Journal of Machine Learning Research 27 (2026) 1-52 Submitted 1/25; Revised 4/26; Published 6/26

Kernel Mean Embedding Deviation Subspace for Unsupervised Learning with Heterogeneous Data

Luoyao Yu LUOYAOYU20@STU.XJTU.EDU.CN
School of Mathematics and Statistics
Xi'an Jiaotong University
China

Lixing Zhu LZHU@BNU.EDU.CN
Center for Statistics and Data Science
Beijing Normal University
China

Ruoqing Zhu RQZHU@ILLINOIS.EDU
Department of Statistics
University of Illinois Urbana-Champaign
USA

Xuehu Zhu* ZHUXUEHU@XJTU.EDU.CN
School of Mathematics and Statistics
Xi'an Jiaotong University
China

Editor: Bharath Sriperumbudur

Abstract

This paper proposes a method for dimension reduction that preserves information in unsupervised learning with high-dimensional heterogeneous data, specifically targeting change point detection and clustering analysis. Our main strategy is to apply a Corrected Kernel Principal Component Analysis (CKPCA) method to construct the so-called kernel mean embedding deviation subspace. The approach efficiently identifies distributional changes in these dimension reduction subspaces for unsupervised dimension reduction. For change point detection, we demonstrate that the locations and number of change points in the dimension-reduced subspaces are identical to those in the original data. Furthermore, we extend this approach to clustering by embedding the original data into nonlinear lower-dimensional spaces, providing enhanced capabilities for clustering analysis. Additionally, we explain the necessity of using CKPCA, as the classical KPCA fails to identify the kernel mean embedding deviation subspace in these problems. Numerical studies on synthetic and real data sets suggest that the dimension reduction versions of existing methods for change point detection and clustering significantly improve the performance of current approaches in finite sample scenarios.

Keywords: dimension reduction functional subspace, unsupervised learning, kernel mean embedding deviation subspace, kernel principal component analysis, change point detection, clustering

*. Corresponding author.

1. Introduction

With the rapid advancement of data collection technologies, modern statistical data sets increasingly face the curse of dimensionality and exhibit substantial heterogeneity. When considering the presence of heterogeneity, the structure of the data may change, leading to statistical challenges such as two-sample tests (Gretton et al., 2012), classification (Rosenblatt, 1958), clustering (Hartigan, 1975), and change point detection (Page, 1954). Meanwhile, the curse of dimensionality often causes classical multivariate statistical methods to break down. Dimension reduction without losing any information from the original data is a crucial technique to address the curse of dimensionality. In regression analysis, this approach is known as sufficient dimension reduction (Li, 1991; Xia et al., 2002; Zhu et al., 2010). The motivation of this paper is to apply a dimension reduction method that preserves information for unsupervised learning with heterogeneous data. Specifically, the proposed method will be applied to change point detection and clustering analysis.

The identification of changes in data structures, such as alterations in mean values, distributions, and clustering, is a critical research domain. This subject has attracted attention across various disciplines including economics, genetics, medicine, image analysis, network data, and public health, as evidenced in the work of ŞErban et al. (2010), Chen and Gupta (2012), Cleynen et al. (2014), Kirch et al. (2015), Bagci et al. (2015), and Gregori et al. (2020).

Given the established methodologies for low-dimensional data (as thoroughly reviewed by Niu et al. (2016)), several nonparametric change point detection methods have been proposed to identify distributional changes. Zou et al. (2014) constructed a nonparametric maximum likelihood approach. Building upon Euclidean distances, Matteson and James (2014) developed the E-Divisive method. The MultiRank method, focusing on rank, was put forth by Lung-Yut-Fong et al. (2015). Arlot et al. (2019) formulated a kernel multiple change point (KCP) algorithm to identify change points, and Madrid et al. (2022) considered a unique algorithm grounded on kernel density estimation.

In the realm of high-dimensional data, the majority of change point detection methods focus on mean changes, with many leveraging cumulative sum (CUSUM) statistics, as initially proposed in Csörgő and Horváth (1997). Aggregation across different dimensions of CUSUM statistics has emerged as a popular and effective approach. Jirak (2015) introduced the coordinate-wise CUSUM-statistics. The sparsified binary segmentation (SBS) method, proposed by Cho and Fryzlewicz (2015), enables the detection of changes in high-dimensional time series. Wang and Samworth (2018) developed a projection-based method under the assumption of sparsity. Enikeeva et al. (2019) introduced a scan-statistic-based algorithm for detecting high-dimensional change points with sparse alternatives. In a different approach, Wang et al. (2022) employed self-normalized U-statistics as an alternative to CUSUM statistics. However, few methods exist for detecting distribution changes in high-dimensional scenarios.

Reducing the dimensionality of data is crucial for addressing the curse of dimensionality. Principal Component Analysis (PCA) is a popularly used method. For instance, Kuncheva and Faithfull (2012) applied PCA to focus on the mean and covariance matrix. Qahtan et al. (2015) fused a semi-parametric log-likelihood change detector with PCA. Jiao et al. (2021) crafted a spectral PCA change point method. Despite these advancements, these

studies lack theoretical evaluations explaining PCA's effectiveness for this problem and do not examine whether the dimension reduction is sufficient not to lose information on the original data's change structure.

On the other hand, clustering tasks are often affected by nonlinearity and dimensionality. These factors can cause many clustering methods to perform poorly. In particular, distance-based methods such as K-means frequently fail to yield satisfactory results. As a result, dimension reduction techniques are frequently employed in clustering analysis, both for visualization and for improving clustering accuracy, such as LLE (Roweis and Saul, 2000), t-SNE (Van der Maaten and Hinton, 2008), and UMAP (McInnes et al., 2018). Among the widely used dimension reduction methods in clustering are PCA and KPCA (Armstrong et al., 2002; Alzate and Suykens, 2008; Abbe et al., 2022). Nevertheless, dimension reduction methods such as PCA used in clustering do not generally address whether the reduced data preserve the clustering label information contained in the original data.

Both change point detection and clustering face the same fundamental difficulty when KPCA/PCA is used for dimension reduction: there is no general guarantee that the reduced data preserves the task-relevant information, namely, change point locations or clustering labels. The key common feature of these two problems is that both are unsupervised learning tasks that aim to recover a latent partition of the observations induced by distributional heterogeneity. Therefore, for both problems, a dimension reduction method is appropriate only if it preserves the information that determines change point locations or clustering labels.

The primary focus of this paper is on dimension reduction for unsupervised learning under distributional heterogeneity, with particular emphasis on change point detection and clustering analysis. Our aim is to develop a unified framework in which dimension reduction preserves all task-relevant information. Specifically, we seek a low-dimensional functional vector f such that the transformed data $\{f(X_i)\}_{i=1}^n$ and the original data $\{X_i\}_{i=1}^n$ share the same change point locations or clustering labels. To achieve this goal, we introduce a novel method, called Corrected Kernel Principal Component Analysis (CKPCA), which identifies low-dimensional functional subspaces that achieve nonlinear dimension reduction in reproducing kernel Hilbert spaces (RKHS) for unsupervised learning with heterogeneous data. We refer to these subspaces as the kernel mean embedding deviation subspace.

Moreover, our analysis further explains why the classical dimension reduction methods KPCA/PCA may lose information in heterogeneous data settings. In the framework of kernel mean embeddings, task-relevant information is contained in Δ^{kernel}, which is constructed from the mean structure, whereas the target operator of KPCA/PCA additionally includes Σ_{pooled}^{kernel}, which is determined by the covariance structure. When Σ_{pooled}^{kernel} dominates Δ^{kernel}, KPCA/PCA may fail to preserve task-relevant information; see Section 2. The proposed correction is designed precisely to address this issue, and can therefore be applied more broadly to unsupervised learning problems with heterogeneous data in which the goal is to recover the latent partition structure.

For change point detection, carrying out further detection in the lower-dimensional subspace significantly enhances the efficacy of existing methods, as demonstrated in our numerical studies. For clustering analysis, an iterative subspace clustering algorithm based on CKPCA is implemented to improve the efficiency of classic clustering approaches such as the K-means method (Hartigan and Wong, 1979), the expectation-maximization algorithm

(Fraley and Raftery, 2002), and the density-based spatial clustering of applications with noise method (Ester et al., 1996).

The remaining sections are organized as follows. Section 2 introduces the concept of the kernel mean embedding deviation subspace and the CKPCA method. Section 3 includes the application of the developed dimension reduction technique in clustering, presents an iterative algorithm and give the theoretical justifications. Section 4 features simulation studies and the analysis of several real data sets. Section 5 discusses the advantages and drawbacks of the new method, along with additional research areas. Appendix discusses nonlinear dimension reduction from the perspective of σ-fields, presents a special case where $\phi(X_t) = X_t$ so that CKPCA reduces to CPCA, and includes simulations under mean changes, an analysis of macroeconomic data, a theoretical analysis of the choice of β_n, the regularity conditions, and the technical proofs of the theorems.

2. Corrected Kernel Principal Component Analysis

Given a set of p-dimensional random vectors $X_i = (X_{i1}, \ldots, X_{ip})^\top \in \mathbb{R}^p$, where $\mu_i = E(X_i)$ and $\Sigma_i = \mathrm{Cov}(X_i)$ for $i = 1, 2, \ldots, n$. Assume that $\{X_i\}_{i=1}^n$ follows unknown distributions $\{P_i\}_{i=1}^n$ without any parametric prior and for s change points $1 \leq z_1 < z_2 < \cdots < z_s \leq n$ in distributions such that

$$P_{z_i+1} = \cdots = P_{z_{i+1}} =: P^{(i+1)} \text{ and } P^{(j)} \neq P^{(j+1)}, \quad \forall\, 0 \leqslant i \leqslant s, 1 \leqslant j \leqslant s, \tag{1}$$

where $z_0 = 0$ and $z_{s+1} = n$.

Similar to Celisse et al. (2018) and Arlot et al. (2019), we employ a nonlinear feature map ϕ to transform X_i as follows: $X_i \to \phi(X_i) = Y_i$ for $i = 1, 2, \ldots, n$, where $\phi : \mathcal{X} \to \mathcal{H}$ and $\mathcal{H}$ is a RKHS generated by a kernel K. We consider the model proposed by Arlot et al. (2019):

$$Y_i = \phi(X_i) = \mu_i^\star + \epsilon_i \in \mathcal{H}, \text{ for } i = 1, \ldots, n, \tag{2}$$

where $\mu_i^\star$ is the "mean" element of $\phi(X_i)$ and $\epsilon_i = Y_i - \mu_i^\star$. According to Ledoux and Talagrand (1991), if $\mathcal{X}$ is separable and $EK(X_i, X_i) < +\infty$, then $\mu_i^\star$ is an element uniquely existed in $\mathcal{H}$:

$$\forall f \in \mathcal{H}, \quad \langle \mu_i^\star, f \rangle_{\mathcal{H}} = E \langle \phi(X_i), f \rangle_{\mathcal{H}},$$

where $\langle \cdot, \cdot \rangle_{\mathcal{H}}$ represents the inner product in $\mathcal{H}$. Furthermore, when K is a characteristic kernel such as the Gaussian kernel (Fukumizu et al., 2004; Sriperumbudur et al., 2011), any change at z_i in the distributions P_t implies a change at z_i in the mean elements $\mu_t^\star$. Therefore, the distributional change in the model (1) can be transformed into a mean change problem:

$$\mu_{z_i+1}^\star = \cdots = \mu_{z_{i+1}}^\star =: \mu_d^{(i+1)} \text{ and } \mu_d^{(j)} \neq \mu_d^{(j+1)}, \quad \forall\, 0 \leqslant i \leqslant s, 1 \leqslant j \leqslant s.$$

Based on this result, we introduce the concept of the kernel mean embedding deviation subspace.

Definition 1 $\mathrm{Span}\{\mu_d^{(i)} - \mu_d^{(j)}, \text{ for } i, j = 1, \ldots, s+1\}$ *is called the kernel mean embedding deviation subspace of the sequence* $\{X_i\}_{i=1}^n$ *and is written as* $\mathrm{S}^d_{\{X_i\}_{i=1}^n}$, *and* $q_d = \dim\{\mathrm{S}^d_{\{X_i\}_{i=1}^n}\}$ *is called the structural dimension of* $\mathrm{S}^d_{\{X_i\}_{i=1}^n}$.

It is worth noting that the structural dimension q_d is unknown and satisfies $q_d \leq s$. The following theorem guarantees the integrity of information of the original data in the lower-dimensional subspace.

Theorem 2 *Under Assumptions 2 and 3 in Appendix, for any basis functions* $\{v_1, v_2, \ldots, v_{q_d}\}$ *of* $\mathrm{S}^d_{\{X_i\}_{i=1}^n}$ *with* $q_d \leq s$, *let* $f(X_i) = \left(\langle v_1, Y_i\rangle_{\mathcal{H}}, \ldots, \langle v_{q_d}, Y_i\rangle_{\mathcal{H}}\right)^\top$. *Both the sequences* $\{f(X_i)\}_{i=1}^n$ *and* $\{X_i\}_{i=1}^n$ *have the same locations of changes.*

Following the result in Section 12 of Li (2018), we define the sample covariance operator as:

$$\Sigma_n^{kernel} = \frac{1}{n}\sum_{i=1}^{n}\left(Y_i - \bar{Y}\right)\otimes\left(Y_i - \bar{Y}\right),$$

where $\bar{Y} = \frac{1}{n}\sum_{i=1}^n Y_i$ and the tensor product $f\otimes g$ is the operator on $\mathcal{H}$ such that $(f\otimes g)h = f\langle g, h\rangle_{\mathcal{H}}$ for all $h \in \mathcal{H}$ and two members f and g of $\mathcal{H}$. When $s = o(n)$, $n_i/n \to c_i$ as $n \to \infty$ and $n_i = z_i - z_{i-1}$, by computing the expectation of Σ_n^{kernel}, we have that

$$\begin{aligned} E(\Sigma_n^{kernel}) &= \sum_{j=1}^{s+1}\frac{1}{n}\sum_{i=z_{j-1}+1}^{z_j} E\left\{\left(Y_i - \bar{Y}\right)\otimes\left(Y_i - \bar{Y}\right)\right\} \\ &\to \sum_{j=1}^{s+1} c_j\Sigma_d^{(j)} + \frac{1}{2}\sum_{i=1}^{s+1}\sum_{j=1}^{s+1} c_i c_j\left(\mu_d^{(i)} - \mu_d^{(j)}\right)\otimes\left(\mu_d^{(i)} - \mu_d^{(j)}\right) \\ &=: \Sigma_{pooled}^{kernel} + \Delta^{kernel} = \Sigma^{kernel}, \end{aligned} \tag{3}$$

where $\Sigma^{kernel} = \Sigma_{pooled}^{kernel} + \Delta^{kernel}$ and $\Sigma_d^{(i)}$ refers to the covariance operator of Y_j for $j = z_{i-1}+1, \ldots, z_i$. Moreover, $\Sigma_{pooled}^{kernel} = \sum_{j=1}^{s+1} c_j\Sigma_d^{(j)}$ and $\Delta^{kernel} = \frac{1}{2}\sum_{i=1}^{s+1}\sum_{j=1}^{s+1} c_i c_j(\mu_d^{(i)} - \mu_d^{(j)})\otimes(\mu_d^{(i)} - \mu_d^{(j)})$. The following theorem states that the eigenfunctions of Δ^{kernel} span the kernel mean embedding deviation subspace. Let $\mathrm{ran}(\Delta^{kernel})$ represent the range of Δ^{kernel}, and $\overline{\mathrm{ran}}(\Delta^{kernel})$ be its closure. We refer to KPCA based on Δ^{kernel} as corrected KPCA.

Theorem 3 *Under Assumptions 2 and 3 in Appendix and the model (2),* $\overline{\mathrm{ran}}(\Delta^{kernel}) = \mathrm{S}^d_{\{X_i\}_{i=1}^n}$. *Further, letting* $v_1, \ldots, v_{q_d}$ *denote the eigenfunctions of* Δ^{kernel} *associated with the nonzero eigenvalues of* Δ^{kernel}, $\mathrm{Span}\{v_1, v_2, \ldots, v_{q_d}\} = \mathrm{S}^d_{\{X_i\}_{i=1}^n}$.

Remark 4 *Theorem 2 and Theorem 3 show that dimension reduction based on the operator* Δ^{kernel} *retains all the information about the changes. Therefore, this dimension reduction approach can address the curse of dimensionality in classical change point methods. We also provide a discussion of information preservation in nonlinear dimension reduction for unsupervised learning from a* σ*-field perspective in the following Appendix A.*

Remark 5 *CKPCA is motivated by the goal of preserving change point information, whereas KPCA is aimed at maximizing variance information. From (3), the target operator of KPCA consists of two components, namely* Σ_{pooled}^{kernel} *and* Δ^{kernel}. *Here,* Δ^{kernel} *is determined by the*

mean structure and contains all the change point information, whereas Σ_{pooled}^{kernel} *is determined by the covariance structure. Note that KPCA can estimate the subspace* $\mathrm{S}^d_{\{X_i\}_{i=1}^n}$ *when* $\Sigma_{pooled}^{kernel} = \sigma I$*, where* I *is the identity operator and* σ *is a constant. However, if* $\Sigma_{pooled}^{kernel} \neq \sigma I$ *for any* σ*, the dimension reduction subspace of KPCA involves both* Σ_{pooled}^{kernel} *and* Δ^{kernel}*. As a result, the low-dimensional data sequence obtained through KPCA may not preserve the change structures of the original data sequence. CKPCA solves this problem and thereby achieves dimension reduction that preserves change point information. We note that CKPCA is developed for settings in which structural changes occur in the data. In the special case where no structural change is present and the data are independent and identically distributed,* $\mathrm{S}^d_{\{X_i\}_{i=1}^n}$ *is empty, and* Δ_n^{kernel} *asymptotically goes to zero.*

The proof of Theorem 3 is given in Appendix F.4. To efficiently use the corrected KPCA via Δ^{kernel}, we employ a localized approach to estimate Σ_{pooled}^{kernel} as follows. Let $r = \lfloor n/\beta_n \rfloor$, where $\lfloor \cdot \rfloor$ denotes the floor operation and β_n is an integer-valued tuning parameter that depends on n. Divide the data into r segments: $\mathcal{S}_m = \{(m-1)\beta_n + 1, \ldots, m\beta_n\}$ for $m = 1, 2, \ldots, r-1$, and $\mathcal{S}_r = \{(r-1)\beta_n + 1, \ldots, n\}$. Compute the covariance matrices for each segment and then average them to obtain the final estimator $\Sigma_{pooled,n}^{kernel}$ of Σ_{pooled}^{kernel}:

$$\Sigma_{pooled,n}^{kernel} = \frac{1}{r}\sum_{m=1}^{r} \hat{\Sigma}_m \text{ with } \hat{\Sigma}_m = \frac{1}{\hat{n}_m - 1}\sum_{i \in \mathcal{S}_m} (Y_i - \bar{Y}_m) \otimes (Y_i - \bar{Y}_m),$$

where $\bar{Y}_m = \frac{1}{\hat{n}_m}\sum_{k \in \mathcal{S}_m} Y_k$ with $\hat{n}_m$ being the cardinality of the sets $\mathcal{S}_m$'s. Δ^{kernel} can be estimated as:

$$\Delta_n^{kernel} = \Sigma_n^{kernel} - \Sigma_{pooled,n}^{kernel}. \tag{4}$$

Let $\hat{v}_j$ and v_j denote the eigenfunctions associated with the eigenvalues $\hat{\lambda}_j$ and λ_j of Δ_n^{kernel} and Δ^{kernel}, respectively. Define $\hat{P}_k$ and P_k as the projection operators onto the subspaces spanned by the kth eigenfunctions of Δ_n^{kernel} and Δ^{kernel}, for $k = 1, \ldots, q_d$. In this paper, we assume that the nonzero eigenvalues of Δ_n^{kernel} and Δ^{kernel} are distinct. Thus, we have $\hat{P}_k = \hat{v}_k \otimes \hat{v}_k$ and $P_k = v_k \otimes v_k$. This assumption is common, as referenced in Hsing and Eubank (2015) and Li and Song (2017). Let $||\cdot||_{HS}$ denote the Hilbert-Schmidt norm. The following theorem provides corresponding theoretical guarantees for the estimator of Δ^{kernel}.

Theorem 6 *Under Assumptions 1–4 in Appendix,* Δ^{kernel} *is a Hilbert-Schmidt operator. Furthermore, if Assumption 5 in Appendix holds and* $\beta_n = O(n^m)$*, we have*

$$||\Delta_n^{kernel} - \Delta^{kernel}||_{HS} = O_p\left(n^{(3\gamma-1)/2} + n^{\gamma-m} + n^{2\gamma+m-1}\right),$$

and

$$||\hat{P}_k - P_k||_{HS} = O_p\left(n^{(3\gamma-1)/2} + n^{\gamma-m} + n^{2\gamma+m-1}\right).$$

Specially, when $m = 1/2$ *and* $\gamma = 0$*, we have*

$$||\Delta_n^{kernel} - \Delta^{kernel}||_{HS} = O_p\left(n^{-1/2}\right) \text{ and } ||\hat{P}_k - P_k||_{HS} = O_p\left(n^{-1/2}\right).$$

The proof of Theorem 6 is given in Appendix F.5. The following theorem states the limiting distribution of $\langle \alpha, (\Delta_n^{kernel} - \Delta^{kernel})\alpha \rangle_{\mathcal{H}}$ for any fixed $\alpha \in \mathcal{H}$ with $||\alpha||_{\mathcal{H}} = 1$. Let $Y^{(i)} = \phi\left(X^{(i)}\right) = \mu_d^{(i)} + \epsilon_d^{(i)}$ for $i = 1, \ldots, s+1$, where random variable $X^{(i)}$ follows the unknown distribution $P^{(i)}$.

Theorem 7 *Suppose Assumptions 1–8 in Appendix hold. Given that $0 < m < 1/2$ and $\gamma = 0$, for any $\alpha \in \mathcal{H}$ satisfying $||\alpha||_{\mathcal{H}} = 1$,*

$$\sqrt{n}\langle \alpha, (\Delta_n^{kernel} - \Delta^{kernel})\alpha \rangle_{\mathcal{H}} \xrightarrow{\mathcal{D}} \mathcal{N}\left(0, \sum_{i=1}^{s+1} 4c_i \mathrm{Var}(\langle \alpha, ((\mu_d^{(i)} - \bar{E}Y) \otimes \epsilon_d^{(i)})\alpha \rangle_{\mathcal{H}})\right),$$

where $\bar{E}Y = \sum_{l=1}^{s+1} c_l \mu_d^{(l)}$. This result leads to

$$\sqrt{n}(\hat{\lambda}_j - \lambda_j) \xrightarrow{\mathcal{D}} \langle v_j, \tilde{\Delta} v_j \rangle_{\mathcal{H}} \text{ and } \sqrt{n}(\hat{v}_j - v_j) \xrightarrow{\mathcal{D}} \sum_{\lambda_k \neq \lambda_j} \frac{1}{\lambda_j - \lambda_k} P_k \tilde{\Delta} v_j, \text{ for } j = 1, \ldots, q_d,$$

where $\langle \alpha, \tilde{\Delta}\alpha \rangle_{\mathcal{H}}$ follows the distribution $\mathcal{N}\left(0, \sum_{i=1}^{s+1} 4c_i \mathrm{Var}(\langle \alpha, ((\mu_d^{(i)} - \bar{E}Y) \otimes \epsilon_d^{(i)})\alpha \rangle_{\mathcal{H}})\right)$.

Remark 8 *Note that the estimator is asymptotically unbiased, and thus, selecting β_n is intrinsically different from selecting a bandwidth in nonparametric estimation that can have an optimal selection when balancing between the bias and variance. If β_n is too small, the estimated covariance for each segment may not be accurate enough, resulting in a lossy estimator of the pooled covariance matrix Σ_{pooled}^{kernel}. Conversely, if β_n is too large and a long segment may contain multiple distributions, the estimator could fail to identify the changes accurately. Based on our analysis, a theoretical optimal parameter β_n cannot be determined, as explained in Appendix. We also conduct simulation experiments to examine the effects of different values of β_n, with results presented in Appendix. Theorem 6 states that when $\beta_n = O(n^m)$ with $0 \leq m \leq 1/2$ (including the case where the sample size β_n is finite), we can ensure $||\hat{P}_k - P_k||_{HS} = O_p\left(n^{-1/2}\right)$. To strike a balance, numerical studies in the later section support choosing $\beta_n = \lfloor \sqrt{n} \rfloor$.*

The proof of Theorem 7 is given in Appendix F.6. According to (4), the eigenvalue problem must be solved to obtain the appropriate lower-dimensional data $\{f(X_i)\}_{i=1}^n$. However, since we lack knowledge about the explicit solution of Y_i, where $i = 1, \ldots, n$, we employ a kernel trick to compute the eigenfunctions of Δ_n^{kernel}. In this approach, we introduce the kernel matrix K, where $K_{ij} = \langle Y_i, Y_j \rangle_{\mathcal{H}} = K(X_i, X_j)$. Here, $K(\cdot, \cdot)$ represents a kernel function. Several kernel functions, including Gaussian, Laplace, and exponential kernels, satisfy the assumptions required for the kernel function in Theorem 6. Then, Δ_n^{kernel} can be expressed as

$$\begin{aligned}
\Sigma_n^{kernel} - \Sigma_{pooled,n}^{kernel} &= \frac{1}{n}\sum_{i=1}^{n}(Y_i - \bar{Y}) \otimes (Y_i - \bar{Y}) - \frac{1}{r}\sum_{m=1}^{r}\hat{\Sigma}_m \\
&= \frac{1}{n}\sum_{i=1}^{n}(Y_i - \bar{Y}) \otimes (Y_i - \bar{Y}) - \frac{1}{r}\sum_{m=1}^{r}\frac{1}{\hat{n}_m - 1}\sum_{i \in \mathcal{S}_m}(Y_i - \bar{Y}_m) \otimes (Y_i - \bar{Y}_m).
\end{aligned}$$

Recall that $\hat{v}_i$ and $\hat{\lambda}_i$ represent the eigenfunction and eigenvalue of Δ_n^{kernel}, respectively. Thus, for any $\hat{v}_i$ and $\hat{\lambda}_i \neq 0$,

$$\hat{\lambda}_i \hat{v}_i = \Delta_n^{kernel} \hat{v}_i = \frac{1}{n}\sum_{i=1}^{n}(Y_i - \bar{Y})\langle Y_i - \bar{Y}, \hat{v}_i\rangle_{\mathcal{H}} - \frac{1}{r}\sum_{m=1}^{r}\frac{1}{\hat{n}_m - 1}\sum_{i\in\mathcal{S}_m}(Y_i - \bar{Y}_m)\langle Y_i - \bar{Y}_m, \hat{v}_i\rangle_{\mathcal{H}},$$

and $\hat{v}_i$ has a linear expression of Y for any given $Y = (Y_1, \ldots, Y_n)$ writing it as $\hat{v}_i = Y\alpha_i$ in form. Define $G_i = \begin{bmatrix} 0_{i\beta_n \times \beta_n} \\ I_{\beta_n \times \beta_n} \\ 0_{(n-(i+1)\beta_n)\times\beta_n} \end{bmatrix}$ and $H_i = \begin{bmatrix} 0_{i\beta_n \times \beta_n} \\ 1_{\beta_n \times \beta_n} \\ 0_{(n-(i+1)\beta_n)\times\beta_n} \end{bmatrix}$, where $I_{p\times p}$ represents the identity matrix, while $1_{p\times p}$ denotes the p-dimensional matrix with all elements equal to 1. We can solve the eigen-decomposition problem by substituting the kernel matrix into the following inference:

$$\begin{aligned}\Delta_n^{kernel}\hat{v}_i = \hat{\lambda}_i\hat{v}_i \Rightarrow &\left\{\frac{1}{n}\sum_{i=1}^{n}\bar{K}_i^{\top}\bar{K}_i - \frac{1}{r}\sum_{m=1}^{r}\frac{1}{\hat{n}_m - 1}\sum_{i\in\mathcal{S}_m}(\bar{K}_i^{(m)})^{\top}\bar{K}_i^{(m)}\right\}\alpha_i = \hat{\lambda}_i K\alpha_i \\ &\Rightarrow (KLK - KUK)\alpha_i = \hat{\lambda}_i K\alpha_i \;\Rightarrow\; K_n\alpha_i = \hat{\lambda}_i\alpha_i,\end{aligned} \tag{5}$$

with $K_n = (L - U)K$, where

$$\bar{K}_i = \left(K_{i1} - \frac{1}{n}\sum_{j=1}^{n}K_{j1}, \ldots, K_{in} - \frac{1}{n}\sum_{j=1}^{n}K_{jn}\right),\ \bar{K}_i^{(m)} = \left(K_{i1} - \frac{1}{\hat{n}_m}\sum_{j\in\mathcal{S}_m}K_{j1}, \ldots, K_{in} - \frac{1}{\hat{n}_m}\sum_{j\in\mathcal{S}_m}K_{jn}\right),$$

$$L = \frac{1}{n}(I_{n\times n} - \frac{1}{n}1_{n\times n})(I_{n\times n} - \frac{1}{n}1_{n\times n})^{\top}, \quad U = \sum_{i=1}^{r}\frac{1}{r(n_i - 1)}(G_i - \frac{1}{\beta_n}H_i)(G_i - \frac{1}{\beta_n}H_i)^{\top}.$$

Here, we only consider that K is a positive definite kernel matrix. Based on the deduction in (5), we have that α_i is an eigenvector of K_n. For any $i = 1, 2, \ldots, n$,

$$f(X_i) = \left(\langle v_1, Y_i\rangle_{\mathcal{H}}, \ldots, \langle v_{q_d}, Y_i\rangle_{\mathcal{H}}\right)^{\top} = (K_i\alpha_1, \ldots, K_i\alpha_{\hat{q}_d})^{\top} = B_n^{\top}K_i^{\top},$$

where $K_i = (K_{i1}, K_{i2}, \ldots, K_{in})$ and $B_n = (\alpha_1, \alpha_2, \ldots, \alpha_{q_d})$. Therefore, the data after dimension reduction is given by $f(X) = (f_1(X), \ldots, f_n(X)) = B_n^{\top}K$. It can be observed that Δ_n^{kernel} and K_n share the same non-zero eigenvalues. To estimate the structural dimension q_d when it is unknown, we employ the thresholding ridge ratio criterion (TRR) as follows:

$$\hat{q}_d := \max_{1\leq k\leq n-1}\left\{k:\ \hat{r}_k = \frac{\hat{\lambda}_{k+1} + c_n}{\hat{\lambda}_k + c_n} \leq \tau\right\}, \tag{6}$$

where $\hat{\lambda}_1 \geq \cdots \geq \hat{\lambda}_n$ are the eigenvalues of the estimated target matrix K_n, c_n is a ridge value approaching zero at a certain rate, and τ is a thresholding value such that $0 < \tau < 1$. Choosing $\tau = 0.5$ is reasonable according to the plug-in principle in Zhu et al. (2020) to avoid overestimation with a large τ and underestimation with a small τ. As the target matrix differs from Zhu et al. (2020), there is no optimal criterion or theoretical result for selection. In this paper, we recommend setting the ridge value to $c_n = 0.2\log(\log(n))\sqrt{1/n}$ for practical purposes. The consistency of $\hat{q}_d$ is stated in the following theorem.

Theorem 9 *Let* $\tilde{\eta}_n = \max\left\{\sqrt{\frac{1}{n}}, \frac{\beta_n}{n}\right\}$. *Under the same conditions in Theorem 6, if* $c_n \to 0$, $\tilde{\eta}_n \to 0$, $c_n/\tilde{\eta}_n \to \infty$ *as* $n \to \infty$, *then* $P(\hat{q}_d = q_d) \to 1$.

The proof of Theorem 9 is given in Appendix F.7. In the specific scenario where $\phi(X_t) = X_t$, KPCA simplifies to PCA. The corresponding Corrected PCA (CPCA) for mean changes is presented in Appendix.

3. Iterative CKPCA in Cluster Analysis

In this section, we extend CKPCA to cluster analysis, aiming to achieve a superior nonlinear low-dimensional embedding. Although the approach can be applied to any clustering algorithm, we use the clustering results from the K-means algorithm as the initial values for the iterative algorithm to demonstrate its effectiveness.

Consider a data set $\{X_i\}_{i=1}^n$ where $X_i = (X_{i1}, \ldots, X_{ip})^\top \in \mathbb{R}^p$ are independent. The data points belong to a union of d categories denoted by $\{\mathcal{C}_k\}_{k=1}^d$, with each category $\mathcal{C}_k$ containing n_k points such that $\sum_{k=1}^d n_k = n$. Assume $n_j/n \to \omega_j$ as $n \to \infty$ be the weight of category $\mathcal{C}_j$. We apply the nonlinear feature map $X_i \to \phi(X_i) = Y_i$, $i = 1, \ldots, n$. Based on Celisse et al. (2018), there exists a ϕ such that if $X_i \in \mathcal{C}_k$ and $X_j \in \mathcal{C}_k$ for $k = 1, \ldots, d$, then $E(Y_i) = E(Y_j)$ holds. For all $Y_j \in \mathcal{C}_k$, let $E(Y_j) = \mu_d^{(k)}$ and $\Sigma_d^{(k)} = E(Y_j - \mu_d^{(k)}) \otimes (Y_j - \mu_d^{(k)})$, for $k = 1, \ldots, d$.

Consistently with the inference in Section 2, define the kernel mean embedding deviation subspace in cluster analysis as $\mathrm{S}^d_{\{X_i\}_{i=1}^n} = \mathrm{Span}\{\mu_d^{(i)} - \mu_d^{(j)},\ \text{for } i, j = 1, \ldots, d\}$ with the structural dimension $q_d = \dim\{\mathrm{S}^d_{\{X_i\}_{i=1}^n}\}$. Consider the "sample covariance operator" Σ_n^{kernel}, and its expectation is, similar to (3),

$$\begin{aligned} E(\Sigma_n^{kernel}) &\to \sum_{j=1}^d \omega_j \Sigma_d^{(j)} + \frac{1}{2}\sum_{i=1}^d \sum_{j=1}^d \omega_i \omega_j (\mu_d^{(i)} - \mu_d^{(j)}) \otimes (\mu_d^{(i)} - \mu_d^{(j)}) \\ &=: \Sigma_{pooled}^{kernel} + \Delta^{kernel} = \Sigma^{kernel}. \end{aligned}$$

Proposition 10 *Under Assumptions 2 and 3 in Appendix, for any basis functions* $\{v_1, v_2, \ldots, v_{q_d}\}$ *of* $\mathrm{S}^d_{\{X_i\}_{i=1}^n}$ *with* $q_d \le d$, *let* $f(X_i) = (\langle v_1, Y_i\rangle_{\mathcal{H}}, \ldots, \langle v_{q_d}, Y_i\rangle_{\mathcal{H}})^\top$. *Both the sequences* $\{f(X_i)\}_{i=1}^n$ *and* $\{X_i\}_{i=1}^n$ *have the same clustering results.*

Remark 11 *The sequences* $\{f(X_i)\}_{i=1}^n$ *and* $\{X_i\}_{i=1}^n$ *having the same clustering results means that their partitions are identical, although the cluster labels may differ. In this case, a relabeling function* g *can be defined to map the cluster labels of* $\{f(X_i)\}_{i=1}^n$ *to those of* $\{X_i\}_{i=1}^n$, *ensuring that the two partitions are equivalent. Specifically, there exists a function* $g : \{1, 2, \ldots, d\} \to \{1, 2, \ldots, d\}$ *that reassigns the cluster labels of* $\{f(X_i)\}_{i=1}^n$ *to match those of* $\{X_i\}_{i=1}^n$. *For every cluster* $i \in \{1, 2, \ldots, d\}$, *the following holds:* $\mathcal{C}_{1,i} = \mathcal{C}_{2,g(i)}$, *where* $\mathcal{C}_{1,i}$ *represents the* i*-th cluster of the sequence* $\{X_i\}_{i=1}^n$, *and* $\mathcal{C}_{2,g(i)}$ *represents the* $g(i)$*-th cluster of the sequence* $\{f(X_i)\}_{i=1}^n$ *after relabeling.*

Σ_{pooled}^{kernel} and Δ^{kernel} can be estimated respectively as follows:

$$\begin{aligned}\Sigma_{pooled,n}^{kernel} &= \sum_{i=1}^{d} \frac{n_i - 1}{n - d} \hat{\Sigma}_i \text{ with } \hat{\Sigma}_i = \frac{1}{n_i - 1} \sum_{j \in \mathcal{C}_i} (Y_j - \bar{Y}_i) \otimes (Y_j - \bar{Y}_i), \\ \Delta_n^{kernel} &= \Sigma_n^{kernel} - \Sigma_{pooled,n}^{kernel} = \frac{1}{n} \sum_{i=1}^{n} (Y_i - \bar{Y}) \otimes (Y_i - \bar{Y}) - \sum_{i=1}^{d} \frac{n_i - 1}{n - d} \hat{\Sigma}_i, \end{aligned} \tag{7}$$

where $\bar{Y}_i = \frac{1}{n_i} \sum_{j \in \mathcal{C}_i} Y_j$. Since the categories $\mathcal{C}_j$ are unknown, we first obtain an initial value by applying a popular clustering method before CKPCA. We then propose an iterative approach, which will be described in detail later.

Given the categories $\mathcal{C}_i$, for each $X_j \in \mathcal{C}_i$, we define G_i as an $n \times n_k$ matrix with a 1 at the $(\mathcal{C}_i(j), j)$th element and zeros elsewhere. Similarly, we define H_i as an $n \times n_k$ matrix with a row of 1s at the $\mathcal{C}_i(j)$th position and zeros elsewhere, where $j = 1, \ldots, n_i$ and $k = 1, 2, \ldots, d$. Following the approach in Section 2, let $\hat{v}_i$ denote an eigenfunction of Δ_n^{kernel}, and we can also express $\hat{v}_i$ as $\hat{v}_i = Y\alpha_i$ for some α_i. Almost identical to (5), for any $\hat{v}_i$ and $\lambda_i \neq 0$, we have:

$$\Delta_n^{kernel} \hat{v}_i = \lambda_i \hat{v}_i \;\Rightarrow\; K_n \alpha_i = \lambda_i \alpha_i, \tag{8}$$

where $K_n = (R - S)K$ and α_i is an eigenvector of K_n. Here,

$$\bar{K}_i = \left(K_{i1} - \frac{1}{n} \sum_{j=1}^{n} K_{j1}, \ldots, K_{in} - \frac{1}{n} \sum_{j=1}^{n} K_{jn} \right), \; \bar{K}_i^{(m)} = \left(K_{i1} - \frac{1}{\hat{n}_m} \sum_{j \in \mathcal{C}_m} K_{j1}, \ldots, K_{in} - \frac{1}{\hat{n}_m} \sum_{j \in \mathcal{C}_m} K_{jn} \right),$$

$$R = \frac{1}{n} \left(I_{n \times n} - \frac{1}{n} 1_{n \times n} \right) \left(I_{n \times n} - \frac{1}{n} 1_{n \times n} \right)^{\top}, \; S = \sum_{i=1}^{d} \frac{1}{n - d} \left(G_i - \frac{1}{n_i} H_i \right) \left(G_i - \frac{1}{n_i} H_i \right)^{\top}.$$

Moreover, the lowered dimensional data is $f(X) = (f(X_1), \ldots, f(X_n)) = B_n^{\top} K$ and $B_n = (\alpha_1, \ldots, \alpha_{\hat{q}_d})$. In order to apply (7), it is important to have information about the categories $\{\mathcal{C}_k\}_{k=1}^{d}$.

However, since information about the categories $\mathcal{C}_j$ is lacking, we need to obtain an initial value by using some popular clustering method. Therefore, we propose an iterative algorithm as follows. First, apply a popular clustering method to the original data X to obtain the initial categories $\{\hat{\mathcal{C}}_i\}_{i=1}^{d}$, with the pre-defined number of categories d. Second, apply (7) and (8) using the results $\{\hat{\mathcal{C}}_i\}_{i=1}^{d}$ to obtain the lowered dimensional data $f(X) = B_n^{\top} K$, where $B_n = (\alpha_1, \ldots, \alpha_{\hat{q}_d})$ represents the eigenvectors associated with the largest $\hat{q}_d$ eigenvalues of the kernel matrix $K_n = (R - S)K$. Finally, the dimension reduction and clustering steps are iteratively applied to the lower-dimensional data until a stopping criterion is satisfied. The Rand Index (RI) (Rand, 1971), which measures the similarity between two adjacent clustering results, is used as the stopping criterion. The iterative algorithm is summarized in Algorithm 1.

We study the theoretical properties of this method, specifically focusing on linear clustering, where step 1 of Algorithm 1 is the K-means algorithm. Using the K-means algorithm stems from the requirement of initial labels, as we currently lack a method that ensures consistent clustering outcomes within the framework of nonlinear dimension reduction. Therefore, we consider the situation where $\phi(X_i) = X_i$ and make reference to the settings and

Algorithm 1 Iterative Subspace Cluster Algorithm.

Require: $X \in \mathbb{R}^{n\times p}$, $\tau = 0.5$, and $c_n = 0.2\log(\log(n))\sqrt{1/n}$.

1: Choose a classical clustering algorithm such as K-means to cluster the original data, then get $\hat{\mathcal{C}}$;
2: Update the target matrix $K_n = (R-S)K$ in (7) and make the eigen-decomposition to get the eigenvalues $\lambda_1 \geq \cdots \geq \lambda_n$ and the corresponding eigenvectors $\alpha_1, \ldots, \alpha_n$;
3: Determine the dimension $\hat{q}_d$ based on TRR as (6) and obtain the basis matrix $B_n = (\alpha_1, \ldots, \alpha_{\hat{q}_d})$, then get the lowered data to be $B_n^\top K$;
4: Repeat step 1 to the lowered dimensional data, then calculate the RI between the clustering result and the previous clustering result;
5: Repeat steps 2-4 until RI exceeds 0.999.

Ensure: $\{\hat{\mathcal{C}}_1, \ldots, \hat{\mathcal{C}}_d\}$.

theories pertaining to the K-means algorithm as discussed in Lu and Zhou (2016). Let X_i represent independent samples drawn from a sub-Gaussian mixture model:

$$X_i = \mu_i + W_i, \quad i = 1, 2, \ldots, n, \tag{9}$$

where W_i are independent sub-Gaussian random vectors with a sub-Gaussian parameter of σ, the location μ_i corresponds to μ_{z_i*}, and $z_i \in [d]$ denotes the cluster label of the ith sample. Redefine $X^{(i)}$ represent the data belonging to the ith class, which satisfies $X^{(i)} = \mu^{(i)} + W^{(i)}$ for $i = 1, 2, \ldots, d$. Consequently, Δ^{kernel} is simplified to $\Delta = \frac{1}{2}\sum_{i=1}^{d}\sum_{j=1}^{d}\omega_i\omega_j(\mu^{(i)} - \mu^{(j)})(\mu^{(i)} - \mu^{(j)})^\top$. The proof of Theorem 12 is given in Appendix F.9.

Theorem 12 *Consider the model (9) and assume that X_i are independent p-dimensional random vectors. Additionally, suppose that Assumptions 13–16 in Appendix hold. Then*

$$\sqrt{n}\alpha^\top(\Delta_n - \Delta)\alpha \xrightarrow{\mathcal{D}} \mathcal{N}\left(0, \sum_{i=1}^{d} 4\omega_i \mathrm{Var}(\alpha^\top(\mu^{(i)} - \bar{E}X)(W^{(i)})^\top\alpha)\right),$$

and when q_d is given,

$$||\Delta_n - \Delta||_F = O_p\left(\sqrt{\frac{p}{n}}\right) \quad \text{and} \quad ||B_n - B||_F = O_p\left(\sqrt{\frac{p}{n}}\right),$$

where $\|\cdot\|_F$ represents the Frobenius norm of a matrix.

4. Numerical Experiments

In this section, we perform various simulation experiments and analyze several real data sets to showcase the effectiveness of the proposed method. The numerical results pertaining to mean changes, as well as the Macroeconomic data, are provided in Appendix in order to conserve space in the main text.

4.1 Simulations on Change Point Detection

The corrected (kernel) PCA improves popular change point methods after dimension reduction. We demonstrate its effect using four popular change point detection methods: the energy-based method (Matteson and James, 2014), the sparsified binary segmentation method (Cho and Fryzlewicz, 2015), the kernel change point algorithm (Arlot et al., 2019), and the change point detection tests using rank statistics (Lung-Yut-Fong et al., 2015), referred to as E-Divisive, SBS, KCP, and Multirank, respectively. To compare the performance of CKPCA with other dimension reduction technologies, we compare three dimension reduction methods: CKPCA, KPCA and the corrected Mahalanobis matrix method proposed by Zhu et al. (2025). For simplicity, we denote the versions of E-Divisive based on the dimension reduction methods as E-Divisive$_C$, E-Divisive$_P$, and E-Divisive$_M$, respectively. Although SBS is applicable to multivariate data, when $\hat{q}_d = 1$, SBS automatically reduces to wild binary segmentation method (WBS) in Fryzlewicz (2014). We also compare CKPCA with three popular high-dimensional methods: the informative sparse projection for estimation of change points (Wang and Samworth, 2018), the double CUSUM statistic method (Cho, 2016), and the method via a geometrically inspired mapping (Grundy et al., 2020), referred to as Inspect, DCBS, and GeomCP, respectively. Since Inspect, DCBS, and GeomCP are applied in high-dimensional scenarios, we did not report their results after dimension reduction. Denoting the estimated number of change points as $\hat{s}$, we evaluate the performance of the estimated change points by measuring the average of $\hat{s}$, the root-mean-square error (RMSE) of $\hat{s}$, and the Rand Index (RI) (Rand, 1971) between the estimated and real segments. Furthermore, we constructed the 95% percentile interval, denoted by PI, using the 2.5% and 97.5% quantiles of the RI. The E-Divisive method is implemented in the R package "ecp", whereas SBS, WBS, and Inspect are implemented in the R packages "hdbinseg", "wbs", and "InspectChangepoint", respectively. The Python code for the Multirank method is provided by the authors of Lung-Yut-Fong et al. (2015), and the Python implementation of KCP is available in the "ruptures" package.

The experiment is repeated 1000 times. The TRR method is used to select $\hat{q}_d$ for CKPCA and the Mahalanobis matrix, with parameters $\tau = 0.5$ and $c_n = 0.2\log(\log(n))\sqrt{1/n}$. For CPCA, we set $c_n = 0.2\log(\log(n))\sqrt{p/n}$. The cumulative variance contribution rate method is employed to select $\hat{q}_d$ for KPCA, with a cumulative variance contribution rate of 0.95. We choose the Gaussian kernel function given by $K(X, X') = \exp\left\{-\|X - X'\|^2/\left(2h^2\right)\right\}$, where h represents the bandwidth. Inspired by the idea of Varon et al. (2015), we select the bandwidth h as $h^2 = mp\left[E\left\{\mathrm{Var}\left(X\right)\right\}\right]$ where $E\left\{\mathrm{Var}\left(X\right)\right\} = \frac{1}{p}\sum_{i=1}^{p}\mathrm{Var}\left(X_i\right)$, $\mathrm{Var}\left(X_i\right)$ denotes the variance of one dimension in the data set, and m denotes a tuning parameter. We recommend $m = 0.8$. More details about the sensitivity of m can be found later.

We evaluate the methods through three scenarios: (1) changes in both distribution and covariance matrix, (2) changes in distribution, and (3) changes in mean. **Examples 1-3** correspond to these scenarios, respectively. **Example 3** which pertains to changes in mean is provided in Appendix to economize space in the main context. We apply PCA and CPCA to detect mean changes, and KPCA and CKPCA to detect distributional changes. Since SBS, WBS, Inspect, DCBS, and GeomCP are designed to detect mean changes, we only consider E-Divisive, KCP, and Multirank in **Examples 1** and **2**. Note that, based on the inference presented in Section 2, CKPCA can transform distributional changes

into mean changes. Therefore, we also explore the performance of SBS after applying the three dimension reduction versions. The sample size is set to be $n = 800$. The data are divided into eight parts, each following a distribution denoted by $\{G_i\}_{i=1}^{8}$. Therefore, the total number of change points is $s = 7$. We conduct experiments on both balanced and imbalanced data sets:

- **Balanced data set.** The change points are located at $100i$ for $i = 1, 2, 3, \ldots, 7$, respectively;
- **Imbalanced data set.** The change points are located at 30, 170, 350, 440, 520, 630, 710, respectively.

Example 1: Changes in both distribution and covariance matrix. The data are generated in the following settings:

- Case 1: $G_1 = G_3 = G_5 = G_7 = N(0_p, \Sigma)$, where $\Sigma = (1.5\mathrm{I}_{p\times p} + \sigma_{ij})$ and $\sigma_{ij} = I(i = j) + bI(i \neq j)$ with $b = 0.5$, and $G_2 = G_4 = G_6 = G_8$ are the p-dimensional uniform distributions on the regions $[-3, 3] \times [-3, 3] \times \cdots \times [-3, 3]$.
- Case 2: The settings of G_i are the same as Case 1, except that $\Sigma = (1.5\mathrm{I}_{p\times p} + \sigma_{ij})$ and $\sigma_{ij} = b^{|i-j|}$ with $b = 0.5$.

In this example, we consider a dimension of $p = 100$ and $p = 200$, with $q_d = 1$. The results are shown in Tables 1 and 2. In Case 1, Multirank$_C$ performs the best, with $\hat{s}$ being close to the true value of 7 and the RI exceeding 0.99. Among the SBS methods, SBS$_C$ outperforms both SBS$_P$ and SBS$_M$. All three versions of dimension reduction demonstrate significant improvements for E-Divisive, with CKPCA exhibiting the most substantial enhancement. The results of Case 2 are similar to Case 1, with Multirank$_C$ still being the top performer. KCP and E-Divisive are less effective, but KCP$_C$ and E-Divisive$_C$ still yield good results. The RI of Multirank is lower than that of Multirank$_C$, and SBS$_C$ outperforms both SBS$_P$ and SBS$_M$.

To assess the sensitivity of the methods to outliers, we introduce imbalanced data with 5% outliers from $G_i + W_i$ between each z_i and z_{i+1}. Here, W_i represents a p-dimensional constant vector. For each i, we randomly select 5% of its elements to take the value 5, while the other elements are set to 0. The results presented in Table 3 demonstrate that the dimension reduction-based methods are relatively robust against imbalanced data and data with outliers.

Example 2: Changes in distribution. The data are generated in the following settings:

- $G_1 = G_3 = G_5 = G_7 = N(0_p, \Sigma)$ and $G_2 = G_4 = G_6 = G_8 = t(df, \Sigma)$ is the p-dimensional t-distribution with the degree $df = a$ and $\Sigma = (\sigma_{ij})$, where $\sigma_{ij} = 0.5^{|i-j|}$, $a = 4, 6$, $p = 100, 200$.

In **Example 2**, the distributions change while the mean and covariance remain constant. We consider different values of $a = 4$ and $a = 6$, representing strong and weak signals, respectively. The results of **Example 2** are presented in Tables 4 and 5. When $a = 6$, E-Divisive and KCP are almost ineffective, while E-Divisive$_C$ and KCP$_C$ still perform well.

Case	p	Method	$\hat{s}$	RMSE	RI	PI	p	Method	$\hat{s}$	RMSE	RI	PI
		E-Divisive$_C$	7.520	0.949	0.992	[0.979, 0.996]		E-Divisive$_C$	7.500	0.908	0.991	[0.978, 0.997]
		E-Divisive$_P$	4.890	2.770	0.830	[0.343, 0.994]		E-Divisive$_P$	2.584	4.812	0.549	[0.124, 0.921]
		E-Divisive$_M$	8.819	2.644	0.915	[0.806, 0.984]		E-Divisive$_M$	6.958	2.038	0.859	[0.666, 0.958]
		E-Divisive	0.769	6.345	0.274	[0.124, 0.781]		E-Divisive	0.688	6.412	0.262	[0.124, 0.759]
		Multirank$_C$	7.002	0.063	0.995	[0.993, 0.996]		Multirank$_C$	7.000	0.000	0.995	[0.991, 0.997]
		Multirank$_P$	0.000	7.000	0.124	[0.124, 0.124]		Multirank$_P$	0.000	7.000	0.124	[0.124, 0.124]
		Multirank$_M$	0.000	7.000	0.124	[0.124, 0.124]		Multirank$_M$	0.000	7.000	0.124	[0.124, 0.124]
1	200	Multirank	0.408	6.808	0.142	[0.124, 0.333]	100	Multirank	0.337	6.811	0.143	[0.124, 0.508]
		KCP$_C$	7.110	0.358	0.998	[0.990, 1.000]		KCP$_C$	7.210	0.500	0.996	[0.986, 1.000]
		KCP$_P$	6.721	1.201	0.968	[0.528, 0.999]		KCP$_P$	1.073	6.198	0.302	[0.124, 0.995]
		KCP$_M$	0.287	6.729	0.147	[0.124, 0.226]		KCP$_M$	0.095	6.913	0.133	[0.124, 0.198]
		KCP	0.000	7.000	0.124	[0.124, 0.124]		KCP	0.000	7.000	0.124	[0.124, 0.124]
		SBS$_C$	8.828	2.607	0.992	[0.973, 1.000]		SBS$_C$	7.825	1.405	0.994	[0.979, 1.000]
		SBS$_P$	0.023	6.979	0.131	[0.124, 0.124]		SBS$_P$	0.047	6.956	0.135	[0.124, 0.325]
		SBS$_M$	6.502	3.270	0.779	[0.301, 0.930]		SBS$_M$	4.883	3.339	0.722	[0.124, 0.914]
		E-Divisive$_C$	9.303	2.762	0.973	[0.952, 0.991]		E-Divisive$_C$	8.861	2.316	0.972	[0.949, 0.991]
		E-Divisive$_P$	4.526	3.495	0.759	[0.124, 0.991]		E-Divisive$_P$	0.644	6.453	0.259	[0.124, 0.786]
		E-Divisive$_M$	10.425	4.043	0.901	[0.811, 0.937]		E-Divisive$_M$	7.721	2.190	0.865	[0.707, 0.928]
		E-Divisive	0.320	6.720	0.198	[0.124, 0.670]		E-Divisive	0.188	6.832	0.172	[0.124, 0.609]
		Multirank$_C$	7.000	0.000	0.987	[0.978, 0.995]		Multirank$_C$	6.993	0.164	0.983	[0.965, 0.994]
		Multirank$_P$	0.000	7.000	0.124	[0.124, 0.124]		Multirank$_P$	0.000	7.000	0.124	[0.124, 0.124]
		Multirank$_M$	0.000	7.000	0.124	[0.124, 0.124]		Multirank$_M$	0.000	7.000	0.124	[0.124, 0.124]
2	200	Multirank	0.081	6.952	0.128	[0.124, 0.124]	100	Multirank	0.131	6.923	0.133	[0.124, 0.213]
		KCP$_C$	8.757	2.143	0.978	[0.959, 0.994]		KCP$_C$	8.577	1.999	0.975	[0.954, 0.992]
		KCP$_P$	0.000	7.000	0.124	[0.124, 0.124]		KCP$_P$	0.000	7.000	0.124	[0.124, 0.124]
		KCP$_M$	0.301	6.716	0.148	[0.124, 0.232]		KCP$_M$	0.119	6.889	0.136	[0.124, 0.215]
		KCP	0.000	7.000	0.124	[0.124, 0.124]		KCP	0.000	7.000	0.124	[0.124, 0.124]
		SBS$_C$	10.473	4.121	0.975	[0.952, 0.993]		SBS$_C$	8.906	2.598	0.976	[0.955, 0.993]
		SBS$_P$	0.027	6.975	0.132	[0.124, 0.266]		SBS$_P$	0.025	6.977	0.132	[0.124, 0.127]
		SBS$_M$	9.672	3.765	0.883	[0.772, 0.933]		SBS$_M$	6.297	2.528	0.812	[0.537, 0.918]

Table 1: Changes in both distribution and covariance matrix in Example 1 with balanced data set

Additionally, Multirank$_C$ outperforms Multirank, and SBS$_C$ outperforms both SBS$_P$ and SBS$_M$. The results when $a = 4$ are similar to those when $a = 6$. All three methods show improvement after applying CKPCA. **Example 3** can be found in Appendix for the purpose of conserving space. Overall, **Examples 1-3** demonstrate that CPCA/CKPCA can enhance the performance of popular change point detection methods, surpassing PCA/KPCA. Furthermore, based on the results from imbalanced cases, outliers, and distinct distributions, CPCA/CKPCA exhibits greater robustness compared to its competitors.

We test the method's sensitivity to bandwidth selection by considering values of $m = 0.4, 0.8, 1.2, 1.6, 2.0$ for **Example 2**. Table 6 presents the E-Divisive results for different bandwidth values when $a = 4$ and $p = 200$. The results demonstrate robustness across bandwidth variations. In Table 6, the best performance in terms of $\hat{s}$ is achieved when $m = 2$, whereas the RI attains its optimal value at $m = 0.4$. We therefore choose $m = 0.8$ as a compromise value, which yields reasonably good results for both $\hat{s}$ and the RI.

To gain intuitive understanding of CKPCA, we plot scatter plots in Figure 1 for the first vector of the original data, $\{B_{P1n}^\top X_i\}_{i=1}^n$, and $\{B_{C1n}^\top X_i\}_{i=1}^n$. Here, $B_{C1n}^\top$ and $B_{P1n}^\top$ represent the first vector of the data after CKPCA and KPCA, respectively. Figure 1 clearly shows

Case	p	Method	$\hat{s}$	RMSE	RI	PI	p	Method	$\hat{s}$	RMSE	RI	PI
		E-Divisive$_C$	7.583	1.049	0.988	[0.961, 0.996]		E-Divisive$_C$	7.579	1.024	0.988	[0.959, 0.996]
		E-Divisive$_P$	3.079	4.186	0.790	[0.671, 0.972]		E-Divisive$_P$	2.616	4.546	0.744	[0.500, 0.903]
		E-Divisive$_M$	8.626	2.648	0.897	[0.779, 0.974]		E-Divisive$_M$	6.978	2.061	0.854	[0.683, 0.950]
		E-Divisive	1.069	6.073	0.362	[0.146, 0.793]		E-Divisive	0.816	6.300	0.316	[0.146, 0.779]
		Multirank$_C$	7.000	0.000	0.994	[0.990, 0.996]		Multirank$_C$	7.000	0.000	0.994	[0.989, 0.997]
		Multirank$_P$	0.000	7.000	0.146	[0.146, 0.146]		Multirank$_P$	0.000	7.000	0.146	[0.146, 0.146]
		Multirank$_M$	0.000	7.000	0.146	[0.146, 0.146]		Multirank$_M$	0.000	7.000	0.146	[0.146, 0.146]
1	200	Multirank	0.290	6.861	0.160	[0.146, 0.338]	100	Multirank	0.173	6.897	0.157	[0.146, 0.299]
		KCP$_C$	7.119	0.375	0.997	[0.976, 1.000]		KCP$_C$	7.243	0.561	0.995	[0.974, 1.000]
		KCP$_P$	5.252	2.544	0.912	[0.719, 0.999]		KCP$_P$	1.286	5.863	0.476	[0.146, 0.904]
		KCP$_M$	0.311	6.706	0.172	[0.146, 0.285]		KCP$_M$	0.088	6.918	0.156	[0.146, 0.233]
		KCP	0.000	7.000	0.146	[0.146, 0.146]		KCP	0.000	7.000	0.146	[0.146, 0.146]
		SBS$_C$	9.242	3.035	0.985	[0.956, 1.000]		SBS$_C$	8.127	1.787	0.990	[0.961, 1.000]
		SBS$_P$	0.021	6.981	0.152	[0.146, 0.146]		SBS$_P$	0.032	6.971	0.156	[0.146, 0.346]
		SBS$_M$	6.809	3.099	0.798	[0.472, 0.920]		SBS$_M$	5.273	3.243	0.747	[0.209, 0.905]
		E-Divisive$_C$	9.803	3.290	0.961	[0.927, 0.990]		E-Divisive$_C$	9.189	2.727	0.960	[0.924, 0.988]
		E-Divisive$_P$	4.114	3.418	0.832	[0.146, 0.980]		E-Divisive$_P$	0.792	6.313	0.340	[0.146, 0.856]
		E-Divisive$_M$	10.549	4.159	0.886	[0.821, 0.924]		E-Divisive$_M$	7.977	2.375	0.856	[0.712, 0.921]
		E-Divisive	0.378	6.669	0.244	[0.146, 0.757]		E-Divisive	0.186	6.834	0.195	[0.146, 0.647]
		Multirank$_C$	6.922	0.335	0.984	[0.960, 0.995]		Multirank$_C$	6.730	0.642	0.973	[0.899, 0.993]
		Multirank$_P$	0.000	7.000	0.146	[0.146, 0.146]		Multirank$_P$	0.000	7.000	0.146	[0.146, 0.146]
		Multirank$_M$	0.000	7.000	0.146	[0.146, 0.146]		Multirank$_M$	0.000	7.000	0.146	[0.146, 0.146]
2	200	Multirank	0.055	6.970	0.150	[0.146, 0.146]	100	Multirank	0.078	6.955	0.151	[0.146, 0.146]
		KCP$_C$	9.069	2.527	0.970	[0.936, 0.995]		KCP$_C$	8.729	2.172	0.967	[0.932, 0.992]
		KCP$_P$	0.000	7.000	0.146	[0.146, 0.146]		KCP$_P$	0.000	7.000	0.146	[0.146, 0.146]
		KCP$_M$	0.300	6.716	0.172	[0.146, 0.285]		KCP$_M$	0.109	6.898	0.156	[0.146, 0.244]
		KCP	0.000	7.000	0.146	[0.146, 0.146]		KCP	0.000	7.000	0.146	[0.146, 0.146]
		SBS$_C$	11.365	5.014	0.963	[0.929, 0.992]		SBS$_C$	9.203	2.936	0.966	[0.930, 0.992]
		SBS$_P$	0.030	6.972	0.156	[0.146, 0.317]		SBS$_P$	0.027	6.975	0.155	[0.146, 0.277]
		SBS$_M$	9.938	3.969	0.872	[0.773, 0.921]		SBS$_M$	6.628	2.512	0.807	[0.501, 0.915]

Table 2: Changes in both distribution and covariance matrix in Example 1 with imbalanced data set

that the changes at the change points become more pronounced after CKPCA, while KPCA does not facilitate change point detection.

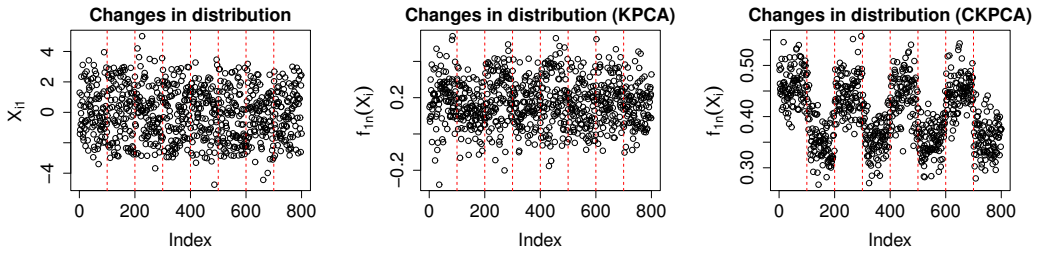

Figure 1: Scatter plots present the first vectors of the original data and the data after KPCA and CKPCA, respectively. The three figures correspond to Case 2 in Example 1 with $p = 200$.

Case	p	Method	$\hat{s}$	RMSE	RI	PI	p	Method	$\hat{s}$	RMSE	RI	PI
		E-Divisive$_C$	7.662	1.126	0.987	[0.956, 0.996]		E-Divisive$_C$	7.655	1.126	0.986	[0.957, 0.996]
		E-Divisive$_P$	3.103	4.173	0.792	[0.671, 0.975]		E-Divisive$_P$	2.586	4.573	0.744	[0.561, 0.903]
		E-Divisive$_M$	8.257	2.450	0.894	[0.759, 0.975]		E-Divisive$_M$	6.756	2.062	0.852	[0.679, 0.949]
		E-Divisive	1.158	5.995	0.379	[0.146, 0.803]		E-Divisive	0.890	6.232	0.333	[0.146, 0.775]
		Multirank$_C$	6.999	0.032	0.994	[0.989, 0.997]		Multirank$_C$	6.992	0.118	0.993	[0.987, 0.997]
		Multirank$_P$	0.000	7.000	0.146	[0.146, 0.146]		Multirank$_P$	0.000	7.000	0.146	[0.146, 0.146]
		Multirank$_M$	0.000	7.000	0.146	[0.146, 0.146]		Multirank$_M$	0.000	7.000	0.146	[0.146, 0.146]
1	200	Multirank	0.335	6.828	0.164	[0.146, 0.479]	100	Multirank	0.270	6.844	0.163	[0.146, 0.541]
		KCP$_C$	7.141	0.432	0.997	[0.974, 1.000]		KCP$_C$	7.274	0.616	0.994	[0.971, 1.000]
		KCP$_P$	4.969	2.796	0.896	[0.715, 0.999]		KCP$_P$	1.150	5.982	0.447	[0.146, 0.900]
		KCP$_M$	0.296	6.721	0.171	[0.146, 0.293]		KCP$_M$	0.103	6.904	0.159	[0.146, 0.285]
		KCP	0.000	7.000	0.146	[0.146, 0.146]		KCP	0.000	7.000	0.146	[0.146, 0.146]
		SBS$_C$	8.601	2.460	0.988	[0.958, 1.000]		SBS$_C$	7.939	1.647	0.990	[0.960, 1.000]
		SBS$_P$	0.034	6.969	0.157	[0.146, 0.365]		SBS$_P$	0.053	6.951	0.162	[0.146, 0.478]
		SBS$_M$	6.654	3.133	0.794	[0.452, 0.917]		SBS$_M$	5.141	3.118	0.754	[0.262, 0.906]
		E-Divisive$_C$	9.904	3.410	0.958	[0.923, 0.989]		E-Divisive$_C$	9.148	2.623	0.959	[0.923, 0.987]
		E-Divisive$_P$	5.304	2.270	0.919	[0.718, 0.987]		E-Divisive$_P$	1.741	5.470	0.537	[0.146, 0.925]
		E-Divisive$_M$	9.880	3.646	0.885	[0.810, 0.924]		E-Divisive$_M$	7.565	2.157	0.855	[0.701, 0.922]
		E-Divisive	1.371	5.772	0.477	[0.146, 0.880]		E-Divisive	0.557	6.504	0.292	[0.146, 0.771]
		Multirank$_C$	6.823	0.532	0.978	[0.908, 0.994]		Multirank$_C$	6.556	0.887	0.963	[0.874, 0.991]
		Multirank$_P$	0.000	7.000	0.146	[0.146, 0.146]		Multirank$_P$	0.000	7.000	0.146	[0.146, 0.146]
		Multirank$_M$	0.000	7.000	0.146	[0.146, 0.146]		Multirank$_M$	0.000	7.000	0.146	[0.146, 0.146]
2	200	Multirank	0.079	6.957	0.150	[0.146, 0.146]	100	Multirank	0.119	6.934	0.154	[0.146, 0.148]
		KCP$_C$	9.036	2.476	0.968	[0.936, 0.993]		KCP$_C$	8.646	2.052	0.965	[0.932, 0.990]
		KCP$_P$	0.000	7.000	0.146	[0.146, 0.146]		KCP$_P$	0.000	7.000	0.146	[0.146, 0.146]
		KCP$_M$	0.304	6.712	0.171	[0.146, 0.264]		KCP$_M$	0.104	6.903	0.158	[0.146, 0.248]
		KCP	0.000	7.000	0.146	[0.146, 0.146]		KCP	0.000	7.000	0.146	[0.146, 0.146]
		SBS$_C$	10.418	4.203	0.965	[0.931, 0.991]		SBS$_C$	8.984	2.717	0.965	[0.929, 0.990]
		SBS$_P$	0.028	6.974	0.155	[0.146, 0.321]		SBS$_P$	0.022	6.980	0.153	[0.146, 0.146]
		SBS$_M$	9.222	3.545	0.866	[0.735, 0.924]		SBS$_M$	6.264	2.554	0.805	[0.510, 0.915]

Table 3: Changes in both distribution and covariance matrix in Example 1 with outliers

4.2 Simulations on Clustering

We conduct a comparative analysis between the iterative subspace cluster algorithm and several well-known clustering methods applied directly on data sets that exhibit clustering structure. Specifically, we consider five commonly used clustering methods: the K-means method (Hartigan and Wong, 1979), the partitioning around medoid method (Reynolds et al., 2006), the expectation-maximization algorithm (Fraley and Raftery, 2002), the density-based spatial clustering of applications with noise method (Ester et al., 1996) and the spectral clustering method (Ng et al., 2001), which are denoted as K-means, PAM, EM, DBSCAN and SC, respectively. We compare two dimension reduction techniques: iterative CKPCA and KPCA. For example, K-means$_C$ and K-means$_P$ represent applying K-means after iterative CKPCA and KPCA, respectively. To assess the performance, we employ the Rand Index (RI) as a measure of similarity between the underlying clusters and the estimated clusters. The average (Mean), standard deviation (SD), and 95% percentile interval (PI) of the RI values are reported. K-means, PAM, EM, and DBSCAN are implemented in the R packages "stats", "cluster", "mclust", and "fpc", respectively, whereas SC is implemented in the Python package "scikit-learn". K-means, PAM, EM, and their dimension-reduced versions use the default settings. For DBSCAN and the dimension-reduced versions, we set `eps=1`. For SC and the dimension-reduced versions, we set `gamma=0.1`. In addition, for SC and SC$_P$, we set `n_components=10`. The bandwidth h is selected following the pro-

p	a	Method	$\hat{s}$	RMSE	RI	PI	p	a	Method	$\hat{s}$	RMSE	RI	PI
		E-Divisive$_C$	7.286	0.623	0.975	[0.947, 0.994]			E-Divisive$_C$	7.145	0.784	0.963	[0.891, 0.992]
		E-Divisive$_P$	6.513	1.087	0.939	[0.808, 0.989]			E-Divisive$_P$	5.990	1.669	0.903	[0.644, 0.985]
		E-Divisive$_M$	2.936	4.852	0.522	[0.124, 0.908]			E-Divisive$_M$	2.263	5.174	0.481	[0.124, 0.884]
		E-Divisive	4.283	3.307	0.754	[0.124, 0.957]			E-Divisive	3.264	4.239	0.642	[0.124, 0.950]
		Multirank$_C$	4.131	4.311	0.618	[0.124, 0.966]			Multirank$_C$	3.721	4.594	0.575	[0.124, 0.965]
		Multirank$_P$	0.144	6.940	0.127	[0.124, 0.124]			Multirank$_P$	0.775	6.684	0.148	[0.124, 0.564]
		Multirank$_M$	0.000	7.000	0.124	[0.124, 0.124]			Multirank$_M$	0.000	7.000	0.124	[0.124, 0.124]
200	4	Multirank	1.362	6.429	0.211	[0.124, 0.770]	100	4	Multirank	0.891	6.632	0.184	[0.124, 0.757]
		KCP$_C$	7.263	0.577	0.991	[0.976, 0.999]			KCP$_C$	7.312	0.669	0.986	[0.965, 0.998]
		KCP$_P$	0.477	6.594	0.214	[0.124, 0.685]			KCP$_P$	0.113	6.897	0.146	[0.124, 0.340]
		KCP$_M$	0.211	6.802	0.144	[0.124, 0.248]			KCP$_M$	0.019	6.982	0.127	[0.124, 0.124]
		KCP	0.000	7.000	0.124	[0.124, 0.124]			KCP	0.000	7.000	0.124	[0.124, 0.124]
		SBS$_C$	8.708	3.383	0.888	[0.149, 0.980]			SBS$_C$	7.869	2.944	0.849	[0.141, 0.975]
		SBS$_P$	0.013	6.988	0.127	[0.124, 0.124]			SBS$_P$	0.019	6.982	0.129	[0.124, 0.124]
		SBS$_M$	9.827	5.783	0.659	[0.151, 0.905]			SBS$_M$	8.399	4.141	0.646	[0.157, 0.891]
		E-Divisive$_C$	7.492	1.089	0.965	[0.916, 0.994]			E-Divisive$_C$	6.218	2.545	0.855	[0.124, 0.988]
		E-Divisive$_P$	4.997	2.659	0.821	[0.340, 0.969]			E-Divisive$_P$	4.222	3.354	0.752	[0.124, 0.962]
		E-Divisive$_M$	4.621	3.761	0.682	[0.124, 0.921]			E-Divisive$_M$	3.419	4.247	0.622	[0.124, 0.902]
		E-Divisive	1.971	5.315	0.480	[0.124, 0.891]			E-Divisive	0.835	6.276	0.300	[0.124, 0.795]
		Multirank$_C$	3.842	4.493	0.589	[0.124, 0.961]			Multirank$_C$	3.001	5.159	0.485	[0.124, 0.961]
		Multirank$_P$	0.000	7.000	0.124	[0.124, 0.124]			Multirank$_P$	0.000	7.000	0.124	[0.124, 0.124]
		Multirank$_M$	0.000	7.000	0.124	[0.124, 0.124]			Multirank$_M$	0.000	7.000	0.124	[0.124, 0.124]
200	6	Multirank	0.775	6.676	0.176	[0.124, 0.756]	100	6	Multirank	0.487	6.786	0.156	[0.124, 0.686]
		KCP$_C$	7.699	1.151	0.984	[0.959, 0.998]			KCP$_C$	7.294	1.292	0.956	[0.761, 0.996]
		KCP$_P$	0.023	6.979	0.128	[0.124, 0.124]			KCP$_P$	0.004	6.996	0.125	[0.124, 0.124]
		KCP$_M$	0.223	6.791	0.147	[0.124, 0.278]			KCP$_M$	0.050	6.954	0.131	[0.124, 0.204]
		KCP	0.000	7.000	0.124	[0.124, 0.124]			KCP	0.000	7.000	0.124	[0.124, 0.124]
		SBS$_C$	7.772	3.323	0.743	[0.137, 0.966]			SBS$_C$	6.406	3.076	0.638	[0.132, 0.963]
		SBS$_P$	0.015	6.986	0.128	[0.124, 0.124]			SBS$_P$	0.020	6.982	0.130	[0.124, 0.124]
		SBS$_M$	8.089	3.990	0.664	[0.151, 0.912]			SBS$_M$	5.971	3.329	0.601	[0.141, 0.885]

Table 4: Changes in distribution in Example 2 with balanced data set

cedure described for change point detection. Experiments are conducted on balanced and imbalanced data sets with three categories: (1) balanced data set with equal sample sizes $n_1 = n_2 = n_3 = n/3$; (2) imbalanced data set with sample sizes $n_1 = 300$, $n_2 = 200$, and $n_3 = 100$. Data X is generated according to the following settings: the kth class contains $\{X_{k,i}\}_{i=1}^{n_k}$, where $X_{k,i} = \sigma_{k,i} w_{k,i}$ for $k = 1, 2, 3$ and $i = 1, \ldots, n_k$, with $\sigma_{k,i}$ sampled from a uniform distribution in the range $[2k-2, 2k-1]$, and $w_{k,i}$ sampled from a uniform distribution on the unit sphere $\mathbb{S}^p$. Here, $p = 10, 100$. Each simulation is repeated 1000 times.

The findings are presented in Tables 7 and 8. Among the various methods considered for balanced data, it is observed that five CKPCA versions exhibit superior performance, achieving a RI of 1. In contrast, PAM's effectiveness is limited, while PAM$_C$ performs well. Notably, CKPCA significantly enhances the performance of most clustering methods, with iterative CKPCA surpassing KPCA. Similar patterns emerge when examining the results for imbalanced data, with CKPCA versions displaying the most favorable performance. Conversely, PAM performs poorly, but its performance is improved by PAM$_C$. These results unequivocally indicate that iterative CKPCA enhances the performance of popular clustering methods and demonstrates exceptional robustness in the context of imbalanced cases.

p	a	Method	$\hat{s}$	RMSE	RI	PI	p	a	Method	$\hat{s}$	RMSE	RI	PI
		E-Divisive$_C$	7.266	0.780	0.971	[0.940, 0.993]			E-Divisive$_C$	7.003	1.066	0.959	[0.886, 0.990]
		E-Divisive$_P$	5.704	1.768	0.925	[0.737, 0.981]			E-Divisive$_P$	5.210	2.283	0.891	[0.667, 0.977]
		E-Divisive$_M$	2.804	4.914	0.519	[0.146, 0.896]			E-Divisive$_M$	2.396	5.069	0.506	[0.146, 0.879]
		E-Divisive	3.475	3.925	0.734	[0.146, 0.946]			E-Divisive	2.625	4.720	0.621	[0.146, 0.937]
		Multirank$_C$	3.622	4.530	0.596	[0.146, 0.957]			Multirank$_C$	3.263	4.818	0.552	[0.146, 0.958]
		Multirank$_P$	0.277	6.909	0.154	[0.146, 0.270]			Multirank$_P$	1.194	6.565	0.180	[0.146, 0.585]
		Multirank$_M$	0.000	7.000	0.146	[0.146, 0.146]			Multirank$_M$	0.000	7.000	0.146	[0.146, 0.146]
200	4	Multirank	1.285	6.501	0.218	[0.146, 0.751]	100	4	Multirank	0.897	6.574	0.201	[0.146, 0.728]
		KCP$_C$	7.269	0.575	0.989	[0.967, 0.999]			KCP$_C$	7.264	0.629	0.983	[0.952, 0.997]
		KCP$_P$	0.714	6.386	0.318	[0.146, 0.827]			KCP$_P$	0.092	6.918	0.164	[0.146, 0.346]
		KCP$_M$	0.215	6.798	0.167	[0.146, 0.309]			KCP$_M$	0.043	6.960	0.153	[0.146, 0.224]
		KCP	0.000	7.000	0.146	[0.146, 0.146]			KCP	0.002	6.998	0.146	[0.146, 0.146]
		SBS$_C$	8.289	3.242	0.868	[0.172, 0.974]			SBS$_C$	7.789	3.320	0.836	[0.164, 0.968]
		SBS$_P$	0.007	6.993	0.148	[0.146, 0.146]			SBS$_P$	0.009	6.992	0.149	[0.146, 0.146]
		SBS$_M$	9.967	5.981	0.653	[0.176, 0.897]			SBS$_M$	8.488	4.273	0.635	[0.194, 0.883]
		E-Divisive$_C$	7.137	1.284	0.956	[0.895, 0.988]			E-Divisive$_C$	5.691	2.903	0.853	[0.146, 0.980]
		E-Divisive$_P$	4.202	3.253	0.811	[0.311, 0.963]			E-Divisive$_P$	3.413	4.000	0.725	[0.146, 0.951]
		E-Divisive$_M$	4.754	3.654	0.686	[0.146, 0.912]			E-Divisive$_M$	3.530	4.160	0.626	[0.146, 0.894]
		E-Divisive	1.526	5.698	0.441	[0.146, 0.882]			E-Divisive	0.623	6.462	0.281	[0.146, 0.792]
		Multirank$_C$	3.427	4.682	0.570	[0.146, 0.955]			Multirank$_C$	2.815	5.159	0.492	[0.146, 0.948]
		Multirank$_P$	0.000	7.000	0.146	[0.146, 0.146]			Multirank$_P$	0.000	7.000	0.146	[0.146, 0.146]
		Multirank$_M$	0.000	7.000	0.146	[0.146, 0.146]			Multirank$_M$	0.000	7.000	0.146	[0.146, 0.146]
200	6	Multirank	0.723	6.675	0.190	[0.146, 0.703]	100	6	Multirank	0.446	6.813	0.171	[0.146, 0.693]
		KCP$_C$	7.593	1.041	0.980	[0.941, 0.998]			KCP$_C$	7.109	1.200	0.956	[0.840, 0.995]
		KCP$_P$	0.010	6.991	0.148	[0.146, 0.146]			KCP$_P$	0.001	6.999	0.146	[0.146, 0.146]
		KCP$_M$	0.203	6.809	0.164	[0.146, 0.274]			KCP$_M$	0.048	6.956	0.155	[0.146, 0.237]
		KCP	0.000	7.000	0.146	[0.146, 0.146]			KCP	0.000	7.000	0.146	[0.146, 0.146]
		SBS$_C$	7.110	3.170	0.701	[0.152, 0.956]			SBS$_C$	5.981	3.097	0.612	[0.152, 0.954]
		SBS$_P$	0.021	6.981	0.152	[0.146, 0.146]			SBS$_P$	0.012	6.989	0.150	[0.146, 0.146]
		SBS$_M$	8.232	4.066	0.662	[0.186, 0.902]			SBS$_M$	6.039	3.136	0.591	[0.168, 0.883]

Table 5: Changes in distribution in Example 2 with imbalanced data set

p	a	h	$\hat{s}$	RMSE	RI	PI
		0.4	7.549	0.999	0.977	[0.947, 0.995]
		0.8	7.286	0.623	0.975	[0.947, 0.994]
200	4	1.2	7.228	0.581	0.973	[0.941, 0.993]
		1.6	7.182	0.694	0.969	[0.931, 0.993]
		2	6.982	1.054	0.956	[0.829, 0.992]

Table 6: The results of Example 2 with different bandwidth values

4.3 Real Data Examples

4.3.1 Genetics Data

We analyze an array comparative genomic hybridization (aCGH) microarray data set, previously analyzed in Stransky et al. (2006) and Blaveri et al. (2006), to detect mean changes in the data structure. The data set comprises 57 individuals with bladder tumors. We use the processed data from the R package "ecp" and select 43 individuals out of the 57, along with 2215 different loci on their genome, resulting in $p = 43$ and $n = 2215$. In the bandwidth formula, we set $m = 0.8$. This empirical study aims to identify unusual chromosomal characteristics.

p	Method	Mean	SD	PI	p	Method	Mean	SD	PI
	DBSCAN$_C$	1.000	0.000	[1.000, 1.000]		DBSCAN$_C$	1.000	0.000	[1.000, 1.000]
	DBSCAN$_P$	0.800	0.006	[0.789, 0.812]		DBSCAN$_P$	0.776	0.003	[0.771, 0.777]
	DBSCAN	0.777	0.000	[0.777, 0.777]		DBSCAN	0.777	0.000	[0.777, 0.777]
	EM$_C$	1.000	0.000	[1.000, 1.000]		EM$_C$	1.000	0.000	[1.000, 1.000]
	EM$_P$	0.737	0.019	[0.728, 0.752]		EM$_P$	0.731	0.015	[0.723, 0.751]
	EM	0.733	0.006	[0.724, 0.748]		EM	0.735	0.005	[0.726, 0.747]
	K-means$_C$	1.000	0.000	[1.000, 1.000]		K-means$_C$	1.000	0.000	[1.000, 1.000]
100	K-means$_P$	1.000	0.000	[1.000, 1.000]	10	K-means$_P$	0.999	0.020	[1.000, 1.000]
	K-means	0.505	0.016	[0.470, 0.536]		K-means	0.565	0.012	[0.539, 0.587]
	PAM$_C$	1.000	0.000	[1.000, 1.000]		PAM$_C$	1.000	0.000	[1.000, 1.000]
	PAM$_P$	1.000	0.000	[1.000, 1.000]		PAM$_P$	0.922	0.015	[0.893, 0.951]
	PAM	0.334	0.000	[0.334, 0.334]		PAM	0.523	0.028	[0.487, 0.583]
	SC$_C$	1.000	0.000	[1.000, 1.000]		SC$_C$	1.000	0.000	[1.000, 1.000]
	SC$_P$	0.833	0.006	[0.823, 0.844]		SC$_P$	0.809	0.003	[0.802, 0.815]
	SC	0.706	0.007	[0.692, 0.720]		SC	0.726	0.008	[0.710, 0.742]

Table 7: Balanced data with $n_1 = n_2 = n_3 = 200$

p	Method	Mean	SD	PI	p	Method	Mean	SD	PI
	DBSCAN$_C$	1.000	0.000	[1.000, 1.000]		DBSCAN$_C$	1.000	0.000	[1.000, 1.000]
	DBSCAN$_P$	0.889	0.000	[0.889, 0.889]		DBSCAN$_P$	0.855	0.012	[0.830, 0.875]
	DBSCAN	0.889	0.000	[0.889, 0.889]		DBSCAN	0.889	0.000	[0.889, 0.889]
	EM$_C$	1.000	0.000	[1.000, 1.000]		EM$_C$	1.000	0.010	[1.000, 1.000]
	EM$_P$	0.800	0.026	[0.780, 0.828]		EM$_P$	0.783	0.013	[0.767, 0.819]
	EM	0.786	0.010	[0.770, 0.812]		EM	0.794	0.011	[0.776, 0.818]
	K-means$_C$	0.978	0.067	[0.770, 1.000]		K-means$_C$	0.983	0.059	[0.772, 1.000]
100	K-means$_P$	0.921	0.108	[0.768, 1.000]	10	K-means$_P$	0.981	0.060	[0.788, 1.000]
	K-means	0.540	0.020	[0.498, 0.579]		K-means	0.622	0.016	[0.590, 0.653]
	PAM$_C$	1.000	0.000	[1.000, 1.000]		PAM$_C$	1.000	0.000	[1.000, 1.000]
	PAM$_P$	1.000	0.000	[1.000, 1.000]		PAM$_P$	0.992	0.006	[0.979, 1.000]
	PAM	0.392	0.000	[0.392, 0.392]		PAM	0.634	0.037	[0.555, 0.699]
	SC$_C$	1.000	0.000	[1.000, 1.000]		SC$_C$	1.000	0.000	[1.000, 1.000]
	SC$_P$	0.781	0.005	[0.768, 0.787]		SC$_P$	0.902	0.001	[0.900, 0.903]
	SC	0.800	0.008	[0.783, 0.815]		SC	0.818	0.009	[0.802, 0.836]

Table 8: Imbalanced data with $n_1 = 300$, $n_2 = 200$, $n_3 = 100$

Given that E-Divisive$_C$ outperformed other methods in previous simulation studies, we employ this method. We discover 34 change points in the dimension-reduced data. Since the dimension q_d is determined as 1 based on the TRR criterion, Figure 2 depicts the change point locations. From Figure 2, it is evident that we can successfully detect the jump locations in the data.

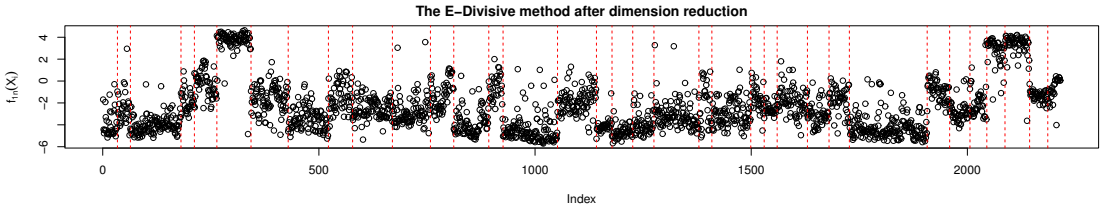

Figure 2: Change point detection for aCGH data, the figure plots the locations detected by the dimension reduction-based E-divisive method.

4.3.2 MNIST Data

We use the MNIST (Modified National Institute of Standards and Technology) data set, available at `http://yann.lecun.com/exdb/mnist/`, for clustering analysis. Specifically, we employ the proposed iterative CKPCA algorithm and compare its performance against the original KPCA version. The evaluation metric used is the Rand Index, which measures the similarity between the real and estimated clustering results. Our analysis focuses on a subset of the MNIST data set consisting of 900 samples, divided into three groups of 300 samples each. Each group corresponds to handwritten digits 6, 8, and 9, which are known to be more challenging to distinguish. Each digit is represented as a grayscale image with dimensions 28×28. Therefore, we have $n = 900$ instances, $p = 784$ dimensions, and $d = 3$ categories, with $n_1 = n_2 = n_3 = 300$. To mitigate sampling randomness, we repeat the experiment 50 times. In our experiments, we set $m = 0.8$ in the bandwidth formula. For DBSCAN, SC, and their dimension-reduced versions, using the same parameter settings as in the simulation study does not yield satisfactory performance. To ensure a fair choice of tuning parameters and a clearer presentation of the results, we carry out a separate preliminary tuning step for this data. Based on the results of the preliminary tuning step, we select `eps=0.01`, `1`, and `1` for DBSCAN_C, DBSCAN_P, and DBSCAN, respectively, and `gamma=0.01`, `10`, and `0.05` for SC_C, SC_P, and SC, respectively. Figure 3 shows the scatter plots of the first two dimensions after PCA, KPCA, and CKPCA. It is observed that applying CKPCA improves the separability of different classes compared to KPCA. The clustering results are presented in Table 9. According to the results, PAM_C achieves the best performance. Additionally, applying the proposed iterative CKPCA algorithm improves the performance of all five clustering methods.

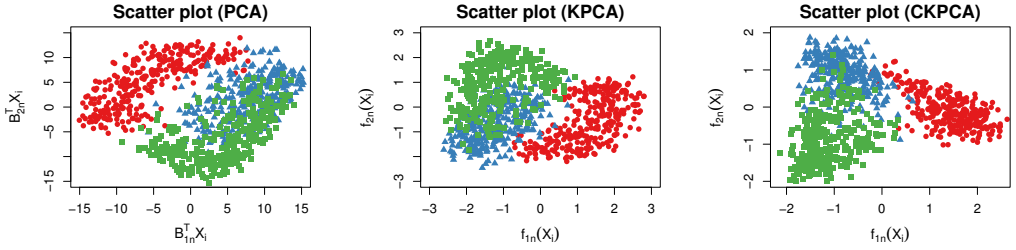

Figure 3: Scatter plots after PCA, KPCA and CKPCA, the three groups of points in red, blue, and green corresponding to the numbers 6, 8, and 9, respectively

Method	RI	Method	RI	Method	RI	Method	RI	Method	RI
DBSCAN$_C$	0.655	EM$_C$	0.739	K-means$_C$	0.844	PAM$_C$	0.876	SC$_C$	0.869
DBSCAN$_P$	0.346	EM$_P$	0.334	K-means$_P$	0.827	PAM$_P$	0.772	SC$_P$	0.857
DBSCAN	0.333	EM	0.338	K-means	0.826	PAM	0.799	SC	0.863

Table 9: The clustering results of real data

5. Conclusion

The motivation of this paper is to apply dimension reduction methods that preserve information for unsupervised learning with heterogeneous data. We develop a Corrected Kernel Principal Component Analysis (CKPCA) method and introduce a notion of kernel mean embedding deviation subspace for identifying distributional changes in data. The identification is implemented in this dimension reduction subspace without loss of information from the original data. As a special case, the Corrected Principal Component Analysis (CPCA) is developed to identify the central mean deviation subspace proposed in Zhu et al. (2025), specifically for detecting mean changes. Finally, we extend CKPCA to cluster analysis, aiming to achieve a superior nonlinear low-dimensional embedding, and develop an iterative subspace cluster algorithm.

This general methodology can readily be extended to other change point detection problems such as missing data (Follain et al., 2021), tensor data (Huang et al., 2022), private data (Berrett and Yu, 2021), and online data (Chen et al., 2021). The asymptotic results apply to both dense and sparse data structures. The main limitation of the current method is its incapability of handling ultra-high dimension data. A possible solution is combining it with simultaneous variable selection, see, for example, Wang et al. (2018), Lin et al. (2019), and Qian et al. (2019). The research is ongoing.

Acknowledgments

Corresponding author (X. Zhu). Email addresses: zhuxuehu@xjtu.edu.cn (X. Zhu). The names of the coauthors are in the alphabetical order. The authors thank the editor, the associate editor, and two anonymous referees for their helpful comments and suggestions, which have greatly improved the manuscript. Xuehu Zhu's research was supported by a grant from National Key R&D Program of China (No.2025YFA1016501), and a grant from the National Scientific Foundation of China (12371276). Lixing Zhu's research was supported by the grants (NSFC12131006, NSFC12471276) from the National Natural Scientific Foundation of China and the grant (CI2023C063YLL) from the Scientific and Technological Innovation Project of China Academy of Chinese Medical Science. The authors declare no competing interests.

Appendix A. Nonlinear Dimension Reduction from the Perspective of σ-Fields

Definition 1 depends on the choice of kernel and is therefore not itself an invariant object. Motivated by the notion of the central σ-field in Li (2018), we consider nonlinear dimension reduction for unsupervised learning from a σ-field perspective. In particular, we take σ-fields as the fundamental invariant quantity in our discussion. Let $\sigma(X)$ denote the σ-field generated by X. For $i = 1, \ldots, s+1$, let $X^{(i)}$ follow the distribution $P^{(i)}$ in (1).

Definition 13 *Suppose that there exist functions $f_1, \ldots, f_q \in \mathcal{H}$ such that $\{f(X_i)\}_{i=1}^n$ and $\{X_i\}_{i=1}^n$ have the same change point locations, where $f(X_i) = \big(f_1(X_i), \ldots, f_q(X_i)\big)^\top$. Then*

$$\mathcal{G}_{\{X_i\}_{i=1}^n} = \sigma\big(f(X^{(1)}), f(X^{(2)}), \ldots, f(X^{(s+1)})\big)$$

is called a changepoint-preserving σ-field of $\{X_i\}_{i=1}^n$. Moreover, define

$$\mathcal{H}(\mathcal{G}_{\{X_i\}_{i=1}^n}) = \overline{\text{Span}}\left\{g \in \mathcal{H} : g \text{ is measurable } \mathcal{G}_{\{X_i\}_{i=1}^n}\right\},$$

where $\overline{\text{Span}}$ denotes the closure of the spanned space. The set $\mathcal{H}\big(\mathcal{G}_{\{X_i\}_{i=1}^n}\big)$ is defined as the changepoint-preserving class of $\{X_i\}_{i=1}^n$.

This definition provides a natural way to characterize information preservation in nonlinear dimension reduction that is invariant to the choice of kernel. $\mathrm{S}^d_{\{X_i\}_{i=1}^n}$ is closely related to the changepoint-preserving σ-field and the changepoint-preserving class. By Theorem 2, let $\{v_1, \ldots, v_{q_d}\}$ be any basis of $\mathrm{S}^d_{\{X_i\}_{i=1}^n}$, and define $f(X_i) = \big(\langle v_1, Y_i\rangle_{\mathcal{H}}, \ldots, \langle v_{q_d}, Y_i\rangle_{\mathcal{H}}\big)^\top$. Then the sequence $\{f(X_i)\}_{i=1}^n$ and the original sequence $\{X_i\}_{i=1}^n$ have the same change point locations. Therefore, $\mathcal{G}^d_{\{X_i\}_{i=1}^n} = \sigma\big(f(X^{(1)}), f(X^{(2)}), \ldots, f(X^{(s+1)})\big)$ is a changepoint-preserving σ-field of $\{X_i\}_{i=1}^n$. Accordingly, $\mathcal{H}\big(\mathcal{G}^d_{\{X_i\}_{i=1}^n}\big)$ is a changepoint-preserving class of $\{X_i\}_{i=1}^n$. Moreover, for any $g \in \mathrm{S}^d_{\{X_i\}_{i=1}^n}$, we have $g \in \mathcal{H}\big(\mathcal{G}^d_{\{X_i\}_{i=1}^n}\big)$, hence $\mathrm{S}^d_{\{X_i\}_{i=1}^n} \subseteq \mathcal{H}\big(\mathcal{G}^d_{\{X_i\}_{i=1}^n}\big)$.

For clustering problems, we can analogously define clustering-preserving σ-fields and clustering-preserving classes.

Appendix B. Corrected Principal Component Analysis

In the special case where $\phi(X_i) = X_i$, KPCA reduces to PCA. Let $\{X_i\}_{i=1}^n$ follow the unknown distributions $\{P_i\}_{i=1}^n$ with s change points $1 \le z_1 < \ldots < z_s \le n$. These change points satisfy the following:

$$\mu_{z_i+1} = \cdots = \mu_{z_{i+1}} =: \mu^{(i)} \text{ and } \mu^{(j)} \neq \mu^{(j+1)}, \quad \forall 0 \leqslant i \leqslant s, 1 \leqslant j \leqslant s. \tag{10}$$

The model (10) represents changes in the mean and corresponds to the previous model (1). Additionally, we simplify the model (2) to the following form:

$$\forall 1 \le i \le n, \quad X_i = \mu_i + \epsilon_i \in \mathbb{R}^p.$$

Next, we consider the sample covariance matrix Σ_n defined as:

$$\Sigma_n = \frac{1}{n}\sum_{i=1}^n \left(X_i - \bar{X}\right)\left(X_i - \bar{X}\right)^\top,$$

where $\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$. Compute the expectation of "the sample covariance" Σ_n as:

$$\begin{aligned} E(\Sigma_n) &= \sum_{j=1}^{s+1} \frac{1}{n} \sum_{i=z_{j-1}+1}^{z_j} E\left\{ \left(X_i - \bar{X}\right)\left(X_i - \bar{X}\right)^\top \right\} \\ &\to \sum_{j=1}^{s+1} c_j \Sigma^{(j)} + \frac{1}{2}\sum_{i=1}^{s+1}\sum_{j=1}^{s+1} c_i c_j (\mu_i - \mu_j)(\mu_i - \mu_j)^\top \equiv: \Sigma_{pooled} + \Delta = \Sigma, \end{aligned} \tag{11}$$

where $n_i/n \to c_i$ as $n \to \infty$, $\Sigma_{pooled} = \sum_{j=1}^{s+1} c_j \Sigma^{(j)}$ and $\Delta = \frac{1}{2}\sum_{i=1}^{s+1}\sum_{j=1}^{s+1} c_i c_j (\mu_i - \mu_j)(\mu_i - \mu_j)^\top$. Referring to Definition 2.1 in Zhu et al. (2025), the central mean deviation subspace of the sequence $\{X_i\}_{i=1}^n$ is defined as $\mathrm{Span}\{\mu^{(i)} - \mu^{(j)},\ \text{for } i, j = 1, \ldots, s+1\}$ and denoted as $S_{\{E(X_i)\}_{i=1}^n}$. The structural dimension of $S_{\{E(X_i)\}_{i=1}^n}$ is represented by $q = \dim\{S_{\{E(X_i)\}_{i=1}^n}\}$. The following theorem establishes the equivalence between $\mathrm{Span}\{\Delta\}$ and $S_{\{E(X_i)\}_{i=1}^n}$.

Theorem 14 *Under the model (10),* $\mathrm{Span}\{\Delta\} = \mathrm{S}_{\{E(X_i)\}_{i=1}^n}$. *Furthermore,* $\mathrm{Span}\{B\} = \mathrm{S}_{\{E(X_i)\}_{i=1}^n}$, *where* $B = (v_1, \ldots, v_q)$ *denotes the matrix consisting of the eigenvectors of* Δ *associated with the nonzero eigenvalues of* Δ, *and the sequences* $\{B^\top X_i\}_{i=1}^n$ *and* $\{X_i\}_{i=1}^n$ *have the same locations of changes.*

The proof of Theorem 14 is given in Appendix F.10. Consistent with CKPCA, we estimate Σ_{pooled} by the localized method:

$$\Sigma_{pooled,n} = \frac{1}{r}\sum_{m=1}^{r} \hat{\Sigma}_m \text{ with } \hat{\Sigma}_m = \frac{1}{\hat{n}_m - 1}\sum_{i \in \mathcal{S}_m} (X_i - \bar{X}_m)(X_i - \bar{X}_m)^\top,$$

where $\bar{X}_m = \frac{1}{\hat{n}_m}\sum_{i \in \mathcal{S}_m} X_i$. Combining (11), Δ_n can be estimated as:

$$\Delta_n = \Sigma_n - \Sigma_{pooled,n}.$$

Then an estimator B_n of the basis matrix B consists of the eigenvectors associated with the largest q eigenvalues of Δ_n. The properties of Δ_n are stated in the following theorem.

Theorem 15 *Under the model (10), assume that* X_i *are* p*-dimensional independent random vectors, and Assumptions 9 – 12 hold. Then, when* q *is given,*

$$||\Delta_n - \Delta||_F = O_p\left(\sqrt{\frac{p}{n}} + \frac{\sqrt{p}\beta_n}{n}\right) \text{ and } ||B_n - B||_F = O_p\left(\sqrt{\frac{p}{n}} + \frac{\sqrt{p}\beta_n}{n}\right),$$

where $||\cdot||_F$ *denotes the Frobenius norm of a matrix.*

Remark 16 *The CPCA method is specifically designed for detecting mean changes. It may have limitations in terms of generality for nonparametric models but is suitable for high-dimensional data. On the other hand, the dimension reduction method presented in Zhu et al. (2025) and CPCA both effectively identify the central mean deviation subspace of the sequence* $\{X_i\}_{i=1}^n$. *Additionally, the estimated matrices obtained from these two methods converge to the same target matrix* Δ *with the same convergence rate. However, CPCA has a lower computational complexity of* $O(p^2 n)$ *compared to the complexity of* $O(p^2 n^2)$ *in Zhu et al. (2025).*

The proof of Theorem 15 is given in Appendix F.11. Additionally, we adopt the TRR criterion, as given by (6), to estimate the structural dimension q. In the high-dimensional situation, if $\beta_n = o(n^m)$ with $0 \leq m \leq 1/2$, we have $||\Delta_n - \Delta||_F = O_p\left(\sqrt{\frac{p}{n}}\right)$. Then we set $c_n = 0.2\log(\log(n))\sqrt{p/n}$.

Appendix C. Part of Numerical Analysis

C.1 The Impact of β_n on the Method

To further demonstrate the robustness against the different β_n, we present the numerical comparisons for Example 1 of the main body with β_n to be $2, \lfloor n^{1/5}\rfloor, \lfloor n^{1/3}\rfloor, \lfloor n^{3/7}\rfloor, \lfloor n^{1/2}\rfloor, \lfloor n^{4/7}\rfloor, \lfloor n^{3/5}\rfloor, \lfloor n^{3/4}\rfloor, \lfloor n/3\rfloor$. For the change point method, we chose E-Divisive as it yielded the best performance to showcase our findings. The experimental results can be found in Figure 4, which show that the performance of β_n values such as $\lfloor n^{1/3}\rfloor, \lfloor n^{3/7}\rfloor, \lfloor n^{1/2}\rfloor, \lfloor n^{4/7}\rfloor$ surpasses that of $\beta_n = 2, \lfloor n^{1/5}\rfloor, \lfloor n^{3/5}\rfloor, \lfloor n^{3/4}\rfloor, \lfloor n/3\rfloor$. We also see that the results are relatively robust against the different β_n. We then recommend $\lfloor n^{1/2}\rfloor$.

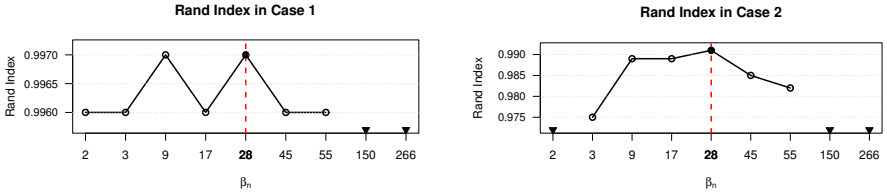

Figure 4: Plots of the Rand index versus β_n for balanced data in Example 1. The values of β_n considered are $2, \lfloor n^{1/5}\rfloor, \lfloor n^{1/3}\rfloor, \lfloor n^{3/7}\rfloor, \lfloor n^{1/2}\rfloor, \lfloor n^{4/7}\rfloor, \lfloor n^{3/5}\rfloor, \lfloor n^{3/4}\rfloor$, and $\lfloor n/3\rfloor$. For visual clarity, the y-axis is truncated to emphasize the region around the optimum.

C.2 Numerical Studies with Changes in Mean

Example 3: Changes in mean. The data are generated from the distributions G_1, G_2, G_2, $G_3, \ldots, G_8$. We consider the multivariate normal distribution situation of $G_i = N(u_i, \Sigma)$, where $\Sigma = (\sigma_{ij})$ is the $p \times p$ covariance matrix with $\sigma_{ij} = 0.5^{|i-j|}$, u_i is a p-dimensional vector satisfying $u_1 = u_3 = u_5 = u_7 = 0$, $u_2 = (u\mathrm{I}_{1\times 5}, \frac{u}{2}\mathrm{I}_{1\times 5}, 0, \ldots, 0)^\top$, $u_4 = (2u\mathrm{I}_{1\times 5}, \frac{u}{3}\mathrm{I}_{1\times 5}, 0, \ldots, 0)^\top$, $u_6 = (\frac{u}{2}\mathrm{I}_{1\times 5}, u\mathrm{I}_{1\times 5}, 0, \ldots, 0)^\top$ and $u_8 = (\frac{u}{3}\mathrm{I}_{1\times 5}, 2u\mathrm{I}_{1\times 5}, 0, \ldots, 0)^\top$ with $u = 0.5$. We set $p = 100, 200$. Different u_i represents the mean shift. The structure dimension q is 3.

The results of **Example 3** are presented in Tables 10 and 11. According to the tables, it is suggested that E-Divisive$_C$ and E-Divisive$_M$ perform the best. Their results are comparable, with rand indices exceeding 0.9. SBS and KCP perform worse compared to the others. None of the seven change point methods, including the three high-dimensional methods, are able to detect the change points without dimension reduction in this example. The four methods show significant improvement after applying the corrected PCA or corrected

Mahalanobis matrix method, which aligns with the theoretical result that the corrected PCA or corrected Mahalanobis matrix reduces dimensionality without losing any information. However, the four methods combined with classical PCA yield unsatisfactory results in this example. Additionally, the dimensionality p has minimal impact on the results after employing the corrected PCA or corrected Mahalanobis matrix method in **Example 3**.

p	Method	$\hat{s}$	RMSE	RI	PI	p	Method	$\hat{s}$	RMSE	RI	PI
	E-Divisive$_C$	7.527	1.352	0.934	[0.820, 0.978]		E-Divisive$_C$	6.851	1.096	0.937	[0.783, 0.986]
	E-Divisive$_P$	1.337	5.861	0.403	[0.124, 0.879]		E-Divisive$_P$	2.690	4.732	0.603	[0.124, 0.959]
	E-Divisive$_M$	7.532	1.370	0.934	[0.818, 0.978]		E-Divisive$_M$	6.850	1.083	0.938	[0.784, 0.984]
	E-Divisive	1.354	5.852	0.407	[0.124, 0.879]		E-Divisive	2.744	4.698	0.609	[0.124, 0.962]
	Multirank$_C$	5.637	2.408	0.844	[0.124, 0.981]		Multirank$_C$	5.258	2.977	0.800	[0.124, 0.985]
	Multirank$_P$	0.304	6.831	0.159	[0.124, 0.758]		Multirank$_P$	1.131	6.350	0.258	[0.124, 0.898]
	Multirank$_M$	3.584	4.618	0.576	[0.124, 0.976]		Multirank$_M$	1.745	5.980	0.345	[0.124, 0.975]
	Multirank	0.027	6.983	0.127	[0.124, 0.124]		Multirank	0.589	6.661	0.193	[0.124, 0.832]
	KCP$_C$	6.781	1.296	0.914	[0.725, 0.979]		KCP$_C$	6.180	1.587	0.904	[0.666, 0.983]
200	KCP$_P$	0.000	7.000	0.124	[0.124, 0.124]	100	KCP$_P$	0.000	7.000	0.124	[0.124, 0.124]
	KCP$_M$	4.357	3.822	0.679	[0.124, 0.973]		KCP$_M$	2.054	5.717	0.395	[0.124, 0.973]
	KCP	0.000	7.000	0.124	[0.124, 0.124]		KCP	0.000	7.000	0.124	[0.124, 0.124]
	SBS$_C$	6.998	1.466	0.922	[0.767, 0.982]		SBS$_C$	6.402	1.484	0.919	[0.671, 0.986]
	SBS$_P$	0.019	6.982	0.132	[0.124, 0.124]		SBS$_P$	0.039	6.964	0.138	[0.124, 0.462]
	SBS$_M$	6.949	1.467	0.921	[0.761, 0.982]		SBS$_M$	6.396	1.513	0.918	[0.668, 0.986]
	SBS	0.324	6.694	0.238	[0.124, 0.598]		SBS	0.321	6.696	0.240	[0.124, 0.597]
	Inspect	0.784	6.284	0.365	[0.124, 0.776]		Inspect	1.499	5.673	0.486	[0.124, 0.871]
	Geomcp	0.003	6.997	0.124	[0.124, 0.124]		Geomcp	0.003	6.997	0.124	[0.124, 0.124]
	DCBS	0.017	6.984	0.128	[0.124, 0.124]		DCBS	0.262	6.755	0.206	[0.124, 0.608]

Table 10: Changes in mean in Example 3 with balanced data set

p	Method	$\hat{s}$	RMSE	RI	PI	p	Method	$\hat{s}$	RMSE	RI	PI
	E-Divisive$_C$	7.616	1.569	0.927	[0.840, 0.976]		E-Divisive$_C$	6.759	1.146	0.936	[0.819, 0.982]
	E-Divisive$_P$	1.415	5.763	0.474	[0.146, 0.904]		E-Divisive$_P$	2.816	4.540	0.690	[0.146, 0.958]
	E-Divisive$_M$	7.615	1.579	0.926	[0.841, 0.975]		E-Divisive$_M$	6.752	1.137	0.936	[0.816, 0.983]
	E-Divisive	1.436	5.746	0.479	[0.146, 0.906]		E-Divisive	2.833	4.526	0.693	[0.146, 0.958]
	Multirank$_C$	5.458	2.273	0.865	[0.146, 0.975]		Multirank$_C$	5.050	2.921	0.817	[0.146, 0.982]
	Multirank$_P$	0.393	6.775	0.192	[0.146, 0.810]		Multirank$_P$	1.033	6.396	0.273	[0.146, 0.901]
	Multirank$_M$	3.333	4.672	0.585	[0.146, 0.969]		Multirank$_M$	1.564	6.023	0.356	[0.146, 0.973]
	Multirank	0.033	6.976	0.152	[0.146, 0.146]		Multirank	0.584	6.664	0.217	[0.146, 0.846]
	KCP$_C$	6.640	1.439	0.908	[0.744, 0.976]		KCP$_C$	6.020	1.663	0.906	[0.701, 0.981]
200	KCP$_P$	0.000	7.000	0.146	[0.146, 0.146]	100	KCP$_P$	0.000	7.000	0.146	[0.146, 0.146]
	KCP$_M$	4.070	4.010	0.669	[0.146, 0.965]		KCP$_M$	1.984	5.722	0.413	[0.146, 0.970]
	KCP	0.000	7.000	0.146	[0.146, 0.146]		KCP	0.000	7.000	0.146	[0.146, 0.146]
	SBS$_C$	6.853	1.561	0.916	[0.761, 0.977]		SBS$_C$	6.186	1.673	0.918	[0.689, 0.984]
	SBS$_P$	0.031	6.971	0.158	[0.146, 0.359]		SBS$_P$	0.043	6.960	0.162	[0.146, 0.541]
	SBS$_M$	6.844	1.532	0.916	[0.767, 0.976]		SBS$_M$	6.168	1.645	0.917	[0.701, 0.983]
	SBS	0.372	6.651	0.287	[0.146, 0.740]		SBS	0.413	6.612	0.303	[0.146, 0.753]
	Inspect	1.064	6.068	0.442	[0.146, 0.874]		Inspect	2.132	5.187	0.610	[0.146, 0.933]
	Geomcp	0.001	6.999	0.146	[0.146, 0.146]		Geomcp	0.006	6.995	0.146	[0.146, 0.146]
	DCBS	0.020	6.982	0.154	[0.146, 0.146]		DCBS	0.289	6.732	0.243	[0.146, 0.699]

Table 11: Changes in mean in Example 3 with imbalanced data set

Location	31	81	117	158	194
Times	01/1994	11/1998	11/2001	4/2005	4/2008
Location	224	273	303	335	
Times	1/2010	11/2014	5/2017	1/2020	

Table 12: The locations and times of the change points in macroeconomic data

Appendix D. Macroeconomic Data

We examine U.S. macroeconomic data, previously analyzed in Barigozzi et al. (2018) for mean change detection and used in Caner and Han (2014) to determine the number of common factors in approximate factor models. The data set comprises 92 monthly macroeconomic variables in the U.S., covering the period from February 1992 to May 2022. The data is available from the St. Louis Federal Reserve Bank website (`https://fred.stlouisfed.org/`). The data set encompasses various facets of the macroeconomy, including real output and income, employment and hours, real retail, manufacturing and trade sales, consumption, housing starts and sales, real inventories and inventory-sales ratios, orders and unfilled orders, stock prices, interest rates, money and credit quantity aggregates, price indexes, average hourly earnings, and miscellaneous. The covariates are transformed to achieve stationarity, and outliers are adjusted as described in Stock and Watson (2002). The data set has a dimension of $p = 92$ and a sample size of $n = 364$. For the analysis of this data set, we adopt a bandwidth value of $m = 2$ as specified in the bandwidth formula.

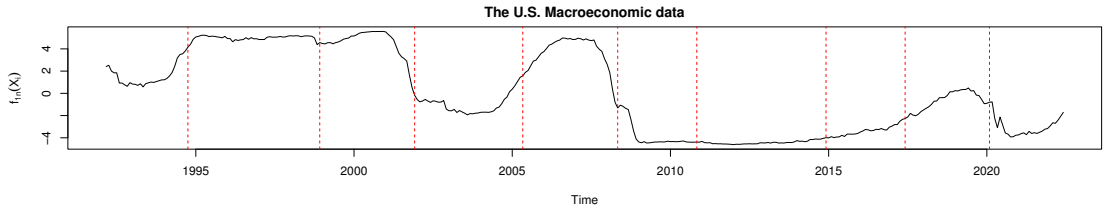

Figure 5: Change point detection after dimension reduction for U.S. macroeconomic data.

The results of the change point detection in the macroeconomic data are shown in Figure 5 and summarized in Table 12. Based on these results, we can analyze the observed periods corresponding to different economic regimes characterized by high or low volatility. Here is an analysis of the detected change points:

- The year 1995 was the first year of the great Internet bull market in the U.S. stock market in the 1990s. From that year until 2000, the U.S. stock market experienced a 5-year bull market.
- The 2001 recession lasted eight months (March-November). It was caused by the dotcom boom and subsequent bust.
- The U.S. economy is growing steadily and rapidly in 2005.

- The Great Recession, which lasted from December 2007 to June 2009. The crisis spread through the economy through the widespread use of derivatives. It leads to the low GDP growth and high unemployment.
- The U.S. economy is emerging from recession after 2009.
- The president's rise to power had a certain impact on the U.S. economy in 2017.
- From the fourth quarter of 2019 to the second quarter of 2020, the U.S. economy shrank at a record annual rate of 19.2 percent on average, which is due to the outbreak of COVID-19.

These identified change points provide insights into significant events and economic trends in the U.S. macroeconomic landscape over the analyzed period.

Appendix E. The Theoretical Analysis of the Selection of β_n

In this section, we analyze the selection of β_n from a theoretical perspective, focusing on the convergence rate and the balance between variance and bias. The results indicate the absence of a universally optimal β_n when change point locations are unknown. Consider that $\beta_n = O\left(n^m\right)$ and $r = O\left(n^{1-m}\right)$.

(1) We have examined the asymptotic properties of $\Delta_n^{kernel} - \Delta^{kernel}$ in Theorem 6. According to Theorem 6, when $0 \leq m \leq 1/2$, it follows that $\|\Delta_n^{kernel} - \Delta^{kernel}\|_{HS} = O_p\left(n^{-1/2}\right)$, and when $1/2 < m \leq 1$, $||\Delta_n^{kernel} - \Delta^{kernel}||_{HS} = O_p\left(n^{m-1}\right)$. This indicates that in the case with $0 \leq m \leq 1/2$, $\|\Delta_n^{kernel} - \Delta^{kernel}\|_{HS}$ has a faster convergence rate compared to the case with $1/2 < m \leq 1$. In other words, $\beta_n = O\left(n^{1/2}\right)$ and $r = O\left(n^{1/2}\right)$ can be optimal rates.

(2) Examine the bias-variance tradeoff concerning $\langle \alpha, \Delta_n^{kernel}\alpha\rangle_{\mathcal{H}}$. We have computed both bias and variance associated with $\langle \alpha, \Delta_n^{kernel}\alpha\rangle_{\mathcal{H}}$. Let us denote s as the number of change points in the sequence $\{Y_i\}_{i=1}^n$, and $z_1, \ldots, z_s$ as the locations of these change points. As s is a constant, we assume each z_i is an element of the set $\mathcal{S}_{m_i}$, defined as $\mathcal{S}_{m_i} = \{(m_i - 1)\beta_n + 1, \ldots, m_i\beta_n\}$. This set is further divided into two subsets: $\mathcal{S}_{m_{i1}} = \{(m_i - 1)\beta_n + 1, \ldots, z_i\}$ and $\mathcal{S}_{m_{i2}} = \{z_i + 1, \ldots, m_i\beta_n\}$, where a_{i1} and a_{i2} represent the number of elements in $\mathcal{S}_{m_{i1}}$ and $\mathcal{S}_{m_{i2}}$, respectively. It is established that $a_{i1} + a_{i2} = \beta_n$. Drawing upon the proofs provided in Theorems 6 and 7, we derive that

$$\text{Bias}(\langle \alpha, \Delta_n^{kernel}\alpha\rangle_{\mathcal{H}}) = O\left(\sum_{i=1}^{s} \frac{a_{1i}a_{2i}}{n\beta_n}\right).$$

To compute the variance $\text{Var}(\langle \alpha, \Delta_n^{kernel}\alpha\rangle_{\mathcal{H}})$, we consider the case where $\text{Var}(\langle \alpha, \epsilon_i\rangle_{\mathcal{H}}) = \sigma^2$ for any $i = 1, \ldots, n$. The proofs of Theorems 6 and 7 show that

$$\text{Var}(\langle \alpha, \Delta_n^{kernel}\alpha\rangle_{\mathcal{H}}) = O\left(\frac{1}{n\beta_n}\right).$$

Thus, altogether,

$$\text{Bias}^2(\langle \alpha, \Delta_n^{kernel}\alpha\rangle_{\mathcal{H}}) + \text{Var}(\langle \alpha, \Delta_n^{kernel}\alpha\rangle_{\mathcal{H}}) = O\left(\left(\sum_{i=1}^{s} \frac{a_{1i}a_{2i}}{n\beta_n}\right)^2\right) + O\left(\frac{1}{n\beta_n}\right).$$

In scenarios involving multiple change points, it is crucial to ascertain that no two change points fall within the same segment. To this end, we assume that the minimum distance between any two change points exceeds $\sqrt{n}$, formalized as $\min_{1\le i\le s}(z_i - z_{i-1}) > \sqrt{n}$. As a result, when dealing with multiple change points, the parameter m must adhere to the condition $m \le 0.5$. Furthermore, the subsequent results elucidate that the optimal value of β varies across different scenarios.

- **The scenario when $a_{11} = \cdots = a_{s1} = 0$ and $a_{12} = \cdots = a_{s2} = \beta_n$.** In this situation, we can obtain the MSE as:

$$\text{Bias}^2(\langle \alpha, \Delta_n^{kernel}\alpha\rangle_{\mathcal{H}}) + \text{Var}(\langle \alpha, \Delta_n^{kernel}\alpha\rangle_{\mathcal{H}}) = O\left(\frac{1}{n\beta_n}\right).$$

  This quantity is monotonically decreasing with respect to β_n. Therefore, we prefer a larger value of β_n as it leads to a smaller MSE. In this scenario, the optimal $\beta_n = O(n^{1/2})$.

- **The scenario when $\max\{a_{11}, \ldots, a_{s1}\} < C$ for some constant $C > 0$.** In this situation, we can obtain:

$$\text{Bias}^2(\langle \alpha, \Delta_n^{kernel}\alpha\rangle_{\mathcal{H}}) + \text{Var}(\langle \alpha, \Delta_n^{kernel}\alpha\rangle_{\mathcal{H}}) = O\left(\frac{1}{n^2}\right) + O\left(\frac{1}{n\beta_n}\right).$$

  The rate is similar to that in the previous case. The MSE is also monotonically decreasing with respect to β_n. Therefore, we prefer a larger value of β_n as it leads to better outcomes in terms of bias and variance tradeoff. Thus, the optimal choice for β_n remains $O(n^{1/2})$.

- **The scenario when $\max\{a_{11}, \ldots, a_{s1}\} = O(\beta_n)$ and $\max\{a_{12}, \ldots, a_{s2}\} = O(\beta_n)$.** In this situation, we can obtain:

$$\text{Bias}^2(\langle \alpha, \Delta_n^{kernel}\alpha\rangle_{\mathcal{H}}) + \text{Var}(\langle \alpha, \Delta_n^{kernel}\alpha\rangle_{\mathcal{H}}) = O\left(\frac{\beta_n^2}{n^2}\right) + O\left(\frac{1}{n\beta_n}\right).$$

  In this scenario, the optimal rate of β_n is $O(n^{1/3})$.

All these results indicate that the optimality of β_n largely depends on the locations of change points. However, these locations are unknown, making it theoretically impossible to directly determine the optimal value of β_n.

Appendix F. Regularity Conditions and Proofs of the Theorems

F.1 Regularity Conditions of Change Point Detection

To investigate the asymptotic properties, we list the following assumptions.

Assumption 1 *Assume $s = O(n^\gamma)$, $0 \le \gamma < 1$, and there exist constants C_1 and C_2 such that $C_1 n^{1-\gamma} \le \min_j n_j \le \max_j n_j \le C_2 n^{1-\gamma}$.*

Assumption 2 *$EK(X^{(i)}, X^{(i)}) < \infty$, for $i = 1, \ldots, s+1$.*

Assumption 3 *The kernel K is characteristic.*

Assumption 4 $\max_{1\leq i\leq s+1}\left\|\mu_d^{(i)}\right\|_{\mathcal{H}}<\infty$.

Assumption 5 $E\left\|\epsilon_d^{(i)}\right\|_{\mathcal{H}}^2<\infty$, *for* $i=1,\ldots,s+1$.

Assumption 6 $0<\min_{1\leq i\leq n}\lambda_{min}(\Sigma_i)\leq\max_{1\leq i\leq n}\lambda_{max}(\Sigma_i)<\infty$.

Assumption 7 $0<\min_{1\leq i\leq n}\lambda_{min}(\mathrm{Var}((E(Y_i)-\bar{E}Y)\otimes\epsilon_i))\leq\max_{1\leq i\leq n}\lambda_{max}(\mathrm{Var}((E(Y_i)-\bar{E}Y)\otimes\epsilon_i))<\infty$.

Assumption 8 $0<\max_{1\leq i\leq n}\lambda_{max}(E((E(Y_i)-\bar{E}Y)\otimes\epsilon_i))^4)<\infty$.

Let $\epsilon_i=X_i-E(X_i)$.

Assumption 9 $0<\min_{1\leq i\leq n}\lambda_{min}(\Sigma_i)\leq\max_{1\leq i\leq n}\lambda_{max}(\Sigma_i)<\infty$.

Assumption 10 $0<\min_{1\leq i\leq n}\lambda_{min}(\mathrm{Var}(\epsilon_i\epsilon_i^\top))\leq\max_{1\leq i\leq n}\lambda_{max}(\mathrm{Var}(\epsilon_i\epsilon_i^\top))<\infty$.

Assumption 11 $0<\max_{1\leq i\leq n}\lambda_{max}(E(\epsilon_i\epsilon_i^\top-\Sigma_i)^4)<\infty$.

Assumption 12 $0\leq\max_{1\leq k\leq s+1}\left|\alpha^\top\mu^{(k)}\right|<\infty$ *for all* $||\alpha||=1$.

Remark 17 *Assumption 2 guarantees that the kernel mean embedding $\mu_i^\star$ exists and is uniquely defined in $\mathcal{H}$. Assumption 3 further ensures that the distributional change point problem can be transformed into a mean change point problem; see Celisse et al. (2018). Assumption 4 furthermore guarantees that Δ^{kernel} is a Hilbert–Schmidt operator. Assumption 5 establishes the boundedness of the residual term, and Assumptions 6, 7, and 8 are used to ensure the Lyapunov condition so that the central limit theorem can be applied to derive the limiting distribution of Δ_n^{kernel}. These assumptions are about the existence of moments and are satisfied in many situations, such as the cases where Σ_i is an identity operator and m-dependent. These assumptions are also commonly used in the theory of reproducing kernel Hilbert space, see relevant references Lee et al. (2013), Li and Song (2017) and Li (2018). Assumptions 9–12 are used in the proofs of the CPCA-related theorems. These assumptions are satisfied in many settings of interest, such as the case where $\Sigma_i=I_p$ and the m-dependent case; see, for example, Dette et al. (2022) and Chen et al. (2010).*

F.2 Regularity Conditions of Clustering

Before introducing the required conditions, we establish some notation refered to Lu and Zhou (2016). For all $k\in[d]$, let $\mathcal{C}_k^{(s)}$ be the estimated cluster k at iteration s. Define $n_k^{(s)}=\left|\mathcal{C}_k^{(s)}\right|$, $n_{kl}^{(s)}=\left|\mathcal{C}_k\cap\mathcal{C}_l^{(s)}\right|$, $\omega_k=n_k/n$. The mis-clustering rate at iteration s can be written as

$$A_s=\frac{1}{n}\sum_{i=1}^{n}\mathbb{I}\left\{\hat{z}_i^{(s)}\neq z_i\right\}=\frac{1}{n}\sum_{k\neq l\in[d]^2}n_{kl}^{(s)}.$$

We define a cluster-wise mis-clustering rate at iteration s as

$$G_s = \max_{l\in[d]} \left\{ \frac{\sum_{k\neq l\in[d]} n_{kl}^{(s)}}{n_l^{(s)}}, \frac{\sum_{k\neq l\in[d]} n_{lk}^{(s)}}{n_l} \right\}.$$

Let $\eta_2 = \min_{k\neq l\in[d]} \left\|\mu^{(k)} - \mu^{(l)}\right\|$ be the signal strength. For $h \in [d]$, let $\hat{\mu}^{(k,s)}$ be the estimated center of cluster k at iteration s. Define our error rate of estimating centers at iteration s as

$$\Lambda_s = \max_{k\in[d]} \frac{1}{\eta_2} \left\|\hat{\mu}^{(k,s)} - \mu^{(k)}\right\|.$$

Define $\lambda = \max_{k\neq l\in[d]} \left\|\mu^{(k)} - \mu^{(l)}\right\| / \eta_2$ and $\alpha_0 = \min_{k\in d} n_k/n$. Similar to the two-cluster case, we define a normalized signal-to-noise ratio

$$r_d = \frac{\eta_2}{\sigma}\sqrt{\frac{\alpha_0}{1+dp/n}}.$$

Assumption 13 *$n\alpha_0^2 \geq Cd\log n$ and $r_d \geq C\sqrt{d}$ for a sufficiently large constant C. Given any (data dependent) initializer satisfying*

$$G_0 < \left(\frac{1}{2} - \frac{6}{\sqrt{r_d}}\right)\frac{1}{\lambda} \quad \text{or} \quad \Lambda_0 \leq \frac{1}{2} - \frac{4}{\sqrt{r_d}},$$

with probability $1-\nu$.

Assumption 14 *Let $\eta_1 = \max_{k\neq l\in[d]} \left\|\mu^{(k)} - \mu^{(l)}\right\|$ and $\eta_2 = \min_{k\neq l\in[d]} \left\|\mu^{(k)} - \mu^{(l)}\right\|$. Assume that $\eta_2 > 4\sigma\sqrt{\log(n)}$, $\eta_1 = O\left(\sqrt{\log(n)}\right)$ and $\eta_2 = O\left(\sqrt{\log(n)}\right)$.*

Assumption 15 *Assume that*

$$0 < \min_{1\leq i\leq n} \lambda_{min}(\mathrm{Var}((E(X_i) - \bar{E}X)W_i^\top)) \leq \max_{1\leq i\leq n} \lambda_{max}(\mathrm{Var}((E(X_i) - \bar{E}X)W_i^\top)) < \infty.$$

Assumption 16 $0 < \max_{1\leq i\leq n} \lambda_{max}(E((E(X_i) - \bar{E}X)W_i^\top))^4) < \infty$.

Remark 18 *These assumptions are satisfied in many clustering methods, such as K-means and spectral clustering. Assumption 13 corresponds to the requirement of the K-means algorithm for initial classifiers in Lu and Zhou (2016). Assumption 14 corresponds to the requirement on the initial clustering labels for the K-means algorithm in Lu and Zhou (2016). This assumption ensures that the mis-clustering rate of the K-means algorithm less than $1/n$, which guarantees the reliability of the initial values for our subsequent iterative algorithm. This assumption is used to establish the asymptotic normality of Δ_n. If one only aims to prove the consistency of Δ_n, the $\sqrt{\log(n)}$ rate in Assumption 14 can be relaxed to merely diverging to infinity. Additionally, Assumptions 15 and 16 are necessary conditions for obtaining the limit distribution of $\Delta_n - \Delta$. See relevant references in the literature such as Lu and Zhou (2016), Liu et al. (2023) and Jiang et al. (2023).*

F.3 Proof of Theorem 2

For any q_d basis functions $\{v_1, v_2, \ldots, v_{q_d}\}$ of $\mathrm{S}^d_{\{X_i\}_{i=1}^n}$ with $q_d \leq s$, we have $\mathrm{Span}\{v_1, \ldots, v_{q_d}\} = \mathrm{S}^d_{\{X_i\}_{i=1}^n}$. Assume there are $\bar{s}$ change points $1 \leq \bar{z}_1 < \bar{z}_2 < \cdots < \bar{z}_{\bar{s}} \leq n$ of the sequence $\{f(X_i)\}_{i=1}^n$ where $f(X_i) = (\langle v_1, Y_i\rangle_{\mathcal{H}}, \ldots, \langle v_{q_d}, Y_i\rangle_{\mathcal{H}})^\top$ such that $E(f(X_{\bar{z}_{k-1}+j})) = \left(\left\langle v_1, \mu^\star_{\bar{z}_{k-1}+j}\right\rangle_{\mathcal{H}}, \ldots, \left\langle v_{q_d}, \mu^\star_{\bar{z}_{k-1}+j}\right\rangle_{\mathcal{H}}\right)^\top$, for $k = 1, \ldots, \bar{s}+1$, $1 \leq j \leq \bar{z}_k - \bar{z}_{k-1}$ and $\left(\langle v_1, \mu^\star_{\bar{z}_k}\rangle_{\mathcal{H}}, \ldots, \langle v_{q_d}, \mu^\star_{\bar{z}_k}\rangle_{\mathcal{H}}\right) \neq \left(\langle v_1, \mu^\star_{\bar{z}_k+1}\rangle_{\mathcal{H}}, \ldots, \langle v_{q_d}, \mu^\star_{\bar{z}_k+1}\rangle_{\mathcal{H}}\right)$.

If the locations of change points in the sequence $\{f(X_i)\}_{i=1}^n$ are not these in the sequence $\{Y_i\}_{i=1}^n$, there exist k such that $\mu^\star_{\bar{z}_k} = \mu^\star_{\bar{z}_k+1}$. Then we have $\left(\langle v_1, \mu^\star_{\bar{z}_k}\rangle_{\mathcal{H}}, \ldots, \langle v_{q_d}, \mu^\star_{\bar{z}_k}\rangle_{\mathcal{H}}\right) = \left(\langle v_1, \mu^\star_{\bar{z}_k+1}\rangle_{\mathcal{H}}, \ldots, \langle v_{q_d}, \mu^\star_{\bar{z}_k+1}\rangle_{\mathcal{H}}\right)$. However, $\bar{z}_k$ is a change point of the sequence $\{f(X_i)\}_{i=1}^n$, which implies that $\left(\langle v_1, \mu^\star_{\bar{z}_k}\rangle_{\mathcal{H}}, \ldots, \langle v_{q_d}, \mu^\star_{\bar{z}_k}\rangle_{\mathcal{H}}\right) \neq \left(\langle v_1, \mu^\star_{\bar{z}_k+1}\rangle_{\mathcal{H}}, \ldots, \langle v_{q_d}, \mu^\star_{\bar{z}_k+1}\rangle_{\mathcal{H}}\right)$. This is a contradiction. Thus, the locations of change points in the sequence $\{f(X_i)\}_{i=1}^n$ are those in the sequence $\{Y_i\}_{i=1}^n$.

On the other hand, if the locations of change points in the sequence $\{Y_i\}_{i=1}^n$ are not those in the sequence $\{f(X_i)\}_{i=1}^n$, there exists a k such that

$$\left(\langle v_1, \mu^\star_{\bar{z}_k} - \mu^\star_{z_k+1}\rangle_{\mathcal{H}}, \ldots, \langle v_{q_d}, \mu^\star_{\bar{z}_k} - \mu^\star_{z_k+1}\rangle_{\mathcal{H}}\right) = 0 \text{ and } \mu^\star_{z_k} \neq \mu^\star_{z_k+1}.$$

Therefore, $\mu^\star_{z_k} - \mu^\star_{z_k+1}$ is vertical to the subspace $\mathrm{Span}\{v_1, \ldots, v_{q_d}\}$, namely

$$\mu^\star_{z_k} - \mu^\star_{z_k+1} \perp \mathrm{Span}\{v_1, \ldots, v_{q_d}\}. \tag{12}$$

By the definition of kernel mean embedding deviation subspace $\mathrm{S}^d_{\{X_i\}_{i=1}^n}$, we have $\mu^\star_{z_k+1} - \mu^\star_{z_k} \in \mathrm{S}^d_{\{X_i\}_{i=1}^n}$. As $\mathrm{Span}(v_1, \ldots, v_{q_d}) = \mathrm{S}^d_{\{X_i\}_{i=1}^n}$, we conclude that

$$\mu^\star_{z_k} - \mu^\star_{z_k+1} \in \mathrm{Span}\{v_1, \ldots, v_{q_d}\}. \tag{13}$$

Altogether the results in (12) and (13), we conclude that $\mu^\star_{z_k} - \mu^\star_{z_k+1} = 0$. This produces the contradiction that z_k is the location of a change point in $\{Y_i\}_{i=1}^n$, namely, $\mu^\star_{z_k} \neq \mu^\star_{z_k+1}$. Therefore, the locations of change points in the sequence $\{Y_i\}_{i=1}^n$ are these in the sequence $\{f(X_i)\}_{i=1}^n$. The proof is finished. ■

F.4 Proof of Theorem 3

Recall that

$$\Delta^{kernel} = \frac{1}{2}\sum_{i=1}^{s+1}\sum_{j=1}^{s+1} c_i c_j \left(\mu_d^{(i)} - \mu_d^{(j)}\right) \otimes \left(\mu_d^{(i)} - \mu_d^{(j)}\right).$$

First, we prove that:

$$\overline{\mathrm{ran}}(\sum_{i=1}^n A_i \otimes A_i) = \mathrm{Span}\{A_1, \ldots, A_n\}. \tag{14}$$

Since $\overline{\mathrm{ran}}(\sum_{i=1}^n A_i \otimes A_i)$ is the closure of $\{(\sum_{i=1}^n A_i \otimes A_i)f : f \in \mathcal{H}\}$. For any element $\beta \in \overline{\mathrm{ran}}(\sum_{i=1}^n A_i \otimes A_i)$, $\beta = (\sum_{i=1}^n A_i \otimes A_i)f = \sum_{i=1}^n A_i\langle A_i, f\rangle_{\mathcal{H}} \in \mathrm{Span}\{A_1, \ldots, A_n\}$. Therefore, $\overline{\mathrm{ran}}(\sum_{i=1}^n A_i \otimes A_i) \subseteq \mathrm{Span}\{A_1, \ldots, A_n\}$.

If $A_1, A_2, \ldots, A_n$ is linearly independent. Take $f = A_j$, $j = 1, 2, \ldots, n$. We have:

$$\beta_j = (\sum_{i=1}^{n} A_i \otimes A_i)A_j = \sum_{i=1}^{n} A_i \langle A_i, A_j \rangle_{\mathcal{H}}.$$

Let G_A denote the gram matrix of $\{A_1, \ldots, A_n\}$. Then, we have:

$$(\beta_1, \ldots, \beta_s)^\top = G_A (A_1, \ldots, A_n)^\top.$$

$\{A_1, \ldots, A_n\}$ is linearly independent and G_A is a positive definite matrix. Therefore, it follows that $\beta_1, \ldots, \beta_n$ is linearly independent and $\dim(\overline{\text{ran}}(\sum_{i=1}^{n} A_i \otimes A_i)) \geq n$. Together with $\overline{\text{ran}}(\sum_{i=1}^{n} A_i \otimes A_i) \subseteq \text{Span}\{A_1, \ldots, A_n\}$ and $\dim(\text{Span}\{A_1, \ldots, A_n\}) = n$, we get (14).

If $A_1, A_2, \ldots, A_n$ is not linearly independent. We set $\dim(\text{Span}\{A_1, \ldots, A_n\}) = q$ and $n_0 = n - q$. Let $A_1, \ldots, A_q$ be linearly independent and denote the basis of $\text{Span}\{A_1, \ldots, A_n\}$. For $l = 1, \ldots, n_0$ and $j = 1, \ldots, q$, let $A_{q+l} = \sum_{i=1}^{q} c_{li} A_i$ and $f = A_j$. Then,

$$\beta_j = (\sum_{i=1}^{n} A_i \otimes A_i)A_j = \sum_{i=1}^{n} A_i \langle A_i, A_j \rangle_{\mathcal{H}} = \sum_{i=1}^{q} A_i \langle A_i, A_j \rangle_{\mathcal{H}} + \sum_{l=1}^{n_0} \left\{ \sum_{i=1}^{q} c_{li} A_i \left(\sum_{k=1}^{q} c_{lk} \langle A_k, A_j \rangle_{\mathcal{H}} \right) \right\}.$$

Moreover, β_j has the following relationship:

$$(\beta_1, \ldots, \beta_q)^\top = G_A (I + D^\top D)(A_1, \ldots, A_q)^\top,$$

where

$$D = \begin{bmatrix} c_{11} & c_{12} & \ldots & c_{1q} \\ c_{21} & c_{22} & \ldots & c_{2q} \\ \vdots & \vdots & & \vdots \\ c_{n_0 1} & c_{n_0 2} & \ldots & c_{n_0 q} \end{bmatrix}_{n_0 \times q}.$$

Both G_A and $I + D^\top D$ are positive definite matrices. Then, $\det(G_A(I + D^\top D)) > 0$. Therefore, $\beta_1, \ldots, \beta_q$ is linearly independent and $\dim(\overline{\text{ran}}(\sum_{i=1}^{n} A_i \otimes A_i)) \geq q$. Together with $\overline{\text{ran}}(\sum_{i=1}^{n} A_i \otimes A_i) \subseteq \text{Span}\{A_1, \ldots, A_n\}$ and $\dim(\text{Span}\{A_1, \ldots, A_n\}) = q$, we also get the conclusion (14).

Let $B_{ij} = \sqrt{c_i c_j}(\mu_d^{(i)} - \mu_d^{(j)})$. Then, $\Delta^{kernel} = \frac{1}{2} \sum_{i=1}^{s+1} \sum_{j=1}^{s+1} B_{ij} \otimes B_{ij}$. For each $i = 1, \ldots, s+1$ and $j = 1, \ldots, s+1$, let $k = (i-1)(s+1) + j$ and $k = 1, \ldots, (s+1)^2$. Thus, for each B_{ij}, we can let $C_k = B_{ij}$. Furthermore, $\{B_{ij}, i = 1, \ldots, s+1, j = 1, \ldots, s+1\}$ is equivalent to $\{C_k, k = 1, \ldots, (s+1)^2\}$. According to (14), we have:

$$\overline{\text{ran}} \left(\sum_{k=1}^{(s+1)^2} C_k \otimes C_k \right) = \text{Span} \left\{ C_1, \ldots, C_{(s+1)^2} \right\}.$$

It suggests that:

$$\overline{\text{ran}} \left(\sum_{i=1}^{s+1} \sum_{j=1}^{s+1} B_{ij} \otimes B_{ij} \right) = \text{Span} \left\{ B_{ij}, i = 1, \ldots, s+1, j = 1, \ldots, s+1 \right\}.$$

By the definition of kernel mean embedding deviation subspace $\text{S}^d_{\{X_i\}_{i=1}^n}$, $\text{Span}\{B_{ij}, i = 1, \ldots, s+1, j = 1, \ldots, s+1\} = \text{S}^d_{\{X_i\}_{i=1}^n}$. Therefore, we can get $\overline{\text{ran}}(\Delta^{kernel}) = \text{S}^d_{\{X_i\}_{i=1}^n}$. The proof is finished. ■

F.5 Proof of Theorem 6

Before proving Theorem 6 and Theorem 7, we first present a few lemmas related to Hilbert-Schmidt operators and the Hilbert-Schmidt norm.

Lemma 19 *Suppose Assumptions 1–4 hold, then Δ^{kernel} is a Hilbert-chmidt operator.*

The proof of Lemma 19. Under Assumptions 2 and 3, the existence of $\mu_d^{(i)}$ within $\mathcal{H}$ is guaranteed. Additionally, the Hilbert-Schmidt norm of Δ^{kernel} can be determined as follows:

$$\begin{aligned}
\left\|\Delta^{kernel}\right\|_{HS} &= \left\| \frac{1}{2} \sum_{i=1}^{s+1} \sum_{j=1}^{s+1} c_i c_j \left(\mu_d^{(i)} - \mu_d^{(j)}\right) \otimes \left(\mu_d^{(i)} - \mu_d^{(j)}\right) \right\|_{HS} \\
&\leq \frac{1}{2} \sum_{i=1}^{s+1} \sum_{j=1}^{s+1} c_i c_j \left\| \left(\mu_d^{(i)} - \mu_d^{(j)}\right) \otimes \left(\mu_d^{(i)} - \mu_d^{(j)}\right) \right\|_{HS} \\
&= \frac{1}{2} \sum_{i=1}^{s+1} \sum_{j=1}^{s+1} c_i c_j \left(\langle \mu_d^{(i)}, \mu_d^{(i)} \rangle_{\mathcal{H}} - \langle \mu_d^{(i)}, \mu_d^{(j)} \rangle_{\mathcal{H}} - \langle \mu_d^{(j)}, \mu_d^{(i)} \rangle_{\mathcal{H}} + \langle \mu_d^{(j)}, \mu_d^{(j)} \rangle_{\mathcal{H}} \right) \\
&\leq \frac{(s+1)^2 \max_j n_j^2}{n^2} = O(1).
\end{aligned}$$

Therefore, it can be concluded that the Hilbert-Schmidt norm of Δ^{kernel} exists, indicating that Δ^{kernel} itself is a Hilbert-Schmidt operator.

Lemma 20 *For $\forall f, g \in \mathcal{H}$, the Hilbert-Schmidt norm of $\|f \otimes g\|_{HS}$ is given by*

$$\|f \otimes g\|_{HS}^2 = \|f\|_{\mathcal{H}}^2 \|g\|_{\mathcal{H}}^2.$$

Lemma 21 *Suppose Assumptions 1–5 hold, we have*

$$E \left\| \frac{1}{n} \sum_{i=1}^{n} \epsilon_i \right\|_{\mathcal{H}}^2 = O\left(\frac{1}{n}\right) \text{ and } E \left\| \frac{1}{n} \sum_{i=1}^{n} \mu_i^\star \right\|_{\mathcal{H}}^2 = O\left(1\right).$$

The proof of Lemma 20 and Lemma 21. For the Hilbert-Schmidt norm of $\|f \otimes g\|_{HS}$, we have

$$\|f \otimes g\|_{HS}^2 = \langle f \otimes g, f \otimes g \rangle_{HS} = \langle f, (f \otimes g) g \rangle_{\mathcal{H}} = \langle f, f \rangle_{\mathcal{H}} \langle g, g \rangle_{\mathcal{H}} = \|f\|_{\mathcal{H}}^2 \|g\|_{\mathcal{H}}^2.$$

Since $\{\epsilon_i\}_{i=1}^n$ are independent, we have

$$E \left\| \frac{1}{n} \sum_{i=1}^{n} \epsilon_i \right\|_{\mathcal{H}}^2 = \frac{1}{n} E \left\| \epsilon_i \right\|_{\mathcal{H}}^2 = O\left(\frac{1}{n}\right).$$

As $\mu_i^\star$ is a constant element in $\mathcal{H}$, we have

$$E \left\| \frac{1}{n} \sum_{i=1}^{n} \mu_i^\star \right\|_{\mathcal{H}}^2 = \left\| \frac{1}{n} \sum_{i=1}^{n} \mu_i \right\|_{\mathcal{H}}^2 = \frac{1}{n^2} \sum_{i=1}^{n} \sum_{j=1}^{n} \langle \mu_i^\star, \mu_j^\star \rangle \leq \max_{1 \leq i,j \leq n} \langle \mu_i^\star, \mu_j^\star \rangle = O\left(1\right).$$

The proof of Lemma 20 and Lemma 21 is completed. ■

The proof of Theorem 6. We divide $\left\|\Delta_n^{kernel} - \Delta^{kernel}\right\|_{HS}$ into two parts:

$$\left\|\Delta_n^{kernel} - \Delta^{kernel}\right\|_{HS} = \left\|\Sigma_n^{kernel} - \Sigma^{kernel} - (\Sigma_{pooled,n}^{kernel} - \Sigma_{pooled}^{kernel})\right\|_{HS},$$

where $\Sigma^{kernel} = \Sigma_{pooled}^{kernel} + \Delta^{kernel}$. Let $\bar{E}Y = \sum_{l=1}^{s+1} c_l \mu_d^{(l)}$, then we have:

$$\begin{aligned}
\Sigma_n^{kernel} &= \frac{1}{n}\sum_{i=1}^{n}\left(Y_i - \bar{Y}\right)\otimes\left(Y_i - \bar{Y}\right) \\
&= \frac{1}{n}\sum_{i=1}^{n}(Y_i - \bar{E}Y)\otimes(Y_i - \bar{E}Y) + \frac{1}{n}\sum_{i=1}^{n}(Y_i - \bar{E}Y)\otimes(\bar{E}Y - \bar{Y}) \\
&\quad + \frac{1}{n}\sum_{i=1}^{n}(\bar{E}Y - \bar{Y})\otimes(Y_i - \bar{E}Y) + \frac{1}{n}\sum_{i=1}^{n}(\bar{E}Y - \bar{Y})\otimes(\bar{E}Y - \bar{Y}) \\
&= H_1 + H_2 + H_3 + H_4.
\end{aligned} \tag{15}$$

We examine the term H_1, which can be decomposed as follows:

$$\begin{aligned}
H_1 &= \frac{1}{n}\sum_{i=1}^{n}(Y_i - \bar{E}Y)\otimes(Y_i - \bar{E}Y) \\
&= \frac{1}{n}\sum_{i=1}^{n}(Y_i - E(Y_i))\otimes(Y_i - E(Y_i)) + \frac{1}{n}\sum_{i=1}^{n}(E(Y_i) - \bar{E}Y)\otimes(Y_i - E(Y_i)) \\
&\quad + \frac{1}{n}\sum_{i=1}^{n}(Y_i - E(Y_i))\otimes(E(Y_i) - \bar{E}Y) + \frac{1}{n}\sum_{i=1}^{n}(E(Y_i) - \bar{E}Y)\otimes(E(Y_i) - \bar{E}Y) \\
&= L_1 + L_2 + L_3 + \Delta^{kernel}.
\end{aligned} \tag{16}$$

Therefore, we decompose $\Sigma_n^{kernel} - \Sigma^{kernel}$ as follows:

$$\Sigma_n^{kernel} - \Sigma^{kernel} = L_1 + L_2 + L_3 + H_2 + H_3 + H_4 - \Sigma_{pooled}^{kernel}.$$

Decompose $\Sigma_{pooled,n}^{kernel} - \Sigma_{pooled}^{kernel}$ as:

$$\begin{aligned}
\Sigma_{pooled,n}^{kernel} - \Sigma_{pooled}^{kernel} &= \frac{1}{r}\sum_{m=1}^{r}(\hat{\Sigma}_m^{kernel} - \Sigma_{pooled}^{kernel}) = (L_1 - \Sigma_{pooled}^{kernel}) - \frac{1}{r}\sum_{m=1}^{r}\frac{1}{\beta_n(\beta_n-1)}\sum_{k\in\mathcal{S}_m}\sum_{l\in\mathcal{S}_m}\epsilon_k\otimes\epsilon_l \\
&\quad + \frac{1}{r}\sum_{m=1}^{r}\frac{1}{\beta_n(\beta_n-1)}\sum_{k\in\mathcal{S}_m}\sum_{l\in\mathcal{S}_m}\delta_{kl}\otimes(\epsilon_k - \epsilon_l) + \frac{1}{r}\sum_{m=1}^{r}\frac{1}{2\beta_n(\beta_n-1)}\sum_{k\in\mathcal{S}_m}\sum_{l\in\mathcal{S}_m}\delta_{kl}\otimes\delta_{kl} \\
&= (L_1 - \Sigma_{pooled}^{kernel}) - G_1 + G_2 + G_3.
\end{aligned} \tag{17}$$

Therefore, we further decompose $\Delta_n^{kernel} - \Delta^{kernel}$ as follows:

$$\begin{aligned}
\Delta_n^{kernel} - \Delta^{kernel} &= \Sigma_n^{kernel} - \Sigma^{kernel} - (\Sigma_{pooled,n}^{kernel} - \Sigma_{pooled}^{kernel}) \\
&= L_2 + L_3 + H_2 + H_3 + H_4 + G_1 - G_2 - G_3.
\end{aligned} \tag{18}$$

Next, we analyze each term in (18). For L_2 and L_3, we have

$$
\begin{aligned}
E\left\|L_2\right\|_{HS} = E\left\|L_3\right\|_{HS} &= E\left\{\left\|\frac{1}{n}\sum_{i=1}^{n}(Y_i - E(Y_i)) \otimes (E(Y_i) - \bar{E}Y)\right\|_{HS}\right\} \\
&\leq \left(\sum_{j=1}^{s+1} E\left\{\left\|\frac{1}{n}\sum_{i=z_{j-1}+1}^{z_j}(Y_i - \mu_d^{(j)})\right\|_{\mathcal{H}}\right\}\right)\left\|\sum_{l=1}^{s+1} c_l(\mu_d^{(j)} - \mu_d^{(l)})\right\|_{\mathcal{H}} \\
&\leq \sum_{j=1}^{s+1} c_j\left(E\left\|\frac{1}{n_j}\sum_{i=z_{j-1}+1}^{z_j}(Y_i - \mu_d^{(j)})\right\|_{\mathcal{H}}^2\right)^{1/2}\left\|\sum_{l=1}^{s+1} c_l(\mu_d^{(j)} - \mu_d^{(l)})\right\|_{\mathcal{H}} \\
&\leq \sum_{j=1}^{s+1} c_j O\left(s n_j^{-1/2}\right) \leq O\left(s^2\frac{\max_j n_j^{1/2}}{n}\right) = O\left(n^{(3\gamma-1)/2}\right).
\end{aligned}
$$

Based on Chebyshev's inequality, we can get that $\left\|L_2\right\|_{HS} = \left\|L_3\right\|_{HS} = O_p\left(n^{(3\gamma-1)/2}\right)$. Regarding H_2, we can draw the following conclusion:

$$
\begin{aligned}
E\left\|H_2\right\|_{HS} &= E\left\|\frac{1}{n}\sum_{i=1}^{n}(Y_i - E(Y_i) + E(Y_i) - \bar{E}Y) \otimes (\bar{E}Y - \bar{Y})\right\|_{HS} \\
&= E\left\|\frac{1}{n}\sum_{i=1}^{n}(Y_i - E(Y_i)) \otimes (\bar{E}Y - \bar{Y})\right\|_{HS} = E\left\{\left\|\frac{1}{n}\sum_{i=1}^{n}\epsilon_i\right\|_{\mathcal{H}}\left\|\frac{1}{n}\sum_{i=1}^{n}\epsilon_i\right\|_{\mathcal{H}}\right\} \\
&\leq \left(E\left\|\frac{1}{n}\sum_{i=1}^{n}\epsilon_i\right\|_{\mathcal{H}}^2\right)^{1/2}\left(E\left\|\frac{1}{n}\sum_{i=1}^{n}\epsilon_i\right\|_{\mathcal{H}}^2\right)^{1/2} = O\left(\frac{1}{n}\right).
\end{aligned}
$$

Similar as H_2, we can also deduce that: $\left\|H_3\right\|_{HS} = O_p\left(\frac{1}{n}\right)$. In the term of H_4, we can derive the following conclusion:

$$
\begin{aligned}
E\left\|H_4\right\|_{HS} &= E\left\|\frac{1}{n}\sum_{i=1}^{n}(\bar{E}Y - \bar{Y}) \otimes (\bar{E}Y - \bar{Y})\right\|_{HS} \\
&= E\left\|(\bar{E}Y - \bar{Y}) \otimes (\bar{E}Y - \bar{Y})\right\|_{HS} = E\left\{\left\|\frac{1}{n}\sum_{i=1}^{n}\epsilon_i\right\|_{\mathcal{H}}\left\|\frac{1}{n}\sum_{i=1}^{n}\epsilon_i\right\|_{\mathcal{H}}\right\} \\
&\leq \left(E\left\|\frac{1}{n}\sum_{i=1}^{n}\epsilon_i\right\|_{\mathcal{H}}^2\right)^{1/2}\left(E\left\|\frac{1}{n}\sum_{i=1}^{n}\epsilon_i\right\|_{\mathcal{H}}^2\right)^{1/2} = O\left(\frac{1}{n}\right).
\end{aligned}
$$

Let z_j belong to $\mathcal{S}_{m_j}$, where $j = 0, \ldots, s$. It follows that

$$\begin{aligned} E\left\|G_1\right\|_{HS} &= E\left\|\frac{1}{r}\sum_{m=1}^{r}\frac{1}{\beta_n(\beta_n-1)}\sum_{k\in\mathcal{S}_m}\sum_{l\in\mathcal{S}_m}\epsilon_k\otimes\epsilon_l\right\|_{HS} \\ &= \frac{1}{n(\beta_n-1)}\left\{\left(\sum_{j=1}^{s}\sum_{m=(m_{j-1}-1)}^{m_j-1}E\left\|\sum_{k\in\mathcal{S}_m}\sum_{l\in\mathcal{S}_m}\epsilon_k\otimes\epsilon_l\right\|_{HS}\right)+\sum_{j=0}^{s}E\left\|\sum_{k\in\mathcal{S}_{m_j}}\sum_{l\in\mathcal{S}_{m_j}}\epsilon_k\otimes\epsilon_l\right\|_{HS}\right\} \\ &= G_{11}+G_{12}. \end{aligned} \tag{19}$$

For any $\mathcal{S}_m$, we have

$$E\left\{\left\|\sum_{k\in\mathcal{S}_m}\sum_{l\in\mathcal{S}_m}\epsilon_k\otimes\epsilon_l\right\|_{HS}\right\}\leq\beta_n^2\left(E\left\|\frac{1}{\beta_n}\sum_{k\in\mathcal{S}_m}\epsilon_k\right\|_{\mathcal{H}}^2\right)^{1/2}\left(E\left\|\frac{1}{\beta_n}\sum_{l\in\mathcal{S}_m}\epsilon_l\right\|_{\mathcal{H}}^2\right)^{1/2}=O\left(\beta_n\right).$$

Since $E\left\|\sum_{k\in\mathcal{S}_{m_{j-1}-1}}\sum_{l\in\mathcal{S}_{m_{j-1}-1}}\epsilon_k\otimes\epsilon_l\right\|_{HS}=\cdots=E\left\|\sum_{k\in\mathcal{S}_{m_j}}\sum_{l\in\mathcal{S}_{m_j}}\epsilon_k\otimes\epsilon_l\right\|_{HS}$, we have

$$\begin{aligned} EG_{11} =& \frac{1}{n(\beta_n-1)}\sum_{j=1}^{s}\sum_{m=(m_{j-1}-1)}^{m_j-1}E\left\|\sum_{k\in\mathcal{S}_m}\sum_{l\in\mathcal{S}_m}\epsilon_k\otimes\epsilon_l\right\|_{HS} \\ =& \frac{1}{n(\beta_n-1)}\sum_{j=1}^{s}(m_j-m_{j-1}-1)E\left\|\sum_{k\in\mathcal{S}_m}\sum_{l\in\mathcal{S}_m}\epsilon_k\otimes\epsilon_l\right\|_{HS} \\ \leq& O_p\left(\frac{1}{n(\beta_n-1)}(s+1)^2\frac{\max_j n_j}{\beta_n}\beta_n\right)=O_p\left(\frac{n^\gamma}{\beta_n}\right). \end{aligned}$$

Similar to G_{11}, we have $G_{12}=O_p\left(n^{2\gamma-1}\right)$. For G_2, we have:

$$G_2=\left\|\frac{1}{r}\frac{1}{\beta_n(\beta_n-1)}\sum_{j=0}^{s}\sum_{k\in\mathcal{S}_{m_j}}\sum_{l\in\mathcal{S}_{m_j}}\delta_{kl}\otimes(\epsilon_k-\epsilon_l)\right\|_{HS}\leq F_1+F_2+F_3+F_4.$$

Regarding F_1, we can deduce the following conclusion:

$$EF_1\leq\frac{1}{n}\sum_{j=0}^{s}\sum_{k\in\mathcal{S}_{m_j}}E\left(\left\|\mu_k^\star\otimes\epsilon_k\right\|_{HS}\right)=\frac{1}{n}\sum_{j=0}^{s}\sum_{k\in\mathcal{S}_{m_j}}E\left(\left\|\mu_k^\star\right\|_{\mathcal{H}}\left\|\epsilon_k\right\|_{\mathcal{H}}\right)=O\left(\beta_n n^{2\gamma-1}\right).$$

Regarding F_2, we can draw the following conclusion:

$$EF_2\leq\frac{\beta_n}{n}\sum_{j=0}^{s}\left(E\left\|\frac{1}{\beta_n}\sum_{k\in\mathcal{S}_{m_j}}\mu_k^\star\right\|_{\mathcal{H}}^2\right)^{1/2}\left(E\left\|\frac{1}{\beta_n}\sum_{l\in\mathcal{S}_{m_j}}\epsilon_l\right\|_{\mathcal{H}}^2\right)^{1/2}=O\left(n^{2\gamma-1}\right).$$

Similar as F_1 and F_2, we have $F_3=O_p\left(n^{2\gamma-1}\right)$ and $F_4=O_p\left(\beta_n n^{2\gamma-1}\right)$. Regarding G_3, we can draw the following conclusion:

$$EG_3=\frac{\beta_n}{2n}\sum_{j=0}^{s}E\left\{\frac{1}{\beta_n(\beta_n-1)}\sum_{k\in\mathcal{S}_{m_j}}\sum_{l\in\mathcal{S}_{m_j}}\left\|\delta_{kl}\right\|_{\mathcal{H}}\left\|\delta_{kl}\right\|_{\mathcal{H}}\right\}=O\left(\beta_n n^{2\gamma-1}\right).$$

Given that $\beta_n = O(n^m)$ and $r = O\left(n^{1-m}\right)$, we can summarize all the obtained results as follows:

$$||\Delta_n^{kernel} - \Delta^{kernel}||_{HS} = O_p\left(n^{(3\gamma-1)/2} + n^{\gamma-m} + n^{2\gamma+m-1}\right).$$

Adopting the similar description as the justification of Corollary 3 in Li and Song (2017), we can conclude that the eigenspaces of Δ_n^{kernel} converge to those of Δ^{kernel} at the same rate, namely:

$$||\hat{P}_k - P_k||_{HS} = O_p\left(n^{(3\gamma-1)/2} + n^{\gamma-m} + n^{2\gamma+m-1}\right).$$

When $\gamma = 0$ and $m = \frac{1}{2}$, we have $||\Delta_n^{kernel} - \Delta^{kernel}||_{HS} = O_p\left(\sqrt{\frac{1}{n}}\right)$ and $||\hat{P}_k - P_k||_{HS} = O_p\left(\sqrt{\frac{1}{n}}\right)$. The proof of Theorem 6 is completed. ■

F.6 Proof of Theorem 7

Based on the result of the proof of Theorem 6, we can divide $\Delta_n^{kernel} - \Delta^{kernel}$ into two parts: $\Delta_n^{kernel} - \Delta^{kernel} = \Sigma_n^{kernel} - \Sigma^{kernel} - (\Sigma_{pooled,n}^{kernel} - \Sigma_{pooled}^{kernel})$. We decompose $\Sigma_n^{kernel} - \Sigma^{kernel}$ as follows: $\Sigma_n^{kernel} - \Sigma^{kernel} = L_1 + L_2 + L_3 + H_2 + H_3 + H_4 - \Sigma_{pooled}^{kernel}$, where L_1, L_2, L_3, H_2, H_3, and H_4 are defined in (15). Decompose $\Sigma_{pooled,n}^{kernel} - \Sigma_{pooled}^{kernel}$ as: $\Sigma_{pooled,n}^{kernel} - \Sigma_{pooled}^{kernel} = (L_1 - \Sigma_{pooled}^{kernel}) - G_1 + G_2 + G_3$, where G_1, G_2, and G_3 are defined in (17). Therefore, we further decompose $\Delta_n^{kernel} - \Delta^{kernel}$ as follows:

$$\begin{aligned}\Delta_n^{kernel} - \Delta^{kernel} =& \Sigma_n^{kernel} - \Sigma^{kernel} - (\Sigma_{pooled,n}^{kernel} - \Sigma_{pooled}^{kernel}) \\ =& L_2 + L_3 + H_2 + H_3 + H_4 + G_1 - G_2 - G_3.\end{aligned}$$

In the investigation of the asymptotic distribution of $\langle \alpha, (L_2 + L_3)\alpha \rangle_{\mathcal{H}}$, it is straightforward to obtain $\langle \alpha, L_2\alpha \rangle_{\mathcal{H}} = \langle \alpha, L_3\alpha \rangle_{\mathcal{H}}$. Consequently, we have $\langle \alpha, (L_2 + L_3)\alpha \rangle_{\mathcal{H}} = 2\langle \alpha, L_2\alpha \rangle_{\mathcal{H}}$. Here, we set $L_2 = \frac{1}{n}\sum_{i=1}^n (E(Y_i) - \bar{E}Y) \otimes \epsilon_i$ and $Z_i = \frac{1}{\sqrt{n}}\langle \alpha, \left((E(Y_i) - \bar{E}Y) \otimes \epsilon_i\right)\alpha \rangle_{\mathcal{H}}$ as well as $\langle \alpha, L_2\alpha \rangle_{\mathcal{H}} = \frac{1}{\sqrt{n}}\sum_{i=1}^n Z_i$. By exploiting the properties of the covariance operator, we can derive the following result:

$$E(Z_i) = 0, \quad \mathrm{Var}(Z_i) = \frac{1}{n}\mathrm{Var}(\langle \alpha, \left((E(Y_i) - \bar{E}Y) \otimes \epsilon_i\right)\alpha \rangle_{\mathcal{H}}),$$

and

$$B^2 = \mathrm{Var}\left(\sum_{i=1}^n Z_i\right) = \frac{1}{n}\sum_{i=1}^n \mathrm{Var}(\langle \alpha, \left((E(Y_i) - \bar{E}Y) \otimes \epsilon_i\right)\alpha \rangle_{\mathcal{H}}).$$

To verify the Lindeberg condition, we have the following conclusion for any $\eta > 0$:

$$\begin{aligned}&\sum_{i=1}^n \frac{1}{B^2}\int_{|Z_i - E(Z_i)| > \eta B} (Z_i - E(Z_i))^2 dF_i \\ \leq& \frac{n}{B^2}\max_i E\left\{(Z_i - E(Z_i))^4\right\}^{1/2}\left\{P(|Z_i - E(Z_i)| > \eta B)\right\}^{1/2} \\ \leq& \max_i \frac{E\left\{(\langle \alpha, ((E(Y_i) - \bar{E}Y) \otimes \epsilon_i)\alpha \rangle_{\mathcal{H}})^4\right\}^{1/2}}{B^2}\left\{\frac{\mathrm{Var}(Z_i)}{\eta^2 B^2}\right\}^{1/2} \\ \leq& \frac{\max_i \lambda_{max}^{1/2}(E((E(Y_i) - \bar{E}Y) \otimes \epsilon_i)^4)\max_i \lambda_{max}^{1/2}(\mathrm{Var}((E(Y_i) - \bar{E}Y) \otimes \epsilon_i)}{\left\{\min_i \lambda_{min}(\mathrm{Var}((E(Y_i) - \bar{E}Y) \otimes \epsilon_i))\right\}^{3/2}\eta\sqrt{n}} = O\left(\frac{1}{\sqrt{n}}\right) \to 0.\end{aligned}$$

Here, $E((E(Y_i) - \bar{E}Y) \otimes \epsilon_i)^4$ and $\mathrm{Var}((E(Y_i) - \bar{E}Y) \otimes \epsilon_i)$ fulfill the following conditions:

$$\langle \alpha, (E((E(Y_i) - \bar{E}Y) \otimes \epsilon_i)^4)\alpha \rangle_{\mathcal{H}} = E\left\{ (\langle \alpha, ((E(Y_i) - \bar{E}Y) \otimes \epsilon_i)\alpha \rangle_{\mathcal{H}})^4 \right\},$$
$$\langle \alpha, (\mathrm{Var}((E(Y_i) - \bar{E}Y) \otimes \epsilon_i))\alpha \rangle_{\mathcal{H}} = \mathrm{Var}(\langle \alpha, ((E(Y_i) - \bar{E}Y) \otimes \epsilon_i)\alpha \rangle_{\mathcal{H}}).$$

Hence we have the following conclusion:

$$\sum_{i=1}^{n} Z_i \xrightarrow{\mathcal{D}} \mathcal{N}\left(0, \sum_{i=1}^{s+1} c_i \mathrm{Var}(\langle \alpha, ((\mu_d^{(i)} - \bar{E}Y) \otimes \epsilon_d^{(i)})\alpha \rangle_{\mathcal{H}})\right).$$

Based on the previous derivations and decompositions, we can now determine the asymptotic distribution as follows:

$$\sqrt{n}\langle \alpha, (L_2 + L_3)\alpha \rangle_{\mathcal{H}} \xrightarrow{\mathcal{D}} \mathcal{N}\left(0, \sum_{i=1}^{s+1} 4c_i \mathrm{Var}(\langle \alpha, ((\mu_d^{(i)} - \bar{E}Y) \otimes \epsilon_d^{(i)})\alpha \rangle_{\mathcal{H}})\right).$$

Based on the result of the proof of Theorem 6, we have $\|H_2\|_{HS} = \|H_3\|_{HS} = \|H_4\|_{HS} = o_p\left(\frac{1}{\sqrt{n}}\right)$. To evaluate $\langle \alpha, G_1\alpha \rangle_{\mathcal{H}}$, we consider

$$G_1 = \frac{1}{r}\sum_{m=1}^{r} \frac{1}{\beta_n(\beta_n - 1)} \sum_{k \in \mathcal{S}_m} \sum_{l \in \mathcal{S}_m} \epsilon_k \otimes \epsilon_l.$$

We set

$$T_m = \frac{1}{\sqrt{r\beta_n(\beta_n - 1)}} \sum_{k \in \mathcal{S}_m} \sum_{l \neq k, l \in \mathcal{S}_m} \langle \alpha, (\epsilon_k \otimes \epsilon_l)\alpha \rangle_{\mathcal{H}}.$$

Subsequently, we establish the relationship $G_1 = \frac{1}{\sqrt{r\beta_n(\beta_n - 1)}} \sum_{m=1}^{r} T_m$, where $E(T_m) = 0$. Additionally, we have

$$\begin{aligned}
\mathrm{Var}(T_m) =& E(T_m^2) = \left\{ \frac{1}{\sqrt{r\beta_n(\beta_n - 1)}} \sum_{k \in \mathcal{S}_m} \sum_{l \neq k, l \in \mathcal{S}_m} \langle \alpha, (\epsilon_k \otimes \epsilon_l)\alpha \rangle_{\mathcal{H}} \right\}^2 \\
=& \frac{2}{r\beta_n(\beta_n - 1)} \sum_{k \in \mathcal{S}_m} \sum_{l \neq k, l \in \mathcal{S}_m} E\{(\langle \alpha, \epsilon_k \rangle_{\mathcal{H}})^2\} E\{(\langle \alpha, \epsilon_l \rangle_{\mathcal{H}})^2\} \\
=& \frac{2}{r\beta_n(\beta_n - 1)} \sum_{k \in \mathcal{S}_m} \sum_{l \neq k, l \in \mathcal{S}_m} \langle \alpha, \Sigma_k \alpha \rangle_{\mathcal{H}} \langle \alpha, \Sigma_l \alpha \rangle_{\mathcal{H}}.
\end{aligned}$$

Consequently, we derive: $\frac{2}{r}\min \lambda_{min}^2(\Sigma_i) \leq \mathrm{Var}(T_m) \leq \frac{2}{r}\max \lambda_{max}^2(\Sigma_i)$, and $2\min \lambda_{min}^2(\Sigma_i) \leq \sum_{m=1}^{r} \mathrm{Var}(T_m) \leq 2\max \lambda_{max}^2(\Sigma_i)$. In addition, we have:

$$\begin{aligned}
E\left(T_m^4\right) =& \frac{C_1}{\{r\beta_n(\beta_n - 1)\}^2} \sum_{k \in \mathcal{S}_m} \sum_{l \neq k, l \in \mathcal{S}_m} E(\langle \alpha, \epsilon_k \rangle_{\mathcal{H}}^4) E(\langle \epsilon_l, \alpha \rangle_{\mathcal{H}}^4) \\
&+ \frac{C_2}{\{r\beta_n(\beta_n - 1)\}^2} \sum_{k_1 \in \mathcal{S}_m} \sum_{\substack{l_1 \neq k_1 \\ l_1 \in \mathcal{S}_m}} \sum_{\substack{k_2 \neq k_1 \neq l_1 \\ k_2 \in \mathcal{S}_m}} \sum_{\substack{l_2 \neq l_1 \neq k_1 \neq l_1 \\ l_2 \in \mathcal{S}_m}} E(\langle \alpha, \epsilon_{k_1} \rangle_{\mathcal{H}}^2) E(\langle \alpha, \epsilon_{l_1} \rangle_{\mathcal{H}}^2) E(\langle \alpha, \epsilon_{k_2} \rangle_{\mathcal{H}}^2) E(\langle \alpha, \epsilon_{l_2} \rangle_{\mathcal{H}}^2) \\
=& O\left(\frac{1}{r^2}\right),
\end{aligned}$$

where C_1 and C_2 are positive integers that don't take a lot of effort to calculate. Then we verify the Lindeberg condition for any $\eta > 0$

$$\begin{aligned}
&\sum_{m=1}^{r} \frac{1}{\mathrm{Var}\left(\sum_{m=1}^{r} T_m\right)} \int_{|T_m| > \sqrt{\eta \mathrm{Var}\left(\sum_{m=1}^{r} T_m\right)}} T_m^2 dF'_m \\
&\leq \frac{1}{\mathrm{Var}\left(\sum_{m=1}^{r} T_m\right)} \sum_{m=1}^{r} E\left(T_m^4\right)^{1/2} P^{1/2}\left(|T_m| > \sqrt{\eta \mathrm{Var}\left(\sum_{m=1}^{r} T_m\right)}\right) \\
&\leq \frac{r \max_{1\leq m\leq r} E\left(T_m^4\right)^{1/2} \max_{1\leq m\leq r} \mathrm{Var}(T_m)^{1/2}}{\mathrm{Var}\left(\sum_{m=1}^{r} T_m\right)^{3/2} \sqrt{\eta}} = O\left(\frac{1}{\sqrt{\eta r}}\right) \to 0.
\end{aligned}$$

Based on the preceding analysis, we can derive the following conclusion: $\langle \alpha, G_1 \alpha \rangle_{\mathcal{H}} = O_p\left(\frac{1}{\sqrt{n\beta_n}}\right)$. Based on the proof of Theorem 6, we have $\|G_2\|_{HS} = \|G_3\|_{HS} = o_p\left(\frac{1}{\sqrt{n}}\right)$. By applying the Weyl inequality, we can derive the subsequent conclusion for any operator M:

$$\langle \alpha, M\alpha \rangle_{\mathcal{H}} \leq \max \lambda_i \leq \|M\|_{HS},$$

where λ_i is the eigenvalue of M. Consequently, the ensuing conclusion can be drawn for the kernel operator Δ_n^{kernel}:

$$\sqrt{n}\langle \alpha, (\Delta_n^{kernel} - \Delta^{kernel})\alpha) \rangle_{\mathcal{H}} \xrightarrow{\mathcal{D}} \mathcal{N}\left(0, \sum_{i=1}^{s+1} 4c_i \mathrm{Var}(\langle \alpha, ((\mu_d^{(i)} - \bar{E}Y) \otimes \epsilon_d^{(i)})\alpha \rangle_{\mathcal{H}})\right).$$

Moreover, we consider the eigenvalues $\hat{\lambda}_j$, λ_j and the corresponding eigenfunctions $\hat{v}_j$, v_j of Δ_n^{kernel} and Δ^{kernel}, respectively. Define the projection operators as $\hat{P}_j = \hat{v}_j \otimes \hat{v}_j$, $P_j = v_j \otimes v_j$, and $\Delta_{0n} = \Delta_n^{kernel} - \Delta^{kernel}$. Since $(\hat{P}_j \Delta_n^{kernel} \hat{P}_j - \lambda_j \hat{P}_j)\hat{v}_j = (\hat{\lambda}_j - \lambda_j)\hat{v}_j$, then $\hat{\lambda}_j - \lambda_j$ can be seen as the eigenvalues of $\hat{P}_j \Delta_n^{kernel} \hat{P}_j - \lambda_j \hat{P}_j$ with the eigenfunction $\hat{v}_j$. In this context, we can establish the following decomposition:

$$\begin{aligned}
&\hat{P}_j \Delta_n^{kernel} \hat{P}_j - \lambda_j \hat{P}_j = \hat{P}_j \Delta_{0n} \hat{P}_j + \hat{P}_j (\Delta^{kernel} - \lambda_j I)\hat{P}_j \\
&= P_j \Delta_{0n} P_j + (\hat{P}_j - P_j)\Delta_{0n}\hat{P}_j + P_j \Delta_{0n}(\hat{P}_j - P_j) + \hat{P}_j(\Delta^{kernel} - \lambda_j I)\hat{P}_j = O_1 + O_2 + O_3 + O_4.
\end{aligned}$$

Based on the Theorem 9.1.1 of Hsing and Eubank (2015), we have

$$\hat{P}_j - P_j = \sum_{\lambda_k \neq \lambda_j} \frac{1}{\lambda_j - \lambda_k}(P_k \Delta_{0n} P_j + P_j \Delta_{0n} P_k) + O_p\left(\|\Delta_{0n}\|_{HS}^2\right).$$

Hence, we get $\|O_1\|_{HS} = O_p\left(\|\Delta_{0n}\|_{HS}\right) = O_p\left(\frac{1}{\sqrt{n}}\right)$, $\|O_2\|_{HS} = O_p\left(\frac{1}{n}\right)$, $\|O_3\|_{HS} = O_p\left(\frac{1}{n}\right)$, and $\|O_4\|_{HS} = O_p\left(\frac{1}{n}\right)$ and O_1 is the main order term. Based on this, we can draw the following conclusion:

$$\sqrt{n}(\hat{\lambda}_j - \lambda_j) \xrightarrow{\mathcal{D}} \langle v_j, \tilde{\Delta} v_j \rangle_{\mathcal{H}},$$

where $\langle \alpha, \tilde{\Delta}\alpha \rangle_{\mathcal{H}}$ follows the distribution $\mathcal{N}\left(0, \sum_{i=1}^{s+1} 4c_i \mathrm{Var}(\langle \alpha, ((\mu_d^{(i)} - \bar{E}Y) \otimes \epsilon_d^{(i)})\alpha \rangle_{\mathcal{H}})\right)$ for any $\alpha \in \mathcal{H}$ satisfying $||\alpha||_{\mathcal{H}} = 1$. For $\hat{v}_j - v_j$, it can be deduced from Theorem 5.1.8 and Theorem 9.1.3 of Hsing and Eubank (2015) that:

$$\sqrt{n}(\hat{v}_j - v_j) \xrightarrow{\mathcal{D}} \sum_{\lambda_k \neq \lambda_j} \frac{1}{\lambda_j - \lambda_k} P_k \tilde{\Delta} v_j.$$

The proof of Theorem 7 is completed. ■

F.7 Proof of Theorem 9

Write $\tilde{\eta}_n = \max\left\{\sqrt{\frac{1}{n}}, \frac{\beta_n}{n}\right\}$. Write $\hat{\lambda}_s(\Delta_n^{kernel})$ and $\lambda_s(\Delta^{kernel})$ as $\hat{\lambda}_s$ and λ_s in short. Based on (5), Δ_n^{kernel} and K_n have the same non-zero eigenvalue. Similar to Zhou et al. (2022), employing the Weyl inequality in operators yields the following results for any fixed s:

$$\left|\hat{\lambda}_s - \lambda_s\right| \leq ||\Delta_n^{kernel} - \Delta^{kernel}||_{HS} = O_p\left(\tilde{\eta}_n\right).$$

For the purpose of dimension reduction, it is reasonable to assume that q_d is smaller than p. Hence, we limit our consideration to the first p eigenvalues for estimating q_d.

$$-||\Delta_n^{kernel} - \Delta^{kernel}||_{HS} \leq \min_{q_d+1\leq s\leq p} \hat{\lambda}_s \leq \max_{q_d+1\leq s\leq p} \hat{\lambda}_s \leq ||\Delta_n^{kernel} - \Delta^{kernel}||_{HS}.$$

Since $\lambda_{q_d} > 0$ and $\lambda_{q_d+1} = 0$, we can obtain

$$\frac{-||\Delta_n^{kernel} - \Delta^{kernel}||_{HS} + c_n}{\lambda_{q_d} + ||\Delta_n^{kernel} - \Delta^{kernel}||_{HS} + c_n} \leq \frac{\hat{\lambda}_{(q_d+1)} + c_n}{\hat{\lambda}_{q_d} + c_n} \leq \frac{||\Delta_n^{kernel} - \Delta^{kernel}||_{HS} + c_n}{\lambda_{q_d} - ||\Delta_n^{kernel} - \Delta^{kernel}||_{HS} + c_n}.$$

Due to the fact that $||\Delta_n^{kernel} - \Delta^{kernel}||_{HS} = O_p\left(\tilde{\eta}_n\right)$ and the conditions $c_n \to 0$ and $c_n/\tilde{\eta}_n \to \infty$, and $c_n/\lambda_{q_d} \to 0$, we have that with a probability going to 1, $\frac{\hat{\lambda}_{(q_d+1)}+c_n}{\hat{\lambda}_{q_d}+c_n} \to 0$. Additionally, since for any $l > q_d$, $\lambda_l = 0$, we achieve

$$\min_{l>q_d} \frac{\hat{\lambda}_{(l+1)} + c_n}{\hat{\lambda}_l + c_n} \geq \frac{\min_{l>q_d} \hat{\lambda}_{q_d} + c_n}{\max_{l>q_d} \hat{\lambda}_{q_d} + c_n} \geq \frac{-||\Delta_n^{kernel} - \Delta^{kernel}||_{HS} + c_n}{||\Delta_n^{kernel} - \Delta^{kernel}||_{HS} + c_n}.$$

Based on the facts $c_n/\tilde{\eta}_n \to \infty$ and $||\Delta_n^{kernel} - \Delta^{kernel}||_{HS} = O_p\left(\tilde{\eta}_n\right)$, we have that with a probability going to 1,

$$\min_{l>q_d} \frac{\hat{\lambda}_{(l+1)} + c_n}{\hat{\lambda}_l + c_n} \to 1 > \tau.$$

Therefore, we conclude that $P(\hat{q}_d = q_d) \to 1$. ■

F.8 Proof of Proposition 10

The similar arguments used to proving Theorem 2 can be used to prove Proposition 10, we then omit the details here.

F.9 Proof of Theorem 12

We divide $\Delta_n - \Delta$ into two parts:

$$\Delta_n - \Delta = \Sigma_n - \Sigma - (\Sigma_{pooled,n} - \Sigma_{pooled}),$$

where $M = \Sigma_{pooled} + \Delta$. Let $\bar{E}X = \sum_{l=1}^{d} \omega_l \mu^{(l)}$, then we decompose Σ_n as:

$$\begin{aligned}
\Sigma_n &= \frac{1}{n}\sum_{i=1}^{n}\left(X_i - \bar{X}\right)\left(X_i - \bar{X}\right)^\top \\
&= \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{E}X)(X_i - \bar{E}X)^\top + \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{E}X)(\bar{E}X - \bar{X})^\top \\
&\quad + \frac{1}{n}\sum_{i=1}^{n}(\bar{E}X - \bar{X})(X_i - \bar{E}X)^\top + \frac{1}{n}\sum_{i=1}^{n}(\bar{E}X - \bar{X})(\bar{E}X - \bar{X})^\top = H_1 + H_2 + H_3 + H_4.
\end{aligned}$$

We can express the decomposition of H_1 as follows:

$$\begin{aligned}
H_1 &= \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{E}X)(X_i - \bar{E}X)^\top \\
&= \frac{1}{n}\sum_{i=1}^{n}(X_i - E(X_i))(X_i - E(X_i))^\top + \frac{1}{n}\sum_{i=1}^{n}(E(X_i) - \bar{E}X)(X_i - E(X_i))^\top \\
&\quad + \frac{1}{n}\sum_{i=1}^{n}(X_i - E(X_i))(E(X_i) - \bar{E}X)^\top + \frac{1}{n}\sum_{i=1}^{n}(E(X_i) - \bar{E}X)(E(X_i) - \bar{E}X)^\top \\
&= L_1 + L_2 + L_3 + \Delta.
\end{aligned}$$

Therefore, we can express the decomposition of $\Sigma_n - M$ as follows:

$$\Sigma_n - M = L_1 + L_2 + L_3 + H_2 + H_3 + H_4 - \Sigma_{pooled}.$$

Next, let us examine the expression $\Sigma_{pooled,n} - \Sigma_{pooled}$. Without loss of generality, we consider the scenario where $C_i \subset \tilde{C}_i$. For the remaining cases, a similar derivation can be obtained. Now, let us proceed with the decomposition of $\Sigma_{pooled,n} - \Sigma_{pooled}$ as follows:

$$\begin{aligned}
&\alpha^\top\left(\Sigma_{pooled,n} - \Sigma_{pooled}\right)\alpha \\
=&\alpha^\top\left\{\sum_{i=1}^{d}\frac{1}{2n\tilde{n}_i}\sum_{k\in\tilde{C}_i}\sum_{l\neq k, l\in\tilde{C}_i}(W_k - W_l + \delta_{kl})(W_k - W_l + \delta_{kl})^\top - \Sigma_{pooled}\right\}\alpha \\
=&\sum_{i=1}^{d}\frac{1}{n\tilde{n}_i}\sum_{k\in\tilde{C}_i}\left(\alpha^\top W_k W_k^\top\alpha - \alpha^\top\Sigma_{pooled}\alpha\right) - \sum_{i=1}^{d}\frac{1}{n\tilde{n}_i}\sum_{k\in\tilde{C}_i}\sum_{l\neq k, l\in\tilde{C}_i}\alpha^\top W_k W_l^\top\alpha \\
&\quad + \sum_{i=1}^{d}\frac{1}{n\tilde{n}_i}\sum_{k\in\tilde{C}_i}\sum_{l\neq k, l\in\tilde{C}_i}\alpha^\top\delta_{kl}(W_k - W_l)^\top\alpha + \sum_{i=1}^{d}\frac{1}{2n\tilde{n}_i}\sum_{k\in\tilde{C}_i}\sum_{l\neq k, l\in\tilde{C}_i}\alpha^\top\delta_{kl}\delta_{kl}^\top\alpha \\
=&\alpha^\top(L_1 - \Sigma_{pooled} - G_1 + G_2 + G_3)\alpha.
\end{aligned}$$

Consequently, we can decompose $\Delta_n - \Delta$ into the following form:

$$\Delta_n - \Delta = L_2 + L_3 + H_2 + H_3 + H_4 + G_1 - G_2 - G_3.$$

We investigate the asymptotic distribution of $\alpha^\top(L_2+L_3)\alpha$, it is easy to get $\alpha^\top L_2\alpha = \alpha^\top L_3\alpha$, then $\alpha^\top(L_2+L_3)\alpha = 2\alpha^\top L_2\alpha$. Let $L_2 = \frac{1}{n}\sum_{i=1}^n (E(X_i)-\bar{E}X)W_i^\top$ and $Z_i = \frac{1}{\sqrt{n}}\alpha^\top(E(X_i)-\bar{E}X)W_i^\top\alpha$ as well as $\alpha^\top L_2\alpha = \frac{1}{\sqrt{n}}\sum_{i=1}^n Z_i$. Due to the properties of the covariance operator, we have the following:

$$E(Z_i)=0, \quad \mathrm{Var}(Z_i) = \frac{1}{n}\mathrm{Var}(\alpha^\top(E(X_i)-\bar{E}X)W_i^\top\alpha),$$

and

$$B^2 = \mathrm{Var}\left(\sum_{i=1}^n Z_i\right) = \frac{1}{n}\sum_{i=1}^n \mathrm{Var}(\alpha^\top(E(X_i)-\bar{E}X)W_i^\top\alpha).$$

To verify the Lindeberg condition, we have the following conclusion for any $\eta>0$:

$$\begin{aligned}
&\sum_{i=1}^n \frac{1}{B^2}\int_{|Z_i-E(Z_i)|>\eta B}(Z_i-E(Z_i))^2 dF_i \\
\leq &\frac{n}{B^2}\max_i E\left\{(Z_i-E(Z_i))^4\right\}^{1/2}\left\{P(|Z_i-E(Z_i)|>\eta B)\right\}^{1/2} \\
\leq &\max_i \frac{E\left\{(\alpha^\top((E(X_i)-\bar{E}X)W_i^\top)\alpha)^4\right\}^{1/2}}{B^2}\left\{\frac{\mathrm{Var}(Z_i)}{\eta^2 B^2}\right\}^{1/2} \\
\leq &\frac{\max_i \lambda_{max}^{1/2}(E((E(X_i)-\bar{E}X)W_i^\top)^4)\max_i \lambda_{max}^{1/2}(\mathrm{Var}((E(X_i)-\bar{E}X)W_i^\top)}{\left\{\min_i \lambda_{min}(\mathrm{Var}((E(X_i)-\bar{E}X)W_i^\top))\right\}^{3/2}\eta\sqrt{n}} = O\left(\frac{1}{\sqrt{n}}\right)\to 0,
\end{aligned}$$

where $E((E(X_i)-\bar{E}X)W_i^\top)^4$ and $\mathrm{Var}((E(X_i)-\bar{E}X)W_i^\top)$ satisfy

$$\begin{aligned}
\alpha^\top(E((E(X_i)-\bar{E}X)W_i^\top)^4)\alpha &= E\left\{(\alpha^\top((X_i-\bar{E}X)W_i^\top)\alpha)^4\right\}, \\
\alpha^\top(\mathrm{Var}((E(X_i)-\bar{E}X)W_i^\top))\alpha &= \mathrm{Var}(\alpha^\top((E(X_i)-\bar{E}X)W_i^\top)\alpha).
\end{aligned}$$

Since $\frac{1}{n}\sum_{i=1}^n \mathrm{Var}(\alpha^\top(E(X_i)-\bar{E}X)W_i^\top\alpha) = \sum_{i=1}^d \omega_i \mathrm{Var}(\alpha^\top(\mu^{(i)}-\bar{E}X)(W^{(i)})^\top\alpha)$, we can derive the resulting asymptotic distribution as follows:

$$\sum_{i=1}^n Z_i \xrightarrow{\mathcal{D}} \mathcal{N}\left(0, \sum_{i=1}^d \omega_i \mathrm{Var}(\alpha^\top(\mu^{(i)}-\bar{E}X)(W^{(i)})^\top\alpha)\right).$$

Moreover, the subsequent conclusion can be derived:

$$\sqrt{n}\alpha^\top(L_2+L_3)\alpha \xrightarrow{\mathcal{D}} \mathcal{N}\left(0, \sum_{i=1}^d 4\omega_i \mathrm{Var}(\alpha^\top(\mu^{(i)}-\bar{E}X)(W^{(i)})^\top\alpha)\right).$$

Therefore, we have $\alpha^\top(L_2+L_3)\alpha = O_p\left(\frac{1}{\sqrt{n}}\right)$. Next, we will provide a proof for the following conclusion: $\alpha^\top H_2\alpha = o_p\left(\frac{1}{\sqrt{n}}\right)$, $\alpha^\top H_3\alpha = o_p\left(\frac{1}{\sqrt{n}}\right)$, $\alpha^\top H_4\alpha = o_p\left(\frac{1}{\sqrt{n}}\right)$, $\alpha^\top G_1\alpha = o_p\left(\frac{1}{\sqrt{n}}\right)$, $\alpha^\top G_2\alpha = o_p\left(\frac{1}{\sqrt{n}}\right)$, and $\alpha^\top G_3\alpha = o_p\left(\frac{1}{\sqrt{n}}\right)$. Based on the given statement, supposing that Assumptions 13 and 14 hold, where $\eta_2 = \min_{k\neq l\in[d]}\left\|\mu^{(k)}-\mu^{(l)}\right\| > 4\sigma\sqrt{\log(n)}$, we can

deduce that $A_s < \frac{1}{n}$. Additionally, under Assumption 13, we have $\alpha_0 > \frac{Cd\log(n)}{n}$. By applying the Cauchy-Schwartz inequality, it can be easily shown that for any vector α with $\|\alpha\| = 1$, we have $\mu_i^\top \alpha \leq O(\log(n))$. Regarding $\alpha^\top H_2 \alpha$, we can draw the following conclusion:

$$\begin{aligned}\alpha^\top H_2\alpha &= \frac{1}{n}\sum_{i=1}^n \alpha^\top (X_i - \bar{E}X)(\bar{E}X - \bar{X})^\top \alpha = \frac{1}{n}\sum_{i=1}^n \alpha^\top (X_i - E(X_i))(\bar{E}X - \bar{X})^\top \alpha \\ &= \alpha^\top (\frac{1}{n}\sum_{i=1}^n W_i)(\frac{1}{n}\sum_{i=1}^n W_i^\top)\alpha \leq |\alpha^\top (\frac{1}{n}\sum_{i=1}^n W_i)||(\frac{1}{n}\sum_{i=1}^n W_i^\top)\alpha| = O_p\left(\frac{1}{n}\right).\end{aligned}$$

Similarly to H_2, we can establish that $\alpha^\top H_3 \alpha = O_p\left(n^{-1}\right)$. As for H_4, we can derive the following conclusion:

$$\alpha^\top H_4\alpha = \sum_{i=1}^n \alpha^\top (\bar{E}X - \bar{X})(\bar{E}X - \bar{X})^\top \alpha \leq |\alpha^\top (\frac{1}{n}\sum_{i=1}^n W_i)||(\frac{1}{n}\sum_{i=1}^n W_i^\top)\alpha| = O_p\left(\frac{1}{n}\right).$$

Regarding G_1, we can draw the following conclusion:

$$\begin{aligned}G_1 &= \sum_{i=1}^d \frac{1}{n\tilde{n}_i}\sum_{k\in\tilde{C}_i}\sum_{l\neq k, l\in\tilde{C}_i} \alpha^\top W_k W_l^\top \alpha \\ &\leq \sum_{i=1}^d \frac{\tilde{n}_i}{n}|\frac{1}{\tilde{n}_i}\sum_{k\in\tilde{C}_i}\alpha^\top W_k||\frac{1}{\tilde{n}_i}\sum_{l\neq k, l\in\tilde{C}_i} W_l^\top \alpha| = O_p\left(\frac{1}{n}\right).\end{aligned}$$

In the case of G_2, we can decompose it into the following form:

$$\begin{aligned}G_2 &= \sum_{i=1}^d \frac{1}{n\tilde{n}_i}\sum_{k\in\tilde{C}_i}\sum_{l\neq k, l\in\tilde{C}_i} \alpha^\top \delta_{kl}(W_k - W_l)^\top \alpha \\ &= \sum_{i=1}^d \frac{1}{n\tilde{n}_i}\sum_{k\in\tilde{C}_i/C_i}\sum_{l\neq k, l\in\tilde{C}_i/C_i} \alpha^\top (\delta_{kl}W_k^\top - \delta_{kl}W_l^\top)^\top \alpha = G_{21} - G_{22}.\end{aligned}$$

Regarding G_{21}, we can deduce the following conclusion:

$$\begin{aligned}G_{21} &= \sum_{i=1}^d \frac{1}{n\tilde{n}_i}\sum_{k\in\tilde{C}_i/C_i}\sum_{l\neq k, l\in\tilde{C}_i/C_i} \alpha^\top \delta_{kl} W_k^\top \alpha \\ &\leq \sum_{i=1}^d \frac{1}{n\tilde{n}_i}\sum_{k\in\tilde{C}_i/C_i}\sum_{l\neq k, l\in\tilde{C}_i/C_i} |\alpha^\top \delta_{kl}||W_k^\top \alpha| \\ &\leq \sum_{i=1}^d \frac{\tilde{n}_i - n_i}{n\tilde{n}_i}\sum_{k\in\tilde{C}_i/C_i} |W_k^\top \alpha| O\left(\sqrt{\log(n)}\right) \\ &= \sum_{i=1}^d O_p\left(\frac{(\tilde{n}_i - n_i)^{3/2}}{n\tilde{n}_i}\sqrt{\log(n)}\right) \\ &\leq \sum_{i=1}^d O_p\left(\frac{(\tilde{n}_i - n_i)^{3/2}}{\alpha_0 n^2}\sqrt{\log(n)}\right) \leq \sum_{i=1}^d O_p\left(\frac{A_s^{3/2}}{\alpha_0}\sqrt{\frac{\log(n)}{n}}\right) < O_p\left(\frac{1}{n}\right).\end{aligned}$$

Similar to G_{21}, for G_{22}, we can conclude the following:

$$G_{22} = \sum_{i=1}^{d} \frac{1}{n\tilde{n}_i} \sum_{k \in \tilde{C}_i/C_i} \sum_{l \neq k, l \in \tilde{C}_i/C_i} \alpha^\top \delta_{kl} W_l^\top \alpha < O_p\left(\frac{1}{n}\right).$$

Regarding G_3, we can draw the following conclusion:

$$\begin{aligned} G_3 &= \sum_{i=1}^{d} \frac{1}{2n\tilde{n}_i} \sum_{k \in \tilde{C}_i} \sum_{l \neq k, l \in \tilde{C}_i} \alpha^\top \delta_{kl} \delta_{kl}^\top \alpha \\ &\leq \sum_{i=1}^{d} \frac{1}{2n\tilde{n}_i} \sum_{k \in \tilde{C}_i/C_i} \sum_{l \neq k, l \in \tilde{C}_i/C_i} |\alpha^\top \delta_{kl}||\delta_{kl}^\top \alpha| \\ &= \sum_{i=1}^{d} \frac{(\tilde{n}_i - n_i)^2}{2n\tilde{n}_i} O\left(\log(n)\right) \leq \frac{A_s^2}{2\alpha_0} O\left(\log(n)\right) < O_p\left(\frac{1}{n}\right). \end{aligned}$$

Therefore, altogether with all the obtained results, we can establish the following:

$$\sqrt{n}\alpha^\top(\Delta_n - \Delta)\alpha \xrightarrow{\mathcal{D}} \mathcal{N}\left(0, \sum_{i=1}^{d} 4\omega_i \mathrm{Var}(\alpha^\top(\mu^{(i)} - \bar{E}X)(W^{(i)})^\top \alpha)\right).$$

Therefore, we can conclude that:

$$\alpha^\top(\Delta_n - \Delta)\alpha = \alpha^\top(\Sigma_n - M)\alpha - \alpha^\top(\Sigma_{pooled,n} - \Sigma_{pooled})\alpha = O_p\left(\frac{1}{\sqrt{n}}\right).$$

By adapting the parallel line as that in the proof of Theorem 2.3 in Zhu et al. (2025), from the results in (20), we get the two following conclusions:

$$||\Delta_n - \Delta||_F = \sqrt{tr(\Delta_n - \Delta)(\Delta_n - \Delta)^\top} = \sqrt{\sum_{k=1}^{p} \lambda_k^2(\Delta_n - \Delta)} = O_p\left(\sqrt{\frac{p}{n}}\right),$$

$$||B_n - B||_F = O_p\left(\sqrt{\frac{p}{n}}\right).$$

The proof of Theorem 12 is completed. ■

F.10 Proof of Theorem 14

Recall that

$$\Delta = \frac{1}{2} \sum_{k=1}^{s+1} \sum_{l=1}^{s+1} c_k c_l \left(\mu^{(k)} - \mu^{(l)}\right)\left(\mu^{(k)} - \mu^{(l)}\right)^\top = \frac{1}{2} A A^\top,$$

where $A = \left(\sqrt{c_1 c_2}\{\mu^{(1)} - \mu^{(2)}\}, \ldots, \sqrt{c_s c_{s+1}}\{\mu^{(s+1)} - \mu^{(s)}\}\right)$, we have $\mathrm{Span}\{\Delta\} \subseteq \mathrm{Span}\{A\}$.

As $\mathrm{rank}(A) = \mathrm{rank}(AA^\top)$ and $\mathrm{Span}\{AA^\top\} = \mathrm{Span}\{A\}$, we conclude that $\mathrm{Span}\{\Delta\} = \mathrm{Span}\{A\}$. By the definition of central mean deviation subspace $S_{\{E(X_i)\}_{i=1}^n}$, $\mathrm{Span}\{A\} = S_{\{E(X_i)\}_{i=1}^n}$. Therefore, we can get $\mathrm{Span}\{\Delta\} = S_{\{E(X_i)\}_{i=1}^n}$.

The rest proof of this theorem can adopted the same description as that in the proof of Theorem 2.1 in Zhu et al. (2025); we omit it here. ■

F.11 Proof of Theorem 15

Here we give a general conclusion, that is, we consider $\alpha^\top(\Delta_n - \Delta)\alpha$ for any $||\alpha|| = 1$. Recall that:

$$\Delta_n - \Delta \;=\; \Sigma_n - \Sigma - (\Sigma_{pooled,n} - \Sigma_{pooled}).$$

Let $\bar{E}X = \sum_{l=1}^{s+1} c_l\mu^{(l)}$, then we have:

$$\begin{aligned}
\Sigma_n &= \frac{1}{n}\sum_{i=1}^{n}\left(X_i - \bar{X}\right)\left(X_i - \bar{X}\right)^\top \\
&= \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{E}X)(X_i - \bar{E}X)^\top + \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{E}X)(\bar{E}X - \bar{X})^\top \\
&\quad + \frac{1}{n}\sum_{i=1}^{n}(\bar{E}X - \bar{X})(X_i - \bar{E}X)^\top + \frac{1}{n}\sum_{i=1}^{n}(\bar{E}X - \bar{X})(\bar{E}X - \bar{X})^\top = H_1 + H_2 + H_3 + H_4.
\end{aligned}$$

We now deal with the fourth terms $\alpha^\top H_i\alpha$ one by one. For any fixed α with $||\alpha|| = 1$, we have:

$$\begin{aligned}
H_1 &= \frac{1}{n}\sum_{j=1}^{s+1}\sum_{i=z_{j-1}+1}^{z_j}(X_i - \sum_{l=1}^{s+1}c_l\mu^{(l)})(X_i - \sum_{l=1}^{s+1}c_l\mu^{(l)})^\top \\
&= \frac{1}{n}\sum_{j=1}^{s+1}\sum_{i=z_{j-1}+1}^{z_j}\Bigg\{(X_i - \mu^{(j)})(X_i - \mu^{(j)})^\top + \sum_{l=1}^{s+1}c_l(X_i - \mu^{(j)})(\mu^{(j)} - \mu^{(l)})^\top \\
&\quad + \sum_{l=1}^{s+1}c_l(\mu^{(j)} - \mu^{(l)})(X_i - \mu^{(j)})^\top + \left(\sum_{l=1}^{s+1}c_l(\mu^{(j)} - \mu^{(l)})\right)\left(\sum_{i=1}^{s+1}c_i(\mu^{(j)} - \mu^{(l)})\right)^\top\Bigg\} \\
&= L_1 + L_2 + L_3 + \Delta.
\end{aligned}$$

For $\alpha^\top L_1\alpha$, Zhu et al. (2025) established the following result, under the Assumptions 9–11:

$$L_1 \to \sum_{j=1}^{s+1}c_j\Sigma_j = \Sigma_{pooled}.$$

Consider the $\alpha^\top L_2\alpha$ and $\alpha^\top L_3\alpha$, we can use Chebyshev's inequality to get the conclusion. For L_2, we have:

$$\begin{aligned}
L_2 &= \frac{1}{n}\sum_{j=1}^{s+1}\sum_{l=1}^{s+1}\sum_{i=z_{j-1}+1}^{z_j}c_l(X_i - \mu^{(j)})(\mu^{(j)} - \mu^{(l)})^\top = \frac{1}{n}\sum_{i=1}^{n}\sum_{l=1}^{s+1}c_l\epsilon_i(E(X_i) - \mu^{(l)})^\top \\
&= \frac{1}{n}\sum_{i=1}^{n}\sum_{l=1}^{s+1}c_l(\epsilon_i E(X_i) - \epsilon_i\mu^{(l)})^\top = L_{21} + L_{22}.
\end{aligned}$$

For L_{21}, the inequality $\alpha^\top L_{21}\alpha \le \frac{s}{n}\sum_{i=1}^{n}\alpha^\top\epsilon_i E(X_i)^\top\alpha$ holds. As s is limited, based on Assumptions 9 and 12, for any $\xi > 0$, we can select M as follows:

$$M = \left\{\frac{\max_{1\le i\le n}|\alpha^\top E(X_i)|^2\max_{1\le i\le n}\lambda_{max}(\Sigma_i)}{\xi}\right\}^{1/2} + 1.$$

Under these conditions, the subsequent inequality is valid.

$$\Pr\left(\sqrt{n}\left|\frac{1}{n}\sum_{i=1}^{n}\alpha^\top\epsilon_i E(X_i)^\top\alpha\right| > M\right)$$
$$\leq \frac{\sum_{i=1}^{n}\left\{\alpha^\top(E(X_i))\right\}^2\alpha^\top E(\epsilon_i\epsilon_i^\top)\alpha}{nM^2} \leq \frac{\max_{1\leq i\leq n}|\alpha^\top E(X_i)|^2\max_{1\leq i\leq n}\lambda_{max}(\Sigma_i)}{M^2} < \xi.$$

Thus, we have $\alpha^\top L_{21}\alpha = O_p\left(\frac{1}{\sqrt{n}}\right)$. Under the Assumption 12, for L_{22}, the following relationship holds:

$$\alpha^\top L_{22}\alpha = \frac{1}{n}\sum_{i=1}^{n}\sum_{l=1}^{s+1}\alpha^\top\epsilon_i c_l(\mu^{(l)})^\top\alpha \leq| \frac{1}{n}\sum_{i=1}^{n}\alpha^\top\epsilon_i \;||\; (\sum_{l=1}^{s+1}c_l\mu^{(l)})^\top \;|= O_p\left(\frac{1}{\sqrt{n}}\right).$$

Thus, we have $\alpha^\top L_2\alpha = O_p\left(\frac{1}{\sqrt{n}}\right)$. For L_3, we can use the similar description as that of L_2 to get $\alpha^\top L_3\alpha = O_p\left(\frac{1}{\sqrt{n}}\right)$.

For H_2, we have the following derivation:

$$\begin{aligned}
H_2 &= \frac{1}{n}\sum_{i=1}^{n}(X_i-\bar{E}X)(\bar{E}X-\bar{X})^\top \\
&= \frac{1}{n}\sum_{j=1}^{s+1}\sum_{i=z_{j-1}+1}^{z_j}\left((X_i-\mu^{(j)})(\sum_{l=1}^{s+1}c_l\left(\bar{X}_l-\mu_l\right))^\top + \sum_{l=1}^{s+1}c_l(\mu^{(j)}-\mu^{(l)})(\sum_{l=1}^{s+1}c_l\left(\bar{X}_l-\mu^{(l)}\right))^\top\right) \\
&= \left\{\frac{1}{n}\sum_{j=1}^{s+1}\sum_{i=z_{j-1}+1}^{z_j}(X_i-\mu^{(j)})\right\}\left\{\frac{1}{n}\sum_{j=1}^{s+1}\sum_{i=z_{j-1}+1}^{z_j}(X_i-\mu^{(j)})\right\}^\top \\
&\quad + \sum_{j=1}^{s+1}\sum_{l=1}^{s+1}c_jc_l(\mu^{(j)}-\mu^{(l)})\left\{\frac{1}{n}\sum_{j=1}^{s+1}\sum_{i=z_{j-1}+1}^{z_j}(X_i-\mu^{(j)})\right\}^\top = L_4+L_5.
\end{aligned}$$

For $\alpha^\top L_4\alpha$, the following holds:

$$\begin{aligned}
\alpha^\top L_4\alpha &= \left\{\frac{1}{n}\sum_{j=1}^{s+1}\sum_{i=z_{j-1}+1}^{z_j}\alpha^\top(X_i-\mu^{(j)})\right\}\left\{\frac{1}{n}\sum_{j=1}^{s+1}\sum_{i=z_{j-1}+1}^{z_j}(X_i-\mu^{(j)})^\top\alpha\right\} \\
&\leq| \frac{1}{n}\sum_{j=1}^{s+1}\sum_{i=z_{j-1}+1}^{z_j}\alpha^\top(X_i-\mu^{(j)}) \;||\; \frac{1}{n}\sum_{j=1}^{s+1}\sum_{i=z_{j-1}+1}^{z_j}(X_i-\mu^{(j)})^\top\alpha \;|= O_p\left(\frac{1}{n}\right).
\end{aligned}$$

Under Assumption 12, considering $\alpha^\top L_5\alpha$, the following holds:

$$\begin{aligned}
\alpha^\top L_5\alpha &= \sum_{j=1}^{s+1}\sum_{l=1}^{s+1}c_jc_l\alpha^\top(\mu^{(j)}-\mu^{(l)})\left\{\frac{1}{n}\sum_{j=1}^{s+1}\sum_{i=z_{j-1}+1}^{z_j}(X_i-\mu^{(j)})^\top\alpha\right\} \\
&\leq| \sum_{j=1}^{s+1}\sum_{l=1}^{s+1}c_jc_l\alpha^\top(\mu^{(j)}-\mu^{(l)}) \;||\; \frac{1}{n}\sum_{j=1}^{s+1}\sum_{i=z_{j-1}+1}^{z_j}(X_i-\mu^{(j)})^\top\alpha \;|= O_p\left(\frac{1}{\sqrt{n}}\right).
\end{aligned}$$

Similar to H_2, the equations for H_3 and H_4 are as follows:

$$\begin{aligned} H_3 &= \frac{1}{n}\sum_{i=1}^{n}(\bar{E}X-\bar{X})(X_i-\bar{E}X)^\top \\ &= \left\{\frac{1}{n}\sum_{j=1}^{s+1}\sum_{i=z_{j-1}+1}^{z_j}(X_i-\mu^{(j)})\right\}\left\{\frac{1}{n}\sum_{j=1}^{s+1}\sum_{i=z_{j-1}+1}^{z_j}(X_i-\mu^{(j)})\right\}^\top \\ &\quad+\left\{\frac{1}{n}\sum_{j=1}^{s+1}\sum_{i=z_{j-1}+1}^{z_j}(X_i-\mu^{(j)})\right\}\sum_{j=1}^{s+1}\sum_{l=1}^{s+1}c_jc_l(\mu^{(j)}-\mu^{(l)})^\top = L_6+L_7. \end{aligned}$$

For H_4, the following holds:

$$\begin{aligned} H_4 &= \frac{1}{n}\sum_{i=1}^{n}(\bar{E}X-\bar{X})(\bar{E}X-\bar{X})^\top \\ &= \left\{\frac{1}{n}\sum_{j=1}^{s+1}\sum_{i=z_{j-1}+1}^{z_j}(X_i-\mu^{(j)})\right\}\left\{\frac{1}{n}\sum_{j=1}^{s+1}\sum_{i=z_{j-1}+1}^{z_j}(X_i-\mu^{(j)})\right\}^\top = L_8. \end{aligned}$$

Adapting a similar argument for the terms L_4 and L_5, we conclude that $\alpha^\top L_6\alpha = O_p\left(\frac{1}{n}\right)$, $\alpha^\top L_7\alpha = O_p\left(\frac{1}{\sqrt{n}}\right)$ and $\alpha^\top L_8\alpha = O_p\left(\frac{1}{n}\right)$.

Therefore, we can draw the following conclusion:

$$\alpha^\top(\Sigma_n-\Delta-\Sigma_{pooled})\alpha = \alpha^\top(\Sigma_n-M)\alpha = O_p\left(\frac{1}{\sqrt{n}}\right).$$

Zhu et al. (2025) proved the asymptotic property of $\alpha^\top(\Sigma_{pooled,n}-\Sigma_{pooled})\alpha$:

$$\begin{aligned} \alpha^\top(\Sigma_{pooled,n}-\Sigma_{pooled})\alpha =& O_p\left(\frac{1}{\sqrt{n}}\right)+O_p\left(\frac{1}{\sqrt{n\beta_n}}\right)+O_p\left(\frac{\sqrt{\beta_n}}{n}\right)+O_p\left(\frac{\beta_n}{n}\right) \\ =& O_p\left(\frac{1}{\sqrt{n}}\right)+O_p\left(\frac{\beta_n}{n}\right). \end{aligned}$$

Therefore, we have:

$$\alpha^\top(\Delta_n-\Delta)\alpha = \alpha^\top(\Sigma_n-\Sigma)\alpha-\alpha^\top(\Sigma_{pooled,n}-\Sigma_{pooled})\alpha = O_p\left(\frac{1}{\sqrt{n}}\right)+O_p\left(\frac{\beta_n}{n}\right). \quad (20)$$

By employing a similar parallel line argument as presented in the proof of Theorem 2.3 in Zhu et al. (2025), and considering the results outlined in (20), we can draw the following two conclusions:

$$\begin{aligned} ||\Delta_n-\Delta||_F = \sqrt{tr(\Delta_n-\Delta)(\Delta_n-\Delta)^\top} =& \sqrt{\sum_{k=1}^{p}\lambda_k^2(\Delta_n-\Delta)} = O_p\left(\sqrt{\frac{p}{n}}\right)+O_p\left(\frac{\sqrt{p}\beta_n}{n}\right), \\ ||B_n-B||_F =& O_p\left(\sqrt{\frac{p}{n}}\right)+O_p\left(\frac{\sqrt{p}\beta_n}{n}\right). \end{aligned}$$

The proof of Theorem 15 is completed. ■

References

Emmanuel Abbe, Jianqing Fan, and Kaizheng Wang. An ℓ_p theory of PCA and spectral clustering. *Annals of Statistics*, 50(4):2359–2385, 2022.

Carlos Alzate and Johan AK Suykens. Multiway spectral clustering with out-of-sample extensions through weighted kernel PCA. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 32(2):335–347, 2008.

Sylvain Arlot, Alain Celisse, and Zaid Harchaoui. A kernel multiple change-point algorithm via model selection. *Journal of Machine Learning Research*, 20:1–56, 2019.

Scott A Armstrong, Jane E Staunton, Lewis B Silverman, Rob Pieters, Monique L den Boer, Mark D Minden, Stephen E Sallan, Eric S Lander, Todd R Golub, and Stanley J Korsmeyer. MLL translocations specify a distinct gene expression profile that distinguishes a unique leukemia. *Nature Genetics*, 30(1):41–47, 2002.

Ibrahim Bagci, Utz Roedig, Ivan Martinovic, Matthias Schulz, and Matthias Hollick. Using channel state information for tamper detection in the Internet of Things. In *Proceedings of the 31st Annual Computer Security Applications Conference*, pages 131–140, 2015.

Matteo Barigozzi, Haeran Cho, and Piotr Fryzlewicz. Simultaneous multiple change-point and factor analysis for high-dimensional time series. *Journal of Econometrics*, 206(1): 187–225, 2018.

Thomas Berrett and Yi Yu. Locally private online change point detection. In *Advances in Neural Information Processing Systems 34*, pages 3425–3437, 2021.

Ekaterini Blaveri, Jeremy L. Brewer, Ritu Roydasgupta, et al. Bladder cancer stage and outcome by array-based comparative genomic hybridization. *The Journal of Urology*, 175 (4):1568–1569, 2006.

Mehmet Caner and Xu Han. Selecting the correct number of factors in approximate factor models: the large panel case with group bridge estimators. *Journal of Business and Economic Statistics*, 32:359–374, 2014.

Alain Celisse, Guillemette Marot, Morgane Pierre-Jean, and Guillem Rigaill. New efficient algorithms for multiple change-point detection with kernels. *Computational Statistics and Data Analysis*, 128:200–220, 2018.

Jie Chen and Arjun K. Gupta. *Parametric Statistical Change Point Analysis: With Applications to Genetics, Medicine, and Finance*. Springer Science & Business Media, 2012.

Song Xi Chen, Li-Xin Zhang, and Ping-Shou Zhong. Tests for high-dimensional covariance matrices. *Journal of the American Statistical Association*, 105(490):810–819, 2010.

Yudong Chen, Tengyao Wang, and Richard J Samworth. Inference in high-dimensional online changepoint detection. *arXiv preprint arXiv:2111.01640*, 2021.

Haeran Cho. Change-point detection in panel data via double cusum statistic. *Electronic Journal of Statistics*, 10(2):2000–2038, 2016.

Haeran Cho and Piotr Fryzlewicz. Multiple-change-point detection for high dimensional time series via sparsified binary segmentation. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 77(2):475–507, 2015.

Alice Cleynen, Sandrine Dudoit, and Stéphane Robin. Comparing segmentation methods for genome annotation based on RNA-Seq data. *Journal of Agricultural, Biological, and Environmental Statistics*, 19(1):101–118, 2014.

Miklós Csörgő and Lajos Horváth. *Limit Theorems in Change-Point Analysis*. John Wiley & Sons, 1997.

Holger Dette, Guangming Pan, and Qing Yang. Estimating a change point in a sequence of very high-dimensional covariance matrices. *Journal of the American Statistical Association*, 117(537):444–454, 2022.

Farida Enikeeva, Zaid Harchaoui, et al. High-dimensional change-point detection under sparse alternatives. *Annals of Statistics*, 47(4):2051–2079, 2019.

Martin Ester, Hans-Peter Kriegel, Jörg Sander, and Xiaowei Xu. A density-based algorithm for discovering clusters in large spatial databases with noise. In *Proceedings of the Second International Conference on Knowledge Discovery and Data Mining*, pages 226–231, 1996.

Bertille Follain, Tengyao Wang, and Richard J Samworth. High-dimensional changepoint estimation with heterogeneous missingness. *arXiv preprint arXiv:2108.01525*, 2021.

Chris Fraley and Adrian E Raftery. Model-based clustering, discriminant analysis, and density estimation. *Journal of the American Statistical Association*, 97(458):611–631, 2002.

Piotr Fryzlewicz. Wild binary segmentation for multiple change-point detection. *Annals of Statistics*, 42(6):2243–2281, 2014.

Kenji Fukumizu, Francis R Bach, and Michael I Jordan. Dimensionality reduction for supervised learning with reproducing kernel Hilbert spaces. *Journal of Machine Learning Research*, 5:73–99, 2004.

Dario Gregori, Danila Azzolina, Corrado Lanera, Ilaria Prosepe, et al. A first estimation of the impact of public health actions against COVID-19 in Veneto (Italy). *Journal of Epidemiology and Community Health*, 74(10):858–860, 2020.

Arthur Gretton, Karsten M Borgwardt, Malte J Rasch, Bernhard Schölkopf, and Alexander Smola. A kernel two-sample test. *Journal of Machine Learning Research*, 13(1):723–773, 2012.

Thomas Grundy, Rebecca Killick, and Gueorgui Mihaylov. High-dimensional changepoint detection via a geometrically inspired mapping. *Statistics and Computing*, 30(4):1155–1166, 2020.

John A. Hartigan. *Clustering Algorithms*. Wiley, 1975.

John A. Hartigan and Manchek A. Wong. A k-means clustering algorithm. *Applied Statistics*, 28:100–108, 1979.

Tailen Hsing and Randall Eubank. *Theoretical Foundations of Functional Data Analysis, with an Introduction to Linear Operators*. John Wiley & Sons, 2015.

Jiaqi Huang, Junhui Wang, Xuehu Zhu, and Lixing Zhu. Two ridge ratio criteria for multiple change point detection in tensors. *arXiv preprint arXiv:2206.13004*, 2022.

Binyan Jiang, Jialiang Li, and Qiwei Yao. Autoregressive networks. *Journal of Machine Learning Research*, 24(227):1–69, 2023.

Shuhao Jiao, Tong Shen, Zhaoxia Yu, and Hernando Ombao. Change-point detection using spectral PCA for multivariate time series. *arXiv preprint arXiv:2101.04334*, 2021.

Moritz Jirak. Uniform change point tests in high dimension. *Annals of Statistics*, 43(6): 2451–2483, 2015.

Claudia Kirch, Birte Muhsal, and Hernando Ombao. Detection of changes in multivariate time series with application to EEG data. *Journal of the American Statistical Association*, 110(511):1197–1216, 2015.

Ludmila I. Kuncheva and William J. Faithfull. PCA feature extraction for change detection in multidimensional unlabelled streaming data. In *Proceedings of the 21st International Conference on Pattern Recognition*, pages 1140–1143, 2012.

Michel Ledoux and Michel Talagrand. *Probability in Banach Spaces: Isoperimetry and Processes*. Springer, 1991.

Kuang Yao Lee, Bing Li, and Francesca Chiaromonte. A general theory for nonlinear sufficient dimension reduction: Formulation and estimation. *Annals of Statistics*, 41(1): 221–249, 2013.

Bing Li. *Sufficient Dimension Reduction: Methods and Applications with R*. CRC Press, 2018.

Bing Li and Jun Song. Nonlinear sufficient dimension reduction for functional data. *Annals of Statistics*, 45(3):1059–1095, 2017.

Ker-Chau Li. Sliced inverse regression for dimension reduction. *Journal of the American Statistical Association*, 86(414):316–327, 1991.

Qian Lin, Zhigeng Zhao, and Jun S. Liu. Sparse sliced inverse regression via lasso. *Journal of the American Statistical Association*, 114(528):1726–1739, 2019.

Tianqi Liu, Yu Lu, Biqing Zhu, and Hongyu Zhao. Clustering high-dimensional data via feature selection. *Biometrics*, 79(2):940–950, 2023.

Yu Lu and Harrison H Zhou. Statistical and computational guarantees of Lloyd's algorithm and its variants. *arXiv preprint arXiv:1612.02099*, 2016.

Alexandre Lung-Yut-Fong, Céline Lévy-Leduc, and Olivier Cappé. Homogeneity and change-point detection tests for multivariate data using rank statistics. *Journal de la Société Française de Statistique*, 154(4):133–162, 2015.

Padilla Madrid, Hernan Oscar, Yi Yu, Daren Wang, and Alessandro Rinaldo. Optimal nonparametric multivariate change point detection and localization. *IEEE Transactions on Information Theory*, 68(3):1922–1944, 2022.

David S Matteson and Nicholas A James. A nonparametric approach for multiple change point analysis of multivariate data. *Journal of the American Statistical Association*, 109 (505):334–345, 2014.

Leland McInnes, John Healy, and James Melville. UMAP: Uniform manifold approximation and projection for dimension reduction. *arXiv preprint arXiv:1802.03426*, 2018.

Andrew Y. Ng, Michael I. Jordan, and Yair Weiss. On spectral clustering: Analysis and an algorithm. In *Advances in Neural Information Processing Systems 14*, pages 849–856, 2001.

Yue S Niu, Ning Hao, and Heping Zhang. Multiple change-point detection: A selective overview. *Statistical Science*, 31(4):611–623, 2016.

E. S. Page. Continuous inspection schemes. *Biometrika*, 41(1–2):100–115, 1954.

Abdulhakim A. Qahtan, Basma Alharbi, Suojin Wang, and Xiangliang Zhang. A PCA-based change detection framework for multidimensional data streams: Change detection in multidimensional data streams. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 935–944, 2015.

Wei Qian, Shanshan Ding, and R. Dennis Cook. Sparse minimum discrepancy approach to sufficient dimension reduction with simultaneous variable selection in ultrahigh dimension. *Journal of the American Statistical Association*, 114(527):1277–1290, 2019.

William M Rand. Objective criteria for the evaluation of clustering methods. *Journal of the American Statistical Association*, 66(336):846–850, 1971.

Alan P. Reynolds, Graeme Richards, Beatriz de la Iglesia, and Victor J. Rayward-Smith. Clustering rules: a comparison of partitioning and hierarchical clustering algorithms. *Journal of Mathematical Modelling and Algorithms*, 5:475–504, 2006.

Frank Rosenblatt. The perceptron: a probabilistic model for information storage and organization in the brain. *Psychological Review*, 65(6):386, 1958.

Sam T Roweis and Lawrence K Saul. Nonlinear dimensionality reduction by locally linear embedding. *Science*, 290(5500):2323–2326, 2000.

Bharath K Sriperumbudur, Kenji Fukumizu, and Gert RG Lanckriet. Universality, characteristic kernels and rkhs embedding of measures. *Journal of Machine Learning Research*, 12(7), 2011.

James H Stock and Mark W Watson. Macroeconomic forecasting using diffusion indexes. *Journal of Business & Economic Statistics*, 20(2):147–162, 2002.

Nicolas Stransky, Céline Vallot, Fabien Reyal, Bernard-Pierrot, et al. Regional copy number–independent deregulation of transcription in cancer. *Nature Genetics*, 38(12): 1386–1396, 2006.

Laurens Van der Maaten and Geoffrey Hinton. Visualizing data using t-SNE. *Journal of Machine Learning Research*, 9(11):2579–2605, 2008.

Carolina Varon, Carlos Alzate, and Johan AK Suykens. Noise level estimation for model selection in kernel PCA denoising. *IEEE Transactions on Neural Networks and Learning Systems*, 26(11):2650–2663, 2015.

Runmin Wang, Changbo Zhu, Stanislav Volgushev, and Xiaofeng Shao. Inference for change points in high-dimensional data via selfnormalization. *Annals of Statistics*, 50(2):781–806, 2022.

Tao Wang, Mengjie Chen, Hongyu Zhao, and Lixing Zhu. Estimating a sparse reduction for general regression in high dimensions. *Statistics and Computing*, 28(1):33–46, 2018.

Tengyao Wang and Richard J Samworth. High dimensional change point estimation via sparse projection. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 80(1):57–83, 2018.

Yingcun Xia, Howell Tong, Wai Keung Li, and Li-Xing Zhu. An adaptive estimation of dimension reduction space. *Journal of the Royal Statistical Society. Series B (Statistical Methodology)*, 64(3):363–410, 2002.

Hang Zhou, Dongyi Wei, and Fang Yao. Theory of functional principal components analysis for discretely observed data. *arXiv preprint arXiv:2209.08768*, 2022.

Liping Zhu, Tao Wang, Lixing Zhu, and Louis Ferré. Sufficient dimension reduction through discretization-expectation estimation. *Biometrika*, 97(2):295–304, 2010.

Xuehu Zhu, Xu Guo, Tao Wang, and Lixing Zhu. Dimensionality determination: A thresholding double ridge ratio approach. *Computational Statistics and Data Analysis*, 146: 106910, 2020.

Xuehu Zhu, Luoyao Yu, Jiaqi Huang, Junmin Liu, and Lixing Zhu. Moment deviation subspaces of dimension reduction for high-dimensional data with change structure. *Statistica Sinica*, 3:737–759, 2025.

Changliang Zou, Guosheng Yin, Long Feng, and Zhaojun Wang. Nonparametric maximum likelihood approach to multiple change-point problems. *Annals of Statistics*, 42(3):970–1002, 2014.

Mihaela ŞErban, Anthony Brockwell, John Lehoczky, and Sanjay Srivastava. Modelling the dynamic dependence structure in multivariate financial time series. *Journal of Time Series Analysis*, 28(5):763–782, 2010.