

5 Simulation Study

In this section, we use Monte Carlo simulation to evaluate performance of point and interval estimators of key parameters and parameter functions, including measures of similarity and agreement, provided by the MPACE and UPACE approaches. This investigation focuses on $J = 2$ measurement methods and takes mean squared error (MSE) of a point estimator and coverage probability of a confidence interval as the measure of accuracy. The data are simulated from the true model (4) along the lines of our real data examples by taking the domain as $\mathcal{T} = [0, 1]$; assuming normality for scores and errors; taking the mean functions of the two methods as $\mu_1(t) = 24 + t$ and $\mu_2(t) = 23 + 2t$; and setting the eigenvalues as $\lambda_k = 100 \times e^{-(k-1)/2}$ for $k \leq 6$ and zero for $k > 6$. The eigenfunctions corresponding to the non-zero eigenvalues are taken as the eigenfunctions estimated from the body temperature data by restricting them to the selected domain $\mathcal{T}$. The grid $\mathbf{t}_{\text{grid}} = \{u : u = 0, 1/49, \ldots, 1\}$ of 50 equally-spaced points between 0 and 1 is used for simulating data as well as point and interval estimation.

We consider a total of four dense and sparse designs. In the dense case, a balanced design with $N_i = 50$ is considered. The observation times in this case are all points on $\mathbf{t}_{\text{grid}}$, and all subjects have the same observation times. In the sparse case, three scenarios with increasing sparsity are considered. Two are balanced designs with $N_i = 30$ and $N_i = 20$, and the third is an unbalanced design with N_i distributed as a Poisson random variable with mean 20. We refer to these four designs as (a), (b), (c), and (d), respectively. The observation times in the sparse cases are drawn from a uniform distribution on $\mathbf{t}_{\text{grid}}$ separately for each subject. Consequently, in the sparse case, the subjects may not have the same observation times. In all the four designs, observations from both measurement methods are simulated at each observation time, ensuring paired data. The observations for different subjects are independent. Three combinations of values are chosen for the error variances of the methods,

namely, $(\tau_1^2, \tau_2^2) = (2, 2)$, $(2, 4)$, and $(4, 4)$, to allow a range of practical scenarios. Three values are chosen for the number of subjects, $n \in \{50, 100, 200\}$. Further, as is common in practice, $p_0 = 0.90$ is taken for TDI and $1 - \alpha = 0.95$ is taken for the confidence intervals and bands. Thus, we consider a total of $4 \times 3 \times 3 = 36$ settings.

For each setting, we simulate a dataset, perform parameter estimation as described in Section 3.1 and Appendix A, and construct 95% confidence intervals and bands as described in Sections 3.2 and 4. The proportion of variation explained that is needed for FPCA is taken to be 0.99 for both MPACE and UPACE. For the smoothing involved in point estimation, `gam` function in `mgcv` package of Wood [19] is used with default settings. For interval estimation, $Q = 250$ bootstrap resamples are used. The entire process from data simulation to interval estimation is repeated 300 times. The results are used to compute estimated MSEs of point estimators of $\log(\tau_1^2)$, $\log(\tau_2^2)$, $\log(\tau_2^2/\tau_1^2)$, $\log\{\text{TDI}(p_0, t)\}$ and $z\{\text{CCC}(t)\}$, with $z(\cdot)$ denoting the Fisher's z-transformation, and estimated coverage probabilities for confidence intervals of these quantities. The coverage probabilities are also computed for $\mu_1(t)$, $\mu_2(t)$, and $\delta(t)$ but these quantities are excluded from the MSE calculation as both MPACE and UPACE use the same point estimators for them. We additionally compute estimated MSE of $\hat{K}$ and estimates of

$$E\left\{\frac{1}{n}\sum_{i=1}^{n}\sum_{k=1}^{\min\{\hat{K},K\}}(\hat{\xi}_{ik}-\xi_{ik})^2\right\} \text{ and } E\left\{\frac{1}{n}\sum_{i=1}^{n}(\hat{Y}_i(t)-Y_i(t))^2\right\},$$

which provide an overall measure of accuracy in prediction of scores and individual curves, respectively. For convenience, these measures are also referred to as MSE. The efficiency of MPACE relative to UPACE is measured by dividing the MSE in case of UPACE by its MPACE counterpart. From a practical viewpoint, if a relative efficiency falls between 0.9 and 1.1, we may consider the two approaches to be equally accurate for estimating that quantity. Now, a note about interval estimation of $\log\{\text{TDI}(p_0, t)\}$ is in order. Our initial simulation studies showed that its confidence band tended to be more accurate without the

bias correction. Therefore, we drop the bias term from (13) when computing the confidence band for this measure in the remainder of this article.

Table 1 presents the MSEs for the two approaches and their relative efficiencies for $(\tau_1^2, \tau_2^2) = (2, 2)$. We see that, with a few exceptions, the efficiency tends to decrease with the sparsity of design. Further, the efficiencies for the curves and scores in all cases are between 0.96 and 1.05, implying that the two approaches may be considered equally accurate for estimating them. Also, the efficiencies for K are between 0.26 and 0.83 in all cases but one. This suggests that UPACE is more accurate than MPACE for estimation of K. Additional investigation shows that MPACE tends to overestimate K. All the efficiencies for z(CCC) are greater than one, implying superiority of MPACE over UPACE. For the remaining quantities, the efficiencies depend on n and sparsity of design. In particular, for dense data (Design (a)), the efficiencies range between 0.96 and 1.20, indicating superiority of MPACE. However, as the level of sparsity increases, MPACE begins to lose its efficiency advantage to UPACE, especially when $n = 50$. But then the advantage of UPACE also shrinks as n increases. For example, for Design (d), the efficiencies range between 0.84 and 0.96 when $n = 50$, clearly indicating superiority of UPACE, but the range becomes 0.98 to 1.03 when $n = 200$, indicating nearly the same efficiency of the two approaches. Qualitatively similar conclusions hold in case of $(\tau_1^2, \tau_2^2) = (4, 4)$ (see Table 2) and also $(2, 4)$, the results for which are omitted. On the whole, these findings indicate that MPACE may be considered slightly more efficient than UPACE for dense data but the converse is true for sparse data with small n. In the other cases, the two may be considered more or less equally efficient. These conclusions remain unaffected by the error variances.

Next, we examine estimated coverage probabilities of the confidence intervals. Table 3 presents the coverage probabilities for confidence intervals of error variances and their ratio, which are free of t. With a few exceptions, the entries are 1-2% higher than the nominal level of 95%, suggesting the intervals are slightly conservative. Both MPACE and UPACE

appear equally accurate and there is little impact of n or the error variances.

For parameter functions that depend on t, Table 4 presents averages of estimated pointwise coverage probabilities of the confidence bands. There is no difference in the entries for μ_1, μ_2, and δ between MPACE and UPACE because both use the same estimates for them. In general, these entries are about 1% higher than 95%. For CCC, the entries are close to 95% for MPACE but about 96-98% for UPACE. For TDI, the entries are close to 95% for MPACE. This is also true for UPACE for $n \geq 100$. These conclusions hold regardless of the values of the error variances and whether the design is dense or sparse. Table 5 presents estimated simultaneous coverage probabilities of the confidence bands. With the exception of TDI, in which case the entries are below 95%, the other entries may be considered close to 95%, especially when $n \geq 100$. In case of TDI, the accuracy of MPACE improves with n and it may be considered acceptable for $n = 200$. Although the accuracy of UPACE also improves with n, but it remains quite liberal even with $n = 200$.

Taken together, our key findings based on the settings considered and their practical implications may be summarized as follows. First, the sparsity of design affects the relative performance of the two approaches in point estimation but not so much in interval estimation. However, the error variances do not seem to have much impact on the performance. Second, for both point and interval estimation, MPACE may be considered to have an edge over UPACE. Finally, we have also evaluated the two variants of MPACE and UPACE algorithms mentioned in Appendix A. However, we did not find any noticeable difference in the results from those presented here. Therefore, these are omitted. The results of an additional simulation study to evaluate the impact of non-normality is presented in online Supplemental Material.

6 Analysis of Body Fat Data

These data from Chinchilli et al. [8] consist of percentage body fat measurements taken over time on a cohort of 112 adolescent girls using skinfold calipers (method 1) and DEXA (method 2) methods. Age at visit is the time variable t here. See [8, 21–23] for more details about the dataset. Upon pre-processing the data which includes retaining only the observation times for which paired observations are available from both methods, we get a total of $2 * 654 = 1308$ observations from $n = 91$ girls. The observations range between 12.7 and 37.4. There are 56 distinct observation times on the domain $\mathcal{T} = [11.2, 16.8]$ years and their numbers per subject range between 4 and 8 with an average of 7.2.

Figure 1 presents the individual longitudinal profiles from the two methods, superimposed with their estimated mean functions (see below). The caliper mean ranges from 23.6 to 24.7, whereas the DEXA mean ranges from 21.4 to 24.3. They also behave differently over the domain. For example, the caliper mean remains essentially flat until age 14, then it decreases slightly until about age 15.5, and begins to increase thereafter. However, the DEXA mean decreases in the beginning, bottoms out around age 13, and increases thereafter with some flattening near the end. Figure 2 shows the age-specific scatterplots for ages 12 through 16. (Note that to draw these plots, the ages have been rounded to the nearest integer. Otherwise, there would be relatively few points in each plot, making it hard to discern any pattern.) The methods appear moderately correlated at these ages, with associated sample correlations $0.80, 0.73, 0.66, 0.67$, and 0.73, respectively. Also, the points do not tightly cluster around the 45° line for any age, implying that the methods do not agree well.

Our next task is to perform an FPCA of these data by fitting the model (9) using both MPACE and UPACE approaches. The smoothing is performed using `gam` function in `mgcv` package of `R` as described in the simulation section. The resulting mean functions are displayed in Figures 1 and 3. The FPCA yields the following estimates for the number of

FPC needed to explain at least 99% of variability, eigenvalues, and error variances:

$$\begin{cases} \hat{K} = 3, (\hat{\lambda}_1, \hat{\lambda}_2, \hat{\lambda}_3) = (97.33, 23.80, 6.11), (\hat{\tau}_1^2, \hat{\tau}_2^2) = (2.64, 2.32), & \text{for MPACE,} \\ \hat{K} = 4, (\hat{\lambda}_1, \hat{\lambda}_2, \hat{\lambda}_3, \hat{\lambda}_4) = (102.43, 12.28, 7.61, 1.27), (\hat{\tau}_1^2, \hat{\tau}_2^2) = (2.61, 2.34), & \text{for UPACE.} \end{cases}$$

MPACE requires one fewer FPC than UPACE but both yield similar estimates for the error variances. Figure 4 presents the estimated eigenfunctions. Ignoring a sign flip as the FPC are unique only up to a sign change, we see that the first three eigenfunctions for caliper and the first two eigenfunctions for DEXA from the two approaches are quite similar. The resulting estimates of standard deviation functions and correlation function for caliper and DEXA are displayed in Figure 3. The standard deviation functions estimated by MPACE and UPACE are somewhat similar, with the function exhibiting a decreasing trend for caliper and an increasing trend for DEXA. However, the two correlation functions seem quite different. In particular, the UPACE estimate shows a decreasing trend throughout, whereas the MPACE estimate shows an initial decreasing trend with minima at age 14, followed by an increasing trend. The latter pattern is consistent with the trend of correlation associated with the scatterplots in Figure 2. Therefore, we use MPACE for rest of the analysis here.

We now proceed as described in Section 4 to compute interval estimates for measures of similarity and agreement using $Q = 500$ bootstrap resamples. The estimate and 95% simultaneous confidence band for mean difference (caliper − DEXA) are displayed in Figure 3. The estimate increases from 1 around age 11 to about 3 around age 13, then starts to decrease and falls slightly under zero around age 15, and then increases to about 0.5 around age 17. The band lies above zero over the age interval from 11.5 to 14.5. The estimate and 95% confidence interval for precision ratio (caliper over DEXA) are 1.14 and $(0.60, 2.57)$, respectively. Taken together, these findings indicate that the methods have the same precision but their means are not the same. Hence the methods cannot be regarded as similar.

For agreement evaluation, Figure 5 presents estimates and 95% one-sided simultaneous confidence bands for CCC and TDI (with $p_0 = 0.90$) functions. A lower band for CCC and

an upper band for TDI are presented. The pattern of increase and decrease of TDI estimate broadly resembles that of the mean difference function in Figure 3. This indicates that the agreement between the methods is best in the beginning. Then, it becomes progressively worse as age increases to about age 13.5, starts to get better till about age 14.5, and gets progressively worse thereafter. The same conclusion can be reached on the basis of CCC also. The TDI upper bound ranges between 6.78 and 9.64 and the CCC lower bound ranges between 0.22 and 0.60. Based on these values, the methods cannot be considered to agree well. It is also clear from the similarity evaluation that this lack of agreement is primarily due to a difference in the means of the two methods. These conclusions are consistent with other analyses of these data reported in [8, 21, 23].

7 Summary and Discussion

To summarize, this article discusses modeling and analysis of functional data arising in a method comparison study. The methodology involves representing the data using a truncated Karhunen-Loève expansion. The unknowns in the model are estimated using two approaches—MPACE and UPACE, both adaptations of existing methods for FPCA of multivariate functional data observed with noise. Confidence intervals for measures of similarity and agreement, obtained by bootstrapping, are used to evaluate similarity and agreement of the measurement methods. A separate FPC decomposition is obtained for each bootstrap resample. Therefore, the variability due to FPC decomposition is also accounted for in the confidence intervals. Although both MPACE and UPACE often have comparable performance, there is evidence in both simulation studies and real data analysis that sometimes MPACE works better than UPACE. Here we use splines for smoothing involved in estimation. However, any other smoothing method can also be used without affecting the general methodology. No parametric assumption is required unless the inference on TDI is needed

in which case normality is assumed for scores and errors.

This article takes a multivariate FPCA approach to model the data. Given that the mixed-effects models are common for modeling univariate method comparison data [6, Chapter 3], an alternative would be to take a functional mixed-effects model approach. For example, Zhou et al. [24] use this to model dependence in paired functional variables. However, this methodology is difficult to implement, especially since no computer program is publicly available to fit their model. Although our methodology works for both dense and sparse functional data, it assumes that observations from all methods are available at each observation time. But this assumption is restrictive. For example, it does not hold for the body fat data. However, it may be possible to relax this assumption. Further research is needed to explore these directions.

Software in the form of `R` code, together with illustration and documentation, is available at `http://www.utdallas.edu/~pankaj/`.

Acknowledgements

The authors thank an anonymous reviewer and the Associate Editor for a constructive review of our manuscript. Their comments have led to an improved article. They thank Prof. Vernon Chinchilli for providing the percentage body fat data and Profs. Runze Li and Mosuk Chow for providing the core body temperature data. They also thank the Texas Advanced Computing Center at The University of Texas at Austin for providing HPC resources for conducting the simulation studies.

References

1. Ramsay JO, Silverman BW. *Functional Data Analysis*. 2nd edn. Springer, 2005.

2. Berrendero JR, Justel A, Svarc M. Principal components for multivariate functional data. *Computational Statistics and Data Analysis* 2011; **55**:2619–2634.

3. Jacques J, Preda C. Model-based clustering for multivariate functional data. *Computational Statistics and Data Analysis* 2014; **71**:92–106.

4. Chiou JM, Chen YT, Yang YF. Multivariate principal component analysis: A normalization approach. *Statistica Sinica* 2014; **24**:1571–1596.

5. Happ C, Greven S. Multivariate functional principal component analysis for data observed on different (dimensional) domains. *Journal of the American Statistical Association* 2018; **113**:649–659.

6. Choudhary PK, Nagaraja HN. *Measuring Agreement: Models, Methods, and Applications*. Wiley, 2017.

7. Bland JM, Altman DG. Statistical methods for assessing agreement between two methods of clinical measurement. *Lancet* 1986; **i**:307–310.

8. Chinchilli VM, Martel JK, Kumanyika S, Lloyd T. A weighted concordance correlation coefficient for repeated measurement designs. *Biometrics* 1996; **52**:341–353.

9. Li R, Chow M. Evaluation of reproducibility for paired functional data. *Journal of Multivariate Analysis* 2005; **93**:81–101.

10. Barnhart HX, Haber MJ, Lin LI. An overview on assessing agreement with continuous measurement. *Journal of Biopharmaceutical Statistics* 2007; **17**:529–569.

11. Lin LI. A concordance correlation coefficient to evaluate reproducibility. *Biometrics* 1989; **45**:255–268. Corrections: 2000, **56**:324-325.

12. Lin LI. Total deviation index for measuring individual agreement with applications in laboratory performance and bioequivalence. *Statistics in Medicine* 2000; **19**:255–270.

13. Dunn G. Regression models for method comparison data. *Journal of Biopharmaceutical Statistics* 2007; **17**:739–756.

14. Yao F, Müller HG, Wang JL. Functional data analysis for sparse longitudinal data. *Journal of the American Statistical Association* 2005; **100**:577–590.

15. Goldsmith J, Greven S, Crainiceanu C. Corrected confidence bands for functional data using principal components. *Biometrics* 2013; **69**:41–51.

16. Goldsmith J, Scheipl F, Huang L, Wrobel J, Gellar J, Harezlak J, McLean MW, Swihart B, Xiao L, Crainiceanu C, Reiss PT. *refund: Regression with Functional Data*, 2016. URL `https://CRAN.R-project.org/package=refund`, R package version 0.1-16.

17. Happ C. *MFPCA: Multivariate Functional Principal Component Analysis for Data Observed on Different Dimensional Domains*, 2018. URL `https://CRAN.R-project.org/package=MFPCA`, R package version 1.2.

18. R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2018. URL `http://www.R-project.org`.

19. Wood SN. *Generalized Additive Models: An Introduction with R*. 2nd edn. Chapman & Hall/CRC, 2017.

20. Hothorn T, Bretz F, Westfall P. Simultaneous inference in general parametric models. *Biometrical Journal* 2008; **50**:346–363.

21. King TS, Chinchilli VM, Carrasco JL. A repeated measures concordance correlation coefficient. *Statistics in Medicine* 2007; **26**:3095–3113.

22. Hiriote S, Chinchilli VM. Matrix-based concordance correlation coefficient for repeated measures. *Biometrics* 2011; **67**:1007–1016.

23. Rathnayake LN, Choudhary PK. Semiparametric modeling and analysis of longitudinal method comparison data. *Statistics in Medicine* 2017; **36**:2003–2015.

24. Zhou L, Huang JZ, Carroll RJ. Joint modelling of paired sparse functional data using principal components. *Biometrika* 2008; **95**:601–619.

25. Bates D, Maechler M. *Matrix: Sparse and Dense Matrix Classes and Methods*, 2017. URL `https://CRAN.R-project.org/package=Matrix`, R package version 1.2-12.

26. Higham N. Computing the nearest correlation matrix — A problem from finance. *IMA Journal of Numerical Analysis* 2002; **22**:329–343.

Appendix A. Two approaches for parameter estimation

This appendix describes two approaches based on FPCA for estimating components of $\boldsymbol{\theta}$ besides the mean functions. Estimation of mean function was discussed in Section 3.1. Both use the centered data $\tilde{\mathbf{Y}}_i(\mathbf{t}_0)$, $i = 1, \ldots, n$ as input.

A.1 Approach 1 (MPACE)

This approach is an adaptation of the PACE methodology for univariate functional data [14, 15] to deal with multivariate functional data. A similar approach has been used by Chiou et al. [4] for normalized functional data. It involves the following steps.

1. Compute the sum of products $\sum_{i=1}^{n} \tilde{\mathbf{Y}}_i(\mathbf{t}_0)\tilde{\mathbf{Y}}_i^T(\mathbf{t}_0)$ using only the non-missing observations in $\tilde{\mathbf{Y}}_i(\mathbf{t}_0)$. Divide each element of this $JN_0 \times JN_0$ matrix by the corresponding number of non-missing terms contributing to the sum. This divisor is n for a balanced design. If at least two observations are available at each $t \in \mathbf{t}_0$, we may subtract 1 from the number of non-missing terms contributing to the sum and use that as the

divisor. Denote the resulting matrix as $\mathbf{V}$. It has a block structure,

$$\mathbf{V} = \begin{pmatrix} \mathbf{V}_{11} & \dots & \mathbf{V}_{1J} \\ \vdots & \ddots & \vdots \\ \mathbf{V}_{J1} & \dots & \mathbf{V}_{JJ} \end{pmatrix}, \tag{A.1}$$

where each of the submatrices is a $N_0 \times N_0$ matrix and $\mathbf{V}_{jl} = \mathbf{V}_{lj}^T$, $j \neq l$. By construction, there is no missing entry in this matrix. The elements of $\mathbf{V}_{jj}$ provide a raw estimate of the autocovariance function $G_{jj}(s,t) + \tau_j^2 I(s = t)$ of $Y_j(t)$ for $s, t \in \mathbf{t}_0$, see (5). Likewise, the elements of $\mathbf{V}_{jl}$ provide a raw estimate of the cross covariance function $G_{jl}(s,t)$, given by (3), for $s, t \in \mathbf{t}_0$.

2. Perform bivariate smoothing of the off-diagonal elements of $\mathbf{V}_{jj}$, separately for each j, to obtain preliminary smooth estimates of the functions $G_{jj}(s,t)$. Evaluate the estimated functions at $s, t \in \mathbf{t}_0$ to get $N_0 \times N_0$ matrices $\tilde{\mathbf{V}}_{jj}$, $j = 1, \dots, J$.

3. Perform bivariate smoothing of the elements of $\mathbf{V}_{jl}$, separately for each (j,l) pair with $l > j$, to obtain preliminary smooth estimates of the functions $G_{jl}(s,t)$. Evaluate the estimated functions at $s, t \in \mathbf{t}_0$ to get $N_0 \times N_0$ matrices $\tilde{\mathbf{V}}_{jl}$, $l > j = 1, \dots, J$.

4. Compute the $JN_0 \times JN_0$ matrix $\tilde{\mathbf{V}}$, an analog of $\mathbf{V}$, by replacing $\mathbf{V}_{jj}$ on the diagonal, $\mathbf{V}_{jl}$ above the diagonal, and $\mathbf{V}_{lj}$ below the diagonal of $\mathbf{V}$ with $\tilde{\mathbf{V}}_{jj}$, $\tilde{\mathbf{V}}_{jl}$, and $\tilde{\mathbf{V}}_{jl}^T$, respectively.

5. Use a quadrature rule (e.g., the trapezoidal rule) that approximates an integral $\int_{\mathcal{T}} f(t)dt$ as $\sum_{q=1}^{N_0} w_q f(t_{0q})$, where the quadrature points $t_{01}, \dots, t_{0N_0}$ are the elements of $\mathbf{t}_0$, to get the associated weights $w_1, \dots, w_{N_0}$. Form a $JN_0 \times JN_0$ diagonal matrix $\mathbf{W}$ with the entire set of weights $(w_1, \dots, w_{N_0})$ repeated J times as the diagonal elements. Compute the $JN_0 \times JN_0$ matrix $\mathbf{U} = \mathbf{W}^{1/2}\tilde{\mathbf{V}}\mathbf{W}^{1/2}$.

6. Perform a spectral decomposition of $\mathbf{U}$ to get the eigenvalues $\hat{\lambda}_k$ and the corresponding $JN_0 \times 1$ orthogonal eigenvectors $\mathbf{u}_k$, $k = 1, \ldots, JN_0$. Replace any negative eigenvalues, which may possibly be nearly zero, with zero. Choose $\hat{K}$ as the smallest number of eigenvalues for which $\left(\sum_{k=1}^{\hat{K}} \hat{\lambda}_k / \sum_{k=1}^{JN_0} \hat{\lambda}_k\right) \geq \pi$, where π is a specified lower bound on the proportion of total variability in the observed curves to be explained by the principal components. Compute the vectors $\hat{\boldsymbol{\phi}}_k(\mathbf{t}_0) = \mathbf{W}^{-1/2}\mathbf{u}_k$ for $k = 1, \ldots, \hat{K}$. The eigenvalues $\hat{\lambda}_k$ provide the estimated score variances and the corresponding vectors $\hat{\boldsymbol{\phi}}_k(\mathbf{t}_0) = (\hat{\boldsymbol{\phi}}_{k1}^T(\mathbf{t}_0), \ldots, \hat{\boldsymbol{\phi}}_{kJ}^T(\mathbf{t}_0))^T$ provide values of the estimated eigenfunctions $\hat{\phi}_{kj}(t)$, $j = 1, \ldots, J$ for $t \in \mathbf{t}_0$.

7. Compute a revised estimate of the covariance functions in (3) as

$$\hat{G}_{jl}(s,t) = \sum_{k=1}^{\hat{K}} \hat{\lambda}_k \hat{\phi}_{kj}(s)\hat{\phi}_{kl}(t), \quad j,l = 1, \ldots, J; \ s,t \in \mathbf{t}_0. \tag{A.2}$$

8. For $j = 1, \ldots, J$, compute $\hat{\tau}_j^2$ by subtracting $\hat{G}_{jj}(t,t)$ given above from the diagonal elements of $\mathbf{V}_{jj}$—given by (A.1) and which estimate $G_{jj}(t,t) + \tau_j^2$—for $t \in \mathbf{t}_0$ and combining the differences to form a single number. One way to accomplish this is to proceed along the lines of the implementations of the PACE methodology in `R` packages `MFPCA` and `refund` [16, 17]. For this, define an interval $\mathcal{T}^* \subset \mathcal{T}$ as

$$\mathcal{T}^* = [\min\{t \in \mathbf{t}_0 : t \geq t_{01} + (t_{0N_0} - t_{01})/4\}, \max\{t \in \mathbf{t}_0 : t \leq t_{0N_0} - (t_{0N_0} - t_{01})/4\}],$$

and let $|\mathcal{T}^*|$ be its length. Also let $t_{01}^*, \ldots, t_{0Q}^*$ be those elements of $\mathbf{t}_0$ that also lie in $\mathcal{T}^*$. Corresponding to each t_{0q}^*, there is a diagonal element of matrix $\mathbf{V}_{jj}$ in (A.1), say, $v_{jj}(t_{0q}^*)$. Then, as in Step 5, let $w_1^*, \ldots, w_Q^*$ be the weights associated with a quadrature rule that takes $t_{01}^*, \ldots, t_{0Q}^*$ as the quadrature points. Finally, take

$$\hat{\tau}_j^2 = \frac{1}{|\mathcal{T}^*|} \sum_{q=1}^{Q} w_q^* \left(v_{jj}(t_{0q}^*) - \hat{G}_{jj}(t_{0q}^*, t_{0q}^*) \right),$$

provided it is positive, otherwise take it to be zero. This is the estimate we use here. An alternative is to proceed as in Goldsmith et al. [15] and use the average difference

between $v_{jj}(t)$ and $\hat{G}_{jj}(t,t)$ computed over the middle 60% of the grid $\mathbf{t}_0$. In either case, some observation times from the two ends of $\mathbf{t}_0$ are discarded to improve stability of the estimate.

9. Estimate the score vector $\boldsymbol{\xi}_i$ by its estimated best linear unbiased predictor under model (11),

$$\hat{\boldsymbol{\xi}}_i = \hat{\boldsymbol{\Lambda}}\hat{\boldsymbol{\Phi}}^T(\mathbf{t}_i)\big\{\hat{\boldsymbol{\Phi}}(\mathbf{t}_i)\hat{\boldsymbol{\Lambda}}\hat{\boldsymbol{\Phi}}^T(\mathbf{t}_i) + \hat{\mathbf{R}}_i\big\}^{-1}\{\mathbf{Y}_i(\mathbf{t}_i) - \hat{\boldsymbol{\mu}}(\mathbf{t}_i)\}.$$

   The estimated matrices here are plug-in estimates of their population counterparts.

The matrix $\tilde{\mathbf{V}}$ in Step 4 is not guaranteed to be positive definite. Therefore, we may replace it by its nearest positive definite approximation, computed using the `nearPD` function in `R` package `Matrix` [25] which implements the algorithm of Higham [26], before continuing with the rest of the steps. This variant of the algorithm is evaluated using simulation in Section 5.

A.2 Approach 2 (UPACE)

This approach is a special case of a general approach for multivariate FPCA proposed by Happ and Greven [5]. In our adaptation here, it begins with the centered data $\tilde{\mathbf{Y}}_i(\mathbf{t}_0)$, $i = 1, \ldots, n$ and involves the following steps.

1. Use the PACE methodology [14] to perform univariate FPCA of data from each measurement method separately. This effectively amounts to considering data from one measurement method at a time, assuming a model for it similar to (9) that is based on a univariate Karhunen-Loève expansion, and fitting the model by applying the algorithm of the previous section by suitably modifying it. Suppose for data from measurement method $j = 1, \ldots, J$, this results in $\hat{K}^{(j)}$ as the smallest number of principal components explaining at least a specified proportion $\pi^{(j)}$ of variability; $\hat{\boldsymbol{\phi}}_k^{(j)}(\mathbf{t}_0)$ as the $N_0 \times 1$

vector of values of the kth estimated eigenfunction for $t \in \mathbf{t}_0$, $k = 1, \ldots, \hat{K}^{(j)}$; $\hat{\tau}^{2(j)}$ as the estimated error variance; and $\hat{\boldsymbol{\xi}}_i^{(j)}$ as the $\hat{K}_j \times 1$ vector of estimated scores for the ith subject. Note that the corresponding true scores have expectation zero and we have a total of $\hat{K}^+ = \hat{K}^{(1)} + \ldots + \hat{K}^{(J)}$ estimated univariate scores for each subject. The $\hat{\tau}^{2(j)}$ resulting from this univariate FPCA also estimate the error variances in $\boldsymbol{\theta}$, i.e., $\hat{\tau}_j^2 = \hat{\tau}^{2(j)}$, $j = 1, \ldots, J$.

2. Arrange the univariate scores as a $n \times \hat{K}^+$ matrix $\hat{\boldsymbol{\Xi}}$ where

$$\hat{\boldsymbol{\Xi}} = \begin{pmatrix} \hat{\boldsymbol{\xi}}_1^{(1)T} & \ldots & \hat{\boldsymbol{\xi}}_1^{(J)T} \\ \vdots & \ldots & \vdots \\ \hat{\boldsymbol{\xi}}_n^{(1)T} & \ldots & \hat{\boldsymbol{\xi}}_n^{(J)T} \end{pmatrix}$$

and form their $\hat{K}^+ \times \hat{K}^+$ covariance matrix $\mathbf{Z} = \hat{\boldsymbol{\Xi}}^T \hat{\boldsymbol{\Xi}}/(n-1)$.

3. Perform a spectral decomposition of $\mathbf{Z}$ to get the eigenvalues $\hat{\lambda}_k$ and the corresponding $\hat{K}^+ \times 1$ orthogonal eigenvectors $\mathbf{z}_k$, $k = 1, \ldots, \hat{K}^+$. Replace any negative eigenvalues, which may possibly be nearly zero, with zero. Choose $\hat{K}$ as the smallest number of eigenvalues for which $(\sum_{k=1}^{\hat{K}} \hat{\lambda}_k / \sum_{k=1}^{\hat{K}^+} \hat{\lambda}_k) \geq \pi$, where π is a specified lower bound on the proportion of variability explained. By construction, $\hat{K} \leq \hat{K}^+$. These $\hat{K}$ and $\hat{\lambda}_k$ estimate the corresponding components K and λ_k of $\boldsymbol{\theta}$.

4. For $k = 1, \ldots, \hat{K}$, represent the $\hat{K}^+ \times 1$ eigenvector $\mathbf{z}_k$ as $\mathbf{z}_k = (\mathbf{z}_k^{(1)T}, \ldots, \mathbf{z}_k^{(J)T})^T$, where $\mathbf{z}_k^{(j)}$ is a $\hat{K}^{(j)} \times 1$ vector with elements $z_{k1}^{(j)}, \ldots, z_{k\hat{K}^{(j)}}^{(j)}$, $j = 1, \ldots, J$. Estimate the values of the eigenfunction $\hat{\phi}_{kj}(t)$ in $\boldsymbol{\theta}$ for $t \in \mathbf{t}_0$ as the $N_0 \times 1$ vector

$$\hat{\boldsymbol{\phi}}_{kj}(\mathbf{t}_0) = \big(\hat{\boldsymbol{\phi}}_1^{(j)}(\mathbf{t}_0), \ldots, \hat{\boldsymbol{\phi}}_{\hat{K}^{(j)}}^{(j)}(\mathbf{t}_0)\big)\mathbf{z}_k^{(j)}.$$

5. Take $\hat{\boldsymbol{\xi}}_i$, the estimated score vector in the model (11), as the ith row of the $n \times \hat{K}$ matrix $\hat{\boldsymbol{\Xi}}(\mathbf{z}_1, \ldots, \mathbf{z}_{\hat{K}})$.

To conclude, both MPACE and UPACE approaches provide estimates of all the components of $\boldsymbol{\theta}$ except the mean functions whose estimation was discussed in Section 3.1. They also provide estimates of the multivariate scores in the model (11). In applications, MPACE involves choosing only one proportion of variation explained—π for multivariate data. On the other hand, UPACE involves choosing $J+1$ such proportions—$\pi^{(1)}, \ldots, \pi^{(J)}$ for univariate data and π for multivariate data. In practice, such proportions are taken to be large, e.g., between 0.95 and 0.99. The specific choice will depend on the application and can be guided by a scree plot [1, Chapter 8]. We have chosen 0.99 in this work. A variant of the UPACE algorithm with $\pi^{(j)} = 1$ for $j = 1, \ldots, J$ is evaluated using simulation in Section 5.

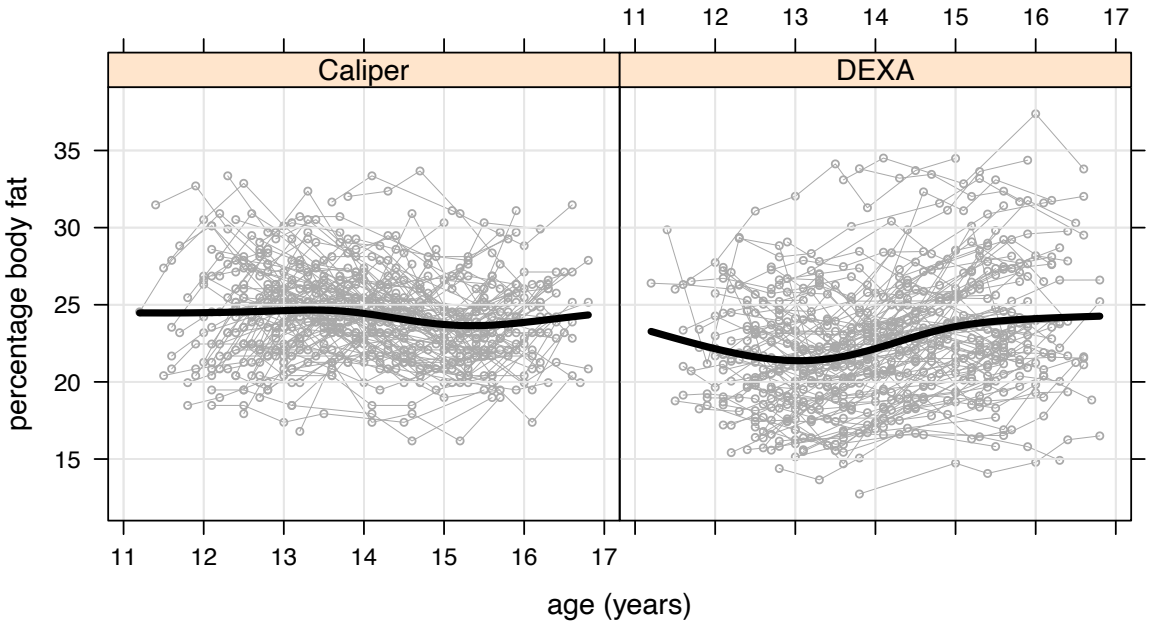


Figure 1: The individual profiles of percentage body fat measurements (in grey) for the two methods superimposed with estimated mean functions (in black).

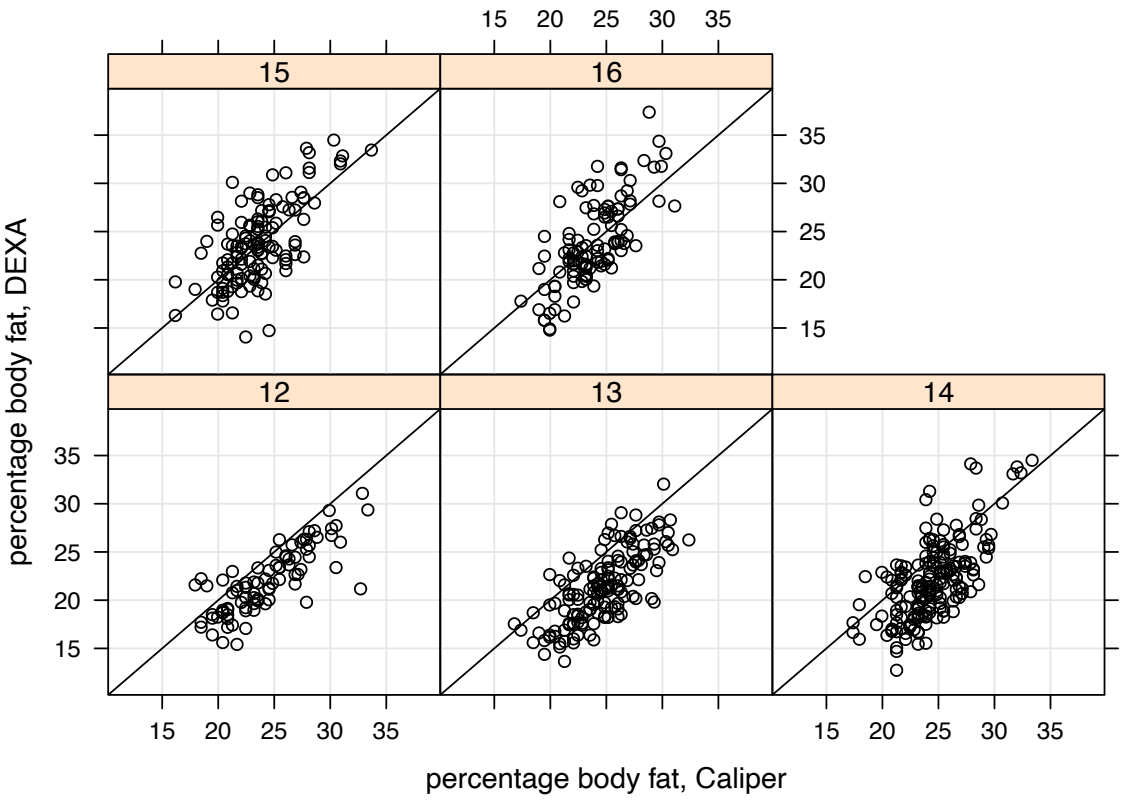


Figure 2: Scatterplots of percentage body fat measurements at ages $t = 12, 13, \ldots, 16$ years. To draw these plots, the ages have been rounded to the nearest integer.

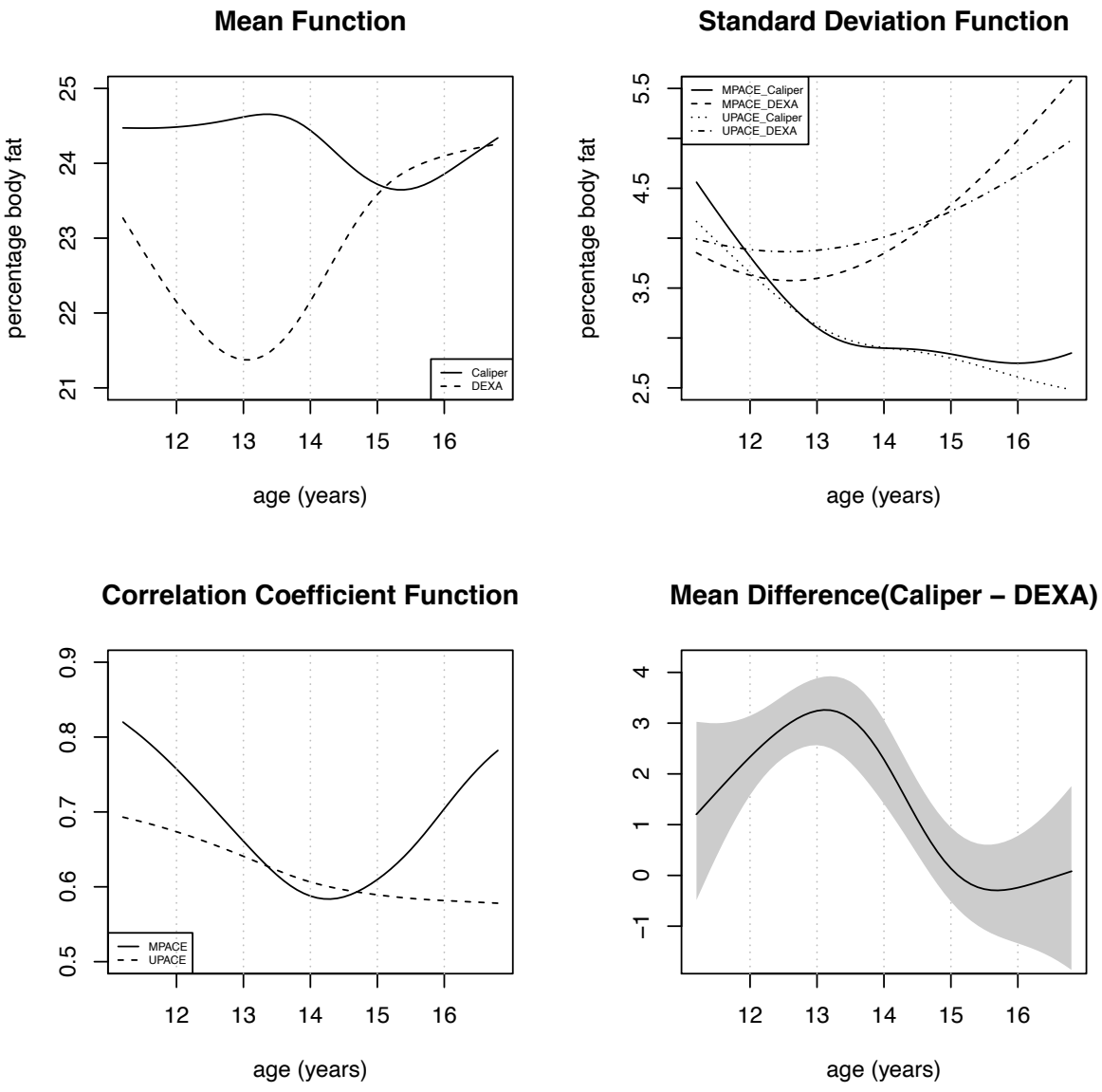


Figure 3: The estimated mean, standard deviation, correlation, and mean difference functions for the two methods for percentage body fat data. The bottom right panel also shows a 95% simultaneous confidence band for the mean difference function.

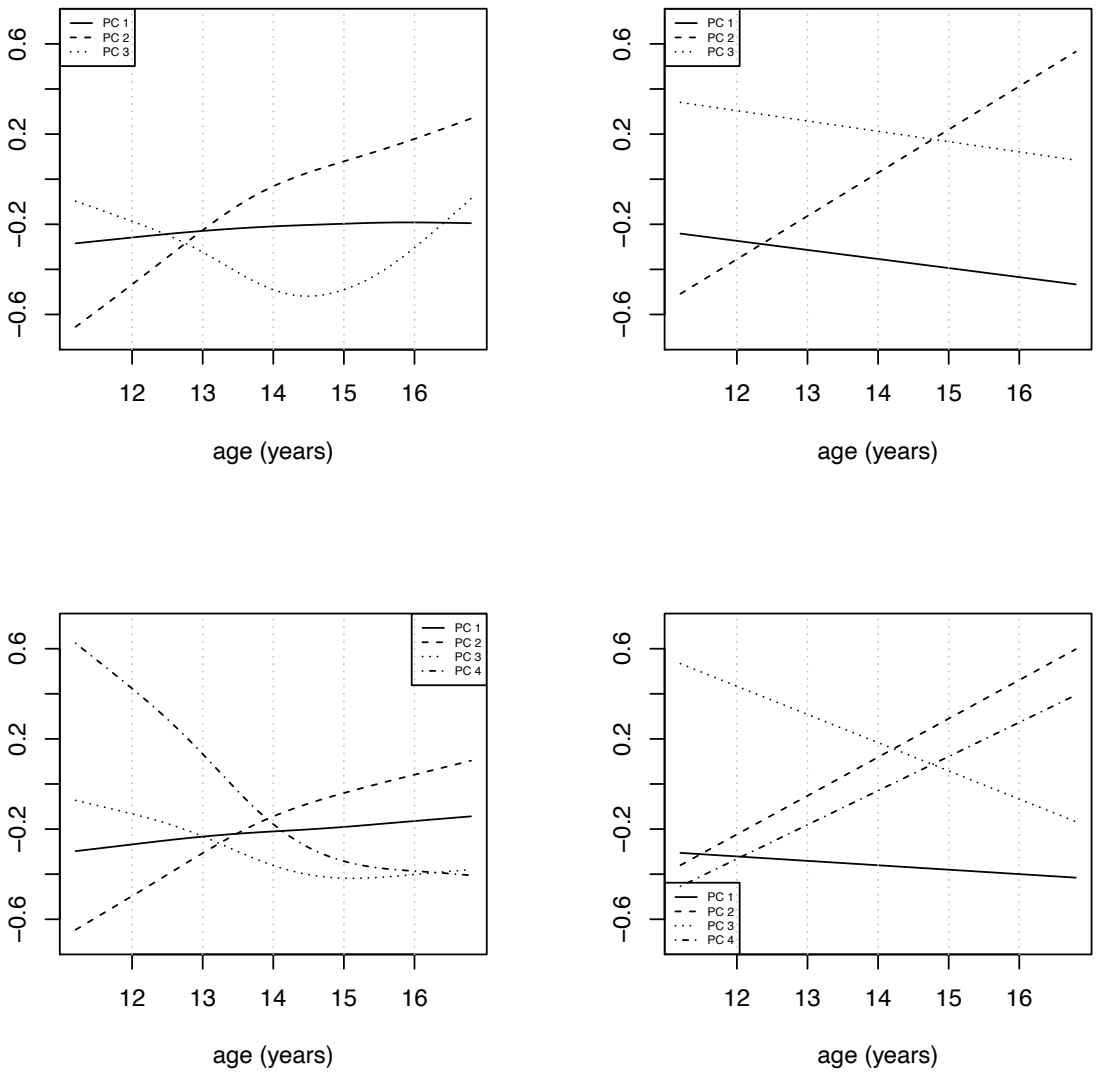


Figure 4: Estimated eigenfunctions for caliper (left panel) and DEXA (right panel) measurements using MPACE (top panel) and UPACE (bottom panel) approaches.

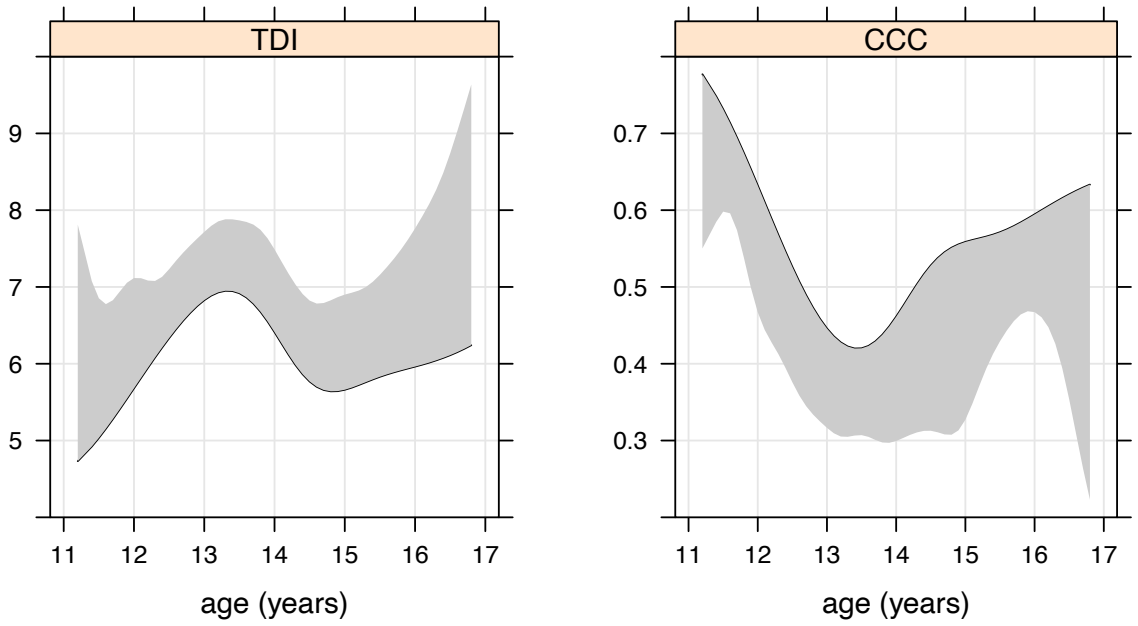


Figure 5: Estimate of TDI function with $p_0 = 0.90$ and its 95% simultaneous upper confidence band (left) and estimate of CCC function and its 95% simultaneous lower confidence band (right) using MPACE.

Table 1: MSEs of estimators of quantities that are free of t and average MSEs of estimators of quantities that depend on t, computed using MPACE (marked as M) and UPACE (marked as U) approaches, and the ratio of the MSEs (marked as U/M) in case of four designs: (a) $N_i = 50$ (dense data), (b) $N_i = 30$ (sparse data), (c) $N_i = 20$ (sparse data), and (d) unbalanced design with mean $N_i = 20$ (sparse data), each with $(\tau_1^2, \tau_2^2) = (2, 2)$.

		(a)			(b)			(c)			(d)		
n	Quantity	M*	U*	U/M	M*	U*	U/M	M*	U*	U/M	M*	U*	U/M
50	Curves	1.83	1.87	1.02	1.74	1.80	1.04	1.64	1.73	1.05	1.65	1.73	1.05
	Scores	9.08	9.01	0.99	8.80	8.80	1.00	8.71	8.78	1.01	8.73	8.76	1.00
	K	1.63	0.70	0.43	4.36	1.63	0.37	5.36	2.11	0.39	4.94	1.92	0.39
	$\log(\tau_1^2)$	1.94	1.90	0.98	6.58	6.05	0.92	15.68	13.59	0.87	15.87	14.72	0.93
	$\log(\tau_2^2)$	1.96	2.00	1.02	6.22	4.97	0.80	12.77	10.26	0.80	10.78	9.10	0.84
	$\log(\tau_1^2/\tau_2^2)$	3.88	3.94	1.02	9.91	9.96	1.01	21.76	21.30	0.98	24.58	23.52	0.96
	log(TDI)	2.53	2.43	0.96	4.07	3.59	0.88	5.09	4.36	0.86	5.11	4.13	0.81
	z(CCC)	4.01	4.48	1.12	6.04	6.44	1.07	7.09	8.43	1.19	7.82	9.51	1.22
100	Curves	1.85	1.88	1.01	1.76	1.82	1.03	1.67	1.75	1.05	1.67	1.74	1.05
	Scores	9.00	8.86	0.98	8.82	8.79	1.00	8.65	8.70	1.01	8.76	8.82	1.01
	K	1.04	0.86	0.83	4.64	1.43	0.31	6.69	2.14	0.32	6.45	2.09	0.32
	$\log(\tau_1^2)$	0.91	0.91	1.00	3.03	2.87	0.94	7.56	7.69	1.02	9.73	9.91	1.02
	$\log(\tau_2^2)$	1.05	1.07	1.02	2.67	2.40	0.90	6.06	5.52	0.91	6.49	5.64	0.87
	$\log(\tau_1^2/\tau_2^2)$	1.88	1.94	1.03	5.02	5.00	1.00	11.74	11.45	0.97	14.31	14.05	0.98
	log(TDI)	1.41	1.48	1.05	1.81	1.74	0.96	2.64	2.48	0.94	2.88	2.52	0.87
	z(CCC)	2.15	2.85	1.33	3.02	4.11	1.36	3.56	5.95	1.67	4.10	6.55	1.60
200	Curves	1.88	1.89	1.01	1.78	1.83	1.03	1.69	1.75	1.04	1.68	1.75	1.04
	Scores	9.20	8.87	0.96	8.85	8.83	1.00	8.73	8.77	1.00	8.69	8.73	1.00
	K	0.43	0.78	1.82	3.97	1.20	0.30	6.69	1.87	0.28	6.32	1.63	0.26
	$\log(\tau_1^2)$	0.51	0.52	1.02	1.57	1.61	1.02	3.38	3.50	1.03	3.62	3.75	1.03
	$\log(\tau_2^2)$	0.47	0.51	1.10	1.19	1.16	0.97	2.62	2.49	0.95	2.99	2.93	0.98
	$\log(\tau_1^2/\tau_2^2)$	1.03	1.05	1.02	2.62	2.65	1.01	5.81	5.96	1.02	5.97	5.88	0.98
	log(TDI)	0.70	0.83	1.20	1.01	1.09	1.08	1.46	1.49	1.02	1.47	1.49	1.01
	z(CCC)	1.13	1.97	1.74	1.70	3.12	1.84	2.15	4.80	2.24	1.96	4.63	2.37

*For all quantities except the curves, scores, and K, the entries represent $10^3 \times$ MSE.

Table 2: MSEs of estimators of quantities that are free of t and average MSEs of estimators of quantities that depend on t, computed using MPACE (marked as M) and UPACE (marked as U) approaches, and the ratio of the MSEs (marked as U/M) in case of four designs: (a) $N_i = 50$ (dense data), (b) $N_i = 30$ (sparse data), (c) $N_i = 20$ (sparse data), and (d) unbalanced design with mean $N_i = 20$ (sparse data), each with $(\tau_1^2, \tau_2^2) = (4, 4)$.

		(a)			(b)			(c)			(d)		
n	Quantity	M*	U*	U/M	M*	U*	U/M	M*	U*	U/M	M*	U*	U/M
50	Curves	3.66	3.74	1.02	3.48	3.60	1.03	3.33	3.49	1.05	3.34	3.50	1.05
	Scores	8.75	8.76	1.00	8.70	8.77	1.01	8.78	8.85	1.01	8.71	8.82	1.01
	K	3.90	1.43	0.37	5.76	2.17	0.38	5.82	1.88	0.32	5.39	1.79	0.33
	$\log(\tau_1^2)$	2.35	2.27	0.97	4.72	4.39	0.93	9.90	9.22	0.93	8.83	8.75	0.99
	$\log(\tau_2^2)$	1.72	1.74	1.01	5.16	4.33	0.84	8.51	7.15	0.84	10.52	8.70	0.83
	$\log(\tau_1^2/\tau_2^2)$	4.24	4.27	1.01	8.64	8.52	0.99	16.28	16.38	1.01	17.34	17.06	0.98
	log(TDI)	1.85	1.70	0.92	2.78	2.38	0.86	3.68	2.88	0.78	3.54	3.09	0.87
	z(CCC)	2.68	3.33	1.24	4.07	4.90	1.21	5.43	6.73	1.24	5.20	6.58	1.27
100	Curves	3.70	3.77	1.02	3.54	3.65	1.03	3.41	3.54	1.04	3.40	3.53	1.04
	Scores	8.82	8.81	1.00	8.81	8.83	1.00	8.66	8.71	1.01	8.67	8.73	1.01
	K	3.64	1.29	0.35	6.57	1.98	0.30	7.20	2.05	0.29	7.06	2.06	0.29
	$\log(\tau_1^2)$	1.10	1.10	1.01	1.97	2.02	1.03	4.06	4.02	0.99	4.35	4.50	1.04
	$\log(\tau_2^2)$	1.13	1.07	0.95	1.91	1.71	0.89	3.79	3.63	0.96	4.06	3.69	0.91
	$\log(\tau_1^2/\tau_2^2)$	2.20	2.23	1.01	3.17	3.20	1.01	6.89	6.89	1.00	7.38	7.27	0.98
	log(TDI)	0.91	0.90	0.99	1.30	1.18	0.91	1.97	1.66	0.84	2.01	1.74	0.87
	z(CCC)	1.37	2.20	1.61	2.04	3.82	1.87	2.91	5.93	2.04	3.07	5.80	1.89
200	Curves	3.74	3.78	1.01	3.58	3.66	1.02	3.43	3.55	1.03	3.44	3.55	1.03
	Scores	8.87	8.79	0.99	8.84	8.85	1.00	8.64	8.71	1.01	8.68	8.73	1.01
	K	3.13	1.03	0.33	6.70	1.66	0.25	8.48	2.03	0.24	7.62	1.82	0.24
	$\log(\tau_1^2)$	0.52	0.52	1.00	0.93	0.98	1.05	2.01	2.08	1.03	2.47	2.37	0.96
	$\log(\tau_2^2)$	0.53	0.53	1.01	1.30	1.16	0.90	2.01	2.09	1.04	1.70	1.65	0.97
	$\log(\tau_1^2/\tau_2^2)$	0.96	0.98	1.03	2.02	2.01	1.00	3.97	4.11	1.04	4.15	4.16	1.00
	log(TDI)	0.53	0.55	1.04	0.71	0.73	1.03	1.07	1.05	0.99	1.05	1.05	1.00
	z(CCC)	0.69	1.52	2.19	1.10	3.01	2.74	1.44	4.63	3.22	1.55	4.94	3.19

*For all quantities except the curves, scores, and K, the entries represent $10^3 \times$ MSE.

Table 3: Estimated coverage probabilities (in %) of 95% confidence intervals for error variances and their ratio in case of four designs: (a) $N_i = 50$ (dense data), (b) $N_i = 30$ (sparse data), (c) $N_i = 20$ (sparse data), and (d) unbalanced design with mean $N_i = 20$ (sparse data).

			(τ_1^2, τ_2^2)											
			$(2,2)$				$(2,4)$				$(4,4)$			
Method	n	ψ	(a)	(b)	(c)	(d)	(a)	(b)	(c)	(d)	(a)	(b)	(c)	(d)
MPACE	50	τ_1^2	97.3	94.3	95.7	97.3	95.7	96.0	96.0	93.3	95.7	97.7	95.7	97.7
		τ_2^2	97.0	96.7	96.3	97.7	96.0	97.0	97.0	97.3	97.7	97.7	97.7	95.3
		τ_1^2/τ_2^2	98.0	97.0	97.7	97.3	95.0	97.0	96.7	96.0	96.7	97.0	95.7	95.7
	100	τ_1^2	98.0	93.0	96.7	96.7	97.7	96.7	96.3	95.0	96.3	97.3	97.7	96.7
		τ_2^2	97.3	96.0	96.7	95.7	97.0	96.3	97.0	97.7	95.7	96.3	98.0	97.0
		τ_1^2/τ_2^2	97.3	96.0	96.3	97.0	96.0	97.0	96.3	96.3	95.3	95.7	97.0	97.7
	200	τ_1^2	96.3	95.3	97.3	97.3	96.3	96.0	97.0	96.0	97.0	97.3	96.0	96.3
		τ_2^2	95.3	97.3	97.0	95.3	97.3	97.0	96.7	97.0	94.3	96.7	95.3	96.7
		τ_1^2/τ_2^2	95.0	96.7	97.0	95.7	97.0	94.7	96.7	95.3	97.7	95.7	96.0	97.0
UPACE	50	τ_1^2	96.3	94.3	94.0	94.7	95.3	93.7	95.0	92.3	96.0	96.0	95.3	96.0
		τ_2^2	96.7	96.3	96.0	96.0	96.3	95.7	97.7	96.0	96.7	95.3	95.7	94.3
		τ_1^2/τ_2^2	98.0	96.7	97.0	98.0	96.0	96.0	96.0	94.7	96.7	96.3	95.7	96.3
	100	τ_1^2	97.3	94.3	96.0	94.3	97.3	94.7	96.3	93.0	96.0	97.0	97.3	94.7
		τ_2^2	97.7	94.7	94.0	95.3	97.0	95.7	97.0	97.7	95.3	95.3	98.0	94.7
		τ_1^2/τ_2^2	97.3	97.0	96.7	97.0	97.0	97.3	96.3	95.3	96.0	96.7	97.0	98.0
	200	τ_1^2	97.3	95.7	97.0	94.7	95.7	94.7	95.7	95.3	97.3	96.7	96.0	97.0
		τ_2^2	94.3	96.3	95.3	94.0	96.0	96.0	95.0	96.3	95.3	96.3	97.0	98.0
		τ_1^2/τ_2^2	93.3	97.0	96.3	96.0	97.0	94.7	95.7	95.7	98.3	95.3	97.0	96.3

Table 4: Average estimated pointwise coverage probabilities (in %) of 95% pointwise confidence bands for function that depend on t in case of four designs, (a) $N_i = 50$ (dense data), (b) $N_i = 30$ (sparse data), (c) $N_i = 20$ (sparse data) and (d) unbalanced design with mean $N_i = 20$ (sparse data).

			(τ_1^2, τ_2^2)											
			$(2,2)$				$(2,4)$				$(4,4)$			
	n	$\psi(t)$	(a)	(b)	(c)	(d)	(a)	(b)	(c)	(d)	(a)	(b)	(c)	(d)
MPACE	50	μ_1	97.1	96.8	96.5	96.3	97.3	96.6	97.3	95.3	96.6	96.0	96.1	96.6
		μ_2	96.3	95.1	96.0	96.7	96.2	96.6	97.5	96.6	96.3	96.1	96.6	97.1
		δ	96.6	94.7	96.2	96.1	96.0	95.9	96.5	95.7	96.5	95.2	95.5	96.1
		TDI	94.1	93.4	93.1	92.9	95.1	95.4	94.6	94.7	95.8	95.3	94.8	94.6
		CCC	95.4	95.6	94.9	95.3	95.5	95.2	94.2	95.2	96.1	95.0	95.1	95.4
	100	μ_1	96.0	97.2	96.4	96.7	97.3	97.3	97.5	97.2	97.2	96.9	96.9	96.4
		μ_2	96.1	96.6	96.1	96.3	97.1	97.0	97.0	97.6	96.1	96.9	97.0	96.2
		δ	96.3	95.8	96.7	96.6	96.2	97.0	96.3	96.9	97.1	96.3	96.6	95.8
		TDI	94.9	95.4	94.6	94.2	96.5	95.3	94.6	94.1	96.2	95.4	94.4	94.6
		CCC	95.9	95.6	95.5	96.4	95.5	95.8	95.7	95.3	96.0	96.5	95.9	95.8
	200	μ_1	96.3	96.9	97.2	96.8	97.4	97.2	96.8	96.8	96.3	97.2	97.7	97.3
		μ_2	95.5	96.9	96.2	97.3	96.2	97.5	96.4	97.2	96.8	97.4	97.6	97.0
		δ	95.6	96.2	95.8	96.9	95.8	96.4	96.0	96.5	95.9	96.4	96.7	96.7
		TDI	96.7	95.9	96.1	95.8	96.8	96.7	95.7	96.0	97.0	96.5	96.5	96.1
		CCC	95.6	94.9	96.1	95.9	96.0	95.8	95.8	95.2	95.8	95.5	95.1	95.9
UPACE	50	μ_1	97.1	96.8	96.5	96.3	97.3	96.6	97.3	95.2	96.6	96.0	96.1	96.6
		μ_2	96.3	95.1	96.0	96.7	96.2	96.6	97.5	96.6	96.3	96.1	95.6	97.1
		δ	96.6	94.7	96.2	96.1	96.0	95.9	96.5	95.7	96.5	95.2	95.5	96.1
		TDI	91.7	91.3	91.0	91.4	93.3	92.9	93.3	93.2	94.3	93.9	93.6	93.8
		CCC	97.1	98.7	97.8	98.5	97.7	99.0	98.0	99.0	98.5	99.1	98.8	99.3
	100	μ_1	96.0	97.2	96.4	96.7	97.3	97.3	97.5	97.2	97.2	96.9	96.9	96.4
		μ_2	96.1	96.6	96.1	96.3	97.1	97.0	97.0	97.6	96.1	96.9	97.0	96.2
		δ	96.3	95.8	96.7	96.6	96.2	97.0	96.3	96.9	97.1	96.3	96.6	95.8
		TDI	93.7	93.2	93.6	93.0	94.2	94.0	93.7	93.4	95.1	94.7	94.4	94.2
		CCC	96.5	98.4	97.8	98.8	97.3	99.0	98.3	98.7	98.2	99.5	98.8	99.5
	200	μ_1	96.3	96.9	97.2	96.8	97.4	97.2	96.8	96.8	96.3	97.2	97.7	97.3
		μ_2	95.5	96.9	96.2	97.3	96.2	97.5	96.4	97.2	96.8	97.4	97.6	97.0
		δ	95.6	96.2	95.8	96.9	95.8	96.4	96.0	96.5	95.9	96.4	96.7	96.7
		TDI	95.0	94.7	94.5	94.3	95.1	94.5	94.8	94.0	95.3	95.1	94.4	94.6
		CCC	96.4	97.6	97.4	97.9	97.3	98.8	97.8	98.2	98.3	99.2	98.0	99.2

Table 5: Estimated simultaneous coverage probabilities (in %) of 95% simultaneous confidence bands in case of four designs, (a) $N_i = 50$ (dense data), (b) $N_i = 30$ (sparse data), (c) $N_i = 20$ (sparse data) and (d) unbalanced design with mean $N_i = 20$ (sparse data).

			(τ_1^2, τ_2^2)											
			$(2,2)$				$(2,4)$				$(4,4)$			
	n	$\psi(t)$	(a)	(b)	(c)	(d)	(a)	(b)	(c)	(d)	(a)	(b)	(c)	(d)
MPACE	50	μ_1	95.7	95.0	94.3	94.3	95.3	94.7	94.7	93.7	95.0	94.7	94.0	94.0
		μ_2	95.0	94.7	94.0	93.3	95.0	93.7	96.7	94.3	95.0	94.7	93.7	95.0
		δ	94.7	94.3	94.3	93.3	94.7	94.0	94.3	93.7	94.7	94.0	94.0	93.3
		TDI	90.0	89.3	88.3	88.3	91.0	91.7	90.0	90.7	93.3	93.0	92.7	92.7
		CCC	93.0	94.7	92.0	91.7	94.0	91.0	91.0	92.7	94.7	93.0	93.7	92.7
	100	μ_1	95.7	96.0	95.3	94.3	96.0	96.7	95.3	95.7	95.7	95.0	95.3	95.3
		μ_2	96.3	94.7	94.0	94.7	94.0	94.7	96.0	95.7	94.7	95.7	94.3	93.3
		δ	95.3	94.7	94.0	93.0	94.3	94.7	93.7	93.3	94.7	93.7	93.3	93.3
		TDI	92.7	93.0	91.7	91.3	93.3	92.7	92.7	91.7	93.7	92.3	91.3	92.0
		CCC	95.0	93.3	92.0	94.7	94.3	94.0	95.3	95.3	95.3	94.3	93.7	94.3
	200	μ_1	95.3	96.7	95.7	95.3	95.7	95.3	95.0	94.0	95.3	95.0	95.0	95.3
		μ_2	95.0	96.0	94.7	94.0	95.3	95.3	94.7	95.7	95.3	94.7	94.7	94.7
		δ	94.3	95.7	94.7	93.7	95.0	95.0	94.3	94.7	95.3	94.3	94.0	94.7
		TDI	94.0	93.7	94.0	93.3	94.7	94.3	93.7	94.0	95.0	94.3	94.3	94.0
		CCC	94.3	92.7	93.0	94.3	96.0	94.0	95.7	94.7	95.3	93.3	95.3	95.3
UPACE	50	μ_1	95.7	95.0	94.0	94.3	95.3	94.3	94.7	94.0	95.0	94.7	94.0	94.0
		μ_2	95.0	94.7	94.0	93.3	95.3	94.0	96.3	94.7	94.7	94.7	93.7	95.3
		δ	94.7	94.3	94.3	93.3	94.7	94.0	94.3	93.7	95.0	94.0	94.0	93.3
		TDI	88.3	87.7	87.3	87.3	89.3	88.0	88.7	89.3	90.7	90.0	89.0	89.3
		CCC	94.3	97.0	96.0	96.7	97.0	99.0	97.3	98.7	98.7	98.7	97.7	98.3
	100	μ_1	95.7	96.0	95.3	94.3	96.3	97.0	95.3	95.7	95.3	95.3	95.3	95.7
		μ_2	95.3	95.0	94.0	94.7	94.0	94.7	96.0	95.3	94.7	95.7	94.3	93.3
		δ	95.0	94.7	94.0	92.7	94.3	94.7	93.7	93.7	95.0	93.7	93.7	93.3
		TDI	89.7	89.3	89.7	89.0	90.3	90.3	89.3	89.0	91.0	90.7	90.3	90.3
		CCC	94.0	97.3	94.3	96.3	94.7	99.3	95.3	97.7	94.7	99.3	97.7	99.7
	200	μ_1	95.3	96.3	95.7	95.3	96.0	95.3	95.3	93.7	95.7	95.0	95.0	95.3
		μ_2	94.7	95.0	94.7	94.0	95.0	95.3	94.3	95.3	95.3	94.7	94.7	94.3
		δ	94.3	94.7	94.3	93.7	95.0	95.0	94.3	94.7	95.3	94.0	94.3	94.3
		TDI	91.3	91.0	91.3	90.7	91.3	90.7	91.0	90.3	92.0	91.7	90.3	90.7
		CCC	94.3	93.3	93.0	93.7	94.7	95.7	94.3	94.3	94.7	94.7	95.7	95.0

Supplemental Material for "Modeling and Analysis of Functional Method Comparison Data" by Galappaththige S. R. de Silva, Lasitha N. Rathnayake and Pankaj K. Choudhary

1 An Additional Simulation Study

In this section, we report results of a limited simulation study to evaluate the impact of non-normality. The setup here is exactly the same as in Section 5 (Simulation Study) of the main article with the exception that the scores ξ_{ik} are now simulated as scaled log-normal variates, specifically, $\xi_{ik}/\sqrt{\lambda_k} \sim \mathcal{LN}(0,1)$, instead of $\xi_{ik}/\sqrt{\lambda_k} \sim \mathcal{N}(0,1)$, introducing skewness in the true data generating model. The model assumed for the data is same as before. This change necessitated dropping TDI from consideration as its definition is tied to normality; computing true values for mean, variance, and covariance functions of $X_1(t)$ and $X_2(t)$ assuming log-normality of the scores; and using these true values instead of $\mu_j(t)$ and $G_{jl}(s,t)$, $j,l = 1,2$ computed assuming normality while computing true values of the parameter functions that involve these moments, such as the CCC. Supplemental Tables 1-3 present the results for $(\tau_1^2, \tau_2^2) = (4,4)$. The corresponding results under normality appear in Tables 2-5 of the main article.

From Supplemental Table 1, we see that the impact of non-normality on the MSE of a point estimator depends on n, the design, and the parameter. For example, the MSE of estimator of K decreases while those of z(CCC) and scores show substantial increase for all designs. A substantial increase is seen for other quantities as well but only for designs (b)-(d), not (a). Much of the increase in the MSEs may be driven by an increase in the biases of the estimators. The impact on the coverage performance of interval estimators also depends on n and the design but not so much on the parameter. The new coverage probability estimates in Supplemental Tables 2 and 3 are generally less than before and the new entries tend to decrease with increase in design sparsity. Nevertheless, the interval estimators for all quantities may be considered to have acceptable accuracy for $n \geq 50$ in case of design (a) and for $n \geq 100$ in case of the other designs. Thus, on the whole, we may conclude that the skewness in the data generally increases the MSE of point estimators but the interval estimators continue to have acceptable accuracy with $n \geq 100$.

2 Analysis of Body Temperature Data

In this section, we proceed along the lines of Section 6 in the main paper to analyze the body temperature data. These data from Li and Chow [1] were collected in a study at the Noll Physiological Research Center of the Pennsylvania State University. The variable of interest here is the core body temperature, which is relatively unaffected by the environmental temperature under normal circumstances. The study was conducted in a chamber set at 36 °C. Each of $n = 12$ study subjects repeated 3 cycles of 10-minute rest followed by 20-minute exercise. For each subject, core body temperature (°C) was measured every minute over a period of 90 minutes at esophagus and rectum. These locations are the two "measurements

Supplemental Table 1: MSEs of estimators of quantities that are free of t and average MSEs of estimators of quantities that depend on t, computed using MPACE (marked as M) and UPACE (marked as U) approaches, and the ratio of the MSEs (marked as U/M) in case of four designs: (a) $N_i = 50$ (dense data), (b) $N_i = 30$ (sparse data), (c) $N_i = 20$ (sparse data), and (d) unbalanced design with mean $N_i = 20$ (sparse data), each with $(\tau_1^2, \tau_2^2) = (4, 4)$.

		(a)			(b)			(c)			(d)		
n	Quantity	M*	U*	U/M	M*	U*	U/M	M*	U*	U/M	M*	U*	U/M
50	Curves	3.70	3.78	1.02	3.70	3.81	1.03	3.54	3.67	1.04	3.62	3.74	1.03
	Scores	71.15	76.34	1.07	62.85	64.42	1.02	61.71	62.74	1.02	60.84	61.33	1.01
	K	0.47	0.19	0.40	1.96	0.82	0.42	2.93	1.33	0.45	2.27	1.43	0.63
	$\log(\tau_1^2)$	2.04	2.55	1.06	36.89	37.94	1.03	52.72	60.67	1.15	99.42	107.83	1.08
	$\log(\tau_2^2)$	2.25	2.39	1.07	21.47	22.72	1.06	35.69	29.82	0.84	47.03	50.05	1.06
	$\log(\tau_1^2/\tau_2^2)$	4.34	4.36	1.00	29.72	28.24	0.95	101.53	94.51	0.93	61.52	54.15	0.88
	z(CCC)	50.00	53.19	1.06	49.13	49.33	1.00	50.55	58.73	1.17	51.89	53.40	1.03
100	Curves	3.74	3.81	1.02	3.69	3.77	1.02	3.54	3.67	1.04	3.60	3.74	1.04
	Scores	68.45	68.45	1.00	60.86	61.04	1.00	61.71	62.74	1.02	55.19	57.31	1.04
	K	0.09	0.08	0.96	1.66	0.76	0.46	2.93	1.33	0.45	2.97	1.60	0.54
	$\log(\tau_1^2)$	1.19	1.09	0.92	22.31	23.51	1.05	35.57	42.89	1.21	49.26	55.24	1.12
	$\log(\tau_2^2)$	1.10	1.25	1.14	16.43	17.40	1.06	19.91	20.86	1.05	17.85	21.57	1.21
	$\log(\tau_1^2/\tau_2^2)$	2.09	2.15	1.03	17.07	16.73	0.98	42.13	39.01	0.93	23.25	21.59	0.93
	z(CCC)	38.06	40.56	1.07	30.82	31.95	1.04	34.83	37.25	1.07	39.08	41.80	1.07
200	Curves	3.76	3.83	1.02	3.67	3.76	1.02	3.52	3.65	1.04	3.55	3.64	1.03
	Scores	65.95	65.00	0.99	64.42	64.01	0.99	67.94	68.77	1.01	67.96	67.62	1.00
	K	0.01	0.02	1.67	1.43	0.59	0.41	3.1	1.35	0.44	2.87	1.63	0.57
	$\log(\tau_1^2)$	0.65	0.70	1.07	17.72	19.02	1.07	34.15	37.36	1.09	25.45	30.73	1.21
	$\log(\tau_2^2)$	0.59	0.67	1.13	10.84	11.55	1.07	19.39	20.13	1.04	16.73	21.58	1.29
	$\log(\tau_1^2/\tau_2^2)$	1.07	1.15	1.08	11.96	11.82	0.99	25.75	24.03	0.93	14.34	14.70	1.02
	z(CCC)	23.61	26.02	1.10	22.36	21.37	0.96	20.67	24.02	1.16	26.11	32.70	1.25

*For all quantities except the curves, scores, and K, the entries represent $10^3 \times$ MSE.

Supplemental Table 2: Estimated coverage probabilities (in %) of 95% confidence intervals for error variances and their ratio in case of four designs, (a) $N_i = 50$ (dense data), (b) $N_i = 30$ (sparse data), (c) $N_i = 20$ (sparse data) and (d) unbalanced design with mean $N_i = 20$ (sparse data), each with $(\tau_1^2, \tau_2^2) = (4, 4)$.

	n	ψ	(a)	(b)	(c)	(d)
MPACE	50	τ_1^2	93.0	93.0	92.7	91.7
		τ_2^2	93.3	92.0	93.3	92.0
		τ_1^2/τ_2^2	94.3	94.7	93.7	93.7
	100	τ_1^2	94.3	94.3	93.0	92.3
		τ_2^2	94.3	93.3	93.0	92.7
		τ_1^2/τ_2^2	95.0	95.3	94.0	94.3
	200	τ_1^2	95.7	94.7	94.3	94.3
		τ_2^2	95.0	94.7	95.0	94.0
		τ_1^2/τ_2^2	95.0	96.0	95.7	95.7
UPACE	50	τ_1^2	93.3	92.0	91.0	91.7
		τ_2^2	93.7	92.3	92.0	91.0
		τ_1^2/τ_2^2	94.7	93.3	93.7	93.7
	100	τ_1^2	94.0	94.0	92.7	92.0
		τ_2^2	95.0	93.3	93.0	92.3
		τ_1^2/τ_2^2	95.3	95.0	94.0	93.7
	200	τ_1^2	95.0	94.0	93.7	93.3
		τ_2^2	94.0	94.3	94.7	93.3
		τ_1^2/τ_2^2	94.0	94.3	94.7	94.7

Supplemental Table 3: Average estimated pointwise coverage probabilities (in %) of 95% pointwise confidence bands and estimated simultaneous coverage probabilities (in %) of 95% simultaneous confidence bands for function that depend on t in case of four designs, (a) $N_i = 50$ (dense data), (b) $N_i = 30$ (sparse data), (c) $N_i = 20$ (sparse data) and (d) unbalanced design with mean $N_i = 20$ (sparse data), each with $(\tau_1^2, \tau_2^2) = (4, 4)$.

			Pointwise				Simultaneous			
	n	$\psi(t)$	(a)	(b)	(c)	(d)	(a)	(b)	(c)	(d)
MPACE	50	$E[Y_1(t)]$	94.9	94.4	93.2	93.5	92.7	92.3	91.0	91.3
		$E[Y_2(t)]$	94.1	94.4	93.0	93.9	91.3	92.3	91.0	91.0
		$E[Y_1(t)] - E[Y_2(t)]$	95.3	95.4	94.3	94.3	94.0	93.3	92.7	92.0
		CCC(t)	94.4	94.8	94.2	93.2	92.3	92.3	92.7	91.7
	100	$E[Y_1(t)]$	96.2	95.7	95.2	94.1	95.3	94.0	93.0	92.3
		$E[Y_2(t)]$	96.6	95.4	95.0	94.3	95.0	94.3	93.0	92.0
		$E[Y_1(t)] - E[Y_2(t)]$	96.6	95.5	95.1	95.3	95.3	94.0	93.7	93.7
		CCC(t)	95.1	95.5	94.3	94.5	94.3	93.7	93.0	93.3
	200	$E[Y_1(t)]$	96.8	96.9	96.4	96.7	95.0	95.3	94.7	94.7
		$E[Y_2(t)]$	96.8	97.4	95.4	95.9	95.3	95.3	94.0	94.0
		$E[Y_1(t)] - E[Y_2(t)]$	96.9	97.1	96.1	95.4	95.3	95.3	94.0	94.3
		CCC(t)	96.7	96.1	95.7	95.0	95.3	94.3	93.7	94.0
UPACE	50	$E[Y_1(t)]$	94.9	94.4	93.2	93.5	93.0	92.3	91.3	91.3
		$E[Y_2(t)]$	94.1	94.4	93.0	93.9	91.3	92.7	91.0	91.0
		$E[Y_1(t)] - E[Y_2(t)]$	95.3	95.4	94.3	94.3	93.7	92.3	92.7	92.0
		CCC(t)	94.0	94.5	95.1	94.9	92.0	92.0	93.0	92.3
	100	$E[Y_1(t)]$	96.2	95.7	95.2	94.1	95.3	93.7	93.0	92.3
		$E[Y_2(t)]$	96.6	95.4	95.0	94.3	95.0	94.3	93.0	92.0
		$E[Y_1(t)] - E[Y_2(t)]$	96.6	95.5	95.1	95.1	95.7	94.0	93.7	94.0
		CCC(t)	94.6	94.7	94.5	94.2	93.7	93.7	93.0	93.3
	200	$E[Y_1(t)]$	96.8	96.9	96.4	96.7	95.3	95.0	94.7	94.7
		$E[Y_2(t)]$	96.8	96.8	95.4	95.9	95.3	94.7	94.7	94.0
		$E[Y_1(t)] - E[Y_2(t)]$	96.8	97.1	96.1	95.4	95.3	95.3	94.0	94.3
		CCC(t)	95.6	95.0	94.4	94.5	94.3	93.7	92.7	93.0

methods," and they are called "Tes" and "Tre," respectively. These are dense functional data. It is known that Tre has a slower response time than Tes to body temperature changes during short durations. Our interest is in evaluating agreement between the two methods. See Li and Chow [1] for further details of the study.

Supplemental Figure 1 displays the observed curves separately for each method. Also superimposed on the curves are their smoothed mean functions. The body temperature is expected to increase during the exercise periods and decrease during the rest periods. Although this effect can be seen in the curves for both methods, the effect is less prominent for Tre because of its slower response time than Tes. The two mean functions clearly appear different. Supplemental Figure 2 presents scatterplots of the data together with the line of equality at $t = 10, 30, 40, 60, 70, 90$ minutes. These times correspond to the end of the rest and exercise periods in the three cycles. The measurements range between 36.5 and 38.5. The curves for both measurement methods show a general upward trend. There is periodicity in the curves due to the three cycles of rest and exercise periods. The periodicity is more prominent for Tes curves than Tre curves. The Tre measurements are almost always a bit higher than the corresponding Tes measurements. The methods appear moderately correlated at these times, with sample correlations ranging between 0.75 and 0.93. The agreement between the methods is less than perfect because in that case all points in the scatterplot would fall on the line of equality. On the whole, there is a small but persistent difference in the measurement methods.

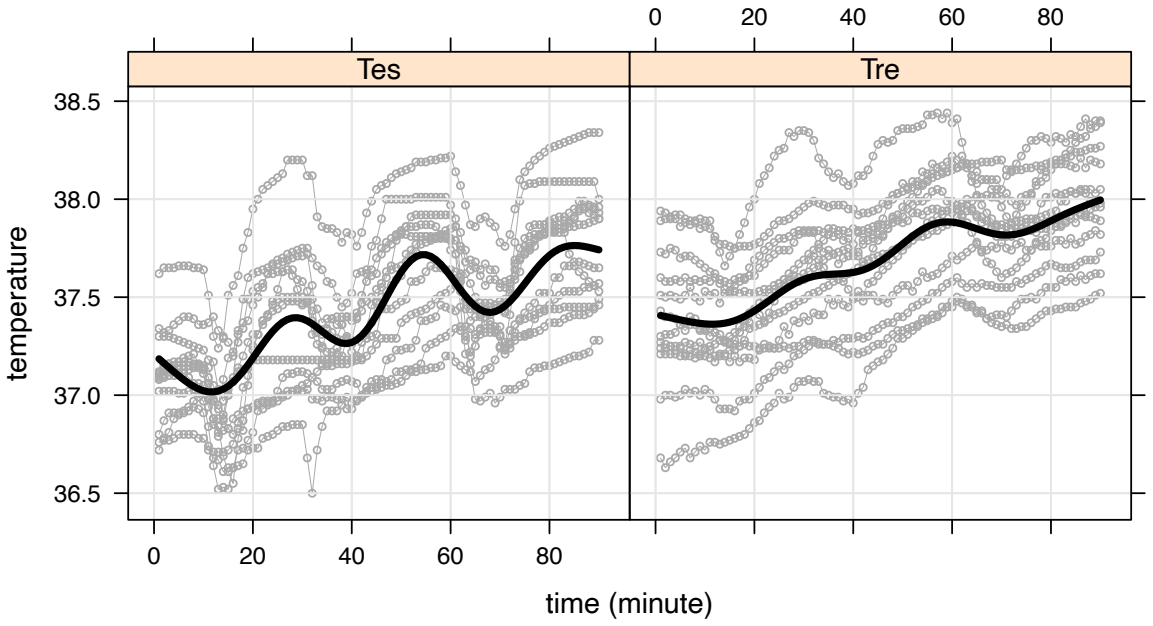
Next, we fit the model given by equation (9) in the main paper using both MPACE and UPACE approaches. As before, the smoothing is performed using `gam` function in `mgcv` package of `R` with default settings. The two smoothed mean functions are displayed in Supplemental Figure 3. Both functions have overall increasing trends—the Tes mean increases from 37.2 at $t = 1$ to 37.7 at $t = 90$, and the Tre mean increases from 37.4 to 38.0 over the same period. The Tre mean lies above the Tes mean throughout. The two functions also appear different. The Tes mean exhibits a marked cyclical behavior that more or less coincides with the cycles of rest and exercise periods in the experiment. In particular, in each cycle, the Tes mean tends to decrease during the rest period and increase during the exercise period. Although the times of troughs and peaks do not correspond exactly to the end of rest and exercise periods, their discrepancy is small. In contrast, the cyclical behavior of the Tre mean function is less apparent. It tends to increase during the exercise period but it does not decrease as much as the Tes mean during the rest period. This difference in the means may be explained by the slower response time of Tre than Tes to body temperature changes during short durations.

The next step in model fitting is to perform an FPCA of these data. The respective estimates of the number of FPC needed to explain at least 99% of variability in the observed curves, eigenvalues, and error variances computed using MPACE are:

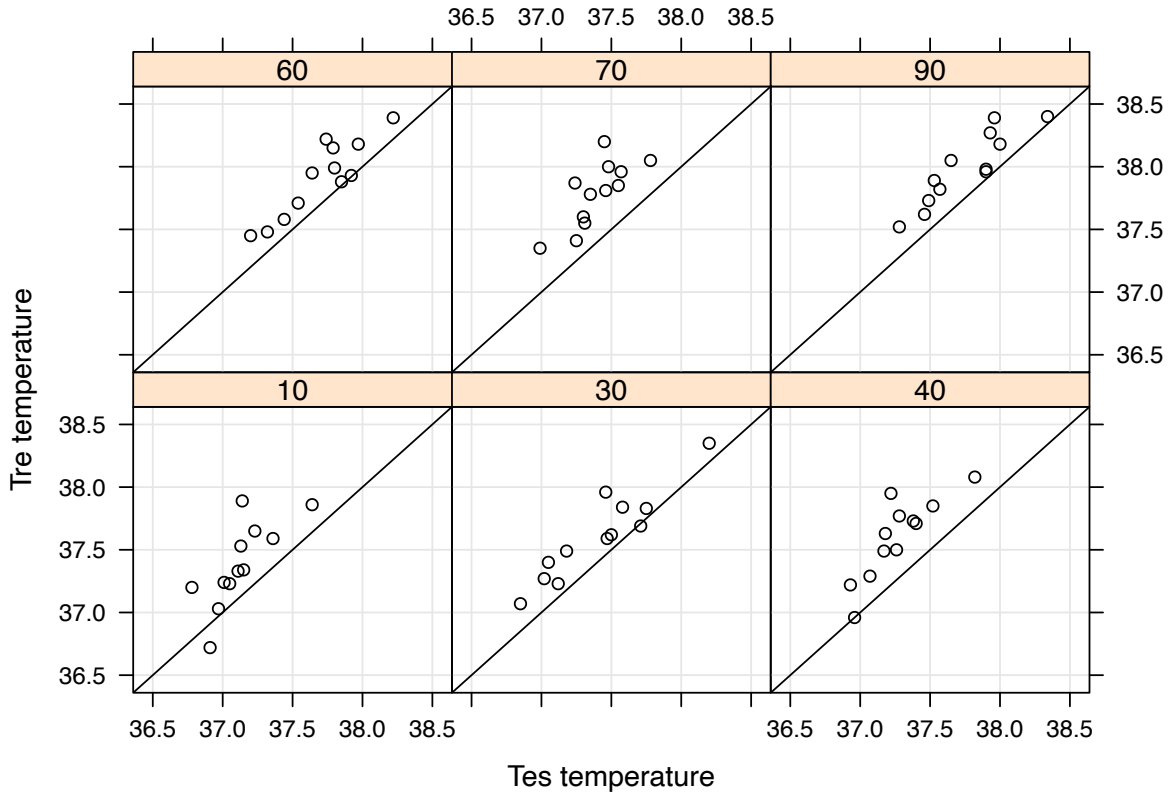
$$\hat{K} = 6, (\hat{\lambda}_1, \hat{\lambda}_2, \hat{\lambda}_3, \hat{\lambda}_4, \hat{\lambda}_5, \hat{\lambda}_6) = (14.29, 1.15, 0.60, 0.22, 0.11, 0.07), (\hat{\tau}_1^2, \hat{\tau}_2^2) = (6.3, 1.4) \times 10^{-3}.$$

The same estimates using UPACE are:

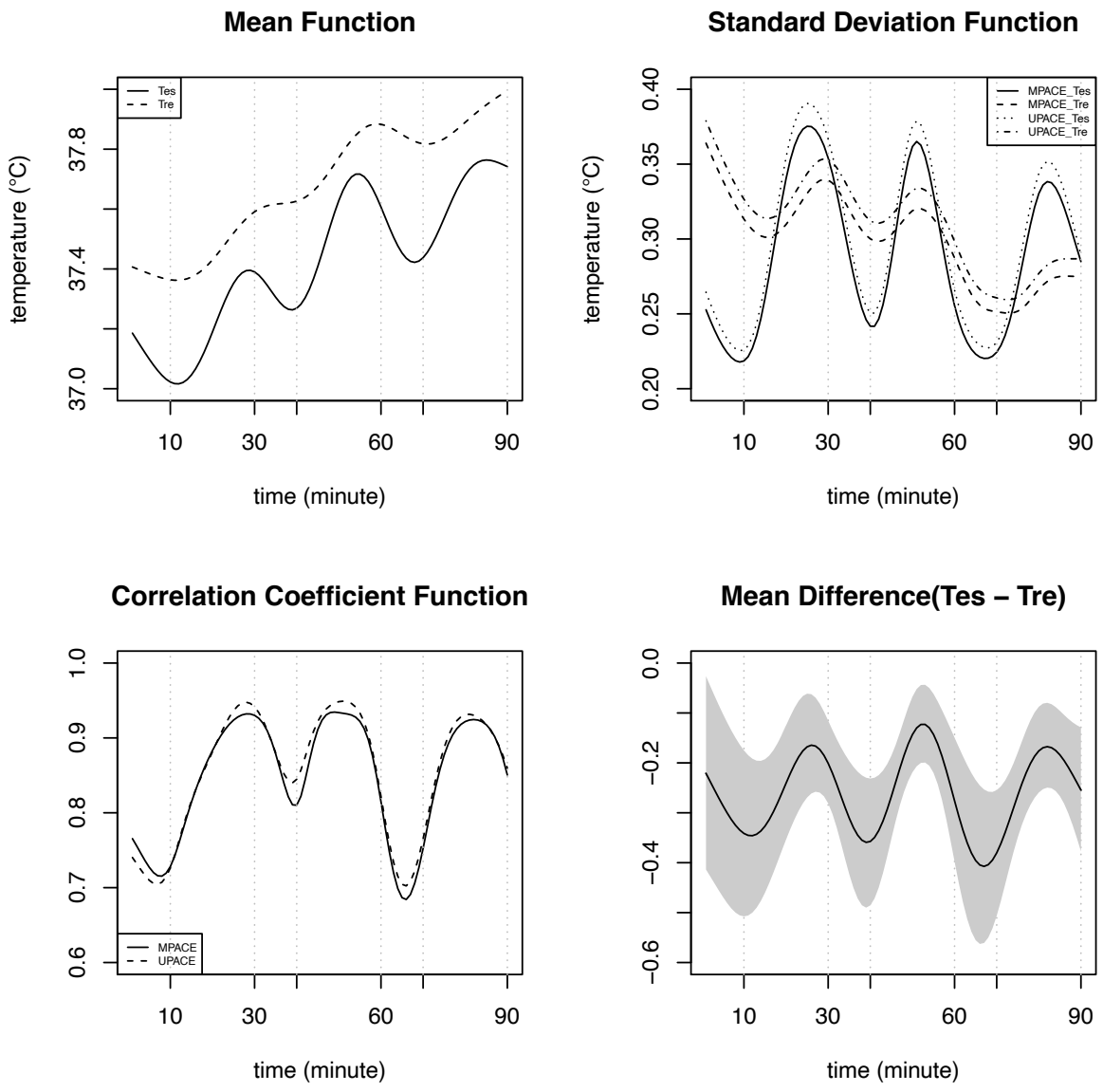
$$\hat{K} = 5, (\hat{\lambda}_1, \hat{\lambda}_2, \hat{\lambda}_3, \hat{\lambda}_4, \hat{\lambda}_5) = (14.28, 1.13, 0.59, 0.20, 0.10), (\hat{\tau}_1^2, \hat{\tau}_2^2) = (6.5, 1.4) \times 10^{-3}.$$



Supplemental Figure 1: The observed individual core body temperature curves (in grey) for the two methods superimposed with estimated mean functions (in black). The vertical broken lines at $t = 10, 40, 70$ mark the beginning of the 20-minute exercise period within each cycle, and those at $t = 30, 60, 90$ mark its end. A 10-minute rest period precedes each exercise period.



Supplemental Figure 2: Scatterplots of body temperatures from two methods at $t = 10, 30, 40, 60, 70, 90$ minutes. These time points mark the end of the 10-minute rest period and the 20-minute exercise period for each of the 3 cycles.



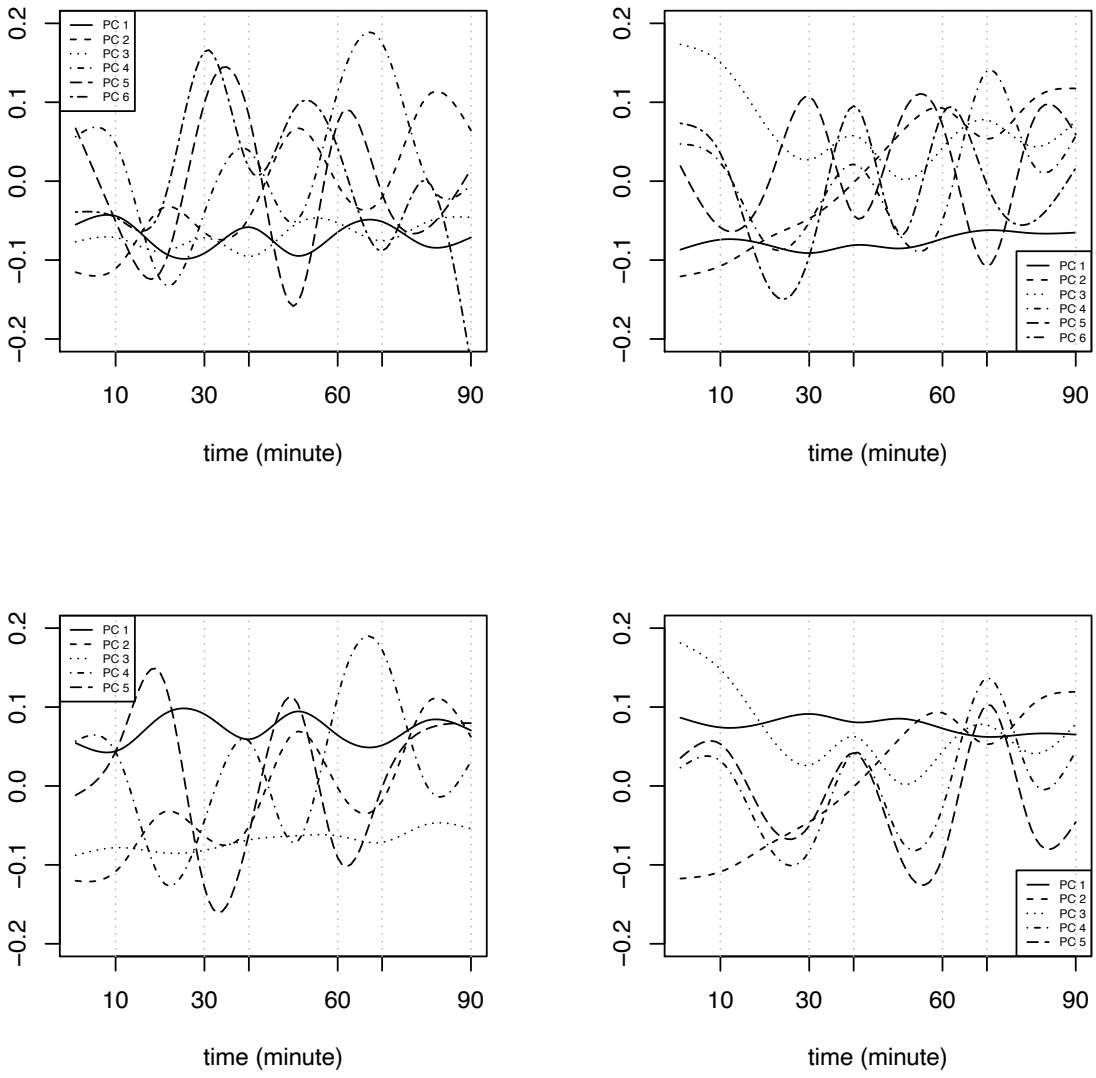
Supplemental Figure 3: The estimated mean, standard deviation, correlation, and mean difference functions for Tre and Tes methods. The bottom right panel also shows a 95% simultaneous confidence band for the mean difference function.

Compared to MPACE, UPACE selects one fewer FPC and its eigenvalues are slightly smaller. On the whole, however, the two sets of estimates are quite similar. Supplemental Figure 4 presents the estimated eigenfunctions for Tes and Tre temperatures ($\hat{\phi}_{kj},\ k = 1, \ldots, \hat{K},\ j = 1, 2$) using the two approaches. The two sets of first five eigenfunctions are similar (ignoring the sign flip for the first component). The eigenfunctions for both temperatures exhibit trend and cyclical behavior. To try to gain further insights, let us focus on UPACE eigenfunctions and examine their behavior.

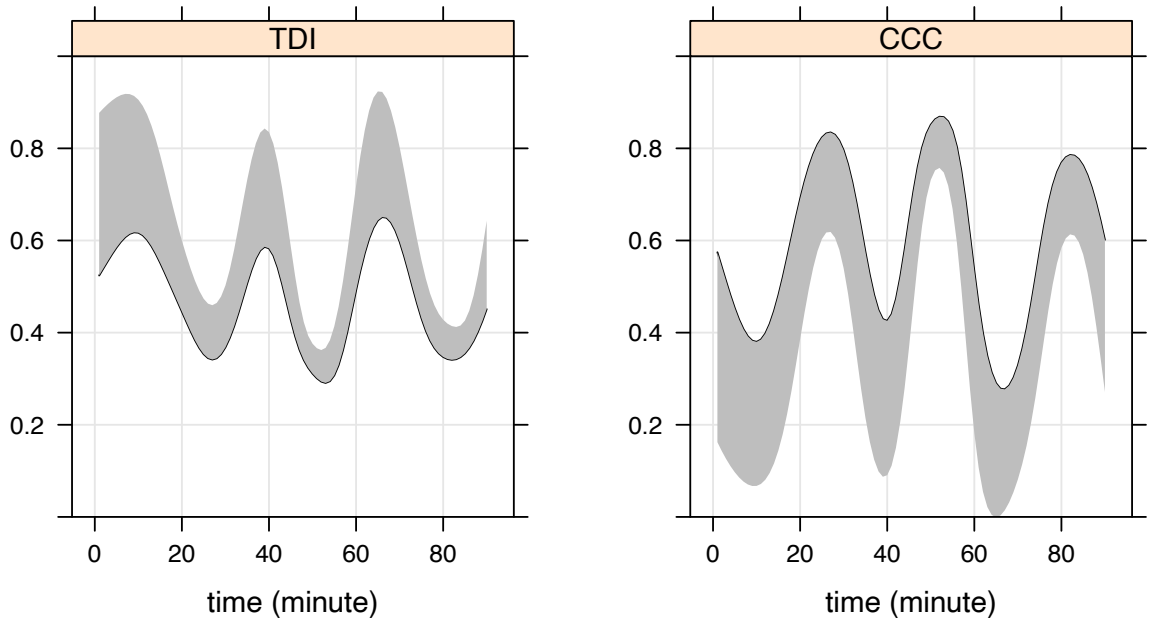
Beginning with Tes, we see that $\hat{\phi}_{11}$ does not have a prominent trend, whereas the others have increasing trends. Likewise, $\hat{\phi}_{31}$ does not have a prominent cyclical behavior, whereas the others do. Both $\hat{\phi}_{11}$ and $\hat{\phi}_{21}$ peak near the middle of the exercise periods and trough near the end of the rest periods. The times of peaks and troughs for $\hat{\phi}_{41}$ are swapped with those $\hat{\phi}_{11}$ and $\hat{\phi}_{21}$, i.e., $\hat{\phi}_{41}$ peaks near the end of the rest periods and troughs near the middle of exercise periods. The function $\hat{\phi}_{51}$ peaks the near the middle of the rest periods and troughs near the middle of the exercise periods, and appears to provide a contrast between the two periods. In case of Tre, $\hat{\phi}_{12}$ has a slight decreasing trend but no prominent cyclical behavior. However, all others exhibit both trend and cyclical behavior. Specifically, $\hat{\phi}_{22}$ increases up to the end of the second exercise period, decreases in the following rest period, and increases again in the third exercise period. On the other hand, $\hat{\phi}_{32}$ decreases up to the end of the first exercise period and then starts exhibiting a cyclical behavior like $\hat{\phi}_{41}$, i.e., it peaks near the end of the rest periods and troughs near the middle of the exercise periods. The function $\hat{\phi}_{42}$ behaves like $\hat{\phi}_{41}$ from the beginning; and $\hat{\phi}_{52}$ is almost like a mirror image of $\hat{\phi}_{42}$. The initial trends in $\hat{\phi}_{32}$ and $\hat{\phi}_{42}$ may be manifestations of the slow response time of Tre to body temperature changes during short durations.

Supplemental Figure 3 also presents the estimates of standard deviation functions $\sigma_j(t)$ and correlation function $\rho(t)$, given by equation (6) in the main paper.Unsurprisingly, both MPACE and UPACE lead to similar estimates. These functions also exhibit periodicity. Essentially, the standard deviation functions of Tes and Tre tend to decrease during the rest periods and increase during the exercise periods. They also cross. There is a downward trend in the function for Tre, which is absent for Tes. The correlation function ranges between 0.68 and 0.95. It tends to trough during the rest periods and peak during the exercise periods. Its values match up well with the raw sample correlations shown in Supplemental Figure 2.

Now, we consider evaluation of similarity. Supplemental Figure 3 presents the estimate and a two-sided 95% simultaneous confidence band for the mean difference function $\delta(t)$. These and other interval estimates reported here use $Q = 500$ bootstrap repetitions. The estimated mean difference function (Tes − Tre) is negative throughout $\mathcal{T}$. The entire confidence band lies below zero. The mean difference function does not have any trend but it varies around -0.25 in a cyclical manner. It tends to decrease during the rest periods and increase during the exercise periods. In absolute value terms, this means that the mean difference tends to increase during the rest periods and decrease during the exercise periods. This cyclical pattern is similar to that of the correlation function. The estimate for precision ratio τ_1^2/τ_2^2 using MPACE is 4.5 and its 95% confidence interval is $(2.2, 9.0)$. These quantities are estimated as 4.6 and $(2.3, 9.7)$, respectively, using UPACE. Although the confidence intervals are somewhat wide, there is indication that Tre is more precise than Tes. These findings indicate that, because of the difference in their mean functions and precisions, the two methods cannot be regarded as similar.



Supplemental Figure 4: Estimated eigenfunctions for Tes (left panel) and Tre (right panel) temperatures using MPACE (top panel) and UPACE (bottom panel) approaches.



Supplemental Figure 5: Estimate of TDI function with $p_0 = 0.90$ and its 95% simultaneous upper confidence band (left) and estimate of CCC function and its 95% simultaneous lower confidence band (right) using MPACE and UPACE approaches.

Next, we consider evaluation of agreement. The probability for TDI is taken as $p_0 = 0.90$. Supplemental Figure 5 presents estimates and 95% one-sided simultaneous confidence bands for CCC and TDI. Lower bands for CCC and upper bands for TDI are presented. Both MPACE and UPACE lead to similar results. The estimates of both CCC and TDI functions as well as their confidence bands continue to exhibit the familiar cyclical pattern. On the basis of both CCC and TDI, we see that the extent of agreement between the methods tends to decrease during the rest periods and increase during the exercise periods. This cyclical pattern is similar to the one observed for the mean difference and correlation functions. The CCC ranges between 0.28 and 0.87 and its lower band ranges between 0 and 0.76. Thus, even during the exercise periods, the CCC represents a rather weak amount of agreement between the methods. This finding is consistent with the conclusion of Li and Chow [1]. The TDI estimate ranges between 0.29 and 0.65 and its upper band ranges between 0.36 and 0.92. The largest upper bound of 0.95 implies that 90% of differences between Tes and Tre methods is within ± 0.95. Thus, if a difference of up to ± 0.95°C is acceptable for the application at hand, the methods may be considered to agree sufficiently well for interchangeable use in that application, but not otherwise. It is also clear from the similarity evaluation that one reason why the methods do not agree well is that their mean functions differ.

References

1. Li R, Chow M. Evaluation of reproducibility for paired functional data. *Journal of Multivariate Analysis* 2005; **93**:81–101.