Modeling and analysis of functional method comparison data

Galappaththige S. R. de Silva, Lasitha N. Rathnayake, and Pankaj K. Choudhary

Department of Mathematical Sciences, University of Texas at Dallas, Richardson, Texas, 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.

1. Introduction

Multivariate functional data arise when repeated measurements of \( J \geq 2 \) variables are taken over time on every subject (Ramsay and Silverman 2005; Berrendero, Justel, and Svarc 2011; Chiou, Chen, and Yang 2014; Jacques and Preda 2014; Happ and Greven 2018). 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 (Choudhary and Nagaraja 2017). 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 (1986), 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. (1996) and body temperature data Li and Chow (2005). 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, Haber, and Lin (2007) and Choudhary and Nagaraja (2017) 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 (1989) and total deviation index (TDI) of Lin (2000). However, in the biomedical literature, the limits of agreement approach of Bland and Altman (1986) is the most popular. These measures are defined in Sec. 4. A reader interested in their comparison may consult Barnhart, Haber, and Lin (2007). 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 (Dunn 2007). Evaluation of similarity is a necessary supplement to evaluation of agreement as it provides information about the sources of disagreement between the methods (Choudhary and Nagaraja 2017, Chapter 1).

In the method comparison literature, we are only aware of Li and Chow (2005) that deals with functional observations. It extends the ideas of Lin (1989) 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 Sec. 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, Haber, and Lin 2007). 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 (2005) for an introduction. A common analytical approach involves performing a functional principal components analysis (FPCA) to obtain a parsimonious representation of the data (Ramsay and Silverman 2005, Chapter 8). The PACE (principal components analysis through conditional expectation) methodology of Yao, Müller, and Wang (2005) 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, Greven, and Crainiceanu (2013) are implemented in the refund (Goldsmith et al. 2016) and MFPCA (Happ 2018) packages for the statistical software system R (R Core Team 2018).

Methodologies for FPCA of multivariate functional data have also been developed, see, e.g., Ramsay and Silverman (2005, Chapter 8), Berrendero, Justel, and Svarc (2011), Jacques and Preda (2014), Chiou, Chen, and Yang (2014), and Happ and Greven (2018). Among these, the approaches of Chiou, Chen, and Yang (2014) and Happ and Greven (2018) 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, Chen, and Yang (2014) and Happ and Greven (2018) differ in their basic premise regarding univariate components of the multivariate observation—in particular, they may have different units in Chiou, Chen, and Yang (2014) and they may be observed on different (dimensional) domains in Happ and Greven (2018)—both first obtain a Karhunen-Loève expansion of the multivariate observations. Thereafter, Chiou, Chen, and Yang (2014) 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 (2018) 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, Müller, and Wang (2005). This methodology is implemented in an R package MFPCA (Happ 2018).

This brings us to our approach for analysis of functional method comparison data. In Sec. 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 Sec. 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, Chen, and Yang (2014). The second approach—termed UPACE—adapts the methodology of Happ and Greven (2018). Bootstrap is used to construct relevant confidence intervals and bands. In Sec. 4, we discuss evaluation of similarity and agreement under the assumed model. Sec. 5 presents a simulation study to evaluate properties of the two estimation approaches. The body fat data are analyzed in Sec. 6. Sec. 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 $T = [a, b], a < b \in \mathbb{R}$. Let the mean and covariance functions of the random functions be denoted by
\[ \mu_j(t) = E(X_j(t)), \quad G_{jl}(s,t) = \text{cov}(X_j(s), X_l(t)), \quad j, l = 1, \ldots, J; \quad s, t \in T. \]

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

### 2.1. Model for population curves

Under certain conditions (Chiou, Chen, and Yang 2014; Happ and Greven 2018), the multivariate Karhunen-Loève Theorem provides a stochastic representation of \( X \) as

\[ X(t) = \mu(t) + \sum_{k=1}^{\infty} \xi_k \phi_k(t), \quad t \in T. \tag{1} \]

Here, \( \phi_k(t) = (\phi_{k1}(t), \ldots, \phi_{kJ}(t))^T \) are orthonormal eigenfunctions, satisfying the property that the inner product of \( \phi_k \) and \( \phi_l \), given as \( \int_T \phi_k(t) \phi_l(t) dt \), equals zero if \( k \neq l \) and one if \( k = l \); and \( \xi_k \)—called “scores”—are uncorrelated random variables with mean zero and variance \( \lambda_k \). The variances \( \lambda_k \) are eigenvalues associated with the eigenfunctions \( \phi_k \) and are non-increasing, i.e., \( \lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_J \). We can write (1) as

\[ X_j(t) = \mu_j(t) + \sum_{k=1}^{\infty} \xi_k \phi_k(t), \quad j = 1, \ldots, J; \quad t \in T. \tag{2} \]

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

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

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 \( \tau_j^2, j = 1, \ldots, J \), and are independent of the true values. Using (2), we can write this model as

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

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

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

and their cross-covariance function is \( \text{cov}(Y_j(s), Y_l(t)) = G_{jl}(s,t), j \neq l = 1, \ldots, J \). Here \( I \) is the indicator function. It follows that, for each \( t \in T \), 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

\[ \delta_{jl}(t) = \mu_j(t) - \mu_l(t), \quad \eta_{jl}^2(t) = \sigma_j^2(t) + \sigma_l^2(t) - 2G_{jl}(t,t). \tag{7} \]
These distributions are used in Sec. 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 \( N_i \) 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 \( \xi_{ik} \) denote the respective counterparts of the population quantities \( X_i(t), \epsilon_i(t), Y_i(t) \), and \( \xi_k \), 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

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

where the errors \( \epsilon_{ij}(t_{im}) \) are independent random variables with mean zero and variance \( \tau_i^2 \). 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 \( \xi_{ik} \) 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

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

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 Sec. 3.

To write (9) in the matrix notation, define the \( N_i \times 1 \) vectors \( t_i = (t_{i1}, \ldots, t_{iN_i})^T, Y_i(t_i) = (Y_{ij}(t_{i1}), \ldots, Y_{ij}(t_{iN_i}))^T, \mu_i(t_i) = (\mu_j(t_{i1}), \ldots, \mu_j(t_{iN_i}))^T, \) and \( \epsilon_i(t_i) = (\epsilon_{ij}(t_{i1}), \ldots, \epsilon_{ij}(t_{iN_i}))^T \). 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

\[
Y_i(t_i) = \begin{pmatrix} Y_{i1}(t_i) \\ \vdots \\ Y_{iJ}(t_i) \end{pmatrix}, \quad \mu_i(t_i) = \begin{pmatrix} \mu_j(t_i) \\ \vdots \\ \mu_j(t_i) \end{pmatrix}, \quad \epsilon_i(t_i) = \begin{pmatrix} \epsilon_{i1}(t_i) \\ \vdots \\ \epsilon_{iJ}(t_i) \end{pmatrix},
\]

and take the \( JN_i \times JN_i \) diagonal matrix \( R_i = \text{diag}\{\tau_{11}^2, \ldots, \tau_{1j}^2, \ldots, \tau_{Jj}^2\} \), where \( \tau_j^2 \) is repeated \( N_i \) times for each \( j \), as the covariance matrix of \( \epsilon_i(t_i) \). Further, define \( \bar{\xi}_i = (\xi_{i1}, \ldots, \xi_{iK})^T \) as the \( K \times 1 \)
vector of scores and $\Lambda = \text{diag}\{\lambda_1, ..., \lambda_K\}$ as its $K \times K$ diagonal covariance matrix; $\phi_{kj}(t_i) = (\phi_{k1}(t_{i1}), ..., \phi_{kN_j}(t_{iN_j}))^T$ as the $N_i \times 1$ vector of values of $k$th eigenfunction $\phi_{kj}$ associated with method $j$; $\Phi_i(t_i) = (\phi_{k1}^T(t_i), ..., \phi_{kJ}^T(t_i))^T$ as the $JN_i \times 1$ vector by stacking the values for all the methods; and $\Phi(t_i) = (\phi_1(t_i), ..., \phi_K(t_i))$ as their $JN_i \times K$ matrix.

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

$$Y_i(t_i) \approx \mu(t_i) + \Phi(t_i)\xi_i + \epsilon_i(t_i), \quad i = 1, ..., n. \quad (11)$$

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

$$E(Y_i(t_i)) \approx \mu(t_i), \quad \text{var}(Y_i(t_i)) \approx \Phi(t_i)\Lambda\Phi^T(t_i) + \mathbf{R}_i. \quad (12)$$

The elements of the first term of $\text{var}(Y_i(t_i))$ 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

$$\theta = \{\mu_1, ..., \mu_J, K, \lambda_1, ..., \lambda_K, \phi_{11}, ..., \phi_{1J}, ..., \phi_{JK}, \tau_{1}^2, ..., \tau_{J}^2\}.$$

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

### 3. Parameter estimation and FPCA

#### 3.1. Parameter estimation

Let $N_0$ be the number of unique observation times in the data and $t_0 = (t_{01}, ..., t_{0N_0})^T$ be the $N_0 \times 1$ vector of these times in increasing order. The elements of $t_0$ form a grid in the domain $\mathcal{T}$. By definition, there is at least one observation from all measurement methods at each time in $t_0$. 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 $\mu_j(t), j = 1, ..., J$. 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 $\tilde{Y}_i(t_0)$ in the same way as $Y_i(t_i)$ are formed in (10). If the subject $i$ does not have an observation for some $t \in t_0$, that observation is set to be missing in $\tilde{Y}_i(t_0)$.

In Appendix A, we describe the two approaches—MPACE and UPACE—for estimating the remaining unknown components of $\theta$ and the multivariate scores $\xi$ in the model (11). Both use the centered data as inputs and involve the PACE methodology of Yao, Müller, and Wang (2005) for univariate functional data. MPACE directly adapts the PACE methodology to deal with multivariate data along the lines of Chiou, Chen, and Yang (2014), whereas UPACE adapts the approach of Happ and Greven (2018). 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 $\tau_j^2$ 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 (Wood 2017), any other smoothing technique—e.g., local linear regression as in Yao, Müller, and Wang (2005) and Chiou, Chen, and Yang (2014)—can also be used without affecting the general methodology. Upon model fitting, the fitted curves are $\hat{Y}_i(t_i) = \hat{\mu}_i(t_i) + \Phi_i(t_i)\hat{\xi}_i, i = 1, ..., n.$
### 3.2. Confidence intervals and bands

Suppose $\psi \equiv \psi(\theta)$ is a function of model parameters of interest. Examples of $\psi$ include the precision ratio $\tau_1^2/\tau_2^2$. Often, the parameter function depends on $t$, i.e., it has the form $\psi(t) \equiv \psi(t, \theta), t \in T$. Examples of $\psi(t)$ include the mean difference $\delta_{ij}(t)$ and the agreement measures defined in next section. Since $\psi$ 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 $t$ of $L$ points in $T$, say, $t_1, \ldots, t_L$. This grid may be the same as the grid $t_0$ formed by the observed time points, used for estimation in Sec. 3. Or it may consist of a subset of these time points. In practice, $L \in [25, 50]$ is often adequate.

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

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

where $z_\alpha$ is the $100\alpha$th percentile of a $\mathcal{N}(0,1)$ distribution. A simultaneous band can be constructed by replacing $z_\alpha$ in (13) by an appropriate percentile (Choudhary and Nagaraja 2017, Chapter 3) that can be computed using the multcomp package of Hothorn, Bretz, and Westfall (2008) in R or via simulation as we do here. We now present a bootstrap methodology to compute $b$ and $S$. 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 $\theta$ from the resampled data to get $\hat{\theta}^*$. 
3. Use $\hat{\theta}^*$ to estimate $\hat{\psi}(t)$ as $\hat{\psi}^*(t)$. This $\hat{\psi}^*(t)$ is a resample of $\hat{\psi}(t)$.
4. Repeat the previous steps $Q$ times to get the resamples $\hat{\psi}_q^*(t), q = 1, \ldots, Q$. Compute the bias vector $b$ as $\sum_{q=1}^Q \hat{\psi}_q^*(t) / Q - \hat{\psi}(t)$, and the covariance matrix $S$ as the sample covariance matrix of the resamples.

In practice, $Q = 500$ is often enough to estimate $b$ and $S$. If there is evidence that a bias correction is not needed, then the term $b_l$ in (13) can be dropped (see Sec. 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 (Goldsmith, Greven, and Crainiceanu 2013). The procedure of this subsection can be easily adapted to construct confidence interval for a parameter function $\psi$ 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 (Dunn 2007). Under the true model (4), \( \delta_\beta(t) \) given by (7) is the mean difference and \( \tau_j^2 / \tau_j^2 \) 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 \( t \in T \).

We specifically consider two agreement measures. One is the concordance correlation coefficient (CCC) due to Lin (1989). 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

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

See Lin (1989) for properties of a CCC. Here we just note that \(|\text{CCC}_{jl}(t)| \leq |\rho_{jl}(t)| \leq 1 \) and \( \text{CCC}_{jl}(t) = \rho_{jl}(t) \) 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 \( \text{CCC}_{jl}(t) = 1 \) for all \( t \).

The other measure is the total deviation index (TDI) due to Lin (2000). For a given large probability \( p_0 \), it is defined as the \( p_0 \)th 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 \( t \in T \), the difference \( D_{jl}(t) \) follows a normal distribution with mean \( \delta_{jl}(t) \) and variance \( \eta_{jl}^2(t) \), given by (7). This implies that the functional TDI can be expressed as

\[
\text{TDI}_{jl}(p_0, t) = 100p_0 \text{th percentile of } |D_{jl}(t)| = \eta_{jl}(t) \left\{ \chi^2_{1,p_0} \left( \frac{\delta_{jl}^2(t)}{\eta_{jl}^2(t)} \right) \right\}^{1/2}, \tag{15}
\]

where \( \chi^2_{1,p_0}(\Delta) \) is the \( 100p_0 \)th percentile of a noncentral \( \chi^2 \) distribution with one degree of freedom and noncentrality parameter \( \Delta \). A TDI is non-negative, and its small value implies good agreement. Agreement between methods \( j \) and \( l \) is perfect when \( \text{TDI}_{jl}(t) = 0 \) 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 \( \tau_j^2 / \tau_l^2 \). 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 Sec. 1, the limits of agreement approach of Bland and Altman (1986) is quite popular in the biomedical literature for agreement evaluation. This involves, under the normality assumption for the differences, computing estimated mean ± 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 \( \hat{\delta}_{jl}(t) \pm 1.96 \hat{\eta}_{jl}(t) \), \( t \in T \).

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 $T = [0,1]$; 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 $T$. The grid $t_{grid} = \{u: u = 0, 1/49, ..., 1\}$ 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 $t_{grid}$, 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 $t_{grid}$ 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.

### 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 $(\tau_1^2, \tau_2^2) = (2, 2)$.

| $n$  | Quantity | (a)   |       | (b)   |       | (c)   |       | (d)   |       |
|------|----------|-------|-------|-------|-------|-------|-------|-------|-------|
| 50   | Curves   | $1.83$| $1.02$| $1.74$| $1.04$| $1.64$| $1.05$| $1.65$| $1.05$|
|     | Scores   | $9.08$| $0.99$| $8.80$| $1.00$| $8.71$| $1.01$| $8.73$| $1.00$|
|     | $K$      | $1.63$| $0.43$| $4.36$| $0.37$| $5.36$| $0.39$| $4.94$| $0.39$|
|     | log ($\tau_1^2$) | $1.94$| $0.98$| $6.58$| $0.92$| $15.68$| $0.87$| $15.87$| $0.93$|
|     | log ($\tau_2^2$) | $1.96$| $1.02$| $6.22$| $0.80$| $12.77$| $0.80$| $10.78$| $0.84$|
|     | log ($\tau_1^2/\tau_2^2$) | $3.88$| $1.02$| $9.91$| $1.01$| $21.76$| $0.98$| $24.58$| $0.96$|
|     | log (TDI) | $2.53$| $0.96$| $4.07$| $0.88$| $5.09$| $0.86$| $5.11$| $0.83$|
|     | $z(1/C)$ | $4.01$| $1.12$| $6.04$| $1.07$| $7.09$| $1.19$| $7.82$| $1.22$|
| 100  | Curves   | $1.85$| $1.01$| $1.76$| $1.03$| $1.67$| $1.05$| $1.67$| $1.05$|
|     | Scores   | $9.00$| $0.98$| $8.82$| $1.00$| $8.65$| $1.01$| $8.76$| $1.01$|
|     | $K$      | $1.04$| $0.83$| $4.64$| $0.31$| $6.69$| $0.32$| $6.45$| $0.32$|
|     | log ($\tau_1^2$) | $0.91$| $1.00$| $3.03$| $0.94$| $7.56$| $1.02$| $9.73$| $1.02$|
|     | log ($\tau_2^2$) | $1.05$| $1.02$| $2.67$| $0.90$| $6.06$| $0.91$| $6.49$| $0.87$|
|     | log ($\tau_1^2/\tau_2^2$) | $1.88$| $1.04$| $5.02$| $1.00$| $11.74$| $0.97$| $14.31$| $0.98$|
|     | log (TDI) | $1.41$| $1.05$| $1.81$| $0.96$| $2.64$| $0.94$| $2.88$| $0.87$|
|     | $z(1/C)$ | $2.15$| $1.33$| $3.02$| $1.36$| $3.56$| $1.67$| $4.10$| $1.60$|
| 200  | Curves   | $1.88$| $1.01$| $1.78$| $1.03$| $1.69$| $1.04$| $1.68$| $1.04$|
|     | Scores   | $9.20$| $0.96$| $8.85$| $1.00$| $8.73$| $1.00$| $8.69$| $1.00$|
|     | $K$      | $0.43$| $1.82$| $3.97$| $0.30$| $6.69$| $0.28$| $6.32$| $0.26$|
|     | log ($\tau_1^2$) | $0.51$| $1.02$| $1.57$| $1.02$| $3.38$| $1.03$| $3.62$| $1.03$|
|     | log ($\tau_2^2$) | $0.47$| $1.10$| $1.19$| $0.97$| $2.62$| $0.95$| $2.99$| $0.98$|
|     | log ($\tau_1^2/\tau_2^2$) | $1.03$| $1.05$| $2.62$| $1.01$| $5.81$| $1.02$| $5.97$| $0.98$|
|     | log (TDI) | $0.70$| $1.20$| $1.01$| $1.09$| $1.46$| $1.02$| $1.47$| $1.01$|
|     | $z(1/C)$ | $1.13$| $1.74$| $1.70$| $1.84$| $2.15$| $2.24$| $1.96$| $2.37$|

*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 $(\tau_1^2, \tau_2^2) = (4, 4)$.

| $n$  | Quantity | (a)  | (b)  | (c)  | (d)  |
|------|----------|------|------|------|------|
|      | 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 |
|      | Scores  | 8.75 | 8.76 | 1.00 | 8.70 | 8.77 | 1.01 | 8.78 | 8.85 | 1.01 |
|      | $K$     | 3.90 | 1.43 | 0.37 | 5.76 | 2.17 | 0.38 | 5.82 | 1.88 | 0.32 |
|      | log $(\tau_1^2)$ | 2.35 | 2.23 | 0.97 | 4.72 | 4.39 | 0.93 | 9.90 | 9.22 | 0.93 |
|      | log $(\tau_2^2)$ | 1.72 | 1.74 | 1.01 | 5.16 | 4.33 | 0.84 | 8.51 | 7.15 | 0.84 |
|      | 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 |
|      | log (TDI) | 1.85 | 1.70 | 0.92 | 2.78 | 2.38 | 0.86 | 3.68 | 2.88 | 0.78 |
|      | $z(CCC)$ | 2.68 | 3.33 | 1.24 | 4.07 | 4.90 | 1.21 | 5.43 | 6.73 | 1.24 |
| 100  | Curves  | 3.70 | 3.77 | 1.02 | 3.54 | 3.65 | 1.03 | 3.41 | 3.54 | 1.04 |
|      | Scores  | 8.82 | 8.81 | 1.00 | 8.81 | 8.83 | 1.00 | 8.66 | 8.71 | 1.01 |
|      | $K$     | 3.64 | 1.29 | 0.35 | 6.57 | 1.98 | 0.30 | 7.20 | 2.05 | 0.29 |
|      | log $(\tau_1^2)$ | 1.10 | 1.10 | 1.01 | 1.97 | 2.02 | 1.03 | 4.06 | 4.02 | 0.99 |
|      | log $(\tau_2^2)$ | 1.13 | 1.07 | 0.95 | 1.91 | 1.71 | 0.89 | 3.79 | 3.63 | 0.96 |
|      | 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 |
|      | log (TDI) | 0.91 | 0.90 | 0.99 | 1.30 | 1.18 | 0.91 | 1.97 | 1.66 | 0.84 |
|      | $z(CCC)$ | 1.37 | 2.20 | 1.61 | 2.04 | 3.82 | 1.87 | 2.91 | 5.93 | 1.24 |
| 200  | Curves  | 3.74 | 3.78 | 1.01 | 3.58 | 3.66 | 1.02 | 3.43 | 3.55 | 1.03 |
|      | Scores  | 8.87 | 8.79 | 0.99 | 8.84 | 8.85 | 1.00 | 8.64 | 8.71 | 1.01 |
|      | $K$     | 3.13 | 1.03 | 0.33 | 6.70 | 1.66 | 0.25 | 8.48 | 2.03 | 0.24 |
|      | log $(\tau_1^2)$ | 0.52 | 0.52 | 1.00 | 0.93 | 0.98 | 1.05 | 2.01 | 2.08 | 1.03 |
|      | log $(\tau_2^2)$ | 0.53 | 0.53 | 1.01 | 1.30 | 1.16 | 0.90 | 2.01 | 2.09 | 1.04 |
|      | 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 |
|      | log (TDI) | 0.53 | 0.55 | 1.04 | 0.71 | 0.73 | 1.03 | 1.07 | 1.05 | 0.99 |
|      | $z(CCC)$ | 0.69 | 1.52 | 2.19 | 1.10 | 3.01 | 2.74 | 1.44 | 4.63 | 3.22 |

For each setting, we simulate a dataset, perform parameter estimation as described in Sec. 3.1 and Appendix A, and construct 95% confidence intervals and bands as described in Secs. 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 (2017) 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 $\log(\tau_1^2)$, $\log(\tau_2^2)$, $\log(\tau_1^2/\tau_2^2)$, $\log\{TDI(p_0, t)\}$ and $z\{CCC(t)\}$, 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

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

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 $\log\{TDI(p_0, t)\}$ is in order. Our initial simulation studies showed that its confidence band tended to be more accurate without the bias correction. Therefore, we
Table 1 presents the MSEs for the two approaches and their relative efficiencies for \((s_1^2, s_2^2) = (2, 2)\). 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 \(z(\text{CCC})\) 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 \((s_1^2, s_2^2) = (4, 4)\) (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 \(\mu_1, \mu_2,\)
and $\delta$ 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 $n/C^2100$. 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 $n/C^2100$.

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 $n=200$.

| $n$ | $\psi(t)$ | $\sigma$ | $\delta$ | TDI | CCC | $\sigma$ | $\delta$ | TDI | CCC |
|-----|-----------|---------|----------|-----|------|---------|----------|-----|------|
| 50  | $\mu_1$   | 97.1    | 96.8     | 96.5  | 96.3 | 97.3    | 96.6     | 97.3 | 95.3 |
|     | $\mu_2$   | 96.3    | 95.1     | 96.0  | 96.7 | 96.2    | 96.6     | 97.5 | 96.6 |
|     | $\delta$  | 96.6    | 94.7     | 96.2  | 96.1 | 96.0    | 95.9     | 96.5 | 95.7 |
|     | TDI       | 94.1    | 93.4     | 93.1  | 92.9 | 95.1    | 95.4     | 94.6 | 94.7 |
|     | CCC       | 95.4    | 95.6     | 94.9  | 95.3 | 95.5    | 95.2     | 94.2 | 95.2 |
| 100 | $\mu_1$   | 96.0    | 97.2     | 96.4  | 96.7 | 97.3    | 97.3     | 97.5 | 97.2 |
|     | $\mu_2$   | 96.1    | 96.6     | 96.1  | 96.3 | 97.1    | 97.0     | 97.0 | 97.6 |
|     | $\delta$  | 96.3    | 95.8     | 96.7  | 96.6 | 96.2    | 97.0     | 96.9 | 96.4 |
|     | TDI       | 94.9    | 95.4     | 94.6  | 94.2 | 96.5    | 95.3     | 95.0 | 95.1 |
|     | CCC       | 95.9    | 95.6     | 95.5  | 96.4 | 95.5    | 95.8     | 95.7 | 95.5 |
| 200 | $\mu_1$   | 96.3    | 96.9     | 97.2  | 96.8 | 97.4    | 97.2     | 96.8 | 97.7 |
|     | $\mu_2$   | 95.5    | 96.9     | 96.2  | 97.3 | 96.2    | 97.5     | 96.4 | 97.2 |
|     | $\delta$  | 95.6    | 96.2     | 95.8  | 96.9 | 95.8    | 96.4     | 96.0 | 96.5 |
|     | TDI       | 96.7    | 95.9     | 96.1  | 95.8 | 96.8    | 96.7     | 96.0 | 96.5 |
|     | CCC       | 95.6    | 94.9     | 96.1  | 95.9 | 96.0    | 95.8     | 95.2 | 95.5 |
| 50  | $\mu_1$   | 97.1    | 96.8     | 96.5  | 96.3 | 97.3    | 96.6     | 97.3 | 95.2 |
|     | $\mu_2$   | 96.3    | 95.1     | 96.0  | 96.7 | 96.2    | 96.6     | 97.5 | 96.6 |
|     | $\delta$  | 96.6    | 94.7     | 96.2  | 96.1 | 96.0    | 95.9     | 96.5 | 95.7 |
|     | TDI       | 94.1    | 93.4     | 93.1  | 92.9 | 95.1    | 95.4     | 94.6 | 94.7 |
|     | CCC       | 95.4    | 95.6     | 94.9  | 95.3 | 95.5    | 95.2     | 94.2 | 95.2 |
| 100 | $\mu_1$   | 96.0    | 97.2     | 96.4  | 96.7 | 97.3    | 97.3     | 97.5 | 97.2 |
|     | $\mu_2$   | 96.1    | 96.6     | 96.1  | 96.3 | 97.1    | 97.0     | 97.0 | 97.6 |
|     | $\delta$  | 96.3    | 95.8     | 96.7  | 96.6 | 96.2    | 97.0     | 96.9 | 97.0 |
|     | TDI       | 94.9    | 95.4     | 94.6  | 94.2 | 96.5    | 95.3     | 95.0 | 95.1 |
|     | CCC       | 95.9    | 95.6     | 95.5  | 96.4 | 95.5    | 95.8     | 95.7 | 95.5 |
| 200 | $\mu_1$   | 96.3    | 96.9     | 97.2  | 96.8 | 97.4    | 97.2     | 96.8 | 97.7 |
|     | $\mu_2$   | 95.5    | 96.9     | 96.2  | 97.3 | 96.2    | 97.5     | 96.4 | 97.2 |
|     | $\delta$  | 95.6    | 96.2     | 95.8  | 96.9 | 95.8    | 96.4     | 96.0 | 96.5 |
|     | TDI       | 96.7    | 95.9     | 96.1  | 95.8 | 96.8    | 96.7     | 96.0 | 96.5 |
|     | CCC       | 95.6    | 94.9     | 96.1  | 95.9 | 96.0    | 95.8     | 95.2 | 95.5 |

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).
6. Analysis of body fat data

These data from Chinchilli et al. (1996) 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 Chinchilli et al. (1996), King, Chinchilli, and Carrasco (2007), Hiriote and Chinchilli (2011), and Rathnayake and Choudhary (2017) 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 \(2/3 \times 1308 = 1308\) observations from \(n = 91\) girls. The observations range between 12.7 and 37.4. There are 56 distinct observation times on the domain \(T = [11.2, 16.8]\) 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

---

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).

| \(n\) | \(\psi(t)\) | \((2, 2)\) | \((2, 4)\) | \((4, 4)\) |
|-------|-------------|-------------|-------------|-------------|
|       | \(\mu_1\)   | \(\mu_2\)   | \(\delta\)  | \(\delta\)  |
| MPACE |             |             |             |             |
| 50    | 95.7        | 95.0        | 94.3        | 94.3        |
|       | 95.0        | 94.7        | 94.0        | 93.3        |
|       | 94.7        | 94.3        | 94.3        | 93.3        |
|       | 94.7        | 94.0        | 94.0        | 93.7        |
| TDI   |             |             |             |             |
| 100   | 93.0        | 94.7        | 92.0        | 91.7        |
|       | 94.0        | 91.0        | 91.0        | 92.7        |
|       | 96.0        | 96.7        | 95.3        | 95.7        |
|       | 94.7        | 94.0        | 94.7        | 93.3        |
| TDI   |             |             |             |             |
| 200   | 95.7        | 96.0        | 95.3        | 94.3        |
|       | 95.0        | 94.7        | 94.0        | 93.3        |
|       | 94.3        | 95.7        | 94.0        | 93.3        |
|       | 95.0        | 95.0        | 94.0        | 93.7        |
| CCC   |             |             |             |             |
| 100   | 92.7        | 93.0        | 91.7        | 91.3        |
|       | 93.3        | 92.7        | 92.7        | 91.7        |
|       | 93.3        | 92.7        | 91.3        | 92.0        |
|       | 93.3        | 92.7        | 91.3        | 92.0        |
| TDI   |             |             |             |             |
| 200   | 95.0        | 94.7        | 94.0        | 94.7        |
|       | 94.7        | 94.0        | 94.7        | 93.3        |
|       | 94.3        | 95.7        | 94.0        | 94.0        |
|       | 95.0        | 94.0        | 95.0        | 94.3        |
| CCC   |             |             |             |             |
| UPACE |             |             |             |             |
| 50    | 95.7        | 95.0        | 94.0        | 94.3        |
|       | 95.0        | 94.7        | 94.0        | 93.3        |
|       | 94.7        | 94.3        | 94.3        | 93.3        |
|       | 94.7        | 94.0        | 94.0        | 93.7        |
| TDI   |             |             |             |             |
| 100   | 95.7        | 96.0        | 95.3        | 94.2        |
|       | 94.7        | 94.0        | 94.7        | 93.7        |
|       | 95.0        | 95.0        | 94.7        | 95.3        |
|       | 94.7        | 94.7        | 94.0        | 94.3        |
| TDI   |             |             |             |             |
| 200   | 95.3        | 96.3        | 95.7        | 95.3        |
|       | 95.0        | 94.7        | 94.0        | 94.3        |
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:
\[ K = 3, \left( \hat{\lambda}_1, \hat{\lambda}_2, \hat{\lambda}_3 \right) = (97.33, 23.80, 6.11), \left( \hat{\tau}_1^2, \hat{\tau}_2^2 \right) = (2.64, 2.32), \]  
for MPACE,

\[ K = 4, \left( \hat{\lambda}_1, \hat{\lambda}_2, \hat{\lambda}_3, \hat{\lambda}_4 \right) = (102.43, 12.28, 7.61, 1.27), \left( \hat{\tau}_1^2, \hat{\tau}_2^2 \right) = (2.61, 2.34), \]  
for UPACE.

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.

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.
We now proceed as described in Sec. 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 Chinchilli et al. (1996), King, Chinchilli, and Carrasco (2007), and Rathnayake and Choudhary (2017).

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 (Choudhary and Nagaraja 2017, Chapter 3), an alternative would be to take a functional mixed-effects model approach. For example, Zhou, Huang, and Carroll (2008) 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.

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.
Appendix A: Two approaches for parameter estimation

This appendix describes two approaches based on FPCA for estimating components of \( \theta \) besides the mean functions. Estimation of mean function was discussed in Sec. 3.1. Both use the centered data \( \tilde{Y}_i(t_0), i = 1, \ldots, n \) as input.

### A.1. Approach 1 (MPACE)

This approach is an adaptation of the PACE methodology for univariate functional data (Yao, Müller, and Wang 2005; Goldsmith, Greven, and Crainiceanu 2013) to deal with multivariate functional data. A similar approach has been used by Chiou, Chen, and Yang (2014) for normalized functional data. It involves the following steps.

1. Compute the sum of products \( \sum_{i=k}^n \tilde{Y}_i(t_0)\tilde{Y}_i^T(t_0) \) using only the non-missing observations in \( \tilde{Y}_i(t_0) \). 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 \( t \in t_0 \), 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 \( \bar{V} \). It has a block structure,

\[
\bar{V} = \begin{pmatrix}
V_{11} & \cdots & V_{1J} \\
\vdots & \ddots & \vdots \\
V_{J1} & \cdots & V_{JJ}
\end{pmatrix},
\tag{A.1}
\]

where each of the submatrices is a \( N_0 \times N_0 \) matrix and \( V_{jl} = V_{lj}^T, j \neq l \). By construction, there is no missing entry in this matrix. The elements of \( V_{jl} \) provide a raw estimate of the autocovariance function \( G_{jl}(s, t) + \tau_j^2 I(s = t) \) of \( Y_j(t) \) for \( s, t \in t_0 \), see (5). Likewise, the elements of \( \bar{V}_{jl} \) provide a raw estimate of the cross covariance function \( \bar{G}_{jl}(s, t) \), given by (3), for \( s, t \in t_0 \).

2. Perform bivariate smoothing of the off-diagonal elements of \( V_{jl} \), separately for each \( j \), to obtain preliminary smooth estimates of the functions \( G_{jl}(s, t) \). Evaluate the estimated functions at \( s, t \in t_0 \) to get \( N_0 \times N_0 \) matrices \( \bar{V}_{jl}, j = 1, \ldots, J \).

3. Perform bivariate smoothing of the elements of \( \bar{V}_{jl} \), separately for each \((j, l)\) pair with \( l > j \), to obtain preliminary smooth estimates of the functions \( \bar{G}_{jl}(s, t) \). Evaluate the estimated functions at \( s, t \in t_0 \) to get \( N_0 \times N_0 \) matrices \( \bar{V}_{jl}, l > j = 1, \ldots, J \).

4. Compute the \( JN_0 \times JN_0 \) matrix \( \bar{V} \), an analog of \( V \), by replacing \( V_{jl} \) on the diagonal, \( \bar{V}_{jl} \) above the diagonal, and \( V_{jl} \) below the diagonal of \( V \) with \( \bar{V}_{jl}, \bar{V}_{jl}, \) and \( \bar{V}_{jl}^T \), respectively.

5. Use a quadrature rule (e.g., the trapezoidal rule) that approximates an integral \( \int f(t)dt \) as \( \sum_{q=1}^{N_0} w_q f(t_0) \), where the quadrature points \( t_0, \ldots, t_{N_0} \) are the elements of \( t_0 \), to get the associated weights \( w_1, \ldots, w_{N_0} \). Form a \( JN_0 \times JN_0 \) diagonal matrix \( \bar{W} \) with the entire set of weights \( (w_1, \ldots, w_{N_0}) \) repeated \( J \) times as the diagonal elements. Compute the \( JN_0 \times JN_0 \) matrix \( \bar{U} = \bar{W}^{1/2} \bar{V} \bar{W}^{1/2} \).

6. Perform a spectral decomposition of \( \bar{U} \) to get the eigenvalues \( \hat{\lambda}_k \) and the corresponding \( JN_0 \times 1 \) orthogonal eigenvectors \( \bar{u}_k, k = 1, \ldots, JN_0 \). Replace any negative eigenvalues, which may possibly be nearly zero, with zero. Choose \( \hat{K} \) as the smallest number of eigenvalues for which \( (\sum_{k=1}^{\hat{K}} \hat{\lambda}_k / \sum_{k=1}^{\hat{K}} \hat{\lambda}_k) \approx \pi \), where \( \pi \) 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 \( \bar{\phi}_k(t_0) = \bar{W}^{-1/2} \bar{u}_k \) for \( k = 1, \ldots, \hat{K} \). The eigenvalues \( \hat{\lambda}_k \) provide the estimated score variances and the corresponding vectors \( \bar{\phi}_k(t_0) = (\bar{\phi}_{kl}(t_0), \ldots, \bar{\phi}_{kJ}(t_0))^T \) provide values of the estimated eigenfunctions \( \bar{\phi}_{kl}(t), j = 1, \ldots, J \) for \( t \in t_0 \).

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

\[
\bar{G}_{jl}(s, t) = \sum_{k=1}^{\hat{K}} \hat{\lambda}_k \bar{\phi}_{kl}(s)\bar{\phi}_{kj}(t), \quad j, l = 1, \ldots, J; \quad s, t \in t_0.
\tag{A.2}
\]

8. For \( j = 1, \ldots, J \), compute \( \hat{\tau}_j^2 \) by subtracting \( \bar{G}_{jl}(t, t) \) given above from the diagonal elements of \( \bar{V}_{jj} \)—given by (A.1) and which estimate \( G_{jl}(t, t) + \tau_j^2 \)—for \( t \in t_0 \) 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 (Goldsmith et al. 2016; Happ 2018). For this, define an interval \( T^+ \subset T \) as

Software in the form of R code, together with illustration and documentation, is available at http://www.utdallas.edu/~pankaj/.
\[ T^* = \left[ \min\{t \in t_0 : t \geq t_{01} + (t_{0N_0} - t_{01})/4\}, \max\{t \in t_0 : t \leq t_{0N_0} - (t_{0N_0} - t_{01})/4\} \right], \]

and let \(|T^*|\) be its length. Also let \(t_{01}, \ldots, t_{0N_0}\) be those elements of \(t_0\) that also lie in \(T^*\). Corresponding to each \(t_{0j}\), there is a diagonal element of matrix \(V_{ij}\) in (A.1), say, \(v_{ij}(t_{0j})\). Then, as in Step 5, let \(w_1^*, \ldots, w_{J_2}^*\) be the weights associated with a quadrature rule that takes \(t_{01}, \ldots, t_{0N_0}\) as the quadrature points. Finally, take

\[ \hat{z}_j^* = \frac{1}{|T^*|} \sum_{j=1}^{Q} w_q^* \left( v_{jq}(t_{0j}) - \hat{G}_{jq}(t_{0j}, t_{0j}) \right), \]

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, Greven, and Crainiceanu (2013) and use the average difference between \(v_{jq}(t)\) and \(\hat{G}_{jq}(t, t)\) computed over the middle 60% of the grid \(t_0\). In either case, some observation times from the two ends of \(t_0\) are discarded to improve stability of the estimate.

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

\[ \hat{z}_i = \hat{A} \hat{\Phi}^T (t_i) \{ \hat{\Phi}(t_i) \hat{\Phi}^T (t_i) + R_i \}^{-1} \{ Y_i(t_i) - \tilde{\mu}(t_i) \}. \]

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

The matrix \(\hat{V}\) 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 (Bates and Maechler 2017) which implements the algorithm of Higham (2002), before continuing with the rest of the steps. This variant of the algorithm is evaluated using simulation in Sec. 5.

### A.2. Approach 2 (UPACE)

This approach is a special case of a general approach for multivariate FPCA proposed by Happ and Greven (2018). In our adaptation here, it begins with the centered data \(\tilde{Y}_i(t_0), i = 1, \ldots, n\) and involves the following steps.

1. Use the PACE methodology (Yao, Müller, and Wang 2005) 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 \(K_j^+\) as the smallest number of principal components explaining at least a specified proportion \(\pi^{(j)}\) of variability; \(\hat{\phi}_k^{(j)}(t_0)\) as the \(N_0 \times 1\) vector of values of the \(k\)th estimated eigenfunction for \(t \in t_0, k = 1, \ldots, K_j^+\); \(\hat{\tau}_j^{(j)}\) as the estimated error variance; and \(\hat{z}_j^{(j)}\) as the \(K_j^+ \times 1\) vector of estimated scores for the \(j\)th subject. Note that the corresponding true scores have expectation zero and we have a total of \(K = K_1^+ + \ldots + K_J^+\) estimated univariate scores for each subject. The \(\hat{\tau}_j^{(j)}\) resulting from this univariate FPCA also estimate the error variances in \(\theta\), i.e., \(\hat{\tau}_j^* = \hat{\tau}_j^{(j)}, j = 1, \ldots, J\).

2. Arrange the univariate scores as a \(n \times K^+\) matrix \(\hat{X}\) where

\[ \hat{X} = \begin{pmatrix} \hat{z}_1^{(1)} & \cdots & \hat{z}_1^{(J)} \\ \vdots & \ddots & \vdots \\ \hat{z}_n^{(1)} & \cdots & \hat{z}_n^{(J)} \end{pmatrix} \]

and form their \(\hat{X}^+ \times \hat{X}^+\) covariance matrix \(Z = \hat{X}^T \hat{X}/(n - 1)\).

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

4. For \(k = 1, \ldots, K\), represent the \(K \times 1\) eigenvector \(\hat{z}_k\) as \(\hat{z}_k = (\hat{z}_k^{(1)} \cdots \hat{z}_k^{(J)})^T\), where \(\hat{z}_k^{(j)}\) is a \(K^+ \times 1\) vector with elements \(\hat{z}_{kj}^{(j)}, j = 1, \ldots, J\). Estimate the values of the eigenfunction \(\hat{\phi}_j(t)\) in \(\theta\) for \(t \in t_0\) as the \(N_0 \times 1\) vector
\[
\hat{\phi}_k(t_0) = (\hat{\phi}_1(t_0), ..., \hat{\phi}_K(t_0))\theta_k.
\]

5. Take \( \hat{\xi}_i \), the estimated score vector in the model (11), as the \( i \)th row of the \( n \times K \) matrix \( \Xi(z_1, ..., z_K) \).

To conclude, both MPACE and UPACE approaches provide estimates of all the components of \( \theta \) except the mean functions whose estimation was discussed in Sec. 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—\( \pi \) for multivariate data. On the other hand, UPACE involves choosing \( J + 1 \) such proportions—\( \pi^{(1)}, ..., \pi^{(J)} \) for univariate data and \( \pi \) 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 (Ramsay and Silverman 2005, Chapter 8). We have chosen 0.99 in this work. A variant of the UPACE algorithm with \( \pi^{(j)} = 1 \) for \( j = 1, ..., J \) is evaluated using simulation in Sec. 5.

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.

ORCID

Pankaj K. Choudhary http://orcid.org/0000-0002-0398-7459

References

Barnhart, H. X., M. J. Haber, and L. I. Lin. 2007. An overview on assessing agreement with continuous measurements. *Journal of Biopharmaceutical Statistics* 17 (4):529–69. doi:10.1080/10543400701376480.

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

Berrendero, J. R., A. Justel, and M. Svarc. 2011. Principal components for multivariate functional data. *Computational Statistics & Data Analysis* 55 (9):2619–34. doi:10.1016/j.csda.2011.03.011.

Bland, J. M., and D. G. Altman. 1986. Statistical methods for assessing agreement between two methods of clinical measurement. The Lancet 327 (8476):307–10. doi:10.1016/S0140-6736(86)90837-8.

Chinchilli, V. M., J. K. Martel, S. Kumanyika, and T. Lloyd. 1996. A weighted concordance correlation coefficient for repeated measurement designs. *Biometrics* 52 (1):341–53. doi:10.2307/2533172.

Chou, J. M., Y. T. Chen, and Y. F. Yang. 2014. Multivariate principal component analysis: A normalization approach. *Statistica Sinica* 24:1571–96. doi:10.5705/ss.2013.305.

Choudhary, P. K., and H. N. Nagaraja. 2017. *Measuring Agreement: Models, Methods, and Applications*. Hoboken, NJ: Wiley.

Dunn, G. 2007. Regression models for method comparison data. *Journal of Biopharmaceutical Statistics* 17 (4):739–56. doi:10.1080/10543400701329513.

Goldsmith, J., S. Greven, and C. Crainiceanu. 2013. Corrected confidence bands for functional data using principal components. *Biometrics* 69 (1):41–51. doi:10.1111/j.1541-0420.2012.01808.x.

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

Happ, C., and S. Greven. 2018. Multivariate functional principal component analysis for data observed on different (dimensional) domains. *Journal of the American Statistical Association* 113 (522):649–59. doi:10.1080/01621459.2016.1273115.

Happ, C. 2018. *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.

Higham, N. 2002. Computing the nearest correlation matrix | A problem from finance. *IMA Journal of Numerical Analysis* 22 (3):329–43. doi:10.1093/imanum/22.3.329.
Hiriote, S., and V. M. Chinchilli. 2011. Matrix-based concordance correlation coefficient for repeated measures. *Biometrics* 67 (3):1007–16. doi:10.1111/j.1541-0420.2010.01549.x.

Hothorn, T., F. Bretz, and P. Westfall. 2008. Simultaneous inference in general parametric models. *Biometrical Journal. Biometrische Zeitschrift* 50 (3):346–63. doi:10.1002/bimj.200810425.

Jacques, J., and C. Preda. 2014. Model-based clustering for multivariate functional data. *Computational Statistics & Data Analysis* 71:92–106. doi:10.1016/j.csda.2012.12.004.

King, T. S., V. M. Chinchilli, and J. L. Carrasco. 2007. A repeated measures concordance correlation coefficient. *Statistics in Medicine* 26 (16):3095–113. doi:10.1002/sim.2778.

Li, R., and M. Chow. 2005. Evaluation of reproducibility for paired functional data. *Journal of Multivariate Analysis* 93 (1):81–101. doi:10.1016/j.jmva.2004.01.010.

Lin, L. I. 1989. A concordance correlation coefficient to evaluate reproducibility. *Biometrics* 45 (1):255–68. Corrections: 2000, 56:324–25. doi:10.2307/2532051.

Lin, L. I. 2000. Total deviation index for measuring individual agreement with applications in laboratory performance and bioequivalence. *Statistics in Medicine* 19 (2):255–70. doi:10.1002/(SICI)1097-0258(20000130)19:2<255::AID-SIM293>3.0.CO;2-8.

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

Ramsay, J. O., and B. W. Silverman. 2005. *Functional Data Analysis*. 2nd ed. New York: Springer.

Rathnayake, L. N., and P. K. Choudhary. 2017. Semiparametric modeling and analysis of longitudinal method comparison data. *Statistics in Medicine* 36 (13):2003–15. doi:10.1002/sim.7261.

Wood, S. N. 2017. *Generalized Additive Models: An Introduction with R*. 2nd ed. Boca Raton, FL: Chapman & Hall/CRC.

Yao, F., H. G. Müller, and J. L. Wang. 2005. Functional data analysis for sparse longitudinal data. *Journal of the American Statistical Association* 100 (470):577–90. doi:10.1198/016214504000001745.

Zhou, L., J. Z. Huang, and R. J. Carroll. 2008. Joint modelling of paired sparse functional data using principal components. *Biometrika* 95 (3):601–19. doi:10.1093/biomet/asn035.