Functional Post-Clustering Selective Inference with Applications to EHR Data Analysis

Zihan Zhu[*] Xin Gai[*] Anru R. Zhang[†]

(November 24, 2023)

Abstract

In the analysis of Electronic Health Records (EHR), clustering patients according to patterns in their data is crucial for uncovering new subtypes of diseases. Existing medical literature often relies on classical hypothesis testing methods to test for differences in means between these clusters. Due to selection bias induced by clustering algorithms, the implementation of these classical methods on post-clustering data often leads to an inflated Type I error. In our study, we introduce a new statistical approach that adjusts for this bias when analyzing data collected over time. Our method extends classical selective inference methods for cross-sectional data to longitudinal data via the utilization of kernel regression. We provide theoretical guarantees for our approach with upper bounds on the selective type-I and type-II errors. We apply the method to simulated data and real-world Acute Kidney Injury (AKI) EHR datasets, thereby illuminating the advantages of our approach.

1 Introduction

Testing for a difference in means between groups of functional data is fundamental to answering research questions across various scientific areas. Recently, there has been an increasing demand for post-clustering inference of functional data, namely, testing the difference between groups discovered via clustering algorithms. In particular, the electronic health records (EHR) system contains a rich source of longitudinal observational data, covering many biochemical markers, making this type of data an ideal choice for identifying patient subphenotypes. With the increasing prevalence of EHR data, longitudinal data clustering methods used to evaluate patient subphenotypes have become more commonly applied in clinical research, especially in the analysis of vital signs, laboratory values, interventions, etc (Manzini et al., 2022; Ramaswamy et al., 2021; Lou et al., 2021; Chen et al., 2022; Zeldow et al., 2021). Post-clustering inference for functional data is a challenging problem and existing testing methods are not applicable. The main challenge of this problem is the selection bias, which would lead to inflated false discoveries if uncorrected, induced by clustering algorithms. In more detail, the clustering forces separation regardless of the underlying truth, and further makes the p-value invalid. In practical applications, empirical observations reveal that

[*] Department of Statistical Science, Duke University. zihan.zhu@duke.edu, xin.gai@duke.edu

[†] Department of Biostatistics & Bioinformatics and Department of Computer Science, Duke University. anru.zhang@duke.edu

applying classical methods directly often leads to spuriously small p-values (Hall and Van Keilegom, 2007; Zhang and Chen, 2007; Horváth and Kokoszka, 2012; Qiu et al., 2021). This is an instance of a broader phenomenon termed *data snooping* (Ioannidis, 2005), which refers to the misuse of data analysis to find patterns in data that can be presented as statistically significant, thus leading to potentially false conclusions.

Selective inference (Fithian et al., 2014) is commonly used to correct the selection bias. The focus of selective inference has so far mainly been on discrete datasets (Lee et al., 2016; Gao et al., 2022). Motivated by applications in EHR data analysis, we consider here a new framework of selective inference that adapts to continuous functional datasets.

In this paper, we develop a valid test for the difference in means between two clusters estimated from the functional data. To handle the continuity of functional datasets, which often contain large timesteps and cannot be treated as discrete data, our method finds the low-rank spectral representation for the continuous data based on kernel ridge regression. To address the selection bias in the inference procedure, we propose a selective inference framework leveraging the clustering information. Mathematically, we define the selective p-value for post-clustering data via conditioning on the observed clustering partition based on the prior literature on selective inference (Fithian et al., 2014; Lee et al., 2016; Yang et al., 2016; Gao et al., 2022; Chen and Witten, 2022). This selective p-value decouples the bias induced by the clustering algorithms and our theoretical guarantees show that this method controls the selective type-I error.

1.1 Applications: Phenotyping based on Electronic Health Records

The application of longitudinal clustering methods to Electronic Health Records (EHR) data has proven to be a powerful tool for phenotypic classification, offering novel insights into patient heterogeneity and disease progression. There are numerous studies that have similarly utilized longitudinal clustering methods with EHR data to identify various patient subtypes and advance clinical research. For instance, researchers studied type 2 diabetes mellitus(T2DM) patients by analyzing their data on various biochemical markers (Manzini et al., 2022). These markers included glycated hemoglobin (HbA1c), body mass index (BMI), both diastolic and systolic blood pressures, among others. By applying longitudinal deep learning clustering methods to this data, they identified seven distinct subtypes of T2DM. In the field of chronic kidney disease (CKD) research (Ramaswamy et al., 2021), a hybrid semimechanistic modeling methodology was introduced to analyze CKD progression. When applied to the EHR data of CKD patients, the model effectively identified five distinct patient subpopulations. Through this pioneering method, the emphasis was placed on harnessing longitudinal data to understand disease progression phenotypes, thereby aiming to streamline individualized treatment strategies for each subgroup. Building on these foundational methodologies and appreciating their transformative impact in the realm of medical research, our focus narrows to a nuanced application. While our method is versatile and can be applied after any clustering approach, we showcase its efficacy using AKI as a representative example in this paper.

Acute kidney injury (AKI) is a potentially life-threatening condition that impacts approximately 20% of hospitalized patients in the United States (Wang et al., 2012). Given this prevalence, early warning of patient outcomes becomes crucial as it can significantly improve prognosis (MacLeod, 2009). Identifying new subphenotypes often serves as the foundation for such early warnings. Importantly, the most direct and insightful indicator for AKI currently available is the temporal trajectory of creatinine. Therefore, we applied our method specifically to the inference after longitudinal clustering of AKI. In conducting this, we utilized EHR data from the *MIMIC-IV* database

Johnson et al. (2020, 2023); Goldberger et al. (2000). As depicted in Figure 3a, there is a notable heterogeneity in the creatinine trajectories among AKI patients. Under such circumstances, our approach has successfully yielded results that are both meaningful and highly credible.

1.2 Main Contributions

Our work presents a post-clustering selective inference framework for functional data and provides theoretical guarantees to control the selective type-I error under the Gaussian distributional assumption. To handle the aforementioned challenges of function data, our framework is comprised of three parts. To begin with, we leverage the low-dimensional embedding to coerce the high-dimensional function data into low-dimensional tensors and complement the miss values simultaneously. The low-dimensional embedding is a linear transformation and preserves normality, thus, the low-dimensional embedding is a random tensor with each slice following the matrix normal distribution. Next, we propose an estimator to evaluate the unknown covariance matrices of the matrix normal distribution and leverage the estimated covariance matrices to conduct the whitening transformation. We then define the selective p-value based on the tensor obtained by the low-dimensional embedding and whitened transformation. Inspired by Gao et al. (2022), the proposed selective p-value leverages the clustering information to reduce the selection bias and further controls the selective type-I error. Moreover, we prove that the proposed p-value is the conditional probability of a scaled chi-square distribution truncated to a subset of $\mathbb{R}$, and we introduce the Monte Carlo approximation to estimate the proposed selective p-value.

Compared with previous works (Gao et al., 2022; Chen and Witten, 2022; Yun and Barber, 2023; Hivert et al., 2022), our work has two major novelties. First, our selective inference framework addresses the functional data with missing values and multiple features, while previous works mostly focus on vector inputs. We complement the missing values by low-dimensional embedding, namely, the basis expansion regression. This is a linear transformation that preserves the null hypothesis and alternative hypothesis, where each record for a feature is transformed into a low-dimensional vector. The transformed data has a tensor structure induced by the multiple features. Thus, we extend the selective inference for matrix inputs (Gao et al., 2022) into the tensor case and define the selective p-value.

Second, we leverage the sample covariance estimator to conduct the whitening transformation. In contrast to previous works, which usually assumed the covariance matrices are scaled identity matrices (Gao et al., 2022; Chen and Witten, 2022; Yun and Barber, 2023; Hivert et al., 2022), this diagonal covariance assumption does not hold in the functional case. As a result, the estimators for the scaled parameter, such as the mean estimator (Gao et al., 2022), fail in the functional setting. To handle this problem, we show that the problem is essentially estimating the covariance of a truncated normal distribution, and we consider the sample covariance estimator. We prove that the selective inference framework controls the selective type-I error (the sample covariance estimator is consistent under the null hypothesis). Moreover, we show that the statistical power converges to 1 and the proposed selective inference framework is asymptotically powerful.

1.3 Related work

Selective inference. In the classic statistical inference, the hypothesis is assumed to be determined prior to observing the dataset. However, in a wide class of supervised and unsupervised learning tasks, such as regression and clustering, the hypothesis can be data-driven. Therefore, the model selection step brings selection bias and classical inference methods are not valid. To handle this problem, Berk et al. (2013); Fithian et al. (2014); Lee et al. (2016) developed the selective inference framework, which is a process of making statistical inferences that account for the selection effect. Following the work of Lee et al. (2016), selective inference has been applied extensively in various problems, such as high-dimensional linear model (Tibshirani et al., 2016; Yang et al., 2016; Loftus and Taylor, 2015; Charkhi and Claeskens, 2018; Taylor and Tibshirani, 2018; Hyun et al., 2021; Jewell et al., 2022). In the recent years, Gao et al. (2022) proposed an elegant selective inference framework to conduct the hypothesis test on post-clustering dataset, and there are a series of following work focusing on the same topic (Chen and Witten, 2022; Zhang et al., 2019; Hivert et al., 2022; Yun and Barber, 2023) In this paper, we develop a selective inference framework for functional data, while most of the existing work concentrates on post-clustering inference for discrete data.

Functional Clustering. In this paper, we study the post-clustering inference for functional data. Functional clustering refers to the process of categorizing or grouping curves, functions, or shapes based on their patterns or structures. There are various approaches for functional clustering and cluster analysis has been well studied in the functional data analysis literature for its practical applications. For instance, Abraham et al. (2003); Serban and Wasserman (2005); Kayano et al. (2010); Coffey et al. (2014); Giacofci et al. (2013) developed two-stage clustering leveraging the functional basis expansion, where the idea is reducing the dimension of functional data by basis expansion regression to implement clustering methods for low-dimensional vectors. In contrast to the functional basis expansion approach that requires a prespecified set of basis functions, Peng and Müller (2008); Chiou and Li (2007) proposed methods that choose the basis by functional principle components (FPC). Besides, there are other lines of research that conduct functional clustering with different approaches, such as leveraging the FPC subspace-projection (Chiou, 2012; Chiou and Li, 2008) and model-based clustering (Banfield and Raftery, 1993; James and Sugar, 2003; Jacques and Preda, 2014; Heinzl and Tutz, 2014).

Cross-covariance matrix. In this paper, we model the multi-feature functional data by the matrix normal distribution. To conduct the whitening transformation in the proposed selective inference framework, we need to estimate the block covariance matrix, which is determined by the Kronecker product of the covariance matrices of the matrix normal distribution. We remark that estimating the block covariance matrix has been widely studied in Dawid (1981); Dutilleul (1999); Yin and Li (2012); Tsiligkaridis and Hero (2013); Zhou (2014); Chen and Liu (2015); Hoff (2015); Ding and Dennis Cook (2018); Hoff et al. (2022).

1.4 Notation and Preliminaries

We introduce some useful notation before proceeding. Throughout this paper, we denote $\mathcal{MN}(\mu, \Sigma_1, \Sigma_2)$ as the matrix normal distribution with the mean μ and covariance matrices Σ_1, Σ_2. For any positive integer n, we denote $\mathbb{S}_+^n$ as the set containing all n-by-n symmetric positive semi-definite matrices. For any matrix $A \in \mathbb{R}^{m\times n}$, we denote $\|A\|_F$ as its Frobenius norm and denote $\text{vec}(A) \in \mathbb{R}^{mn}$ as the vectorization of A. For any Hilbert space $\mathcal{H}$, we denote $\|\cdot\|_{\mathcal{H}}$ as the associated norm. For any functions f, g, define $f \odot g$ as their cartesian product, namely, for any $(f \odot g)(x, y) = f(x) \cdot g(y)$ where x, y are in the domains of f, g respectively. For any two matrices $A \in \mathbb{R}^{m\times n}, B \in \mathbb{R}^{p\times q}$, define

$$A \otimes B = \begin{bmatrix} a_{11}B & \cdots & a_{1n}B \\ \vdots & \ddots & \vdots \\ a_{m1}B & \cdots & a_{mn}B \end{bmatrix} \in \mathbb{R}^{(pm)\times(qn)}$$

as their Kronecker product. For any positive integer n and any n-mode tensor $\mathcal{A} \in \mathbb{R}^{i_1\times\cdots\times i_n}$, define $\mathcal{A}[j, :, \cdots, :]$ as its jth mode-1 slice, $\mathcal{A}[:, j, :, \cdots, :]$ as its jth mode-2 slice, and so on. For any matrix $U \in \mathbb{R}^{i_m\times K}$, where $i_1, \ldots, i_n, K$ are positive integers, define $U \times_m \mathcal{A} \in \mathbb{R}^{i_1\times\cdots i_{m-1}\times K\times i_{m+1}\times i_n}$ as their m-mode tensor product.

1.5 Roadmap

The rest of this paper is organized as follows. We first demonstrate the problem setting of post-clustering inference for functional data in Section 2. We then present our method based on kernel ridge regression and selective inference technique in Section 3. Next, we derive theoretical guarantees for our method to show that it controls the selective type-I error in Section 4. Finally, we conduct numerical experiments for synthetic data to verify our theory and apply it to real datasets on Acute Kidney Injury (AKI) EHR in Section 5.

2 Problem Settings

In this section, we define the generative model of input datasets and introduce the problem formulation of post-clustering inference for functional data, as well as the challenges and our approach. To elaborate, the electronic health record (EHR) contains records of diverse features for different patients, where each record of a feature is functional data.

Suppose there are T time points in total, n subjects (or patients), and m features. Specifically, we observe $W_{ij}(t), t \in \Omega_{ij}$, where $i \in [m]$ is the subject index, $j \in [n]$ is the feature index, $t \in [0, T]$ is the time point of the measurements, and Ω_{ij} is the set of time points for subject i and feature j. We remark that EHR data usually contains missing values and there might be few time points in Ω_{ij}. Given the collection of records for features and time points that the data were observed for subject (or patient) i, we aim to discover the endotypes of the subjects, i.e., if these subjects form subclusters. For this goal, we leverage kernel ridge regression and apply clustering algorithms to find the subclusters (refer to Section 3).

We consider the model for m patients with n features and T total time points. In more detail, we observe $\mathcal{W} = (W_i)_{i\in[m]}$ for each patient $i \in [m]$, where $W_i := (W_{ij})_{j\in[n]}$ is the observed data of the i th patient n curves within a certain time period recording their physical features. Let $\Omega = (\Omega_i)_{i\in[m]}$ be the corresponding time points of the record $\mathcal{W}$, where $\Omega_i := (\Omega_{ij})_{j\in[n]}$. Here $\Omega_{ij} := (t_{ijk})_{k\in[r_{ij}]} \in \mathbb{R}^{r_{ij}}$ is the record of time points for the j th feature of the i th patient and r_{ij} is the number of time points for this record, and $W_{ij} := (W_{ij}(t_{ijk})) \in \mathbb{R}^{r_{ij}}$ is the record for the j feature of the i patient. For all the $i \in [m], j \in [n]$ and $k \in [r_{ij}]$, we remark that $t_{ijk} \in [0, T]$. In summary, the data for each patient i contains n features, where the record of each feature j is a vector W_{ij} associated with the time points Ω_{ij}.

2.1 Model Setup

Now we present a basic assumption on the data-generating process. Intuitively, we assume each physical feature follows a Gaussian process, which indicates the feature record W_{ij} follows a multivariate normal distribution in the discrete regime. Considering the similarity between patients and for simplicity of analysis, we assume these normal distributions have the same covariance matrix Σ_1, Σ_2. Here Σ_1 is the covariance matrix induced by the kernel of the Gaussian process for each feature and Σ_2 presents the covariance between different features. Intuitively, this assumption supposes the observed record of each patient on the same feature following Gaussian processes only with the difference in the mean functions, and the covariances between any two features are the same for different patients.

Assumption 1 (Distributional Assumption)**.** *Suppose $\mathcal{W} = (W_i)_{i\in[m]}$, where W_i, W_j are independent random matrices for any $i \neq j$. Suppose that $W_i \in \mathbb{R}^{n\times T}$ follows a matrix normal distribution with Gaussian noise:*

$$W_i = Z_i + \epsilon_i, \qquad Z_i \sim \mathcal{MN}(\mu_i, \Sigma_1, \Sigma_2), \qquad \epsilon_i \sim \mathcal{MN}(0, \mathrm{diag}(\sigma_j^2)_{j\in[n]}, I_T) \qquad \forall i \in [m], \tag{1}$$

where ϵ_i are matrices with i.i.d. standard Gaussian noise, σ_j^2 is the variance of noise terms for the jth feature, $\mu_i \in \mathbb{R}^{n\times T}$, $\Sigma_1 \in \mathbb{S}_+^n$, and $\Sigma_2 \in \mathbb{S}_+^T$ are the same as Assumption 1.

Note that the above assumption considers all the features within T time points (i.e. T is the total time points) and supposes they follow the multivariate normal distribution with T coordinates. As a result, with a slight abuse of notation, the data W_i in Assumption 1 is a $n \times T$ matrix. However, we remark that the observed data W_i contains many missing values, and we only observe the realizations of these multivariate normal distributions in coordinates corresponding with Ω_{ij}. That is, we only observe the records for coordinates Ω_{ij} in the data W_i. Also, Assumption 1 considers the matrix normal noise terms ϵ_i, which fits the real-world situation that observed records are often noisy. Specifically, we assume the covariance matrix of each ϵ_i is a diagonal matrix, and the i th coordinate of noise terms is σ_i^2. We further remark that all the parameters $\{\mu_i\}_{i\in[m]}, \Sigma_1, \Sigma_2, \{\sigma_i\}_{i\in[m]}$ are unknown and we only observe the record $\{W_i\}_{i\in[m]}$ as well as the time points $\{\Omega_i\}_{i\in[m]}$.

2.2 Problem Formulation

Given a dataset $\mathcal{X}$ following model (1), one might impose a clustering algorithm to split the patients into two clusters $\mathcal{C}(\mathcal{X}) := \mathcal{C}_1 \cup \mathcal{C}_2$, where $\{\mathcal{C}_1, \mathcal{C}_2\}$ forms a partition of $[m]$ and record the number of patients in each cluster. As aforementioned, we are interested in testing the difference of means between clusters $\mathcal{C}_1, \mathcal{C}_2$, namely, the difference between the mean matrices of the matrix normal distribution defined in Assumption 1. For any partition of $\mathcal{G} \subset [m]$, define

$$\bar{\mu}_{\mathcal{G}} = \frac{\sum_{i\in\mathcal{G}} \mu_i}{|\mathcal{G}|}, \qquad \bar{W}_{\mathcal{G}} = \frac{\sum_{i\in\mathcal{G}} W_i}{|\mathcal{G}|},$$

where $\bar{\mu}_{\mathcal{G}}$ denotes the population mean and $\bar{W}_{\mathcal{G}}$ denotes the sample mean for the data within the partition $\mathcal{G}$: $\mathcal{C}_1, \mathcal{C}_2$, respectively. Next, we cast the problem described above into the following hypothesis testing problem:

$$H_0^{\{\mathcal{C}_1,\mathcal{C}_2\}} : \bar{\mu}_{\mathcal{C}_1} = \bar{\mu}_{\mathcal{C}_2} \text{ versus } H_1 : \bar{\mu}_{\mathcal{C}_1} \neq \bar{\mu}_{\mathcal{C}_2}. \tag{2}$$

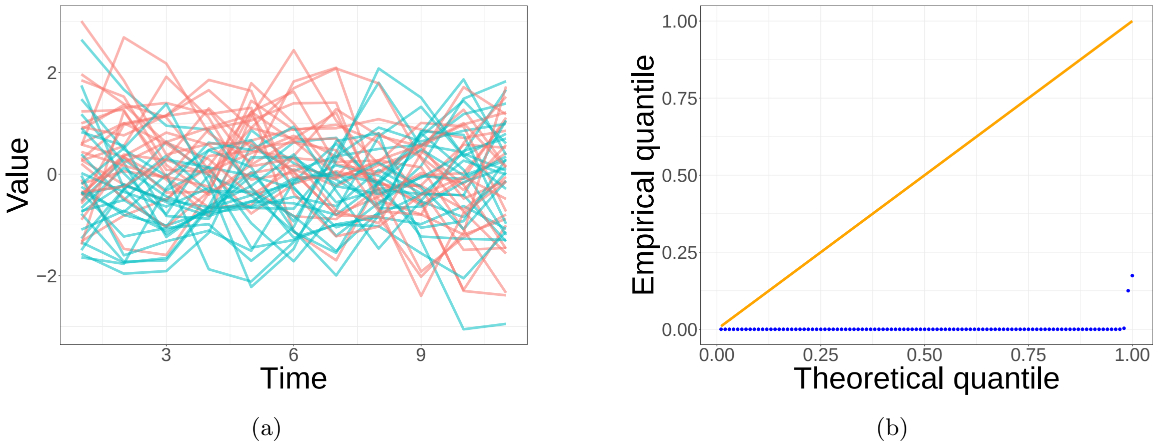


Figure 1: **Failure of the two-sample t-test.** We sample 100 datasets following the model (1) with the common mean $\mu_i = 0$ and the zero noise $\epsilon_i = 0$ for all $i \in [100]$. Each dataset contains the record of 500 patients, where the record for each patient contains one curve with 11 time points ($m = 500, n = 1, T = 11$). We use the k-means algorithm to obtain two clusters. **(a)** shows the first 100 curves of the first dataset labeled by the clustering results. **(b)** is the quantile plot of the p-values for the 100 datasets obtained by the two-sample t-test.

A simple approach for (2) is the Wald test. Suppose $\{w_i\}_{i\in[m]}$ is an observed dataset satisfying the generating process (1), and $\bar{w}_{\mathcal{C}_1}, \bar{w}_{\mathcal{C}_2}$ are sample means for the clusters obtain by certain algorithms. A straightforward formulation for the p-value is

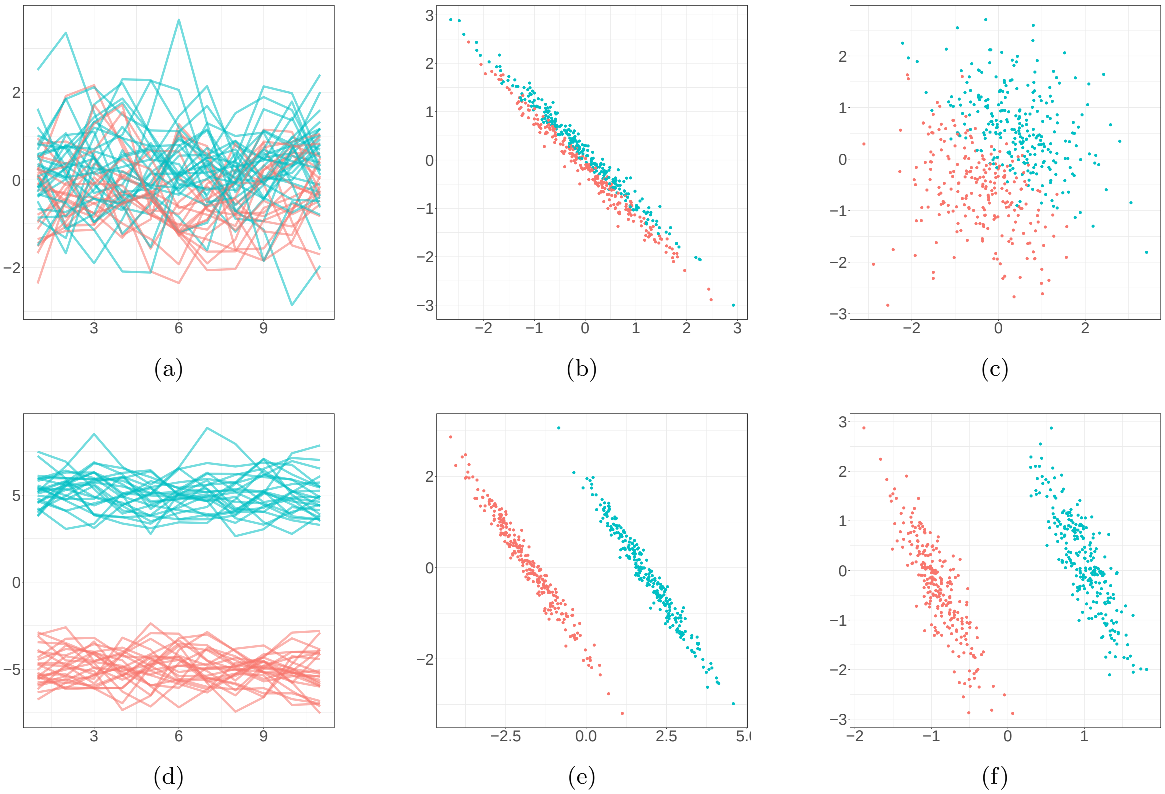
$$\mathbb{P}_{H_0^{\{\mathcal{C}_1,\mathcal{C}_2\}}}(\|\bar{W}_{\mathcal{C}_1} - \bar{W}_{\mathcal{C}_2}\|_F \geq \|\bar{w}_{\mathcal{C}_1} - \bar{w}_{\mathcal{C}_2}\|_F), \tag{3}$$

where $\bar{W}_{\mathcal{C}_1}$ is the sample mean of a realization for $W_i \sim \mathcal{MN}(\mu_i, \Sigma_1, \Sigma_2)$ for any $i \in \mathcal{C}_1$ and the same for $\bar{W}_{\mathcal{C}_2}$. However, this test fails to control the type-I error, namely, one might find the p-value as the trend to be 0 or 1, which is problematic in real practice.

We propose two ideas to handle this challenge. First, we use basis expansion regression to find a low-dimensional expression for the functional data. Specifically, we choose a certain basis, such as Hermite functions, and extract the coefficients of the basis regression; see Section 3.1 for more details. Second, we derive a selective p-value for the low-dimensional expression of the model (2) following Gao et al. (2022); Chen and Witten (2022), the selection procedure leverages the informative of post-clustering data and ensures that this selective p-value can control the type-I error. We provide theoretic results to guarantee that when $r_{ij} \to \infty$ (namely, the observed time points go to infinite), the proposed selective p-value controls the type-I error.

3 Selective Inference for Functional Data Clustering

In this section, we present the method of selective inference for functional data and propose a new statistic that controls the type-I error, i.e., the selective p-value, for testing the difference between post-clustering matrix data following the model (1), inspired by Gao et al. (2022). Our method is comprised of three parts, first, we impose the basis expansion regression to embed the



(a) (b) (c)

(d) (e) (f)

Figure 2: **Overall procedure. (a)** shows the first 100 curves of the dataset with the common mean $\mu_i = 0$ and the zero noise $\epsilon_i = 0$ labeled by the clustering results. **(b)** is the scatter plot (first two coordinates) of the low-dimensional embedding for this dataset. **(c)** is the scatter of the whitened dataset. **(d)** shows the first 100 curves of the dataset with the mean $\mu_i = -5$ or $\mu_i = 5$ and the zero noise term labeled by the clustering results. **(e)** and **(f)** are the scatter plots (first two coordinates) for the low-dimensional embedding and whitened dataset, respectively.

functional data W_i into low-dimensional vectors. Next, we estimate unknown covariance matrices of the embedded vectors. We then leverage these covariance matrices to conduct the whitening transformation and further use the Monte Carlo method (importance sampling) to approximate the proposed selective p-value.

Our method is based on the selective inference framework, which has been studied thoroughly in Benjamini and Bogomolov (2014); Fithian et al. (2014); Loftus and Taylor (2015); Taylor and Tibshirani (2015); Yang et al. (2016); Lee et al. (2016); Hyun et al. (2018). The key idea of this framework is adjusting the inferential process to account for the selection that has occurred, thereby providing statistically valid conclusions. In the context of hypothesis test (2), we want to test $\bar{\mu}_{\mathcal{C}_1} = \bar{\mu}_{\mathcal{C}_2}$ given $\mathcal{C}_1, \mathcal{C}_2 \in \mathcal{C}(\mathcal{W})$, where $\mathcal{W} := (w_i)_{i\in[m]}$ are realizations of patients follows the Assumption 1. Thus, we derive the selection procedure by constructing a p-value conditioning on the event $\{\mathcal{C}_1, \mathcal{C}_2 \in \mathcal{C}(\mathcal{W})\}$, namely,

$$\mathbb{P}_{H_0^{\{\mathcal{C}_1,\mathcal{C}_2\}}}\left(\|\bar{W}_{\mathcal{C}_1} - \bar{W}_{\mathcal{C}_2}\|_F \geq \|\bar{w}_{\mathcal{C}_1} - \bar{w}_{\mathcal{C}_2}\|_F \,\middle|\, \mathcal{C}_1, \mathcal{C}_2 \in \mathcal{C}(\mathcal{W})\right). \tag{4}$$

We remark this p-value is well defined when w_i is a $n \times T$ matrix for all $i \in [m]$, that is, there is no missing value in w_i. However, in the real-world application, w_i might contain missing values and $\bar{w}_{\mathcal{C}_1}, \bar{w}_{\mathcal{C}_2}$ are not well defined. Because the time records Ω_{ij} might be different for all patient i and feature j, and thus we cannot take the summation $\sum_{i=1}^m w_i$. For the case involving missing values, we complement these missing values by the following basis expansion regression and also find low-dimensional representations to calculate the selective p-value.

3.1 Low-dimensional embedding

The goal is to find a low-dimensional representation of the functional data, namely, represent (1) with a low-dimensional Gaussian matrix model (because time points T in the original model might be large and numerically infeasible). For the functional data, it is natural to consider the basis expansion regression, and we further take the coefficients as low-dimensional representations.

Recall the model (2), where W_i denotes the observed data of the ith patient with each row j denoting the jth feature. Given the record time points T and a record of feature $W_{ij} \in \mathbb{R}^{r_{ij}}$, where r_{ij} is the number of time points for this record. We choose q basis functions $\{\phi_s\}_{s\in[q]}$, where q is a user-specified positive integer, and impose the following ridge regression:

$$\underset{\alpha\in\mathbb{R}^q}{\arg\min} \left\{ \left\| W_{ij} - \sum_{s=1}^{q} \alpha_s \phi_s((t_{ij1}, \dots, t_{ijr_{ij}})/T) \right\|^2 + \lambda \|\alpha\|^2 \right\}, \tag{5}$$

where λ is a user-specified regularization term and $(t_{ij1}, \dots, t_{ijr_{ij}})/T$ is the linear transformation that scales the record of times into $[0, 1]$. We remark that T is the maximum record time of the data-collection period such as 120 hours in the examples mentioned earlier, thus $[0, 1]$ is the uniformly scaled time period for all patients i and features j. We use $\alpha_{ij} \in \mathbb{R}^q$ to denote the solution of (5) with respect to W_{ij}. Define the matrix $\Phi_{ij} := (\phi_s(t_{ijk}/T))_{s\in[q],k\in[r_{ij}]} \in \mathbb{R}^{q\times r_{ij}}$ where the $(s, k) \in [q] \times [r_{ij}]$ entry is $\phi_s(t_{ijk}/T)$ and define $K_{ij} := \Phi_{ij}\Phi_{ij}^\top \in \mathbb{R}^{q\times q}$, then α_{ij} has the following closed-form expression:

$$\alpha_{ij} = (K_{ij} + \lambda I_q)^{-1}\Phi_{ij}W_{ij}, \tag{6}$$

and the basis expansion function is $\widehat{\mu}_{ij} = \sum_{s=1}^q \alpha_{ijs}\phi_s$. Intuitively, the estimation error $\widehat{\mu}_{ij}(t_{ij}) - \mu_{ij}$ goes to zeros if certain regularity assumptions hold and $r_{ij} \to \infty$.

The linear transformation (6) embeds the functional data W_{ij} into a q-dimensional vector α_{ij}. Therefore, implementing this transformation for all $i \in [m]$ and $j \in [n]$ transforms the functional data $\mathcal{W}$ into a collection of q-dimensional vectors, which contain a tensor structure. To elaborate, given the basis $\{\phi_s\}_{s\in[q]}$ and the time record Ω, we define the linear map $H : \mathcal{W} \to \mathbb{R}^{m\times n\times q}$, where $H(\mathcal{W})_{i,j,:} := \alpha_{ij} = (K_{ij} + \lambda I_q)^{-1}\Phi_{ij}W_{ij}$, namely, the $(i, j, k) \in [m] \times [n] \times [q]$ entry of $H(\mathcal{W})$ is α_{ijk}. We will propose the selective p-value for the tensor $H(\mathcal{W})$ to conduct selective inference in Section 3.3.

We remark that the linear transformation H preserves the normality assumption. In more detail, the following lemma shows that each slide of $H(\mathcal{W})[i, :, :] \in \mathbb{R}^{n\times q}$ follows a matrix normal distribution for all $i \in [m]$. Also, the proposed low-dimensional embedding is scalable for non-functional features, such as weight and height in practical applications. We can impose Gaussian assumptions for these scalar features and incorporate them into the tensor $H(\mathcal{W})$, where the linearity of H implies the normality of all entries.

Lemma 1. *Suppose that $\mathcal{W}$ follows the model* (1)*, then $W_i \in \mathbb{R}^{n\times T}$ and Φ_{ij}, K_{ij} are the same for all $i \in [m], j \in [n]$. Define $\Phi := \Phi_{11} \in \mathbb{R}^{n\times T}, K := K_{11} \in \mathbb{R}^{q\times q}$, the proposed low-dimension representation $H(\mathcal{W})$ is a random tensor, where each slice is a random matrix satisfying the following matrix normal distribution*

$$\begin{aligned} H(\mathcal{W})[i,:,:] \sim \mathcal{MN}(\mu_i\Phi^\top(K+\lambda I_q)^{-1}, \Sigma_1, (K+\lambda I_q)^{-1}\Phi\Sigma_2\Phi^\top(K+\lambda I_q)^{-1}) \\ + \mathcal{MN}(0, \operatorname{diag}(\sigma_j^2)_{j\in[n]}, (K+\lambda I_q)^{-1}K(K+\lambda I_q)^{-1})). \end{aligned} \tag{7}$$

If $\mathcal{W}$ contains missing values characterized by the time record Ω. Define the blocked diagonal matrix $D_i := \operatorname{diag}((K_{ij}+\lambda I_q)^{-1}\Phi_{ij})_{j\in[n]} \in \mathbb{R}^{nq\times(\sum_{j=1}^n r_{ij})}$, then the vectorization of $H(\mathcal{W})[i,:,:]$ satisfies multivariate normal distribution corresponding with Ω_i:

$$vec(H(\mathcal{W})[i,:,:]) \sim \mathcal{N}(vec((\mu_{ij}^{\Omega_i}\Phi_{ij}^\top(K_{ij}+\lambda I_q)^{-1})_{j\in[n]}), D_i\left[\Sigma_2\otimes\Sigma_1 + I_T\otimes\operatorname{diag}(\sigma_j^2)_{j\in[n]}\right]^{\Omega_i})D_i^\top), \tag{8}$$

where $\mu_{ij}^{\Omega_i} := (\mu_{ij}(t_{ijk}))_{k\in[r_{ij}]}$ is the subvector of μ_{ij} characterized by the index t_{ij}, and $\left[\Sigma_2\otimes\Sigma_1 + I_T\otimes\operatorname{diag}(\sigma_j^2)_{j\in[n]}\right]^{\Omega_i})$ is the submatrix of $\Sigma_2\otimes\Sigma_1 + I_T\otimes\operatorname{diag}(\sigma_j^2)_{j\in[n]}$ characterized by the time record Ω_i (i.e. the submatrix that each coordinate corresponds to an observed time point).

Proof. See Appendix A.4 for detailed proofs. □

Lemma 1 implies that the linear transformation H maintains the normality. If there is no missing value, (7) indicates this linear transformation does not change the validity of the hypothesis test (2), because $\bar{\mu}_{\mathcal{C}_1} = \bar{\mu}_{\mathcal{C}_2} \iff \overline{H(\mu)}_{\mathcal{C}_1} = \overline{H(\mu)}_{\mathcal{C}_2}$. If there are missing values, intuitively, $D_i \to ((K+\lambda I_q)^{-1}\Phi)\otimes I_n$ when $T\to\infty$ and $r_{ij}/T \to 1$ for all $j\in[n]$, and then (8) is equivalent to (7). This further implies the linear transformation H approximately maintains the validity of the hypothesis test if the number of total time points is large and there are few missing values.

3.2 Covariance Estimation and Whitening

As aforementioned, the identity matrices assumption in Lemma 3 is not satisfied in general. Therefore, it is natural to impose the whitening transformation on the distribution (7) to change the covariance matrices into identity matrices. To elaborate, define $\Lambda = \left[(K+\lambda I_q)^{-1}\Phi\Sigma_2\Phi^\top(K+\lambda I_q)^{-1}\right]\otimes\Sigma_1 + \left[(K+\lambda I_q)^{-1}K(K+\lambda I_q)^{-1}\right]\otimes\operatorname{diag}(\sigma_j^2)_{j\in[n]}$, the distribution (7) would be $\operatorname{vec}(H(\mathcal{W})[i,:,:]) \sim \mathcal{N}(H(\mu_i), \Lambda)$. In this section, We will estimate the unknown covariance matrix Λ to impose the whitening transformation.

To estimate Λ, we notice that if the null hypothesis $H_0^{\{\mathcal{C}_1,\mathcal{C}_2\}}$ holds, we have $\bar{\mu}_{\mathcal{C}_1} = \bar{\mu}_{\mathcal{C}_2}$ which further implies $\overline{H(\mu)}_{\mathcal{C}_1} = \overline{H(\mu)}_{\mathcal{C}_2}$. As a result, given the fact that $\operatorname{vec}(H(\mathcal{W})[i,:,:]) \sim \mathcal{N}(H(\mu_i), \Lambda)$ for all $i\in[m]$, it is natural to use the sample covariance of the low-dimensional representations $H(\mathcal{W})$ as an estimator for the covariance matrix Λ, that is:

$$\widehat{\Lambda} := \frac{1}{m-1}\left[\sum_{i=1}^m(\operatorname{vec}(H(\mathcal{W})[i,:,:]) - \operatorname{vec}(\overline{H(\mathcal{W})}))(\operatorname{vec}(H(\mathcal{W})[i,:,:] - \operatorname{vec}(\overline{H(\mathcal{W})}))^\top\right],$$

where $\overline{H(\mathcal{W})}$ is the sample mean $\sum_{i=1}^m H(\mathcal{W})[i,:,:]/m$. Intuitively, if $H_0^{\{\mathcal{C}_1,\mathcal{C}_2\}}$ holds and the sample size $m\to\infty$, we have $\widehat{\Lambda}\to\Lambda$.

If the alternative hypothesis $H_1^{\{\mathcal{C}_1,\mathcal{C}_2\}}$ holds, i.e. the mean of two clusters are different ($\bar{\mu}_{\mathcal{C}_1} \neq \bar{\mu}_{\mathcal{C}_2}$), the proposed sample covariance estimator satisfies the following equation:

$$\begin{aligned}\widehat{\Lambda} = \Bigg(\sum_{j=1}^{2}\Bigg[&\sum_{i\in\mathcal{C}_j}(\mathrm{vec}(H(\mathcal{W})[i,:,:]) - \mathrm{vec}(\overline{H(\mathcal{W})}_{\mathcal{C}_j}))(\mathrm{vec}(H(\mathcal{W})[i,:,:]) - \mathrm{vec}(\overline{H(\mathcal{W})}_{\mathcal{C}_j}))^\top \\ &+ |\mathcal{C}_j|(\mathrm{vec}(\overline{H(\mathcal{W})}_{\mathcal{C}_j}) - \mathrm{vec}(\overline{H(\mathcal{W})}))(\mathrm{vec}(\overline{H(\mathcal{W})}_{\mathcal{C}_j}) - \mathrm{vec}(\overline{H(\mathcal{W})}))^\top\Bigg]\Bigg)/(m-1).\end{aligned} \tag{9}$$

Suppose $|\mathcal{C}_1|, |\mathcal{C}_2| \to \infty$ and $|\mathcal{C}_1|/|\mathcal{C}_2| \to c$, the above equation implies that $\widehat{\Lambda} \to \Lambda + c(\overline{H(\mu)}_{\mathcal{C}_1} - \overline{H(\mu)}_{\mathcal{C}_2})(\overline{H(\mu)}_{\mathcal{C}_1} - \overline{H(\mu)}_{\mathcal{C}_2})^\top/(c+1)^2$ by simple calculation. Therefore, the sample covariance estimator has a constant bias $c(\overline{H(\mu)}_{\mathcal{C}_1} - \overline{H(\mu)}_{\mathcal{C}_2})(\overline{H(\mu)}_{\mathcal{C}_1} - \overline{H(\mu)}_{\mathcal{C}_2})^\top/(c+1)^2$ which depends on c and $\overline{H(\mu)}_{\mathcal{C}_1} - \overline{H(\mu)}_{\mathcal{C}_2}$.

Whitening. In this step, we use linear transformation to change the covariance matrices of $H(\mathcal{W})[i,:,:]$ into identity matrices. Note that $\mathrm{vec}(H(\mathcal{W})[i,:,:]) \sim \mathcal{N}(\mathrm{vec}(H(\mu_i)), \Lambda)$ where $H(\mu_i) = \mu_i \Phi^\top (K + \lambda I_q)^{-1}$ and Λ is defined in Section 3.2, we consider the following linear transformation for each slices of $H(\mathcal{W})$:

$$\mathrm{vec}(H(\mathcal{W})[i,:,:]) \to \widehat{\Lambda}^{-\frac{1}{2}}\mathrm{vec}(H(\mathcal{W})[i,:,:]). \tag{10}$$

Intuitively, if $m \to \infty$, we have $\widehat{\Lambda} \to \Lambda$ and the covariance of the transformed matrix $\widehat{\Lambda}^{-\frac{1}{2}}\mathrm{vec}(H(\mathcal{W})[i,:,:])$ is approximately the identity matrix I_{nq}:

$$\widehat{\Lambda}^{-\frac{1}{2}}\mathrm{vec}(H(\mathcal{W})[i,:,:]) \sim \mathcal{N}(\widehat{\Lambda}^{-\frac{1}{2}}\mathrm{vec}(H(\mathcal{W})[i,:,:]), I_{nq}). \tag{11}$$

Thus, if the null hypothesis $H_0^{\{\mathcal{C}_1,\mathcal{C}_2\}}$ holds, the transformed data has the same population mean for partitions $\mathcal{C}_1$ and $\mathcal{C}_2$, i.e. $\widehat{\Lambda}^{-\frac{1}{2}}\mathrm{vec}(\overline{H(\mathcal{W})}_{\mathcal{C}_1} - \overline{H(\mathcal{W})}_{\mathcal{C}_2}) \sim \mathcal{N}(0, (1/|\mathcal{C}_1| + 1/|\mathcal{C}_2|)I_{nq})$. For convenience, we define the composition of the low-dimensional embedding H and whitening transformation (10) as the linear transformation $L : \mathcal{W} \to \mathbb{R}^{m\times n\times q}$, where $\mathrm{vec}(L(\mathcal{W})[i,:,:]) \coloneqq \widehat{\Lambda}^{-\frac{1}{2}}\mathrm{vec}(H(\mathcal{W})[i,:,:])$ for all $i \in [m]$. As aforementioned, when $m \to \infty$, we have

$$\overline{L(\mathcal{W})}_{\mathcal{C}_1} - \overline{L(\mathcal{W})}_{\mathcal{C}_2} \sim \sqrt{1/|\mathcal{C}_1| + 1/|\mathcal{C}_2|}\mathcal{MN}\left(0, I_n, I_q\right). \tag{12}$$

We will define the selective p-value (Definition 2) based on the transformed data $L(\mathcal{W})$, (17) implies that the proposed selective p-value can be rewritten as a truncated survival function of the $c \cdot \chi_{nq}$ distribution where c is a positive constant (Lemma 3).

3.3 Selective p-value

Suppose w is a realization of the model (1) and $L(w)$ is the aforementioned transformed data based on the low-dimensional embedding and whitening steps. Next, the user might apply clustering algorithms on $L(w)$ to separate patients into 2 clusters where the corresponding non-intersect partitions for $[m]$ are denoted by $\mathcal{C}_1, \mathcal{C}_2$, and our goal is to test the group mean $\overline{L(\mathcal{W})}_{\mathcal{C}_1} = \overline{L(\mathcal{W})}_{\mathcal{C}_2}$. While the classical hypothesis test methods fail to control the type-I error due to the selection bias, we consider the following selective p-value conditioned on the clustering information:

$$\mathbb{P}_{H_0^{\{\mathcal{C}_1,\mathcal{C}_2\}}}\left(\|\overline{L(\mathcal{W})}_{\mathcal{C}_1} - \overline{L(\mathcal{W})}_{\mathcal{C}_2}\|_F \geq \|\overline{L(w)}_{\mathcal{C}_1} - \overline{L(w)}_{\mathcal{C}_2}\|_F \,\middle|\, \mathcal{C}_1, \mathcal{C}_2 \in \mathcal{C}(L(\mathcal{W}))\right), \tag{13}$$

Intuitively, the proposed p-value is the conditional probability of the difference in cluster means among all realizations of the model (1) that has the same partition $\mathcal{C}_1, \mathcal{C}_2$ as the observed data. Fithian et al. (2014) shows that the selective p-value leverages the model selection information to reduce the selection bias and thus controls the selective type-I error.

Definition 1. *(Post-clustering selective type-I error). Suppose that $\mathcal{W}$ follows the model (1) and w is a realization of $\mathcal{W}$. Suppose that a specific clustering algorithm has been applied to w, yielding a non-intersect partition $\mathcal{C}_1, \mathcal{C}_2$ satisfies $\mathcal{C}_1 \cup \mathcal{C}_2 = \{1, 2, \ldots, m\}$. Let $H_0^{\{\mathcal{C}_1,\mathcal{C}_2\}}$ be the null hypothesis defined as (2), we say a test of $H_0^{\{\mathcal{C}_1,\mathcal{C}_2\}}$ based on $\mathcal{W}$ controls the selective type-I error for clustering at level α if*

$$\mathbb{P}_{H_0^{\{\mathcal{C}_1,\mathcal{C}_2\}}} \left(\textit{reject } H_0^{\{\mathcal{C}_1,\mathcal{C}_2\}} \textit{ based on } \mathcal{W} \textit{ at level } \alpha \,\middle|\, \mathcal{C}_1, \mathcal{C}_2 \in \mathcal{C}(L(\mathcal{W})) \right) \leq \alpha \tag{14}$$

for any $\alpha \in [0, 1]$.

However, while the proposed selective p-value (13) controls the selective type-I error, this p-value cannot be directly calculated, because the condition $\mathcal{C}_1, \mathcal{C}_2 \in \mathcal{C}(L(\mathcal{W}))$ is numerically infeasible. To elaborate, the conditional probability (13) is the probability mass function $(\|\overline{L(\mathcal{W})}_{\mathcal{C}_1} - \overline{L(\mathcal{W})}_{\mathcal{C}_2}\|_F \geq \|\overline{L(w)}_{\mathcal{C}_1} - \overline{L(w)}_{\mathcal{C}_2}\|_F)$ of the matrix normal distribution truncated onto the set $\{\mathcal{W} \in \mathbb{R}^{m\times n\times T} : \mathcal{C}_1, \mathcal{C}_2 \in \mathcal{C}(L(\mathcal{W}))\}$, and directly computing this set is impossible. To handle this problem, we propose the following orthogonal decomposition for tensor inspired by Gao et al. (2022). To begin with, we define the indicator vector:

$$\nu(\mathcal{C}_1, \mathcal{C}_2) := \left(\frac{\mathbb{1}_{i\in\mathcal{C}_1}}{|\mathcal{C}_1|} - \frac{\mathbb{1}_{i\in\mathcal{C}_2}}{|\mathcal{C}_2|} \right)_{i\in[m]},$$

where each coordinate of $\nu(\mathcal{C}_1, \mathcal{C}_2)$ is $1/|\mathcal{C}_1|$ if $i \in \mathcal{C}_1$ and $1/|\mathcal{C}_2|$ if $i \in \mathcal{C}_2$. Given this definition, we can rewrite the difference of cluster means by the tensor mode product. In more detail, we have $\nu(\mathcal{C}_1, \mathcal{C}_2) \times_1 L(\mathcal{W}) = \overline{L(\mathcal{W})}_{\mathcal{C}_1} - \overline{L(\mathcal{W})}_{\mathcal{C}_2}$.

Lemma 2. *(Orthogonal decomposition). For any tensor $\mathcal{A} \in \mathbb{R}^{m\times n\times q}$ and any partition of $[m]$ denoted by $\mathcal{C}_1, \mathcal{C}_2$, we have the following decomposition:*

$$\mathcal{A} = \pi^{\perp}_{\nu(\mathcal{C}_1,\mathcal{C}_2)} \times_1 \mathcal{A} + \left(\frac{\|\bar{\mathcal{A}}_{\mathcal{C}_1} - \bar{\mathcal{A}}_{\mathcal{C}_2}\|_F}{1/|\mathcal{C}_1| + 1/|\mathcal{C}_2|} \right) \nu(\mathcal{C}_1, \mathcal{C}_2) \otimes dir(\bar{\mathcal{A}}_{\mathcal{C}_1} - \bar{\mathcal{A}}_{\mathcal{C}_2})^{\top}, \tag{15}$$

where $\bar{\mathcal{A}}_{\mathcal{C}_i} = \sum_{j\in\mathcal{C}_i} \mathcal{A}_{j,:,:}/|\mathcal{C}_i|$ is the mean of mode-1 slices corresponding to the partition $\mathcal{C}_i$, $\times_1$ denotes the tensor mode-1 product, $\pi^{\perp}_{\nu} = I - \frac{\nu\nu^{\top}}{\|\nu\|^2}$ is an orthogonal projection matrix, and $dir(\omega) = \frac{\omega}{\|\omega\|_F}\mathbb{1}_{\{\omega\neq 0\}}$ is the direction of ω (here ω is a matrix, $\|\omega\|_F$ is its Frobenius norm, and $\mathbb{1}_{\{\omega\neq 0\}}$ is the indicator function takes the value 0 when all the entries in ω are zero and takes the value 1 otherwise).

Proof. See Appendix A.5 for detailed proof. □

To compute (13), which is conditional on $\mathcal{C}_1, \mathcal{C}_2 \in \mathcal{C}(L(\mathcal{W}))$, the orthogonal decomposition (Lemma 2) implies that we need additional information on $\pi^{\perp}_{\nu(\mathcal{C}_1,\mathcal{C}_2)}$ and $\mathrm{dir}(\overline{L(\mathcal{W})}_{\mathcal{C}_1} - \overline{L(\mathcal{W})}_{\mathcal{C}_2})$. Given this intuition, we consider the following selective p-value with additional conditions on $\pi^{\perp}_{\nu(\mathcal{C}_1,\mathcal{C}_2)}$ and $\mathrm{dir}(\overline{L(\mathcal{W})}_{\mathcal{C}_1} - \overline{L(\mathcal{W})}_{\mathcal{C}_2})$.

Definition 2. *(Selective p-value). Suppose that $\mathcal{W}$ follows the model* (1) *and w is a realization of $\mathcal{W}$. Suppose that a specific clustering algorithm is applied on $L(w)$, yielding a non-intersect partition $\mathcal{C}_1, \mathcal{C}_2$ satisfies $\mathcal{C}_1 \cup \mathcal{C}_2 = \{1, 2, \ldots, m\}$. Let $H_0^{\{\mathcal{C}_1,\mathcal{C}_2\}}$ be the null hypothesis defined as (2), we propose the following selective p-value:*

$$
\begin{aligned}
p_{selective} = \mathbb{P}_{H_0^{\{\mathcal{C}_1,\mathcal{C}_2\}}} \Bigg(\|\overline{L(\mathcal{W})}_{\mathcal{C}_1} - \overline{L(\mathcal{W})}_{\mathcal{C}_2}\|_F \geq \|\overline{L(w)}_{\mathcal{C}_1} - \overline{L(w)}_{\mathcal{C}_2}\|_F \Bigg| \mathcal{C}_1, \mathcal{C}_2 \in \mathcal{C}(L(\mathcal{W})), \\
\pi^{\perp}_{\nu(\mathcal{C}_1,\mathcal{C}_2)} \times_1 L(\mathcal{W}) = \pi^{\perp}_{\nu(\mathcal{C}_1,\mathcal{C}_2)} \times_1 L(w), dir(\overline{L(\mathcal{W})}_{\mathcal{C}_1} - \overline{L(\mathcal{W})}_{\mathcal{C}_2}) = dir(\overline{L(w)}_{\mathcal{C}_1} - \overline{L(w)}_{\mathcal{C}_2}) \Bigg),
\end{aligned}
\tag{16}
$$

Next, we show that the selective p-value (16) is numerically feasible. To elaborate, under the null hypothesis $H_0^{\{\mathcal{C}_1,\mathcal{C}_2\}}$, (12) implies that $\|\overline{L(\mathcal{W})}_{\mathcal{C}_1} - \overline{L(\mathcal{W})}_{\mathcal{C}_2}\|_F$ follows the distribution $\sqrt{1/|\mathcal{C}_1| + 1/|\mathcal{C}_2|}\chi_{nq}$ when the sample size m is large (i.e. $\widehat{\Lambda} \to \Lambda$ when $m \to \infty$). In this case, the following lemma shows this selective p-value is characterized by a survival function of the distribution $\sqrt{1/|\mathcal{C}_1| + 1/|\mathcal{C}_2|} \cdot \chi_{nq}$ truncated to a set $\mathcal{S}(w; \mathcal{C}_1, \mathcal{C}_2) \subset \mathbb{R}$, and it allows us to use the Monte Carlo methods to approximate the selective p-value numerically (Section 3.4).

Lemma 3. *Suppose that $\mathcal{W}$ follows the model* (1) *and w is a realization of $\mathcal{W}$, then $L(w)$ is a realization of $L(\mathcal{W})$ and the proposed selective p-value* (16) *can be rewritten as follows:*

$$
p_{selective} = 1 - \mathbb{F}\left(\|\overline{L(w)}_{\mathcal{C}_1} - \overline{L(w)}_{\mathcal{C}_2}\|_F; \sqrt{\frac{1}{|\mathcal{C}_1|} + \frac{1}{|\mathcal{C}_2|}}, \mathcal{S}(w; \mathcal{C}_1, \mathcal{C}_2) \right), \tag{17}
$$

where $\mathbb{F}(t; c, \mathcal{S})$ denotes the cumulative distribution of a $c \cdot \chi_{nq}$ random variable truncated to the set $\mathcal{S}$ defined by

$$
\mathcal{S}(w; \mathcal{C}_1, \mathcal{C}_2) := \left\{ \varphi \geq 0 : \mathcal{C}_1, \mathcal{C}_2 \in \mathcal{C}\left(\pi^{\perp}_{\nu(\mathcal{C}_1,\mathcal{C}_2)} \times_1 L(w) + \left[\frac{\varphi}{\frac{1}{|\mathcal{C}_1|} + \frac{1}{|\mathcal{C}_2|}} \right] \nu(\mathcal{C}_1, \mathcal{C}_2) \otimes dir(\overline{L(w)}_{\mathcal{C}_1} - \overline{L(w)}_{\mathcal{C}_2})^{\top} \right) \right\}.
$$

Proof. See Appendix A.6 for detailed proof. □

3.4 Numerical approximation of Selective p-value

Now we describe the procedure to calculate the selective p-value (16), which is equivalent to the truncated survival function (17). In the following, we will use the Monte Carlo method to approximate the truncation set $\mathcal{S}(w; \mathcal{C}_1, \mathcal{C}_2)$ and further calculate the survival function.

To begin with, we briefly talk about the intuitive explanation for $\mathcal{S}(w; \mathcal{C}_1, \mathcal{C}_2)$. Given a partition $\mathcal{C}_1$ and $\mathcal{C}_2$ obtained by a certain clustering algorithm, we consider the linear transformation $F : \mathbb{R} \to \mathbb{R}^{m \times n \times q}$:

$$
F(\varphi) := \pi^{\perp}_{\nu(\mathcal{C}_1,\mathcal{C}_2)} \times_1 L(w) + \left[\frac{\varphi}{\frac{1}{|\mathcal{C}_1|} + \frac{1}{|\mathcal{C}_2|}} \right] \nu(\mathcal{C}_1, \mathcal{C}_2) \otimes \mathrm{dir}\left(\overline{L(w)}_{\mathcal{C}_1} - \overline{L(w)}_{\mathcal{C}_2} \right)^{\top}. \tag{18}
$$

Intuitively, the transformation F operates the orthogonal projection $\pi^{\perp}_{\nu(\mathcal{C}_1,\mathcal{C}_2)} \times_1 L(w)$ along a "vector" $\nu(\mathcal{C}_1, \mathcal{C}_2) \otimes \mathrm{dir}(\overline{L(w)}_{\mathcal{C}_1} - \overline{L(w)}_{\mathcal{C}_2})^{\top}$ with the length φ. And the set $\mathcal{S}(w; \mathcal{C}_1, \mathcal{C}_2)$ contains all the "length" $\varphi \in \mathbb{R}$ such that the transformed tensor has the same clustering outputs as $L(w)$ (i.e. the partition is equal to $\mathcal{C}_1, \mathcal{C}_2$). Recalling Lemma 2, when $\varphi = \|\overline{L(w)}_{\mathcal{C}_1} - \overline{L(w)}_{\mathcal{C}_2}\|_F$, we have

$$L(w) = \pi^{\perp}_{\nu(\mathcal{C}_1,\mathcal{C}_2)} \times_1 L(w) + \left[\frac{\|\overline{L(w)}_{\mathcal{C}_1} - \overline{L(w)}_{\mathcal{C}_2}\|_F}{\frac{1}{|\mathcal{C}_1|} + \frac{1}{|\mathcal{C}_2|}}\right] \nu(\mathcal{C}_1, \mathcal{C}_2) \otimes \operatorname{dir}\left(\overline{L(w)}_{\mathcal{C}_1} - \overline{L(w)}_{\mathcal{C}_2}\right)^{\top},$$

Therefore, we obtain $\|\overline{L(w)}_{\mathcal{C}_1} - \overline{L(w)}_{\mathcal{C}_2}\|_F \in \mathcal{S}(w; \mathcal{C}_1, \mathcal{C}_2)$. This further implies that when $\varphi > \mathcal{S}(w; \mathcal{C}_1, \mathcal{C}_2)$, the transformation F "push away" the sets $\{L(w)[i,:,:]\}_{i\in\mathcal{C}_1}$ and $\{L(w)[i,:,:]\}_{i\in\mathcal{C}_2}$ along the vector $\nu(\mathcal{C}_1, \mathcal{C}_2) \otimes \operatorname{dir}(\overline{L(w)}_{\mathcal{C}_1} - \overline{L(w)}_{\mathcal{C}_2})^{\top}$, and vice versa. If φ is too large or small, the sets $\{L(w)[i,:,:]\}_{i\in\mathcal{C}_1}$ and $\{L(w)[i,:,:]\}_{i\in\mathcal{C}_2}$ will lead to different clustering results (i.e. the partition would not be $\mathcal{C}_1, \mathcal{C}_2$). As a result, elements in the set $\mathcal{S}(w; \mathcal{C}_1, \mathcal{C}_2)$ might concentrate near $\|\overline{L(w)}_{\mathcal{C}_1} - \overline{L(w)}_{\mathcal{C}_2}\|_F$.

Monte Carlo Approximation. We use the Monte Carlo method to approximate the survival function (17), which is the survival function of the distribution $\sqrt{1/|\mathcal{C}_1| + 1/|\mathcal{C}_2|} \cdot \chi_{nq}$ truncated on the set $\mathcal{S}(w; \mathcal{C}_1, \mathcal{C}_2)$. Although this set does not have a closed-form expression, we can sample some $\varphi \in \mathbb{R}$ and check if $\varphi \in \mathcal{S}(w; \mathcal{C}_1, \mathcal{C}_2)$ to approximate this set. Mathematically, we rewrite (17) as follows:

$$p_{selective} = \frac{\mathbb{P}(\varphi \geq \|\overline{L(w)}_{\mathcal{C}_1} - \overline{L(w)}_{\mathcal{C}_2}\|_F, \mathcal{C}_1, \mathcal{C}_2 \in \mathcal{C}(F(\varphi)))}{\mathbb{P}(\mathcal{C}_1, \mathcal{C}_2 \in \mathcal{C}(F(\varphi)))} = \frac{\mathbb{E}[\mathbb{1}_{\{\varphi \geq \|\overline{L(w)}_{\mathcal{C}_1} - \overline{L(w)}_{\mathcal{C}_2}\|_F, \mathcal{C}_1, \mathcal{C}_2 \in \mathcal{C}(F(\varphi))\}}]}{\mathbb{E}[\mathbb{1}_{\{\mathcal{C}_1, \mathcal{C}_2 \in \mathcal{C}(F(\varphi))\}}]},$$

where φ follows the distribution $\sqrt{1/|\mathcal{C}_1| + 1/|\mathcal{C}_2|} \cdot \chi_{nq}$, $\mathbb{P}$ is the corresponding probability mass function, and $\mathbb{E}$ is the expectation with respect to φ.

To reduce the computational complexity, we use the importance sampling technique to approximate this conditional probability. As aforementioned, $\mathcal{S}(w; \mathcal{C}_1, \mathcal{C}_2)$ might concentrate near $\|\overline{L(w)}_{\mathcal{C}_1} - \overline{L(w)}_{\mathcal{C}_2}\|_F$. Therefore, we Define $g(x) = f_1(x)/f_2(x)$, where f_1 is the density of $\sqrt{1/|\mathcal{C}_1| + 1/|\mathcal{C}_2|} \cdot \chi_{nq}$ and f_2 is the density function of $\mathcal{N}\left(\|\overline{L(w)}_{\mathcal{C}_1} - \overline{L(w)}_{\mathcal{C}_2}\|_F, 1/|\mathcal{C}_1| + 1/|\mathcal{C}_2|\right)$. For a positive integer S, we sample S values $\gamma_1, \ldots, \gamma_S \sim \mathcal{N}(\|\overline{L(w)}_{\mathcal{C}_1} - \overline{L(w)}_{\mathcal{C}_2}\|_F, 1/|\mathcal{C}_1| + 1/|\mathcal{C}_2|)$, then the approximation of selective p-value would be:

$$p_{selective} \approx \frac{\sum \pi_i \mathbb{1}_{\{\gamma_i \geq \|\overline{L(w)}_{\mathcal{C}_1} - \overline{L(w)}_{\mathcal{C}_2}\|_F, \mathcal{C}_1, \mathcal{C}_2 \in \mathcal{C}(F(\gamma_i))\}}}{\sum \pi_i \mathbb{1}_{\{\mathcal{C}_1, \mathcal{C}_2 \in \mathcal{C}(F(\gamma_i))\}}}, \qquad \pi_i = \frac{f_1(\gamma_i)}{f_2(\gamma_i)}. \tag{19}$$

Overall Procedures We summarize the three steps for computing the selective p-value as an overall procedure in Algorithm 1. We will keep the notation as defined in the previous sections.

4 Theoretical Guarantees

In this section, we present theoretical results for the proposed selective p-value. We are going to prove that the p-value (16) controls the selective type-I error when the covariance matrix Λ is known (Theorem 1). Moreover, we present the asymptotic result for the statistical power of the proposed selective inference framework. Specifically, we show that the power converges to 1 when the sample size and the difference in clusters mean increase (Theorem 3).

Algorithm 1: Selective inference for functional data

```
Step I: Low-dimensional embedding;
Input: Data of $m$ patients $\mathcal{W}$, time record $\Omega$, basis functions $\{\phi_s\}_{s\in[q]}$, regularization
       term $\lambda$.
1. Compute the matrices $K_{ij}$ and $\Phi_{ij}$ by $\Omega$ and $\{\phi_s\}_{s\in[q]}$;
2. (Basis expansion regression). $X_i \leftarrow (K_{ij} + \lambda I_q)^{-1}\Phi_{ij}W_{ij}$, for $i \in [m]$;
Output: Low-dimensional embedding $\{X_i\}_{i\in[m]}$.

Step II: Covariance estimation;
Input: Low-dimensional embedding $\{X_i\}_{i\in[m]}$.
3. Compute the sample covariance $\widehat{\Lambda} := \frac{\sum_{i=1}^{m}(\mathrm{vec}(X_i)-\mathrm{vec}(\bar{X}))(\mathrm{vec}(X_i)-\mathrm{vec}(\bar{X}))^\top}{m-1}$;
Output: Estimated covariance matrix $\widehat{\Lambda}$.

Step III: Whitening and Clustering;
Input: Low-dimensional embedding $\{X_i\}_{i\in[m]}$, covariance matrix $\widehat{\Lambda}$.
4. (Whitening). Conduct the linear transformation $\mathrm{vec}(Y_i) \leftarrow (\widehat{\Lambda})^{-\frac{1}{2}}\mathrm{vec}(X_i)$;
5. (Clustering). Apply certain clustering algorithm on $\{Y_i\}_{i\in[m]}$ and obtain a partition
   $\mathcal{C}_1, \mathcal{C}_2$, where the number of clusters is 2.;
Output: Whitened data $\{Y_i\}_{i\in[m]}$, partition $\mathcal{C}_1, \mathcal{C}_2$.

Step IV: Numerical approximation of the selective $p$-value;
Input: Whitened data $\{Y_i\}_{i\in[m]}$, partition $\mathcal{C}_1, \mathcal{C}_2$, sampling horizon $S$.
for $s = 1 \to S$ do
    6. Generate $\gamma_s \sim \sqrt{\frac{1}{|\mathcal{C}_1|} + \frac{1}{|\mathcal{C}_2|}} \cdot \chi_{nq}$, compute $\pi_s = \frac{f_1(\gamma_s)}{f_2(\gamma_s)}$;
    7. Apply the same clustering algorithm to obtain the partition $\mathcal{C}(F(\gamma_s))$;
end
Output: Selective $p$-value $\frac{\sum_{s=1}^{S} \pi_s \mathbb{1}_{\{\omega_s \geq \|\bar{Y}_{\mathcal{C}_1} - \bar{Y}_{\mathcal{C}_2}\|, \mathcal{C}_1, \mathcal{C}_2 \in \mathcal{C}(F(\gamma_s))\}}}{\sum_{s=1}^{S} \pi_s \mathbb{1}_{\{\mathcal{C}_1, \mathcal{C}_2 \in \mathcal{C}(F(\gamma_s))\}}}$.
```

To begin with, we present the theoretical guarantee about the selective type-I error. Intuitively, (17) shows that the selective p-value is a truncated continuous survival function of the distribution $\sqrt{1/|\mathcal{C}_1| + 1/|\mathcal{C}_2|} \cdot \chi_{nq}$ when the covariance matrix Λ is known (i.e. we use Λ to conduct the whiten transformation). Therefore, the selective p-value would follow the uniform distribution and thus control the selective type-I error.

Theorem 1. *(Selective Type-I error control). Suppose that $\mathcal{W}$ follows the model (1) and the covariance matrices are known. Suppose that w is a realization of $\mathcal{W}$, and $\mathcal{C}_1, \mathcal{C}_2$ is the non-intersect partition of $[m]$ obtained by a clustering algorithm $\mathcal{C}(\cdot)$ on $L(w)$. If the null hypothesis $H_0^{\{\mathcal{C}_1,\mathcal{C}_2\}}$ holds, then for all $\alpha \in [0,1]$ and, the selective type-I error is controlled by α:*

$$\mathbb{P}_{H_0^{\{\mathcal{C}_1,\mathcal{C}_2\}}}\left(p(\mathcal{W};\mathcal{C}_1,\mathcal{C}_2) \leq \alpha \middle| \, \mathcal{C}_1, \mathcal{C}_2 \in \mathcal{C}(L(\mathcal{W}))\right) = \alpha. \tag{20}$$

Proof. The key is to prove that the selective p-value follows the uniform distribution under the clustering condition, namely, $p(\mathcal{W};\mathcal{C}_1,\mathcal{C}_2)\big|\ \mathcal{C}_1,\mathcal{C}_2 \in \mathcal{C}(L(\mathcal{W})) \sim \text{Unif}(0,1)$. See Appendix A.1 for detailed proof. □

We remark that the statement in Theorem 1 cannot be verified by simulation efficiently because there is a small probability of generating a $\mathcal{W}$ follows the model (1) that satisfies $\mathcal{C}_1,\mathcal{C}_2 \in \mathcal{C}(L(\mathcal{W}))$. Next, we present an analogous theorem that can be verified by simulation studies, where we present the numerical results in Section 5.

Theorem 2. *Suppose that $\mathcal{W}$ follows the model* (1) *and the covariance matrices are known, suppose that $\mathcal{C}_1^{\mathcal{W}},\mathcal{C}_2^{\mathcal{W}}$ is a non-intersect partition of $[m]$ obtained by a clustering algorithm $C(\cdot)$ on $L(\mathcal{W})$. Define $\bar{\mu}_{\mathcal{C}_1^{\mathcal{W}}},\bar{\mu}_{\mathcal{C}_2^{\mathcal{W}}}$ as the population means corresponding to the two clusters, then the following property holds for all $\alpha \in [0,1]$:*

$$\mathbb{P}\left(p(\mathcal{W};\mathcal{C}_1^{\mathcal{W}},\mathcal{C}_2^{\mathcal{W}}) \le \alpha\big|\ \bar{\mu}_{\mathcal{C}_1^{\mathcal{W}}} = \bar{\mu}_{\mathcal{C}_2^{\mathcal{W}}}\right) = \alpha. \tag{21}$$

Proof. See Appendix A.2 for detailed proof. □

Now we present the theorem for the statistical power, we briefly talk about the intuition that the statistical power will converge to 1 asymptotically. Under the alternative hypothesis $H_1^{\{\mathcal{C}_1,\mathcal{C}_2\}}$, if $|\mathcal{C}_1|,|\mathcal{C}_2| \to \infty$, $|\mathcal{C}_1|/|\mathcal{C}_2| \to c$, and $\|\bar{\mu}_{\mathcal{C}_1} - \bar{\mu}_{\mathcal{C}_2}\|_F \to \infty$, then the low-ranking embedding (6) satisfies $\|\overline{H(\mu)}_{\mathcal{C}_1} - \overline{H(\mu)}_{\mathcal{C}_2}\|_F \to \infty$. Recalling the covariance estimator $\widehat{\Lambda}$ defined by (9), we have $\widehat{\Lambda} \to \Lambda + c(\overline{H(\mu)}_{\mathcal{C}_1} - \overline{H(\mu)}_{\mathcal{C}_2})(\overline{H(\mu)}_{\mathcal{C}_1} - \overline{H(\mu)}_{\mathcal{C}_2})^\top/(c+1)^2$. Intuitively, combining this with the whitening transformation (11) yields that $\|\overline{L(\mathcal{W})}_{\mathcal{C}_1} - \overline{L(\mathcal{W})}_{\mathcal{C}_2}\|_F \to (c+1)/\sqrt{c}$. Therefore, given a realization w, the asymptotic property together with Lemma 3 indicates that the selective p-value is $1 - \mathbb{F}((c+1)/\sqrt{c}; \sqrt{1/|\mathcal{C}_1| + 1/|\mathcal{C}_2|}, \mathcal{S}(w;\mathcal{C}_1,\mathcal{C}_2))$, which converges to 1 when the sample size m increases.

Theorem 3. *(Statistical power). Suppose that $\mathcal{W}$ follows the model* (1) *and w is a realization of $\mathcal{W}$. Suppose that $\mathcal{C}_1,\mathcal{C}_2$ is the non-intersect partition of $[m]$ obtained by a clustering algorithm $\mathcal{C}(\cdot)$ on $L(w)$. If the alternative hypothesis $H_1^{\{\mathcal{C}_1,\mathcal{C}_2\}}$ holds, if the sample size and difference between clusters mean increase, and the clusters are asymptotically balanced (i.e. $|\mathcal{C}_1|/|\mathcal{C}_2| \to c \in (0,1)$), then for all $\alpha \in [0,1]$ the statistical power of the proposed selective inference framework leveraging* (17) *converges to 1:*

$$\lim_{m\to\infty}\ \lim_{\|\bar{\mu}_{\mathcal{C}_1}-\bar{\mu}_{\mathcal{C}_2}\|_F\to\infty} \mathbb{P}_{H_1^{\{\mathcal{C}_1,\mathcal{C}_2\}}}\left(p(\mathcal{W};\mathcal{C}_1,\mathcal{C}_2) \le \alpha\big|\ \mathcal{C}_1,\mathcal{C}_2 \in \mathcal{C}(L(\mathcal{W}))\right) = 1.$$

Proof. See Appendix A.3 for detailed proof. □

5 Simulation Studies

In this section, we conduct experiments on synthetic data to verify our theory. We check the selective type-I error in Section 5.1 and the statistical power in Section 5.2. Next, we study the robustness of the proposed selective inference framework in Section 5.3. Due to the page limit, we present the basic setup and discussion of the experiments in this section, and put the figures of experiments in Appendix B.

5.1 Selective type-I error under a global null

In this section, we present numerical results to verify Theorem 2. To elaborate, we generate data under a global null and compute the corresponding selective p-value.

Basic Setup. We generate a dataset containing 100 data following the model (1), where each data contains $m = 50000$ patients. Namely, for all $k \in [100]$ and $i \in [m]$, we generate $X_i^{(k)} = Z_i^{(k)} + \epsilon_i^{(k)}$ where $Z_i^{(k)} \sim \mathcal{MN}(\mu_i, \Sigma_1, \Sigma_2)$ and $\epsilon_i^{(k)} \sim \sigma \cdot \mathcal{MN}(0, I_n, I_q)$ (here ϵ_i is a $n \times T$ matrix with each entry as i.i.d. noise terms following $\mathcal{N}(0, \sigma^2)$). Specifically, we set the number of features n as 2, and the number of total time points T as 15. Also, we set $\mu_i = 0_{n\times T}$, $\sigma^2 = 0.5$, $\Sigma_1 = I_n$, and set Σ_2 as the covariance matrix of a certain kernel K (i.e. $\Sigma_2 = (K(i/T, j/T))_{i,j\in[T]}$). To elaborate, we conduct the simulation for three different kernels: the rational quadratic kernel, period kernel, and truncated local period kernel:

(i) Rational quadratic kernel.

$$K(x, y) = \left(1 + \frac{(x-y)^2}{\ell^2}\right)^{-1/2}.$$

(ii) Period kernel.

$$K(x, y) = e^{-8\sin^2(2\pi|x-y|)}.$$

(iii) Truncated local period kernel.

$$K(x, y) = \mathbb{1}_{\{1/3<|x-y|<2/3\}} \cdot e^{-8\sin^2(2\pi|x-y|)} e^{-2(x-y)^2} + \mathbb{1}_{\{|x-y|\le 1/3 \text{ or } |x-y|\ge 2/3\}} \cdot 0.01.$$

Next, we set the basis functions $\{\phi_s\}_{s\in[q]}$ as the eigenfunctions of the Gaussian RBF (Radial Basis Function) kernel $K(x, y) = e^{-\frac{\rho}{1-\rho^2}(x-y)^2}$ where $\rho \in (0, 1)$. By the Mercer expansion (Fasshauer and McCourt, 2012), the i-th eigenfunction is

$$\phi_i(x) = \frac{1}{\sqrt{N_i}} H_i(x) e^{-\frac{\rho}{1+\rho}x^2}, \tag{22}$$

here $N_i = 2^i i! \sqrt{\frac{1-\rho}{1+\rho}}$ and $H_i(x)$ is the i-th order physicist's Hermite polynomial. In this experiment, we set $\rho = 0.99$ and set the truncation number q as 3, namely, we use the first three eigenfunctions to conduct the low-dimensional embedding.

Next, we apply the proposed selective inference framework to the generated datasets. Figure 4 shows quantile plots of the selective p-value for datasets corresponding to the aforementioned three kernels, it shows that the selective p-value follows the uniform distribution under the global null hypothesis, which validates the statement of Theorem 2.

5.2 Statistical Power

In this section, we present the numerical results to verify Theorem 3. In more detail, we generate datasets following the model (1) under the alternative hypothesis and compute the corresponding statistical power. To verify Theorem 3, we compute the selective p-value with respect to datasets generated by different cluster mean $\|\bar{\mu}_{\mathcal{C}_1} - \bar{\mu}_{\mathcal{C}_2}\|_F$ and sample size m.

To check the statistical power with respect to different sample sizes, we consider the sample size as $m = 10 \cdot k$ where $k \in \{3, 4, \ldots, 10\}$. Set $n = 1$, $T = 15$, $\sigma^2 = 0.1$, and $\Sigma_1 = I_n$, for each sample size m, we generate a dataset containing m record following a model of alternative hypothesis: $\mu_i = (-10)_{n\times T}$ for $i \leq m/2$ and $\mu_i = (10)_{n\times T}$ for $i > m/2$. We use the same basis (22) with the parameter $q = 3$ to conduct the low-dimensional embedding. Figure 5(b) presents the statistical power with the same mean and increasing sample sizes, it shows that the statistical power increases as the sample sizes increase.

To check the statistical power with respect to different cluster means, we set the sample size as $m = 60$ and the other parameters are the same as in the previous paragraph.For each $k \in \{0.3, 0.4, \ldots, 1\}$, we generate a dataset with m records following the sample means $\mu_i = (k)_{n\times T}$ for $i \leq m/2$ and $\mu_i = (-k)_{n\times T}$ for $i > m/2$. Figure 5(c) presents the statistical power with the same sample size and the increasing difference between cluster means, it shows that the statistical power increases as the difference between cluster means increases.

5.3 Empirical Robustness Analysis

Robustness to missing values. We consider the practical situations, where there are missing values on observed records. To elaborate, we set $n = 1, T = 15, \sigma^2 = 0.1, \Sigma_1 = I_n$, and $m = 100$. For three kernels described in Section 5.1, we generate 100 datasets under the global null and randomly drop 50% points for each record as the missing values. We set $q = 3$ and use the same basis (22) to compute the p-value by the proposed selective inference framework. Figure 6 shows the results for these three kernels, where the left column presents the first 5 records and the right column presents the QQ-plot of the selective p-value. We find that Figures 6(b), 6(d) are close to the uniform QQ-line and Figure 6(f) is slightly deviate. This implies that the selective p-value is robust to missing values in general.

Robustness to misspecification. We consider the specification cases and compute the selective p-value under a global null. In more detail, we consider three misspecification cases: Brown motion (Figure 7(a)), uniform random walk (Figure 7(c)), and Poisson process (Figure 7(e)). We present the QQ-plot of the selective p-value in Appendix B.

6 Phenotyping of Acute Kidney Injury (AKI) based on EHR

Now we present a real-data application of our selective inference framework. Acute Kidney Injury (AKI) is a common clinical syndrome, which is notably complex in its treatment process and frequently leads to high mortality rates and adverse outcomes. The pathology of AKI is characterized by a high degree of heterogeneity, posing significant challenges to the formulation of treatment plans. Consequently, the identification of new AKI subtypes is crucial. The severity of disease in AKI patients tends to vary over time, making the problem of hypothesis testing for functional disease subtypes of significant practical importance.

In this section, we used the MIMIC-IV EHR dataset from PhysioNet Johnson et al. (2020, 2023); Goldberger et al. (2000), which encompasses deidentified medical data, encompasses information on in excess of 40,000 patients who were admitted to the Intensive Care Units (ICU) at Beth Israel Deaconess Medical Center from 2008 to 2019. The database provides information from various angles, including vital signs, medications, laboratory measurements, diagnostic codes, and hospital length of stay. Spanning over a decade, this dataset is rich in individual patient-level information and is freely accessible, making it feasible for clinical research worldwide. Examples include identifying disease subtypes, predicting patient outcomes, and exploring effective therapeutic measures.

To avoid systemic bias, we only used data from patients with AKI admitted to the ICU. Initially, we preprocessed the data based on the framework provided by Song et al. (2020), excluding patients with: 1) End Stage Renal Disease, 2) Burns, 3) Renal Dialysis. Subsequently, according to the clinical practice guidelines for Acute Kidney Injury designated by Kidney Disease Improving Global Outcomes (KDIGO), we defined three subtypes of AKI as follows Khwaja (2012):

- Stage-1 AKI: Serum Creatinine (SCr) value rises to 1.5-1.9 times the baseline value within 7 days.
- Stage-2 AKI: SCr value rises to 2.0-2.9 times the baseline value within 7 days.
- Stage-3 AKI: SCr value rises to 3 times the baseline value or more within 7 days or the maximum SCr value over 2 days is greater than 4.0mg/dl.

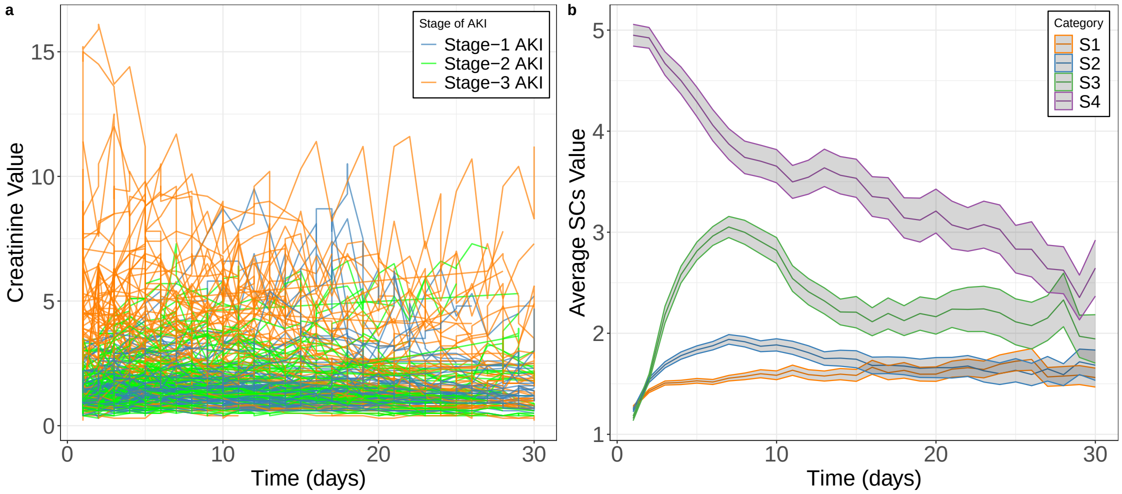


Figure 3: a. Real trajectories for 100 randomly selected patients from each category. b. Trajectories (Mean ± 1.96× standard deviation/$\sqrt{\text{sample size}}$) of the four AKI subtypes.

Here we define the SCr baseline value as the earliest recorded value for the patient. Considering the heterogeneity in the definition of ”Stage-3 AKI”, we separated the first criterion (using growth rate as an indicator) and the second criterion (using absolute value as an indicator) since their longitudinal data shapes are likely different. For a clear presentation of the analysis results, we named Stage-1 AKI as ”S1”, Stage-2 AKI as ”S2”, the first criterion of Stage-3 AKI as ”S3”, and the second criterion as ”S4”.The specific shape of this longitudinal data is shown in Figure 3. We then used hierarchical clustering based on squared Euclidean distance to cluster each category combination, specifying the number of clusters as 2. In this clustering scenario, we compared the p-values under two distinct test methods. The first is the test (2), *i.e.*, $H_0^{\{\mathcal{C}_1,\mathcal{C}_2\}} : \bar{\mu}_{\mathcal{C}_1} = \bar{\mu}_{\mathcal{C}_2}$. The second method we considered is the Wald test method.

	“No clusters”				“Clusters”		
Cluster pairs	$S1$	$S2$	$S3$	$S4$	$(S1, S4)$	$(S2, S4)$	$(S3, S4)$
Our p-value	0.19253	0.18459	0.89726	0.53099	0.00868	0.16164	0.54320
Wald p-value	$< 10^{-307}$	$< 10^{-307}$	$< 10^{-307}$	$< 10^{-307}$	$< 10^{-307}$	$< 10^{-307}$	$< 10^{-307}$

Table 1: Comparison of p-values under different clustering scenarios.

As evident from Table 1, in the ”No clusters” scenario, the p-values from our test are relatively high, while those from the Wald test are notably low. This elevated p-value in our approach correctly refrains from rejecting the null hypothesis. Given that, in the ”No clusters” scenario, we individually clustered and tested the four subtypes which, in reality, all belong to the same category, it implies that our method successfully identified the inherent homogeneity among these subtypes. In contrast, the Wald test, with its low p-values, could mislead researchers into believing that these subtypes are distinct, suggesting that the Wald test may not be as reliable in this specific context.

In the ”Clusters” scenario, the p-value for the combination of S1 and S4 is notably low. This correctly identifies the heterogeneity of this combined class. Clinically, considering the AKI definitions, S1 and S4 present significant differences in both shape and mean, making the rejection of the null hypothesis appropriate.

Examining the combination of S2 and S4, we observe that its p-value is the second smallest. Although it is not sufficiently low to decisively reject the null hypothesis, this may be attributed to the inherent heterogeneity of S2 itself (as evidenced by its standalone p-value of 0.1846, which is notably smaller than the other three classes). However, it’s noteworthy that this p-value is still lower than in the ”No clusters” scenario, indicating a certain degree of heterogeneity between S2 and S4.

Lastly, for the combination of S3 and S4, the corresponding p-value is relatively high, suggesting that S3 and S4 likely represent the same subtype. This is consistent with clinical understanding, since both S3 and S4 are classified as Stage-3 AKI. Many individuals meet the criteria for both classifications simultaneously, resulting in minimal inter-individual longitudinal data variation, not significant enough to classify them as distinct subtypes.

References

Abraham, C., Cornillon, P.-A., Matzner-Løber, E. and Molinari, N. (2003). Unsupervised curve clustering using b-splines. *Scandinavian journal of statistics*, **30** 581–595.

Banfield, J. D. and Raftery, A. E. (1993). Model-based gaussian and non-gaussian clustering. *Biometrics* 803–821.

Benjamini, Y. and Bogomolov, M. (2014). Selective inference on multiple families of hypotheses. *Journal of the Royal Statistical Society: Series B: Statistical Methodology* 297–318.

Berk, R., Brown, L., Buja, A., Zhang, K. and Zhao, L. (2013). Valid post-selection inference. *The Annals of Statistics* 802–837.

Charkhi, A. and Claeskens, G. (2018). Asymptotic post-selection inference for the akaike information criterion. *Biometrika*, **105** 645–664.

Chen, A. et al. (2022). Learning longitudinal patterns and subtypes of pediatric crohn disease treated with infliximab via trajectory cluster analysis. *Journal of Pediatric Gastroenterology and Nutrition*, **74** 383–388.

Chen, X. and Liu, W. (2015). Statistical inference for matrix-variate gaussian graphical models and false discovery rate control. *arXiv preprint arXiv:1509.05453*.

Chen, Y. T. and Witten, D. M. (2022). Selective inference for k-means clustering. *arXiv preprint arXiv:2203.15267*.

Chiou, J.-M. (2012). Dynamical functional prediction and classification, with application to traffic flow prediction.

Chiou, J.-M. and Li, P.-L. (2007). Functional clustering and identifying substructures of longitudinal data. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, **69** 679–699.

Chiou, J.-M. and Li, P.-L. (2008). Correlation-based functional clustering via subspace projection. *Journal of the American Statistical Association*, **103** 1684–1692.

Coffey, N., Hinde, J. and Holian, E. (2014). Clustering longitudinal profiles using p-splines and mixed effects models applied to time-course gene expression data. *Computational Statistics & Data Analysis*, **71** 14–29.

Dawid, A. P. (1981). Some matrix-variate distribution theory: notational considerations and a bayesian application. *Biometrika*, **68** 265–274.

Ding, S. and Dennis Cook, R. (2018). Matrix variate regressions and envelope models. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, **80** 387–408.

Dutilleul, P. (1999). The mle algorithm for the matrix normal distribution. *Journal of statistical computation and simulation*, **64** 105–123.

Fasshauer, G. E. and McCourt, M. J. (2012). Stable evaluation of gaussian radial basis function interpolants. *SIAM Journal on Scientific Computing*, **34** A737–A762.

Fithian, W., Sun, D. and Taylor, J. (2014). Optimal inference after model selection. *arXiv preprint arXiv:1410.2597*.

Gao, L. L., Bien, J. and Witten, D. (2022). Selective inference for hierarchical clustering. *Journal of the American Statistical Association* 1–11.

Giacofci, M., Lambert-Lacroix, S., Marot, G. and Picard, F. (2013). Wavelet-based clustering for mixed-effects functional models in high dimension. *Biometrics*, **69** 31–40.

Goldberger, A. L., Amaral, L. A., Glass, L., Hausdorff, J. M., Ivanov, P. C., Mark, R. G., Mietus, J. E., Moody, G. B., Peng, C.-K. and Stanley, H. E. (2000). Physiobank, physiotoolkit, and physionet: components of a new research resource for complex physiologic signals. *circulation*, **101** e215–e220.

Hall, P. and Van Keilegom, I. (2007). Two-sample tests in functional data analysis starting from discrete data. *Statistica Sinica* 1511–1531.

Heinzl, F. and Tutz, G. (2014). Clustering in linear-mixed models with a group fused lasso penalty. *Biometrical Journal*, **56** 44–68.

Hivert, B., Agniel, D., Thiébaut, R. and Hejblum, B. P. (2022). Post-clustering difference testing: valid inference and practical considerations. *arXiv preprint arXiv:2210.13172*.

Hoff, P., McCormack, A. and Zhang, A. R. (2022). Core shrinkage covariance estimation for matrix-variate data. *arXiv preprint arXiv:2207.12484*.

Hoff, P. D. (2015). Multilinear tensor regression for longitudinal relational data. *The annals of applied statistics*, **9** 1169.

Horváth, L. and Kokoszka, P. (2012). *Inference for functional data with applications*, vol. 200. Springer Science & Business Media.

Hyun, S., G'sell, M. and Tibshirani, R. J. (2018). Exact post-selection inference for the generalized lasso path.

Hyun, S., Lin, K. Z., G'Sell, M. and Tibshirani, R. J. (2021). Post-selection inference for changepoint detection algorithms with application to copy number variation data. *Biometrics*, **77** 1037–1049.

Ioannidis, J. P. (2005). Why most published research findings are false. *PLoS medicine*, **2** e124.

Jacques, J. and Preda, C. (2014). Model-based clustering for multivariate functional data. *Computational Statistics & Data Analysis*, **71** 92–106.

James, G. M. and Sugar, C. A. (2003). Clustering for sparsely sampled functional data. *Journal of the American Statistical Association*, **98** 397–408.

Jewell, S., Fearnhead, P. and Witten, D. (2022). Testing for a change in mean after changepoint detection. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, **84** 1082–1104.

Johnson, A., Bulgarelli, L., Pollard, T., Horng, S., Celi, L. A. and Mark, R. (2020). Mimic-iv. *PhysioNet. Available online at: https://physionet. org/content/mimiciv/1.0/(accessed August 23, 2021).*

Johnson, A. E., Bulgarelli, L., Shen, L., Gayles, A., Shammout, A., Horng, S., Pollard, T. J., Hao, S., Moody, B., Gow, B. et al. (2023). Mimic-iv, a freely accessible electronic health record dataset. *Scientific data*, **10** 1.

Kayano, M., Dozono, K. and Konishi, S. (2010). Functional cluster analysis via orthonormalized gaussian basis expansions and its application. *Journal of classification*, **27** 211–230.

Khwaja, A. (2012). Kdigo clinical practice guidelines for acute kidney injury. *Nephron Clinical Practice*, **120** c179–c184.

Lee, J. D., Sun, D. L., Sun, Y. and Taylor, J. E. (2016). Exact post-selection inference, with application to the lasso.

Loftus, J. R. and Taylor, J. E. (2015). Selective inference in regression models with groups of variables. *arXiv preprint arXiv:1511.01478.*

Lou, J. et al. (2021). Learning latent heterogeneity for type 2 diabetes patients using longitudinal health markers in electronic health records. *Statistics in Medicine*, **40** 1930–1946.

MacLeod, A. (2009). Ncepod report on acute kidney injury—must do better. *The Lancet*, **374** 1405–1406.

Manzini, E. et al. (2022). Longitudinal deep learning clustering of type 2 diabetes mellitus trajectories using routinely collected health records. *Journal of Biomedical Informatics*, **135** 104218.

Peng, J. and Müller, H.-G. (2008). Distance-based clustering of sparsely observed stochastic processes, with applications to online auctions.

Qiu, Z., Chen, J. and Zhang, J.-T. (2021). Two-sample tests for multivariate functional data with applications. *Computational Statistics & Data Analysis*, **157** 107160.

Ramaswamy, R. et al. (2021). Ckd subpopulations defined by risk-factors: A longitudinal analysis of electronic health records. *CPT: Pharmacometrics & Systems Pharmacology*, **10** 1343–1356.

Serban, N. and Wasserman, L. (2005). Cats: clustering after transformation and smoothing. *Journal of the American Statistical Association*, **100** 990–999.

Song, X., Yu, A. S., Kellum, J. A., Waitman, L. R., Matheny, M. E., Simpson, S. Q., Hu, Y. and Liu, M. (2020). Cross-site transportability of an explainable artificial intelligence model for acute kidney injury prediction. *Nature communications*, **11** 5668.

Taylor, J. and Tibshirani, R. (2018). Post-selection inference for-penalized likelihood models. *Canadian Journal of Statistics*, **46** 41–61.

Taylor, J. and Tibshirani, R. J. (2015). Statistical learning and selective inference. *Proceedings of the National Academy of Sciences*, **112** 7629–7634.

Tibshirani, R. J., Taylor, J., Lockhart, R. and Tibshirani, R. (2016). Exact post-selection inference for sequential regression procedures. *Journal of the American Statistical Association*, **111** 600–620.

Tsiligkaridis, T. and Hero, A. O. (2013). Covariance estimation in high dimensions via kronecker product expansions. *IEEE Transactions on Signal Processing*, **61** 5347–5360.

Wang, H. E., Muntner, P., Chertow, G. M. and Warnock, D. G. (2012). Acute kidney injury and mortality in hospitalized patients. *American Journal of Nephrology*, **35** 349–355.

Yang, F., Foygel Barber, R., Jain, P. and Lafferty, J. (2016). Selective inference for group-sparse linear models. *Advances in neural information processing systems*, **29**.

Yin, J. and Li, H. (2012). Model selection and estimation in the matrix normal graphical model. *Journal of multivariate analysis*, **107** 119–140.

Yun, Y. and Barber, R. F. (2023). Selective inference for clustering with unknown variance. *arXiv preprint arXiv:2301.12999*.

Zeldow, B. et al. (2021). Functional clustering methods for longitudinal data with application to electronic health records. *Statistical Methods in Medical Research*, **30** 655–670.

Zhang, J. M., Kamath, G. M. and David, N. T. (2019). Valid post-clustering differential analysis for single-cell rna-seq. *Cell systems*, **9** 383–392.

Zhang, J.-T. and Chen, J. (2007). Statistical inferences for functional data.

Zhou, S. (2014). Gemini: Graph estimation with matrix variate normal instances. *The Annals of Statistics*, **42** 532–562.

A Additional Proofs

A.1 Proof of Theorem 1

To begin with, for a realization w of the model (1) and the corresponding partition $\mathcal{C}_1, \mathcal{C}_2$ obtain by a clustering algorithm on $L(w)$. For any $\alpha \in [0,1]$, we consider the following conditional probability:

$$\mathbb{P}_{H_0^{\{\mathcal{C}_1,\mathcal{C}_2\}}}\left(p(\mathcal{W};\mathcal{C}_1,\mathcal{C}_2) \leq \alpha \middle| \mathcal{C}_1,\mathcal{C}_2 \in \mathcal{C}(L(\mathcal{W})), \pi^{\perp}_{\nu(\mathcal{C}_1,\mathcal{C}_2)} \times_1 L(\mathcal{W}) = \pi^{\perp}_{\nu(\mathcal{C}_1,\mathcal{C}_2)} \times_1 L(w), \right. \\ \left. \operatorname{dir}(\overline{L(\mathcal{W})}_{\mathcal{C}_1} - \overline{L(\mathcal{W})}_{\mathcal{C}_2}) = \operatorname{dir}(\overline{L(w)}_{\mathcal{C}_1} - \overline{L(w)}_{\mathcal{C}_2})\right) \tag{23}$$

Recalling the definition of the selective p-value (16) and its equivalent form (17). Given the partition $\mathcal{C}_1, \mathcal{C}_2$ and any realization $\mathcal{W}$ of the model (1) satisfies $\mathcal{C}_1, \mathcal{C}_2 \in \mathcal{C}(L(\mathcal{W}))$, we have

$$p(\mathcal{W};\mathcal{C}_1,\mathcal{C}_2) = 1 - \mathbb{F}\left(\|\overline{L(\mathcal{W})}_{\mathcal{C}_1} - \overline{L(\mathcal{W})}_{\mathcal{C}_2}\|_F; \sqrt{\frac{1}{|\mathcal{C}_1|} + \frac{1}{|\mathcal{C}_2|}}, \mathcal{S}(\mathcal{W};\mathcal{C}_1,\mathcal{C}_2)\right).$$

Given this equation, we rewrite (23) as follows:

$$\mathbb{P}_{H_0^{\{\mathcal{C}_1,\mathcal{C}_2\}}}\left(1 - \mathbb{F}\left(\|\overline{L(\mathcal{W})}_{\mathcal{C}_1} - \overline{L(\mathcal{W})}_{\mathcal{C}_2}\|_F; \sqrt{\frac{1}{|\mathcal{C}_1|} + \frac{1}{|\mathcal{C}_2|}}, \mathcal{S}(\mathcal{W};\mathcal{C}_1,\mathcal{C}_2)\right) \leq \alpha \middle| \mathcal{C}_1,\mathcal{C}_2 \in \mathcal{C}(L(\mathcal{W})), \right. \\ \left. \pi^{\perp}_{\nu(\mathcal{C}_1,\mathcal{C}_2)} \times_1 L(\mathcal{W}) = \pi^{\perp}_{\nu(\mathcal{C}_1,\mathcal{C}_2)} \times_1 L(w), \operatorname{dir}(\overline{L(\mathcal{W})}_{\mathcal{C}_1} - \overline{L(\mathcal{W})}_{\mathcal{C}_2}) = \operatorname{dir}(\overline{L(w)}_{\mathcal{C}_1} - \overline{L(w)}_{\mathcal{C}_2})\right) \tag{24}$$

Given the conditions $\pi^{\perp}_{\nu(\mathcal{C}_1,\mathcal{C}_2)} \times_1 L(\mathcal{W}) = \pi^{\perp}_{\nu(\mathcal{C}_1,\mathcal{C}_2)} \times_1 L(w), \operatorname{dir}(\overline{L(\mathcal{W})}_{\mathcal{C}_1} - \overline{L(\mathcal{W})}_{\mathcal{C}_2}) = \operatorname{dir}(\overline{L(w)}_{\mathcal{C}_1} - \overline{L(w)}_{\mathcal{C}_2})$, the two sets $\mathcal{S}(\mathcal{W};\mathcal{C}_1,\mathcal{C}_2)$ and $\mathcal{S}(w;\mathcal{C}_1,\mathcal{C}_2)$ are equivalent:

$$\begin{aligned}\mathcal{S}(\mathcal{W};\mathcal{C}_1,\mathcal{C}_2) &= \left\{\varphi \geq 0 : \mathcal{C}_1,\mathcal{C}_2 \in \mathcal{C}\left(\pi^{\perp}_{\nu(\mathcal{C}_1,\mathcal{C}_2)} \times_1 L(\mathcal{W}) + \left[\frac{\varphi}{\frac{1}{|\mathcal{C}_1|} + \frac{1}{|\mathcal{C}_2|}}\right] \nu(\mathcal{C}_1,\mathcal{C}_2) \otimes \operatorname{dir}(\overline{L(\mathcal{W})}_{\mathcal{C}_1} - \overline{L(\mathcal{W})}_{\mathcal{C}_2})^{\top}\right)\right\} \\ &= \left\{\varphi \geq 0 : \mathcal{C}_1,\mathcal{C}_2 \in \mathcal{C}\left(\pi^{\perp}_{\nu(\mathcal{C}_1,\mathcal{C}_2)} \times_1 L(w) + \left[\frac{\varphi}{\frac{1}{|\mathcal{C}_1|} + \frac{1}{|\mathcal{C}_2|}}\right] \nu(\mathcal{C}_1,\mathcal{C}_2) \otimes \operatorname{dir}(\overline{L(w)}_{\mathcal{C}_1} - \overline{L(w)}_{\mathcal{C}_2})^{\top}\right)\right\} \\ &= \mathcal{S}(w;\mathcal{C}_1,\mathcal{C}_2).\end{aligned}$$

Moreover, the random variable $\|\overline{L(\mathcal{W})}_{\mathcal{C}_1} - \overline{L(\mathcal{W})}_{\mathcal{C}_2}\|_F$ is independent of $\pi^{\perp}_{\nu(\mathcal{C}_1,\mathcal{C}_2)} \times_1 L(\mathcal{W})$ and $\mathrm{dir}(\overline{L(w)}_{\mathcal{C}_1} - \overline{L(w)}_{\mathcal{C}_2})$. Therefore, the conditional probability (24) is equal to

$$\begin{aligned}
&\mathbb{P}_{H_0^{\{\mathcal{C}_1,\mathcal{C}_2\}}}\left(1-\mathbb{F}\left(\|\overline{L(\mathcal{W})}_{\mathcal{C}_1} - \overline{L(\mathcal{W})}_{\mathcal{C}_2}\|_F; \sqrt{\frac{1}{|\mathcal{C}_1|}+\frac{1}{|\mathcal{C}_2|}}, \mathcal{S}(w;\mathcal{C}_1,\mathcal{C}_2)\right) \leq \alpha \middle| \mathcal{C}_1,\mathcal{C}_2 \in \mathcal{C}\left(\pi^{\perp}_{\nu(\mathcal{C}_1,\mathcal{C}_2)} \times_1 L(w) + \right.\right.\\
&\left.\left[\frac{\|\overline{L(\mathcal{W})}_{\mathcal{C}_1} - \overline{L(\mathcal{W})}_{\mathcal{C}_2}\|_F}{1/|\mathcal{C}_1|+1/|\mathcal{C}_2|}\right]\nu(\mathcal{C}_1,\mathcal{C}_2)\otimes \mathrm{dir}(\overline{L(w)}_{\mathcal{C}_1} - \overline{L(w)}_{\mathcal{C}_2})^{\top}\right), \pi^{\perp}_{\nu(\mathcal{C}_1,\mathcal{C}_2)} \times_1 L(\mathcal{W}) = \pi^{\perp}_{\nu(\mathcal{C}_1,\mathcal{C}_2)} \times_1 L(w),\\
&\left.\mathrm{dir}(\overline{L(\mathcal{W})}_{\mathcal{C}_1} - \overline{L(\mathcal{W})}_{\mathcal{C}_2}) = \mathrm{dir}(\overline{L(w)}_{\mathcal{C}_1} - \overline{L(w)}_{\mathcal{C}_2})\right)\\
=&\mathbb{P}_{H_0^{\{\mathcal{C}_1,\mathcal{C}_2\}}}\left(1-\mathbb{F}\left(\|\overline{L(\mathcal{W})}_{\mathcal{C}_1} - \overline{L(\mathcal{W})}_{\mathcal{C}_2}\|_F; \sqrt{\frac{1}{|\mathcal{C}_1|}+\frac{1}{|\mathcal{C}_2|}}, \mathcal{S}(w;\mathcal{C}_1,\mathcal{C}_2)\right) \leq \alpha \middle| \mathcal{C}_1,\mathcal{C}_2 \in \mathcal{C}\left(\pi^{\perp}_{\nu(\mathcal{C}_1,\mathcal{C}_2)} \times_1 L(w) + \right.\right.\\
&\left.\left.\left[\frac{\|\overline{L(\mathcal{W})}_{\mathcal{C}_1} - \overline{L(\mathcal{W})}_{\mathcal{C}_2}\|_F}{1/|\mathcal{C}_1|+1/|\mathcal{C}_2|}\right]\nu(\mathcal{C}_1,\mathcal{C}_2)\otimes \mathrm{dir}(\overline{L(w)}_{\mathcal{C}_1} - \overline{L(w)}_{\mathcal{C}_2})^{\top}\right)\right)\\
=&\mathbb{P}_{H_0^{\{\mathcal{C}_1,\mathcal{C}_2\}}}\left(1-\mathbb{F}\left(\|\overline{L(\mathcal{W})}_{\mathcal{C}_1} - \overline{L(\mathcal{W})}_{\mathcal{C}_2}\|_F; \sqrt{\frac{1}{|\mathcal{C}_1|}+\frac{1}{|\mathcal{C}_2|}}, \mathcal{S}(w;\mathcal{C}_1,\mathcal{C}_2)\right) \leq \alpha \middle| \|\overline{L(\mathcal{W})}_{\mathcal{C}_1} - \overline{L(\mathcal{W})}_{\mathcal{C}_2}\|_F \in \mathcal{S}(w;\mathcal{C}_1,\mathcal{C}_2)\right)
\end{aligned} \tag{25}$$

Therefore, (25) indicates that the conditional probability (23) is the survival function of the truncated $\sqrt{1/|\mathcal{C}_1|+1/|\mathcal{C}_2|}\cdot\chi_{nq}$ distribution. Namely, we have

$$\begin{aligned}
\mathbb{P}_{H_0^{\{\mathcal{C}_1,\mathcal{C}_2\}}}\Big(p(\mathcal{W};\mathcal{C}_1,\mathcal{C}_2) \leq \alpha \Big| \mathcal{C}_1,\mathcal{C}_2 \in \mathcal{C}(L(\mathcal{W})), \pi^{\perp}_{\nu(\mathcal{C}_1,\mathcal{C}_2)} \times_1 L(\mathcal{W}) = \pi^{\perp}_{\nu(\mathcal{C}_1,\mathcal{C}_2)} \times_1 L(w),&\\
\mathrm{dir}(\overline{L(\mathcal{W})}_{\mathcal{C}_1} - \overline{L(\mathcal{W})}_{\mathcal{C}_2}) = \mathrm{dir}(\overline{L(w)}_{\mathcal{C}_1} - \overline{L(w)}_{\mathcal{C}_2})\Big) = \alpha.&
\end{aligned} \tag{26}$$

Now we use (26) to compute the selective p-value. By the law of iterated expectation, we rewrite the selective type-I error as follows:

$$\begin{aligned}
&\mathbb{P}_{H_0^{\{\mathcal{C}_1,\mathcal{C}_2\}}}\left(p(\mathcal{W};\mathcal{C}_1,\mathcal{C}_2)\leq\alpha \middle| \mathcal{C}_1,\mathcal{C}_2 \in \mathcal{C}(L(\mathcal{W}))\right) = \mathbb{E}_{H_0^{\{\mathcal{C}_1,\mathcal{C}_2\}}}\left(\mathbb{1}_{p(\mathcal{W};\mathcal{C}_1,\mathcal{C}_2)\leq\alpha} \middle| \mathcal{C}_1,\mathcal{C}_2 \in \mathcal{C}(L(\mathcal{W}))\right)\\
&= \mathbb{E}_{H_0^{\{\mathcal{C}_1,\mathcal{C}_2\}}}\Big(\mathbb{E}\big[\mathbb{1}_{p(\mathcal{W};\mathcal{C}_1,\mathcal{C}_2)\leq\alpha} \big| \mathcal{C}_1,\mathcal{C}_2 \in \mathcal{C}(L(\mathcal{W})), \pi^{\perp}_{\nu(\mathcal{C}_1,\mathcal{C}_2)} \times_1 L(\mathcal{W}) = \pi^{\perp}_{\nu(\mathcal{C}_1,\mathcal{C}_2)} \times_1 L(w),\\
&\qquad \mathrm{dir}(\overline{L(\mathcal{W})}_{\mathcal{C}_1} - \overline{L(\mathcal{W})}_{\mathcal{C}_2}) = \mathrm{dir}(\overline{L(w)}_{\mathcal{C}_1} - \overline{L(w)}_{\mathcal{C}_2})\big] \Big| \mathcal{C}_1,\mathcal{C}_2 \in \mathcal{C}(L(\mathcal{W}))\Big).
\end{aligned}$$

Plugging in (26), we obtain

$$\begin{aligned}
&\mathbb{P}_{H_0^{\{\mathcal{C}_1,\mathcal{C}_2\}}}\left(p(\mathcal{W};\mathcal{C}_1,\mathcal{C}_2)\leq\alpha \middle| \mathcal{C}_1,\mathcal{C}_2 \in \mathcal{C}(L(\mathcal{W}))\right)\\
&= \mathbb{E}_{H_0^{\{\mathcal{C}_1,\mathcal{C}_2\}}}\left(\alpha \middle| \mathcal{C}_1,\mathcal{C}_2 \in \mathcal{C}(L(\mathcal{W}))\right) = \alpha.
\end{aligned}$$

A.2 Proof of Theorem 2

By the Bayes rule, we rewrite the conditional probability $\mathbb{P}\left(p(\mathcal{W};\mathcal{C}_1^{\mathcal{W}},\mathcal{C}_2^{\mathcal{W}})\leq\alpha\middle|\,\bar{\mu}_{\mathcal{C}_1^{\mathcal{W}}}=\bar{\mu}_{\mathcal{C}_2^{\mathcal{W}}}\right)$ as follows:

$$\mathbb{P}\left(p(\mathcal{W};\mathcal{C}_1^{\mathcal{W}},\mathcal{C}_2^{\mathcal{W}})\leq\alpha\middle|\,\bar{\mu}_{\mathcal{C}_1^{\mathcal{W}}}=\bar{\mu}_{\mathcal{C}_2^{\mathcal{W}}}\right)=\mathbb{P}\left(p(\mathcal{W};\mathcal{C}_1^{\mathcal{W}},\mathcal{C}_2^{\mathcal{W}})\leq\alpha,\bar{\mu}_{\mathcal{C}_1^{\mathcal{W}}}=\bar{\mu}_{\mathcal{C}_2^{\mathcal{W}}}\right)/\mathbb{P}(\bar{\mu}_{\mathcal{C}_1^{\mathcal{W}}}=\bar{\mu}_{\mathcal{C}_2^{\mathcal{W}}}).\tag{27}$$

We notice that $\bar{\mu}_{\mathcal{C}_1^{\mathcal{W}}}=\bar{\mu}_{\mathcal{C}_2^{\mathcal{W}}}$ can be rewritten as follows:

$$\mathcal{C}(L(\mathcal{W}))=\{\mathcal{C}_1,\mathcal{C}_2\},\qquad\bar{\mu}_{\mathcal{C}_1}=\bar{\mu}_{\mathcal{C}_2}.$$

Therefore, we decompose $\mathbb{P}\left(p(\mathcal{W};\mathcal{C}_1^{\mathcal{W}},\mathcal{C}_2^{\mathcal{W}})\leq\alpha,\bar{\mu}_{\mathcal{C}_1^{\mathcal{W}}}=\bar{\mu}_{\mathcal{C}_2^{\mathcal{W}}}\right)$ into the sum of non-intersect partition $\mathcal{C}_1,\mathcal{C}_2$. Namely, the probability can be rewritten as follows:

$$\mathbb{P}\left(p(\mathcal{W};\mathcal{C}_1^{\mathcal{W}},\mathcal{C}_2^{\mathcal{W}})\leq\alpha,\bar{\mu}_{\mathcal{C}_1^{\mathcal{W}}}=\bar{\mu}_{\mathcal{C}_2^{\mathcal{W}}}\right)=\sum_{\substack{\mathcal{C}_1\cup\mathcal{C}_2=[m],\\ \bar{\mu}_{\mathcal{C}_1}=\bar{\mu}_{\mathcal{C}_2}}}\mathbb{P}\left(p(\mathcal{W};\mathcal{C}_1,\mathcal{C}_2)\leq\alpha,\mathcal{C}(L(\mathcal{W}))=\{\mathcal{C}_1,\mathcal{C}_2\}\right).$$

By the Bayes rule, the above equality implies that

$$\begin{aligned}&\mathbb{P}\left(p(\mathcal{W};\mathcal{C}_1^{\mathcal{W}},\mathcal{C}_2^{\mathcal{W}})\leq\alpha,\bar{\mu}_{\mathcal{C}_1^{\mathcal{W}}}=\bar{\mu}_{\mathcal{C}_2^{\mathcal{W}}}\right)\\&\qquad=\sum_{\substack{\mathcal{C}_1\cup\mathcal{C}_2=[m],\\ \bar{\mu}_{\mathcal{C}_1}=\bar{\mu}_{\mathcal{C}_2}}}\mathbb{P}\left(p(\mathcal{W};\mathcal{C}_1,\mathcal{C}_2)\leq\alpha\middle|\mathcal{C}(L(\mathcal{W}))=\{\mathcal{C}_1,\mathcal{C}_2\}\right)\cdot\mathbb{P}(\mathcal{C}(L(\mathcal{W}))=\{\mathcal{C}_1,\mathcal{C}_2\}).\end{aligned}$$

Theorem 1 implies that $\mathbb{P}\left(p(\mathcal{W};\mathcal{C}_1,\mathcal{C}_2)\leq\alpha\middle|\mathcal{C}(L(\mathcal{W}))=\{\mathcal{C}_1,\mathcal{C}_2\}\right)=\alpha$ for any partition $\mathcal{C}_1,\mathcal{C}_2$. Thus, we plug in this result and obtain that

$$\begin{aligned}\mathbb{P}\left(p(\mathcal{W};\mathcal{C}_1^{\mathcal{W}},\mathcal{C}_2^{\mathcal{W}})\leq\alpha,\bar{\mu}_{\mathcal{C}_1^{\mathcal{W}}}=\bar{\mu}_{\mathcal{C}_2^{\mathcal{W}}}\right)&=\alpha\cdot\sum_{\substack{\mathcal{C}_1\cup\mathcal{C}_2=[m],\\ \bar{\mu}_{\mathcal{C}_1}=\bar{\mu}_{\mathcal{C}_2}}}\mathbb{P}(\mathcal{C}(L(\mathcal{W}))=\{\mathcal{C}_1,\mathcal{C}_2\})\\&=\alpha\cdot\mathbb{P}(\bar{\mu}_{\mathcal{C}_1^{\mathcal{W}}}=\bar{\mu}_{\mathcal{C}_2^{\mathcal{W}}}).\end{aligned}$$

Combining this equation with (27) directly yields (11) and complete the proof.

A.3 Proof of Theorem 3

To begin with, recall (17), we rewrite the survival function $\mathbb{F}(\cdot)$ as follows:

$$\mathbb{F}\left(\|\overline{L(w)}_{\mathcal{C}_1}-\overline{L(w)}_{\mathcal{C}_2}\|_F;\sqrt{\frac{1}{|\mathcal{C}_1|}+\frac{1}{|\mathcal{C}_2|}},\mathcal{S}(w;\mathcal{C}_1,\mathcal{C}_2)\right)=\mathbb{P}\left(\varphi\geq\|\overline{L(w)}_{\mathcal{C}_1}-\overline{L(w)}_{\mathcal{C}_2}\|_F\middle|\varphi\in\mathcal{S}(w;\mathcal{C}_1,\mathcal{C}_2)\right),\tag{28}$$

where φ follows the distribution $\sqrt{1/|\mathcal{C}_1|+1/|\mathcal{C}_2|}\cdot\chi_{nq}$.

Next, we study the asymptotic behaviour of $\|\overline{L(\mathcal{W})}_{\mathcal{C}_1}-\overline{L(\mathcal{W})}_{\mathcal{C}_2}\|_F$ under the alternative hypothesis when $\|\bar{\mu}_{\mathcal{C}_1}-\bar{\mu}_{\mathcal{C}_2}\|_F\to\infty$ and $m\to\infty$. Recall the low-dimensional embedding (6) and Lemma 1, we obtain that

$$\|\overline{H(\mu)}_{\mathcal{C}_1}-\overline{H(\mu)}_{\mathcal{C}_2}\|_F=\|(\bar{\mu}_{\mathcal{C}_1}-\bar{\mu}_{\mathcal{C}_2})\Phi^\top(K+\lambda I_q)^{-1}\|_F\to\infty,$$

where K and Φ are defined in Lemma 1. Recall the covariance estimator $\widehat{\Lambda}$ defined by (9), we have $\widehat{\Lambda} \to \Lambda + c(\overline{H(\mu)}_{\mathcal{C}_1} - \overline{H(\mu)}_{\mathcal{C}_2})(\overline{H(\mu)}_{\mathcal{C}_1} - \overline{H(\mu)}_{\mathcal{C}_2})^\top/(c+1)^2$. Therefore, the whitening transformation (11) implies that

$$\|\overline{L(\mathcal{W})}_{\mathcal{C}_1} - \overline{L(\mathcal{W})}_{\mathcal{C}_2}\|_F = \|\widehat{\Lambda}^{-1/2} \cdot \mathrm{vec}(\overline{H(\mathcal{W})}_{\mathcal{C}_1} - \overline{H(\mathcal{W})}_{\mathcal{C}_2})\|_F \to (c+1)/\sqrt{c}.$$

Therefore, a proper clustering algorithm would output the labels corresponding to the partition $\mathcal{C}_1$ and $\mathcal{C}_2$, namely, $\mathbb{P}(\mathcal{C}_1, \mathcal{C}_2 \in \mathcal{C}(L(\mathcal{W}))) \to 1$. Moreover, recall the selective p-value (17), which is the survival function of the distribution $\sqrt{1/|\mathcal{C}_1| + 1/|\mathcal{C}_2|} \cdot \chi_{nq}$ truncated to the set $\mathcal{S}(w; \mathcal{C}_1, \mathcal{C}_2)$. Recall that $\mathcal{S}(w; \mathcal{C}_1, \mathcal{C}_2)$ is comprised of all the $\varphi \ge 0$ that have the same clustering output $\mathcal{C}_1, \mathcal{C}_2$ on the perturbed data $F(\varphi)$ defined in (18). Since $\|\overline{L(\mathcal{W})}_{\mathcal{C}_1} - \overline{L(\mathcal{W})}_{\mathcal{C}_2}\|_F$ equals to $(c+1)/\sqrt{c}$ asymptotically, $\mathcal{S}(w; \mathcal{C}_1, \mathcal{C}_2)$ concentrates near $(c+1)/\sqrt{c}$. As a result, any $\gamma_i \in \mathcal{N}(\|\overline{L(w)}_{\mathcal{C}_1} - \overline{L(w)}_{\mathcal{C}_2}\|_F, 1/|\mathcal{C}_1| + 1/|\mathcal{C}_2|)$ generated by the importance sampling step (19) are close to $(c+1)/\sqrt{c}$, which further induces that $\mathcal{C}(F(\gamma_i)) = \mathcal{C}_1, \mathcal{C}_2$. The above asymptotic analysis shows that

$$\mathbb{P}\left(\varphi \ge \|\overline{L(w)}_{\mathcal{C}_1} - \overline{L(w)}_{\mathcal{C}_2}\|_F \middle| \varphi \in \mathcal{S}(w; \mathcal{C}_1, \mathcal{C}_2)\right) \to \mathbb{P}\left(\varphi \ge (c+1)/\sqrt{c}\right) = 0,$$

the last inequality holds because $\varphi \sim \sqrt{1/|\mathcal{C}_1| + 1/|\mathcal{C}_2|} \cdot \chi_{nq}$ and $1/|\mathcal{C}_1| + 1/|\mathcal{C}_2| \to 0$. Using the above equality, we finally obtain that

$$\begin{aligned}
&\lim_{m\to\infty} \lim_{\|\bar{\mu}_{\mathcal{C}_1} - \bar{\mu}_{\mathcal{C}_2}\|_F \to \infty} \mathbb{P}_{H_1^{\{\mathcal{C}_1, \mathcal{C}_2\}}}\left(p(\mathcal{W}; \mathcal{C}_1, \mathcal{C}_2) \le \alpha \middle| \, \mathcal{C}_1, \mathcal{C}_2 \in \mathcal{C}(L(\mathcal{W}))\right) \\
&\quad = 1 - \lim_{m\to\infty} \lim_{\|\bar{\mu}_{\mathcal{C}_1} - \bar{\mu}_{\mathcal{C}_2}\|_F \to \infty} \mathbb{F}\left(\|\overline{L(w)}_{\mathcal{C}_1} - \overline{L(w)}_{\mathcal{C}_2}\|_F; \sqrt{\frac{1}{|\mathcal{C}_1|} + \frac{1}{|\mathcal{C}_2|}}, \mathcal{S}(w; \mathcal{C}_1, \mathcal{C}_2)\right) = 1.
\end{aligned}$$

A.4 Proof of Lemma 1

Recalling Assumption 1, where $\mathrm{vec}(W_i) \sim \mathcal{N}(\mathrm{vec}(\mu_i), \Sigma_2 \otimes \Sigma_1 + I_T \otimes \mathrm{diag}(\sigma_j^2)_{j\in[n]})$. Given (6), we have

$$H(\mathcal{W})[i,:,:] = W_i \Phi^\top (K + \lambda I_q)^{-1}.$$

Therefore, by the property of the vectorization operator that $\mathrm{vec}(AXB) = (B^\top \otimes A)\mathrm{vec}(X)$, we have

$$\mathrm{vec}(H(\mathcal{W})[i,:,:]) = \mathrm{vec}(W_i \Phi^\top (K + \lambda I_q)^{-1}) = (((K + \lambda I_q)^{-1}\Phi) \otimes I_n)\mathrm{vec}(W_i),$$

This further implies that $\mathrm{vec}(H(\mathcal{W})[i,:,:])$ follows the multivariate normal distribution

$$\begin{aligned}
&\mathcal{N}((((K + \lambda I_q)^{-1}\Phi) \otimes I_n)\mathrm{vec}(\mu_i), (((K + \lambda I_q)^{-1}\Phi) \otimes I_n)(\Sigma_2 \otimes \Sigma_1)(((K + \lambda I_q)^{-1}\Phi) \otimes I_n)^\top) \\
&\ + \mathcal{N}(0, (((K + \lambda I_q)^{-1}\Phi) \otimes I_n)(I_T \otimes \mathrm{diag}(\sigma_j^2)_{j\in[n]})(((K + \lambda I_q)^{-1}\Phi) \otimes I_n)^\top) \\
&\ = \mathcal{N}((((K + \lambda I_q)^{-1}\Phi) \otimes I_n)\mathrm{vec}(\mu_i), \left[(K + \lambda I_q)^{-1}\Phi\Sigma_2\Phi^\top(K + \lambda I_q)^{-1}\right] \otimes \Sigma_1) \\
&\ + \mathcal{N}(0, \left[(K + \lambda I_q)^{-1}K(K + \lambda I_q)^{-1}\right] \otimes \mathrm{diag}(\sigma_j^2)_{j\in[n]}),
\end{aligned}$$

which directly implies (7).

If $\mathcal{W}$ is a realization of the model (1) with missing values, the vectorization of W_i follows the marginal distribution of the above normal distribution, namely, it follows a normal distribution corresponds with the time record Ω_i:

$$\mathrm{vec}(W_i) \sim \mathcal{N}(\mathrm{vec}((\mu_{ij}^{\Omega_i})_{j\in[n]}), \left[\Sigma_2 \otimes \Sigma_1 + I_T \otimes \mathrm{diag}(\sigma_j^2)_{j\in[n]}\right]^{\Omega_i}),$$

where $\left[\Sigma_2 \otimes \Sigma_1 + I_T \otimes \mathrm{diag}(\sigma_j^2)_{j\in[n]}\right]^{\Omega_i} \in \mathbb{R}^{(\sum_{j=1}^n r_{ij})\times(\sum_{j=1}^n r_{ij})}$ is the submatrix of $\Sigma_2 \otimes \Sigma_1 + I_T \otimes \mathrm{diag}(\sigma_j^2)_{j\in[n]}$ that characterized by the time record Ω_i. Moreover, since $H(\mathcal{W})_{i,j,:} = W_{ij}\Phi_{ij}^\top(K_{ij} + \lambda I_q)^{-1}$, we have $H(\mathcal{W})[i,:,:] = (W_{ij}\Phi_{ij}^\top(K_{ij}+\lambda I_q)^{-1})_{j\in[n]}$. Therefore, we have

$$\mathrm{vec}(H(\mathcal{W})[i,:,:]) = \mathrm{diag}((K_{ij}+\lambda I_q)^{-1}\Phi_{ij})_{j\in[n]}\mathrm{vec}(W_i),$$

which further yields (8).

A.5 Proof of Lemma 2

To begin with, we have

$$\mathcal{A} = \pi^\perp_{\nu(\mathcal{C}_1,\mathcal{C}_2)} \times_1 \mathcal{A} + (I - \pi^\perp_{\nu(\mathcal{C}_1,\mathcal{C}_2)}) \times_1 \mathcal{A}.$$

By the definition of $\pi^\perp_{\nu(\mathcal{C}_1,\mathcal{C}_2)}$, we have $I - \pi^\perp_{\nu(\mathcal{C}_1,\mathcal{C}_2)} = \nu(\mathcal{C}_1,\mathcal{C}_2)\nu(\mathcal{C}_1,\mathcal{C}_2)^\top/\|\nu(\mathcal{C}_1,\mathcal{C}_2)\|^2$ and $\|\nu(\mathcal{C}_1,\mathcal{C}_2)\|^2 = 1/|\mathcal{C}_1| + 1/|\mathcal{C}_2|$. As a result, we can rewrite the second term in the above equation as follows:

$$\begin{aligned}(I - \pi^\perp_{\nu(\mathcal{C}_1,\mathcal{C}_2)}) \times_1 \mathcal{A} &= \frac{\nu(\mathcal{C}_1,\mathcal{C}_2)\nu(\mathcal{C}_1,\mathcal{C}_2)^\top}{1/|\mathcal{C}_1| + 1/|\mathcal{C}_2|} \times_1 \mathcal{A} \\ &= \frac{\nu(\mathcal{C}_1,\mathcal{C}_2)}{1/|\mathcal{C}_1| + 1/|\mathcal{C}_2|} \otimes (\bar{\mathcal{A}}_{\mathcal{C}_1} - \bar{\mathcal{A}}_{\mathcal{C}_2})^\top,\end{aligned} \tag{29}$$

where the last equation holds by the property of tensor mode product. The equation (29) further leads to (15) and finishes the proof.

A.6 Proof of Lemma 3

Combine the definition (16) with the orthogonal decomposition (15), we have

$$\begin{aligned}
p_{selective} = \mathbb{P}_{H_0^{\{\mathcal{C}_1,\mathcal{C}_2\}}}\Bigg(&\|\overline{L(\mathcal{W})}_{\mathcal{C}_1} - \overline{L(\mathcal{W})}_{\mathcal{C}_2}\|_F \geq \|\overline{L(w)}_{\mathcal{C}_1} - \overline{L(w)}_{\mathcal{C}_2}\|_F \Bigg| \mathcal{C}_1,\mathcal{C}_2 \in \mathcal{C}\Bigg(\pi^\perp_{\nu(\mathcal{C}_1,\mathcal{C}_2)} \times_1 L(\mathcal{W}) + \\
&\left[\frac{\|\overline{L(\mathcal{W})}_{\mathcal{C}_1} - \overline{L(\mathcal{W})}_{\mathcal{C}_2}\|_F}{1/|\mathcal{C}_1| + 1/|\mathcal{C}_2|}\right]\nu(\mathcal{C}_1,\mathcal{C}_2) \otimes \mathrm{dir}(\overline{L(\mathcal{W})}_{\mathcal{C}_1} - \overline{L(\mathcal{W})}_{\mathcal{C}_2})^\top\Bigg), \pi^\perp_{\nu(\mathcal{C}_1,\mathcal{C}_2)} \times_1 L(\mathcal{W}) = \pi^\perp_{\nu(\mathcal{C}_1,\mathcal{C}_2)} \times_1 L(w), \\
&\mathrm{dir}(\overline{L(\mathcal{W})}_{\mathcal{C}_1} - \overline{L(\mathcal{W})}_{\mathcal{C}_2}) = \mathrm{dir}(\overline{L(w)}_{\mathcal{C}_1} - \overline{L(w)}_{\mathcal{C}_2})\Bigg) \\
= \mathbb{P}_{H_0^{\{\mathcal{C}_1,\mathcal{C}_2\}}}\Bigg(&\|\overline{L(\mathcal{W})}_{\mathcal{C}_1} - \overline{L(\mathcal{W})}_{\mathcal{C}_2}\|_F \geq \|\overline{L(w)}_{\mathcal{C}_1} - \overline{L(w)}_{\mathcal{C}_2}\|_F \Bigg| \mathcal{C}_1,\mathcal{C}_2 \in \mathcal{C}\Bigg(\pi^\perp_{\nu(\mathcal{C}_1,\mathcal{C}_2)} \times_1 L(w) + \\
&\left[\frac{\|\overline{L(\mathcal{W})}_{\mathcal{C}_1} - \overline{L(\mathcal{W})}_{\mathcal{C}_2}\|_F}{1/|\mathcal{C}_1| + 1/|\mathcal{C}_2|}\right]\nu(\mathcal{C}_1,\mathcal{C}_2) \otimes \mathrm{dir}(\overline{L(w)}_{\mathcal{C}_1} - \overline{L(w)}_{\mathcal{C}_2})^\top\Bigg), \pi^\perp_{\nu(\mathcal{C}_1,\mathcal{C}_2)} \times_1 L(\mathcal{W}) = \pi^\perp_{\nu(\mathcal{C}_1,\mathcal{C}_2)} \times_1 L(w), \\
&\mathrm{dir}(\overline{L(\mathcal{W})}_{\mathcal{C}_1} - \overline{L(\mathcal{W})}_{\mathcal{C}_2}) = \mathrm{dir}(\overline{L(w)}_{\mathcal{C}_1} - \overline{L(w)}_{\mathcal{C}_2})\Bigg).
\end{aligned}$$

Next, we show that $\|\overline{L(\mathcal{W})}_{\mathcal{C}_1} - \overline{L(\mathcal{W})}_{\mathcal{C}_2}\|_F$ is independent of $\pi^\perp_{\nu(\mathcal{C}_1,\mathcal{C}_2)} \times_1 L(\mathcal{W})$ and $\mathrm{dir}(\overline{L(\mathcal{W})}_{\mathcal{C}_1} - \overline{L(\mathcal{W})}_{\mathcal{C}_2})$. To begin with, we remark that $\mathrm{vec}(\pi^\perp_{\nu(\mathcal{C}_1,\mathcal{C}_2)} \times_1 L(\mathcal{W})) = (I_q \otimes \pi^\perp_{\nu(\mathcal{C}_1,\mathcal{C}_2)})\mathrm{vec}(L(\mathcal{W}))$, where $(I_q \otimes \pi^\perp_{\nu(\mathcal{C}_1,\mathcal{C}_2)})$ is the orthogonal projection matrix that projects $\mathrm{vec}(L(\mathcal{W}))$ onto a subspace orthogonal to $I_q \otimes \nu(\mathcal{C}_1, \mathcal{C}_2)$. It follows from the properties of the multivariate normal distributions that $(I_q \otimes \pi^{\perp}_{\nu(\mathcal{C}_1,\mathcal{C}_2)})\text{vec}(L(\mathcal{W}))$ is independent to $(I_q \otimes \nu(\mathcal{C}_1, \mathcal{C}_2))\text{vec}(L(\mathcal{W}))$, which is equivalent to the statement that $\text{vec}(\pi^{\perp}_{\nu(\mathcal{C}_1,\mathcal{C}_2)^\top} \times_1 L(\mathcal{W}))$ is independent to $\text{vec}(\nu(\mathcal{C}_1, \mathcal{C}_2)^\top \times_1 L(\mathcal{W})) = \text{vec}(\overline{L(\mathcal{W})}_{\mathcal{C}_1} - \overline{L(\mathcal{W})}_{\mathcal{C}_2})$. Besides, we notice that $\text{vec}(L(\mathcal{W})[i, :, :]) \sim \mathcal{N}(\text{vec}(L(\mu_i)), I_{nq})$, which implies that $\text{vec}(\overline{L(\mathcal{W})}_{\mathcal{C}_1} - \overline{L(\mathcal{W})}_{\mathcal{C}_2})$ follows the scaled standard normal distribution $\mathcal{N}(0, (1/|\mathcal{C}_1| + 1/|\mathcal{C}_2|)I_{nq})$. As a result, the independence of the length and direction of a standard multivariate distribution implies that $\text{vec}(\overline{L(\mathcal{W})}_{\mathcal{C}_1} - \overline{L(\mathcal{W})}_{\mathcal{C}_2})$ is independent to $\text{dir}(\text{vec}(\overline{L(\mathcal{W})}_{\mathcal{C}_1} - \overline{L(\mathcal{W})}_{\mathcal{C}_2})) = \text{vec}(\text{dir}(\overline{L(\mathcal{W})}_{\mathcal{C}_1} - \overline{L(\mathcal{W})}_{\mathcal{C}_2}))$.

Thus, $\|\overline{L(\mathcal{W})}_{\mathcal{C}_1} - \overline{L(\mathcal{W})}_{\mathcal{C}_2}\|_F$ is independent of $\pi^{\perp}_{\nu(\mathcal{C}_1,\mathcal{C}_2)} \times_1 L(\mathcal{W})$ and $\text{dir}(\overline{L(\mathcal{W})}_{\mathcal{C}_1} - \overline{L(\mathcal{W})}_{\mathcal{C}_2})$, which implies that

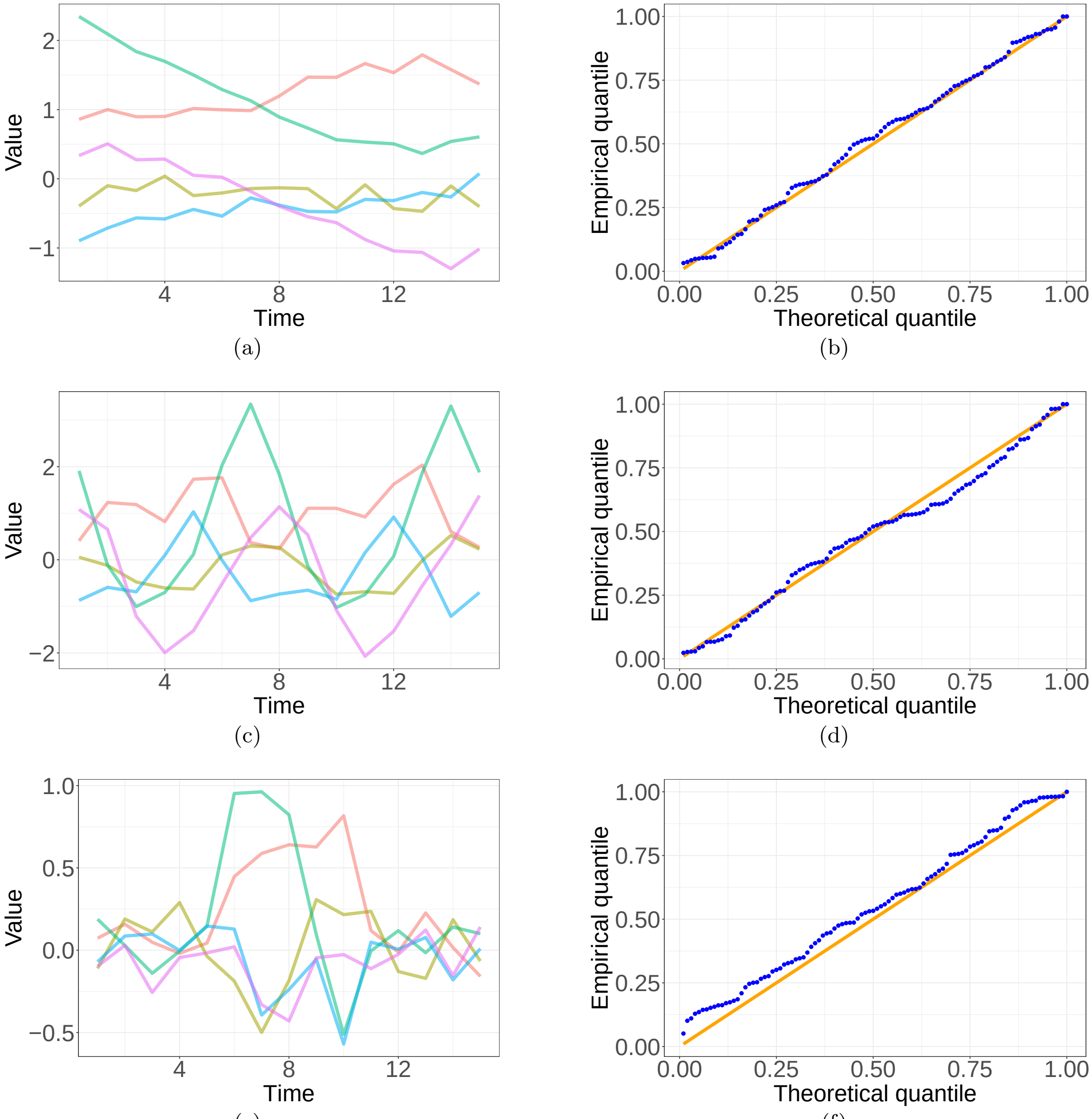
$$p_{selective} = \mathbb{P}_{H_0^{\{\mathcal{C}_1,\mathcal{C}_2\}}}\Bigg(\|\overline{L(\mathcal{W})}_{\mathcal{C}_1} - \overline{L(\mathcal{W})}_{\mathcal{C}_2}\|_F \geq \|\overline{L(w)}_{\mathcal{C}_1} - \overline{L(w)}_{\mathcal{C}_2}\|_F \Bigg| \mathcal{C}_1, \mathcal{C}_2 \in \mathcal{C}\Bigg(\pi^{\perp}_{\nu(\mathcal{C}_1,\mathcal{C}_2)} \times_1 L(w) + \left[\frac{\|\overline{L(\mathcal{W})}_{\mathcal{C}_1} - \overline{L(\mathcal{W})}_{\mathcal{C}_2}\|_F}{1/|\mathcal{C}_1| + 1/|\mathcal{C}_2|}\right] \nu(\mathcal{C}_1, \mathcal{C}_2) \otimes \text{dir}(\overline{L(w)}_{\mathcal{C}_1} - \overline{L(w)}_{\mathcal{C}_2})^\top\Bigg)\Bigg).$$

Define $\varphi = \|\overline{L(\mathcal{W})}_{\mathcal{C}_1} - \overline{L(\mathcal{W})}_{\mathcal{C}_2}\|_F$ and the set $\mathcal{S}(w, \mathcal{C}_1, \mathcal{C}_2) = \{\varphi \geq 0 : \mathcal{C}_1, \mathcal{C}_2 \in \mathcal{C}(\pi^{\perp}_{\nu(\mathcal{C}_1,\mathcal{C}_2)} \times_1 L(w) + (\varphi/(1/|\mathcal{C}_1| + 1/|\mathcal{C}_2|))\, \nu(\mathcal{C}_1, \mathcal{C}_2) \otimes \text{dir}(\overline{L(w)}_{\mathcal{C}_1} - \overline{L(w)}_{\mathcal{C}_2})^\top)\}$, the selective p-value has the form $p_{selective} = \mathbb{P}_{H_0^{\{\mathcal{C}_1,\mathcal{C}_2\}}}(\varphi \geq \|\overline{L(w)}_{\mathcal{C}_1} - \overline{L(w)}_{\mathcal{C}_2}\|_F | \mathcal{C}_1, \mathcal{C}_2 \in \mathcal{S}(w, \mathcal{C}_1, \mathcal{C}_2))$. Finally, since $\varphi = \|\overline{L(\mathcal{W})}_{\mathcal{C}_1} - \overline{L(\mathcal{W})}_{\mathcal{C}_2}\|_F$ and $\text{cov}(\text{vec}(L(\mathcal{W})[i, :, :])) = I_{nq}$, the random variable φ follows the χ_{nq} distribution and thus finishes the proof.

B Supplementary Figures

In this section, we present the auxiliary figures for both numerical simulation and EHR-dataset application in Section 5 and Section 6.

Q-Q plot of the selective p-value under the global null



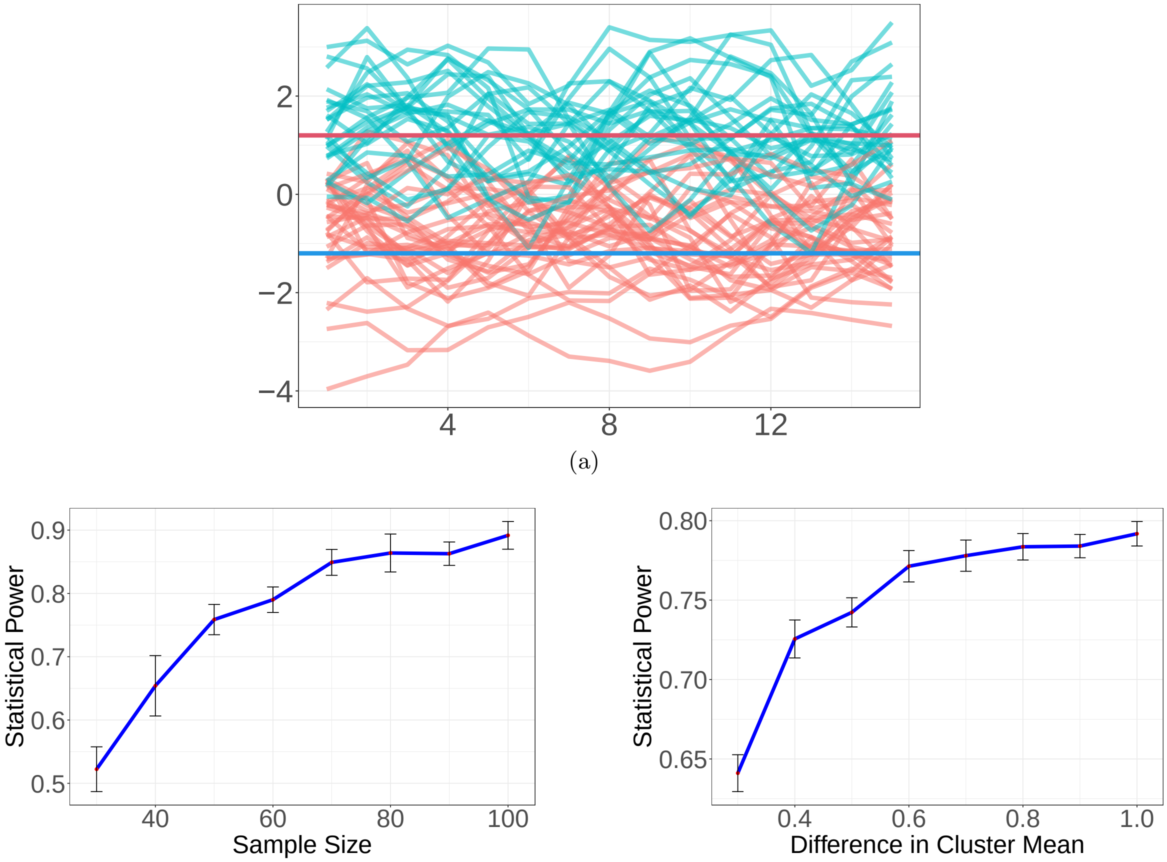
(a) (b)

(c) (d)

(e) (f)

Figure 4: **Left column**: Records of the first feature and the first 5 patients for the dataset generated with 15 time points. **Right column**: Quantile plots of the selective p-value for the corresponding kernel with 100 generated datasets, where **(b)** is the result of **RQ Kernel**, **(d)** is the result of **PE Kernel**, and **(f)** is the result of **LPE Kernel**.

Statistical Power



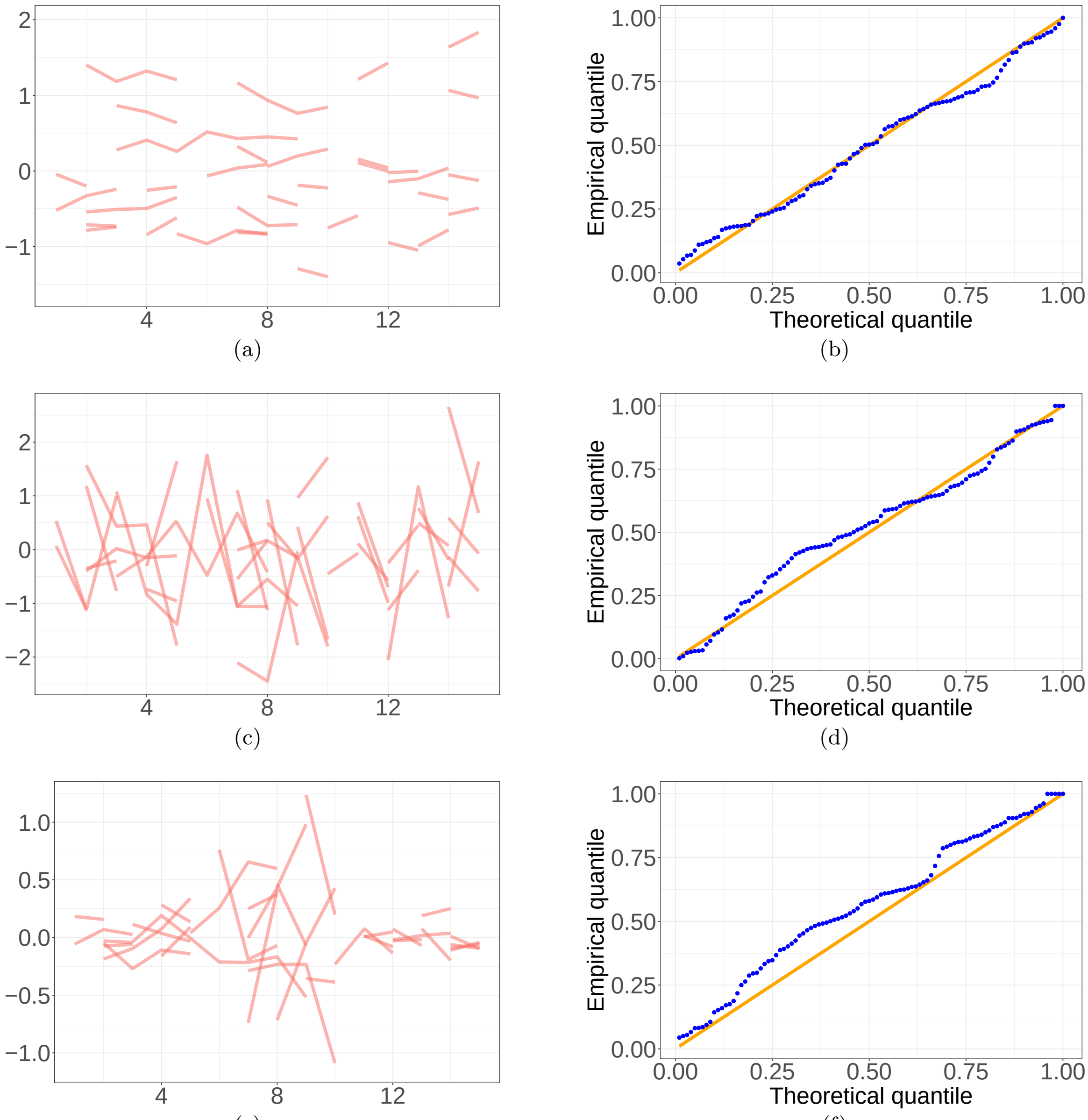
(a)

(b)

(c)

Figure 5: **(a)**: Example of the dataset generated under $H_1^{\{\mathcal{C}_1,\mathcal{C}_2\}}$ with 15 time points and sample size $m = 60$, where sample means are $\mu_i = (1.1)_{n\times T}$ for $i \leq 50$ and $\mu_i = (-1.1)_{n\times T}$ for $i > 50$. **(b)**: Statistical power with sample size $m \in \{30, 40, \cdots, 100\}$, where sample means are $\mu_i = (10)_{n\times T}$ for $i \leq m/2$ and $\mu_i = (-10)_{n\times T}$ for $i > m/2$. **(c)**: Statistical power with sample size $m = 60$ and sample means $\mu_i = (k)_{n\times T}$ for $i \leq m/2$ and $\mu_i = (-k)_{n\times T}$ for $i > m/2$, where $k \in \{0.3, 0.4, \cdots, 1\}$.

Q-Q plot of the selective *p*-value under the global null with missing values



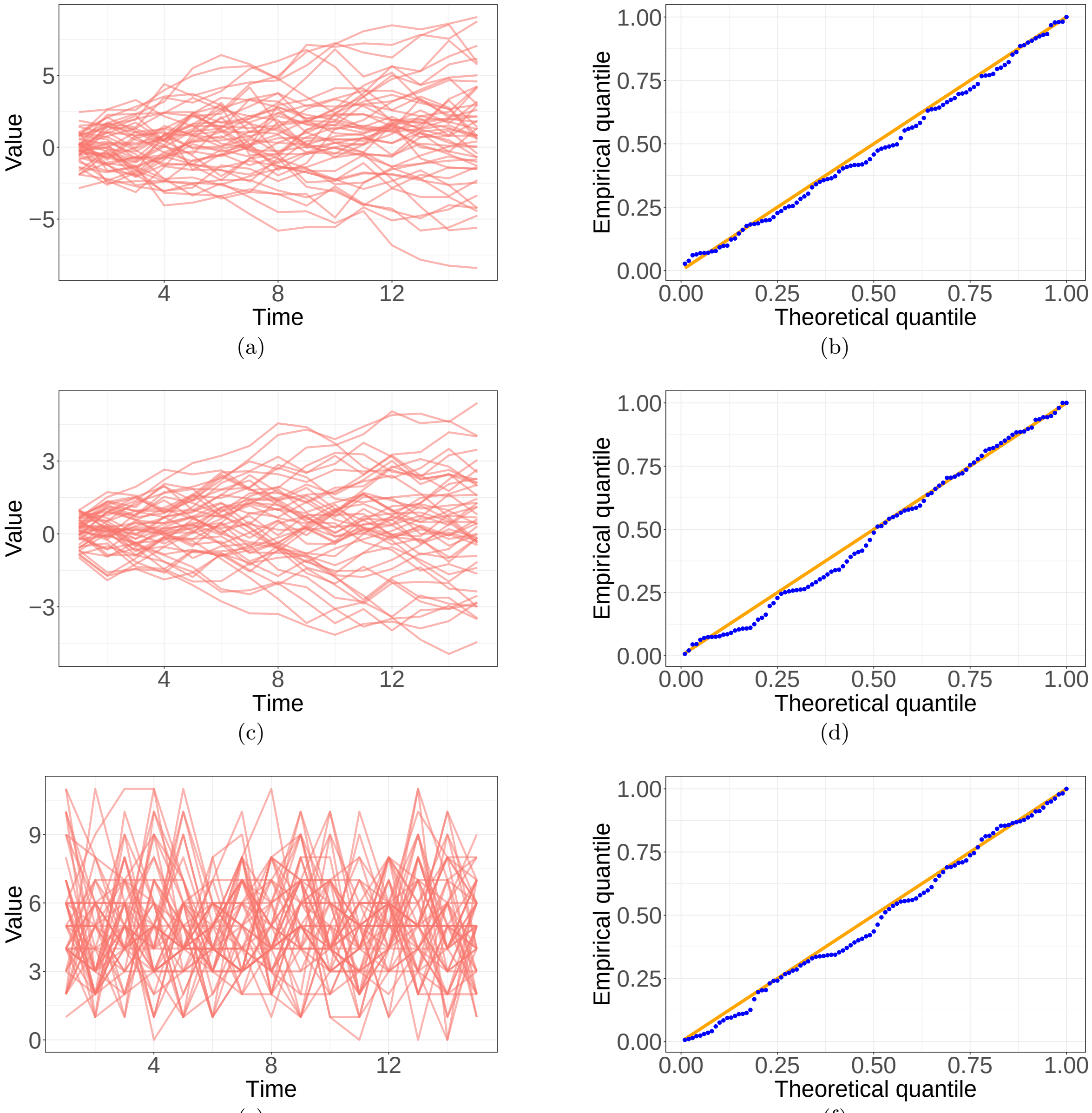
(a) (b)

(c) (d)

(e) (f)

Figure 6: **Left column**: Records of the first feature of the first 5 patients for the generated dataset with 15 time points, we randomly drop 50% data points as missing values. **Right column**: Quantile plots of the selective *p*-value for the corresponding kernel with 100 generated datasets, where **(b)** is the result of **RQ Kernel**, **(d)** is the result of **PE Kernel**, and **(f)** is the result of **LPE Kernel**.

Q-Q plot of the selective p-value under global null (misspecification cases)



(a) (b)

(c) (d)

(e) (f)

Figure 7: **Left column**: Records of the first feature of the first 5 patients for the dataset generated with 15. **Right column**: quantile plots of the selective p-value. **(a)**: Each record is generated independently under the Brownian motion. **(c)**: Each record is generated independently under the uniform random walk. **(e)**: Each record is generated independently under the Poisson process.