Localized Functional Principal Component Analysis

Kehui Chen and Jing Lei

University of Pittsburgh and Carnegie Mellon University

January 15, 2015

Abstract

We propose localized functional principal component analysis (LFPCA), looking for orthogonal basis functions with localized support regions that explain most of the variability of a random process. The LFPCA is formulated as a convex optimization problem through a novel Deflated Fantope Localization method and is implemented through an efficient algorithm to obtain the global optimum. We prove that the proposed LFPCA converges to the original FPCA when the tuning parameters are chosen appropriately. Simulation shows that the proposed LFPCA with tuning parameters chosen by cross validation can almost perfectly recover the true eigenfunctions and significantly improve the estimation accuracy when the eigenfunctions are truly supported on some subdomains. In the scenario that the original eigenfunctions are not localized, the proposed LFPCA also serves as a nice tool in finding orthogonal basis functions that balance between interpretability and the capability of explaining variability of the data. The analyses of a country mortality data reveal interesting features that cannot be found by standard FPCA methods.

KEYWORDS: functional principal component analysis, domain selection, interpretability, orthogonality, deflation, convex optimization

1. INTRODUCTION

Functional principal component analysis has emerged as a major tool to explore the source of variability in a sample of random curves and has found wide applications in functional regression, curve classification, and clustering (Castro et al., 1986; Rice and Silverman, 1991; Cardot, 2000; Yao et al., 2005; Ramsay and Silverman, 2005; Hall and Hosseini-Nasab, 2006). In this paper, we consider functional principal component analysis with localized support regions. That is, for a smooth random function X, we look for orthogonal basis functions with localized support regions that explain most of the variance. The main motivation of the localized functional principal component analysis (LFPCA) is to find a parsimonious linear representation of the data that balances the interpretability and the capability of explaining variance of the stochastic process.

The proposed method outputs localized basis functions whose localization level is controlled by a localization tuning parameter. We propose two methods to select the localization parameter, corresponding to two useful applications of our proposed method. First, one can choose the localization parameter by maximizing the explained variance of the random process computed by V-fold cross-validation. Our simulation shows that when the eigenfunctions truly have localized support regions, the proposed LFPCA with a localization parameter chosen by cross-validation significantly improves the estimation accuracy of the eigenfunctions compared to standard FPCA methods. On the other hand, when the original eigenfunctions are not localized, the localization parameter chosen by cross-validation is expected to be very close to zero and it is confirmed by our numerical studies that the performance of the proposed LFPCA is almost identical to standard FPCA methods. The second method of choosing the localization parameter is to seek the most localized basis functions that explain a fixed level of variance. This method is particularly useful when the standard eigenfunctions are not localized, and it makes sense for the proposed LFPCA not to target at the standard eigenfunctions but to balance between interpretability and the capability of explaining variance of the stochastic process. Details can be found in Section 3.2. We

consider a a country mortality data application to illustrate the second method of choosing the localization parameter. In the country mortality data, the mortality rates at age 60 were recorded from year 1960–2006 for 27 countries around the world. The first three localized basis functions, explaining more than 85% variance, correspond to variational modes around mid 1990s, 1980s, and 1960s, respectively. Another example concerning a growth curve data is presented in the Online Supplementary File. The first two localized basis functions explain more than 85% variance, and clearly indicate that the main variational modes of female height growth are around age 12 and around age 5, perfectly matching the knowledge of growth spurts. These interesting features cannot be revealed by standard FPCA.

Domain localization has been studied by several authors in the functional regression model: $E(Y|X) = a + \int_{\mathcal{T}} X(t)\beta(t)dt$, where the coefficient function $\beta(t)$ is desired to be zero outside a subdomain $\mathcal{T}_0 \in \mathcal{T}$ with the purpose of improved interpretability (James et al., 2009; Zhao et al., 2012; Zhou et al., 2013). Most of these methods turn the problem into a variable selection problem and use LASSO type penalties. In a recent thesis work Lin (2013), interpretable functional principal component analysis is studied, which has a very similar flavor as our proposed LFPCA. In their work, an ℓ_0 penalty is added on eigenfunctions and a greedy algorithm based on basis expansion of curves is developed to solve the non-convex optimization problem. The formulation of localization or domain selection in the context of functional principal component analysis is quite challenging for at least two reasons. First, the eigen problem together with a localization penalty is usually not convex, and in general it is an NP-hard problem to find a global optimum. Second, in order to obtain a sequence of mutually orthogonal eigen-components, a commonly taken procedure is to deflate the empirical covariance operator at step j by removing the effect of the previous $j-1$ components (White, 1958; Mackey, 2008). But with the localization penalty in the objective function, this procedure can not guarantee the orthogonality of such sequentially obtained eigen-components. In sequential estimation of principal components, being orthogonal to the first component is a natural requirement when looking for the second component, otherwise the maximization over second direction is not well-defined since the solution would still

be the first direction. From a dimension reduction perspective, the orthogonality is also appealing since the resulting k dimensional orthogonal basis leads to very simple calculation for subsequent inferences.

The main contribution of this paper is three-fold. First, we formulate the LFPCA as a convex optimization problem with explicit constraint on the orthogonality of eigen-components. Second, we provide an efficient algorithm to obtain the global maximum of this convex problem. Third, we carefully investigate the estimation error from the discretized data version to the functional continuous version, as well as the complex interaction between the eigen problem and the localization penalty, and prove consistency of the estimated eigenfunctions. The starting point of our method is a sup-norm consistent estimator of the covariance operator, up to a constant shift on the diagonal. For dense and equally spaced observations with or without measurement error, the proposed method can be directly carried out on the sample covariance, i.e., without the need to perform basis expansion, smoothing of the individual curves, or smoothing of the estimated covariance operator. For other designs of functional data, the proposed method is still applicable when an appropriate covariance estimator is available.

Our formulation of LFPCA borrows ideas from recent developments in sparse principal component analysis. In Vu et al. (2013); Lei and Vu (2015), a similar convex framework based on Fantope Projection and Selection has been proposed to estimate a k dimensional sparse principal subspace of a high dimensional random vector (see also d'Aspremont et al. (2007) for $k = 1$). These sparse subspace methods are useful when the union of the support regions of several leading eigenvectors is sparse. In sparse PCA settings, the notion of sparsity requires the proportion of non-zero entries in the leading eigenvectors to vanish as the dimensionality increases and therefore it makes sense to consider the union of the support regions of several leading eigenvectors to be sparse. However, in functional data settings, the length ratio of a support subdomain over the entire domain is determined by the random curve model and usually a constant, and the union of several leading subdomains can be as large as the entire domain. This is also the reason that we use the notion "localized" instead

of "sparse". It has remained challenging to obtain sparse eigenvectors sequentially that each one is allowed to have a different support region. A particular challenge is the interaction between orthogonality and the sparse penalty. Besides the difference between functional PCA and sparse PCA, one main extension developed in our method is the construction of a deflated Fantope to estimate individual eigen-components sequentially, with possibly different support regions and guaranteed orthogonality. This deflated Fantope formulation is of independent interest in many other structured principal component analysis.

The rest of this paper is organized as follows. In section 2, we introduce the formulation of localized functional principal component analysis. Section 3 derives the solution to the optimization problem and describes the algorithm as well as the selection of tuning parameters. Section 4 contains the consistency results. Section 5 and Section 6 present numerical experiments and data examples to illustrate our method. Section 7 contains some discussions and extensions. Technical details and additional materials are provided in the Online Supplementary File.

2. LFPCA THROUGH DEFLATED FANTOPE LOCALIZATION

We consider a square integrable random process $X(t) : \mathcal{T} \mapsto \mathbb{R}$ over a compact interval $\mathcal{T} \subset \mathbb{R}$, with mean and covariance functions $\mu(t) = EX(t)$ and $\Gamma(s,t) = \text{Cov}(X(s), X(t))$, and covariance operator $(\Gamma f)(t) = \int_{s \in \mathcal{T}} f(s)\Gamma(t,s)ds$. Under the minimal assumption that $\Gamma(s,t)$ is continuous over (s,t), this operator Γ has orthonormal eigenfunctions $\phi_j(t), j = 1, 2, \cdots$, with nonincreasing eigenvalues λ_j, satisfying $\Gamma\phi_j = \lambda_j\phi_j$. The well known Karhunen-Loève expansion then gives the representation

$$X(t) = \mu(t) + \sum_{j=1}^{\infty} \xi_j \phi_j(t), \tag{1}$$

where ξ_j, $j \geq 1$, is a sequence of uncorrelated random variables satisfying $E(\xi_j) = 0$ and $\text{var}(\xi_j) = \lambda_j$, with an explicit representation $\xi_j = \int_{t \in \mathcal{T}} (X(t) - \mu(t))\phi_j(t)dt$. A key inference task in functional principal component analysis (FPCA) is to estimate the leading eigenfunctions $\phi_j(t)$, $1 \leq j \leq k$, from n independent sample curves.

4

In practice, the underlying sample curves $X_i(t)$, $1 \leq i \leq n$, are usually recorded at a grid of points and may be contaminated with additive measurement errors. We start with dense and equally spaced observations

$$Y_{il} = X_i(t_l) + \epsilon_{il} = \mu(t_l) + \sum_{j=1}^{\infty} \xi_{ij} \phi_j(t_l) + \epsilon_{il}, \quad i = 1, \ldots, n, \quad l = 1, \ldots, p, \qquad (2)$$

where ϵ_{il} are independent noises with mean zero and variance σ^2, and t_l, $1 \leq l \leq p$, are grid points in $\mathcal{T}$ at which observations are recorded. Starting from such discrete and possibly noisy data, there are different ways of introducing smoothness in the estimation of FPCA. Rice and Silverman (1991); Silverman (1996); Huang et al. (2008), among others, studied approaches where a smoothness penalty on eigenfunctions is integrated in the optimization step of eigen-decomposition. Let S be the $p \times p$ sample covariance matrix of the observed vector Y, and v be a p dimensional vector. Rice and Silverman (1991) used a roughening matrix $D = \Delta^T \Delta$ where $\Delta \in \mathbb{R}^{(p-2) \times p}$ is a second-differencing operator:

$$\Delta_{ij} = \begin{cases} 1, & \text{if } j \in \{i, i+2\}, \\ -2, & \text{if } j = i+1, \\ 0, & \text{otherwise}. \end{cases}$$

A smoothed eigenfunction estimator is obtained by solving the following eigen problem:

$$\max v^T (S - \rho_1 D) v, \quad \text{s.t. } \|v\|_2 = 1, \qquad (3)$$

where ρ_1 is a smoothing parameter, and $\|v\|_2$ is the Euclidean norm of v.

A straightforward approach to localize the estimated eigenfunctions is to add another localization penalty:

$$\max v^T (S - \rho_1 D) v - \rho_2 \|v\|_1, \quad \text{s.t. } \|v\|_2 = 1, \qquad (4)$$

where ρ_2 is a tuning parameter, and $\|v\|_1$ is the ℓ_1 norm of v. However, this is not a convex problem and there are no known algorithms that can efficiently find a global optimum even for the first eigen-component.

Here we propose a novel sequential estimation procedure based on the idea of estimating the rank one projection matrix vv^T. Let $\langle A, B \rangle = \text{trace}(A^T B)$ for matrices A, B of compatible dimensions. Denote $\|H\|_{1,1}$ the matrix ℓ_1 norm, which is the sum absolute value of all entries in H. Starting from $\widehat{\Pi}_0 = 0$, for each $j = 1, ..., k$, the jth localized eigen-component is estimated as follows.

$$H_j = \arg\max \langle S - \rho_1 D, H \rangle - \rho_2 \|H\|_{1,1}, \ \text{s.t.} \ H \in \mathcal{D}_{\widehat{\Pi}_{j-1}},$$
$$\widehat{v}_j = \text{ the first eigenvector of } H_j, \tag{5}$$
$$\widehat{\Pi}_j = \widehat{\Pi}_{j-1} + \widehat{v}_j \widehat{v}_j^T,$$

where, for any $p \times p$ projection matrix Π,

$$\mathcal{D}_\Pi := \{H : 0 \preceq H \preceq I, \ \text{trace}(H) = 1, \ \text{and} \ \langle H, \Pi \rangle = 0\},$$

and, for symmetric matrices A, B, "$A \preceq B$" means that $B - A$ is positive semidefinite.

Problem (5) is a convex relaxation of (4) with integrated orthogonality constraints on the estimated localized eigen-components. With the estimated $\widehat{v}_j$, we can easily obtain an estimate $\widehat{\phi}_j(t)$ of the localized eigenfunction $\phi_j(t)$ by standard interpolation techniques such as linear interpolation, plus an optional final step of re-orthogonalization and re-normalization.

With appropriately chosen tuning parameters, the performance of the proposed method ties to the maximum entry-wise error of the discretized covariance estimator S (see Section 4 for details). Our presentation will focus on a sample covariance S that is computed from dense and equally spaced observations, but the proposed method is not restricted to a dense regular design as long as a reasonable covariance estimate can be obtained. More discussions can be found in Section 7.

To solve for $\widehat{v}_j$ and $\widehat{\phi}_j(t)$, the key step in (5) is to solve

$$\max_H \langle S - \rho_1 D, H \rangle - \rho_2 \|H\|_{1,1}, \ \text{s.t.} \ H \in \mathcal{D}_\Pi, \tag{6}$$

where $\Pi = \widehat{\Pi}_{j-1}$ at step j. In next section we present an algorithm that solves problem (6), with a discussion on the choice of tuning parameters ρ_1 and ρ_2.

In the sparse PCA literature, d'Aspremont et al. (2007); Vu et al. (2013) have considered the following problem,

$$\max_{H}\langle S, H\rangle - \rho\|H\|_{1,1}, \text{ s.t. } H \in \mathcal{F}^d, \tag{7}$$

where the convex set $\mathcal{F}^d := \{H : 0 \preceq H \preceq I, \text{ trace}(H) = d\}$ is called the *Fantope* of degree d (Dattorro, 2005), and is the convex hull of all rank d projection matrices. While the convex relaxation given in (7) allows us to estimate d-dimensional sparse principal subspaces, it does not lead to a sequence of mutually orthogonal eigenvectors with different support regions.

To ensure orthogonality among the estimated eigenvectors, we consider $\mathcal{D}_\Pi := \{H : H \in \mathcal{F}^1, \text{ and } \langle H, \Pi\rangle = 0\}$, which we call the *deflated Fantope*. It can be naturally generalized to $\mathcal{D}_\Pi^d := \{H : H \in \mathcal{F}^d, \text{ and } \langle H, \Pi\rangle = 0\}$ to estimate mutually orthogonal principal subspaces. Such a feasibility deflation technique is quite different from the commonly suggested matrix deflation techniques in sequential estimation of eigenvectors (see Mackey (2008) for example).

3. ALGORITHM

3.1 Deflated Fantope Localization using ADMM

The main difficulty in solving problem (6) is the complex interaction between the ℓ_1 penalty and the deflated Fantope constraint. To overcome this difficulty, we write (6) in an equivalent form to separate the ℓ_1 penalty and deflated Fantope constraint:

$$\min_{H,Z} \mathbb{I}_{\mathcal{D}_\Pi}(H) - \langle S - \rho_1 D, H\rangle + \rho_2\|Z\|_{1,1},$$
$$\text{s.t. } H - Z = 0, \tag{8}$$

where $\mathbb{I}_{\mathcal{D}_\Pi}$ is the convex indicator function, which is ∞ outside $\mathcal{D}_\Pi$ and 0 inside $\mathcal{D}_\Pi$. Problem (8) is a convex global variable consensus optimization, which can be solved using alternating direction method of multipliers (ADMM, Boyd et al. (2011)). We describe in Algorithm 1 an ADMM algorithm that solves (8) and hence (6). It extends the FPS algorithm in Vu et al. (2013) to the deflated Fantope.

The two matrix operators used in the algorithm are defined as follows.

Algorithm 1 Deflated Fantope Localization using ADMM

Require: $S = S^T$, Π, D, $\rho_1, \rho_2 \geq 0$, $\tau > 0$, $\epsilon > 0$

   $Z^{(0)} \leftarrow 0, W^{(0)} \leftarrow 0$ ▷ Initialization

   repeat $r = 1, 2, \ldots$

      $H^{(r)} \leftarrow \mathcal{P}_{\mathcal{D}_\Pi} \left[Z^{(r-1)} - W^{(r-1)} + (S - \rho_1 D)/\tau \right]$ ▷ Deflated Fantope projection

      $Z^{(r)} \leftarrow \mathcal{S}_{\rho_2/\tau} \left(H^{(r)} + W^{(r-1)} \right)$ ▷ Elementwise soft thresholding

      $W^{(r)} \leftarrow W^{(r-1)} + H^{(r)} - Z^{(r)}$ ▷ Dual variable update

   until $\|H^{(r)} - Z^{(r)}\|_F^2 \vee \tau^2 \|Z^{(r)} - Z^{(r-1)}\|_F^2 \leq \epsilon^2$ ▷ Stopping criterion

   return $Z^{(r)}$

(i) Soft-thresholding operator: for any $a > 0$,

$$\mathcal{S}_a(x) = \text{sign}(x) \max(|x| - a, 0).$$

(ii) Deflated-Fantope-projection operator: For any $p \times p$ symmetric matrix A and projection matrix Π,

$$\mathcal{P}_{\mathcal{D}_\Pi}(A) := \arg \min_{B \in \mathcal{D}_\Pi} \|A - B\|_F^2$$

is the Frobenius norm projection of A onto the deflated Fantope $\mathcal{D}_\Pi$.

A non-trivial subroutine in Algorithm 1 is to calculate the deflated-Fantope-projection $\mathcal{P}_{\mathcal{D}_\Pi}(A)$ for a symmetric matrix A. The following lemma gives a close-form characterization of the deflated-Fantope-projection operator.

Lemma 3.1. *Let $\Pi = VV^T$, where V is a $p \times d$ matrix with orthonormal columns. Let U be a $p \times (p-d)$ matrix that forms an orthogonal complement basis of V. Then*

$$\mathcal{P}_{\mathcal{D}_\Pi}(A) = U \left[\sum_{i=1}^{p-d} \gamma_i^+(\theta) \eta_i \eta_i^T \right] U^T,$$

where $(\gamma_i, \eta_i)_{i=1}^{p-d}$ are eigenvalue-eigenvector pairs of $U^T A U$: $U^T A U = \sum_{i=1}^{p-d} \gamma_i \eta_i \eta_i^T$, and $\gamma_i^+(\theta) = \min(\max(\gamma_i - \theta, 0), 1)$, with θ chosen such that $\sum_{i=1}^{p-d} \gamma_i^+(\theta) = 1$.

The next theorem ensures the convergence of our algorithm to a global optimum of problem (6). The proofs of Lemma 3.1 and Theorem 3.2 are given in the Online Supplementary

File.

Theorem 3.2. *In Algorithm 1, $Z^{(r)} \to H^*$, $H^{(r)} \to H^*$ as $r \to \infty$, where H^* is a global optimum of problem* (6).

In the proof of Theorem 3.2 we will see that the auxiliary number τ used in Algorithm 1 plays a role that is similar to the step size commonly seen in iterative convex optimization solvers. The particular choice of τ does not affect the theoretical convergence of the ADMM algorithm. There are some general guidelines on the practical choice of τ and we refer to Boyd et al. (2011) for further details.

3.2 Choice of Tuning Parameters

The optimization problem (5) involves two tuning parameters: ρ_1 controls the roughness of eigenfunctions; and ρ_2 controls the localization of the eigenfunctions. We present a two-step approach where ρ_1 is chosen first and kept the same for all $1 \leq j \leq k$, and then ρ_2 is determined sequentially for each eigenfunction $\phi_j(t)$ and denoted by $\rho_{2,j}$. The two parameters can be chosen together by straightforward modification but computationally it would be a bit intensive.

The choice of the smoothing parameter ρ_1 has been discussed in Rice and Silverman (1991), and they recommended cross-validation or manual selection. In our simulation and data analysis, we have used V-fold cross-validation. First the data is divided into V folds, denoted by $\mathcal{P}_1, \mathcal{P}_2, \ldots, \mathcal{P}_V$. Let $H_j^{(-v)}(\rho_1, \rho_2)$ be the estimated H_j in (5) using data other than $\mathcal{P}_v$ with tuning parameters ρ_1 and ρ_2. Let S^v be the discrete covariance estimated from data $\mathcal{P}_v$. The smoothing parameter is chosen by maximizing the cross-validated inner product of $H_1^{(-v)}$ and S^v:

$$\widehat{\rho}_1 = \arg\max_{\rho \in \mathcal{A}_1} \sum_{v=1}^{V} \langle H_1^{(-v)}(\rho, 0), S^v \rangle, \tag{9}$$

where $\mathcal{A}_1$ is a candidate set of ρ_1 and empirically we found that a sequence between 0 and p times the largest eigenvalue of S works well.

In the following, we present two methods for the choice of ρ_2 given a pre-chosen smoothing parameter ρ_1^*. The first method is to choose $\rho_{2,j}$ by maximizing the cross-validated inner product of $H_j^{(-v)}$ and S^v:

$$\widehat{\rho}_{2,j} = \arg\max_{\rho \in \mathcal{A}_{2,j}} \sum_{v=1}^{V} \langle H_j^{(-v)}(\rho_1^*, \rho), S^v \rangle \,, \; j = 1, 2, ..., k \,, \tag{10}$$

where $\mathcal{A}_{2,j}$ is a candidate set for $\rho_{2,j}$, and we propose to use a sequence between 0 and the 95% quantile of absolute values of off-diagonal entries in S_j, with $S_j = (I - \widehat{\Pi}_{j-1})S(I - \widehat{\Pi}_{j-1})$.

The V-fold cross-validation approach is expected to give a $\widehat{\rho}_2$ that indicates the true localization level of the eigenfunctions. The criterion in our proposed cross-validation corresponds to maximizing $\langle \Sigma, \widehat{H}(\rho_2) \rangle$, where Σ is the discretized population covariance and is substituted by the test sample covariance in practice. When the true eigenvector v is localized, $\langle \Sigma, \widehat{H}(\rho_2) \rangle$ shall be maximized at approximately the value of ρ_2^* which corresponds to the ideal localization level of the eigenfunction, i.e., $\|\widehat{H}(\rho_2^*)\|_{1,1} \approx \|vv^T\|_{1,1}$. We shall expect $\langle \Sigma, \widehat{H}(\rho_2) \rangle$ as a function of ρ_2 to be (i) monotonically increasing on $[0, \rho_2^*]$ as the search area gradually expands to cover the true eigenvector, and (ii) monotonically decreasing on $[\rho_2^*, \infty)$ as the search area goes unnecessarily larger so that the estimation becomes more noisy. When the true eigenvector v is not localized, then we shall expect $\langle \Sigma, \widehat{H}(\rho_2) \rangle$ to be monotonically decreasing as ρ_2 increases. The numerical study confirms the good performance of the cross-validation method. See Section 5 for more details.

In some applications, we may not want to target the standard eigenfunction, but instead we may want to find orthogonal linear expansions that balance the interpretability (localization) and the capability of explaining the variance of the process. We therefore propose a second method of choosing ρ_2, which is based on the notion of fraction of variance explained (FVE). For a p-dimensional vector v of unit length and a sup-norm consistent estimator S of the covariance operator,

$$FVE(v) = v^T S v / totV(S), \tag{11}$$

where $totV(S)$ is the sum of positive eigenvalues of S. We note that for dense and equally spaced observations with measurement error, the $FVE(v)$ defined above is not directly ap-

plicable to a sample covariance S because the sample covariance is sup-norm consistent up to a shift σ^2 on the diagonal, where σ^2 is the error variance. To avoid serious bias by the nugget effect on the diagonal, one may use the eigenvalues from the smoothed covariance. Another practical way is to approximate $totV(S)$ by the sum of the first M leading eigenvalues of S, for a finite number M. For a reasonable error level, the nugget effect bias $M\sigma^2$ is small compared to the sum of the first M eigenvalues of S, which is of order p (Kneip et al., 2011), while the remaining true eigenvalues beyond M are usually very small because the smoothness of $X(t)$ ensures fast decay of eigenvalues. In numerical experiments where FVE is needed for determining the number of principal components to be included, we use $M = \min(20, p - 2)$.

To sequentially select the sparsity parameter ρ_2 for the jth eigenfunction. Suppose that we have estimated $\widehat{v}_i$ for $1 \leq i \leq j - 1$. Let $\widehat{v}_j(\rho_1^*, \rho)$ be the solution of (5) by using a fixed ρ_1^* and $\rho_2 = \rho$, with $\widehat{\Pi}_{j-1}$ being the projector of the subspace spanned by $(\widehat{v}_i : 1 \leq i \leq j - 1)$. We can define

$$rFVE(\rho) = \frac{FVE(\widehat{v}_j(\rho_1^*, \rho))}{FVE(\widehat{v}_j(\rho_1^*, 0))} = \frac{\widehat{v}_j^T(\rho_1^*, \rho)S\widehat{v}_j(\rho_1^*, \rho)}{\widehat{v}_j^T(\rho_1^*, 0)S\widehat{v}_j(\rho_1^*, 0)}, \tag{12}$$

and choose $\rho_{2,j}$ as

$$\max\{\rho \in \mathcal{A}_{2,j} : rFVE(\rho) \geq 1 - a\}, \tag{13}$$

where $a \in [0, 1)$ is the proportion of FVE that one chooses to sacrifice in return of localization. For any $a \in [0, 1)$, a ρ satisfying (13) always exists, because $rFVE(\rho) = 1$ for $\rho = 0$ and $rFVE(\rho) \in [0, 1)$ for $\rho > 0$. Equation (12) also suggests that $rFVE(\rho)$ can be calculated without computing $totV(S)$. Although the first localized basis function explains less variance than the standard eigenfunction, the lost proportion is likely to be picked up by the second component, and we are still able to explain a large proportion of the total variance with a small number of components. We illustrate this method with real data analyses in Section 6.

4. ASYMPTOTIC PROPERTIES

In this section we establish the ℓ_2 consistency of the proposed estimator in an asymptotic setting where both the sample size n and the number of grid points p increase. We will first provide sufficient conditions on the tuning parameters ρ_1 and ρ_2 such that the LFPCA estimate is consistent. Our second result provides further insights on how the localization penalty ρ_2 affects the rate of convergence. We make the following assumptions.

A1. The input matrix S in (5) satisfies sup-norm consistency up to a constant shift on the diagonal: for some constant $\alpha \geq 0$ and a sequence $e_n = o(1)$,

$$\max_{1 \leq l, l' \leq p} |S(l, l') - \Gamma(t_l, t_{l'}) - \alpha \mathbf{1}(l = l')| = O_P(e_n), \quad \text{as} \ (n, p) \to \infty.$$

Remark: Assumption (A1) puts a mild condition on the input matrix S that can be satisfied by many standard estimators. Consider functional data with dense and equally spaced observations. If S is the sample covariance estimator from the raw data, standard large deviation bounds such as Bernstein's inequality (Van Der Vaart and Wellner (1996), Chapter II) imply that Assumption (A1) holds with $e_n = \sqrt{\log p/n}$ if $\log p/n \to 0$ and the random curve $X(t)$ as well as the observation error in model (2) has sub-Gaussian tails; see also Kneip et al. (2011). In this case $\alpha = \sigma^2$, the noise variance. If a smoothed covariance estimator S is used, the sup-norm rate can be $\sqrt{\log n/n}$ (Li et al., 2010). The convergence in other norms such as Frobenius norm can be found in (Hall and Hosseini-Nasab, 2006; Hall et al., 2006; Bunea and Xiao, 2015). In general, the consistency result does not really depend on the observational design as long as we can get an estimate of the covariance operator whose sup-norm error vanishes as n and p increase. More discussions about cases where a sample covariance is not feasible can be found in Section 7.

A2. There is a positive integer k such that the eigenvalues of Γ satisfies $\lambda_1 >, \ldots, > \lambda_k > \lambda_{k+1} \geq, \ldots, \geq 0$, with positive eigen-gap $\delta := \min_{1 \leq j \leq k}(\lambda_j - \lambda_{j+1}) > 0$.

A3. Γ is Lipschitz continuous:

$$|\Gamma(s, t) - \Gamma(s', t')| \leq L \max(|s - s'|, |t - t'|), \quad \forall \ s, s', t, t'.$$

A4. The k leading eigenfunctions of Γ have Lipschitz first derivatives:

$$|\phi_j'(t) - \phi_j'(s)| \le L|t - s|, \quad \forall\, 1 \le j \le k\,.$$

Theorem 4.1 (ℓ_2 consistency). *Under assumptions (A1-A4), if $\rho_1/p^5 \to 0$ and $\rho_2 \to 0$ as $(n,p) \to \infty$, then for k as defined in A2, we have*

$$\sup_{1 \le j \le k} \|\widehat{\phi}_j(t) - \phi_j(t)\|_2 \xrightarrow{P} 0\,.$$

The proof of Theorem 4.1 is given in the Online Supplementary File. Here we outline the proof, highlighting some key technical challenges.

Let u_j be the unit vector obtained by discretizing and re-normalizing the eigenfunction $\phi_j(t)$. We will prove Theorem 4.1 by proving

$$\sup_{1 \le j \le k} \|\widehat{v}_j - u_j\|_2 \xrightarrow{P} 0.$$

Let Σ be the discretized covariance operator, with a possible constant shift a on the diagonal. Roughly speaking, $\widehat{v}_j$ and u_j approximate the jth eigenvector of S and Σ, respectively. We hope to establish the following inequality based on the standard Davis-Kahan $\sin\Theta$ theorem (Bhatia (1997), Theorem VII.3.1),

$$\|\widehat{v}_j\widehat{v}_j^T - u_ju_j^T\|_F \le \frac{c}{\delta p}\|S - \Sigma\|_F \le \frac{c}{\delta}\|S - \Sigma\|_{\infty,\infty}\,, \tag{14}$$

provided that S and Σ have eigengap of order δp (Lemma C.2), where $\|A\|_F = \langle A, A \rangle^{1/2}$ is the Frobenius norm and $\|A\|_{\infty,\infty}$ is the maximum absolute value of all entries in A.

However, to rigorously obtain an approximated version of (14) is non-trivial. First, u_j is not an eigenvector of Σ because of the discretization error. The discretization error will be explicitly tracked in all subsequent analysis when comparing u_j with $\widehat{v}_j$, for example, in the characterization of population PCA problem (Lemma C.4). Second, $\widehat{v}_j$ is obtained by solving a penalized eigenvector problem over the deflated Fantope, and hence is not directly comparable to its ideal theoretical counterpart u_j, which may not be in the feasible set of problem (5). To overcome this difficulty, we will consider a modified version of u_j that

13

is feasible for (5) but still possesses similar smoothness as well as proximity to the true eigenvector of Σ. Furthermore, the sequential estimation procedure (5) involves deflation based on estimated projection matrix $\widehat{\Pi}_{j-1}$, which carries over estimation error from previous steps. We will use an induction argument to control the sequential error accumulation.

Due to the sequential error accumulation, the convergence rate involves ρ_1 and ρ_2 in a complex way. To provide insights for our localized estimation procedure, the following theorem shows how the rate of convergence depends on ρ_2 when the smoothing parameter $\rho_1 = 0$. The proof is included in the Online Supplementary File.

Theorem 4.2 (Rate of convergence). *Under assumptions (A1-A4), if $\rho_1 = 0$ and $\rho_2 \to 0$ as $(n, p) \to \infty$, then for k as defined in A2, we have*

$$\sup_{1 \leq j \leq k} \|\widehat{\phi}_j(t) - \phi_j(t)\|_2 = O_P(e_n + \rho_2 + p^{-1}).$$

The three parts in the rate of convergence correspond to covariance estimation error, bias caused by localization penalty, and discretization error, respectively. According to the discussion after Assumption A1, the sup-norm covariance estimation error e_n can be made as small as $\sqrt{\log p / n}$ or $\sqrt{\log n / n}$ depending on the estimating method used and observation scheme. Thus if p grows at the same or higher order than $\sqrt{n}$, and $\rho_2 = O(e_n)$, our LFPCA estimate achieves an error rate of e_n within a logarithm factor from the standard FPCA error rate.

5. NUMERICAL STUDY

To illustrate our methods for localized functional principal component analysis, we conduct simulations under two scenarios, *Simulation I: localized case* and *Simulations II: non-localized case*. For *Simulation I*, data $\{Y_{il}, i = 1, \ldots, n, l = 1, \ldots, p\}$ are generated according to model (2), where t_l are equally spaced observational points on $[0, 1]$. We set $\mu(t) = 0$, $\xi_{ij} \sim N(0, \lambda_j)$, independent, with λ_j taken from $(4^2, \ 3^2, \ 2.5^2, \ 1.25^2, \ 1, \ 0.75^2, \ 0.5^2, \ 0.25^2)$ and $\lambda_j = 0$ for $j > 8$, and the measurement errors $\epsilon_{il} \overset{iid}{\sim} N(0, \sigma^2)$. We generate the eigenfunctions as follows. Let $\widetilde{\phi}_1(t) = B_3(t)$, $\widetilde{\phi}_2(t) = B_6(t)$ and $\widetilde{\phi}_3(t) = B_9(t)$, where $B_b(t)$ is the bth cubic B-spline basis
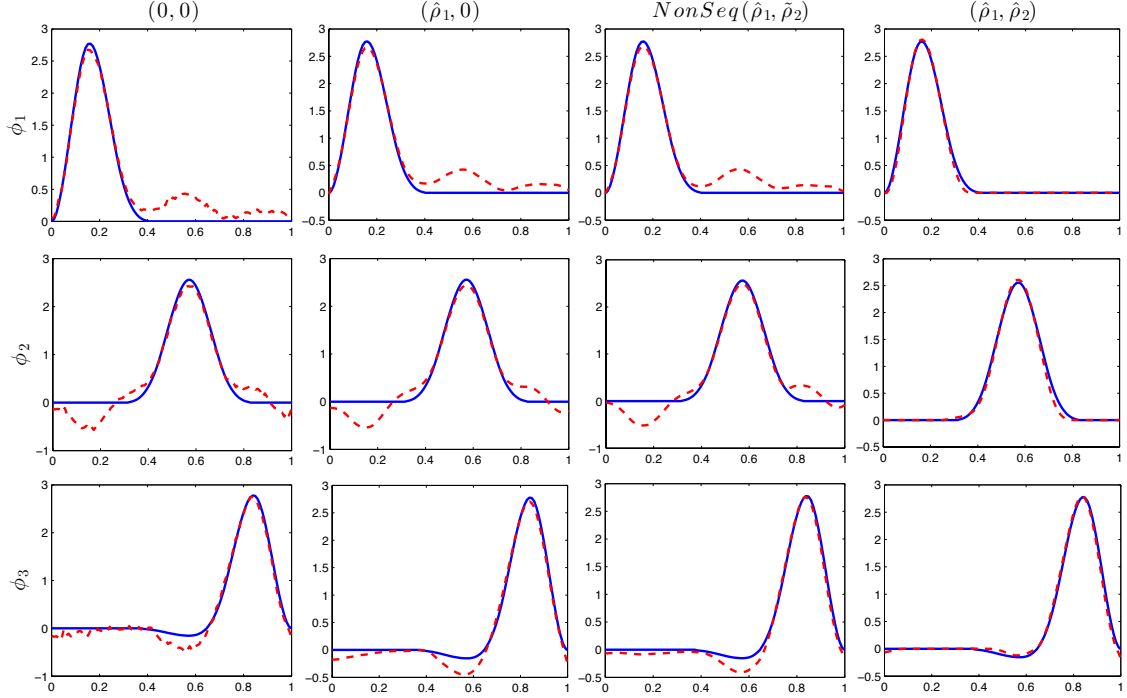
14

Figure 1: True (blue-solid) and estimated (red-dashed) eigenfunctions $\phi_j(t)$, $j = 1, 2, 3$, in one run of *Simulation I*, with $n = 100$, $p = 100$ and $\sigma = 1$, by four different methods as described in Section 5. Tuning parameters are chosen by 5-fold cross-validation.

on $[0, 1]$, with 8 equally spaced interior knots. For $j > 3$, $\widetilde{\phi}_j(t) = \sqrt{2}\cos((j + 1)\pi t)$ for odd values of j and $\widetilde{\phi}_j(t) = \sqrt{2}\sin(j\pi t)$ for even values of j, $0 \leq t \leq 1$. Then $\phi_j(t)$, $1 \leq j \leq 8$, are obtained by applying Gram-Schmidt orthonormalization on the set of $\widetilde{\phi}_j(t)$, $1 \leq j \leq 8$. For *Simulation II*, we use $\phi_j(t) = \sqrt{2}\cos((j + 1)\pi t)$ for odd values of j and $\phi_j(t) = \sqrt{2}\sin(j\pi t)$ for even values of j, $0 \leq t \leq 1$, for $1 \leq j \leq 8$. The rest is the same as *simulation I*.

We investigate the performance of the proposed LFPCA under varying combinations of sample size n, number of observations per curve p, and noise level σ^2. More importantly, we compare the estimates given by four different methods (i) $(\rho_1, \rho_2) = (0, 0)$ corresponds to the ordinary PCA estimation directly obtained from the sample covariance; (ii) $(\rho_1, \rho_2) = (\widehat{\rho}_1, 0)$ corresponds to the smoothed eigenfunction estimation without localization, where $\widehat{\rho}_1$ was chosen by 5-fold cross-validation as discussed in Section 3.2. Empirically, we found these estimated eigenfunctions almost identical to those estimated from a smoothed covariance
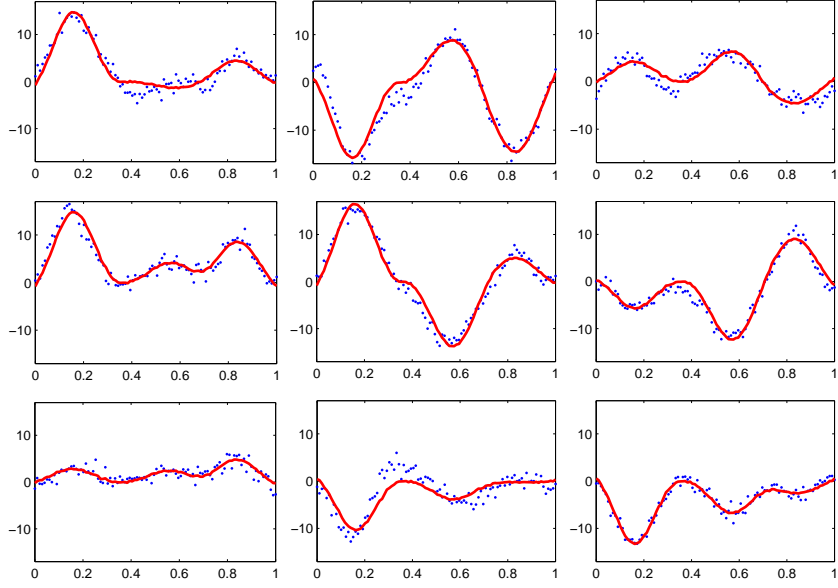
Figure 2: Noisy observations and recovered functions $\widehat{X}_i(t)$ (red-solid) for nine randomly selected subjects, as obtained in one run of *simulation I* with $n = 100$, $p = 100$ and $\sigma = 1$, using $(\widehat{\rho}_1, \widehat{\rho}_2)$ chosen by 5-fold cross-validation.

function or pre-smoothed individual curves; (iii) *NonSeq* corresponds to the subspace method developed in Vu et al. (2013). For comparison purpose, we incorporate the roughening matrix, i.e., use $S - \rho_1 D$ as input matrix in (7) with $\widehat{\rho}_1$ chosen by 5-fold cross-validation, the same as used in (ii) and (iv). The sparse tuning parameter is chosen by 5-fold cross validation $\widetilde{\rho}_2 = \arg\max_{\rho \in \mathcal{A}_2} \sum_{v=1}^{V} \langle H^{(-v)}(\rho_1, \rho), S^v \rangle$. We also note that their proposed method only outputs k basis vectors of the k-dimensional subspace, and one needs to rotate that basis to obtain eigenvectors. (iv) $(\rho_1, \rho_2) = (\widehat{\rho}_1, \widehat{\rho}_2)$ corresponds to the proposed LFPCA with tuning parameters selected by 5-fold cross-validation as detailed in (9) and (10). Each setting is repeated 200 times to assess the average performance. The number of included components k is chosen to account for at least 85% of the total variance, i.e. $\sum_{j=1}^{k} FVE(\widehat{v}_j) \geq 85\%$, where FVE is defined in (11). The selected number of k is quite robust among all simulation settings, and the average number over 200 simulations is 3.01.

The estimated eigenfunctions $\phi_j(t)$, $j = 1, 2, 3$, from a typical run of *Simulation I* with $p = 100$, $n = 100$, and $\sigma = 1$ are visualized in Figure 1. The four columns correspond to
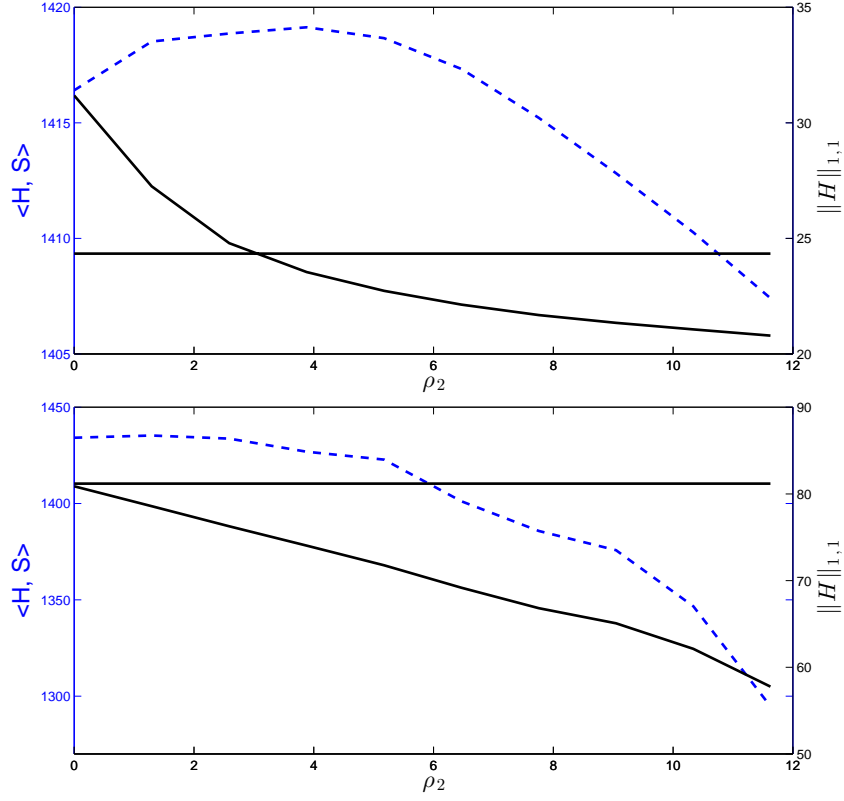
Figure 3: Performance of the 5-fold cross-validation to choose ρ_2 for ϕ_1. The top panel is for *Simulation I: localized case*: The peak of the cross-validated inner product $\langle H, S \rangle$ (dashed curve, left y label) corresponds to $\widehat{\rho}_2 = 3.8$. The estimated ϕ_1 is more localized as ρ_2 increases, and an ideal ρ_2 is where the $\|\cdot\|_{1,1}$ of estimated H (solid curve, right y label) meets that of true discretized projection matrix corresponding to ϕ_1 (indicated by the horizontal line). The bottom panel is for *Simulation II: non-localized case*: $\widehat{\rho}_2 = 1.2$.

results given by the four methods described above. One can clearly see the improvement by adding smoothing and localization penalties. The 5-fold cross-validation choice $(\widehat{\rho}_1, \widehat{\rho}_2)$ leads to almost perfect recovery of the true eigenfunctions. The fitted curves $\widehat{X}_i(t) = \widehat{\mu}(t) + \sum_{j=1}^{k} \widehat{\xi}_{ij} \widehat{\phi}_j(t)$ for nine randomly chosen subjects are shown in Figure 2, where $\widehat{\phi}_j(t)$ are obtained using $(\widehat{\rho}_1, \widehat{\rho}_2)$. It demonstrates accurate recovery of the true curves $X_i(t)$.

To better quantify the performance of estimating $\phi_j(t)$ we report the ℓ_2 distance $\|\phi_j(t) - \widehat{\phi}_j(t)\|_2$. The medians of the errors over 200 simulation runs are reported in Table 1. The results are quite similar for different levels of p and σ, so only results for $p = 100$ and $\sigma = 1$

		Simulation I: Localized			*Simulation II: Non-localized*		
		$n = 50$	$n = 100$	$n = 200$	$n = 50$	$n = 100$	$n = 200$
ϕ_1	$(0, 0)$	0.22 (0.09)	0.18 (0.06)	0.13 (0.04)	0.22 (0.08)	0.15 (0.05)	0.12 (0.04)
	$(\widehat{\rho}_1, 0)$	0.22 (0.09)	0.18 (0.07)	0.13 (0.04)	0.21 (0.08)	0.15 (0.06)	0.11 (0.04)
	NonSeq	0.22 (0.09)	0.18 (0.07)	0.13 (0.04)	0.21 (0.08)	0.15 (0.06)	0.11 (0.03)
	$(\widehat{\rho}_1, \widehat{\rho}_2)$	0.12 (0.05)	0.10 (0.03)	0.06 (0.02)	0.22 (0.08)	0.15 (0.05)	0.13 (0.03)
ϕ_2	$(0, 0)$	0.42 (0.16)	0.30 (0.13)	0.20 (0.07)	0.42 (0.16)	0.29 (0.12)	0.19 (0.07)
	$(\widehat{\rho}_1, 0)$	0.41 (0.16)	0.30 (0.13)	0.20 (0.07)	0.41 (0.15)	0.28 (0.11)	0.19 (0.06)
	NonSeq	0.40 (0.16)	0.30 (0.12)	0.20 (0.07)	0.41 (0.15)	0.28 (0.11)	0.19 (0.07)
	$(\widehat{\rho}_1, \widehat{\rho}_2)$	0.26 (0.17)	0.14 (0.07)	0.11 (0.05)	0.41 (0.17)	0.28 (0.11)	0.20 (0.07)
ϕ_3	$(0, 0)$	0.37 (0.12)	0.27 (0.11)	0.18 (0.07)	0.31 (0.10)	0.26 (0.09)	0.18 (0.06)
	$(\widehat{\rho}_1, 0)$	0.36 (0.12)	0.27 (0.11)	0.18 (0.08)	0.30 (0.10)	0.26 (0.09)	0.17 (0.06)
	NonSeq	0.33 (0.14)	0.27 (0.11)	0.18 (0.07)	0.30 (0.10)	0.26 (0.09)	0.18 (0.06)
	$(\widehat{\rho}_1, \widehat{\rho}_2)$	0.24 (0.15)	0.14 (0.08)	0.09 (0.04)	0.31 (0.11)	0.26 (0.09)	0.18 (0.06)

are reported with various sample size n. The errors are found to decline with increasing sample size n, as expected. For *Simulation I: localized case*, the proposed LFPCA with $\widehat{\rho}_2$ chosen by cross-validation significantly outperforms other methods. The $\widehat{\rho}_2$ chosen by cross-validation well approximates the true localization level of the eigenfunctions; see Figure 3 for an illustration and detailed discussions can be found in section 3.2. The *NonSeq*, a subspace method proposed by Vu et al. (2013), does not perform well since the union of the subdomains under consideration is the entire domain, not 'sparse' at all in their setting. The results demonstrate the advantage of our proposed sequential method. For *Simulation II: non-localized case*, 5-fold cross-validation method combined with the proposed LFPCA choose $\widehat{\rho}_{2j}$ to be very close to 0, and as expected, the ℓ_2 errors of the four methods are almost identical.
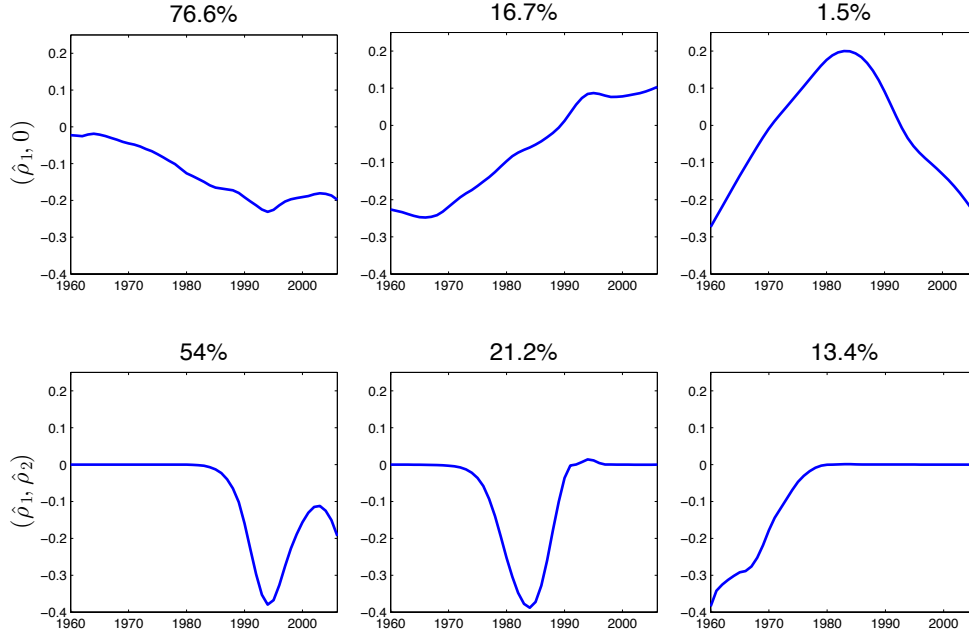
Figure 4: Top Row: Estimated eigenfunctions for the mortality data, $\widehat{\rho}_1$ is chosen by 5-fold cross validation; Bottom Row: Estimated orthogonal basis functions, $\widehat{\rho}_2$ is chosen to maintain rFVE at 70% in (12), and the number of components $k = 3$ is chosen to explain at least 85% of the total variance.

6.   APPLICATION TO COUNTRY MORTALITY DATA

The analysis of human mortality is important in assessing the future demographic prospects of societies, and quantifying differences between countries with regard to the overall public health measure. Functional data analysis approaches have been previously applied to study mortality data (Hyndman et al., 2007; Chiou and Müller, 2009; Chen and Müller, 2012). To study the variational modes of mortality rates across countries over years, we applied the proposed LFPCA method to period life tables for 27 countries, with rates of mortality at age 60 available for each of the calendar years from 1960 to 2006. The data were obtained from the Human Mortality Database (downloaded on March 1, 2011), maintained by University of California, Berkeley (USA), and Max Planck Institute for Demographic Research (Germany). The data is available at `www.mortality.org` or `www.humanmortality.de`, with detailed description in Wilmoth et al. (2007).

Let $X_i(t)$ denote the mortality rate in the ith country for subjects at age 60 during calendar year t, where $1960 \le t \le 2006$. We directly compute the sample covariance matrix S from the observed data and apply the proposed algorithm to solve problem (5). The $\widehat{\rho}_1$ chosen by 5-fold cross-validation (9) is always used to ensure a relatively smooth estimate of the eigenfunction, and the solution path along different levels of localization is investigated. The 5-fold cross-validation method gave $\widehat{\rho}_2 = 0$, indicating the eigenfunctions are not exactly localized. The estimated eigenfunctions without localization penalty are given in the top row of Figure 4. We then choose $\widehat{\rho}_2$ by the second method as defined in (13) with $a = 30\%$. The estimated localized basis functions, as visualized in the bottom row of Figure 4, reveal several interesting features. The first localized basis function $\phi_1(t)$, explaining 54% of the total variance, indicates that a big variation of the mortality functions $X_i(t)$ around their mean function happens around mid 1990s. The second basis function $\phi_2(t)$ with a mode around 1980s accounts for 21.2% of the total variation. The third basis function $\phi_3(t)$ characterizes variation of mortality around 1960s. Although the first localized basis function explains less variance than the first leading eigenfunction, only retaining 70% of the capability in return of localization, the lost proportion is picked up by the second component. The second component could have explained 30.3% of the variance without localization. Therefore, we only need three localized eigenfunctions to account for more than 85% of the total variance. Plots of estimated eigenvalues using other values of a are provided in the Online Supplementary File.

7. DISCUSSION

In this paper, we propose a localized functional principal component analysis through a De-flated Fantope Localization method, where sequentially obtained eigenfunctions have guaranteed orthogonality and are allowed to be supported on different localized subdomains. As mentioned in Section 2, the deflated Fantope $\mathcal{D}_\Pi$ can easily be generalized to a d-dimensional version $\mathcal{D}_\Pi^d$. In some applications, one might be interested to estimate mutually orthogonal principal subspaces with dimensions $d_1, \ldots, d_k$, and each principal subspace is only supported

on a subdomain $\mathcal{T}_j \subset \mathcal{T}$.

Throughout the paper, we mainly focus on a dense and regular functional data design where p equally spaced observations are recorded on each curve. Most commonly seen functional data have this design and a sample covariance can be easily computed from the discrete and possibly noisy observations. Our proposed formula (5) takes the sample covariance S and outputs smooth and localized estimates of eigenfunctions. In fact the proposed method puts rather minimal requirement on the input matrix S and does not rely on the design of observations. Consider the discretized version of Γ by evaluating on $p \times p$ equally spaced grid points. The estimation error of the eigenfunctions directly ties to the maximum entry-wise error of the input matrix S. Here we briefly discuss several scenarios where a sample covariance is not feasible. (i) For dense but irregularly observed functional data, one can simply smooth each curve (Ramsay and Silverman, 2005) or interpolate between points to get p equally spaced observations, and then a $p \times p$ covariance estimate S can be computed. This is what we have done for the Berkeley growth data given in the Online Supplementary File. (ii) For sparse functional data where the observations are recorded at random and sparse time points, individual smoothing or interpolation is impossible. But a uniformly consistent covariance estimation is possible by, for example, two-dimensional smoothing methods (Yao et al., 2005; Li et al., 2010). Our proposed LFPCA can then be applied by taking S as the discretized version of the smooth covariance estimator. (iii) For ultra dense and noisy data, the independent measurement errors accumulate if one uses sample covariance computed from the raw measurements. Moreover, using a large $p \times p$ matrix is not computationally efficient. We recommend performing pre-smoothing or pre-binning on individual curves and choosing a grid with a moderate size p.

We proposed two methods of choosing the localization parameter ρ_2. When the cross-validation method chooses $\rho_2 = 0$, it roughly means that the true eigenfunctions are not localized. Then for a fixed number of a, we find a set of orthogonal basis functions that retain the ability to explain a fair amount of variance (at least a $(1-a)$ proportion) and are localized. In this case, the outcome would depend on the choice of a and they should not

be interpreted as the true eigenfunctions. Rather, these localized basis functions and the corresponding projection scores have ready interpretation with domain knowledge.

REFERENCES

Bhatia, R. (1997), *Matrix analysis*, vol. 169, Springer.

Boyd, S., Parikh, N., Chu, E., Peleato, B., and Eckstein, J. (2011), "Distributed optimization and statistical learning via the alternating direction method of multipliers," *Foundations and Trends® in Machine Learning*, 3, 1–122.

Bunea, F. and Xiao, L. (2015), "On the sample covariance matrix estimator of reduced effective rank population matrices, with applications to fPCA," *Bernoulli*, to appear.

Cardot, H. (2000), "Nonparametric estimation of smoothed principal components analysis of sampled noisy functions," *Journal of Nonparametric Statistics*, 12, 503–538.

Castro, P. E., Lawton, W. H., and Sylvestre, E. A. (1986), "Principal Modes of Variation for Processes with Continuous Sample Curves," *Technometrics*, 28, 329–337.

Chen, K. and Müller, H.-G. (2012), "Modeling repeated functional observations," *Journal of the American Statistical Association*, 107, 1599–1609.

Chiou, J.-M. and Müller, H.-G. (2009), "Modeling hazard rates as functional data for the analysis of cohort lifetables and mortality forecasting," *Journal of the American Statistical Association*, 104, 572–585.

d'Aspremont, A., El Ghaoui, L., Jordan, M., and Lanckriet, G. (2007), "A Direct Formulation of Sparse PCA using Semidefinite Programming," *SIAM Review*, 49.

Dattorro, J. (2005), *Convex Optimization & Euclidean Distance Geometry*, Meboo Publishing USA, v2012.01.28.

Gasser, T., Müller, H.-G., Köhler, W., Prader, A., Largo, R., and Molinari, L. (1985), "An analysis of the mid-growth and adolescent spurts of height based on acceleration," *Annals of Human Biology*, 12, 129–148.

Hall, P. and Hosseini-Nasab, M. (2006), "On properties of functional principal components analysis," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 68,

109–126.

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

Huang, J. Z., Shen, H., Buja, A., et al. (2008), "Functional principal components analysis via penalized rank one approximation," *Electronic Journal of Statistics*, 2, 678–695.

Hyndman, R. J., Ullah, S., et al. (2007), "Robust forecasting of mortality and fertility rates: a functional data approach," *Computational Statistics & Data Analysis*, 51, 4942–4956.

James, G. M., Wang, J., and Zhu, J. (2009), "Functional linear regression that's interpretable," *The Annals of Statistics*, 2083–2108.

Kneip, A., Sarda, P., et al. (2011), "Factor models and variable selection in high-dimensional regression analysis," *The Annals of Statistics*, 39, 2410–2447.

Lei, J. and Vu, V. Q. (2015), "Sparsistency and Agnostic Inference in Sparse PCA," *The Annals of Statistics*, to appear.

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

Lin, Z. (2013), "Some perspectives of smooth and locally sparse estimators," *Master thesis, Simon Fraser University, Canada*.

Mackey, L. (2008), "Deflation Methods for Sparse PCA," in *NIPS*, vol. 21, pp. 1017–1024.

Mühl, A., Herkner, K., and Swoboda, W. (1991), "[The mid-growth spurt–a pre-puberty growth spurt. Review of its significance and biological correlations]," *Padiatrie und Padologie*, 27, 119–123.

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

Rice, J. A. and Silverman, B. W. (1991), "Estimating the mean and covariance structure nonparametrically when the data are curves," *Journal of the Royal Statistical Society. Series B (Methodological)*, 233–243.

Sheehy, A., Gasser, T., Molinari, L., and Largo, R. (1999), "An analysis of variance of

the pubertal and midgrowth spurts for length and width," *Annals of human biology*, 26, 309–331.

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

Tuddenham, R. and Snyder, M. (1954), "Physical growth of California boys and girls from birth to age 18," *Calif. Publ. Child Deve.*, 1, 183–364.

Van Der Vaart, A. W. and Wellner, J. A. (1996), *Weak Convergence*, Springer.

Vu, V. Q., Cho, J., Lei, J., and Rohe, K. (2013), "Fantope Projection and Selection: A near-optimal convex relaxation of sparse PCA," in *Advances in Neural Information Processing Systems*, pp. 2670–2678.

Vu, V. Q. and Lei, J. (2013), "Minimax sparse principal subspace estimation in high dimensions," *The Annals of Statistics*, 41, 2905–2947.

White, P. A. (1958), "The computation of eigenvalues and eigenvectors of a matrix," *Journal of the Society for Industrial & Applied Mathematics*, 6, 393–437.

Wilmoth, J. R., Andreev, K., Jdanov, D., Glei, D. A., Boe, C., Bubenheim, M., Philipov, D., Shkolnikov, V., and Vachon, P. (2007), "Methods protocol for the human mortality database," *University of California, Berkeley, and Max Planck Institute for Demographic Research, Rostock. URL: http://mortality. org [version 31/05/2007]*.

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

Zhao, Y., Ogden, R. T., and Reiss, P. T. (2012), "Wavelet-based LASSO in functional linear regression," *Journal of Computational and Graphical Statistics*, 21, 600–617.

Zhou, J., Wang, N.-Y., and Wang, N. (2013), "Functional linear model with zero-value coefficient function at sub-regions," *Statistica Sinica*, 23, 25.