Published in final edited form as:
Proc AAAI Conf Artif Intell. 2019 ; 33: 4320–4327. doi:10.1609/aaai.v33i01.33014320.

MFPCA: Multiscale Functional Principal Component Analysis

Zhenhua Lin,
University of California, Davis, One Shields Avenue, Davis, CA 95616, linzh@ucdavis.edu

Hongtu Zhu
University of North Carolina at Chapel Hill, Chapel Hill, NC 27599, htzhu@email.unc.edu

Abstract

We consider the problem of performing dimension reduction on heteroscedastic functional data where the variance is in different scales over entire domain. The aim of this paper is to propose a novel multiscale functional principal component analysis (MFPCA) approach to address such heteroscedastic issue. The key ideas of MFPCA are to partition the whole domain into several subdomains according to the scale of variance, and then to conduct the usual functional principal component analysis (FPCA) on each individual subdomain. Both theoretically and numerically, we show that MFPCA can capture features on areas of low variance without estimating high-order principal components, leading to overall improvement of performance on dimension reduction for heteroscedastic functional data. In contrast, traditional FPCA prioritizes optimizing performance on the subdomain of larger data variance and requires a practically prohibitive number of components to characterize data in the region bearing relatively small variance.

1 Introduction

Functional principal component analysis (FPCA) is a key tool for performing dimension reduction on functional data that features infinite dimensionality and emerges in many machine learning applications (Ghebreab, Smeulders, and Adriaans 2008; Ramsay and Silverman 2005; Ferraty and Vieu 2006; Hsing and Eubank 2015). In classic FPCA referred to as *single-scale* FPCA in this paper, a single eigen-analysis is conducted for observed functions on the *entire* domain. Specifically, FPCA is built on eigen-analysis of the covariance function of functional data, analogous to the covariance matrix, from which we derive functional principal components. For the purpose of dimension reduction, only the principal components corresponding to the first few largest eigenvalues are retained. Related works include early development of FPCA (Rao 1958; Dauxois, Pousse, and Romain 1982; Besse and Ramsay 1986), and more recent advances such regularization techniques (Rice and Silverman 1991; Silverman 1996), estimation and theory (Yao, Müller, and Wang 2005; Hall, Müller, and Wang 2006; Li and Hsing 2010; Zhang and Wang 2016) for functional data that are sparsely observed, and interpretability (Chen and Lei 2015; Lin, Wang, and Cao 2016), to name a few. All of these works adopt the single-scale paradigm, which is briefly described in the next section.

Such one-size-fits-all scheme, as we show in Section 2, inevitably comes with a side effect that the leading principal components prioritize producing good approximation quality for functions on the subdomain that holds large data variance. However, for heteroscedastic functional data, where the variance of the data has substantially different scales over the domain, single-scale FPCA approach often needs a large number of principal components in order to characterize the behavior of the data on the subdomain with relatively small variance. However, it is notoriously difficult to accurately estimate high-order principal components. For example, for a fixed sample size, the estimation error for the kth principal components is approximately of the order k^2 (Mas and Ruymgaart 2015). This quadratic decay in estimation quality prohibits one from obtaining reliable estimates of principal components for a moderate sample size except for the first few leading ones. Consequently, for heteroscedastic functional data, single-scale FPCA may not be able to discover useful features in the area of low variance.

To address the issue of FPCA for heteroscedastic functional data, we consider a question, which states

“Could we modify standard FPCA to efficiently deal with heteroscedastic functional data?”

Our solution to the above question is the development of a novel *multiscale* FPCA framework. The key ideas of MFPCA are to partition the whole domain into several subdomains in the way that the variance of the data within a subdomain is approximately on the same scale, and then to conduct FPCA on each subdomain separately. Compared with the existing methods in the literature, three major methodological contributions in this paper are as follows:

- Our MFPCA approach is simple and yet powerful. Specifically, compared to single-scale FPCA, the multiscale approach alleviates the issue of estimating high-order principal components and leads to overall improvement of data representation, since fewer components are required for sufficient approximation of the data within each subdomain.
- Our MFPCA is able to discover useful patterns in the area of low variance, as demonstrated in the analysis of the brain microstructure in relation to multiple sclerosis using diffusion tensor imaging techniques; see Section 4 for details.
- Theoretically, we show that MFPCA has larger capacity of representation for functional data and yields higher estimation quality for principal components than singlescale FPCA. The theoretical analysis is complemented by numerical simulation in Section 4.

2 Classic Functional Principal Component Analysis

Without loss of generality, let X be a random process defined on an interval I such that $\mathbf{E}\|X\|^2_{\mathscr{L}_2} < \infty$ and $\mathbf{E}X = 0$. FPCA is built on the concept of Karhunen-Loève expansion of X. Specifically, $X(t)$ admits the following representation

$$X(t) = \sum_{k=1}^{\infty} \xi_k \phi_k(t), \qquad (1)$$

where ξ_k are centered uncorrelated random variables, $\phi_k(\cdot)$ are normalized eigenfunctions of the covariance function $\mathscr{C}(s,t) = \mathrm{Cov}\{X(s), X(t)\}$, and $\mathrm{Var}(\xi_k)$, often denoted by λ_k, is the eigenvalue of $\mathscr{C}$ corresponding to ϕ_k, i.e.,

$$\int_{I} \mathscr{C}(\,\cdot\,, t)\phi_k(t)\mathrm{d}t = \lambda_k \phi_k(\,\cdot\,)\,. \tag{2}$$

It is also conventionally assumed that $\lambda_1 > \lambda_2 > \cdots > 0$. The eigenfunctions $\phi_1(\cdot)$, $\phi_2(\cdot), \ldots$, referred to as principal components in the context of FPCA, form an orthonormal basis of $\mathscr{L}^2(I)$.

The above FPCA is a single-scale approach in the sense that both (1) and (2) are defined on the entire interval I. It has the following global optimality of Karhunen-Loève expansion for approximation of X using the first few eigenfunctions. For any fixed $K > 0$, let $\mathbf{B} = \{B_1, \ldots, B_K\}$ be a collection of K orthonormal real-valued functions in $\mathscr{L}^2(I)$, and define P_{B} the projection operator of the space spanned by $B_1, \ldots, B_K$, i.e., $P_{\mathrm{B}}X = \sum_{k=1}^{K} \langle B_k, X \rangle B_k$, where $\langle B_k, X \rangle = \int_I B_k(t)X(t)\mathrm{d}t$. In principle, $P_{\mathrm{B}}X$ is the approximation of X using the basis functions $B_1, \ldots, B_K$. The approximation error is often measured by $\varepsilon(\mathbf{B}) = \mathbf{E}(\|X - P_{\mathbf{B}}X\|^2_{\mathscr{L}^2})$. It can then be shown that $\boldsymbol{\varepsilon}(\boldsymbol{\Phi}) \le \boldsymbol{\varepsilon}(\mathbf{B})$ for $\boldsymbol{\Phi} = \{\phi_1, \ldots, \phi_K\}$ and any $\mathbf{B}$ of K orthonormal functions in $\mathscr{L}^2(I)$.

Although such optimality of single-scale FPCA is attractive for approximating functional data, it has an undesired side-effect for functional data with different scales of variability on the interval I. To elaborate, we divide I into J equal subintervals $I_1, \ldots, I_J$ that form a partition of I, in the sense that $\bigcup_{j=1}^{J} I_j = I$ and $I_j \cap I_k = \varnothing$ for $1 \le j \ne k \le J$. Denote $X^{(j)}(t)$ as the restriction of $X(t)$ to the subinterval I_j, i.e., $X^{(j)}(t) = X(t)$ if $t \in I_j$ and $X^{(j)}(t) = 0$ if $t \notin I_j$.

Then, we have

$$\begin{aligned}
\varepsilon(\mathbf{B}) &= \mathbf{E}\left(\|X - P_{\mathbf{B}}X\|^2_{\mathscr{L}^2}\right) \\
&= \sum_{j=1}^{J} \mathbf{E}\left(\left\|X^{(j)} - (P_{\mathbf{B}}X)^{(j)}\right\|^2_{\mathscr{L}^2}\right) \\
&= \mathbf{E}\left(\|X\|^2_{\mathscr{L}^2}\right) \times \\
&\quad \sum_{j=1}^{J} \frac{\mathbf{E}\left(\left\|X^{(j)} - (P_{\mathbf{B}}X)^{(j)}\right\|^2_{\mathscr{L}^2}\right)}{\mathbf{E}\left(\left\|X^{(j)}\right\|^2_{\mathscr{L}^2}\right)} \frac{\mathbf{E}\left(\left\|X^{(j)}\right\|^2_{\mathscr{L}^2}\right)}{\mathbf{E}\left(\|X\|^2_{\mathscr{L}^2}\right)},
\end{aligned}$$

where $(P_{\mathrm{B}}X)^{(j)}$ denotes the restriction of $P_{\mathrm{B}}X$ to I_j. Let $w_j = \mathbf{E}\left(\left\|X^{(j)}\right\|^2_{\mathscr{L}^2}\right) \big/ \mathbf{E}\left(\left\|X\right\|^2_{\mathscr{L}^2}\right)$ for all $j \le J$, which are constants that are independent of $\mathbf{B}$. We interpret the weights w_j as the variance density of X on I_j. Then, finding $\mathbf{B}$ to optimize $\boldsymbol{\varepsilon}(\mathbf{B})$, which the single-scale FPCA does, prioritizes minimizing the approximation error for the pieces with larger weights, i.e., with relatively larger variance density. Therefore, in order to capture features of pieces that have small weights, one needs a large number of principal components. As the principal components ϕ_k are unknown, one needs to estimate them from a finite sample. However, it is difficult to estimate the high-order principal components, in light of the fact that any estimate for ϕ_k based on n independently and identically distributed (i.i.d.) samples of X cannot have an estimation error less than ck^2/n for some constant $c > 0$ (Mas and Ruymgaart 2015). Thus, if one applies single-scale FPCA to $X(t)$ that exhibits multiscale variability over I, then the features in the area with low variance density might be concealed by those in the region of high variance density.

3 Multiscale Karhunen-Loève Expansion

To overcome the aforementioned shortcoming of singlescale FPCA, for heteroscedastic functional data, we adopt the following simple divide-and-conquer strategy to conduct FPCA: divide the interval I into several subintervals according to the scale of variance on I, and then perform FPCA on each subinterval. To be more precise, let $I_1,\ldots,I_J$ and $X^{()},\ldots,X^{(J)}$ be defined as in Section 2. Applying single-scale FPCA to each $X^{(j)}$ on I_j, we can obtain the Karhunen-Loève approximation $X^{(j)}(t) \approx \sum_{k=1}^{K_j} \xi_k^{(j)} \phi_k^{(j)}(t)$, where $\phi_k^{(j)}$ is the kth leading eigenfunction of the multiscale covariance function $\mathscr{C}^{(j)}$ that is defined by $\mathscr{C}^{(j)}(s,t) = \mathrm{Cov}\{X^{(j)}(s)X^{(j)}(t)\}$ on the square $I_j \times I_j$. The approximation for X is then obtained by

$$X(t) \approx \sum_{j=1}^{J} \sum_{k=1}^{K_j} \xi_k^{(j)} \phi_k^{(j)}(t)\,. \tag{3}$$

Although there are $\sum_{j=1}^{J} K_j$ terms in (3), for each $t \in I_j$, only K_j of them are nonzero as $\phi_k^{(j)}$ is zero outside I_j. The number of principal components, K_j required for a good approximation of $X^{(j)}$ can vary among the subintervals, which allows for the adaptive approximation of X over different subintervals. For example, if sample paths of X are rough on I_j, then a relatively larger number of principal components can be used to obtain a good approximation, while for relatively smooth pieces, $X^{(k)}$, a small number might be sufficient. In contrast, single-scale FPCA uses the same number of principal components for each $t \in I$ and hence does not enjoy such extra adaptivity. Note that (3) automatically takes into account the fact that the sub-domains where the data have small variance contribute less than others to approximating the process $X(t)$, by observing that the principal component scores $\xi_k^{(j)}$ have different scales of variance on different sub-domains. For those sub-domains with small variance, the scale of the corresponding principal component scores is also small. Nevertheless, the features of a small scale could be influential in supervised learning tasks, as demonstrated in Section 4.

As stated previously, single-scale FPCA has the global optimality property, which can be equivalently stated as $\mathbf{E}\|P_\Phi X\|_{\mathscr{L}^2}^2 \geq \mathbf{E}\|P_{\mathrm{B}} X\|_{\mathscr{L}^2}^2$, where $\mathbf{\Phi}$ and $\mathbf{B}$ are defined in Section 2. The quantity $\mathbf{E}\|P_\Phi X\|_{\mathscr{L}^2}^2$ is often interpreted as the amount of variance explained by the principal components in $\mathbf{\Phi}$. The optimality property essentially asserts that the first K eigenfunctions explain the variance of X more than any other K orthonormal functions for any $K \geq 1$. However, by exploring local adaptivity, our MFPCA enjoys larger capacity of representation of functional data, i.e., it can explain larger variance than single-scale FPCA when the same number of principal components is used to approximate $X(t)$ for each $t \in I$. The following proposition illustrates this point for $K = 1$.

Proposition 1. Let $I1,\ldots, I_j$ *form a disjoint partition of I. Suppose that X is a random process satisfying* $\mathbf{E}\|X\|^2_{\mathscr{L}^2} < \infty$ *and (without loss of generality)* $\mathbf{E}X = 0$. *Let* $\mathscr{S}_j$ *be the collection of function f such that* $\|f\|^2_{\mathscr{L}^2} = 1$ *and* $\mathrm{support}(f) \subset I_j$. *Then*

$$\mathbf{E}\|P_{\phi_1}X\|^2_{\mathscr{L}^2} = \mathbf{E}\left\langle X, \phi_1\right\rangle^2 \le \sup_{\psi^{(j)} \in \mathcal{S}_j} \sum_{j=1}^{J} \mathbf{E}\left\langle X, \psi^{(j)}\right\rangle^2 .$$

The equality is possible only if $\mathbf{E}\langle X, \phi_1^{(j)}\rangle = \mathbf{E}\langle X, \phi_1^{(k)}\rangle$ *for all* $1 \le j, k \le J$, *where* $\phi_1^{(j)}$ *denotes the restriction of the single-scale eigenfunction* $\boldsymbol{\phi}_1$ *onto* I_j.

Proof. It is easy to see that the conclusion holds for $J > 2$ if it is true for $J = 2$. Suppose $\|\phi_1^{(1)}\|_{\mathscr{L}^2} > 0$ and $\|\phi_1^{(2)}\|_{\mathscr{L}^2} > 0$; otherwise, the conclusion is trivial. Set $\tilde{\phi}_1 = \phi_1^{(1)} / \|\phi_1^{(1)}\|_{\mathscr{L}^2}$ and $\tilde{\phi}_2 = \phi_1^{(2)} / \|\phi_1^{(2)}\|_{\mathscr{L}^2}$. Then

$$\begin{aligned}
&\mathbf{E}\langle X, \phi_1\rangle^2 = \mathbf{E}\langle X, \phi_1^{(1)} + \phi_1^{(2)}\rangle^2 \\
&= \mathbf{E}\langle X, \phi_1^{(1)}\rangle^2 + \mathbf{E}\langle X, \phi_1^{(2)}\rangle^2 + 2\mathbf{E}(\langle X, \phi_1^{(1)}\rangle\langle X, \phi_1^{(2)}\rangle) \\
&= \|\phi_1^{(1)}\|^2_{\mathscr{L}^2}\mathbf{E}\langle X, \tilde{\phi}_1\rangle^2 + \|\phi_1^{(2)}\|^2_{\mathscr{L}^2}\mathbf{E}\langle X, \tilde{\phi}_2\rangle^2 \\
&\quad + 2\|\phi_1^{(1)}\|_{\mathscr{L}^2}\|\phi_1^{(2)}\|_{\mathscr{L}^2}\mathbf{E}(\langle X, \tilde{\phi}_1\rangle\langle X, \tilde{\phi}_2\rangle) \\
&\le \|\phi_1^{(1)}\|^2_{\mathscr{L}^2}\mathbf{E}\langle X, \tilde{\phi}_1\rangle^2 + \|\phi_1^{(2)}\|^2_{\mathscr{L}^2}\mathbf{E}\langle X, \tilde{\phi}_2\rangle^2 \\
&\quad + \|\phi_1^{(1)}\|^2_{\mathscr{L}^2}\mathbf{E}\langle X, \tilde{\phi}_2\rangle^2 + \|\phi_1^{(2)}\|^2_{\mathscr{L}^2}\mathbf{E}\langle X, \tilde{\phi}_1\rangle^2 \\
&= \mathbf{E}\langle X, \tilde{\phi}_1\rangle^2 + \mathbf{E}\langle X, \tilde{\phi}_2\rangle^2,
\end{aligned}$$

where the last equality is due to $\|\phi_1^{(1)}\|^2_{\mathscr{L}^2} + \|\phi_1^{(2)}\|^2_{\mathscr{L}^2} = 1$. The equality holds if $\mathbf{E}\langle X, \phi_1^{(1)}\rangle = \mathbf{E}\langle X, \phi_1^{(2)}\rangle$. The fact that $\tilde{\phi}_1 \in \mathscr{S}_1$ and $\tilde{\phi}_2 \in \mathscr{S}_2$ concludes the proof. □

Principal components are unknown and need to be estimated from data in practice. Given n i.i.d. samples $X_1, X_2,\ldots, X_n$, the single-scale covariance function $\mathscr{C}$ is estimated by its empirical version $\widehat{\mathscr{C}}(s, t) = n^{-1}\sum_{i=1}^{n} X_i(s)X_i(t)$. The kth single-scale principal component $\boldsymbol{\phi}_k$ is then estimated by the kth principal component $\hat{\phi}_k$ of $\hat{C}$. The quality of the estimator $\hat{\phi}_k$ is quantified by the error $\mathbf{E}\|P_{\hat{\phi}_k} - P_{\phi_k}\|^2_\infty$, where $\|\cdot\|_\infty$ denotes the operator norm on $\mathscr{L}^2(\boldsymbol{I})$. It is shown that $\mathbf{E}\|P_{\hat{\phi}_k} - P_{\phi_k}\|^2_\infty \ge ck^2 / n$ for all $k \ge 1$ and a universal constant $c > 0$ (Mas and Ruymgaart 2015). In Section 2, we point out that singlescale FPCA focuses on the region of high variance density. For heteroscedastic functional data, if one needs to learn features in the area of low variance density, then a potentially large number of principal components are required. However, due to the quadratic growth in estimation error, it is notoriously difficult to obtain reliable estimates for high-order principal components.

The difficulty of handling high-order principal components might be alleviated or even avoided if one takes a multiscale perspective. To give a more concrete example, we assume $X = \sum_{j=1}^{J}\sum_{k=1}^{j}\xi_k^{(j)}\phi_k^{(j)} + X_\perp$, where $\phi_k^{(j)}$ are orthonormal, $X_\perp$ is orthogonal to all $\phi_k^{(j)}$, and $\xi_k^{(j)}$ are uncorrelated and centered random variables with variance $\lambda_k^{(j)} \equiv \mathbf{E}(\xi_k^{(j)})^2 \asymp \{J(J+1)/2 - j(j-1)/2 + (k-j)\}^{-\alpha}$ for some constant $\alpha > 1$, i.e., $\lambda_1^{(J)} > \lambda_2^{(J)} > \cdots \lambda_J^{(J)} > \lambda_1^{(J-1)} > \lambda_2^{(J-1)} > \cdots > \lambda_1^{(1)}$. Also, if $\varphi \in \mathscr{L}^2(I)$ and $\|\varphi\|_{\mathscr{L}^2} = 1$, then $\mathbf{E}\langle X_\perp, \varphi\rangle^2 < \min\{\lambda_k^{(j)}: j = 1, \ldots, J, k = 1, \ldots, j\} = \lambda_1^{(1)}$. In other words, $\phi_k^{(j)}$ are the first $J(J + 1)/2$ principal components of X. If one applies single-scale FPCA to obtain an estimate $\hat{\phi}_{k,S}^{(j)}$ for each principal component $\phi_k^{(j)}$, then the estimation error for $\phi_1^{(1)}$ is $\mathbf{E}\|P_{\hat{\phi}_{1,S}^{(1)}} - P_{\phi_{1,S}^{(1)}}\|_\infty^2 \geq cJ^2(J+1)^2/(4n)$. For multiscale FPCA, given the partition $I_1, I_2, \ldots, I_J$, the multiscale estimator for $\mathscr{C}^{(j)}$ is given by its empirical version $\widehat{\mathscr{C}}^{(j)}(s,t) = \frac{1}{n}\sum_{i=1}^n X_i^{(j)}(s)X_i^{(j)}(t)$ and the principal component $\phi_k^{(j)}$ is estimated by the kth leading principal components $\hat{\phi}_{k,M}^{(j)}$ of $\widehat{\mathscr{C}}^{(j)}$. In particular, $\hat{\phi}_{1,M}^{(1)}$ is the first leading principal component of $\widehat{\mathscr{C}}^{(1)}$. The estimation error of $\phi_{1,M}^{(1)}$ is $\mathbf{E}\|P_{\hat{\phi}_{1,M}^{(1)}} - P_{\phi_{1,M}^{(1)}}\|_\infty^2 \leq c\log^2 n/n$ (Mas and Ruymgaart 2015), which contrasts with the error $cJ^2(J+1)^2/(4n)$ for the single-scale estimate $\hat{\phi}_{1,S}^{(1)}$ when $J \gg \log n$.

To conduct the dimension reduction in (3), X is approximated by its projection onto the principal components $\phi_k^{(j)}$ for $k = 1,2,\ldots, K_j$ and $j = 1,2,\ldots, J$ simultaneously. This projection, denoted by P, is equivalent to the projection onto the linear space $span\{\phi_k^{(j)}: k = 1, 2, \ldots, K_j, j = 1, 2, \ldots, J\}$. In practice, it is estimated by the projection $\hat{P}_M$ onto the space $\mathrm{span}\{\hat{\phi}_{k,M}^{(j)}: k = 1, 2, \ldots, K_j, j = 1, 2, \ldots, J\}$. The estimation error of $\hat{P}_M$ is bounded in the following theorem.

Theorem 2. *Let* $\lambda_1^{(j)} > \lambda_2^{(j)} > \cdots$ *be eigenvalues of* $\mathscr{C}^{(j)}$, *which satisfy* $c_1 j^{-1-\alpha} \leq \lambda_k^{(j)} \leq c_2 j^{-1-\alpha}$ *for some constants* c_1, $c_2 > 0$ *and for all j and k. Then, we have*

$$E\|\hat{P}_M - P\|_\infty^2 \leq \frac{2c\log^2 n}{n}\sum_{j=1}^J K_j^2, \tag{4}$$

where $\hat{P}_M$ *is the multiscale estimate of P based on n i.i.d. samples, and* $c > 0$ *is a universal constant.*

Proof. Let $P^{(j)}$ be the projection onto the linear space $\mathrm{span}\{\phi_k^{(j)}: k = 1, 2, \ldots, K_j\}$. Note that $\mathbf{E}\|\hat{P}_M^{(j)} - P^{(j)}\|_\infty^2 \leq cK_j^2\log^2 n/n$ (Mas and Ruymgaart 2015), where $\hat{P}_M^{(j)}$ is the multiscale estimate of $P^{(j)}$. Then (4) follows from the fact $\mathbf{E}\|\hat{P}_M - P\|_\infty^2 \leq 2\sum_{j=1}^J \mathbf{E}\|\hat{P}_M^{(j)} - P^{(j)}\|_\infty^2$. □

Apply the theorem to the above example, we see that the estimation error of $\hat{P}_M$ is bounded by $2cn^{-1}\log^2 n\sum_{j=1}^J K_j^2 \leq 2cJ^3n^{-1}\log^2 n$. In particular, if we use $P^{(j)}$ and $P_k^{(j)}$ to denote the projection operators for the linear spaces $\mathrm{span}\{\phi_k^{(j)}: k = 1, 2, \ldots, K_j\}$ and $\mathrm{span}\{\phi_k^{(j)}\}$, respectively, and denote their multiscale estimates by $\hat{P}_M^{(j)}$ and $\hat{P}_{k,M}^{(j)}$, respectively, then $\sup_{Q \in \mathscr{P}}\mathbf{E}\|\hat{Q}_M - Q\|_\infty^2 \quad \leq \quad 2cJ^3\log^2 n/n$, where $\mathscr{P} = \{P\} \cup \{P^{(j)}: j = 1, \ldots, J\} \cup \{P_k^{(j)}: k = 1, 2, \ldots, K_j, j = 1, 2, \ldots, J\}$, and $\hat{Q}_M$ is the multiscale estimator for Q. However, for single-scale FPCA, one has $\sup_{Q \in \mathscr{P}}\mathbf{E}\|\hat{Q}_S - Q\|_\infty^2 \quad \geq \quad \mathbf{E}\left\|\hat{P}_{1,S}^{(1)} - P_1^{(1)}\right\|_\infty^2 > cJ^4/(4n)$ which is larger than the multiscale one when $J \gg \log^2 n$, where $\hat{Q}_S$ is the single-scale estimator for Q. Note that the bounds in the above derivation are quite loose. The practical advantage of the multiscale approach could emerge for small J, as numerically illustrated by the simulation studies presented in the next section.

To apply multiscale FPCA, one needs to find a good partition of I. A simple and yet effective approach is to segment I according to the variance function $V(t) = \mathrm{Var}[X(t)]$ and in practice its empirical version $\hat{V}(t) = (n-1)^{-1}\sum_{i=1}^{n}\{X_i(t)\}^2$. Practical functional data are often only recorded at some discrete points of I subject to potential measurement noise, rather than fully observed. There are two types of functional data, including dense data and sparse data. Dense functional data refer to the scenario that each X_i is recorded on a common, regular and dense grid $t_1,\ldots,t_m$ of I, where m denotes the number of observations for each subject X_i; while sparse functional data refer to the case that X_i are measured on an irregular and subject-specific sparse grid $t_{i1},\ldots, t_{im_i}$, where m_i is the number of measurements for X_i. For dense data, the variance function at t_j can be estimated by $\hat{V}(t_j) = \frac{1}{n-1}\sum_{i=1}^{n}\{X_i(t_j)\}^2$. For sparse data, such a simple estimate does not work, and we adopt local linear smoothing (Zhang and Wang 2016) on the observations to obtain the estimate $\hat{V}(t) = \hat{b}_{t0}$ with

$$(\hat{b}_{t0}, \hat{b}_{t1}) = \underset{(b_{t0}, b_{t1}) \in \mathbb{R}^2}{\arg\min} \sum_{i=1}^{n} \frac{1}{m_i} \sum_{j=1}^{m_i} \mathscr{K}\left(\frac{t_{ij}-t}{h}\right) \times [\{X_i(t_{ij})\}^2 - b_{t0} - b_{t1}(t_{ij}-t)\}]^2,$$

where $\mathscr{K}$ is a kernel function supported on $[-1,1]$, and $h > 0$ is the bandwidth to be chosen by cross-validation. Therefore, for both dense and sparse functional data, we can obtain the estimate of $V(t)$ on a dense and regular grid $t_1,\ldots,t_m$ of I. To derive a partition on I for MFPCA, we apply a multiple change point detection method (Niu and Zhang 2012; Frick, Munk, and Sieling 2014) on $\hat{V}(t_1), \ldots, \hat{V}(t_m)$ to discover the change points of $\hat{V}(t)$ and partition the interval I according to those points, or we may apply the propagation-separation method to partition I into disjoint regions (Polzehl and Spokoiny 2006; Spokoiny and Vial 2009; Zhu, Fan, and Kong 2014). These change point detection methods also provide data-driven approaches to select the number of change points, which determines the number J of subintervals of the partition. Alternatively, one can visualize the empirical variance function and then determine the number of change points manually, which is often sufficient and effective in practice. The number K_j of principal components can be selected by a threshold (e.g. 95%) on the fraction of variance explained (FVE) for the jth subinterval, $j = 1,2,\ldots,J$. The complete algorithm is given below.

Algorithm 1 (MFPCA). *Suppose that $X_1,\ldots,X_n$ are functional data on a common interval I, either given in the dense form or sparse form.*

1. *Obtain the estimate $\hat{V}(t)$ on a dense and regular grid $t_1 < \ldots < t_m$ of I.*
2. *Apply a multiple change point detection method on $\hat{V}(t_1), \ldots, \hat{V}(t_m)$. Suppose there are $J — 1$ change points that are denoted by $T_1 < \ldots < T_{J-1}$. Partition I into J subintervals according to $T_1 < \ldots < \mathrm{T}_{J-1}$. Denote the subintervals by $I_1,\ldots,I_J$.*

3. *Apply single-scale FPCA on each I_j, with data $X_1^{(j)}, \ldots, X_n^{(j)}$, recalling that $X_i^{(j)}$ denotes the restriction of X_i to I_j. Suppose* $\widehat{\lambda}_k^{(j)}$ for $j = 1, 2, \ldots, K_j$ *and* $j = 1, 2, \ldots, J$ *are the first $K = \sum_{j=1}^{J} K_j$ empirical eigenvalues from all subintervals I_j and $\widehat{\phi}_k^{(j)}$ are their corresponding eigenfunctions. Each function X_i is approximated by $X_i(t) \approx \sum_{j=1}^{J} \sum_{k=1}^{K_j} \widehat{\xi}_{i,k}^{(j)} \widehat{\phi}_k^{(j)}(t)$, where the scores $\widehat{\xi}_{i,k}^{(j)}$ are computed by* $\int_{I_j} \widehat{\phi}_k^{(j)}(t) X_i^{(j)}(t) \mathrm{d}t$.

The number of overall principal components, K, can be determined by a fractional variance explained threshold, i.e., the smallest number of leading components required to account for at least 95% of the variance of the data. In practice, the multiscale features might be buried by the first singlescale principal component ϕ_1, i.e., the variance function of X does not exhibit multiscale features, but the variance function of $X - P_{\phi_1} X$ might. In this case, we can estimate the first single-scale principal component, denoted by $\widehat{\phi}_{1,S}$, and then apply the above multiscale algorithm to the residuals $X_i - P_{\widehat{\phi}_{1,S}} X_i$. Indeed, this procedure can be applied recursively to each partition to form hierarchical FPCA which is a multiscale system. In such a hierarchical structure, different layers represent FPCA that is performed at different scales, with the bottom layer corresponding to the finest.

4 Numerical Illustration

We illustrate the estimation of eigenfunctions ϕ and projection P_Φ via $n = 50$ and $n = 200$ simulated samples from a Gaussian process on the interval [0,1] with mean function $\mu(t) = e^{-(t-0.1)^2/0.003}$ — $2e^{-(t+0.1)^2/0.008} + e^{-(t-0.95)^2/0.01} - 0.5e^{-(t-0.7)^2/0.012}$, variance function $\sigma_0^2(t) = e^{-(t-0.05)^2/0.01}(t + 0.05) + 1.5e^{-(0.95-t)^2/0.015}(1.05 - t)$ and Matérn correlation function $\rho(s,t) = 2^{1-\nu}(\sqrt{2\nu}\,|s-t|)^{\nu} B_\nu(\sqrt{2\nu}\,|s-t|)/\Gamma(\nu)$, where B_ν is the modified Bessel function of the second kind of order ν, Γ is the gamma function, and we set $\nu = 0.1$. The mean function and variance function are specifically designed to mimic the shape of the mean function and variance function in the data application that follows. For the purpose of comparison, we also include the case of variance function $\sigma_1^2(t) \equiv 1$, which might favor single-scale FPCA. We repeat each simulation setting $N = 100$ times independently. The estimation quality of an estimator $\widehat{\phi}$ for an eigenfunction ϕ is quantified by the Monte Carlo root mean squared error (RMSE), defined by $(\widehat{\phi}) = \frac{1}{N}\sum_{r=1}^{N} \|\widehat{\phi}_{(r)} - \phi\|_{\mathscr{L}^2}$, while for an estimator $\widehat{P}$ of the projection P, the Monte Carlo RMSE is given by $(\widehat{P}) = \frac{1}{N}\sum_{r=1}^{N} \|\widehat{P}_{(r)} - P\|_\infty$, where $\widehat{\phi}_{(r)}$ and $\widehat{P}_{(r)}$ are the estimates produced in the rth simulation replication. The results, shown in Figure 1 and 2, suggest that the multiscale approach produces better estimation quality for eigenfunctions and projections, with more prominent advantage when data exhibit a multiscale feature, such as in the case of $\sigma^2(t) = \sigma_0^2(t)$.

We apply multiscale FPCA to analyze the brain microstructure in the corpus callosum of healthy subjects and patients with multiple sclerosis (MS). MS is a common demyelinating disease that is often caused by immune-mediated inflammation. More precisely, demyelination refers to damage to myelin, an insulating material that covers nerve cells, protects axons and helps nerve signal to travel faster. Patient with MS suffer from demyelination that occurs in the white matter of the brain and can lose mobility or even cognitive function (Jongen, Ter Horst, and Brands 2012). Myelin damage in the brain can be examined by diffusion tensor imaging (DTI), which produces highresolution images of white matter tissues by tracing water diffusion within the tissues. Fractional anisotropy of water diffusion can be determined from DTI and evaluated in relation to MS (Ibrahim et al. 2011).

The DTI dataset we used in the following analysis was collected at Johns Hopkins University and the Kennedy-Krieger Institute. It consists of data from $n_1 = 340$ MS patients and $n_2 = 42$ healthy subjects. All fractional anisotropy profiles are recorded on a common grid of 93 points, as shown in the left panel of Figure 3. It seems that these profiles have larger variance at both ends relative to that in the middle. This is confirmed by the cross-sectional variance that is plotted in the right panel in Figure 3. We see that the variance function can be divided into three regions, including [0, 20], [20, 70] and [70,93], with relatively small variance density at [20,70]. This motivates us to apply the proposed multiscale FPCA on these regions separately. The first three estimated eigenfunctions of each region are shown in Figure 4, as well as the first three single-scale eigenfunctions. For comparison, the restriction of these single-scale eigenfunctions to each region is normalized to unit $\mathscr{L}^2$ norm. We observe that both the multiscale FPCA and single-scale FPCA yield similar results for the first principal component on each region, but differ substantially for the second and third components. For example, the first principal component on the region [20,70] is almost a straight line, which can be interpreted as a size component, i.e., subjects with a positive score have larger fractional anisotropy than an average subject on locations between 20 and 70. However, for the second component, the multiscale analysis is able to capture additional features of the data on the region [20,70], while the single-scale analysis still yields almost a straight line that is similar to the depiction of its first component within that region. This agrees with our analysis in Section 2: as fractional anisotropy has low variance density on [20,70], single-scale FPCA prioritizes other regions.

It is of interest to see whether the additional patterns uncovered by MFPCA are useful for subsequent analysis, such as predicting MS status based on fractional anisotropy. To answer this question, we study classification of MS patients and healthy subjects using a random forest classifier based on the scores derived from the first K principal components. To account for the imbalance of the data, we adopt a simple undersampling strategy to train the classifier as follows. We randomly sample 42 subjects among all 342 MS patients without replacement, and combine those data with the data from the 42 healthy subjects to form a new dataset, which we then randomly divide into two equal halves. The classifier is trained on one half, while the correct classification rate is computed on the other half. We repeat the above procedure 100 times independently for each $K = 1,2,\ldots, 12$, where 12 is the number of components required to explain over 95% of the variance of the data. In reality, K could be determined by cross-validation or BIC criterion. We also compare the proposed method to wavelet transformation which is capable of localizing signals in time domain. Specifically, each fractional anisotropy profile is represented by a set of coefficients with respect to Daubechies' least asymmetric wavelets (Daubechies 1992), and wavelet coefficients are fed into a random forest classifier. Different vanishing moments are considered, namely, from db1 to db6. Roughly speaking, wavelets of more vanishing moments can represent more complex function with a sparser set of wavelet coefficients.

The results, shown in Table 1, clearly suggest that MFPCA outperforms single-scale FPCA for all values of K except $K = 3$. We note that the first three multiscale components, $\phi_1^{(2)}$, $\phi_1^{(3)}$ and $\phi_1^{(1)}$, are the first component from each region. As previously stated, the first multiscale principal component and the single-scale counterpart for each region are almost identical, which might explain why the performance of classification for both methods is quite close when $K \leq 3$. A possible explanation for the observation that the multiscale method performs worse when $K = 3$ is that the third multiscale principal component might not be strongly related to MS status. In such case, including the component in the model does not reduce prediction bias, but increases prediction variability and thus reduces the correct classification rate.

As high-order components come into play, in particular when the fifth component $\phi_2^{(3)}$, the sixth component $\phi_2^{(2)}$ and the seventh component $\phi_3^{(1)}$ are included, the advantage of MFPCA becomes more prominent. In contrast, for singlescale FPCA, the classification performance barely improves when more principal components are added. This demonstrates that MFPCA is able to discover useful features that might not be captured by single-scale FPCA in regions of low variance density. The MFPCA method also outperforms the wavelet method, likely due to the fact that principal components derived from MFPCA are innately data-driven, while wavelet bases are not. Such data-driven feature allows a parsimonious representation of functional data and is attractive in practice. For instance, in the above simulated functional data, on average 12 principal components are sufficient to account for 95% of variation of data, while over 50 wavelets are required to achieve a similar level of representation.

5 Concluding Remark

We have presented a multiscale FPCA method for functional data and demonstrated that it is numerically superior to its single-scale counterpart and the well known wavelet transformation, thanks to its data-driven nature and the ability to localize signals in time domain. It is worth of noting that while principal component scores from the same segment are uncorrelated, those from different segments could be correlated. Such correlation indeed accounts for the correlation structure among different segments. This is distinct from block-diagonal structures where functional data from different segments are uncorrelated or even independent for Gaussian processes. It is also of interest to apply the proposed MFPCA method to machine learning tasks other than classification, such as regression and clustering, which is one of our future research topics.

Acknowledgments

Dr. Zhu's work was partially supported by NIH grants MH086633 and MH116527.

References

Besse P, and Ramsay JO 1986. Principal components analysis of sampled functions. Psychometrika 51(2):285–311.

Chen K, and Lei J 2015. Localized functional principal component analysis. Journal of the American Statistical Association 110:1266–1275. [PubMed: 26806987]

Daubechies I 1992. Ten Lectures on Wavelets. CBMSNSF Regional Conference Series in Applied Mathematics. SIAM.

Dauxois J; Pousse A; and Romain Y 1982. Asymptotic theory for the principal component analysis of a vector random function: some applications to statistical inference. Journal of Multivariate Analysis 12(1):136–154.

Ferraty F, and Vieu P 2006. Nonparametric Functional Data Analysis: Theory and Practice. New York: Springer-Verlag.

Frick K; Munk A; and Sieling H 2014. Multiscale change-point inference. Journal of the Royal Statistical Society: Series B (Statistical Methodology) 76:495–580.

Ghebreab S; Smeulders A; and Adriaans P 2008. Predicting brain states from fmri data: Incremental functional principal component regression. In Platt JC; Koller D; Singer Y; and Roweis ST, eds., Advances in Neural Information Processing Systems 20. Curran Associates, Inc. 537–544.

Hall P; Müller HG; and Wang JL 2006. Properties of principal component methods for functional and longitudinal data analysis. The Annals of Statistics 34:1493–1517.

Hsing T, and Eubank R 2015. Theoretical Foundations ofFunctional Data Analysis, with an Introduction to Linear Operators. Wiley.

Ibrahim I; Tintera J; Skoch A; F., J.; P., H.; Martinkova P; Zvara K; and Rasova K 2011. Fractional anisotropy and mean diffusivity in the corpus callosum of patients with multiple sclerosis: the effect of physiotherapy. Neuroradiology 53(11):917–926. [PubMed: 21556863]

Jongen P; Ter Horst A; and Brands A 2012. Cognitive impairment in multiple sclerosis. Minerva Medica 103(2):73–96. [PubMed: 22513513]

Li Y, and Hsing T 2010. Uniform convergence rates for nonparametric regression and principal component analysis in functional/longitudinal data. The Annals of Statistics 38:3321–3351.

Lin Z; Wang L; and Cao J 2016. Interpretable functional principal component analysis. Biometrics 72(3):846–854. [PubMed: 26683051]

Mas A, and Ruymgaart F 2015. High-dimensional principal projections. Complex Analysis and Operator Theory 9(1):35–63.

Niu YS, and Zhang H 2012. The screening and ranking algorithm to detect DNA copy number variations. The Annals of Applied Statistics 6(3):1306–1326. [PubMed: 24069112]

Polzehl J, and Spokoiny VG 2006. Propagation-separation approach for local likelihood estimation. Probab. Theory Relat. Fields 135:335–362.

Ramsay JO, and Silverman BW 2005. Functional Data Analysis. Springer Series in Statistics. New York: Springer, 2nd edition.

Rao CR 1958. Some statistical methods for comparison of growth curves. Biometrics 14(1):1–17.

Rice JA, and Silverman BW 1991. Estimating the mean and covariance structure nonparametrically when the data are curves. Journal of the Royal Statistical Society. Series B 53(1):233–243.

Silverman BW 1996. Smoothed functional principal com¬ponents analysis by choice of norm. The Annals of Statistics 24(1):1–24.

Spokoiny V, and Vial C 2009. Parameter tuning in point-wise adaptation using a propagation approach. Ann. Statist 37:2783–2807.

Yao F; Müller HG; and Wang J-L 2005. Functional data analysis for sparse longitudinal data. Journal of the American Statistical Association 100:577–590.

Zhang X, and Wang JL 2016. From sparse to dense functional data and beyond. The Annals of Statistics 44:2281–2321.

Zhu H; Fan J; and Kong L 2014. Spatially varying coefficient models with applications in neuroimaging data with jumping discontinuity. Journal of American Statistical Association 109:977–990.

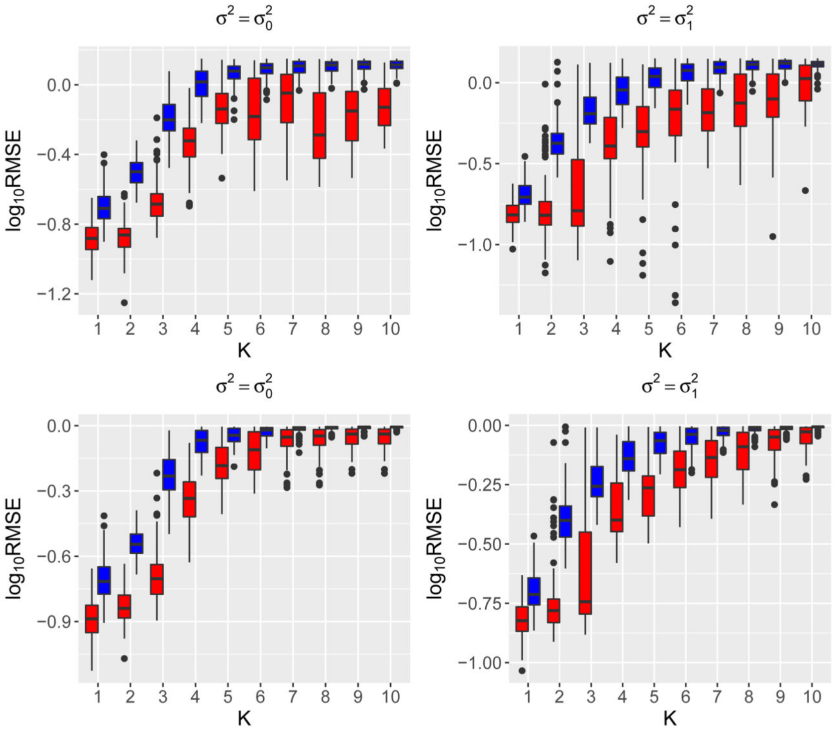

Figure 1:

$\log_{10}$(RMSE) of the multiscale (red) and singlescale (blue) estimators for the first 10 eigenfunctions (top panels) and projection (bottom panels) onto the first k leading components for $k = 1, 2, \ldots, 10$ for $n = 50$.

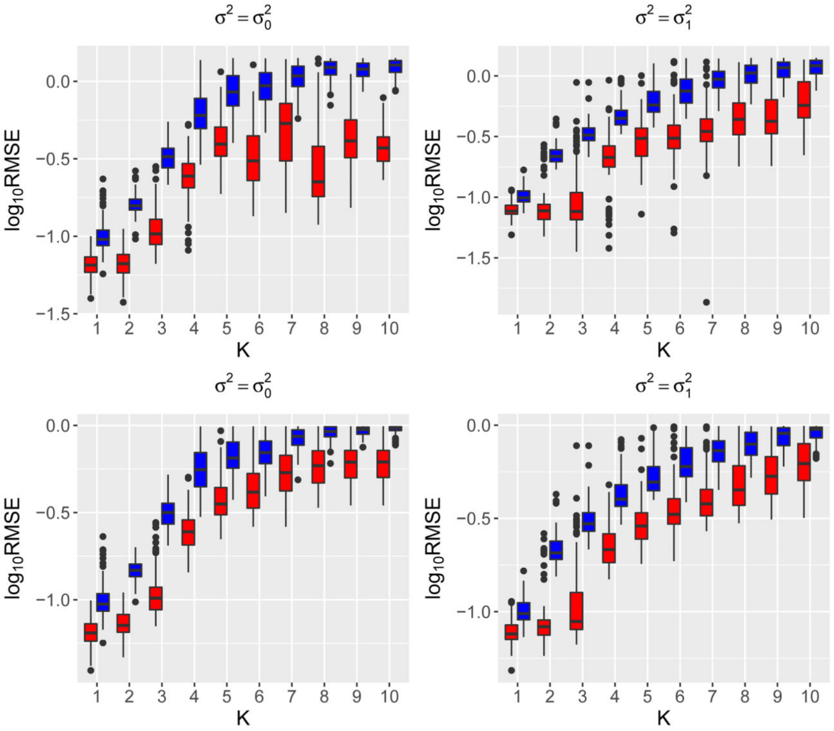

Figure 2:

$\log_{10}$(RMSE) of the multiscale (red) and singlescale (blue) estimators for the first 10 eigenfunctions (top panels) and projection (bottom panels) onto the first *k* leading components for *k* = 1, 2,…, 10 for *n* = 200.

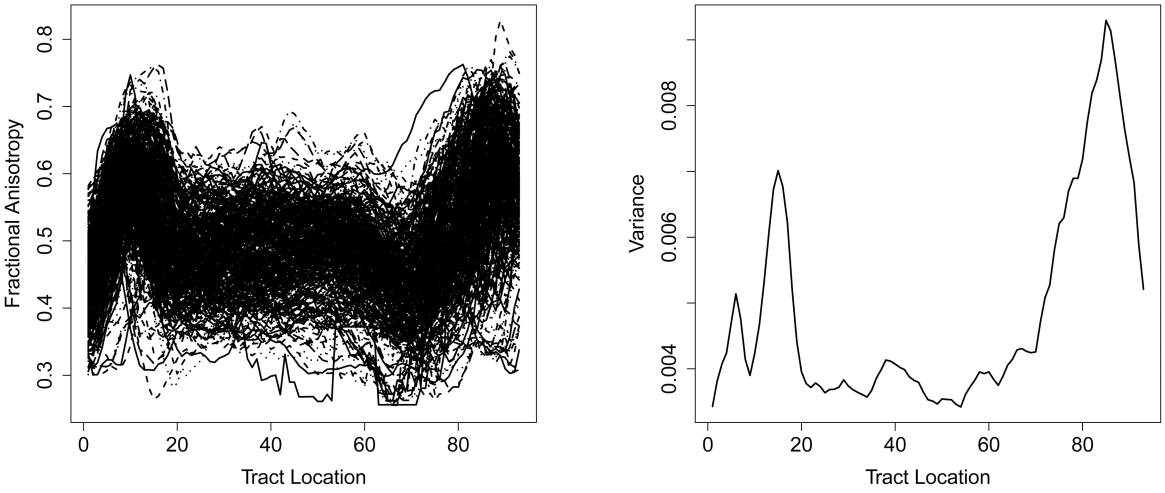

Figure 3:

Left panel: fractional anisotropy profile of 382 subjects. Right panel: empirical variance function of fractional anisotropy profile.

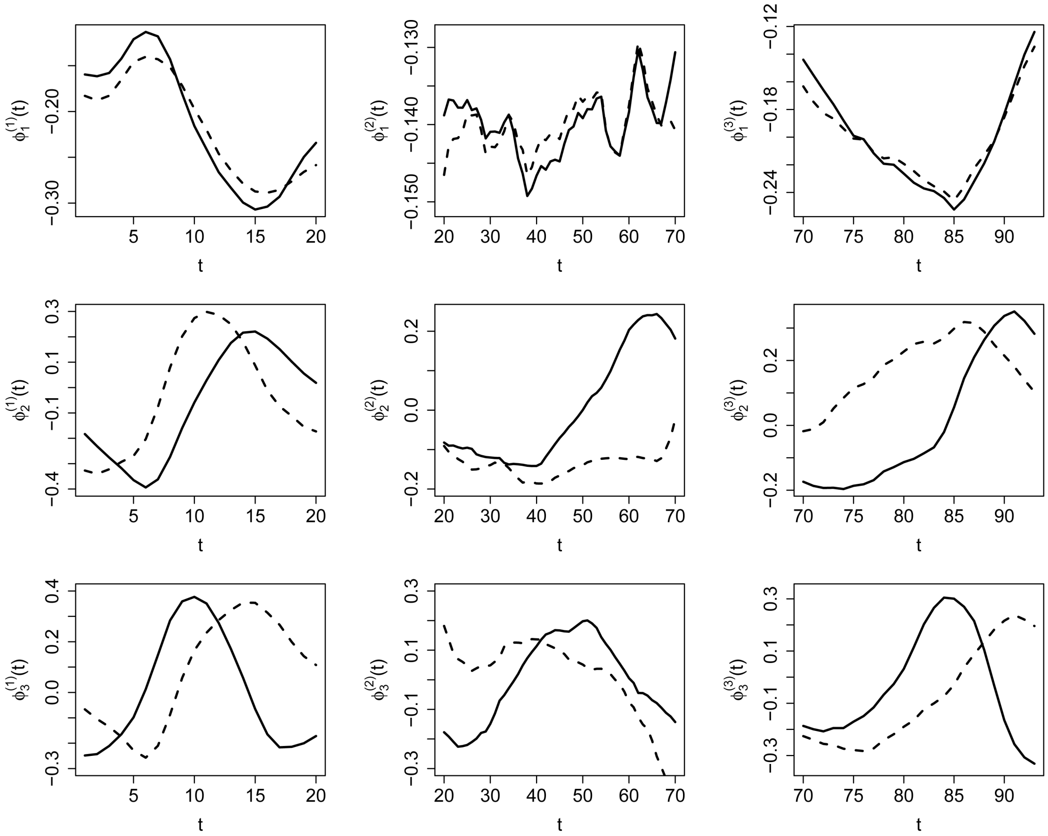

Figure 4:

The first 3 principal components of each partition by multiscale FPCA (solid) and single-scale (dashed) FPCA.

Table 1:

Correct classification rates of the random forest classifier trained on wavelet basis coefficients, multiscale principal component scores and single-scale principal component scores, respectively.

K	1	2	3	4	5	6
Multiscale	66.2	71.0	69.9	70.5	72.9	73.4
Single-scale	64.7	69.8	70.5	69.8	70.7	70.9

K	7	8	9	10	11	12
Multiscale	74.5	73.5	73.9	72.5	72.8	72.6
Single-scale	71.1	70.9	71.1	69.5	69.7	69.9

	db1	db2	db3	db4	db5	db6
wavelet	69.5	70.8	68.6	70.1	69.5	69.1