Testing Regression Models with Kernel Fisher Discriminant Analysis[*]

Yuhao Li
Xi'an Jiaotong-Liverpool University
yuhao.li@xjtlu.edu.cn

Xiaojun Song
Peking University
sxj@gsm.pku.edu.cn

Abstract

This paper introduces a novel approach to enhancing the goodness-of-fit test for regression models using kernel Fisher discriminant analysis. The proposed method incorporates the covariance structure of integrated regression functions into the test statistics. Unlike existing test statistics, the new approach uniformly weights the components associated with the leading eigenvalues of the covariance operator and downweights the remaining ones. This allows for greater testing power by focusing on a user-tunable number of components. Additionally, under certain assumptions regarding the convergence speed of the regularization term, the test statistic can be made pivotal.

Keywords: Regression Models, RKHS, KFDA, Covariance operator, Regularization

JEL Classification: C12, C52

[*] This paper is a work in progress. Please do not cite without permission. This paper is funded by the Research Development Fund (RDF-23-02-022) of the Xi'an Jiaotong-Liverpool University

1. Introduction

Regression models are widely used in empirical research, and rigorously testing these models is crucial. This testing ensures that the models accurately reflect the relationships proposed by economic theory and the underlying data-generating processes.

The majority of existing tests are based on the integrated regression function (IRF), which extends classical goodness-of-fit tests for cumulative distribution function (CDF) specifications to the testing of regression models. For further details, refer to González-Manteiga & Crujeiras (2013) for an in-depth review. Within this framework, two primary approaches to the testing problem can be identified. The first approach relies on local smoothing methods for regression, while the second involves the construction of empirical regression processes, often referred to as the integrated conditional moment (ICM) approach. Noteworthy works aligned with the first approach include Hardle & Mammen (1993), Zheng (1996), Li & Wang (1998), Dette (1999), and Hsiao et al. (2007), among others. Works following the second approach include Bierens (1982), Delgado (1993), Andrews (1997), Bierens & Ploberger (1997), Stute (1997), Delgado et al. (2006), and Sant'Anna & Song (2019), to name a few.

These two approaches have long been considered complementary (see Fan & Li (2000) for a detailed discussion): ICM tests exhibit greater power than local smoothing tests against Pitman-type local alternatives and are also insensitive to the dimension of the covariates. In contrast, local smoothing tests demonstrate greater power against high-frequency alternatives than ICM tests and are generally pivotal. However, none of the existing works address scenarios where the dimension of the covariates is large and deviations from the null occur in high-frequency directions simultaneously.

Addressing this challenge requires more than just identifying a missing piece in a puzzle; it necessitates a comprehensive analysis of the power properties of the underlying test statistics. In this paper, we reframe the testing problem as a classification problem, and utilize recently developed kernel mean embedding techniques (see Muandet et al. (2017) for a detailed review) from the machine learning community to create a test statistic based on a maximized kernel Fisher discriminant ratio (KFD ratio). Our test statistic combines the strengths of the two aforementioned approaches while avoiding their limitations. Specifically, our test statistic has the following properties.

Good Performance Against High-Frequency Alternatives. We conduct spectral analysis and demonstrate that test statistics based on the IRF can be resolved into a linear combination of infinitely many orthogonal directional test statistics, with descending weights $\{\lambda_j\}_{j\geq 1}$ attributed to each of these directional statistics. Specifically, higher frequency directions are assigned lower weights. The KFD ratio addresses the higher frequency directions by adjusting the weights to $\{(\lambda_j)/(\lambda_j+\gamma_n)\}_{j\geq 1}$, where γ_n represents a regularization parameter. In comparison, local smoothing test statistics also accommodate the higher frequency deviations by employing identical weights to all spectral directions as the bandwidth goes to zero, $h \to 0$.

$n^{-1/2}$-rate against local alternatives. When the regularization parameter is fixed at a constant: $\gamma_n := \gamma$, our test statistic can detect local deviations that converge in probability to the null model at a rate of $n^{-1/2}$.

Pivotal with vanishing regularization parameter. When $\gamma_n \to 0$, our test statistic, under some conditions, will converge to a standard normal distribution. In this scenario, our statistic remains responsive to high-frequency deviations; however, it only exhibits non-trivial power against local alternatives that converge to the null at a rate of $n^{-1/4}$.

Insensitive to covariate dimension. Through the appropriate selection of reproducing kernels, our statistic remains insensitive to the dimension of the covariates, provided that this dimension is fixed.

The structure of the paper is as follows. In the next section, we introduce key concepts related to the reproducing kernel Hilbert space (RKHS), including the mean element and covariance operator. Section 3 discusses the main components of the proposed test statistic, covering its computation and the effects of estimation. We also perform a spectral analysis to demonstrate the statistic's sensitivity to high-frequency deviations. Section 4 presents the asymptotic results, detailing the null distributions under both fixed and vanishing regularization parameters, as well as consistency and local alternative results. In Section 5, we propose a multiplier bootstrap algorithm for determining critical values when the regularization parameter is fixed. Section 6 provides simulation results. In Section 7, we compare the proposed test statistic with existing ones and explore its connections to these statistics. Finally, Section 8 concludes the paper.

2. Mean Element and Covariance Operator in RKHS

2.1. Some Operator-Theoretic Tools

A linear operator T is said to be *bounded* if there is a number C such that $\|Tf\|_{\mathcal{H}} \leq C\ \|f\|_{\mathcal{H}}$ for all $f \in \mathcal{H}$. The operator norm of T is then defined as the minimum of such numbers C, that is

$$\|T\| = \sup_{\|f\|_{\mathcal{H}} \leq 1} \|Tf\|_{\mathcal{H}}$$

.

Futhermore, a bounded linear operator $\|T\|$ is said to be Hilbert-Schmidt, if the Hilbert-Schmidt norm

$$\|T\|_{HS} = \left\{\sum_{l=1}^{\infty} < Te_l, Te_l >_{\mathcal{H}}\right\}^{\frac{1}{2}} = \left\{\sum_{l=1}^{\infty} \lambda_l^2\right\}^{\frac{1}{2}} < \infty$$

where λ_l and e_l are the eigenvalues and eigenfunctions of the operator T.

The tensor product operator $u \otimes v$ for $u, v \in \mathcal{H}$ is defined for all

$$(u \otimes v)f = \langle v, f \rangle_{\mathcal{H}} u$$

2.2. Mean Element and Covariance Operator

Given a sample $\{Z_1, ..., Z_n\}$, where $Z_i = \left(Y_i, X_i^\top\right)^\top$, the dependent variable Y depends on the explanatory variables X by:

$$
\begin{aligned}
Y &= \mathbb{E}(Y|X) + \varepsilon \\
&= \mathcal{M}_{\theta_0}(X) + \varepsilon
\end{aligned}
$$

where $\mathcal{M}_{\theta_0}$ is a parametrical model indexed by a vector of parameters θ_0. Let $\hat{\theta}$ be a consistent esitmator of θ_0 under the null (i.e., the model is correctly specified). The residual vector is denoted by $\hat{\varepsilon} = \left(\varepsilon_1\left(\hat{\theta}\right), ..., \varepsilon_i\left(\hat{\theta}\right)\right)^\top$, where

$$\varepsilon_i\left(\hat{\theta}\right) = Y_i - \mathcal{M}_{\hat{\theta}}(X_i)$$

We will denote $\varepsilon_i\left(\hat{\theta}\right)$ by $\hat{\varepsilon}_i$ and $\varepsilon_i(\theta_0)$ by ε_i whenever there is no ambiguity.

Let k be a bounded reproducing kernel, i.e.,

$$\sup_{(x,y)\in\mathcal{X}\times\mathcal{X}} k(x,y) < \infty$$

If $\int k^{1/2}(x,x)\mathbb{P}(dx) < \infty$, then the mean element μ is defined as the unique element in $\mathcal{H}$ satisfying for all functions $f \in \mathcal{H}$,

$$\langle \mu, f \rangle_{\mathcal{H}} = \mathbb{P}f := \int \varepsilon f d\mathbb{P}$$

or alternatively,

$$\mu = \mathbb{E}(\varepsilon k(X, \cdot))$$

where under the null $\varepsilon := \varepsilon_{\theta_0}$

If furthermore $\int k(x,x)\mathbb{P}(dx) < \infty$, then the uncentered covariance operator $\Sigma_{\mathbb{P}}$ is defined as the unique linear operator onto $\mathcal{H}$ satisfying for all $f, g \in \mathcal{H}$,

$$\langle f, \Sigma_{\mathbb{P}} g \rangle_{\mathcal{H}} := \int (\varepsilon f)(\varepsilon' g) d\mathbb{P}$$

Or alternatively,

$$\Sigma := \mathbb{E}\left(\varepsilon^2 k(X, \cdot) \otimes k(X, \cdot)\right)$$

We now define an important concept denoted as $\Sigma^{-1/2}$. For a compact operator Σ, the range $\mathcal{R}\left(\Sigma^{1/2}\right)$ of $\Sigma^{1/2}$ is characterized as

$$\mathcal{R}\left(\Sigma^{1/2}\right) = \left\{ f \in \mathcal{H}, \sum_{l=1}^{\infty} \lambda_l \langle f, e_l \rangle_{\mathcal{H}}^2 < \infty, f \perp \mathcal{N}\left(\Sigma^{1/2}\right) \right\}$$

where $\{\lambda_l\}_{l\geq 1}$ and $\{e_l\}_{l\geq 1}$ are eigenvalues and eigenvectors of the covariance operator Σ. And

$$\mathcal{N}(\Sigma) = \{f \in \mathcal{H}, \Sigma f = 0\}$$

is the null-space of Σ.

Defining

$$\mathcal{R}^{-1}\left(\Sigma^{1/2}\right) = \left\{ g \in \mathcal{H}, g = \sum_{l=1}^{\infty} \lambda_l^{-1/2} \langle f, e_l \rangle_{\mathcal{H}} e_l, f \in \mathcal{R}\left(\Sigma^{1/2}\right) \right\}$$

Thus, $\Sigma^{1/2}$ is a one-to-one mapping between $\mathcal{R}^{-1}\left(\Sigma^{1/2}\right)$ and $\mathcal{R}\left(\Sigma^{1/2}\right)$. Thus, restricting the domain of $\Sigma^{1/2}$ to $\mathcal{R}^{-1}\left(\Sigma^{1/2}\right)$, we may define its inverse for all $f \in \mathcal{R}\left(\Sigma^{1/2}\right)$ as

$$\Sigma^{-1/2} f = \sum_{l=1}^{\infty} \lambda_l^{-1/2} \langle f, e_l \rangle_{\mathcal{H}} e_l$$

The empirical estimates respectively of the mean element and the covariance operator are then defined as follows:

$$\begin{aligned} \hat{\mu} &:= \frac{1}{n} \sum_{i=1}^{n} k(X_i, \cdot) \hat{\varepsilon}_i \\ &:= \mathbb{E}_n\left(\hat{\varepsilon} k(X, \cdot)\right) \end{aligned} \tag{1}$$

$$\begin{aligned} \hat{\Sigma} &:= \frac{1}{n} \sum_{i=1}^{n} \hat{\varepsilon}_i^2 k(X_i, \cdot) \otimes k(X_i, \cdot) \\ &:= \mathbb{E}_n\left(\hat{\varepsilon}^2 k(X, \cdot) \otimes k(X, \cdot)\right) \end{aligned} \tag{2}$$

Remark. By the reproducing property, the empirical mean element has the form:

$$\langle \hat{\mu}, f \rangle = \frac{1}{n} \sum_{i=1}^{n} \hat{\varepsilon}_i f(X_i), \forall f \in \mathcal{H}$$

On the other hand, by the fact that $(u \otimes v) f = \langle v, f \rangle_{\mathcal{H}} u$, we have

$$\begin{aligned} \hat{\Sigma} g &= \left(\frac{1}{n} \sum_{i=1}^{n} \hat{\varepsilon}_i^2 k(X_i, \cdot) \otimes k(X_i, \cdot) \right) g \\ &= \frac{1}{n} \sum_{i=1}^{n} \hat{\varepsilon}_i^2 \langle k(X_i, \cdot), g(\cdot) \rangle_{\mathcal{H}} k(X_i, \cdot) \\ &= \frac{1}{n} \sum_{i=1}^{n} \hat{\varepsilon}_i^2 g(X_i) k(X_i, \cdot) \\ &= \mathbb{E}_n\left(\hat{\varepsilon}^2 g(X) k(X)\right) \end{aligned}$$

Hence,

$$\begin{aligned} \langle f, \hat{\Sigma} g \rangle_{\mathcal{H}} &= \frac{1}{n} \sum_{i=1}^{n} \hat{\varepsilon}_i^2 f(X_i) g(X_i) \\ &= \mathbb{E}_n\left(\hat{\varepsilon}^2 f(X) g(X)\right) \end{aligned}$$

3. The KFDA Test Statistic

3.1. Testing Problem as a Classification Problem

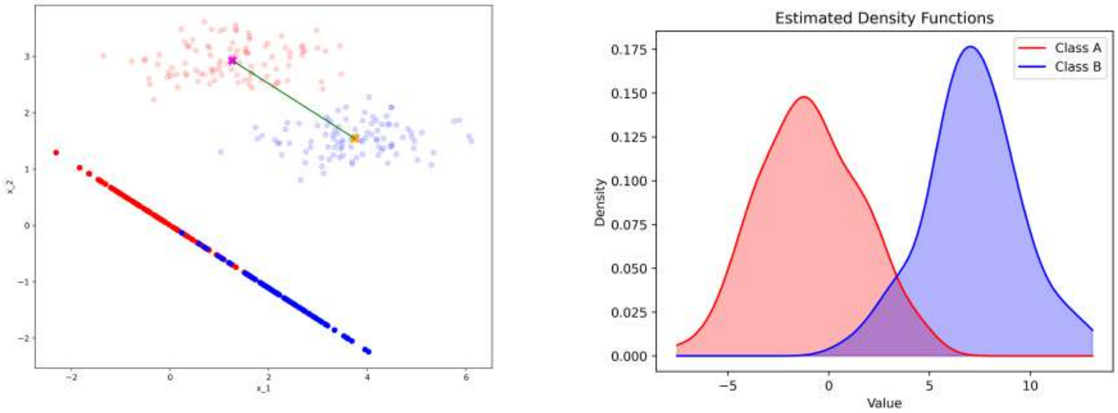
The testing problem can be reframed as a classification problem, which can provide insights into the power of the test. Under the alternative hypothesis, one can construct two classes of data: $\left\{y_i f_{x_i}(\cdot)\right\}$ and $\left\{\mathcal{M}_{\hat{\theta}}(x_i) f_{x_i}(\cdot)\right\}$, where $f_x(\cdot)$ is a member of an RKHS $\mathcal{H}$ with infinite dimensions. The first class consists of data generated by the true underlying distribution, while the second class consists of data generated by the estimated model. To enhance the power, it is essential to find a classification rule $f_x(\cdot)$ that effectively separates these two classes:

$$\max_{f \in \mathcal{H}} \| \ \mathbb{E}_n\left(\left(y - \mathcal{M}_{\hat{\theta}}(x)\right) f_x(\cdot)\right) \ \|_{\mathcal{H}}$$

where $\mathbb{E}_n\left(\left(y - \mathcal{M}_{\hat{\theta}}(x)\right) f_x(\cdot)\right)$ is the sample mean average of the two classes, and can be understood as "signals" of the deviation.

However, as Fisher (1936) argued, maximizing the signal alone is not sufficient; the noise of the test statistic should also be minimized for optimal performance. The key idea of Fisher's discriminant analysis can be illustrated by the following simple linear classification example.

Consider two classes of 2-dimensional random samples: $\left\{X_1, X_2 \in \mathbb{R}^2\right\}$, where X_1 and X_2 are generated from two different distributions. One might first project the 2-dimensional data into a 1-dimensional space using a linear projection[2]. A straightforward approach is to find a projection direction that maximizes the separation of the two classes' means, as illustrated in Figure 1a. However, this projection is not optimal, as the distributions of the projected data might overlap, as shown in Figure 1b. In our context, a significant overlap in the distributions of these two classes would result in a high type II error rate.



(a) Simple Linear Projection

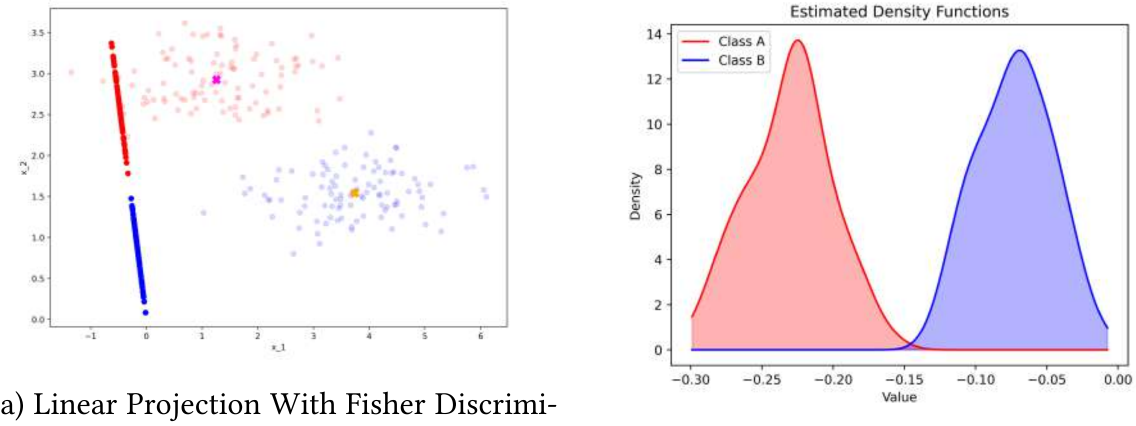
(b) Distributions of Projected Data

Figure 1: Simple Classification

In contrast, Fisher discriminant analysis seeks a projection direction that not only maximizes the separation between the means of the two classes but also minimizes the within-class variance. This is depicted in Figure 2a. Although the absolute difference between the means of the

[2] The example here is only for illustration; in our context, we will map data from a low-dimensional space to the RKHS, which is an infinite-dimensional Hilbert space

two classes is smaller compared to the simple projection, the distributions of the projected data are more distinctly separated, as shown in Figure 2b. Fisher discriminant analysis achieves this by finding a projection direction that maximizes the ratio of between-class variance to within-class variance. In our context, this approach reduces the type II error rate by minimizing the overlap between the distributions of the two classes in the projected space.



(a) Linear Projection With Fisher Discriminant Analysis

(b) Distributions of Projected Data

Figure 2: Distinct Classification

3.2. Structure of the Statistic

Let $\Sigma_B := \mu \otimes \mu$ be the between class covariance operator. Let $\{\gamma_n\}_{n\geq 1}$ be a sequence of strictly positive numbers. The *Maximum Kernel Fisher Discriminant Ratio* serves as a basis of the proposed test statistic:

$$n \max_{f \in \mathcal{H}} \frac{\langle f, \hat{\Sigma}_B f \rangle_{\mathcal{H}}}{\langle f, \left(\hat{\Sigma} + \gamma_n I\right)^{-1} f \rangle_{\mathcal{H}}} \tag{3}$$

where I denotes the identity operator. The goal of the Fisher discriminant analysis is give a large separation of the class means while also keeping the in-class variance small.

The above optimization problem is equivalent to:

$$n \max_{f \in \mathcal{H}} \langle f, \hat{\Sigma}_B f \rangle_{\mathcal{H}}, \quad s.t$$

$$\langle f, \left(\hat{\Sigma} + \gamma_n I\right)^{-1} f \rangle_{\mathcal{H}} = 1$$

By the Lagrange multiplier argument, it is easy to show that

$$f^* \propto \left(\hat{\Sigma} + \gamma_n I\right)^{-1} \hat{\mu}$$

The maximized kernel Fisher discriminant ratio is:

$$n \left\| \left(\hat{\Sigma} + \gamma_n I\right)^{-1/2} \hat{\mu} \right\|_{\mathcal{H}}^2 \tag{4}$$

Theorem 1 (Equivalence to the Null). Let the reporducing kernel $k(\cdot,\cdot)$ be integrally strictly positive definite:

$$\int\int f(x)f(y)k(x,y)d\eta(x)d\eta(y) > 0$$

where $\eta(\cdot)$ is a valid probability measure. The null hypothesis holds true, i.e.,

$$\mathbb{E}(\varepsilon|X) = 0$$

if and only if

$$\left\|(\Sigma + \gamma_n I)^{-1/2}\mu\right\|_{\mathcal{H}}^2 = 0$$

Proof: See Section A.1. □

3.3. Computation of the Statistic

The kernel trick would facilitate the computation of the test statistic. Specifically, denote $\hat{\boldsymbol{G}} : \mathbb{R}^n \to \mathcal{H}$, a vector in $\mathcal{H}$:

$$\hat{\boldsymbol{G}} = (\hat{\varepsilon}_1 k(X_1,\cdot), ..., \hat{\varepsilon}_n k(X_n,\cdot))$$

Let $\widehat{\boldsymbol{K}} = \hat{\boldsymbol{G}}^\top \hat{\boldsymbol{G}}$ be the Gram matrix given by

$$\widehat{\boldsymbol{K}}(i,j) = \hat{\varepsilon}_i k(X_i, X_j)\hat{\varepsilon}_j \quad \text{for} \quad i,j \in \{1,...,n\}$$

Finally, define the vector $\boldsymbol{m}_n = (\boldsymbol{m}_{n,i})_{1\leq i\leq n}$ with $\boldsymbol{m}_{n,i} = 1/n$.

With these notations introduced above, we have

$$\hat{\mu} = \hat{\boldsymbol{G}}\boldsymbol{m}_n \tag{5}$$

$$\hat{\Sigma} = \frac{1}{n}\hat{\boldsymbol{G}}\hat{\boldsymbol{G}}^\top \tag{6}$$

and finally,

$$\begin{aligned}
& n\left\|\left(\hat{\Sigma} + \gamma_n I\right)^{-1/2}\hat{\mu}\right\|_{\mathcal{H}}^2 \\
&= \langle\hat{\mu}, \left(\hat{\Sigma} + \gamma_n I\right)^{-1}\hat{\mu}\rangle_{\mathcal{H}} \\
&= n\boldsymbol{m}_n^\top \hat{\boldsymbol{G}}^\top \left(\frac{1}{n}\hat{\boldsymbol{G}}\hat{\boldsymbol{G}}^\top + \gamma I\right)^{-1}\hat{\boldsymbol{G}}\boldsymbol{m}_n \\
&= n\gamma_n^{-1}\boldsymbol{m}_n^\top \hat{\boldsymbol{G}}^\top \left\{I - n^{-1}\hat{\boldsymbol{G}}\left(\gamma_n I + n^{-1}\hat{\boldsymbol{G}}^\top\hat{\boldsymbol{G}}\right)^{-1}\hat{\boldsymbol{G}}^\top\right\}\hat{\boldsymbol{G}}\boldsymbol{m}_n \\
&= n\gamma_n^{-1}\left\{\boldsymbol{m}_n^\top\widehat{\boldsymbol{K}}\boldsymbol{m}_n - n^{-1}\boldsymbol{m}_n^\top\widehat{\boldsymbol{K}}\left(\gamma_n I + n^{-1}\widehat{\boldsymbol{K}}\right)^{-1}\widehat{\boldsymbol{K}}\boldsymbol{m}_n\right\}
\end{aligned} \tag{7}$$

where the third equality comes from the matrix inversion lemma.

3.4. Dealing with the Estimation Effects

The estimation effects arise because the empirical mean element and covariance operator depend on the parameter estimators $\hat{\theta}$, which are estimated from the same data used for testing. To mitigate these effects, we introduce a projection mean difference in the RKHS.

Let

$$g_i(\theta) := \nabla_\theta \varepsilon_i(\theta)$$

be the gradient of $\varepsilon_i(\theta)$ and let $\bar{\theta} = \delta\theta + (1 - \delta)\theta, \delta \in (0, 1)$. The term $n\widehat{\boldsymbol{K}}_n$ can be decomposed into:

$$\begin{aligned} n\widehat{\boldsymbol{K}}_n(i,j) &= n\left(\varepsilon_i + g_i^\top(\bar{\theta})\left(\hat{\theta} - \theta_0\right)\right)k(X_i, X_j)\left(\varepsilon_j + g_j^\top(\bar{\theta})\left(\hat{\theta} - \theta_0\right)\right) \\ &= \underbrace{n\varepsilon_i k(X_i, X_j)\varepsilon_j}_{A_1(i,j)} + \sqrt{n}\underbrace{\left(\varepsilon_i k(X_i, X_j) g_j^\top(\bar{\theta}) + \varepsilon_j k(X_i, X_j) g_i^\top(\bar{\theta})\right)}_{A_2(i,j)}\sqrt{n}\left(\hat{\theta} - \theta_0\right) \\ &\quad + \sqrt{n}\left(\hat{\theta} - \theta_0\right)^\top \underbrace{g_i(\bar{\theta})k(X_i, X_j)g_j^\top(\bar{\theta})}_{A_3(i,j)}\sqrt{n}\left(\hat{\theta} - \theta_0\right) \\ &= nA_1(i,j) + \sqrt{n}A_2(i,j)O_p(1) + \left(o_p(1)\right)^\top \sqrt{n}A_3(i,j)O_p(1) \end{aligned}$$

and

$$\begin{aligned} n\boldsymbol{m}_n^\top \widehat{\boldsymbol{K}} \boldsymbol{m}_n &= \frac{n}{n^2}\sum_{i,j=1}^{n} A_1(i,j) + \frac{\sqrt{n}}{n^2}\sum_{i,j=1}^{n} A_2(i,j)O_p(1) + \left(o_p(1)\right)^\top \frac{\sqrt{n}}{n^2}\sum_{i,j=1}^{n} A_3(i,j)O_p(1) \\ &= \underbrace{nA_1}_{O_p(1)} + \underbrace{\sqrt{n}A_2}_{O_p(1)}O_p(1) + \underbrace{\left(o_p(1)\right)^\top \underbrace{\sqrt{n}A_3}_{O_p(1)}O_p(1)}_{o_p(1)} \end{aligned}$$

where under the null, the last equality comes from A_1 is a degenerate V-statistic, and A_2 and A_3 are non degenerate.

From this decomposition, it is clear that the estimation effects are introduced via the second term i.e., A_2. To eliminate this estimation effect, we introduce a projection mean difference $\left\{\hat{\varepsilon}_i k_p(x_i, \cdot)\right\}$ in RKHS:

$$\hat{\varepsilon}_i k_p(x_i, \cdot) = \hat{\varepsilon}_i k(x_i, \cdot) - g_i^\top\left(\hat{\theta}\right)\hat{\Gamma}^{-1}\mathbb{E}_n\left(g\left(\hat{\theta}\right)\hat{\varepsilon}k(X, \cdot)\right) \tag{8}$$

where $\hat{\Gamma} = \mathbb{E}_n\left(g\left(\hat{\theta}\right)g^\top\left(\hat{\theta}\right)\right)$, and $\Gamma = \mathbb{E}(g(\theta_0)g^\top(\theta_0))$ is assumed to be non-singular.

Let $\hat{\boldsymbol{G}}_p = \left(\hat{\varepsilon}_1 k_p(x_1, \cdot), ..., \hat{\varepsilon}_n k_p(x_n, \cdot)\right)$ and $\hat{\boldsymbol{g}}$ be a $n \times d$ matrix with i-th row $g_i^\top\left(\hat{\theta}\right)$, the vectorized Equation 8 reads:

$$\hat{\boldsymbol{G}}_p^\top = \hat{\boldsymbol{\varepsilon}}\boldsymbol{k_p}(\boldsymbol{X}, \cdot) = \hat{\boldsymbol{\varepsilon}}\boldsymbol{k}(\boldsymbol{X}, \cdot) - \hat{\boldsymbol{g}}\left(\hat{\boldsymbol{g}}^\top\hat{\boldsymbol{g}}\right)^{-1}\hat{\boldsymbol{g}}^\top\hat{\boldsymbol{\varepsilon}}\boldsymbol{k}(\boldsymbol{X}, \cdot)$$

then it is easy to show that

$$\begin{aligned}\hat{\boldsymbol{G}}_p^\top &= \underbrace{\boldsymbol{\varepsilon k}(\boldsymbol{X},\cdot) - \hat{\boldsymbol{g}}\left(\hat{\boldsymbol{g}}^\top\hat{\boldsymbol{g}}\right)^{-1}\hat{\boldsymbol{g}}^\top\boldsymbol{\varepsilon k}(\boldsymbol{X},\cdot)}_{\boldsymbol{G}_{p,n}} - \hat{\boldsymbol{g}}\left(\hat{\boldsymbol{g}}^\top\hat{\boldsymbol{g}}\right)^{-1}\hat{\boldsymbol{g}}^\top\bar{\boldsymbol{g}}\left(\hat{\theta}-\theta_0\right)\boldsymbol{k}(\boldsymbol{X},\cdot)\\ &\quad+\bar{\boldsymbol{g}}\left(\hat{\theta}-\theta_0\right)\boldsymbol{k}(\boldsymbol{X},\cdot)\\ &= \boldsymbol{G}_p - \hat{\boldsymbol{g}}\left(\hat{\boldsymbol{g}}^\top\hat{\boldsymbol{g}}\right)^{-1}\hat{\boldsymbol{g}}^\top\left(\hat{\boldsymbol{g}} + O_p\left(n^{-1/2}\right)\right)\left(\hat{\theta}-\theta_0\right)\boldsymbol{k}(\boldsymbol{X},\cdot) + \left(\hat{\boldsymbol{g}} + O_p\left(n^{-1/2}\right)\right)\left(\hat{\theta}-\theta_0\right)\boldsymbol{k}(\boldsymbol{X},\cdot)\\ &= \boldsymbol{G}_p + O_p\left(n^{-1}\right)\end{aligned}\tag{9}$$

Thus,

$$\begin{aligned}\hat{\mu}_p := \hat{\boldsymbol{G}}_p\boldsymbol{m}_n &= \boldsymbol{G}_{p,n}\boldsymbol{m}_n + O_p\left(n^{-1}\right)\\ &= \mu_{p,n} + O_p\left(n^{-1}\right)\end{aligned}\tag{10}$$

Similary, one can show

$$\boldsymbol{m}_n^\top\widehat{\boldsymbol{K}}_p\boldsymbol{m}_n = \boldsymbol{m}_n^\top\boldsymbol{K}_{p,n}\boldsymbol{m}_n + O_p\left(n^{-1}\right)$$

where

$$\begin{aligned}\widehat{\boldsymbol{K}}_p &= \hat{\boldsymbol{G}}_p^\top\hat{\boldsymbol{G}}_p\\ \boldsymbol{K}_{p,n} &= \boldsymbol{G}_{p,n}^\top\boldsymbol{G}_{p,n}\end{aligned}$$

Let $\hat{\Sigma}_p = n^{-1}\hat{\boldsymbol{G}}_p\hat{\boldsymbol{G}}_p^\top$, $\Sigma_{n,p} = n^{-1}\boldsymbol{G}_{p,n}\boldsymbol{G}_{p,n}^\top$. Note that

$$\begin{aligned}\hat{\Sigma}_p &= \mathbb{E}_n\left(\hat{\varepsilon}^2 k_p(x,\cdot)\otimes k_p(x,\cdot)\right)\\ &= \mathbb{E}_n\left(\left(\varepsilon + g\left(\bar{\theta}\right)^\top\left(\hat{\theta}-\theta_0\right)\right)^2 k_p(x,\cdot)\otimes k_p(x,\cdot)\right)\end{aligned}$$

Further analysis of it reveals:

$$\begin{aligned}\hat{\Sigma}_p &= \mathbb{E}_n\left(\left(\varepsilon^2 + 2\varepsilon g\left(\bar{\theta}\right)^\top\left(\hat{\theta}-\theta_0\right) + g\left(\bar{\theta}\right)^\top\left(\hat{\theta}-\theta_0\right)\left(\hat{\theta}-\theta_0\right)^\top g\left(\bar{\theta}\right)\right)k_p(x,\cdot)\otimes k_p(x,\cdot)\right)\\ &= \mathbb{E}_n\left(\varepsilon^2 k_p(x,\cdot)\otimes k_p(x,\cdot)\right) + O_p\left(n^{-1/2}\right)^\top\mathbb{E}_n\left(\varepsilon g\left(\bar{\theta}\right)k_p(x,\cdot)\otimes k_p(x,\cdot)\right)\\ &\quad+O_p\left(n^{-1/2}\right)^\top\mathbb{E}_n\left(g\left(\bar{\theta}\right)k_p(x,\cdot)\otimes g\left(\bar{\theta}\right)^\top k_p(x,\cdot)\right)O_p\left(n^{-1/2}\right)\\ &= \Sigma_{n,p} + O_p\left(n^{-1/2}\right)^\top B_{1,n} + O_p\left(n^{-1/2}\right)^\top B_{2,n}O_p\left(n^{-1/2}\right)\end{aligned}$$

Thus,

$$\left\|\hat{\Sigma}_p - \Sigma_{n,p}\right\| = \left\|O_p\left(n^{-1/2}\right)^\top B_{1,n} + O_p\left(n^{-1/2}\right)^\top B_{2,n}O_p\left(n^{-1/2}\right)\right\| = O_p\left(n^{-1/2}\right)\tag{11}$$

The results presented in Equation 10 and Equation 11, with assumptions listed below, lead to Lemma 1, which would be helpful when performing the asymptotic analyses.

> **Assumption 1.1**: (i) The parameter space Θ is a compact subset of $\mathbb{R}^d$; (ii) the true parameter θ_0 is an interior point of Θ; (iii) $\|\hat{\theta}-\theta_0\| = O_p\left(n^{-1/2}\right)$.

Assumption 1.2: The residual $\varepsilon(\theta)$ is twice continuously differentiable with respect to θ, with its first derivative $g(\theta) = \nabla_\theta \varepsilon(\theta)$ satisfying $\mathbb{E}[\sup_{\theta \in \Theta} \|g(\theta)\|] < \infty$ and its second derivative satisfying $\mathbb{E}[\sup_{\theta \in \Theta} \|\nabla_\theta g(\theta)\|] < \infty$. Furthermore, the matrix $\Gamma = \mathbb{E}(g(\theta) g^\top(\theta))$ is nonsingular in a neibourhood of θ_0.

Assumption 1.1 is weaker than related conditions in the literature. We only impose $\sqrt{n}\left(\hat{\theta} - \theta_0\right) = O_p(1)$, but do not require it to admit an asymptotically linear representation. This could be useful in the context of non-standarad estimation procedures, such as the Lasso. Assumption 1.2 is standard in the literature and imposes regularity conditions on the smoothness of the residual function.

Lemma 1. Let $\hat{\mu}_p$ be the projected mean element as presented in Equation 10, assume Assumption 1.1 and Assumption 1.2, then

$$\begin{aligned} & n \left\| \left(\hat{\Sigma}_p + \gamma_n I\right)^{-1/2} \hat{\mu}_p \right\|_{\mathcal{H}}^2 \\ & \quad = n \left\| \left(\Sigma_{p,n} + \gamma_n I\right)^{-1/2} \mu_{p,n} \right\|_{\mathcal{H}}^2 + o_p(1) \end{aligned} \tag{12}$$

Proof: See Section A.2. □

Thus, for asymptotic analysis, it would be suffice to focuse on

$$n \left\| \left(\Sigma_{p,n} + \gamma_n I\right)^{-1/2} \mu_{p,n} \right\|_{\mathcal{H}}^2$$

In the literature of goodness-of-fit test, similar idea of projecting out the estimation effects can be found in Bickel et al. (2006), Escanciano (2009), Escanciano & Goh (2014), and Sant'Anna & Song (2019).

Finally, the proposed test statistic is updated as:

$$\begin{aligned} \hat{T}_n(\gamma_n) &= n \left\| \left(\hat{\Sigma}_p + \gamma_n I\right)^{-1/2} \hat{\mu}_p \right\|_{\mathcal{H}}^2 \\ &= n \gamma_n^{-1} \{ \boldsymbol{m}_n^\top \widehat{\boldsymbol{K}}_p \boldsymbol{m}_n - n^{-1} \boldsymbol{m}_n^\top \widehat{\boldsymbol{K}}_p \left(\gamma_n I + n^{-1} \widehat{\boldsymbol{K}}_p\right)^{-1} \widehat{\boldsymbol{K}}_p \boldsymbol{m}_n \end{aligned}$$

3.5. Spectral Analysis on the (Projected) KFDA statistics

We now analyze the weights of the projected KFDA statistics. Let $\{e_l\}_{l \geq 1}$ be the eigenfunctions of Σ_p, and $\left\{\lambda_l\left(\Sigma_p\right)\right\}_{l \geq 1}$ are the corresponding eigenvalues. Furthermore, let

$$f_l = \lambda_l^{-1/2}\left(\Sigma_p\right)(\varepsilon e_l) \tag{13}$$

Notice that

$$\begin{aligned}\lambda_k(\Sigma_p)\delta_{k,l} = \langle e_l, \Sigma_p e_k\rangle_{\mathcal{H}} &= \langle(\varepsilon e_k), (\varepsilon' e_l)\rangle_{L^2(\mathbb{P})} \\ &= \lambda_k^{1/2}(\Sigma_p)\lambda_l^{1/2}(\Sigma_p)\langle f_k, f_l\rangle_{L^2(\mathbb{P})}\end{aligned}$$

where $\delta_{k,l}$ is the Kronecker's delta. Hence $\{f_k\}_{k\geq 1}$ is an orthonormal system of $L^2(\mathbb{P})$, and $\mathbb{P}$ is a joint probability measure of (ε, X).

$$\begin{aligned}& n\left\|\left(\Sigma_{n,p} + \gamma_n I\right)^{-1/2}\mu_{p,n}\right\|_{\mathcal{H}}^2 \\ &= n\left\|\sum_{j\geq 1}\left(\lambda_j(\Sigma_{n,p}) + \gamma_n\right)^{-1/2}\langle\mu_{p,n}, e_j\rangle_{\mathcal{H}} e_j(\cdot)\right\|_{\mathcal{H}}^2 \\ &\overset{(1)}{=} n\left\|\sum_{j\geq 1}\left(\lambda_j(\Sigma_{n,p}) + \gamma_n\right)^{-1/2}\left(\lambda_j(\Sigma_p)^{1/2}\mathbb{E}_n(f_j(Z))\right)e_j(\cdot)\right\|_{\mathcal{H}}^2 \\ &= n\sum_{j\geq 1}\left(\frac{\lambda_j(\Sigma_p)}{\lambda_j(\Sigma_{n,p}) + \gamma_n}\right)\left(\mathbb{E}_n(f_j(Z))\right)^2 \\ &\overset{(2)}{=} n\sum_{j\geq 1}\left(\frac{\lambda_j(\Sigma_p)}{\lambda_j(\Sigma_p) + \gamma_n}\right)\left(\mathbb{E}_n(f_j(Z))\right)^2 + o_p(1)\end{aligned} \tag{14}$$

To derive equality (1), simply note that

$$\begin{aligned}\langle\mu_{p,n}, e_j\rangle_{\mathcal{H}} &= \mathbb{E}_n(\varepsilon e_j(X)) \\ &= \lambda_j(\Sigma_p)^{1/2}\mathbb{E}_n(f_j(Z))\end{aligned}$$

Equality (2) comes directly from Lemma 6:

$$\left|\lambda_j(\Sigma_{n,p}) - \lambda_j(\Sigma_p)\right| = o_p(1)$$

and the continuous mapping theorem.

Under the null, we have

$$\mathbb{E}(f_l(Z)) = \lambda_l^{-1/2}\mathbb{E}(\varepsilon e_l(X)) = \lambda_l^{-1/2}\langle\mu_p, e_l\rangle_{\mathcal{H}} = 0$$

and

$$\begin{aligned}\mathbb{V}(f_l(Z)) &= \mathbb{E}(f_l^2(Z)) - \mathbb{E}(f_l(Z))^2 \\ &= \langle f_l, f_l\rangle_{L^2(\mathbb{P})} \\ &= 1\end{aligned}$$

Equation 14 indicates how our proposed test statistic "assigns" different weights to different orthogonal directions, represeted by $\{e_j\}_{j\geq 1}$. As long as $\gamma_n > 0$ is (strictly) positive constant, $\sum_{j\geq 1}\lambda_j(\Sigma_p)/(\lambda_j(\Sigma_p) + \gamma_n) < \infty$. For non-vanishing ratios, its value are all approximately to one, overcoming the weakness of ICM test statistics, i.e., downweighting the higher frequency directions. More importantly, γ is user chosen and researchers have the flexibility of controlling how sensitive the test statistic is against high frequency alternatives.

4. Asymptotic Results

We first discuss the scenario where the regularization parameter γ_n is fixed. Subsequently, we consider the case where γ_n approaches zero at a specific rate.

To establish the asymptotic theories, the following assumptions are also needed.

- **Assumption 1.3**: The kernel k is integral strictly positive definite (ISPD):

$$\int\int f(x)k(x,x')f(x')d\eta(x)d\eta(x') > 0, \quad \forall \|f\|_{L^2(\eta)} \neq 0$$

- **Assumption 1.4**: (i) The eigenvalues $\left\{\lambda_j\left(\Sigma_p\right)\right\}_{j\geq 1}$ satisfy

$$\sum_{j=1}^{\infty} \lambda_j^{1/2}\left(\Sigma_p\right) < \infty$$

. (ii) There are infinitely many strictly positive eigenvalues $\left\{\lambda_j\left(\Sigma_p\right)\right\}_{j\geq 1}$ of Σ_p.

4.1. Fixed Regularization Parameter

First, we discuss the case where $\gamma_n := \gamma > 0$ is fixed. Theorem 2, Theorem 3 and Theorem 4 provide the asymptotic results under the null, fixed alternatives, and local alternatives, respectively.

Theorem 2 (Null Limiting Distribution under Fixed Regularization). Assume Assumption 1.3 and Assumption 1.4 (i), and assume in addition that H_0 holds, the random variable (vector) ε and $g\left(\bar{\theta}\right)$ are both bounded in probability, and that $\gamma_n := \gamma > 0$. Then,

$$\hat{T}_n(\gamma) \xrightarrow{d} T_{\infty}\left(\Sigma_p, \gamma\right) := \sum_{j=1}^{\infty} \frac{\lambda_j\left(\Sigma_p\right)}{\lambda_j\left(\Sigma_p\right) + \gamma} W_j^2 \tag{15}$$

where $W_j, j \geq 1$ are independent standard normal variables.

Proof: See Section A.3. □

Note that for all $\gamma > 0$, the weights $\left(\lambda_j\left(\Sigma_p\right) + \gamma\right)^{-1} \lambda_j\left(\Sigma_p\right)$ are summable.

Theorem 3. Under the fixed alternatives, and assume Assumption 1.3. If $\gamma_n := \gamma$, then for any $t > 0$,

$$\mathbb{P}_{H_1}\left(\hat{T}_n(\gamma_n) > t\right) \to 1$$

Proof: See Section A.5 □

We now consider the limiting power of the proposed test statistic under a local alternative:

$$\mathbb{H}_{1,n} : Y_i - \mathcal{M}_{\theta_0}(X_i) = \varepsilon_i + n^{-\alpha/2} R(X_i) := \tilde{\varepsilon}_i$$

where $\mathbb{E}(\varepsilon K(X,\cdot)) = \mathbf{0}$, and $\mathbb{E}(R(X)k(X,\cdot)) \neq \mathbf{0}$.

Theorem 4. Under the local alternative $\mathbb{H}_{1,n}$, let

$$\eta := \mathbb{E}\big(R(X)k_p(X,\cdot)\big) \in \mathcal{R}\Big(\Sigma_p^{1/2}\Big)$$

If $\gamma_n = \gamma > 0$ is fixed and $\alpha = 1$, we have

$$\hat{T}_n(\gamma_n) \xrightarrow{d} T_\infty\big(\Sigma_p, \gamma\big) + D_1$$

where

$$D_1 := \left\| \big(\Sigma_p + \gamma I\big)^{-1/2} \eta \right\|_{\mathcal{H}}^2$$

.

Proof: See Section A.6. □

4.2. Vanishing Regularization Parameter

The null limiting distribution $\sum_{l=1}^{\infty} \lambda_l/(\lambda + \gamma_n) W_l^2$ can be interpreted as a χ_v^2 distribution with degrees of freedom $v = \sum_{l=1}^{\infty} \lambda_l/(\lambda + \gamma_n)$. It is well known that the distribution of a χ_v^2 with a large degree of freedom v can be approximated by a normal distribution with mean v and variance $2v$. As γ_n approaches zero, the degrees of freedom of the test statistic tend to infinity. Consequently, the limiting distribution of the test statistic should also approach a standard normal distribution, provided the statistic is appropriately studentized.

To achieve such studentization, define a quantity:

$$d_r(\Sigma, \gamma) := \left\{ \sum_{l=1}^{\infty} \left(\frac{\lambda_l}{\lambda_l + \gamma} \right)^r \right\}^{1/r} \tag{16}$$

where $\{\lambda_l\}_{l \geq 1}$ are eigenvalues associated with Σ.

Specifically, two quantities of such type is relevant to this paper:

$$d_1(\Sigma, \gamma) = \sum_{l=1}^{\infty} \frac{\lambda_l}{\lambda_l + \gamma}$$

and

$$d_2(\Sigma, \gamma) = \left\{ \sum_{l=1}^{\infty} \left(\frac{\lambda_l}{\lambda_l + \gamma} \right)^2 \right\}^{1/2}$$

The proposed test statistic is studentized as:

$$\hat{T}_n(\gamma_n) = \frac{n \left\| \left(\hat{\Sigma}_p + \gamma_n I \right)^{-1/2} \hat{\mu}_p \right\|_{\mathcal{H}}^2 - d_1\left(\hat{\Sigma}_p, \gamma_n \right)}{\sqrt{2} d_2\left(\hat{\Sigma}_p, \gamma_n \right)} \tag{17}$$

Theorem 5, Theorem 6 and Theorem 7 provide the asymptotic results under the null, fixed alternatives, and local alternatives, respectively.

Theorem 5 (Null limiting Distribution under Vanishing Regularization). Under the null, assume Assumption 1.3 and Assumption 1.4. Assume in addition that the regularization term γ_n satisfies

$$\gamma_n + d_2^{-1}\left(\Sigma_p, \gamma_n\right) d_1\left(\Sigma_p, \gamma_n\right) \gamma_n^{-1} n^{-1/2} \longrightarrow 0$$

Then,

$$\hat{T}_n(\gamma_n) \xrightarrow{d} \mathcal{N}(0, 1)$$

Proof: See Section A.4. □

Remark. Contrary to the case where $\gamma_n \equiv \gamma$, the limiting distribution does not depend on the kernel, nor on the sequence of regularization parameters $\{\gamma_n\}_{n \geq 1}$.

However, notice that $d_2^{-1}(\Sigma_W, \gamma_n) d_1(\Sigma_W, \gamma_n) \gamma_n^{-1} n^{-1/2} \to 0$ requires that $\{\gamma_n\}_{n \geq 1}$ goes to zero at a rate slower than $n^{-1/2}$. In addition, this condition also implies that the decay rate of γ_n is also affected by the kernel as well as the underlying distribution.

Theorem 6. Under the fixed alternatives, and assume Assumption 1.3. If $\gamma_n + d_2^{-1}\left(\Sigma_p, \gamma_n\right) d_1\left(\Sigma_p, \gamma_n\right) \gamma_n^{-1} n^{-1/2} \to 0$, then for any $t > 0$,

$$\mathbb{P}_{H_1}\left(\hat{T}_n(\gamma_n) > t \right) \to 1$$

Proof: See Section A.5 □

Recall under the local alterantive, we have

$$\mathbb{H}_{1,n} : Y_i - \mathcal{M}_{\theta_0}(X_i) = \varepsilon_i + n^{-\alpha/2} R(X_i) := \tilde{\varepsilon}_i$$

The next theorem states the limiting distribution of the proposed test statistic with vanishing regularization parameter under the local alternative.

Theorem 7. Under the local alternative $\mathbb{H}_{1,n}$, let

$$\eta := \mathbb{E}\big(R(X)k_p(X,\cdot)\big) \in \mathcal{R}\Big(\Sigma_p^{1/2}\Big)$$

if

$$\gamma_n + d_2^{-1}(\Sigma_p,\gamma_n)d_1(\Sigma_p,\gamma_n)\gamma_n^{-1}n^{-1/2} \to 0$$

and $\alpha = 1/2$, then we have

$$\hat{T}_n(\gamma_n) \xrightarrow{d} \mathcal{N}(0,1) + D_2$$

where

$$D_2 := \Delta\left\|\Sigma_p^{-1/2}\eta\right\|_{\mathcal{H}}^2$$

and $\frac{n^{1/2}}{d_2(\Sigma_p,\gamma_n)} \to \Delta < \infty$.

Proof: See Section A.6. □

Remark. When $\eta_p \notin \mathcal{R}\Big(\Sigma_p^{1/2}\Big)$, the asymptotic distribution of $\hat{T}_n(\gamma_n)$ for a vanishing γ_n is, in general, not well defined.

5. Critical Value via Multiplier Bootstrap for Fixed γ

When the regularization term $\gamma_n := \gamma$ is fixed, the corresponding test statistic $\hat{T}_n(\gamma)$ is non-pivotal. In this section, we propose a simple-to-use multiplier bootstrap procedure to approximate the null distribution. Its implementation is listed below:

1. Generate a sequence of i.i.d random variables $\{v_i\}_{i=1,\ldots,n}$ with mean zero and unit variance. Random variables with such properties could include,e.g, Rademacher random variable, standard normal or Bernoulli random variable with $\mathbb{P}(v = 1-\kappa) = \kappa/\sqrt{5}$ and $\mathbb{P}(v = \kappa) = 1 - \kappa/\sqrt{5}$, where $\kappa = \big(\sqrt{5}+1\big)/2$.

2. Compute

$$\Big(\hat{T}_n^*(\gamma)\Big)_b = \frac{n\left\|\Big(\hat{\Sigma}_p^* + \gamma I\Big)^{-1/2}\hat{\mu}_p^*\right\|_{\mathcal{H}}^2 - d_1\Big(\hat{\Sigma}_p^*,\gamma\Big)}{d_2\Big(\hat{\Sigma}_p^*,\gamma\Big)}$$

   where

$$\hat{\mu}_p^* = \widehat{\boldsymbol{G}^*}\boldsymbol{m}_n$$
$$\widehat{\boldsymbol{G}^*} = \big(v_1\hat{\varepsilon}_1 k_p(X_1,\cdot),\ldots,v_n\hat{\varepsilon}_n k_p(X_n,\cdot)\big)$$

   and $\hat{\Sigma}_p^*$ is generated using $\hat{\mu}_p^*$.

3. Repeat Steps 1 and 2 B times, and collect $\Big\{\Big(\hat{T}_n^*(\gamma)\Big)_b, b = 1,\ldots,B\Big\}$

4. Define a confidence level α, obtain the $(1-\alpha)$-th quantile of $\left\{\left(\hat{T}_n^*(\gamma)\right)_b, b = 1, ..., B\right\}$, $c_{n,\alpha}^*$
5. Reject the null if $\hat{T}_n(\gamma) > c_{n,\alpha}^*$, and fail to reject otherwise.

The multiplier bootstrapped procedure has several attractive properties. First, it does not require computing new parameter estimates at each bootstrap draw,reducing the computational intensity of the proposed procedure. Second, due to the employment of the projection, its implementation does not require using estimators that admit an asymptotic linear representation. These computational conveniences are important when the dimension is high.

The next theorem establishes the asymptotic validity of the proposed bootstrap procedure.

Theorem 8. Assme that $T(\gamma, \theta) < \infty$ for all $\theta \in \Theta$. Then, we have

$$\hat{T}_n^*(\gamma) \xrightarrow{d^*} T_\infty(\Sigma_p, \gamma)$$

with probability one under the bootstrap law. Here $\xrightarrow{d^*}$ denotes the weak convergence under the bootstrap law, i.e., conditional on the sample $[Z]_n = \{Z_i\}_{i=1,...,n}$

Proof: See Section A.7 □

6. Simulation Studies

In this section, we conduct a series of simulation studies to evaluate the finite sample performance of the proposed test statistic. We compare the proposed test statistic with the MMD-Gaussian, Bierens, and MMD-IMQ tests. The MMD-Gaussian test is based on the maximum mean discrepancy (MMD) statistic with a Gaussian kernel, while the MMD-IMQ test is based on the MMD statistic with the inverse multi-quadratic kernel. The Bierens test is based on the empirical process, but can be regarded as a MMD-Gaussian test with the kenerl parameter 1/2.

6.1. Comparision Test Statistics

The MMD test statistic is defined as (refer to the next section for a detailed discussion):

$$\begin{aligned} n\hat{T}_n &= n\|\hat{\mu}_p\|_{\mathcal{H}} \\ &= \frac{1}{n}\sum_{i,j} \hat{\varepsilon}_i k_p(x_i, x_j)\hat{\varepsilon}_j \end{aligned} \tag{18}$$

For the MMD-Gaussian, the original kernel (before projection) is defined as:

$$k(x, x') = \exp(-\delta \parallel x - x' \parallel_2^2)$$

where δ is the bandwidth parameter, $\| \cdot \|_2$ denotes the L_2 norm. Bierens' statistic can also be understood as an MMD-Gaussian test with the kernel parameter $\delta = 1/2$.

The MMD-IMQ test is based on the kernel:

$$k(x, x') = \left(1 + \| \ x - x' \ \|_2^2\right)^{-1.5}$$

The simulation studies are conducted after accounting for the estimation effects, as outlined in Equation 10.

6.2. Data Generating Processes

We consider the following data generating processes (DGPs), which are grouped into four categories: (1) Null DGPs, (2) Fixed Alternatives, (3) Frequency Alternatives, and (4) Local Alternatives.

The null DGPs are defined as:

$$Y = X\beta + \sigma u$$

where for all null DGPs, X is an $n \times p$ matrix, $\beta = \mathbf{1}_p$ is a p-dimensional vector of ones, $\sigma = 1$, and u is an $n \times 1$ vector of independent standard normal random variables. The specifications for different null DGPs are as follows:

- For DGP1, $X = (\mathbf{1}_n, X_1, X_2)$, where X_1 and X_2 are independent standard normal random variables.
- For DGP2, $X = (\mathbf{1}_n, X_1, X_2, X_3, X_4, X_5)$, where for $i = 1, 2, 3$, X_i are independent uniformly distributed random variables on $[0, 1]$, and for $i = 4, 5$, X_i are independent normal random variables with mean zero and standard deviation 1.
- For DGP3, $X = \left(\mathbf{1}_n, \tilde{X}\right)$, where $\tilde{X}$ is an $n \times 10$ matrix with columns $\{X_i\}_{i=1,...,10}$. For $i = 1, ..., 5$, X_i are independent uniformly distributed random variables on $[0, 1]$, and for $i = 6, ..., 10$, X_i are independent normal random variables with mean zero and standard deviation 1.
- For DGP4, $X = \left(\mathbf{1}_n, \tilde{X}\right)$, where $\tilde{X}$ is an $n \times 20$ matrix with columns $\{X_i\}_{i=1,...,20}$. For $i = 1, ..., 10$, X_i are independent uniformly distributed random variables on $[0, 1]$, and for $i = 11, ..., 20$, X_i are independent normal random variables with mean zero and standard deviation 1.

DGPs 5 to 8 are fixed alternatives, and are defined as:

$$Y = X\beta + D + \sigma u$$

where D is a $n \times 1$ vector of deviation specifications. The parameters β and the error terms u are the same as in the null DGPs. The specifications for different fixed alternatives are as follows:

- For DGP5, $X = (\mathbf{1}_n, X_1, X_2)$, where X_1 and X_2 are independent normal random variables with mean zeros and standard deviation equal to 10. $\sigma = 5$, and $D = \|\tilde{X}\|_2$, where $\tilde{X} = [X_1, X_2]$.
- For DGP6, $X = (\mathbf{1}_n, X_1, X_2, X_3, X_4, X_5)$, where for $i = 1, 2, 3$, X_i are independent uniformly distributed random variables on $[0, 10 \times i]$, and for $i = 4, 5$, X_i are independent normal random variables with mean zero and standard deviation $10 \times (i - 3)$. $\sigma = 6$, and $D = \|\tilde{X}\|_2$, where $\tilde{X} = [X_1, ..., X_5]$.

- For DGP7, $X = \left(\mathbf{1}_n, \tilde{X}\right)$, where $\tilde{X}$ is an $n \times 10$ matrix with columns $\{X_i\}_{i=1,\ldots,10}$. For $i = 1, \ldots, 5$, X_i are independent uniformly distributed random variables on $[0, 1 + 0.1 \times (i - 1)]$, and for $i = 6, \ldots, 10$, X_i are independent normal random variables with mean zero and standard deviation $1 + 0.1 \times (i - 5)$. D = $\|\tilde{X}\|_2$. $\sigma = 7$, and $D = \|\tilde{X}\|_2$.
- For DGP8, $X = \left(\mathbf{1}_n, \tilde{X}\right)$, where $\tilde{X}$ is an $n \times 20$ matrix with columns $\{X_i\}_{i=1,\ldots,20}$. For $i = 1, \ldots, 10$, X_i are independent uniformly distributed random variables on $[0, 1 + 0.1 \times (i - 1)]$, for $i = 11, \ldots, 15$, X_i are independent normal random variables with mean zero and standard deviation $1 + 0.1 \times (i - 11)$, and for $i = 16, \ldots, 20$, X_i are independent normal random variables with mean zero and standard deviation $1 + 0.1 \times (i - 16)$. $\sigma = 8$, $D = D_1 + D_2$, where $D_1 = \|\tilde{X}_{1:10}\|_2$ and $D_2 = \|\tilde{X}_{11:20}\|_2$, and $\tilde{X}_{1:10}$ and $\tilde{X}_{11:20}$ are the first and second half of the columns of $\tilde{X}$, respectively.

> *Remark.* DGPs 5-6 have relatively low dimension and small error term standard deviation, while DGPs 7-8 have higher dimension and larger error term standard deviation. We would expect that the first DGPs are relatively easier to detect than the latter DGPs.

Frequency alternatives are represented in DGPs 9-11, and are defined as:

$$Y = X\beta + S + \sigma u$$

where S is a $n \times 1$ vector of frequency deviation specifications. The parameters β and the error terms u are the same as in the null DGPs. $X = \left(\mathbf{1}_n, \tilde{X}\right)$, where $\tilde{X}$ is an $n \times 20$ matrix with columns $\{X_i\}_{i=1,\ldots,20}$. For $i = 1, \ldots, 10$, X_i are independent uniformly distributed random variables on $[0, 1]$, and for $i = 11, \ldots, 20$, X_i are independent normal random variables with mean zero and standard deviation 1. $\sigma = 1$, and $S = 5 \times \Pi_{i=1}^{20} \sin(b \times X_i)$. The specifications for different frequency alternatives are as follows:

- For DGP9, the parameter $b = 1$.
- For DGP10, the parameter $b = 2$.
- For DGP11, the parameter $b = 7$.

> *Remark.* DGPs 9-11 have the same dimension and error term standard deviation, but different frequency deviation specifications. We would expect that the first DGPs are relatively easier to detect than the latter DGPs.

Finally, the local alternatives are represented in DGPs 12-14, and are defined as:

$$Y = X\beta + n^{-1/2}R + \sigma u$$

where $n^{-1/2}R$ is a $n \times 1$ vector of deviation specifications. The parameters β and the error terms u are the same as in the null DGPs. The specifications for different fixed alternatives are as follows:

- For DGP12, $X = (\mathbf{1}_n, X_1, X_2)$, where X_1 and X_2 are independent normal random variables with mean zeros and standard deviation equal to 10. $\sigma = \sqrt{0.1 + X_1^2 + X_2^2}$, and $n^{-1/2}R = n^{-1/2}\ \|\tilde{X}\|_2$, where $\tilde{X} = [X_1, X_2]$.

- For DGP13, $X = (\mathbf{1}_n, X_1, X_2, X_3, X_4, X_5)$, where for $i = 1, 2, 3$, X_i are independent uniformly distributed random variables on $[0, 10 \times i]$, and for $i = 4, 5$, X_i are independent normal random variables with mean zero and standard deviation $10 \times (i - 3)$. $\sigma = \sqrt{0.1 + \sum_{i=1}^{3} X_i + \sum_{j=4}^{5} X_j^2}$, and $n^{-1/2} R = n^{-1/2} \ \|\tilde{X}\|_2$, where $\tilde{X} = [X_1, ..., X_5]$.
- For DGP14, $X = \left(\mathbf{1}_n, \tilde{X}\right)$, where $\tilde{X}$ is an $n \times 10$ matrix with columns $\{X_i\}_{i=1,...,10}$. For $i = 1, ..., 5$, X_i are independent uniformly distributed random variables on $[0, 1 + 0.1 \times (i - 1)]$, and for $i = 6, ..., 10$, X_i are independent normal random variables with mean zero and standard deviation $1 + 0.1 \times (i - 5)$. D = $\|\tilde{X}\|_2$. $\sigma = \sqrt{0.1 + \sum_{i=1}^{5} X_i + \sum_{j=6}^{10} X_j^2}$, and $n^{-1/2} R = n^{-1/2} \ \|\tilde{X}\|_2$.
- For DGP15, $X = \left(\mathbf{1}_n, \tilde{X}\right)$, where $\tilde{X}$ is an $n \times 20$ matrix with columns $\{X_i\}_{i=1,...,20}$. For $i = 1, ..., 10$, X_i are independent uniformly distributed random variables on $[0, 1 + 0.1 \times (i - 1)]$, for $i = 11, ..., 15$, X_i are independent normal random variables with mean zero and standard deviation $1 + 0.1 \times (i - 11)$, and for $i = 16, ..., 20$, X_i are independent normal random variables with mean zero and standard deviation $1 + 0.1 \times (i - 16)$. $\sigma = \sqrt{0.1 + \sum_{i=1}^{5} X_i + \sum_{j=6}^{20} X_j^2}$, and $n^{-1/2} R = n^{-1/2} \ \|\tilde{X}\|_2$.

  Remark. DGPs 12-13 have relatively small deviations, while DGPs 14-15 have larger deviations. We would expect that the first DGPs are relatively more diffcoult to detect than the latter DGPs.

6.3. Simulation Results

We choose the Gaussian kernel in our KFDA statistic with the bandwidth parameter $\delta = (10 \times \tilde{c})^{-1}$, where $\tilde{c}$ is caculated using the average value of the principle component values for the explanatory matrix X. The bandwidth parameter in MMD-Gaussian is set by the heuristic rule: $\delta = 1/(2\sigma^2)$, where $\sigma = \text{median}\|x_i - x_j\|_2 : i, j = 1, ..., n$.

The regularization parameter in the proposed test statistic is set at $\gamma = 0.5$.

Table 1 and Table 2 present the empirical size and power of the proposed test statistic, as well as the MMD-Gaussian, Bierens, and MMD-IMQ tests. The results are based on 200 and 400 sample size respectively. The empirical size and power are calculated based on 1000 replications. Within each replication, the bootstrap size is 500. The significance levels are set at 0.1, 0.05, and 0.01.

Table 1: Rejection Rates Compariation of KFDA, MMD-Gaussian, and Bierens with MMD-IMQ, $N = 200$

N=200	**KFDA,$\gamma_n = 0.5$**			**MMD-Gaussian**			**Bierens**			**MMD-IMQ**		
Size	**0.1**	**0.05**	**0.01**	**0.1**	**0.05**	**0.01**	**0.1**	**0.05**	**0.01**	**0.1**	**0.05**	**0.01**
DGP1	0.111	0.059	0.011	0.089	0.040	0.008	0.110	0.059	0.009	0.081	0.037	0.010
DGP2	0.123	0.068	0.014	0.109	0.057	0.010	0.095	0.050	0.009	0.057	0.015	0.002
DGP3	0.140	0.063	0.011	0.125	0.054	0.014	0.000	0.000	0.000	0.000	0.000	0.000
DGP4	0.256	0.06	0.001	0.183	0.086	0.016	0.000	0.000	0.000	0.000	0.000	0.000
Power	**0.1**	**0.05**	**0.01**	**0.1**	**0.05**	**0.01**	**0.1**	**0.05**	**0.01**	**0.1**	**0.05**	**0.01**
Fixed Alternatives												
DGP5	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	0.920	0.999	0.995	0.864
DGP6	1.000	1.000	1.000	1.000	1.000	1.000	0.000	0.000	0.000	0.000	0.000	0.000
DGP7	0.540	0.413	0.155	0.401	0.275	0.099	0.082	0.020	0.002	0.009	0.000	0.000
DGP8	1.000	1.000	0.995	0.370	0.241	0.072	0.000	0.000	0.000	0.000	0.000	0.000
Frequency Alternatives												
DGP9	0.483	0.202	0.005	0.152	0.069	0.008	0.001	0.000	0.000	0.001	0.000	0.000
DGP10	0.474	0.194	0.009	0.195	0.078	0.015	0.001	0.000	0.000	0.000	0.000	0.000
DGP11	0.473	0.191	0.012	0.184	0.078	0.010	0.003	0.000	0.000	0.000	0.000	0.000
Local Alternatives												
DGP12	0.222	0.136	0.039	0.214	0.130	0.039	0.219	0.138	0.038	0.208	0.124	0.031
DGP13	0.383	0.279	0.119	0.390	0.272	0.087	0.231	0.132	0.031	0.193	0.091	0.013
DGP14	0.990	0.979	0.891	0.999	0.962	0.859	0.000	0.000	0.000	0.000	0.000	0.000
DGP15	1.000	1.000	0.999	1.000	1.000	1.000	0.000	0.000	0.000	0.000	0.000	0.000

Table 2: Rejection Rates Compariation of KFDA, MMD-Gaussian, and Bierens with MMD-IMQ, $N = 400$

N=400	**KFDA,$\gamma_n = 0.5$**			**MMD-Gaussian**			**Bierens**			**MMD-IMQ**		
Size	**0.1**	**0.05**	**0.01**	**0.1**	**0.05**	**0.01**	**0.1**	**0.05**	**0.01**	**0.1**	**0.05**	**0.01**
DGP1	0.098	0.046	0.012	0.113	0.067	0.010	0.103	0.059	0.014	0.112	0.065	0.014
DGP2	0.099	0.048	0.012	0.105	0.046	0.011	0.101	0.056	0.009	0.074	0.021	0.001
DGP3	0.114	0.051	0.010	0.110	0.052	0.008	0.000	0.000	0.000	0.000	0.000	0.000
DGP4	0.136	0.045	0.006	0.137	0.066	0.011	0.000	0.000	0.000	0.000	0.000	0.000
Power	**0.1**	**0.05**	**0.01**	**0.1**	**0.05**	**0.01**	**0.1**	**0.05**	**0.01**	**0.1**	**0.05**	**0.01**
Fixed Alternatives												
DGP5	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
DGP6	1.000	1.000	1.000	1.000	1.000	1.000	0.000	0.000	0.000	0.000	0.000	0.000
DGP7	0.630	0.494	0.234	0.611	0.496	0.249	0.197	0.098	0.017	0.082	0.022	0.002
DGP8	1.000	1.000	1.000	0.466	0.352	0.147	0.000	0.000	0.000	0.000	0.000	0.000
Frequency Alternatives												
DGP9	0.384	0.183	0.024	0.105	0.058	0.009	0.000	0.000	0.000	0.000	0.000	0.000
DGP10	0.374	0.154	0.016	0.141	0.057	0.007	0.000	0.000	0.000	0.000	0.000	0.000
DGP11	0.339	0.143	0.010	0.113	0.048	0.007	0.000	0.000	0.000	0.000	0.000	0.000
Local Alternatives												
DGP12	0.222	0.130	0.028	0.195	0.126	0.037	0.215	0.138	0.038	0.200	0.129	0.044
DGP13	0.344	0.250	0.112	0.340	0.231	0.093	0.241	0.143	0.047	0.222	0.129	0.028
DGP14	0.975	0.958	0.887	0.987	0.973	0.891	0.000	0.000	0.000	0.000	0.000	0.000
DGP15	1.000	1.000	1.000	1.000	1.000	1.000	0.000	0.000	0.000	0.000	0.000	0.000

The proposed test statistic demonstrates reasonably accurate size control across most of the null DGPs with a small sample size ($N = 200$). The only exception is DGP4, where the statistic shows an inflated size. However, this issue quickly resolves as the sample size increases to $N = 400$. DGP4 seems to be the most challenging for all test statistics in the small sample case. The MMD-Gaussian test statistic also exhibits inflated size, while the other statistics experience under-size distortion.

The statistic also exhibits exceptional performance across all DGPs. Notably, it is particularly effective in detecting frequency alternatives, which are characterized by both high frequency and high dimensionality. Conventional methods, such as the Bierens test, are insensitive to high-frequency deviations, while local smoothing test statistics suffer from the curse of dimensionality.

DGPs 7-8 and 12-13 are characterized by small signal-to-noise ratios ρ:

$$\rho = \frac{\left\|\mu_p\right\|_{\mathcal{H}}^2}{s}$$

where $s^2 = 4V_Z\left(\mathbb{E}_{Z'}\left(\varepsilon k_p(x, x')\varepsilon'\right)\right)$.

MMD type statistics, under the alternative, are essentially non-degenerate V (or U) statistics. Thus, using standard V (or U) statistic theory, it can be shown that a small signal-to-noise ratio would lead to a low power of the test, see e.g., Muandet et al. (2020), Li & Song (2022). The proposed test statistic, however, is able to detect these small deviations with high power. Intuitively, this is because the construction of the Fisher discriminant ratio is built on the minimization of the within-class variance, which helps to lower the noise level and boost the signal-to-noise ratio.

We also investigate the finite sample performance of the proposed test statistic under different regularization parameters. We consider the following regularization parameters: $\gamma = 0.1, 0.5, 1$. The results are presented in Figure 3 and Figure 4.

When γ is small, the proposed test statistic exhibits a slower decay rate of eigenvalues $\lambda_i/(\lambda_i + \gamma)$, which amplifies the impact of higher frequency directions. This can lead to an oversize problem under the null hypothesis, as the magnitudes of oscillations in higher frequencies are large. Conversely, when γ is large, the test statistic has a faster decay rate, placing more weight on lower frequency directions. This can result in a loss of power under alternative hypotheses. Selecting the optimal γ is crucial for performance but is a non-trivial task and is beyond the scope of this paper.

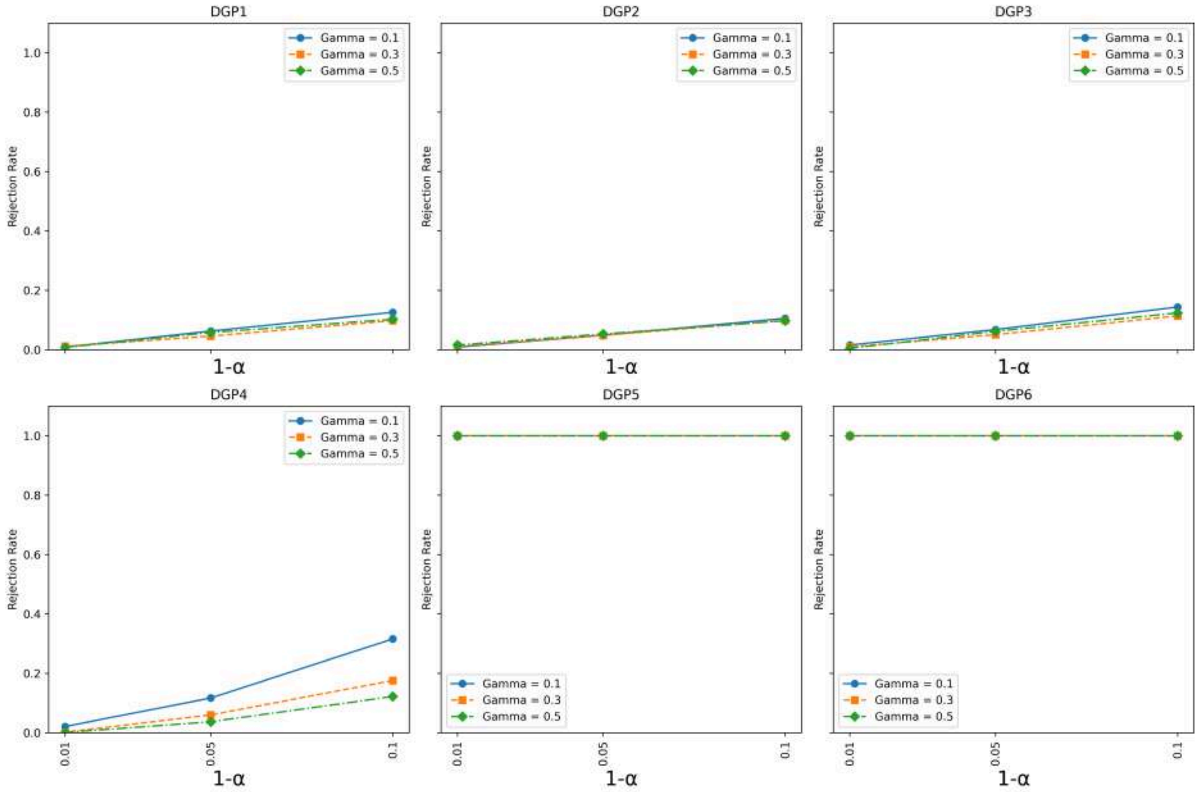


Figure 3: KFDA Rejection Rates under Different Regularization Parameters, DGP1-6

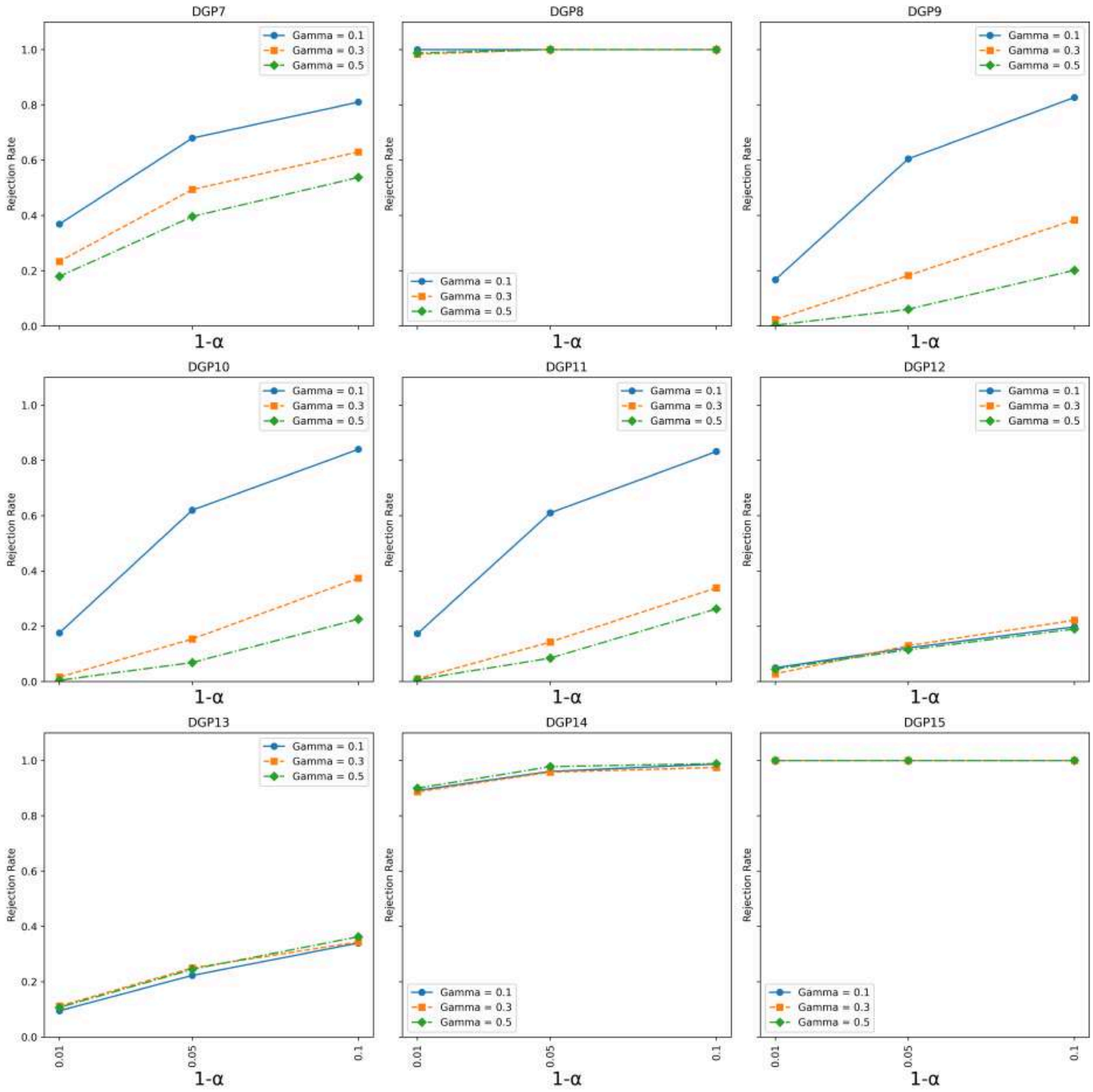


Figure 4: KFDA Rejection Rates under Different Regularization Parameters, DGP7-15

7. Discussion

In this section, we will discuss how our proposed test statistic is related to those found in the existing literature. These discussions will provide further insight into the ways in which our statistic has been enhanced.

7.1. ICM based Test Statistics

Most of the ICM based test statistics are based on an V (U)-statistic. For example, the ICM test statistic of Bierens (1982) can be written as

$$n\hat{T}_n = \frac{1}{n}\sum_{i,j}\hat{\varepsilon}_i \exp\left(-\frac{\|x_i - x_j\|_2^2}{2}\right)\hat{\varepsilon}_j$$

Let's consider the optimization problem of the numerator part of the KFD ratio, i.e.,

$$\hat{T}_n = \max_{\|f\|_{\mathcal{H}} \leq 1} \left\langle f, \hat{\Sigma}_B f \right\rangle_{\mathcal{H}}$$

it is easy to show that the solution of this problem is

$$f = \frac{\hat{\mu}_p}{\left\|\hat{\mu}_p\right\|_{\mathcal{H}}}$$

Hence

$$\begin{aligned} \hat{T}_n &= \frac{\left\langle \hat{\mu}_p, \hat{\Sigma}_B \hat{\mu}_p \right\rangle_{\mathcal{H}}}{\left\|\hat{\mu}_p\right\|_{\mathcal{H}}^2} \\ &= \left\|\hat{\mu}_p\right\|_{\mathcal{H}}^2 \\ &= \frac{1}{n^2} \sum_{i,j} \hat{\varepsilon}_i k_p(x_i, x_j) \hat{\varepsilon}_j \end{aligned} \tag{19}$$

Equation 19 implies that some IRF based test statistics can be expressed using a between class covariance operator in a RKHS. A spectral analysis of Equation 19 reveals that for ICM statistics,

$$\begin{aligned} \left\|\hat{\mu}_p\right\|_{\mathcal{H}}^2 &= \left\| \sum_{j \geq 1} \left\langle \hat{\mu}_p, e_j \right\rangle_{\mathcal{H}} e_j(\cdot) \right\|_{\mathcal{H}}^2 \\ &= \left\| \sum_{j \geq 1} \lambda_j(\Sigma_p)^{1/2} \mathbb{E}_n(f_j(Z)) e_j(\cdot) \right\|_{\mathcal{H}}^2 \\ &= \sum_{j \geq 1} \lambda_j(\Sigma_p) \left(\mathbb{E}_n(f_j(Z))\right)^2 \end{aligned}$$

where $e_j(\cdot)$, $f_j(\cdot)$ and $\lambda_j(\Sigma_p)$ are the same as in Equation 14. The eigenvalues $\left\{\lambda_j(\Sigma_p)\right\}_{j \geq 1}$ would decrease to zero, which explain why ICM statistics perform poorly when detecting high frequency deviations.

7.2. Maximum Mean Discrepancy based Test Statistics

Equation 19 is precisely the maximum mean discrepancy (MMD) based statistics for regression models. MMD statistics are first introduced for the nonparametric two sample test, see Gretton et al. (2012a), Gretton et al. (2012b). Similar concept is then developped for regression model checks, see Muandet et al. (2020) and Li & Song (2022). Thus, MMD statistics can be understood as special cases of the ICM statistic.

MMD based statistics are insensitive to the dimension of the covariate, provided this dimension is fixed. For example, Muandet et al. (2020) have shown that for any $0 < \delta < 1$, with probability at least $1 - \delta$, one has

$$\left\|\hat{\mu}_p - \mu_p\right\|_{\mathcal{H}} \leq \frac{2C_\theta \log\left(\frac{2}{\delta}\right)}{n} + \sqrt{\frac{2\sigma_\theta^2 \log\left(\frac{2}{\delta}\right)}{n}}$$

where C_θ is constant number such that $\left\|\varepsilon_\theta k_p(X, \cdot)\right\|_{\mathcal{H}} < C_\theta < \infty$ almost surely, and $\sigma_\theta^2 := \mathbb{E}\left(\left\|\varepsilon_\theta k_p(X, \cdot)\right\|_{\mathcal{H}}^2\right)$.

This non-asymptotic upper bound states that $\hat{\mu}_p$ converges at a rate of $n^{-1/2}$ that is independent of the dimension of X. This property is appealing because inferences based on $\hat{\mu}_p$ (including MMD and our statistic) become less susceptible to the *curse of dimensionality*.

7.3. The Local Smooth Test Statistics

The local smooth test proposed by Zheng (1996) is constructed as:

$$\hat{T}_n = \frac{1}{n^2} \sum_{i,j} \frac{1}{h^d} \hat{\varepsilon}_i k\left(\frac{x_i - x_j}{h}\right) \hat{\varepsilon}_j$$

where h is the bandwidth of the kernel k. To connect our proposed test statistic to the local smooth test, we consider the Gaussian kernel. As $h \to 0$, the local smooth kernel matrix $k\left(\frac{x_i - x_j}{h}\right)$ converges to an identity matrix, where its eigenvalues are essentially the same for different frequency directions. A similar phenomenon can be observed in the proposed test statistic when $\gamma_n \to 0$. In this spectral sense, we believe that the proposed test statistic bridges the gap between the local smooth test and the ICM test statistics.

7.4. The Pivotal Property

Local smoothing test statistics exhibit a normal null distribution, whereas the null distribution for ICM statistics is notably more intricate. When a projection, such as the one presented in this paper, is utilized, the null distribution takes the form of a linear combination of χ^2 distributions weighted by corresponding eigenvalues.

Recent literature has focused on establishing a pivotal property for ICM statistics. Raiola (2024) partition the covariate space into disjoint sub-cells to construct χ^2 statistics within each cell. This partitioning approach was first developed by Delgado & Vainora (2022) in the application of testing conditional distribution models.

Jiang & Tsyawo (2024) propose the use of a modified residual term in a generalized martingale difference divergence (GMDD) metric, ensuring that the test statistic is first-order non-degenerate and thus guarantees the pivotal property.

Our approach to achieving pivotality differs from existing methods. Firstly, under the null hypothesis, our statistic converges to a standard normal distribution, whereas others converge to a χ^2 distribution. Secondly, the method proposed by Raiola (2024) is not omnibus due to the partition of the covariate space, whereas ours is omnibus. While the approach suggested by Jiang & Tsyawo (2024) is omnibus, its power properties against high-frequency deviations are not clear. In contrast, our statistic has a theoretical basis for demonstrating good performance against high-frequency alternatives.

7.5. Kernel Fisher Discriminant Analysis in Testing

The primary goal of a Fisher discriminant analysis (FDA) (Fisher, 1936) is to find a subspace (or sub-minifold) which separates the classes as much as possible while the data also become as spread as possible. The kernel FDA (Mika et al., 1999) performs this goal in a reproducing kernel Hilbert space. Kernel FDA has found wide-ranging applications in machine learning practices,, as detailed in the review by Ghojogh et al. (2019).

The utilization of kernel FDA in hypothesis testing is relatively limited. To the best of the authors' knowledge, Harchaoui et al. (2007) were the first to propose the use of such a framework in the context of a nonparametric two-sample testing problem. Balasubramanian et al. (2021) also employ a similar idea to explore the optimality of kernel-embedding based test statistics.

8. Conclusion

This paper introduces a novel method for enhancing the goodness-of-fit testing of regression models through the application of Kernel Fisher Discriminant Analysis (KFDA). By leveraging the covariance structure of integrated regression functions, the proposed test statistic is designed to improve upon existing methodologies by modifying the weights associated to each component of the test statistic. This approach can be used to address a significant gap in the literature, particularly in scenarios where the dimensionality of the covariates is high and deviations from the null hypothesis occur in high-frequency directions.

The proposed test statistic is constructed to uniformly weight components associated with the leading eigenvalues of the covariance operator and downweight the remaining components. This strategy allows the test to gain greater power by concentrating on a user-tunable number of components. Moreover, under specific assumptions about the convergence speed of the regularization term, the test statistic becomes pivotal, providing a valuable property for practitioners.

The paper also presents asymptotic results for the proposed test statistic under both fixed and vanishing regularization parameters. Under a fixed regularization parameter, the test statistic achieves an $n^{-1/2}$ rate against local alternatives, making it capable of detecting subtle deviations from the null hypothesis. With a vanishing regularization parameter, the test statistic converges to a standard normal distribution, maintaining sensitivity to high-frequency deviations but exhibiting non-trivial power against local alternatives that converge to the null at a slower rate.

Another noteworthy aspect of the proposed methodology is its insensitivity to the dimension of the covariates, provided that this dimension is fixed. This property is achieved through the appropriate selection of reproducing kernels, making the test statistic robust in high-dimensional settings.

To facilitate practical implementation, the paper provides a multiplier bootstrap algorithm for finding critical values when the regularization parameter is fixed. Simulation results are presented to validate the theoretical findings and demonstrate the effectiveness of the proposed test statistic in various scenarios.

References

Andrews, D. W. (1997). A conditional Kolmogorov test. *Econometrica: Journal of the Econometric Society*, 1097–1128.

Balasubramanian, K., Li, T., & Yuan, M. (2021). On the optimality of kernel-embedding based goodness-of-fit tests. *Journal of Machine Learning Research*, *22*(1), 1–45.

Bickel, P. J., Ritov, Y., & Stoker, T. M. (2006). Tailor-made tests for goodness of fit to semiparametric hypotheses. *The Annals of Statistics*, *34*(2), 721–741.

Bierens, H. J. (1982). Consistent model specification tests. *Journal of Econometrics*, *20*(1), 105–134.

Bierens, H. J., & Ploberger, W. (1997). Asymptotic theory of integrated conditional moment tests. *Econometrica: Journal of the Econometric Society*, 1129–1151.

Delgado, M. A. (1993). Testing the equality of nonparametric regression curves. *Statistics & Probability Letters*, *17*(3), 199–204.

Delgado, M. A., & Vainora, J. (2022). Conditional Distribution Model Specification Testing Using Chi-Square Goodness-of-Fit Tests. *Arxiv Preprint Arxiv:2210.00624.*

Delgado, M. A., Dominguez, M. A., & Lavergne, P. (2006). Consistent tests of conditional moment restrictions. *Annales D'Économie Et De Statistique*, 33–67.

Dette, H. (1999). A consistent test for the functional form of a regression based on a difference of variance estimators. *The Annals of Statistics*, *27*(3), 1012–1040.

Escanciano, J. C. (2009). *Simple bootstrap tests for conditional moment restrictions.*

Escanciano, J. C., & Goh, S.-C. (2014). Specification analysis of linear quantile models. *Journal of Econometrics*, *178*, 495–507.

Fan, Y., & Li, Q. (2000). Consistent model specification tests: Kernel-based tests versus Bierens' ICM tests. *Econometric Theory*, *16*(6), 1016–1041.

Fisher, R. A. (1936). The use of multiple measurements in taxonomic problems. *Annals of Eugenics*, *7*(2), 179–188.

Ghojogh, B., Karray, F., & Crowley, M. (2019). Fisher and kernel Fisher discriminant analysis: Tutorial. *Arxiv Preprint Arxiv:1906.09436.*

González-Manteiga, W., & Crujeiras, R. M. (2013). An updated review of goodness-of-fit tests for regression models. *Test*, *22*, 361–411.

Gretton, A., Borgwardt, K. M., Rasch, M. J., Schölkopf, B., & Smola, A. (2012a). A kernel two-sample test. *The Journal of Machine Learning Research*, *13*(1), 723–773.

Gretton, A., Sejdinovic, D., Strathmann, H., Balakrishnan, S., Pontil, M., Fukumizu, K., & Sriperumbudur, B. K. (2012b). Optimal kernel choice for large-scale two-sample tests. *Advances in Neural Information Processing Systems*, *25.*

Harchaoui, Zaid, Eric, M., & Bach, F. (2007). Testing for homogeneity with kernel Fisher discriminant analysis. *Advances in Neural Information Processing Systems*, *20.*

Hardle, W., & Mammen, E. (1993). Comparing nonparametric versus parametric regression fits. *The Annals of Statistics*, 1926–1947.

Hsiao, C., Li, Q., & Racine, J. S. (2007). A consistent model specification test with mixed discrete and continuous data. *Journal of Econometrics*, *140*(2), 802–826.

Jiang, F., & Tsyawo, E. S. (2024). A Consistent ICM-based χ^2 Specification Test. *Arxiv Preprint Arxiv:2208.13370*.

Li, Q., & Wang, S. (1998). A simple consistent bootstrap test for a parametric regression function. *Journal of Econometrics*, *87*(1), 145–165.

Li, Y., & Song, X. (2022). *Consistent Test for Conditional Moment Restriction Models in Reproducing Kernel Hilbert Spaces*.

Mika, S., Ratsch, G., Weston, J., Scholkopf, B., & Mullers, K.-R. (1999). Fisher discriminant analysis with kernels. *Neural Networks for Signal Processing IX: Proceedings of the 1999 IEEE Signal Processing Society Workshop (Cat. No. 98th8468)*, 41–48.

Muandet, K., Fukumizu, K., Sriperumbudur, B., Schölkopf, B., & others. (2017). Kernel mean embedding of distributions: A review and beyond. *Foundations and Trends® in Machine Learning*, *10*(1–2), 1–141.

Muandet, K., Jitkrittum, W., & Kübler, J. (2020). Kernel conditional moment test via maximum moment restriction. *Conference on Uncertainty in Artificial Intelligence*, 41–50.

Raiola, A. (2024). *Testing Conditional Moment Restrictions: A Partitioning Approach*.

Sant'Anna, P. H., & Song, X. (2019). Specification tests for the propensity score. *Journal of Econometrics*, *210*(2), 379–404.

Stute, W. (1997). Nonparametric model checks for regression. *The Annals of Statistics*, 613–641.

Zheng, J. X. (1996). A consistent test of functional form via nonparametric estimation techniques. *Journal of Econometrics*, *75*(2), 263–289.

Appendix A. Proofs

A.1 Proof of Theorem 1

Proof: Note that

$$0 \leq \left\| (\Sigma + \gamma_n I)^{-1/2} \mu \right\|_{\mathcal{H}}^2 \leq \left\| (\Sigma + \gamma_n I)^{-1/2} \right\| \times \|\mu\|_{\mathcal{H}}$$

Thus, the problem can be reduced to showing that

$$\mathbb{E}(\varepsilon | X) = 0$$

if and only if

$$\|\mu\|_{\mathcal{H}}^2 = \mathbb{E}(\varepsilon k(X, X') \varepsilon') = 0$$

The "if" direction is relatively straightforward. For the "only if" direction, note that by the Mecer's theorem, we have

$$\mathbb{E}(\varepsilon k(X, X') \varepsilon') = \mathbb{E}\left(\varepsilon \sum_{j \geq 1} \xi_j \varphi_j(X) \varphi_j(X') \varepsilon' \right) = 0$$

where $\{\xi_j\}$ and $\{\varphi_j\}$ are eigenvalues and eigenfunctions of the following integral operator:

$$(Tf)(x) = \int f(x) k(x, s) d\eta(s)$$

Note that $\{\varphi_j(X)\}$ are also basis functions of the space of functions in $L^2(\eta)$ with respect to the measure η defined on the domain of X. Thus,

$$\mathbb{E}(\varepsilon | X) = \sum_{i \geq 1} \alpha_i \varphi_i(X)$$

In addition,

$$\mathbb{E}\left(\varepsilon \sum_{j \geq 1} \xi_j \varphi_j(X) \varphi_j(X') \varepsilon' \right) = 0$$

implies that ε is orthogonal to all these basis functions:

$$\mathbb{E}(\varepsilon \varphi_i(X)) = 0, \quad \forall i$$

To find the the coefficients α_i, we use the orthogonality conditions:

$$\mathbb{E}(\varepsilon \varphi_i(X)) = \mathbb{E}(\mathbb{E}(\varepsilon | X) \varphi_i(X)) = \mathbb{E}\left(\left(\sum_{j \geq 1} \alpha_j \varphi_j(X) \right) \varphi_i(X) \right) = \alpha_i \mathbb{E}(\varphi_i^2(X)) = 0$$

Since $\mathbb{E}(\varphi_i^2(X)) = 1$, we conclude

$$\alpha_i = 0, \quad \forall i$$

That completes the proof.

□

A.2 Proof of Lemma 1

Proof:

$$\begin{aligned}\left\|\left(\hat{\Sigma}_p+\gamma_n I\right)^{-1/2}\hat{\mu}_p\right\|_{\mathcal{H}} &= \left\langle\hat{\mu}_p,\left(\hat{\Sigma}_p+\gamma_n I\right)^{-1}\hat{\mu}_p\right\rangle_{\mathcal{H}}\\ &= \left\langle\hat{\mu}_p,\left(\left(\hat{\Sigma}_p+\gamma_n I\right)^{-1}-\left(\Sigma_{p,n}+\gamma_n I\right)^{-1}+\left(\Sigma_{p,n}+\gamma_n I\right)^{-1}\right)\hat{\mu}_p\right\rangle_{\mathcal{H}}\\ &= \underbrace{\left\langle\hat{\mu}_p,\left(\left(\hat{\Sigma}_p+\gamma_n I\right)^{-1}-\left(\Sigma_{p,n}+\gamma_n I\right)^{-1}\right)\hat{\mu}_p\right\rangle_{\mathcal{H}}}_{A}\\ &\quad+\underbrace{\left\langle\hat{\mu}_p,\left(\Sigma_{p,n}+\gamma_n I\right)^{-1}\hat{\mu}_p\right\rangle_{\mathcal{H}}}_{B}\end{aligned}$$

Let’s first analyze A:

$$\begin{aligned}A &\le \left\|\hat{\mu}_p\right\|_{\mathcal{H}}\left\|\left(\left(\hat{\Sigma}_p+\gamma_n I\right)^{-1}-\left(\Sigma_{p,n}+\gamma_n I\right)^{-1}\right)\hat{\mu}_p\right\|_{\mathcal{H}}\\ &\le \left\|\left(\hat{\Sigma}_p+\gamma_n I\right)^{-1}-\left(\Sigma_{p,n}+\gamma_n I\right)^{-1}\right\|\times\left\|\hat{\mu}_p\right\|_{\mathcal{H}}^2\end{aligned}$$

Let

$$\begin{aligned}\hat{C} &= \hat{\Sigma}_p+\gamma_n I\\ C_n &= \Sigma_{p,n}+\gamma_n I\end{aligned}$$

Note that

$$\left\|\hat{C}^{-1}-C_n^{-1}\right\| = \left\|C_n^{-1}\left(C_n-\hat{C}\right)\hat{C}^{-1}\right\| \le \left\|C_n^{-1}\right\|\times\left\|C_n-\hat{C}\right\|\times\left\|\hat{C}^{-1}\right\|$$

and

$$\begin{aligned}\left\|C_n^{-1}\right\| &\le \gamma_n^{-1}\\ \left\|\hat{C}^{-1}\right\| &\le \gamma_n^{-1}\end{aligned}$$

in addition, by Equation 11,

$$\left\|C_n-\hat{C}\right\| = O_p\left(n^{-1/2}\right)$$

Using Equation 10, we have

$$\begin{aligned}A &\le O_p\left(n^{-1/2}\right)\left\|\hat{\mu}_p\right\|_{\mathcal{H}}^2\\ &= O_p\left(n^{-1/2}\right)\left(\left\|\mu_{p,n}\right\|_{\mathcal{H}}^2+2O_p\left(n^{-1}\right)+O_p\left(n^{-2}\right)\right)\end{aligned}$$

We now analyze B:

$$\begin{aligned} B &= \left\| \left(\Sigma_{p,n} + \gamma_n I\right)^{-1/2} \left(\mu_{p,n} + O_p\left(n^{-1}\right)\right) \right\|_{\mathcal{H}}^2 \\ &= \left\| \left(\Sigma_{p,n} + \gamma_n I\right)^{-1/2} \mu_{p,n} \right\|_{\mathcal{H}}^2 \\ &\quad + 2O_p\left(n^{-1}\right) + O_p\left(n^{-2}\right) \end{aligned}$$

Putting everything together, we have

$$\begin{aligned} & n\left\| \left(\hat{\Sigma}_p + \gamma_n I\right)^{-1/2} \hat{\mu}_p \right\|_{\mathcal{H}}^2 \\ &= n\left\| \left(\Sigma_{p,n} + \gamma_n I\right)^{-1/2} \mu_{p,n} \right\|_{\mathcal{H}}^2 + 2O_p\left(n^{-1}\right) + O_p\left(n^{-2}\right) \\ &\quad + O_p\left(n^{-1/2}\right) \left\| \mu_{p,n} \right\|_{\mathcal{H}}^2 \\ &= n\left\| \left(\Sigma_{p,n} + \gamma_n I\right)^{-1/2} \mu_{p,n} \right\|_{\mathcal{H}}^2 + o_p(1) \end{aligned}$$

□

A.3 Proof of Theorem 2

Proof: By Equation 12, it suffices to analyze

$$n\left\| \left(\Sigma_{n,p} + \gamma I\right)^{-1/2} \mu_{p,n} \right\|_{\mathcal{H}}^2 = \sum_{j \geq 1} \left(\frac{\lambda_j\left(\Sigma_{n,p}\right)}{\lambda_j\left(\Sigma_{n,p}\right) + \gamma} \right) \underbrace{\left(\sqrt{n}\mathbb{E}_n\left(f_j(Z)\right)\right)^2}_{\substack{\text{degenearte V-statistic} \\ \text{under the null}}}$$

where the equality comes from Equation 14.

Note that under the null, by the central limit theory

$$\left(\sqrt{n}\mathbb{E}_n\left(f_j(Z)\right)\right)^2 \xrightarrow{p} W_j^2$$

where W_j is a standard normal distributed random variable. By Lemma 7, we have

$$\left|\lambda_j\left(\Sigma_{n,p}\right) - \lambda_j\left(\Sigma_p\right)\right| = O_p\left(n^{-1/2}\right), \forall j \geq 1 \tag{20}$$

and finally, by the continuous mapping theorem,

$$\frac{\lambda_j\left(\Sigma_{n,p}\right)}{\lambda_j\left(\Sigma_{n,p}\right) + \gamma} \xrightarrow{p} \frac{\lambda_j\left(\Sigma_p\right)}{\lambda_j\left(\Sigma_p\right) + \gamma}$$

Putting everything together, we have proved what have been claimed.

□

A.4 Proof of Theorem 5

Proof: First, we analyze the asymptotic behavior of $d_r\left(\hat{\Sigma}_p, \gamma\right), r = 1, 2$.

Note that

$$\left\|\hat{\Sigma}_p - \Sigma_{p,n}\right\| = O_p\left(n^{-1/2}\right)$$

Using Lemma 6, and the assumptions that $\varepsilon = O_p(1)$, $g\left(\bar{\theta}\right) = O_p(1)$ and $|k|_\infty < \infty$, we have

$$\left|\lambda_j\left(\hat{\Sigma}_p\right) - \lambda_j\left(\Sigma_{p,n}\right)\right| \leq \left\|\left(\hat{\Sigma}_p - \Sigma_{p,n}\right)e_j\right\|_{\mathcal{H}} = O_p\left(n^{-1/2}\right), \forall j \geq 1$$

Using the continuous mapping theorem again, we have

$$\left|d_r\left(\hat{\Sigma}_p, \gamma\right) - d_r\left(\Sigma_{p,n}, \gamma\right)\right| \xrightarrow{p} 0, r = 1, 2$$

Using Lemma 8, we are able to prove that

$$\left|d_2\left(\Sigma_{p,n}, \gamma\right) - d_2\left(\Sigma_p, \gamma\right)\right| = O_p\left(n^{-1/2}\right)$$

Similarly, using Equation 23 in Lemma 3, we are able to prove that

$$\left|d_1\left(\Sigma_{p,n}, \gamma\right) - d_1\left(\Sigma_p, \gamma\right)\right| \xrightarrow{p} 0$$

Finally, using the triangle inequality

$$\begin{aligned}\left|d_1\left(\hat{\Sigma}_p, \gamma\right) - d_1\left(\Sigma_p, \gamma\right)\right| &\leq \left|d_1\left(\hat{\Sigma}_p, \gamma\right) - d_1\left(\Sigma_{p,n}, \gamma\right)\right| + \left|d_1\left(\Sigma_{p,n}, \gamma\right) - d_1\left(\Sigma_p, \gamma\right)\right| \\ &\xrightarrow{p} 0\end{aligned}$$

Next, by Equation 12, it suffices to analyze

$$n\left\|\left(\Sigma_{n,p} + \gamma_n I\right)^{-1/2}\mu_{p,n}\right\|_{\mathcal{H}}^2$$

We first will show that

$$\begin{aligned}n\left\|\left(\Sigma_{n,p} + \gamma_n I\right)^{-1/2}\mu_{p,n}\right\|_{\mathcal{H}}^2 = n&\left\|\left(\Sigma_p + \gamma_n I\right)^{-1/2}\mu_{p,n}\right\|_{\mathcal{H}}^2 \\ &+ O_p\left(d_1\left(\Sigma_p, \gamma_n\right)\gamma_n^{-1} n^{-1/2}\right)\end{aligned}$$

This asymptotic approximation leaves $\mu_{p,n}$ as the only stochastic term. We then use the Berry-Esseen inequality to show the asymptotic normal distribution.

Using a similar argument in the proof of Lemma 4, we can show that

$$\left|\left\|\left(\Sigma_{p,n} + \gamma_n I\right)^{-1/2}\mu_{p,n}\right\|_{\mathcal{H}}^2 - \left\|\left(\Sigma_p + \gamma_n I\right)^{-1/2}\mu_{p,n}\right\|_{\mathcal{H}}^2\right| \leq C_1 C_2 D$$

where

$$
\begin{aligned}
C_1 &= \left\| \left(\Sigma_{p,n} + \gamma_n I\right)^{-1/2} \mu_{p,n} \right\|_{\mathcal{H}} \\
C_2 &= \left\| \left(\Sigma_p + \gamma_n I\right)^{-1/2} \mu_{p,n} \right\|_{\mathcal{H}} \\
D &= \left\| \left(\Sigma_{p,n} + \gamma_n I\right)^{-1/2} \left(\Sigma_{p,n} - \Sigma_p\right) \left(\Sigma_p + \gamma_n I\right)^{-1/2} \right\|
\end{aligned}
$$

As before, since $\left\| \left(\Sigma_{p,n} + \gamma_n I\right)^{-1} \left(\Sigma_p + \gamma_n I\right) \right\| = 1 + o_p(1)$, it is suffices to focus on C_2^2.

Note that,

$$
\mathbb{E}\left\| \left(\Sigma_p + \gamma_n I\right)^{-1/2} \mu_{p,n} \right\|_{\mathcal{H}}^2 = \operatorname{Tr}\left\{ \left(\Sigma_p + \gamma_n I\right)^{-1} \mathbb{E}\left(\mu_{p,n} \otimes \mu_{p,n}\right) \right\}
$$

where, by Lemma 5

$$
\begin{aligned}
&\mathbb{E}\left(\mu_{p,n} \otimes \mu_{p,n}\right) \\
&= \frac{1}{n} \mathbb{E}\left(\varepsilon^2 k_p(X, \cdot) \otimes k_p(X, \cdot)\right) \\
&\quad \underbrace{- \frac{1}{n} \mathbb{E}\left(\varepsilon k_p(X, \cdot)\right) \otimes \mathbb{E}\left(\varepsilon k_p(X, \cdot)\right) + \mathbb{E}\left(\varepsilon k_p(X, \cdot)\right) \otimes \mathbb{E}\left(\varepsilon k_p(X, \cdot)\right)}_{=0 \text{ under the null}}
\end{aligned}
$$

Thus,

$$
\operatorname{Tr}\left\{ \left(\Sigma_p + \gamma_n I\right)^{-1} \mathbb{E}\left(\mu_{p,n} \otimes \mu_{p,n}\right) \right\} \overset{\text{null}}{=} n^{-1} \operatorname{Tr}\left\{ \left(\Sigma_p + \gamma_n I\right)^{-1} \Sigma_p \right\}
$$

and,

$$
\mathbb{E}\left\| \left(\Sigma_p + \gamma_n I\right)^{-1/2} \mu_{p,n} \right\|_{\mathcal{H}}^2 \overset{\text{null}}{=} n^{-1} \operatorname{Tr}\left\{ \left(\Sigma_p + \gamma_n I\right)^{-1} \Sigma_p \right\} = n^{-1} d_1\left(\Sigma_p, \gamma_n\right)
$$

As for D, we notice that $\left\| \left(\Sigma_{p,n} + \gamma_n\right)^{-1/2} \right\| \leq \gamma_n^{-1/2}$ and $\left\| \left(\Sigma_p + \gamma_n\right)^{-1/2} \right\| \leq \gamma_n^{-1/2}$.

In addtion, by Lemma 8, we have

$$
\left\| \Sigma_{p,n} - \Sigma_p \right\|_{\text{HS}} = O_p\left(n^{-1/2}\right)
$$

Putting everything together, we have

$$
n \left| \left\| \left(\Sigma_{p,n} + \gamma_n I\right)^{-1/2} \mu_{p,n} \right\|_{\mathcal{H}}^2 - \left\| \left(\Sigma_p + \gamma_n I\right)^{-1/2} \mu_{p,n} \right\|_{\mathcal{H}}^2 \right| = O_p\left(d_1\left(\Sigma_p, \gamma_n\right) \gamma_n^{-1} n^{-1/2}\right)
$$

This asymptotic approximation allows us to focus on $n \left\| \left(\Sigma_p + \gamma_n I\right)^{-1/2} \mu_{p,n} \right\|_{\mathcal{H}}^2$ and apply the Berry-Esseen inequality to obtain the asymptotic normality result.

Using Equation 14, one can show that

$$n\left\|\left(\Sigma_p + \gamma_n I\right)^{-1/2}\mu_{p,n}\right\|_{\mathcal{H}}^2 = \sum_{j\geq 1}\left(\frac{\lambda_j\left(\Sigma_p\right)}{\lambda_j\left(\Sigma_p\right)+\gamma_n}\right)\left(\sqrt{n}\mathbb{E}_n\left(f_j(Z)\right)\right)^2 + o_p(1)$$

Let

$$S_{m,n} = \sum_{j=1}^{m}\left(\frac{\lambda_j\left(\Sigma_p\right)}{\lambda_j\left(\Sigma_p\right)+\gamma_n}\right)\left(\sqrt{n}\mathbb{E}_n\left(f_j(Z)\right)\right)^2$$

be a partial sum and notice that under the null,

$$S_{m,n} \xrightarrow{d} S_{m,\infty} = \sum_{j=1}^{m}\left(\frac{\lambda_j\left(\Sigma_p\right)}{\lambda_j\left(\Sigma_p\right)+\gamma_n}\right)W_j^2$$

where $\left\{W_j\right\}_j$ are i.i.d standard normal distributed random variables.

Let

$$Y_j = \left(\frac{\lambda_j\left(\Sigma_p\right)}{\lambda_j\left(\Sigma_p\right)+\gamma_n}\right)\left(W_j^2 - 1\right)$$

It is easy to check that

$$\mathbb{E}\left(Y_j\right) = 0$$

$$\sigma_j^2 := \mathrm{Var}\left(Y_j\right) = 2\left(\frac{\lambda_j\left(\Sigma_p\right)}{\lambda_j\left(\Sigma_p\right)+\gamma_n}\right)^2$$

Let

$$\bar{S}_{m,\infty} = \frac{\sum_{j=1}^{m} Y_j}{\sqrt{\sum_{j=1}^{m}\sigma_j^2}}$$

By the Berry-Esseen theorem, we have

$$\sup_{t\in\mathbb{R}}\left|\mathbb{P}\left(\bar{S}_{m,\infty} < t\right) - \Phi(t)\right| \leq C\frac{\sum_{j=1}^{m}\rho_j}{\left(\sum_{j=1}^{m}\sigma_j^2\right)^{3/2}}$$

where $\rho_j = \mathbb{E}\left(\left|Y_j\right|^3\right)$, and $\Phi(\cdot)$ is the cumulative distribution fucntion of the standard normal.

Remark.

1. $\sqrt{\sum_{j=1}^{\infty}\sigma_j^2} = \sqrt{2}d_2\left(\Sigma_p, \gamma_n\right)$
2. The rate of convergence is determined by the ratio $\sum_{j=1}^{m}\rho_j / \left(\sum_{j=1}^{m}\sigma_j^2\right)^{3/2}$, which is $O\left(m^{-1/2}\right)$.

□

A.5 Proof of Theorem 3 and Theorem 6

Proof: When $\gamma_n := \gamma$ is fixed, by Equation 14, we have

$$\left\| \left(\Sigma_{p,n} + \gamma I\right)^{-1/2} \mu_{p,n} \right\|_{\mathcal{H}}^2 = \sum_{j \geq 1} \frac{\lambda_j\left(\Sigma_{p,n}\right)}{\lambda_j\left(\Sigma_{p,n}\right) + \gamma} \left(\mathbb{E}_n\left(f_j(Z)\right)\right)^2$$

Note that

$$0 < \sum_{j \geq 1} \frac{\lambda_j\left(\Sigma_{p,n}\right)}{\lambda_j\left(\Sigma_{p,n}\right) + \gamma} \left(\mathbb{E}_n\left(f_j(Z)\right)\right)^2 < \sum_{j \geq 1} \left(\mathbb{E}_n\left(f_j(Z)\right)\right)^2 = 1$$

Thus,

$$n \left\| \left(\Sigma_{p,n} + \gamma I\right)^{-1/2} \mu_{p,n} \right\|_{\mathcal{H}}^2 \longrightarrow \infty$$

as $n \longrightarrow \infty$, but at the same time $d_r\left(\Sigma_p, \gamma\right) < \infty, r = 1, 2$.

We now discuss the situation where γ_n vanishes to zero. We first prove that

$$d_2\left(\Sigma_p, \gamma_n\right)^{-1} \left\| \left(\Sigma_{p,n} + \gamma_n I\right)^{-1/2} \mu_{p,n} \right\|_{\mathcal{H}}^2 \xrightarrow{p} d_2\left(\Sigma_p, \gamma_\infty\right)^{-1} \left\| \left(\Sigma_p + \gamma_\infty I\right)^{-1/2} \mu_p \right\|_{\mathcal{H}}^2$$

where $\gamma_\infty := \lim_{n \to \infty} \gamma_n = 0$. It is easy to show that

$$d_2\left(\Sigma_p, \gamma_n\right)^{-1} \left| \left\| \left(\Sigma_{p,n} + \gamma_n I\right)^{-1/2} \mu_{p,n} \right\|_{\mathcal{H}}^2 - \left\| \left(\Sigma_p + \gamma_\infty I\right)^{-1/2} \mu_{p,n} \right\|_{\mathcal{H}}^2 \right| \leq d_2\left(\Sigma_p, \gamma_n\right)^{-1} C_1$$

where

$$\begin{aligned} C_1 := & \left\| \left(\Sigma_p + \gamma_n I\right)^{-1/2} \mu_{p,n} \right\|_{\mathcal{H}} \left\| \left(\Sigma_{p,n} + \gamma_n I\right)^{-1/2} \mu_{p,n} \right\|_{\mathcal{H}} \\ & \times \left\| \left(\Sigma_{p,n} + \gamma_n I\right)^{-1/2} \left(\Sigma_{p,n} - \Sigma_p\right) \left(\Sigma_p + \gamma_n I\right)^{-1/2} \right\| \end{aligned}$$

Using the proof results of Theorem 5 (in Section A.4), we have

$$d_2\left(\Sigma_p, \gamma_n\right)^{-1} C_1 = n^{-1} d_2\left(\Sigma_p, \gamma_n\right)^{-1} O_p\left(d_1\left(\Sigma_p, \gamma_n\right) \gamma_n^{-1} n^{-1/2}\right) = o_p(1)$$

Thus, from now on, we will focus on investigating

$$\lim_{n \to \infty} \frac{n \left\| \left(\Sigma_p + \gamma_\infty I\right)^{-1/2} \mu_{p,n} \right\|_{\mathcal{H}}^2 - d_1\left(\Sigma_p, \gamma_\infty\right)}{d_2\left(\Sigma_p, \gamma_\infty\right)} \tag{21}$$

Note that

$$n \left\| \left(\Sigma_p + \gamma_\infty I\right)^{-1/2} \mu_p \right\|_{\mathcal{H}}^2 = \sum_{j \geq 1} \left(\sqrt{n} \mathbb{E}_n\left(f_j(Z)\right)\right)^2$$

where $f_j(\cdot)$ is defined in Equation 13. Under the fixed alterantive, we have

$$\begin{aligned}\mathbb{E}\big(f_j(Z)\big) &= C_j \\ \mathbb{V}\big(f_j(Z)\big) &= 1 - C_j^2\end{aligned}$$

Denote

$$\begin{aligned}S_{m,n} &= \sum_{j=1}^{m} \left(\mathbb{E}_n\big(f_j(Z)\big)\right)^2 \\ &= \sum_{j=1}^{m} \left(\mathbb{E}_n\big(f_j(Z) - C_j\big) + C_j\right)^2 \\ &= \sum_{j=1}^{m} \left(\mathbb{E}_n\left(\tilde{f}_j(Z)\right) + C_j\right)^2 \\ &= \sum_{j=1}^{m} \left(\mathbb{E}_n\left(\tilde{f}_j(Z)\right)\right)^2 + 2\sum_{j=1}^{m} C_j \mathbb{E}_n\left(\tilde{f}_j(Z)\right) + \sum_{j=1}^{m} C_j^2\end{aligned}$$

and,

$$nS_{m,n} = \sum_{j=1}^{m} \left(\sqrt{n}\mathbb{E}_n\left(\tilde{f}_j(Z)\right)\right)^2 + 2\sum_{j=1}^{m} C_j n\mathbb{E}_n\left(\tilde{f}_j(Z)\right) + n\sum_{j=1}^{m} C_j^2$$

Hence, Equation 21 is equivalent to

$$\lim_{n\to\infty}\lim_{m\to\infty} \frac{nS_{m,n} - m}{\sqrt{2m}}$$

Fix an m, then

$$\begin{aligned}\lim_{n\to\infty} \frac{nS_{m,n} - m}{\sqrt{2m}} \overset{d}{=} & \frac{\sum_{j=1}^{m}\big(1 - C_j^2\big)\big(W_j^2 - 1\big)}{\sqrt{2m}} - \sum_{j=1}^{m} \frac{C_j^2}{\sqrt{2m}} \\ & + \lim_{n\to\infty} \frac{2\sum_{j=1}^{m} C_j n\mathbb{E}_n\left(\tilde{f}_j(Z)\right)}{\sqrt{2m}} + \lim_{n\to\infty} n\frac{\sum_{j=1}^{m} C_j^2}{\sqrt{2m}}\end{aligned}$$

Let C_1 and C_2 be the lower bounds of $\left\{C_j\mathbb{E}_n\left(\tilde{f}_j(Z)\right)\right\}_j$ and $\left\{C_j^2\right\}_j$, respectively, Let C_3 and C_4 be the uppoer bounds of $\left\{C_j\right\}_j$ and $\left\{C_j^2\right\}_j$. Then

$$\begin{aligned}\lim_{n\to\infty} \frac{nS_{m,n} - m}{\sqrt{2m}} \overset{d}{\geq} & (1 - C_4)\frac{\sum_{j=1}^{m}\big(W_j^2 - 1\big)}{\sqrt{2m}} - C_3\frac{m}{\sqrt{2m}} \\ & + \lim_{n\to\infty} nC_1\sqrt{2m} + \lim_{n\to\infty} nC_2\frac{m}{\sqrt{2m}}\end{aligned}$$

As $m \to \infty$, the first term converges in distribution to $(1 - C_4)\mathcal{N}(0, 1)$, but the rest of the term goes to infinity in probability.

Hence,

$$\hat{T}_n(\gamma_n) \xrightarrow{p} \infty$$

□

A.6 Proof of Theorem 4 and Theorem 7

Proof: Denote

$$\begin{aligned}
&\hat{\tilde{\varepsilon}}_i = \hat{\varepsilon}_i + n^{-\alpha/2} R(X_i), \quad \tilde{\varepsilon}_i = \varepsilon_i + n^{-\alpha/2} R(X_i) \\
&\hat{\tilde{\mu}}_p = \mathbb{E}_n\left(\hat{\tilde{\varepsilon}} k_p(X, \cdot)\right), \quad \tilde{\mu}_{p,n} = \mathbb{E}_n\left(\tilde{\varepsilon} k_p(X, \cdot)\right) \\
&\hat{\tilde{\Sigma}}_p = \mathbb{E}_n\left(\hat{\tilde{\varepsilon}} k_p(X, \cdot) \otimes \hat{\tilde{\varepsilon}} k_p(X, \cdot)\right) \\
&\tilde{\Sigma}_{p,n} = \mathbb{E}_n\left(\tilde{\varepsilon} k_p(X, \cdot) \otimes \tilde{\varepsilon} k_p(X, \cdot)\right) \\
&\tilde{\Sigma}_p = \mathbb{E}\left(\tilde{\varepsilon} k_p(X, \cdot) \otimes \tilde{\varepsilon} k_p(X, \cdot)\right)
\end{aligned}$$

Using the similar arguments used in the proof of Lemma 1, we can show

$$\begin{aligned}
n\left\| \left(\hat{\tilde{\Sigma}}_p + \gamma_n I\right)^{-1/2} \hat{\tilde{\mu}}_p \right\|_{\mathcal{H}}^2 = n\left\| \left(\tilde{\Sigma}_{p,n} + \gamma_n I\right)^{-1/2} \tilde{\mu}_{p,n} \right\|_{\mathcal{H}}^2 \\
+2\underbrace{\left\langle \tilde{\mu}_{p,n}, \left(\tilde{\Sigma}_{p,n} + \gamma_n I\right)^{-1} \mathbb{E}_n\left(k_p(X, \cdot)\right)\right\rangle_{\mathcal{H}}}_{A} \\
+\underbrace{O_p\left(n^{1/2}\right) \left\| \tilde{\mu}_{p,n} \right\|_{\mathcal{H}}^2}_{B} + O_p\left(n^{-1/2}\right)
\end{aligned}$$

and note that

$$\begin{aligned}
A &= \left\langle \mu_{p,n} + n^{-\alpha/2} \underbrace{\mathbb{E}_n\left(R(X) k_p(X, \cdot)\right)}_{\eta_n(X, \cdot)}, \left(\tilde{\Sigma}_{p,n} + \gamma_n I\right)^{-1} \mathbb{E}_n\left(k_p(X, \cdot)\right) \right\rangle_{\mathcal{H}} \\
&= n^{-\alpha/2} \left\langle \eta_n, \left(\tilde{\Sigma}_{p,n} + \gamma_n I\right)^{-1} \mathbb{E}_n\left(k_p(X, \cdot)\right)\right\rangle_{\mathcal{H}} + o_p(1) \\
&= o_p(1)
\end{aligned}$$

and

$$\begin{aligned}
B &= O_p\left(n^{1/2}\right) \left\| \mu_{p,n} + n^{-\alpha/2} \eta_n \right\|_{\mathcal{H}}^2 \\
&= O_p\left(n^{1/2}\right) \left\| \mu_{p,n} \right\|_{\mathcal{H}}^2 + O_p\left(n^{1/2-\alpha/2}\right) \left\langle \mu_{p,n}, \eta_n \right\rangle_{\mathcal{H}} \\
&\quad + O_p\left(n^{1/2-\alpha}\right) \left\| \eta_n \right\|_{\mathcal{H}}^2 \\
&= o_p(1)
\end{aligned}$$

Thus,

$$n\left\| \left(\hat{\tilde{\Sigma}}_p + \gamma_n I\right)^{-1/2} \hat{\tilde{\mu}}_p \right\|_{\mathcal{H}}^2 = n\left\| \left(\tilde{\Sigma}_{p,n} + \gamma_n I\right)^{-1/2} \tilde{\mu}_{p,n} \right\|_{\mathcal{H}}^2 + o_p(1)$$

Next, we show that as long as $\alpha > 1/2$, we have

$$n\left\|\left(\tilde{\Sigma}_{p,n}+\gamma_n I\right)^{-1/2}\tilde{\mu}_{p,n}\right\|_{\mathcal{H}}^2 = n\left\|\left(\tilde{\Sigma}_{p}+\gamma_n I\right)^{-1/2}\tilde{\mu}_{p,n}\right\|_{\mathcal{H}}^2 + o_p(1)$$

The argument goes as follows,

$$\begin{aligned} n\left\|\left(\tilde{\Sigma}_{p,n}+\gamma_n I\right)^{-1/2}\tilde{\mu}_{p,n}\right\|_{\mathcal{H}}^2 &= n\left\langle\tilde{\mu}_{p,n}, \left(\tilde{\Sigma}_{p,n}+\gamma_n I\right)^{-1} - \left(\tilde{\Sigma}_{p}+\gamma_n I\right)^{-1}\tilde{\mu}_{p,n}\right\rangle_{\mathcal{H}} \\ &+n\left\|\left(\tilde{\Sigma}_{p}+\gamma_n I\right)^{-1/2}\tilde{\mu}_{p,n}\right\|_{\mathcal{H}}^2 \end{aligned}$$

and,

$$\begin{aligned} &n\left\langle\tilde{\mu}_{p,n}, \left(\tilde{\Sigma}_{p,n}+\gamma_n I\right)^{-1} - \left(\tilde{\Sigma}_{p}+\gamma_n I\right)^{-1}\tilde{\mu}_{p,n}\right\rangle_{\mathcal{H}} \\ &\quad\le n\left\|\left(\tilde{\Sigma}_{p,n}+\gamma_n I\right)^{-1}\right\| \times \left\|\tilde{\Sigma}_{p,n}-\tilde{\Sigma}_p\right\| \\ &\qquad\times \left\|\left(\tilde{\Sigma}_{p}+\gamma_n I\right)^{-1}\right\| \times \left\|\tilde{\mu}_{p,n}\right\|_{\mathcal{H}}^2 \\ &\quad\le \gamma_n^{-2} n O_p\left(n^{-1/2}\right)\Big(\left\|\mu_{p,n}\right\|_{\mathcal{H}}^2 + n^{-\alpha/2}\left\langle\mu_{p,n},\eta_n\right\rangle_{\mathcal{H}} \\ &\qquad+n^{-\alpha}\left\|\eta_n\right\|_{\mathcal{H}}^2\Big) \\ &\quad= \gamma_n^{-2}\left(\underbrace{O_p\left(n^{1/2}\right)\left\|\mu_{p,n}\right\|_{\mathcal{H}}^2}_{O_p(n^{-1/2})} + \underbrace{O_p\left(n^{1/2-\alpha/2}\right)\left\langle\mu_{p,n},\eta_n\right\rangle_{\mathcal{H}}}_{o_p(1)} + \underbrace{O_p\left(n^{1/2-\alpha}\right)\left\|\eta_n\right\|_{\mathcal{H}}^2}_{o_p(1)}\right) \end{aligned}$$

Using a similar argument, we can show

$$n\left\|\left(\tilde{\Sigma}_{p}+\gamma_n I\right)^{-1/2}\tilde{\mu}_{p,n}\right\|_{\mathcal{H}}^2 = n\left\|\left(\Sigma_{p}+\gamma_n I\right)^{-1/2}\tilde{\mu}_{p,n}\right\|_{\mathcal{H}}^2 + o_p(1)$$

Thus, it suffices to analyze

$$nd_2\left(\Sigma_p,\gamma_n\right)^{-1}\left(\left\|\left(\Sigma_{p}+\gamma_n I\right)^{-1/2}\tilde{\mu}_{p,n}\right\|_{\mathcal{H}}^2 - d_1\left(\Sigma_p,\gamma_n\right)\right)$$

Note that

$$\tilde{\mu}_{p,n} = \mu_{p,n} + n^{-\alpha/2}\eta_n$$

For a fixed regularization parameter, we have:

$$\begin{aligned} n\left\|\left(\Sigma_{p}+\gamma_n I\right)^{-1/2}\tilde{\mu}_{p,n}\right\|_{\mathcal{H}}^2 &= n\left\|\left(\Sigma_{p}+\gamma_n I\right)^{-1/2}\mu_{p,n}\right\|_{\mathcal{H}}^2 + 2nn^{-\alpha/2}\left\langle\mu_{p,n},\eta_n\right\rangle_{\mathcal{H}} \\ &+n^{1-\alpha}\left\|\left(\Sigma_{p}+\gamma_n I\right)^{-1/2}\eta_n\right\|_{\mathcal{H}}^2 \\ &= C_1 + C_2 + C_3 \end{aligned}$$

For vanishing regularization parameter, we have:

$$
\begin{aligned}
n\frac{\left\|\left(\Sigma_p+\gamma_n I\right)^{-1/2}\tilde{\mu}_{p,n}\right\|_{\mathcal{H}}^2-d_1\left(\Sigma_p,\gamma_n\right)}{d_2\left(\Sigma_p,\gamma_n\right)} &= n\frac{\left\|\left(\Sigma_p+\gamma_n I\right)^{-1/2}\mu_{p,n}\right\|_{\mathcal{H}}^2-d_1\left(\Sigma_p,\gamma_n\right)}{d_2\left(\Sigma_p,\gamma_n\right)} \\
&\quad+\frac{2nn^{-\alpha/2}\left\langle\mu_{p,n},\eta_n\right\rangle_{\mathcal{H}}}{d_2\left(\Sigma_p,\gamma_n\right)} \\
&\quad+\frac{n^{1-\alpha}\left\|\left(\Sigma_p+\gamma_n I\right)^{-1/2}\eta_n\right\|_{\mathcal{H}}^2}{d_2\left(\Sigma_p,\gamma_n\right)} \\
&= C_1+C_2+C_3
\end{aligned}
$$

Note that

$$
C_1 \xrightarrow{d} \begin{cases} T_\infty\left(\Sigma_p,\gamma\right) & \text{if } \gamma_n \text{ is fixed} \\ \mathcal{N}(0,1) & \text{if } \gamma_n + d_2^{-1}\left(\Sigma_p,\gamma_n\right)d_1\left(\Sigma_p,\gamma_n\right)\gamma_n^{-1}n^{-1/2} \to 0 \end{cases}
$$

and

$$
C_2 = o_p(1)
$$

When $\gamma_n := \gamma$ is fixed and $\alpha = 1$, we have

$$
C_3 \xrightarrow{p} \frac{\left\|\left(\Sigma_p+\gamma I\right)^{-1/2}\eta\right\|_{\mathcal{H}}^2}{d_2\left(\Sigma_p,\gamma\right)}
$$

When $\gamma_n \to 0$ with speed such that $\frac{n^{1-\alpha}}{d_2\left(\Sigma_p,\gamma_n\right)} \longrightarrow \Delta$, then

$$
C_3 \xrightarrow{p} \Delta\left\|\Sigma_p^{-1/2}\eta\right\|_{\mathcal{H}}^2 \begin{cases} < \infty & \text{if } \eta \in \mathcal{R}\left(\Sigma_p^{1/2}\right) \\ = \infty & \text{otherwise} \end{cases}
$$

□

A.7 Proof of Theorem 8

Proof: Notice that $\mathbb{E}\left(\hat{\mu}_p^*\right) = 0$. and using a similar argument to Lemma 1, it suffices to focus on

$$
n\left\|\left(\Sigma_{p,n}^*+\gamma I\right)^{-1/2}\mu_{p,n}^*\right\|_{\mathcal{H}}^2
$$

where

$$
\mu_{p,n}^* = \boldsymbol{G}_{p,n}^*\boldsymbol{m}_n
$$

and

$$
\boldsymbol{G}_{p,n}^* = \left(v_1\varepsilon_1 k_p(X_1,\cdot),...,v_n\varepsilon_n k_p(X_n,\cdot)\right)
$$

Further notice that, by Lemma 8

$$\left\|\Sigma^*_{p,n} - \Sigma^*_p\right\| = O_p\left(n^{-1/2}\right)$$

thus, using the same argument used in the proof of Lemma 1, we have

$$n\left\|\left(\Sigma^*_{p,n} + \gamma I\right)^{-1/2}\mu^*_{p,n}\right\|^2_{\mathcal{H}} = n\left\|\left(\Sigma^*_p + \gamma I\right)^{-1/2}\mu^*_{p,n}\right\|^2_{\mathcal{H}} + O_p\left(n^{-1/2}\right)$$

where $\left(\Sigma^*_p + \gamma I\right)^{-1/2}$ is deterministic and one only needs to focus on the asymptotic behaviors of

$$T_n = n\left\|\mu^*_{p,n}\right\|^2_{\mathcal{H}}$$

Let

$$\Sigma^*_{p,n} = \mathbb{E}_n\left(v\varepsilon k_p(X,\cdot) \otimes v\varepsilon k_p(X,\cdot)\right)$$

and $\{e_k\}_{k\geq 1}$ be the eigenvectors of $\Sigma^*_{p,n}$, while let $\left\{\lambda_k\left(\Sigma^*_{p,n}\right)\right\}_{k\geq 1}$ be the corresponding eigenvalues. Notice that

$$\begin{aligned} \lambda_k\left(\Sigma^*_{p,n}\right)\delta_{k,l} = \left\langle e_k, \Sigma^*_{p,n} e_k\right\rangle_{\mathcal{H}} &= \langle v\varepsilon e_k, v'\varepsilon' e_l\rangle_{L^2(\mathbb{P}_n)} \\ &= \lambda_k^{1/2}\left(\Sigma^*_{p,n}\right)\lambda_l^{1/2}\left(\Sigma^*_{p,n}\right)\langle v f_k, v' f_l\rangle_{L^2(\mathbb{P}_n)} \end{aligned}$$

where

$$f_k = \lambda_k^{-1/2}\left(\Sigma^*_{p,n}\right)\varepsilon e_k$$

Thus, $\{v f_k\}_{k\geq 1}$ is an orthonormal system of $L^2(\mathbb{P}_n)$, and

$$1 = \|v f_k\|^2_{L^2(\mathbb{P}_n)} = \frac{1}{n}\sum_{i=1}^n v_i^2 f_k^2(z_i) = \frac{1}{n}\sum_{i=1}^n v_i^2 \frac{1}{n}\sum_{i=1}^n f_k^2(z_i)$$

Thus,

$$\frac{1}{n}\sum_{i=1}^n f_k^2(z_i) \xrightarrow{p} 1$$

Hence,

$$\begin{aligned} T_n &= n\left\|\sum_{k\geq 1}\left\langle \mu^*_{p,n}, e_k\right\rangle_{\mathcal{H}} e_k(\cdot)\right\|^2_{\mathcal{H}} \\ &= n\left\|\sum_{k\geq 1}\mathbb{E}_n(v\varepsilon e_k(X))e_k(\cdot)\right\|^2_{\mathcal{H}} \\ &= \sum_{k\geq 1}\lambda_k\left(\Sigma^*_{p,n}\right)\left(\sqrt{n}\mathbb{E}_n(v f_k(Z))\right)^2 \end{aligned}$$

Note that for all $k \geq 1$

$$\mathbb{E}_{[Z]_n}(\mathbb{E}_n(vf_k(Z))) = 0$$

and,

$$\mathbb{V}_{[Z]_n}\left(\sqrt{n}\mathbb{E}_n(vf_k(Z))\right) = \frac{1}{n}\sum_{i=1}^{n} f_k^2(z_i) \xrightarrow{p} 1$$

By CLT and continuous mapping theorem, we have

$$\left(\sqrt{n}\mathbb{E}_n(vf_k(Z))\right)^2 \xrightarrow{d^*} W_k^2$$

where $W_k \sim \mathcal{N}(0,1)$.

We now show that

$$\lambda_k(\Sigma_{p,n}^*) \xrightarrow{p} \lambda_k(\Sigma_p)$$

Note that

$$\left\|\Sigma_{p,n}^* - \Sigma_p\right\| \leq \left\|\Sigma_{p,n}^* - \Sigma_{p,n}\right\| + \left\|\Sigma_{p,n} - \Sigma_p\right\|$$

where by Lemma 8, we already known that

$$\left\|\Sigma_{p,n} - \Sigma_p\right\| = O_p\left(n^{-1/2}\right)$$

Since

$$\begin{aligned}
\Sigma_{p,n}^* &= \mathbb{E}_n\left(v\varepsilon k_p(X,\cdot) \otimes v\varepsilon k_p(X,\cdot)\right) \\
&= \mathbb{E}_n\left(\mathbb{E}_{[Z]_n,n}(v^2)\varepsilon k_p(X,\cdot) \otimes \varepsilon k_p(X,\cdot)\right) \\
&\overset{(1)}{=} \mathbb{E}_n\left(\varepsilon k_p(X,\cdot) \otimes \varepsilon k_p(X,\cdot)\right) + O_p\left(n^{-1/2}\right) \\
&= \Sigma_{p,n} + O_p\left(n^{-1/2}\right)
\end{aligned}$$

where equality (1) arises from

$$\begin{aligned}
\mathbb{E}_{[Z]_n,n}(v^2) = \mathbb{E}_n(v^2) &= \mathbb{E}(v^2) + O_p\left(n^{-1/2}\right) \\
&= 1 + O_p\left(n^{-1/2}\right)
\end{aligned}$$

Thus,

$$\left\|\Sigma_{p,n}^* - \Sigma_{p,n}\right\| = O_p\left(n^{-1/2}\right)$$

Using Lemma 7, we can conclude

$$\lambda_k(\Sigma_{p,n}^*) \xrightarrow{p} \lambda_k(\Sigma_p), \forall k \geq 1$$

Putting everything together, we have

$$T_n \xrightarrow{d^*} Y = \sum_{k \geq 1} \lambda_k(\Sigma_p) W_k^2$$

□

Appendix B. Some Useful Lemmas

B.1 Perturbation Results on Covariance Operator

Lemma 2. Let A be a compact self-adjoint operator, with $\{\lambda_l\}_{l\geq 1}$ the eigenvalues of A and $\{e_l\}_{l\geq 1}$ an orthonormal system of eigenvectors of A. Then for all integer $k > 1$, using the convention $l_{k+1} = l_1$,

$$\sum_{l=1}^{\infty}\langle e_l, (AB)^k e_l\rangle = \sum_{l_1=1}^{\infty}\sum_{l_2=1}^{\infty}\dots\sum_{l_k=1}^{\infty}\left\{\left(\prod_{j=1}^{k}\lambda_{l_j}\right)\left(\prod_{j}^{k}\langle e_{l_j} Be_{l_{j+1}}\rangle\right)\right\}$$

Proof: Let k be some integer, fixed throughout the proof. The proof is by unduction, i.e., we shall prove that for all $l \in \{1, \dots, k\}$,

$$\begin{aligned}&\sum_{l=1}^{\infty}\langle e_l, (AB)^k e_l\rangle \\ &\quad = \sum_{l_1=1}^{\infty}\sum_{l_2=1}^{\infty}\dots\sum_{l_m=1}^{\infty}\left\{\left(\prod_{j=1}^{l-1}\lambda_{l_j}\right)\left(\prod_{j}^{l-1}\langle e_{l_j} Be_{l_{j+1}}\rangle\right)\langle e_{l_m}, (AB)^{k-l+1}e_{l_1}\rangle\right\}, \quad \mathcal{P}(1)\end{aligned}$$

First, for $l = 2$, using that $A^*e_{l_1} = Ae_{l_1} = \lambda_{l_1}e_{l_1}$, and $B^*e_{l_1} = \sum_{l_2=1}^{\infty}\langle e_{l_1}, Be_{l_2}\rangle e_{l_2}$, we have

$$\begin{aligned}\sum_{l_1=1}^{\infty}\left\langle e_{l_1}, AB(AB)^{k-1}e_{l_1}\right\rangle &= \sum_{l_1=1}^{\infty}\lambda_{l_1}\left\langle B^*e_{l_1}, (AB)^{k-1}e_{l_1}\right\rangle \\ &= \sum_{l_1=1}^{\infty}\lambda_{l_1}\left\langle\sum_{l_2=1}^{\infty}\langle e_{l_1}, Be_{l_2}\rangle e_{l_2}, (AB)^{k-1}e_{l_1}\right\rangle \\ &= \sum_{l_1=1}^{\infty}\sum_{l_2=1}^{\infty}\lambda_{l_1}\langle e_{l_1}, Be_{l_2}\rangle\left\langle e_{l_2}, (AB)^{k-1}e_{l_1}\right\rangle, \quad \mathcal{P}(2)\end{aligned}$$

Assume the statement $\mathcal{P}(1)$ is true with $l < k - 1$. Let us now marginalize out, first A then B in $(AB)^{k-l+1}$, for $(l + 1)$-th time, by summing over an index l_{m+1}. Using the same arguments as above, that is $A^*e_{l_m} = \lambda_{l_m}e_{l_m}$ and $B^*e_{l_m} = \sum_{l_{m+1}}^{\infty}\langle e_{l_m}, Be_{l_{m+1}}\rangle e_{l_{m+1}}$, we have

$$
\begin{aligned}
&\sum_{l=1}^{\infty}\langle e_l, (AB)^k e_l\rangle \\
&= \sum_{l_1=1}^{\infty}\cdots\sum_{l_m=1}^{\infty}\left\{\left(\prod_{j=1}^{l-1}\lambda_{l_j}\right)\left(\prod_{j}^{l-1}\langle e_{l_j}Be_{l_{j+1}}\rangle\right)\left\langle e_{l_m}, AB(AB)^{k-l}e_{l_1}\right\rangle\right\}, \quad \mathcal{P}(m) \\
&= \sum_{l_1=1}^{\infty}\cdots\sum_{l_m=1}^{\infty}\left\{\left(\prod_{j=1}^{l-1}\lambda_{l_j}\right)\lambda_{l_m}\left(\prod_{j}^{l-1}\langle e_{l_j}Be_{l_{j+1}}\rangle\right)\left\langle B^* e_{l_m}, (AB)^{k-l}e_{l_1}\right\rangle\right\} \\
&= \sum_{l_1=1}^{\infty}\cdots\sum_{l_m=1}^{\infty}\sum_{l_{m+1}=1}^{\infty}\left\{\left(\prod_{j=1}^{l}\lambda_{l_j}\right)\left(\prod_{j}^{l-1}\langle e_{l_j}Be_{l_{j+1}}\rangle\right)\left\langle e_{l_m}Be_{l_{m+1}}\right\rangle\left\langle e_{l_{m+1}}, (AB)^{k-l}e_{l_1}\right\rangle\right\}
\end{aligned}
$$

which proves $\mathcal{P}(m+1)$.

The proof is concluded by a k-induction. □

[3] An operator T on a Hilbert space $\mathcal{H}$ is said to be trace-class if it is compact (meaning it maps bounded sets to relatively compact sets, having the property of taking any bounded sequence to a sequence with a convergent subsequence) and the sum of its eigenvalues is finite. Formally, if T has a complete orthonormal set of eigenvectors $\{e_l\}_{l\geq 1}$ with corresponding eigenvalues $\{\lambda_l\}_{l\geq 1}$, then T is trace-class if and only if

$$\sum_{l=1}^{\infty}|\lambda_l| < \infty$$

The trace of this operator is defined as

$$\mathrm{Tr}(T) = \sum_{l=1}^{\infty}\lambda_l$$

[4] A trace-class perturbation operator refers to a situation in which a small or infinitesimal change to an operator is made, and this change itself is a trace-class operator. If the perturbation Δ (the change you introduce) is a trace-class operator, a few key points are relevant:

- **Stability**: Trace-class operators are compact, which implies they do not drastically change the overall structure of the spectrum of an operator. This stability can be important for understanding the qualitative behavior of solutions or eigenstates under perturbation.
- **Finite Trace**: The fact that the sum of the absolute values of the eigenvalues of Δ is finite ensures that the perturbation does not introduce unbounded energy or drastic changes that could make the system ill-behaved.
- **Spectral Theory**: In the context of spectral theory, trace-class perturbations can lead to explicit formulas for the change in eigenvalues or the resolvent of the operator, which is crucial for understanding the dynamics of the perturbed system.

Lemma 3. Let $\gamma > 0$ and S a trace-class operator[3]. Denote $\{\lambda_l\}_{l\geq 1}$ and $\{e_l\}_{l\geq 1}$ respectively the positive eigenvalues and the corresponding eigenvectors of S. Consider $d_r(T, \gamma)$ for $r = 1, 2$, with T a compact operator. If Δ is a trace-class perturbation operator[4] such that

$$\left\|(S + \gamma I)^{-1}\Delta\right\| < 1$$

and

$$\|\Delta\|_{\mathcal{C}_1} = \sum_{l=1}^{\infty}\|\Delta e_l\| < \gamma$$

then

$$|d_r(S + \Delta, \gamma) - d_r(S, \gamma)| \leq \frac{\gamma^{-1}\ \|\Delta\|_{\mathcal{C}_1}}{1 - \gamma^{-1}\ \|\Delta\|_{\mathcal{C}_1}}, \quad \text{for } r = 1, 2 \tag{22}$$

If $d_2(S, \gamma)\|S^{-1/2}\Delta S^{-1/2}\|_{\mathrm{HS}} < 1$, then

$$|d_1(S + \Delta, \gamma) - d_1(S, \gamma)| \leq \frac{d_2(S, \gamma)\left\|S^{-1/2}\Delta S^{-1/2}\right\|_{\mathrm{HS}}}{1 - d_2(S, \gamma)\left\|S^{-1/2}\Delta S^{-1/2}\right\|_{\mathrm{HS}}} \tag{23}$$

$$|d_2(S + \Delta, \gamma) - d_2(S, \gamma)| \leq \frac{\left\|S^{-1/2}\Delta S^{-1/2}\right\|_{\mathrm{HS}}}{1 - \left\|S^{-1/2}\Delta S^{-1/2}\right\|_{\mathrm{HS}}} \tag{24}$$

Proof: If $\|(S + \gamma I)^{-1}\Delta\| < 1$, then we may write

$$\begin{aligned}(S + \Delta + \gamma I)^{-1}(S + \Delta) &= \left(I + (S + \gamma I)^{-1}\Delta\right)^{-1}(S + \gamma I)^{-1}(S + \Delta) \\ &= \sum_{k=0}^{\infty}(-1)^k\{(S + \gamma I)^{-1}\Delta\}^k(S + \gamma I)^{-1}(S + \Delta) \\ &= (S + \gamma I)^{-1}S + \sum_{k=1}^{\infty}(-1)^k\{(S + \gamma I)^{-1}\Delta\}^k((S + \gamma I)^{-1}S - I)\end{aligned}$$

Note that the first equality holds true because:

$$\begin{aligned}&\left(I + (S + \gamma I)^{-1}\Delta\right)^{-1}(S + \gamma I)^{-1}(S + \Delta + \gamma I) \\ &= \left(I + (S + \gamma I)^{-1}\Delta\right)^{-1} + \left(I + (S + \gamma I)^{-1}\Delta\right)^{-1}(S + \gamma I)^{-1}\Delta\end{aligned}$$

Let

$$A = I + (S + \gamma I)^{-1}\Delta$$

and

$$B = (S + \gamma I)^{-1}\Delta$$

we have,

$$A - B = I$$

and by the Sherman-Morrison-Woodbury formula,

$$(A - B)^{-1} = A^{-1} + A^{-1}B(A - B)^{-1}$$

we have

$$\left(I + (S + \gamma I)^{-1}\Delta\right)^{-1} + \left(I + (S + \gamma I)^{-1}\Delta\right)^{-1}(S + \gamma I)^{-1}\Delta = I$$

Hence

$$(S + \Delta + \gamma I)^{-1} = \left(I + (S + \gamma I)^{-1}\Delta\right)^{-1}(S + \gamma I)^{-1}$$

The second equality comes from the Neuman Series,

$$(I - T)^{-1} = \sum_{k=0}^{\infty} T^k$$

and its variation:

$$(I + T)^{-1} = (I - (-T))^{-1} = \sum_{k=0}^{\infty} (-1)^k T^k$$

In our case, $T = (S + \gamma I)^{-1}\Delta$.

Finally, the third equality comes from:

$$\begin{aligned}
&\sum_{k=0}^{\infty} (-1)^k \left\{(S + \gamma I)^{-1}\Delta\right\}^k (S + \gamma I)^{-1}(S + \Delta) \\
=\ &\sum_{k=0}^{\infty} (-1)^k \left\{(S + \gamma I)^{-1}\Delta\right\}^k (S + \gamma I)^{-1}S + \sum_{k=0}^{\infty} (-1)^k \left\{(S + \gamma I)^{-1}\Delta\right\}^k (S + \gamma I)^{-1}\Delta \\
=\ &(S + \gamma I)^{-1}S + \sum_{k=1}^{\infty} (-1)^k \left\{(S + \gamma I)^{-1}\Delta\right\}^k (S + \gamma I)^{-1}S + \sum_{k=0}^{\infty} (-1)^k \left\{(S + \gamma I)^{-1}\Delta\right\}^{k+1} \\
=\ &(S + \gamma I)^{-1}S + \sum_{k=1}^{\infty} (-1)^k \left\{(S + \gamma I)^{-1}\Delta\right\}^k (S + \gamma I)^{-1}S + \sum_{k=1}^{\infty} (-1)^k \left\{(S + \gamma I)^{-1}\Delta\right\}^k (-I) \\
=\ &(S + \gamma I)^{-1}S + \sum_{k=1}^{\infty} (-1)^k \left\{(S + \gamma I)^{-1}\Delta\right\}^k \left((S + \gamma I)^{-1}S - I\right)
\end{aligned}$$

Since the trace is continuous in the space of trace-class operators, and using $\left\|(S + \gamma I)^{-1}S - I\right\| < 1$, we get, by linearity of the trace,

$$\begin{aligned}|d_1(S+\Delta,\gamma)-d_1(S,\gamma)| &= \left|\operatorname{Tr}\{(S+\Delta+\gamma I)^{-1}(S+\Delta)\}-\operatorname{Tr}\{(S+\gamma I)^{-1}(S)\}\right| \\ &= \sum_{k=1}^{\infty}\left|\operatorname{Tr}\left\{\{(S+\gamma I)^{-1}\Delta\}^k\{(S+\gamma I)^{-1}S-I\}\right\}\right| \\ &\leq \sum_{k=1}^{\infty}\left|\operatorname{Tr}\left\{\{(S+\gamma I)^{-1}\Delta\}^k\right\}\right|\end{aligned} \tag{25}$$

Now applying Lemma 2 with $B=\Delta$, and $A=(S+\gamma I)^{-1}$, we obtain

$$\begin{aligned}\operatorname{Tr}\left\{((S+\gamma I)^{-1}\Delta)^k\right\} &= \sum_{l=1}^{\infty}\left\langle e_l,((S+\gamma I)^{-1}\Delta)^k e_l\right\rangle \\ &= \sum_{l_1=1}^{\infty}\cdots\sum_{l_k=1}^{\infty}\left\{\left(\prod_{j=1}^{k}\left(\lambda_{l_j}+\gamma\right)^{-1}\right)\left(\prod_{j=1}^{k}\langle e_{l_j},\Delta e_{l_{j+1}}\rangle\right)\right\}\end{aligned}$$

Since for all $1\leq j\leq k$, we have $\left|\langle e_{l_j},\Delta e_{l_{j+1}}\rangle\right|\leq\left\|\Delta e_{l_j}\right\|$ and $\left(\lambda_{l_j}+\gamma\right)^{-1}\leq\gamma^{-1}$, the upper bound in Equation 25 is the sum of a geometric series whose ratio is

$$\gamma^{-1}\sum_{l=1}^{\infty}\|\Delta e_l\|=\gamma^{-1}\|\Delta\|_{\mathcal{C}_1}$$

where $\gamma^{-1}\|\Delta\|_{\mathcal{C}_1}<1$ by assumption, which completes the proof in Equation 22 when $r=1$. A similar reasoning as above allows to prove Equation 22 when $r=2$.

Now, let's prove the upper bound in Equation 23. Using that

$$\left|\operatorname{Tr}\left\{((S+\gamma I)^{-1}\Delta)^k\right\}\right|=\left|\operatorname{Tr}\left[\left\{\left(S^{1/2}(S+\gamma I)^{-1}S^{1/2}\right)\left(S^{-1/2}\Delta S^{-1/2}\right)\right\}^k\right]\right|$$

and apply Lemma 2 again, but with $B=S^{-1/2}\Delta S^{-1/2}$, and $A=S^{1/2}(S+\gamma I)^{-1}S^{1/2}$, yielding

$$\begin{aligned}\operatorname{Tr}\left\{((S+\gamma I)^{-1}\Delta)^k\right\} &= \sum_{l=1}^{\infty}\left\langle e_l,((S+\gamma I)^{-1}\Delta)^k e_l\right\rangle \\ &= \sum_{l_1=1}^{\infty}\cdots\sum_{l_k=1}^{\infty}\left\{\left(\prod_{j=1}^{k}\left(\lambda_{l_j}+\gamma\right)^{-1}\lambda_l\right)\left(\prod_{j=1}^{k}\langle e_{l_j},(S^{-1/2}\Delta S^{-1/2})e_{l_{j+1}}\rangle\right)\right\}\end{aligned}$$

Then, using that

$$\left|\langle e_{l_j},(S^{-1/2}\Delta S^{-1/2})e_{l_{j+1}}\rangle\right|\leq\left\|(S^{-1/2}\Delta S^{-1/2})e_{l_j}\right\|$$

and applying Hölder inequality[5], we obtain

[5] Let $(\Omega,\mathcal{F},\mathbb{P})$ be a measure space and let $p,q\in[1,\infty)$ with $1/p+1/q=1$. Then for all measurable real or complex valued functions f and g on Ω,

$$\|fg\|_1\leq\|f\|_p\|g\|_q$$

where $\|f\|_p=\int_\Omega|f|^p\ d\mathbb{P}$. In case of a counting measure, we have

$$
\begin{aligned}
&\left|\mathrm{Tr}\left\{\left((S+\gamma I)^{-1}\Delta\right)^k\right\}\right| \\
&\leq \left\{\sum_{l=1}^{\infty}(\lambda_l+\gamma)^{-2}\lambda_l^2\right\}^{k/2}\left\{\sum_{l_1=1}^{\infty}\cdots\sum_{l_k=1}^{\infty}\left(\prod_{j=1}^{k}\langle e_{l_j},(S^{-1/2}\Delta S^{-1/2})e_{l_{j+1}}\rangle^2\right)\right\}^{1/2} \\
&\leq (d_2(S,\gamma))^k\left\|S^{-1/2}\Delta S^{-1/2}\right\|_{\mathrm{HS}}^k
\end{aligned}
$$

Finally, going back to Equation 25, the upper bound is the sum of a geometric series whose ratio is $d_2(S,\gamma)\left\|S^{-1/2}\Delta S^{-1/2}\right\|_{\mathrm{HS}}$, where $d_2(S,\gamma)\left\|S^{-1/2}\Delta S^{-1/2}\right\|_{\mathrm{HS}} < 1$ by assumption, which completes the proof of Equation 23. As for Equation 24, observe that

$$
\begin{aligned}
|d_2(S+\Delta,\gamma)-d_2(S,\gamma)| &\leq \sum_{k=1}^{\infty}\left\|\left\{(S+\gamma I)^{-1}\Delta\right\}^k\left\{(S+\gamma I)^{-1}S-I\right\}\right\|_{\mathrm{HS}} \\
&\leq \sum_{k=1}^{\infty}\left\|\left\{(S+\gamma I)^{-1}\Delta\right\}^k\right\|_{\mathrm{HS}} \\
&\leq \sum_{k=1}^{\infty}\left\|\left\{S^{-1/2}\Delta S^{-1/2}\right\}\right\|_{\mathrm{HS}}^k
\end{aligned}
$$

where we used the inequality $\|AB\|_{\mathrm{HS}} \leq \|A\|_{\mathrm{HS}}\ \|B\|_{\mathrm{HS}}$, and $\|(S+\gamma I)^{-1}S-I\| \leq 1$ and $\|(S+\gamma I)^{-1}S\| \leq 1$ □

Lemma 4. Let Σ_A and Σ_B be two compact and self-adjoint operators in the RKHS $\mathcal{H}$. Assume that for $f \in \mathcal{H}$,

$$
\begin{aligned}
\|\Sigma_A^{-1/2}f\|_{\mathcal{H}} &< \infty \\
\|\Sigma_B^{-1/2}f\|_{\mathcal{H}} &< \infty
\end{aligned}
$$

Then,

$$
\left|\|\Sigma_A^{-1/2}f\|_{\mathcal{H}}^2-\|\Sigma_B^{-1/2}f\|_{\mathcal{H}}^2\right| \leq \left\|\Sigma_A^{-1/2}f\right\|_{\mathcal{H}}\left\|\Sigma_B^{-1/2}f\right\|_{\mathcal{H}}\left\|\Sigma_A^{-1/2}(\Sigma_A-\Sigma_B)\Sigma_B^{-1/2}\right\|
$$

Proof: Let

$$
\begin{aligned}
A &:= \Sigma_A \\
B &:= \Sigma_A-\Sigma_B \\
A-B &:= \Sigma_B
\end{aligned}
$$

Using the Sherman-Morrison-Woodbury formula argument, we have

$$
(A-B)^{-1} = A^{-1}+A^{-1}B(A-B)^{-1}
$$

$$
\sum_{k=1}^{n}|x_k y_k| \leq \left(\sum_{k=1}^{n}|x_k|^p\right)^{\frac{1}{p}}\left(\sum_{k=1}^{n}|y_k|^q\right)^{\frac{1}{q}}
$$

Thus,

$$\begin{aligned}\left|\|\Sigma_A^{-1/2}f\|_{\mathcal{H}}^2-\|\Sigma_B^{-1/2}f\|_{\mathcal{H}}^2\right| &= \left|\langle f, A^{-1}f\rangle_{\mathcal{H}}-\langle f,(A-B)^{-1}f\rangle_{\mathcal{H}}\right|\\ &= \left|\langle f, \Sigma_A^{-1}(\Sigma_A-\Sigma_B)\Sigma_B^{-1}f\rangle_{\mathcal{H}}\right|\\ &\overset{(1)}{=} \left|\langle \Sigma_A^{-1/2}f, \Sigma_A^{-1/2}(\Sigma_A-\Sigma_B)\Sigma_B^{-1}f\rangle_{\mathcal{H}}\right|\\ &\overset{(2)}{\le} \left\|\Sigma_A^{-1/2}f\right\|_{\mathcal{H}}\left\|\Sigma_A^{-1/2}(\Sigma_A-\Sigma_B)\Sigma_B^{-1/2}\Sigma_B^{-1/2}f\right\|_{\mathcal{H}}\\ &\overset{(3)}{\le} \left\|\Sigma_A^{-1/2}f\right\|_{\mathcal{H}}\left\|\Sigma_B^{-1/2}f\right\|_{\mathcal{H}}\left\|\Sigma_A^{-1/2}(\Sigma_A-\Sigma_B)\Sigma_B^{-1/2}\right\|\end{aligned}$$

here, equality (1) arises from the self-adjoint property of $\Sigma_A^{-1/2}$; inequality (2) is a direct consequence of the Cauchy-Schwarz inequality; inequality (3) is due to the fact that for any operator T in $\mathcal{H}$, we have

$$\|\ Tf\ \|_{\mathcal{H}} \le \|T\|\ \|f\|_{\mathcal{H}}$$

□

Lemma 5. Let $q_z(\cdot)\in\mathcal{H}$, indexed by z, be a member of an RHKS $\mathcal{H}$, and Σ be a compact and self-adjoint operator in this RKHS. Then,

$$\begin{aligned}\mathbb{E}\left\|\Sigma^{-1/2}\mathbb{E}_n(q_z(\cdot))\right\|_{\mathcal{H}}^2 = \frac{1}{n}\operatorname{Tr}\{\Sigma^{-1}\mathbb{E}(q_z(\cdot)\otimes q_z(\cdot))\} - \frac{1}{n}\operatorname{Tr}\{\Sigma^{-1}\mathbb{E}(q_z(\cdot))\otimes\mathbb{E}(q_z(\cdot))\}\\ +\operatorname{Tr}\{\Sigma^{-1}\mathbb{E}(q_z(\cdot))\otimes\mathbb{E}(q_z(\cdot))\}\end{aligned}$$

Proof:

$$\mathbb{E}\left\|\Sigma^{-1/2}\mathbb{E}_n(q_z(\cdot))\right\|_{\mathcal{H}}^2 = \operatorname{Tr}\{\Sigma^{-1}\mathbb{E}(\mathbb{E}_n(q_z(\cdot))\otimes\mathbb{E}_n(q_z(\cdot)))\}$$

Note that

$$\begin{aligned}\mathbb{E}(\mathbb{E}_n(q_z(\cdot))\otimes\mathbb{E}_n(q_z(\cdot))) &= \mathbb{E}\left(\frac{1}{n^2}\sum_i\sum_j q_{z_i}(\cdot)\otimes q_{z_j}(\cdot)\right)\\ &= \frac{1}{n^2}[n\mathbb{E}(q_z(\cdot)\otimes q_z(\cdot))+n(n-1)\mathbb{E}(q_z(\cdot))\otimes\mathbb{E}(q_z(\cdot))]\end{aligned}$$

The second equality comes from the fact that in the double summation, when $i=j$, we have n pairs of $q_{z_i}(\cdot)\otimes q_{z_i}(\cdot)$; and when $i\neq j$, we have $n(n-1)$ pairs of $q_{z_i}(\cdot)\otimes q_{z_j}(\cdot)$.

Thus,

$$\begin{aligned}\mathrm{Tr}\{\Sigma^{-1}\mathbb{E}(\mathbb{E}_n(q_z(\cdot))\otimes\mathbb{E}_n(q_z(\cdot)))\} = \frac{1}{n}\ &\mathrm{Tr}\{\Sigma^{-1}\mathbb{E}(q_z(\cdot)\otimes q_z(\cdot))\}\\ -\frac{1}{n}\ &\mathrm{Tr}\{\Sigma^{-1}\mathbb{E}(q_z(\cdot))\otimes\mathbb{E}(q_z(\cdot))\}\\ +\ &\mathrm{Tr}\{\Sigma^{-1}\mathbb{E}(q_z(\cdot))\otimes\mathbb{E}(q_z(\cdot))\}\end{aligned}$$

□

B.2 Results on Eigenvalues

Lemma 6. Let A be a self-adjoint compact operator on $\mathcal{H}$, and let $\{\psi_l\}_{l\geq 1}$ be an orthonormal basis of $\mathcal{H}$ consisting of a sequence of eigenfunctions of A corresponding to the eigenvalues $\{\lambda_l(A)\}$ of this latter operator, so that

$$\langle\psi_l, A\psi_l\rangle_{\mathcal{H}} = \lambda_l(A)$$

Then

$$|\lambda_l(A)| \leq \|A\psi_l\|_{\mathcal{H}}$$

In addition, for any orthonormal basis $\{\varphi_l\}_{l\geq 1}$ of $\mathcal{H}$, we have

$$\sum_{l=1}^{\infty}|\lambda_l(A)| \leq \sum_{l=1}^{\infty}\|A\varphi_l\|_{\mathcal{H}}$$

Proof:

For any orthonormal basis $\{\varphi_q\}_{q\geq 1}$, we have

$$\begin{aligned}A\psi_l &= \sum_q\langle A\psi_l, \varphi_q\rangle_{\mathcal{H}}\varphi_q\\ &= \sum_q\langle\psi_l, A\varphi_q\rangle_{\mathcal{H}}\varphi_p\end{aligned}$$

where the last equality arises from the self-adjoint property of A. Then,

$$\begin{aligned}|\lambda_l(A)| = |\langle\psi_l, A\psi_l\rangle_{\mathcal{H}}| &\leq \sum_{q=1}^{\infty}|\langle\varphi_q, A\psi_l\rangle_{\mathcal{H}}||\langle\varphi_q, \psi_l\rangle_{\mathcal{H}}|\\ &\leq \left(\sum_{q=1}^{\infty}\left|\langle\varphi_q, A\psi_l\rangle_{\mathcal{H}}\right|^2\right)^{1/2}\left(\sum_{q=1}^{\infty}\left|\langle\varphi_q, \psi_l\rangle_{\mathcal{H}}\right|^2\right)^{1/2}\\ &= \|A\psi_l\|_{\mathcal{H}}\end{aligned}$$

where the first inequality comes from the triangle inequality. The second inequality comes from the Hölder inequality. The third equality comes the Parseval's identity:

$$1 = \left\|\varphi_q\right\|_{\mathcal{H}}^2 = \sum_{l\geq 1}\left\langle\varphi_q, \psi_l\right\rangle_{\mathcal{H}}^2$$

Similarly,

$$\begin{aligned}\sum_{l=1}^{\infty}|\lambda_l(A)| = \sum_{l=1}^{\infty}|\langle\psi_l, A\psi_l\rangle_{\mathcal{H}}| &\leq \sum_{q=1}^{\infty}\sum_{l=1}^{\infty}|\langle A\varphi_q, \psi_l\rangle_{\mathcal{H}}||\langle\varphi_q, \psi_l\rangle_{\mathcal{H}}| \\ &\leq \sum_{q=1}^{\infty}\left(\sum_{l=1}^{\infty}\left|\langle A\varphi_q, \psi_l\rangle_{\mathcal{H}}\right|^2\right)^{1/2}\left(\sum_{l=1}^{\infty}\left|\langle\varphi_q, \psi_l\rangle_{\mathcal{H}}\right|^2\right)^{1/2} \\ &= \sum_{q=1}^{\infty}\|A\varphi_q\|_{\mathcal{H}}\end{aligned}$$

□

Lemma 7. Assume A1. Let $\{X_1^n, ..., X_n^n\}$ be a triangular array of i.i.d random variables, whose mean element and covariance operator are respectively (μ^n, Σ^n). If, for all n, all the eigenvalues $\lambda_l(\Sigma^n)$ of Σ^n are non-negative, and if there exists $D > 0$ such that for all n, we have

$$\sum_{l=1}^{\infty}\sqrt{\lambda_l(\Sigma^n)} < D$$

and the residual random variable is bounded in probability: $\varepsilon = O_p(1)$.

Then,

$$\sum_{l=1}^{\infty}|\lambda_l(\Sigma_n - \Sigma^n)| = O_p\left(n^{-1/2}\right)$$

Proof: Lemma 6 shows that, for any orthonormal basis $\{e_l\}_{l\geq 1}$ in the RKHS $\mathcal{H}$:

$$\sum_{l=1}^{\infty}|\lambda_l(\Sigma_n - \Sigma^n)| \leq \sum_{l=1}^{\infty}\|(\Sigma_n - \Sigma^n)e_l\|_{\mathcal{H}}$$

We take the orthonormal family of eigenvectors $\{e_l\}_{l\geq 1}$ of the covariance operator Σ^n. Then, it suffices to show that $\sum_{l=1}^{\infty}\|(\Sigma_n - \Sigma^n)e_l\|_{\mathcal{H}} = O_p\left(n^{-1/2}\right)$. Note that,

$$(\Sigma_n - \Sigma^n)e_l = n^{-1}\sum_{i=1}^{n}\zeta_{l,n,i}$$

where

$$\zeta_{l,n,i} := \varepsilon_i^2 k(X_i, \cdot)e_l(X_i) - \mathbb{E}^n\left(\varepsilon^2 k(X_1, \cdot)e_l(X_1)\right)$$

Thus,

$$\left\{\mathbb{E}^n \|(\Sigma_n - \Sigma^n)e_l\|_{\mathcal{H}}^2\right\}^{1/2} = \left\{\mathbb{E}^n \left\| n^{-1} \sum_{i=1}^{n} \zeta_{l,n,i} \right\|_{\mathcal{H}}^2 \right\}^{1/2}$$
$$= A_1$$

Let’s consider A_1. We have,

$$A_1^2 = n^{-1}\mathbb{E}^n\left(\|\zeta_{l,n,i}\|_{\mathcal{H}}^2\right) \leq n^{-1}\mathbb{E}^n\{\|\varepsilon^2 k(X_1,\cdot)\|_{\mathcal{H}}^2 \mid e_l(X_1)|^2\} \leq n^{-1}C\ |k|_\infty\ \mathbb{E}^n[\![e_l(X_1)|^2]$$

By the Minkowski inequality[6], this shows that

$$\left\{\mathbb{E}^n\left(\sum_{l=1}^{\infty}\|(\Sigma_n - \Sigma^n)e_l\|\right)_{\mathcal{H}}^2\right\}^{1/2} \leq 2C\ |k|_\infty^{1/2}\ n^{-1/2}\sum_{l=1}^{\infty}\{\mathbb{E}^n[|e_l(X_1)|^2]\}^{1/2}$$

Let’s investigate $\{\mathbb{E}^n[|e_l(X_1)|^2]\}^{1/2}$. Recall that

$$\mathbb{E}^n[|e_l(X_1)|^2] = \langle e_l, \Sigma^n e_l\rangle_{\mathcal{H}} = \lambda_l(\Sigma^n)$$

Thus, $\sum_{l=1}^{\infty}\{\mathbb{E}^n[|e_l(X_1)|^2]\}^{1/2} = \sum_{l=1}^{\infty}\sqrt{\lambda_l(\Sigma^n)} < D$ by assumption. Finally, putting everything together, we have

$$\sum_{l=1}^{\infty}|\lambda_l(\Sigma_n - \Sigma^n)| \leq 2\ |k|_\infty^{1/2}\ n^{-1/2}\sum_{l=1}^{\infty}\{\mathbb{E}^n[\![e_l(X_1)|^2]\}^{1/2} = O_p\left(n^{-1/2}\right)$$

□

Lemma 8. Assume A1. Let $\{Z_{1,n}, ..., Z_{n,n}\}$ be a triangular array, whose elements and covariance operators are respectively (μ^n, Σ^n).

If

$$\sup_{n\geq 0}\sum_{l=1}^{\infty}\sqrt{\lambda_l(\Sigma^n)} < \infty$$

then

$$\|\Sigma_n - \Sigma^n\|_{\text{HS}} = O_p\left(n^{-1/2}\right)$$

[6] The Minkowski inequality establishes that the L^p spaces are normed vector space. Let S be a measurable space, let $1 \leq p < \infty$ and let f and g be elements of $L^p(S)$, and we have the triangle inequality:

$$\|f + g\|_p \leq \|f\|_p + \|g\|_p$$

The Minkowski inequality can be specialized to sequences and vectors by using the counting measure:

$$\left(\sum_{k=1}^{n}|x_k + y_k|^p\right)^{1/p} \leq \left(\sum_{k=1}^{n}|x_k|^p\right)^{1/p} + \left(\sum_{k=1}^{n}|y_k|^p\right)^{1/p}$$

Proof: This lemma is a direct consequence of Lemma 7. Specifically, applying Lemma 7 states:

$$\sum_{l=1}^{\infty} |\lambda_l(\Sigma_n) - \lambda_l(\Sigma^n)| = O_p\left(n^{-1/2}\right)$$

Now, using that

$$\underbrace{\|\Sigma_n - \Sigma^n\|_{\text{HS}} \leq \sum_{l=1}^{\infty} | \ \lambda_l(\Sigma_n - \Sigma^n)|}_{\text{using the fact that } \|\lambda\|_q \leq \|\lambda\|_p \text{ , for all } q>p} \leq \sum_{l=1}^{\infty} \left\|(\Sigma_n - \Sigma^n)e_p\right\|_{\mathcal{H}}$$

Then,

$$\|\Sigma_n - \Sigma^n\|_{\text{HS}} = O_p\left(n^{-1/2}\right)$$

□