Learning Shared and Source-specific Subspaces across Multiple Data Sources for Functional Data

Chi Zhang, Peijun Sang, Yingli Qin
Department of Statistics and Actuarial Science, University of Waterloo

Abstract

In the era of big data, integrating multi-source functional data to extract a subspace that captures the shared subspace across sources has attracted considerable attention. In practice, data collection procedures often follow source-specific protocols. Directly averaging sample covariance operators across sources implicitly assumes homogeneity, which may bias the recovery of both shared and source-specific variation patterns. To address this issue, we propose a projection-based data integration method that explicitly separates the shared and source-specific subspaces. The method first estimates source-specific projection operators via smoothing to accommodate the nonparametric nature of functional data. The shared subspace is then isolated by examining the eigenvalues of the averaged projection operator across all sources. If a source-specific subspace is of interest, we re-project the associated source-specific covariance estimator onto the subspace orthogonal to the estimated shared subspace, and estimate the source-specific subspace from the resulting projection. We further establish the asymptotic properties of both the shared and source-specific subspace estimators. Extensive simulation studies demonstrate the effectiveness of the proposed method across a wide range of settings. Finally, we illustrate its practical utility with an example of air pollutant data.

Keywords— Functional data analysis, Functional principal component analysis, Multi-source analysis, Shared subspaces.

1 Introduction

Functional data analysis (FDA) has gained increasing prominence in modern statistics, driven by advances in data collection technologies. It provides a nonparametric framework for analyzing discrete observations taken from realizations of a continuous random function, often defined over time or space. The inherently infinite-dimensional nature of functional data necessitates the use of dimension reduction techniques, which transform infinite-dimensional random functions into finite-dimensional random vectors and thereby facilitate subsequent analysis. Among these techniques, functional principal component analysis (FPCA) is one of the most widely used; see, for example, Chapter 8 of Ramsay and Silverman (2005) or Chapter 11 of Kokoszka and Reimherr (2017). Analogous to principal component analysis, FPCA aims to extract a parsimonious subspace that explains the most relevant variation around a mean function in a data-adaptive manner.

The theoretical foundations of FPCA have been extensively developed over the past two decades (Yao et al., 2005; Hall and Hosseini-Nasab, 2006; Zhou et al., 2024). FPCA is also commonly adopted as a regularization tool when inverting the covariance operator is required. For example, the inverse of the covariance operator is unbounded without regularization in functional linear regression (Hall and Horowitz, 2007; Zhou et al., 2023) and functional generalized linear models (Dou et al., 2012). While FPCA methods are well established both in theory and applications, comparatively less attention has been devoted to settings involving multiple functional data sources, where one seeks to aggregate information and uncover shared structure. Yet, such scenarios are increasingly common in modern applications, where data collected at different sites often present heterogeneity. For instance, in the air pollutants study discussed in Section 5, measurements are obtained from five cities with distinct geographical locations and climate conditions.

When functional data are collected from multiple sources, heterogeneity across sources poses additional challenges for analysis. A naive extension of single-source FPCA is to compute source-specific sample covariance operators and then perform FPCA on their average. However, this averaging procedure implicitly assumes homogeneity across sources, and dominant patterns present only in particular sources may introduce severe bias in identifying the true shared structure. To illustrate, we construct a toy example in Example 2.2 with two sources. The covariance operator of the second source is generated by applying a simple rotation to the basis functions of the first source, while retaining the same eigenvalues. Even though the two sources have identical sample sizes and eigenvalues, the eigenfunction estimated from the averaged covariance operator deviates substantially from the truth, as shown in Figure 1.

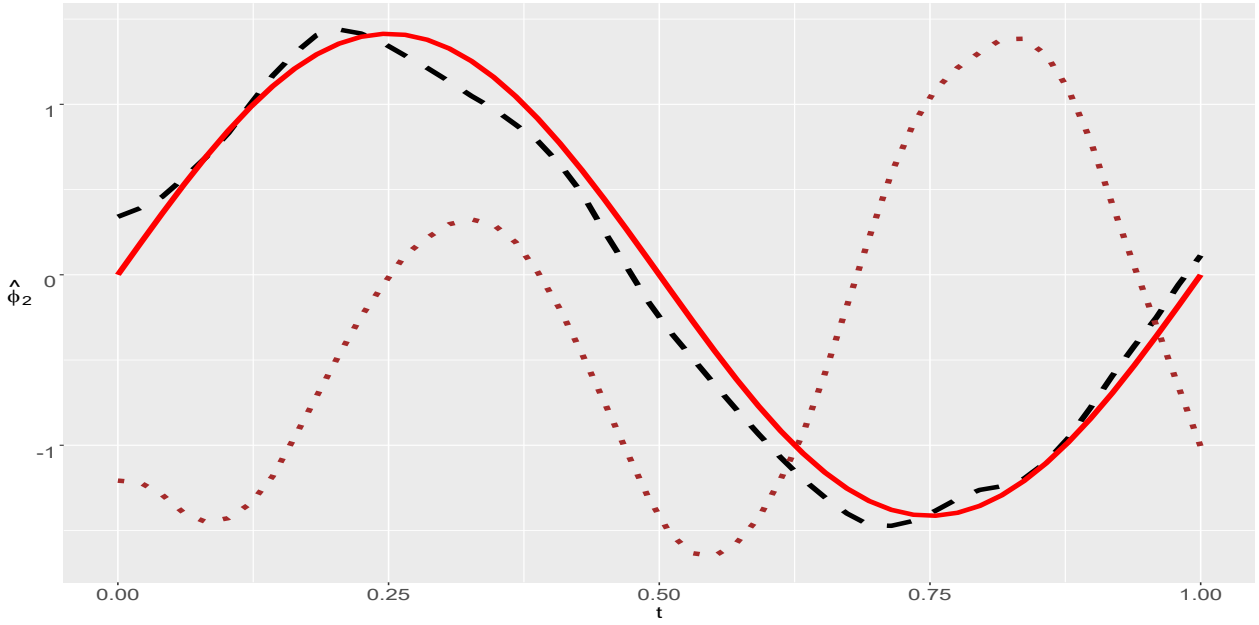

Figure 1: Data are generated from model (4) and Example 2.2. The solid line represents the true shared eigenfunction across two sites. The dashed line shows the estimated shared eigenfunction obtained using the proposed method in Section 2.3. The dotted line corresponds to the eigenfunction estimated from the averaged covariance operator.

Beyond naive averaging, much of the existing literature on multi-source functional data has focused on mean estimation under the framework of transfer learning. For example, Cai et al. (2024) studied phase transition phenomena in mean function estimation based on discretely sampled

functional data and established the corresponding minimax rates. Another line of work concerns common FPCA (CFPCA) and partially common FPCA (PCFPCA). CFPCA assumes that the leading eigenfunctions for all sources are the same, with differences arising only in the eigenvalues (Benko et al., 2009; Boente et al., 2010), whereas PCFPCA allows some eigenfunctions to be shared and others to remain source-specific (Jiao et al., 2024). For instance, Jiao et al. (2024) considered a forest-structured hierarchy of communities, where sources in each layer were partitioned into communities such that the data sources within the same community partially shared a subset of basis functions.

However, existing CFPCA and PCFPCA methods are designed for fully or densely observed functional data, which may be too restrictive in many real-world applications. Moreover, although the parameter of interest is often a low-dimensional shared subspace, most existing approaches rely on linear combinations of covariance functions estimated from each source. Such strategies remain prone to bias when the eigenvalues associated with source-specific directions dominate those of the shared subspace across sources, regardless of how the averaging weights are chosen. A more natural perspective is to focus directly on the subspace itself: if the shared subspace is the primary object of interest, one can safely discard the eigenvalues during the data integration process and retain only the eigenfunctions. This perspective motivates the projection-based approach we develop in this paper.

To our knowledge, no existing methods directly pursue this subspace-focused strategy in the functional data literature to extract the shared subspace. The idea is inspired by Li et al. (2024), which aims to enhance the estimation efficiency of the shared structure under the framework of transfer learning for multivariate data. They identified the shared space by solving an optimization problem on the Grassmann manifold, where the manifold is often defined for finite-dimensional Euclidean space.

We generalize the method to functional data, where the underlying space is infinite-dimensional. Specifically, we first estimate the sample covariance function for each source using a local linear smoother, thereby accommodating the nonparametric nature of functional data. Aggregating observations within each source allows us to borrow strength across subjects, enabling estimation even for sparse functional data. Discretizing the smoothed covariance estimates yields estimated eigenpairs and corresponding sample projection operators for each source. We then construct a weighted average projection estimator, $\widehat{\boldsymbol{P}}_w$, from the source-wise sample projection operators. By utilizing the spectral properties of $\widehat{\boldsymbol{P}}_w$, we identify the shared subspace, which is then used to estimate the source-specific subspace by re-projecting the estimated covariance operator of the specific source onto the orthogonal complement of the shared subspace.

Furthermore, we establish theoretical convergence guarantees for both the shared and source-specific estimators. The convergence rate of the estimator of the shared projection operator highlights the gain in estimation efficiency achieved by effectively leveraging information from multiple data sources. These theoretical insights align with the numerical results from simulation studies under various settings, where the proposed method consistently outperforms single-source estimators.

The remainder of this article is organized as follows. In Section 2.1, we introduce the data generation mechanism and standard FPCA for one source. Then, we rigorously define the shared structure for functional data in Section 2.2. The proposed method is detailed in Section 2.3. We develop asymptotic results for the proposed estimator in Section 3. Detailed proofs are relegated to the online supplement. We showcase the finite-sample performance of the proposed method under various simulation settings in Section 4. In Section 5 we apply the proposed method to a real world example. Finally, we conclude the paper in Section 6.

2 Estimation via Data Integration

2.1 Standard FPCA

Before presenting the proposed method for multiple data sources, we begin by reviewing the standard FPCA for a single source and introducing the notation. Let $X(t)$ be a continuous and square-integrable stochastic process defined on a compact interval $\mathcal{T} = [0, 1]$ with $\mathbb{E}[X(t)] = 0$ for $t \in \mathcal{T}$. Denote $\mathcal{H} = \mathcal{L}^2(\mathcal{T})$, the Hilbert space of square-integrable functions on $\mathcal{T}$, equipped with inner product $\langle f, g\rangle_2 = \int_{t\in\mathcal{T}} f(t)g(t)\,\mathrm{d}t$. Moreover, $\|f\|_2^2 = \langle f, f\rangle_2$. For a fixed m, FPCA seeks an m-dimensional subspace $\mathcal{H}_0 \subset \mathcal{H}$ such that the projection of X onto $\mathcal{H}_0$ provides the optimal mean-squared approximation to X. More precisely, FPCA identifies m orthonormal functions $e_1, e_2, \ldots, e_m \in \mathcal{H}$ that minimize the expected mean-squared distance between X and its m-dimensional projection, which is formalized as follows:

$$\widetilde{\boldsymbol{P}} = \underset{\boldsymbol{P}\in\mathcal{P}_m}{\arg\min}\ \mathbb{E}\|X - \boldsymbol{P}X\|_2^2. \tag{1}$$

Here, $\mathcal{P}_m = \{\boldsymbol{P} \in \mathcal{B}(\mathcal{H}) \mid \boldsymbol{P}^2 = \boldsymbol{P}, \mathrm{rank}(\boldsymbol{P}) = m, \boldsymbol{P} \text{ is self-adjoint}\}$, and $\mathcal{B}(\mathcal{H})$ denotes the set of bounded linear operators on $\mathcal{H}$. In addition, any $\boldsymbol{P} \in \mathcal{P}_m$ can be written as $\boldsymbol{P} = \sum_{\nu=1}^m e_\nu \otimes e_\nu$, where $f \otimes g(\cdot) = \langle f, \cdot\rangle\, g$ is a rank-one operator on $\mathcal{H}$ for $f, g \in \mathcal{H}$.

For two operators $\boldsymbol{A}, \boldsymbol{B} \in \mathcal{B}(\mathcal{H})$, define $\langle \boldsymbol{A}, \boldsymbol{B}\rangle_{\mathrm{HS}} = \sum_\nu \langle \boldsymbol{A}(e_\nu), \boldsymbol{B}(e_\nu)\rangle_2$, where $\{e_\nu : \nu = 1, 2, \ldots\}$ is any orthonormal basis of $\mathcal{H}$. This definition does not depend on the choice of basis and induces the Hilbert–Schmidt norm $\|\boldsymbol{A}\|_{\mathrm{HS}}^2 = \langle \boldsymbol{A}, \boldsymbol{A}\rangle_{\mathrm{HS}}$. With this notation, the optimization problem in (1) is equivalent to

$$\widetilde{\boldsymbol{P}} = \underset{\boldsymbol{P}\in\mathcal{P}_m}{\arg\min}\ \mathbb{E}\|X - \boldsymbol{P}X\|_2^2 - \mathbb{E}\|X\|_2^2 = \underset{\boldsymbol{P}\in\mathcal{P}_m}{\arg\max}\ \langle \boldsymbol{G}, \boldsymbol{P}\rangle_{\mathrm{HS}}\,, \tag{2}$$

where $\boldsymbol{G}$ is the covariance operator induced by the covariance function of X. Specifically, for any $f \in \mathcal{H}$,

$$(\boldsymbol{G}f)(t) = \int_{s\in\mathcal{T}} G(s, t)f(s)\,\mathrm{d}s$$

with $G(s, t) = \mathbb{E}[X(s)X(t)]$ for $s, t \in \mathcal{T}$. The continuity of X together with the definition of G implies that G is continuous, symmetric, and positive semi-definite. By Mercer's theorem,

$$G(s, t) = \sum_{\nu=1}^{\infty} \lambda_\nu \phi_\nu(s)\phi_\nu(t), \quad t, s \in \mathcal{T}, \tag{3}$$

where (λ_ν, ϕ_ν) is the νth eigenvalue-eigenfunction pair of $\boldsymbol{G}$ satisfying $\boldsymbol{G}\phi_\nu = \lambda_\nu\phi_\nu$, and $\{\phi_\nu\}_{\nu\geq 1}$ forms an orthonormal basis of $\mathcal{H}$. Combining equations (2) and (3), it follows immediately that $e_i = \phi_i$ for $i = 1, \ldots, m$. Therefore, it implies that we can use the leading m eigenfunctions of $\boldsymbol{G}$ to approximate X.

In practice, we can hardly get access to fully observed sample paths without any measurement error for X. Suppose that there are n i.i.d. sample paths of X, with N_i observations for subject i, observed at times T_{ij} with additive measurement error:

$$Y_{ij} = X_i(T_{ij}) + \varepsilon_{ij}, \quad i = 1, 2, \ldots, n,\ j = 1, 2, \ldots, N_i.$$

Here N_i are i.i.d. copies of a random variable N; T_{ij} are subject-specific observation times in $\mathcal{T}$; and ε_{ij} are independent with mean zero and variance σ_ε^2. Moreover, N, T, ε, and X are assumed to be independent. This model is widely adopted in modeling longitudinal observations under the

framework of functional data; see Yao et al. (2005), Li and Hsing (2010), Zhang and Wang (2016), Zhou et al. (2024) and references therein.

We first estimate the covariance function $G(t, s)$ by pooling observations across subjects, following the idea in Yao et al. (2005), where a local linear smoother is employed. Next, we estimate the eigenvalue–eigenfunction pairs $(\lambda_\nu, \phi_\nu), \nu = 1, 2, \ldots, m$, from the discretization of the smoothed covariance function; see Rice and Silverman (1991) and Ramsay and Silverman (2005) for more details. The choice of m can be determined by the fraction of variance explained (FVE), information-criterion-based methods (Yao et al., 2005; Li et al., 2013), or methods integrating information from both eigenvalues and eigenfunctions (Zhang et al., 2025).

2.2 Multi-source Model

Suppose that there are K sources, and let $\mathcal{K} = \{1, 2, \ldots, K\}$ denote the index set. For any $k \in \mathcal{K}$, we assume that the observations $Y_{ij}^{[k]}$ collected at $T_{ij}^{[k]}$ are generated from the following:

$$Y_{ij}^{[k]} = X_i^{[k]}\left(T_{ij}^{[k]}\right) + \varepsilon_{ij}^{[k]}, \quad i = 1, 2, \ldots, n_k, \; j = 1, 2, \ldots, N_i^{[k]}. \tag{4}$$

Analogous to the single-source model, we assume $N_i^{[k]}$ are i.i.d. copies of $N^{[k]}$, and $T_{ij}^{[k]}$'s are observation times in $\mathcal{T}$ for source k. Measurement errors $\varepsilon_{ij}^{[k]}$ are i.i.d with mean zero and variance $(\sigma_\varepsilon^{[k]})^2$. Moreover, $N^{[k]}$, $T^{[k]}$, $\varepsilon^{[k]}$ and $X^{[k]}$ are independent.

The Karhunen–Loève expansion of $X^{[k]}(t)$ for $k = 1, \ldots, K$, is

$$X^{[k]}(t) = \sum_{\nu=1}^{\infty} \xi_\nu^{[k]} \phi_\nu^{[k]}(t), \quad t \in \mathcal{T}, \tag{5}$$

where $\xi_\nu^{[k]} = \int_{\mathcal{T}} X^{[k]}(t)\phi_\nu^{[k]}(t)\,\mathrm{d}t$ are uncorrelated mean-zero random variables with variance $\lambda_\nu^{[k]}$. It is common to assume that the underlying process $X^{[k]}$ is well-approximated by the leading m_k eigenfunctions (Yao et al., 2005). Moreover, the number of eigenfunctions that can be consistently estimated depends on the sample size and the number of observations per subject (Zhou et al., 2024). Throughout this paper, we focus on the leading m_k functional principal components for source k. More precisely, we assume $\mathrm{rank}(\boldsymbol{P}_k) = m_k$ for $k \in \mathcal{K}$. This assumption is standard in the CFPCA/PCFPCA literature, where m_k is typically fixed and independent of n_k; see, for example, Benko et al. (2009) and Jiao et al. (2024). To better characterize the inherent infinite-dimensional nature of functional data, we instead allow m_k to diverge as n_k grows. Finally, let $\mathcal{H}_0^{[k]}$ denote the subspace spanned by the leading m_k eigenfunctions of source k.

As shown in (3), the covariance structure is determined by its eigenfunctions, which capture variability patterns, and the corresponding eigenvalues, which quantify their magnitudes. Ideally, if the sample sizes and the number of observations per subject for different sources are roughly the same, the number of selected eigenfunctions is about the same. Furthermore, if $\phi_\nu^{[k]} = \phi_\nu$ for all $k \in \mathcal{K}$ and $\nu \in \mathbb{N}_+$, the shared subspace can be obtained from the leading eigenfunctions of the averaged sample covariance operators. In practice, however, these assumptions are often too restrictive. Real-world multi-source datasets may exhibit heterogeneity in covariance structures or imbalanced sample sizes, which introduces additional challenges for data integration in the following ways:

- The number of eigenfunctions that can be consistently estimated depends on the sample size and the average of the number of observations per subject. Consequently, m_k may vary across sources when n_k or the average $N_i^{[k]}$ differ substantially.

- For some source k, the space spanned by its leading m_k eigenfunctions can be decomposed into a shared subspace and a source-specific subspace. If the latter is not properly isolated, it may introduce bias into the recovery of the shared subspace.
- In some sources, the eigenfunctions associated with the shared subspace may correspond to relatively small eigenvalues, whereas source-specific eigenfunctions may be associated with larger eigenvalues. This imbalance can substantially degrade the performance of naive averaging approaches.

To integrate information across sources and address these challenges, we assume that there is a shared subspace $\mathcal{H}_s$ with $\dim(\mathcal{H}_s) = m_s$, spanned by the same eigenfunctions across all sources. Note that, we allow m_s to diverge as the sample sizes increase. Thus, the shared structure can be formalized as follows: Let $\mathcal{I}_s^{[k]}$ contain the indices of the eigenfunctions that span $\mathcal{H}_s$ for source $k \in \mathcal{K}$. Then by (5), we rewrite $X^{[k]}(t)$ as

$$X^{[k]}(t) = \sum_{\nu \in \mathcal{I}_s^{[k]}} \xi_\nu^{[k]} \phi_\nu^{[k]}(t) + \sum_{\nu \notin \mathcal{I}_s^{[k]}} \xi_\nu^{[k]} \phi_\nu^{[k]}(t), \quad t \in \mathcal{T}.$$

Equivalently, for each $k \in \mathcal{K}$

$$\mathcal{H}_s^{[k]} \triangleq \left\{ f = \sum_{\nu=1}^{m_s} \alpha_\nu \phi_{\pi_k(\nu)} \mid \alpha_1, \alpha_2, \ldots, \alpha_{m_s} \in \mathbb{R} \right\} = \mathcal{H}_s,$$

where $\pi_k : \{1, 2, \ldots, m_s\} \to \mathcal{I}_s^{[k]}$ is a bijective function. This shared structure is similar to that defined in Jiao et al. (2024). Accordingly, define the projection operators

$$\boldsymbol{P}_s^{[k]} = \sum_{\nu \in \mathcal{I}_s^{[k]}} \phi_\nu^{[k]} \otimes \phi_\nu^{[k]}, \quad \boldsymbol{P}_p^{[k]} = \sum_{\nu \notin \mathcal{I}_s^{[k]},\ \nu \le m_k} \phi_\nu^{[k]} \otimes \phi_\nu^{[k]}, \tag{6}$$

where $\boldsymbol{P}_s^{[k]}$ projects onto the shared subspace and $\boldsymbol{P}_p^{[k]}$ onto the source-specific subspace.

The following toy example illustrates the shared structure.

Example. Consider the Fourier basis functions:

$$\phi_1(t) = 1, \quad \phi_\nu(t) = \begin{cases} \sqrt{2}\sin(\nu\pi t) \text{ if } \nu \ge 2 \text{ and } \nu \text{ is even}, \\ \sqrt{2}\cos\{(\nu-1)\pi t\} \text{ if } \nu \ge 2 \text{ and } \nu \text{ is odd}. \end{cases}$$

Consider two sources with covariance operators $\boldsymbol{G}^{[k]} = \sum_{\nu=1}^{\infty} \lambda_\nu^{[k]} \phi_\nu^{[k]} \otimes \phi_\nu^{[k]}$ for $k = 1, 2$. For source 1, set $\lambda_\nu^{[1]} = \nu^{-2}$ and $\phi_\nu^{[1]} = \phi_\nu$ for $\nu \in \mathbb{N}_+$. For source 2, we assume $\lambda_2^{[2]} = \lambda_3^{[1]}$, $\lambda_3^{[2]} = \lambda_2^{[1]}$, and $\lambda_\nu^{[2]} = \lambda_\nu^{[1]}$ for all other $\nu \in \mathbb{N}_+$. Define the eigenfunctions as

$$\phi_1^{[2]} = 2^{-1/2}(\phi_4 + \phi_5), \quad \phi_2^{[2]} = \phi_2, \quad \phi_3^{[2]} = \phi_3, \quad \phi_4^{[2]} = \phi_1,$$

and, for $\nu \ge 5$,

$$\phi_\nu^{[2]} = 2^{-1/2}(\phi_{2\nu-4} + \phi_{2\nu-3}).$$

If $m_1 = m_2 = 3$, then the shared dimension is $m_s = 2$. In this case, $\pi_1(\nu) = \nu$ for $\nu \le m_s$, while $\pi_2(1) = 3$ and $\pi_2(2) = 2$.

Remark 1. In this construction, the eigenfunction associated with the largest eigenvalue always corresponds to the source-specific variation, whereas the shared subspace lies in the tail and is associated with relatively small eigenvalues. Consequently, weighting sample covariance operators and extracting leading eigenfunctions to identify the shared subspace may still perform poorly regardless of the choice of weights.

2.3 Multi-source FPCA

To identify the shared structure across multiple sources and generalize the standard FPCA, we consider focusing exclusively on the eigenfunctions while discarding the eigenvalues during the integration step. If source-specific variation is of further interest, analysis can be refined within the source of interest based on the estimated shared structure.

Motivated by the above idea and Example 2.2, we propose the following projection-based estimator:

$$\widehat{\boldsymbol{P}}_w = n_t^{-1} \sum_{k \in \mathcal{K}} n_k \widetilde{\boldsymbol{P}}^{[k]}, \tag{7}$$

where $n_t = \sum_{k \in \mathcal{K}} n_k$ and $\widetilde{\boldsymbol{P}}^{[k]} = \sum_{\nu=1}^{m_k} \tilde{\phi}_\nu^{[k]} \otimes \tilde{\phi}_\nu^{[k]}$. Here $\tilde{\phi}_\nu^{[k]}$ is the νth eigenfunction estimated from the smoothed covariance function of the kth source. Note that, consistency of $\tilde{\phi}_\nu^{[k]}$ implies the consistency of $\widetilde{\boldsymbol{P}}^{[k]}$.

The target of interest is the leading m_s eigenfunctions of $\widehat{\boldsymbol{P}}_w$, or equivalently, the subspace they span. Let $\widehat{\boldsymbol{P}}_s$ denote the projection operator onto this estimated shared subspace. Then, it follows from (2) that

$$\begin{aligned} \widehat{\boldsymbol{P}}_s &= \underset{\boldsymbol{P} \in \mathcal{P}_{m_s}}{\arg\max} \frac{1}{n_t} \sum_{k \in \mathcal{K}} n_k \left\langle \widetilde{\boldsymbol{P}}^{[k]}, \boldsymbol{P} \right\rangle_{\mathrm{HS}} \\ &= \underset{\boldsymbol{P} \in \mathcal{P}_{m_s}}{\arg\min} \frac{1}{n_t} \sum_{k \in \mathcal{K}} n_k \|\widetilde{\boldsymbol{P}}^{[k]} - \boldsymbol{P}\|_{\mathrm{HS}}^2, \end{aligned}$$

where $\mathcal{P}_{m_s}$ denotes the collection of rank-m_s projection operators in $\mathcal{H}$. Therefore, the projection operator of the shared structure minimizes the weighted average of Hilbert–Schmidt distances between $\widetilde{\boldsymbol{P}}^{[k]}$ and $\boldsymbol{P}$ across all sources.

Note that at the population level,

$$\boldsymbol{P}_w = n_t^{-1} \sum_{k \in \mathcal{K}} n_k \boldsymbol{P}^{[k]} = \boldsymbol{P}_s + \frac{1}{n_t} \sum_{k \in \mathcal{K}} n_k \boldsymbol{P}_p^{[k]}, \tag{8}$$

which implies that the eigenvalues corresponding to the shared subspace are always equal to one. While the largest eigenvalue of $\sum_{k \in \mathcal{K}} n_k \boldsymbol{P}_p^{[k]} / n_t$ is strictly smaller than one. Since $\widetilde{\boldsymbol{P}}^{[k]}$ is consistent, $\widehat{\boldsymbol{P}}_w$ is also consistent. Consequently, the rank of the shared subspace can be identified from the scree plot of $\widehat{\boldsymbol{P}}_w$; see Figure 2 for illustration. Further discussions of the eigengap between the shared subspace and $\sum_{k \in \mathcal{K}} n_k \boldsymbol{P}_p^{[k]} / n_t$ are provided in Remark 2 and Example 3. Let $(\hat{\theta}_\nu, \hat{\psi}_\nu)$ denote the νth eigen-pair of $\widehat{\boldsymbol{P}}_w$. Therefore, we have

$$\widehat{\boldsymbol{P}}_s = \sum_{\nu=1}^{m_s} \hat{\psi}_\nu \otimes \hat{\psi}_\nu. \tag{9}$$

If source-specific variation for source k is of interest and $m_k > m_s$, the corresponding projector $\boldsymbol{P}_p^{[k]}$, which is defined in (6), can be estimated using only the kth source data after obtaining $\widehat{\boldsymbol{P}}_s$. The space associated with $\widehat{\boldsymbol{P}}_s$ can be viewed as an average of the shared subspace across K sources. Consequently, the basis functions generating $\widehat{\boldsymbol{P}}_s$ are not necessarily orthogonal to those generating the source-specific variation. To preserve orthogonality, we need to define a projection operator $\widehat{\boldsymbol{P}}_d$ that projects functions onto $\widehat{\mathcal{H}}_d^{[k]} = \widetilde{\mathcal{H}}_0^{[k]} \ominus \widehat{\mathcal{H}}_s$, where $\widetilde{\mathcal{H}}_0^{[k]}$ and $\widehat{\mathcal{H}}_s$ are two subspaces associated with $\widetilde{\boldsymbol{P}}^{[k]}$ and $\widehat{\boldsymbol{P}}_s$, respectively. Define $\widehat{\boldsymbol{T}}^{[k]} = \widetilde{\boldsymbol{P}}^{[k]}(\mathrm{id} - \widehat{\boldsymbol{P}}_s)\widetilde{\boldsymbol{P}}^{[k]}$, where id is the identity

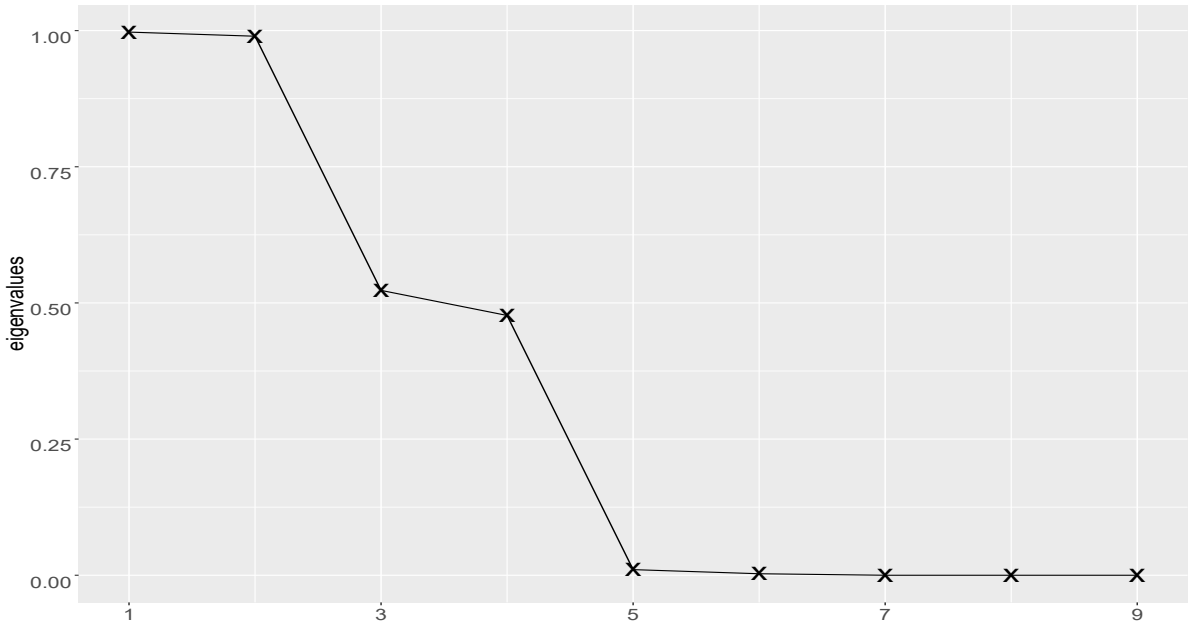

Figure 2: Scree plot of $\widehat{\boldsymbol{P}}_w$ based on Example 2.2 with $m_1 = m_2 = 3$. Data are generated from two sources, each with 200 subjects and 25 observations per subject.

operator that maps any functions in $\mathcal{H}$ to itself. Denote by $\widehat{\boldsymbol{P}}_d$ the projection operator generated by the eigenfunctions of $\widehat{\boldsymbol{T}}^{[k]}$ associated with the eigenvalue one. Note that $\widehat{\boldsymbol{T}}^{[k]}(f) = f$ for any $f \in \widehat{\mathcal{H}}_d^{[k]}$, which implies $\widehat{\boldsymbol{P}}_d(f) = f$. The construction of $\widehat{\boldsymbol{P}}_d$ implicitly forces any function not in $\widetilde{\mathcal{H}}_0^{[k]} \cap \widehat{\mathcal{H}}_s^{\perp}$ to be zero as we set these eigenvalues to be 0. Thus, $\widehat{\boldsymbol{P}}_d(f) = 0$ for any $f \in \mathcal{H} \ominus \widehat{\mathcal{H}}_d^{[k]}$. To ensure the eigenfunctions in the source-specific subspace are identifiable, we then project $\widetilde{\boldsymbol{G}}^{[k]}$, the estimated covariance function for the kth source, onto $\widehat{\boldsymbol{P}}_d^{[k]}$. The source-specific subspace $\widehat{\boldsymbol{P}}_p^{[k]}$ is then represented by the leading $m_k - m_s$ eigenfunctions of

$$\widehat{\boldsymbol{G}}_p^{[k]} = \widehat{\boldsymbol{P}}_d^{[k]} \widetilde{\boldsymbol{G}}^{[k]} \widehat{\boldsymbol{P}}_d^{[k]}.$$

The refined estimator of projection operator for the kth source is given by $\widehat{\boldsymbol{P}}^{[k]} = \widehat{\boldsymbol{P}}_p^{[k]} + \widehat{\boldsymbol{P}}_s$. The entire procedure is summarized in Algorithm 1.

3 Statistical Theory

This section is devoted to the theoretical justification of the proposed method. The shared structure is formalized in Section 2.2. In what follows, let $\|\boldsymbol{A}\|_{\text{op}} = \sup_{f \in \mathcal{H}} \|\boldsymbol{A}(f)\|_2 / \|f\|_2$ denote the operator norm of $\boldsymbol{A}$. For a compact operator, the operator norm equals its largest eigenvalue. For any continuous function f defined on $[0, 1]$, $\|f\|_\infty = \sup_{t \in [0,1]} |f(t)|$. Let $a_n \lesssim b_n$ denote $a_n \leq C b_n$ for all $n \in \mathbb{N}^+$ with some positive constant C, and $\gtrsim$ is defined similarly. Lastly, $a_n \asymp b_n$ if $a_n \lesssim b_n$ and $b_n \lesssim a_n$. We need the following assumptions to develop theoretical properties for the proposed estimator.

Assumption 1. The shared space $\mathcal{H}_s$ is spanned by the same basis functions across all sources. Furthermore, we assume $m_s \leq \min_{k \in \mathcal{K}} m_k$.

Algorithm 1 Multi-source FPCA

1. For each source k, estimate the covariance function $\widetilde{G}^{[k]}(t,s)$ and eigenpairs $\{(\tilde{\lambda}_\nu^{[k]}, \tilde{\phi}_\nu^{[k]}) \mid \nu = 1, 2, \ldots, m_k\}$ using the method detailed in Section 2.1. Moreover, compute $\widetilde{\boldsymbol{P}}^{[k]} = \sum_{\nu=1}^{m_k} \tilde{\phi}_\nu^{[k]} \otimes \tilde{\phi}_\nu^{[k]}$.

2. Construct the weighted average projection operator $\widehat{\boldsymbol{P}}_w$ as in equation (7), and perform eigendecomposition of its discretized version to obtain eigenpairs $(\hat{\theta}_\nu, \hat{\psi}_\nu)$.

3. Identify the rank m_s of $\mathcal{H}_s$ based on the eigenvalues of $\widehat{\boldsymbol{P}}_w$. The projection operator of the shared subspace is determined by using equation (9).

4.1 (optional) If source-specific variation for source k is of interest and $m_k > m_s$, obtain $\widehat{\boldsymbol{P}}_d^{[k]}$ by acquiring the eigenspace of $\widetilde{\boldsymbol{P}}^{[k]}(\mathrm{id} - \widehat{\boldsymbol{P}}_s)\widetilde{\boldsymbol{P}}^{[k]}$ associated with the eigenvalue equals one. Then, we re-project $\widetilde{\boldsymbol{G}}^{[k]}$ onto the space associated with $\widehat{\boldsymbol{P}}_d^{[k]}$ to obtain $\widehat{\boldsymbol{G}}_p^{[k]} = \widehat{\boldsymbol{P}}_d^{[k]} \widetilde{\boldsymbol{G}}^{[k]} \widehat{\boldsymbol{P}}_d^{[k]}$.

4.2 Extract the leading $m_k - m_s$ eigenfunctions of $\widehat{\boldsymbol{G}}_p^{[k]}$ to estimate the source-specific subspace and its associated projection operator $\widehat{\boldsymbol{P}}_p^{[k]}$.

Assumption 2. For the source-specific subspaces, we assume

$$\left\| \frac{1}{n_t} \sum_{k \in \mathcal{K}} n_k \boldsymbol{P}_p^{[k]} \right\|_{\mathrm{op}} \leq 1 - d,$$

where $d > 0$ is a constant that is independent of sample sizes.

Remark 2. Note that the largest eigenvalue of $\boldsymbol{P}_w$ defined in (8) satisfies $\lambda_1(P_w) = 1$. Assumption 2 essentially provides a lower bound on the discrepancy between the largest two eigenvalues of P_w. The magnitude of $1 - d$ characterizes the degree of similarity among the source-specific subspaces across sources. A smaller d indicates greater challenges in estimating the shared space. In particular, if $d = 0$, then an eigenfunction is shared by all sources, and by Assumption 1, this eigenfunction must belong to $\mathcal{H}_s$.

Example. Let us revisit Example 2.2 and further assume $n_1 = n_2$. It is easy to see that $\left\| 2^{-1} \sum_{k \in \mathcal{K}} \boldsymbol{P}_p^{[k]} \right\|_{\mathrm{op}} = 1/2$, and $d = 1/2$. If instead, we let $\phi_1^{[2]} = 2^{-1/2}(\phi_1 + \phi_4)$ and $\phi_4^{[2]} = \phi_5$ while keeping $m_1 = m_2 = 3$, then $\|2^{-1} \sum_{k \in K} \boldsymbol{P}_p^{[k]}\|_{\mathrm{op}} = 1/2 + 2^{1/2}/4$, and $d = 1/2 - 2^{1/2}/4$. Thus, d is smaller in second case as $\phi_1^{[2]}$ is partially aligned with $\phi_1^{[1]} = \phi_1$, indicating an overlap between two source-specific subspaces. In contrast, if $\phi_1^{[2]}$ is completely aligned with $\phi_1^{[1]}$, then $m_s = 3$ and this eigenfunction belongs to the shared space.

Assumptions 3 - 8 should be understood as they hold for any $k \in \mathcal{K}$. For brevity, we omit the superscript $[k]$ or k. These assumptions ensure that the covariance function and eigenpairs are well estimated for each source.

Assumption 3. There exists a positive constant c_0 such that $\mathbb{E}\xi_\nu^4 \le c_0(\lambda_\nu)^2$.

Assumption 4. The second order partial derivatives of $G(s,t)$, $\partial G(s,t)/\partial t\partial s$, $\partial G(s,t)/\partial s^2$, and $\partial G(s,t)/\partial t^2$, are all bounded on $\mathcal{T}\times\mathcal{T}$.

Assumption 5. The eigenvalues of $\boldsymbol{G}$ follow a polynomial decay rate. That is, there exists a positive constant $a>1$ such that, for any $\nu\in\mathbb{N}_+$,

$$\nu^{-a} \gtrsim \lambda_\nu \gtrsim \lambda_{\nu+1} \gtrsim \nu^{-(a+1)}$$

Assumption 6. Suppose $\|\phi_\nu\|_\infty = O(1)$ for $\nu\in\mathbb{N}_+$, and there exists a positive constant c

$$\|\phi_\nu^{(\ell)}\|_\infty \lesssim \nu^{c/2}\|\phi_\nu^{(\ell-1)}\|_\infty \quad \text{for } \ell=1,2,$$

where $f^{(\ell)}$ denotes the ℓth derivative of f.

Assumption 7. $\mathbb{E}(\|X\|_\infty^{2\alpha})$ and $\mathbb{E}(|\varepsilon|^{2\alpha})$ are bounded for some $\alpha>3$.

Assumption 8. The observation times T_{ij} are independent of X and ε. Moreover, T_{ij}'s are iid with a density function bounded away from zero and infinity on $\mathcal{T}$.

Assumptions 3 and 4 are widely adopted in the functional data literature when smoothing methods are employed; see Yao et al. (2005) and Cai and Yuan (2011) for example. Assumption 5 considers polynomial decay rates for eigenvalues, where the resulting lower bounds of eigengaps facilitate a theoretical framework for the estimated functional principal components (Hall and Horowitz, 2007; Zhou et al., 2023). Assumption 6 was considered in Zhou et al. (2024) to control the smoothness of eigenfunctions and their derivatives. In particular, it is easy to see that $\|\phi_\nu^{(2)}\|_\infty \lesssim \nu^c$, which implies that higher order derivatives can grow up to a polynomial rate with respect to the index ν of the eigenfunction. The moments condition in Assumption 7 ensures the uniform convergences of the estimated covariance function, which can be found, for example, in Li and Hsing (2010) and Zhang and Wang (2016).

We first present the convergence rate of $\widehat{\boldsymbol{P}}_s$ in terms of the operator norm. Note that the projection operator is the aggregation of the projection operator for each source. Thus, Theorems 3.1 and 3.2 are based on the FPCA theory for a single source.

Theorem 3.1. *Suppose Assumptions 1 and 2 hold. For each source k, suppose Assumptions 3 - 8 hold with respective constants, and denote*

$$\bar{N}_2^{[k]} = \left\{ n_k^{-1}\sum_{i=1}^{n_k}(N_i^{[k]})^{-2} \right\}^{-1/2}.$$

For any $k\in\mathcal{K}$, we further assume $m_k^{2a_k+2}/n_k = o(1)$, $\bar{N}_2^{[k]} \gtrsim m_k$, and for $\nu \le m_k$,

$$\frac{m_k^{2a_k+2}}{n_k(\bar{N}_2^{[k]})^2 h_k^2(\nu)} = o(1),\ \ h_k^4(\nu)\max\{m_k^{2a_k+2c_k}, m_k^{4a_k}\log(n_k)\} = o(1),$$

where

$$h_k(\nu) = (n_k\bar{N}_2^{[k]})^{-1/5}\nu^{(a_k-2c_k-2)/5}(1+\nu^{a_k}/\bar{N}_2^{[k]})^{1/5}.$$

Under these assumptions, we have

$$\|\widehat{\boldsymbol{P}}_s - \boldsymbol{P}_s\|_{op} = O_p(\kappa_{n_t}).$$

Here,

- *When* $\bar{N}_2^{[k]} \gtrsim m_k^{a_k}$, *we further assume* $c_k + 2 < 4a_k$ *for all* $k \in \mathcal{K}$. *Then,*

$$\kappa_{n_t} = \frac{\left(\sum_{k\in\mathcal{K}} m_k^4\right)^{1/2}}{n_t^{1/2}} + \frac{\left(\sum_{k\in\mathcal{K}} m_k^{c_k/2+3}\right)^{2/5}}{n_t^{2/5}}.$$

If $\bar{N}_2^{[k]} \gtrsim n_k^{1/4} m_k^{a_k+c_k/2-2}$ *for all* $k \in \mathcal{K}$, *then*

$$\kappa_{n_t} = \frac{\left(\sum_{k\in\mathcal{K}} m_k^4\right)^{1/2}}{n_t^{1/2}}.$$

- *When* $\bar{N}_2^{[k]} \asymp m_k$ *for all* $k \in \mathcal{K}$, *we further assume* $\max_{k\in\mathcal{K}} 2(a_k+1)/(a_k+c_k/2+1) > 1$. *Then,*

$$\kappa_{n_t} = \frac{\left(\sum_{k\in\mathcal{K}} m_k^{2a_k+2}\right)^{1/2}}{n_t^{1/2}} + \frac{\left(\sum_{k\in\mathcal{K}} m_k^{2a_k+c_k/2+1}\right)^{2/5}}{n_t^{2/5}}.$$

Remark 3. The additional conditions on a_k and c_k arise from aggregating estimated eigenfunctions within the shared space, which are very mild. For example, standard bases used in FDA, including Fourier, Legendre, and wavelet, satisfy Assumption 6 with $c_k = 2$. The polynomial decay rate for eigenvalues in Assumption 5 normally assumes $a_k > 1$. Consequently, both $c_k + 2 < a_k$ and $\max_{k\in\mathcal{K}} 2(a_k+1)/(a_k+c_k/2+1) > 1$ hold.

Remark 4. To illustrate the efficiency gain from data integration, consider the following example. For simplicity, we focus on the dense case, where $\kappa_{n_t} = \left(\sum_{k\in\mathcal{K}} m_k^4\right)^{1/2} / n_t^{1/2}$. Suppose there exists a source k_0 such that $n_{k_0} \asymp n_t$ and $n_k/n_{k_0} = o(1)$ for all $k \neq k_0$. When $m_k \asymp m$ for all k, we have $\kappa_{n_t} \asymp m^2/n_{k_0}^{1/2}$. In contrast, if only source k is available for some $k \neq k_0$, the convergence rate for estimating the projection operator of the shared subspace is $O_p[(m_s m)/n_k^{1/2}]$; see Proposition 1 in the supplementary material. This rate is based on the oracle setting where the true indices of the shared eigenfunctions are known for source k. The ratio between the two convergence rates is $(mn_k)^{1/2}/(m_s n_{k_0})$. For example, if $m_s \asymp m$, the ratio is $o(1)$. This demonstrates the improvement in estimation efficiency achieved through data integration. This theoretical result aligns with the numerical findings reported in Table 1 in Section 4.

Theorem 3.2. *Under Assumptions for Theorem 3.1, if* $m_k > m_s$, *then the estimator of the kth source-specific subspace satisfies*

$$\left\| \widehat{\boldsymbol{P}}_p^{[k]} - \boldsymbol{P}_p^{[k]} \right\|_{op} = O_p(\zeta_{n_k}).$$

Here,

- *When* $\bar{N}_2^{[k]} \gtrsim m_k^{a_k}$,

$$\zeta_{n_k} = \frac{(m_k - m_s)^2}{n_k^{1/2}} + \left\{ \frac{(m_k - m_s)^{c_k/2+3}}{n_k} \left(\frac{m_k - m_s}{m_k}\right)^{a_k} \right\}^{2/5}.$$

If we further assume $\bar{N}_2^{[k]} \gtrsim n_k^{1/4} m_k^{a_k+c_k/2-2}$, *then*

$$\zeta_{n_t} = \frac{(m_k - m_s)^2}{n_k^{1/2}}.$$

- *When* $\bar{N}_2^{[k]} \asymp m_k$,

$$\zeta_{n_k} = \frac{(m_k - m_s)^{a_k+1}}{n_k^{1/2}} + \frac{(m_k - m_s)^{(4a_k+c_k+2)/5}}{n_k^{2/5}}.$$

Remark 5. Let the source-specific subspace for source k_0 be denoted by $\mathcal{H}_d^{[k_0]} = \mathcal{H}_0^{[k_0]} \ominus \mathcal{H}_s$. When $m_{k_0} > m_s$, it is not surprising that the convergence rate in Theorem 3.2 is of the same order as that of a single-source estimator, unless $\dim(\mathcal{H}_d^{[k_0]}) \ll \dim(\mathcal{H}_s)$. In particular, we focus on the dense case discussed in Remark 4, and suppose $m_k \asymp m$ for all $k \in \mathcal{K}$. In the oracle setting where the true indices of the eigenfunctions spanning $\mathcal{H}_d^{[k_0]}$ are known, the convergence rate for estimating the projection operator of the source-specific subspace is $O_p[(m-m_s)m/n_{k_0}^{1/2}]$ based on Proposition 1 in the supplementary material, while Theorem 3.2 gives $\zeta_{n_{k_0}} = (m-m_s)^2/n_{k_0}^{1/2}$. Thus, the ratio between the two convergence rates is $1/(1+r)$, where $r = \dim(\mathcal{H}_s)/\dim(\mathcal{H}_d^{[k_0]})$. This result implies that if $\dim(\mathcal{H}_d^{[k_0]}) \asymp \dim(\mathcal{H}_s)$, the efficiency gain is asymptotically constant; however, when $\dim(\mathcal{H}_d^{[k_0]}) \ll \dim(\mathcal{H}_s)$, the gain becomes substantial.

For example, in the extreme case where $\dim(\mathcal{H}_d^{[k_0]}) = 1$ and $m_s \asymp m \to \infty$ as $n_k \to \infty$ for all $k \in \mathcal{K}$, the estimated subspace $\widehat{\mathcal{H}}_d^{[k_0]}$ must be orthogonal to the estimated shared subspace, and accurate estimation of $\mathcal{H}_s$ markedly improves the recovery of $\mathcal{H}_d^{[k_0]}$. Conversely, if $\dim(\mathcal{H}_s) = 1$ but $\dim(\mathcal{H}_d^{[k_0]}) \to \infty$ as $n_k \to \infty$, improvements in estimating $\mathcal{H}_s$ have negligible effect on the estimation of $\mathcal{H}_d^{[k_0]}$. In summary, although the information for estimating the source-specific subspace resides solely within source k_0, the orthogonality constraint between $\mathcal{H}_s$ and $\mathcal{H}_0^{[k_0]} \ominus \mathcal{H}_s$ can help enhance estimation efficiency.

4 Simulation

In this section, we perform simulation studies to investigate the finite sample performance of the proposed method, denoted as Multi-FPCA. We consider three sources in total. To showcase the superior performance of the proposed method, we consider three different sample size settings: 1.) $n_1 = 50$ and $n_2 = n_3 = 200$; 2.) $n_1 = 100, n_2 = n_3 = 200$; 3.) $n_1 = 100, n_2 = n_3 = 400$. In addition, we consider the number of observations per subject $N_i = N \in \{10, 25, 50\}$, where $N = 10$ and 50 are referred to as sparse and dense functional data, respectively. The intermediate case, $N = 25$, represents a transition state between sparse and dense, hereby referred to as neither.

We generate data based on model (4) and the Karhunen–Loève expansion in (5). Observation times $T_{ij}^{[k]}$ are uniformly generated in $[0, 1]$. For brevity, we assume the true mean function is zero. In terms of the covariance structure, we first denote the ordinary Fourier basis as follows:

$$\phi_1(t) = 1, \quad \phi_\nu(t) = \begin{cases} \sqrt{2}\sin(\nu\pi t) \text{ if } \nu \geq 2 \text{ and } \nu \text{ is even}, \\ \sqrt{2}\cos\{(\nu-1)\pi t\} \text{ if } \nu \geq 2 \text{ and } \nu \text{ is odd}. \end{cases}$$

The eigen-systems for the three sources are specified as follows:

Source 1: Set $\lambda_1^{[1]} = 24, \lambda_2^{[1]} = 12, \lambda_3^{[1]} = 6$, and $\lambda_\nu = 0$ for $\nu \geq 4$. The eigenfunctions are $\phi_\nu^{[1]} = \phi_\nu$ for $\nu \in \mathbb{N}_+$. The variance of the measurement error, σ_ε^2, is 0.49.

Source 2: Set $\lambda_1^{[2]} = 12, \lambda_2^{[2]} = 10, \lambda_3^{[2]} = 8, \lambda_4^{[2]} = 4$, and $\lambda_\nu = 0$ for $\nu \geq 5$. As for the eigenfunctions, $\phi_1^{[2]} = \phi_3$, $\phi_2^{[2]} = (\sin\theta)\phi_4 + (\cos\theta)\phi_5$, $\phi_3^{[2]} = \phi_2$, $\phi_4^{[2]} = (\sin\theta)\phi_6 + (\cos\theta)\phi_7$ with $\theta = \pi/4$.

The variance of the measurement error, σ_ε^2, is 0.16.

Source 3: Set $\lambda_1^{[3]} = 20, \lambda_2^{[3]} = 10, \lambda_3^{[3]} = 5$, and $\lambda_\nu = 0$ for $\nu \geq 4$. The eigenfunctions are $\phi_1^{[3]} = (\sin\omega)\phi_4 + (\cos\omega)\phi_5$, $\phi_2^{[3]} = \phi_2$, $\phi_3^{[3]} = \phi_3$ with $\omega = \pi/3$. The variance of the error, σ_ε^2, is 0.16.

The above settings imply $m_s = 2$.

To illustrate the benefit of integrating information, we compare Multi-FPCA with an estimator using only source 1 data. For the latter, we first estimate $\widetilde{\boldsymbol{P}}^{[1]}$ with the method in Section 2.1. We then compute $\widetilde{\boldsymbol{P}}_s^{[1]}$ and $\widetilde{\boldsymbol{P}}_p^{[1]}$ by providing the true indices of the shared eigenfunctions in the first source. In particular, $\widetilde{\boldsymbol{P}}_s^{[1]} = \tilde{\phi}_2^{[1]} \otimes \tilde{\phi}_2^{[1]} + \tilde{\phi}_3^{[1]} \otimes \tilde{\phi}_3^{[1]}$. It is worth noting that the true indices are provided in computing $\widetilde{\boldsymbol{P}}_s^{[1]}$ and $\widetilde{\boldsymbol{P}}_p^{[1]}$, while the indices of the shared eigenfunctions are always estimated from the data for Multi-FPCA. The estimation procedure for the Multi-FPCA is detailed in Section 2.3 and summarized in Algorithm 1. 500 independent simulation trials were run for each setting mentioned above.

To assess the performance of these methods, we consider the following metric:

$$\frac{1}{M}\sum_{i=1}^{M}\|\widehat{\boldsymbol{P}}_i - \boldsymbol{P}\|,$$

where $M = 500$ and $\|\cdot\|$ denotes either the operator norm or Hilbert–Schmidt norm. Here $\widehat{\boldsymbol{P}}_i$ denotes the generic estimate of the target (sub)space of interest, i.e., the entire space, the shared subspace or the source-specific subspace, based on Multi-FPCA or using the source 1 data only from the ith simulation run, and $\boldsymbol{P}$ is the corresponding true (sub)space.

We report both the averaged norm differences and the associated standard error. The results of estimating the shared subspace and the source-specific subspace for source 1 are displayed in Table 1 and Table 2, respectively. Table 3 summarizes the results for estimating the entire space for source 1.

Table 1: Mean and standard error (in parenthesis) comparison of the norm differences for the *shared* subspace across 500 Monte Carlo runs between Multi-FPCA and using the source 1 data only.

(n_1, n_2, n_3)	N	Multi-FPCA		Source 1 only	
		$\|\cdot\|_{\text{op}}$	$\|\cdot\|_{\text{HS}}$	$\|\cdot\|_{\text{op}}$	$\|\cdot\|_{\text{HS}}$
$(50, 200, 200)$	10	0.188(0.06)	0.294(0.09)	0.306(0.10)	0.475(0.15)
	25	0.117(0.03)	0.188(0.04)	0.251(0.09)	0.388(0.13)
	50	0.089(0.02)	0.144(0.03)	0.247(0.11)	0.377(0.14)
$(100, 200, 200)$	10	0.160(0.04)	0.254(0.06)	0.238(0.07)	0.368(0.11)
	25	0.101(0.02)	0.166(0.04)	0.196(0.07)	0.306(0.10)
	50	0.077(0.01)	0.126(0.02)	0.181(0.07)	0.279(0.09)
$(100, 400, 400)$	10	0.142(0.05)	0.224(0.06)	0.237(0.07)	0.368(0.10)
	25	0.088(0.02)	0.142(0.03)	0.189(0.06)	0.296(0.09)
	50	0.067(0.02)	0.108(0.02)	0.176(0.06)	0.272(0.08)

Table 2: Mean and standard error (in parenthesis) comparison of the norm differences for the *source-specific* subspace across 500 Monte Carlo runs between Multi-FPCA and using the source 1 data only.

(n_1, n_2, n_3)	N	Multi-FPCA		Source 1 only	
		$\|\cdot\|_{\text{op}}$	$\|\cdot\|_{\text{HS}}$	$\|\cdot\|_{\text{op}}$	$\|\cdot\|_{\text{HS}}$
$(50, 200, 200)$	10	0.122(0.07)	0.167(0.08)	0.208(0.11)	0.287(0.15)
	25	0.091(0.03)	0.126(0.04)	0.201(0.10)	0.279(0.15)
	50	0.085(0.05)	0.117(0.05)	0.216(0.12)	0.299(0.16)
$(100, 200, 200)$	10	0.103(0.07)	0.142(0.08)	0.152(0.09)	0.209(0.11)
	25	0.078(0.02)	0.109(0.03)	0.153(0.07)	0.212(0.1)
	50	0.067(0.02)	0.093(0.02)	0.153(0.08)	0.212(0.11)
$(100, 400, 400)$	10	0.094(0.08)	0.128(0.08)	0.158(0.10)	0.216(0.11)
	25	0.073(0.02)	0.102(0.03)	0.150(0.07)	0.207(0.10)
	50	0.065(0.02)	0.090(0.02)	0.148(0.07)	0.205(0.09)

Table 3: Mean and standard error (in parenthesis) comparison of the norm differences for the *entire* space across 500 Monte Carlo runs between Multi-FPCA and using the source 1 data only.

(n_1, n_2, n_3)	N	Multi-FPCA		Source 1 only	
		$\|\cdot\|_{\text{op}}$	$\|\cdot\|_{\text{HS}}$	$\|\cdot\|_{\text{op}}$	$\|\cdot\|_{\text{HS}}$
$(50, 200, 200)$	10	0.191(0.08)	0.307(0.10)	0.261(0.11)	0.403(0.14)
	25	0.126(0.03)	0.213(0.05)	0.182(0.05)	0.296(0.08)
	50	0.103(0.05)	0.176(0.05)	0.152(0.06)	0.257(0.07)
$(100, 200, 200)$	10	0.159(0.07)	0.256(0.08)	0.207(0.09)	0.320(0.11)
	25	0.105(0.02)	0.179(0.04)	0.145(0.04)	0.237(0.06)
	50	0.084(0.01)	0.144(0.02)	0.115(0.03)	0.196(0.04)
$(100, 400, 400)$	10	0.149(0.08)	0.239(0.09)	0.207(0.09)	0.320(0.11)
	25	0.097(0.02)	0.162(0.03)	0.139(0.04)	0.229(0.06)
	50	0.078(0.02)	0.132(0.02)	0.116(0.03)	0.196(0.04)

Generally, the estimation accuracy improves for both methods as sample sizes or the number of observations per subject increases in terms of the norm differences. Tables 1 - 3 evidently illustrate the superior performance of Multi-FPCA. From Table 1, we find Multi-FPCA always outperforms the method using only single source data even if the true indices of the shared eigenfunctions are provided for the latter case. This confirms that Multi-FPCA effectively recovers the shared structure by leveraging information from additional sources, which also helps to reduce the estimation uncertainty. The improved shared-space estimation further enhances the recovery of the source-specific subspace as shown in Table 2. These two improved estimators, $\widehat{\boldsymbol{P}}_s$ and $\widehat{\boldsymbol{P}}_p^{[1]}$, together lead

to a better estimate for the entire subspace, as shown in Table 3.

5 Real Data Illustration

Air pollution has become one of the major environmental concerns in China due to its rapid industrial expansion and urbanization over the past four decades (Guan et al., 2016). Fine particulate matter (PM) with an aerodynamic diameter less than 2.5 µm, often referred to as $PM_{2.5}$, is one of the main pollutants in multiple cities in China (Huang et al., 2014). Cardiovascular and respiratory diseases, and even lung cancer, are associated with long-term exposure to $PM_{2.5}$ (Pope III et al., 2002; Hoek et al., 2013; Lelieveld et al., 2015).

In this study, our objective is to identify the shared variation pattern of the daily $PM_{2.5}$ levels of five Chinese major cities: Beijing, Chengdu, Guangzhou, Shanghai, and Shenyang. These cities are located in economically developed regions, collectively contributing over 10% of China's Gross Domestic Product and approximately 6% of its total population. The data were collected by U.S. diplomatic posts located in the five cities. The U.S. Embassy in Beijing started releasing hourly $PM_{2.5}$ readings in April 2008, followed by the consulates in Guangzhou, Shanghai, Chengdu, and Shenyang in November and December 2011, June 2012, and April 2013, respectively.

To begin, we stratify the data by human activity patterns and seasonal effects. Holidays and weekends are excluded to ensure that the selected days reflect homogeneous human activity patterns. Given the substantial influence of meteorological conditions on daily $PM_{2.5}$ levels, we analyze the data separately for winter months (November to February of the following year) and summer months (May to August). This partitioning accounts for the winter heating regulated by local governments in northern cities such as Beijing and Shenyang, where heating typically begins in November, following the approach of Liang et al. (2015). Finally, daily measurements are aggregated into weekly averages to mitigate temporal dependence.

The numbers of selected FPCs are 4, 4, 6, 4, 5 for Beijing, Chengdu, Guangzhou, Shanghai, and Shenyang, respectively, in winter, and 5, 6, 4, 5, 5 in summer. Figure 3 displays the eigenvalues obtained from the weighted average of the projection operator for the winter and summer months. In general, the alignment of variability patterns across five cities is stronger in summer than in winter, suggesting human activities patterns are much more similar in summer.

The top two eigenvalues are close to one for both seasons, implying two dominant shared components. The associated eigenfunctions are shown in Figure 4. In both seasons, $\hat{\psi}_2$ increases after the morning rush hour, peaks at approximately 4:30 PM, and declines slightly thereafter, suggesting that it captures afternoon traffic emissions. The shape of $\hat{\psi}_1$ differs between winter and summer, likely reflecting meteorological effects. Overall, these findings indicate the existence of a shared structure in the variation pattern of daily $PM_{2.5}$ levels across the five cities, despite their substantial geographical and climatic diversity.

6 Concluding Remarks

In this paper, we propose a novel procedure for performing FPCA across multiple data sources using projection operators. It adapts to both sparse and dense functional data. The primary goal is to recover the shared variation pattern across sources. Specifically, we estimate a projection operator from each source via local linear regression. Then we construct a weighted averaged projection operator by aggregating the projection operator of each source. With the help of this estimator, we can unveil the variation pattern across sources even if the eigenvalues of the eigenfunctions

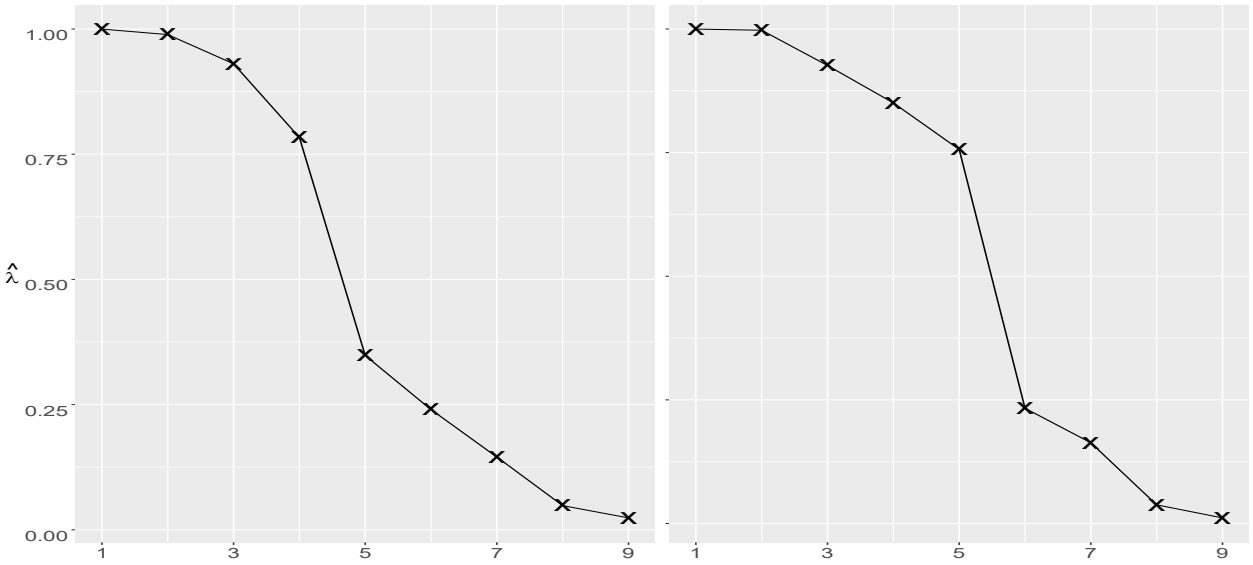

Figure 3: The left panel shows the estimated eigenvalues of $\widehat{\boldsymbol{P}}_{\text{winter}}$ for five cities in the winter months, and the right panel demonstrates the estimated eigenvalues of $\widehat{\boldsymbol{P}}_{\text{summer}}$ in the summer months. Both projected operators are calculated based on equation (7).

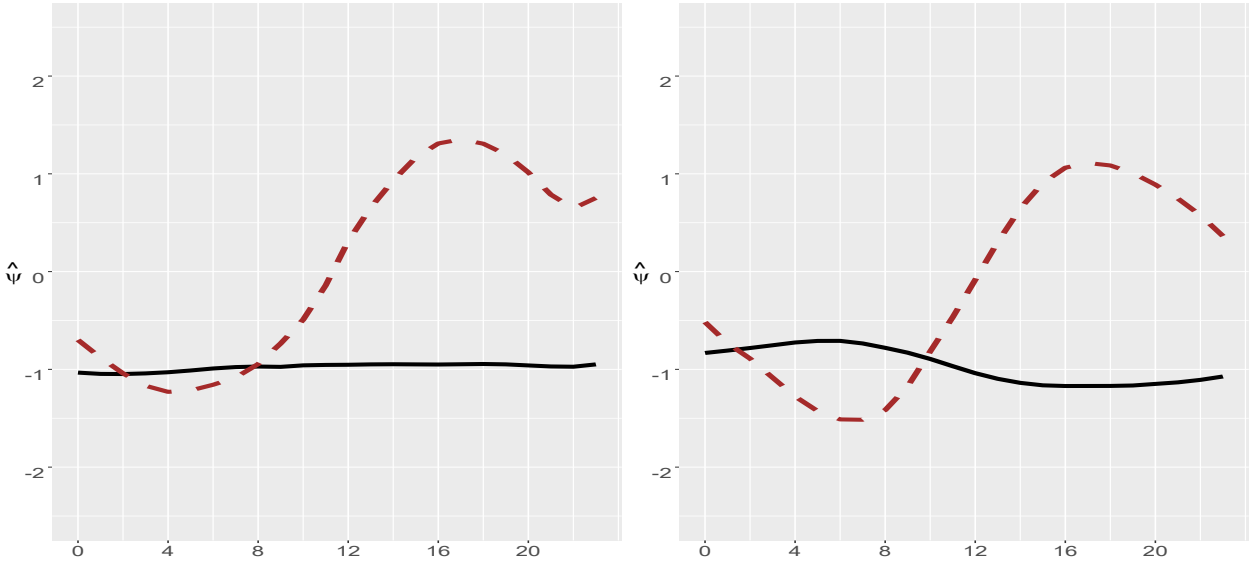

Figure 4: The left panel displays the leading two estimated eigenfunctions of $\widehat{\boldsymbol{P}}_{\text{winter}}$, and the right panel presents those of $\widehat{\boldsymbol{P}}_{\text{summer}}$. In both panels, the solid and dashed lines represent $\hat{\psi}_1$ and $\hat{\psi}_2$, respectively, obtained from the corresponding $\widehat{\boldsymbol{P}}_s$.

spanned this subspace are relatively small within individual sources. When the source-specific variation is of interest, we further propose a re-projection method to estimate the source-specific subspace. We establish asymptotic properties for both the shared space estimator and the source-specific variation estimator. Extensive numerical studies demonstrate the superior performance of the proposed methods. Lastly, we illustrate our method through analyzing the air pollutant data in five major cities in China.

References

Benko, M., Härdle, W., and Kneip, A. (2009). Common functional principal components. *The Annals of Statistics*, 37(1):1 –34.

Boente, G., Rodriguez, D., and Sued, M. (2010). Inference under functional proportional and common principal component models. *Journal of Multivariate Analysis*, 101(2):464–475.

Cai, T. T., Kim, D., and Pu, H. (2024). Transfer learning for functional mean estimation: Phase transition and adaptive algorithms. *The Annals of Statistics*, 52(2):654–678.

Cai, T. T. and Yuan, M. (2011). Optimal estimation of the mean function based on discretely sampled functional data: Phase transition. *The Annals of Statistics*, 39(5):2330–2355.

Dou, W. W., Pollard, D., and Zhou, H. H. (2012). Estimation in functional regression for general exponential families. *The Annals of Statistics*, 40(5):2421–2451.

Guan, W.-J., Zheng, X.-Y., Chung, K. F., and Zhong, N.-S. (2016). Impact of air pollution on the burden of chronic respiratory diseases in China: time for urgent action. *The Lancet*, 388(10054):1939–1951.

Hall, P. and Horowitz, J. L. (2007). Methodology and convergence rates for functional linear regression. *The Annals of Statistics*, 35(1):70–91.

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(1):109–126.

Hoek, G., Krishnan, R. M., Beelen, R., Peters, A., Ostro, B., Brunekreef, B., and Kaufman, J. D. (2013). Long-term air pollution exposure and cardio-respiratory mortality: a review. *Environmental Health*, 12(1):43.

Huang, R.-J., Zhang, Y., Bozzetti, C., Ho, K.-F., Cao, J.-J., Han, Y., Daellenbach, K. R., Slowik, J. G., Platt, S. M., Canonaco, F., Zotter, P., Wolf, R., Pieber, S. M., Bruns, E. A., Crippa, M., Ciarelli, G., Piazzalunga, A., Schwikowski, M., Abbaszade, G., Schnelle-Kreis, J., Zimmermann, R., An, Z., Szidat, S., Baltensperger, U., Haddad, I. E., and Prévôt, A. S. H. (2014). High secondary aerosol contribution to particulate pollution during haze events in China. *Nature*, 514(7521):218–222.

Jiao, S., Frostig, R., and Ombao, H. (2024). Filtrated common functional principal component analysis of multigroup functional data. *The Annals of Applied Statistics*, 18(2):1160–1177.

Kokoszka, P. and Reimherr, M. (2017). *Introduction to Functional Data Analysis*. Chapman and Hall/CRC, New York.

Lelieveld, J., Evans, J., Fnais, M., Giannadaki, D., and Pozzer, A. (2015). The contribution of outdoor air pollution sources to premature mortality on a global scale. *Nature*, 525:367–71.

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(6):3321–3351.

Li, Y., Wang, N., and Carroll, R. J. (2013). Selecting the number of principal components in functional data. *Journal of the American Statistical Association*, 108(504):1284–1294.

Li, Z., Qin, K., He, Y., Zhou, W., and Zhang, X. (2024). Knowledge transfer across multiple principal component analysis studies. DOI: 10.48550/arXiv.2403.07431.

Liang, X., Zou, T., Guo, B., Li, S., Zhang, H., Zhang, S., Huang, H., and Chen, S. X. (2015). Assessing Beijing's $PM_{2.5}$ pollution: severity, weather impact, APEC and winter heating. *Proceedings of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 471(2182):20150257.

Pope III, C. A., Burnett, R. T., Thun, M. J., Calle, E. E., Krewski, D., Ito, K., and Thurston, G. D. (2002). Lung cancer, cardiopulmonary mortality, and long-term exposure to fine particulate air pollution. *JAMA*, 287(9):1132–1141.

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

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: Statistical Methodology*, 53(1):233–243.

Yao, F., Müller, H.-G., and Wang, J.-L. (2005). Functional data analysis for sparse longitudinal data. *Journal of the American Statistical Association*, 100(470):577–590.

Zhang, C., Sang, P., and Qin, Y. (2025). Order determination for functional data. DOI: 10.48550/arXiv.2503.03000.

Zhang, X. and Wang, J.-L. (2016). From sparse to dense functional data and beyond. *The Annals of Statistics*, 44(5):2281–2321.

Zhou, H., Wei, D., and Yao, F. (2024). Theory of functional principal component analysis for noisy and discretely observed data. DOI: 10.48550/arXiv.2209.08768, accepted by the Annals of Statistics.

Zhou, H., Yao, F., and Zhang, H. (2023). Functional linear regression for discretely observed data: from ideal to reality. *Biometrika*, 110(2):381–393.