Testing for Homogeneity with Kernel Fisher Discriminant Analysis

Zaid Harchaoui, Francis Bach, Eric Moulines

Testing for Homogeneity with Kernel Fisher Discriminant Analysis

Zaïd Harchaoui zaid.harchaoui@enst.fr
LTCI, TELECOM ParisTech & CNRS
46, rue Barrault
75634 Paris cedex 13, France

Francis R. Bach francis.bach@mines.org
INRIA - Willow Project-Team
Laboratoire d'Informatique de l'École Normale Supérieure
45, rue d'Ulm
75230 Paris, France

Éric Moulines eric.moulines@enst.fr
LTCI, TELECOM ParisTech & CNRS
46, rue Barrault
75634 Paris cedex 13, France

Editor:

Abstract

We propose to investigate test statistics for testing homogeneity in reproducing kernel Hilbert spaces. Asymptotic null distributions under null hypothesis are derived, and consistency under fixed and local alternatives is assessed. Finally, experimental evidence of the performance of the proposed approach on both artificial data and a speaker verification task is provided.

Keywords: statistical hypothesis testing, reproducing kernel Hilbert space, covariance operator

1. Introduction

An important problem in statistics and machine learning consists in testing whether the distributions of two random variables are identical under the alternative that they may differ in some ways. More precisely, let $\{X_1^{(1)}, \ldots, X_{n_1}^{(1)}\}$ and $\{X_1^{(2)}, \ldots, X_{n_2}^{(2)}\}$ be independent random variables taking values in an arbitrary input space $\mathcal{X}$, with common distributions $\mathbb{P}_1$ and $\mathbb{P}_2$, respectively. The problem consists in testing the null hypothesis of homogeneity $\mathbf{H}_0 : \mathbb{P}_1 = \mathbb{P}_2$, against the alternative $\mathbf{H}_A : \mathbb{P}_1 \neq \mathbb{P}_2$. This problem arises in many applications, ranging from computational anatomy (Grenander and Miller, 2007) to speaker segmentation (Bimbot et al., 2004). We shall allow the input space $\mathcal{X}$ to be quite general, including for example finite-dimensional euclidean spaces but also function spaces, or more sophisticated structures such as strings or graphs (see Shawe-Taylor and Cristianini, 2004) arising in applications such as bioinformatics (see recently Borgwardt et al., 2006).

Traditional approaches to this problem are based on cumulative distribution functions (cdf), and use a certain distance between the empirical cdf obtained from the two samples. Popular procedures are the two-sample Kolmogorov-Smirnov tests or the Cramer-Von Mises

tests (Lehmann and Romano, 2005), that have been frequently used to address these issues, at least for low-dimensional data. Although these tests are popular due to their simplicity, they are known to be insensitive to certain characteristics of the distributions, such as densities containing high-frequency components or local features such as narrow bumps. The low-power of the traditional cdf-based test statistics can be improved on by using test statistics based on probability density estimators. Tests based on kernel density estimators have been studied by Anderson et al. (1994) and Allen (1997), using respectively the L^2 and L^1 distances between densities. More recently, the use of wavelet estimators has been proposed and thoroughly analyzed. Adaptive versions of these tests, that is where smoothing parameters for the density estimator are obtained from the data, have been considered by Butucea and Tribouley (2006).

Recently, Gretton et al. (2006) cast the two-sample homogeneity test in a kernel-based framework, and have shown that their test statistics, coined Maximum Mean Discrepancy (MMD) yields as a particular case the L^2-distance between kernel density estimators. We propose here to further enhance such an approach by directly incorporating the covariance structure of the probability distributions into our test statistics, yielding in some sense to a chi-square divergence between the two distributions. For discrete distributions, it is well-known that such a normalization yield test statistics with greater power (Lehmann and Romano, 2005).

The paper is organized as follows. In Section 2 and Section 3, we state the main definitions and we build our test statistics upon kernel Fisher discriminant analysis. In Section 4, we give the asymptotic distribution of our test statistic under the null hypothesis, and establish the consistency and the limiting distribution of the test for both fixed and a class of local alternatives. In Section 5, we first investigate the limiting power of our test statistics against directional then non-directional sequences of local alternatives in a particular setting, that is when $\mathbb{P}_1$ is the uniform distribution and $\mathbb{P}_2$ is a one-frequency contamination of $\mathbb{P}_1$ on the Fourier basis and the reproducing kernel belongs to the class of periodic spline kernels, and then compare our test statistics with the MMD test statistics in terms of limiting power. In Section 6 we provide experimental evidence of the performance of our test statistic on a speaker identification task. Detailed proofs are presented in the last sections.

2. Mean and covariance in reproducing kernel Hilbert spaces

We first highlight the main assumptions on the reproducing kernel, and then introduce operator-theoretic tools for defining the mean element and the covariance operator associated with a reproducing kernel.

2.1 Reproducing kernel Hilbert spaces

Let $(\mathcal{X}, d)$ be a separable measurable metric space, and denote by $\mathfrak{X}$ the associated σ-algebra. Let X be $\mathcal{X}$-valued random variable, with probability measure $\mathbb{P}$, and the expectation with respect to $\mathbb{P}$ is denoted by $\mathbb{E}$. Consider a Hilbert space $(\mathcal{H}, \langle \cdot, \cdot \rangle_{\mathcal{H}})$ of functions from $\mathcal{X}$ to $\mathbb{R}$. The Hilbert space $\mathcal{H}$ is a reproducing kernel Hilbert space (RKHS) if at each $x \in \mathcal{X}$, the point evaluation operator $\delta_x : \mathcal{H} \to \mathbb{R}$, which maps $f \in \mathcal{H}$ to $f(x) \in \mathbb{R}$, is a bounded linear functional. To each point $x \in \mathcal{X}$, there corresponds an element $\Phi(x) \in \mathcal{H}$ such that

$\langle \Phi(x), f \rangle_{\mathcal{H}} = f(x)$ for all $f \in \mathcal{H}$, and $\langle \Phi(x), \Phi(y) \rangle_{\mathcal{H}} = k(x,y)$, where $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ is a positive definite kernel (Aronszajn, 1950). In this situation, $\Phi(\cdot)$ is the Aronszajn-map, and we denote by $\|f\|_{\mathcal{H}} = \langle f, f \rangle_{\mathcal{H}}^{1/2}$ the associated norm. It is assumed from now on that $\mathcal{H}$ is a separable Hilbert space. Note that this is always the case if $\mathcal{X}$ is a separable metric space and if the kernel is continuous (see Steinwart et al., 2006a). We make the following two assumptions on the kernel:

(**A1**) The kernel k is bounded, that is $|k|_\infty \stackrel{\text{def}}{=} \sup_{(x,y) \in \mathcal{X} \times \mathcal{X}} k(x,y) < \infty$.

(**A2**) For all probability distributions $\mathbb{P}$ on $(\mathcal{X}, \mathfrak{X})$, the RKHS associated with $k(\cdot,\cdot)$ is dense in $L^2(\mathbb{P})$.

Note that some of our results (such as the limiting distribution under the null distribution) are valid without assumption (A2), while consistency results against fixed or local alternatives do need (A2). Assumption (A2) is true in particular for the gaussian kernel on $\mathbb{R}^d$ as shown in (Steinwart et al., 2006b, Theorem 2), and that $\mathcal{X}$ may be a discrete space (Steinwart et al., 2006b, Corollary 3).

2.2 Mean element and covariance operator

We shall need some operator-theoretic tools (see Aubin, 2000), to define mean elements and covariance operators. 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 $\|T\| = \sup_{\|f\|_{\mathcal{H}} \leq 1} \|Tf\|_{\mathcal{H}}$. Furthermore, a bounded linear operator T is said to be Hilbert-Schmidt, if the Hilbert-Schmidt-norm $\|T\|_{\text{HS}} = \{\sum_{p=1}^{\infty} \langle Te_p, Te_p \rangle_{\mathcal{H}}\}^{1/2}$ is finite, where $\{e_p\}_{p \geq 1}$ is any complete orthonormal basis of $\mathcal{H}$. Note that $\|T\|_{\text{HS}}$ is independent of the choice of the orthonormal basis. We shall make frequent use of tensor product notations. The tensor product operator $u \otimes v$ for $u, v \in \mathcal{H}$ is defined for all $f \in \mathcal{H}$ as $(u \otimes v) f = \langle v, f \rangle_{\mathcal{H}} u$.

We now introduce the mean element and covariance operator (see Blanchard et al., 2007). If $\int k^{1/2}(x,x) \mathbb{P}(dx) < \infty$, the mean element $\mu_{\mathbb{P}}$ is defined as the unique element in $\mathcal{H}$ satisfying for all functions $f \in \mathcal{H}$,

$$\langle \mu_{\mathbb{P}}, f \rangle_{\mathcal{H}} = \mathbb{P}f \stackrel{\text{def}}{=} \int f d\mathbb{P} \,. \tag{1}$$

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

$$\langle f, \Sigma_{\mathbb{P}} g \rangle_{\mathcal{H}} \stackrel{\text{def}}{=} \int (f - \mathbb{P}f)(g - \mathbb{P}g) d\mathbb{P} \,, \tag{2}$$

that is $\langle f, \Sigma_{\mathbb{P}} g \rangle_{\mathcal{H}}$ is the covariance between $f(X)$ and $g(X)$ where X is distributed according to $\mathbb{P}$. Note that the mean element and covariance operator are well-defined when (A1) is satisfied. Moreover, when assumption (A2) is satisfied, then the map from $\mathbb{P} \mapsto \mu_{\mathbb{P}}$ is injective. Note also that the operator $\Sigma_{\mathbb{P}}$ is a self-adjoint nonnegative trace-class operator. In the sequel, the dependence of $\mu_{\mathbb{P}}$ and $\Sigma_{\mathbb{P}}$ in $\mathbb{P}$ is omitted whenever there is no risk of confusion.

We now define what we later denote by $\Sigma^{-1/2}$ in our proofs. For a compact operator Σ, the range $\mathcal{R}(\Sigma^{1/2})$ of $\Sigma^{1/2}$ is defined as $\mathcal{R}(\Sigma^{1/2}) = \{\Sigma^{1/2} f,\ f \in \mathcal{H}\}$, and may be characterized by $\mathcal{R}(\Sigma^{1/2}) = \{f \in \mathcal{H},\ \sum_{p=1}^{\infty} \lambda_p \langle f, e_p\rangle_{\mathcal{H}}^2 < \infty,\ f \perp \mathcal{N}(\Sigma^{1/2})\}$, where $\{\lambda_p, e_p\}_{p\geq 1}$ are the nonzero eigenvalues and eigenvectors of Σ, and $\mathcal{N}(\Sigma) = \{f \in \mathcal{H},\ \Sigma f = 0\}$ is the null-space of Σ, that is functions which are constant in the support of $\mathbb{P}$. Defining $\mathcal{R}^{-1}(\Sigma^{1/2}) = \{g \in \mathcal{H},\ g = \sum_{p=1}^{\infty} \lambda_p^{-1/2} \langle f, e_p\rangle_{\mathcal{H}} e_p,\ f \in \mathcal{R}(\Sigma^{1/2})\}$, we observe that $\Sigma^{1/2}$ is a one-to-one mapping between $\mathcal{R}^{-1}(\Sigma^{1/2})$ and $\mathcal{R}(\Sigma^{1/2})$. Thus, restricting the domain of $\Sigma^{1/2}$ to $\mathcal{R}^{-1}(\Sigma^{1/2})$, we may define its inverse for all $f \in \mathcal{R}(\Sigma^{1/2})$ as $\Sigma^{-1/2} f = \sum_{p=1}^{\infty} \lambda_p^{-1/2} \langle f, e_p\rangle_{\mathcal{H}} e_p$. The null space may be reduced to the null element (in particular for the gaussian kernel), or may be infinite-dimensional. Similarly, there may be infinitely many strictly positive eigenvalues (true nonparametric case) or finitely many (underlying finite-dimensional problems).

Given a sample $\{X_1, \ldots, X_n\}$, the empirical estimates respectively of the mean element and the covariance operator are then defined as follows:

$$\hat{\mu} \stackrel{\text{def}}{=} n^{-1} \sum_{i=1}^{n} k(X_i, \cdot)\ , \tag{3}$$

$$\hat{\Sigma} \stackrel{\text{def}}{=} n^{-1} \sum_{i=1}^{n} k(X_i, \cdot) \otimes k(X_i, \cdot) - \hat{\mu} \otimes \hat{\mu}\ . \tag{4}$$

By the reproducing property, they lead, on the one hand, to empirical means as from (3) we have $\langle \hat{\mu}, f\rangle = n^{-1} \sum_{i=1}^{n} f(X_i)$ for all $f \in \mathcal{H}$, and on the other hand, to empirical covariances as from (4) we have $\langle f, \hat{\Sigma} g\rangle_{\mathcal{H}} = n^{-1} \sum_{i=1}^{n} f(X_i) g(X_i) - \{n^{-1} \sum_{i=1}^{n} f(X_i)\}\{n^{-1} \sum_{i=1}^{n} g(X_i)\}$ for all $f, g \in \mathcal{H}$.

3. KFDA-based test statistic

Our two-sample homogeneity test can be formulated as follows. Let $\{X_1^{(1)}, \ldots, X_{n_1}^{(1)}\}$ and $\{X_1^{(2)}, \ldots, X_{n_2}^{(2)}\}$ two independent identically distributed samples (iid) respectively from $\mathbb{P}_1$ and $\mathbb{P}_2$, having mean and covariance operators given by (μ_1, Σ_1) and (μ_2, Σ_2). We build our test statistics using a (regularized) kernelized version of the Fisher discriminant analysis. Denote by $\Sigma_W \stackrel{\text{def}}{=} (n_1/n)\Sigma_1 + (n_2/n)\Sigma_2$ the pooled covariance operator, where $n \stackrel{\text{def}}{=} n_1 + n_2$, corresponding to the within-class covariance matrix in the finite-dimensional setting (see Hastie et al., 2001).

3.1 Maximum Kernel Fisher Discriminant Ratio

Let us denote $\Sigma_B \stackrel{\text{def}}{=} (n_1 n_2/n^2)(\mu_2 - \mu_1) \otimes (\mu_2 - \mu_1)$ the between-class covariance operator. For $a = 1, 2$, denote by $(\hat{\mu}_a, \hat{\Sigma}_a)$ respectively the empirical estimates of the mean element and the covariance operator, defined as previously stated in (3) and (4). Denote $\hat{\Sigma}_W \stackrel{\text{def}}{=} (n_1/n)\hat{\Sigma}_1 + (n_2/n)\hat{\Sigma}_2$ the empirical pooled covariance estimator, and $\hat{\Sigma}_B \stackrel{\text{def}}{=} (n_1 n_2/n^2)(\hat{\mu}_2 - \hat{\mu}_1) \otimes (\hat{\mu}_2 - \hat{\mu}_1)$ the empirical between-class covariance operator. Let $\{\gamma_n\}_{n\geq 0}$ be a sequence of strictly positive numbers. The *maximum kernel Fisher discriminant ratio* serves as a

basis of our test statistics:

$$n \max_{f \in \mathcal{H}} \frac{\left\langle f, \hat{\Sigma}_B f \right\rangle_{\mathcal{H}}}{\left\langle f, (\hat{\Sigma}_W + \gamma_n \mathrm{I}) f \right\rangle_{\mathcal{H}}} = n_1 n_2 / n \ \left\| (\hat{\Sigma}_W + \gamma_n \mathrm{I})^{-1/2} (\hat{\mu}_2 - \hat{\mu}_1) \right\|_{\mathcal{H}}^2 , \tag{5}$$

where I denotes the identity operator. Note that if the input space is Euclidean, *e.g.*, $\mathcal{X} = \mathbb{R}^d$, the kernel is linear $k(x, y) = x^T y$ and $\gamma_n = 0$, this quantity matches the so-called Hotelling's T^2-statistic in the two-sample case (Lehmann and Romano, 2005).

We shall make the following assumptions respectively on Σ_1 and Σ_2

(**B1**) For $u = 1, 2$, the eigenvalues $\{\lambda_p(\Sigma_u)\}_{p \geq 1}$ satisfy $\sum_{p=1}^{\infty} \lambda_p^{1/2}(\Sigma_u) < \infty$.

(**B2**) For $u = 1, 2$, there are infinitely many strictly positive eigenvalues $\{\lambda_p(\Sigma_u)\}_{p \geq 1}$ of Σ_u.

The statistical analysis conducted in Section 4 shall demonstrate, in the case $\gamma_n \to 0$, the need to respectively recenter and rescale (a standard statistical transformation known as *studentization*) the maximum Fisher discriminant ratio, in order to get a theoretically well-grounded test statistic. These roles, recentering and rescaling, will be played respectively by $d_1(\Sigma_W, \gamma)$ and $d_2(\Sigma_W, \gamma)$, where for a given compact operator Σ with decreasing eigenvalues λ_p, the quantity $d_r(\Sigma, \gamma)$ is defined for all $q \geq 1$ as

$$d_r(\Sigma, \gamma) \stackrel{\text{def}}{=} \left\{ \sum_{p=1}^{\infty} (\lambda_p + \gamma)^{-r} \lambda_p^r \right\}^{1/r} . \tag{6}$$

3.2 Computation of the test statistics

In practice the test statistics may be computed thanks to the kernel trick, adapted to the kernel Fisher discriminant analysis as outlined in (Shawe-Taylor and Cristianini, 2004, Chapter 6). Let us consider the two samples $\{X_1^{(1)}, \ldots, X_{n_1}^{(1)}\}$ and $\{X_{n_1}^{(2)}, \ldots, X_{n_2}^{(2)}\}$, with $n_1 + n_2 = n$. Denote by $\mathbf{G}_n^{(u)} : \mathbb{R}^{n_u} \mapsto \mathcal{H}$, $u = 1, 2$, the linear operators which associates to a vector $\boldsymbol{\alpha}^{(u)} = [\alpha_1^{(u)}, \ldots, \alpha_{n_u}^{(u)}]^T$ the vector in $\mathcal{H}$ given by $\mathbf{G}_n^{(u)} \boldsymbol{\alpha}^{(u)} = \sum_{j=1}^{n_u} \alpha_j^{(u)} k(X_j^{(u)}, \cdot)$. This operator may be presented in a matrix form

$$\mathbf{G}_n^{(u)} = \left[k(X_1^{(u)}, \cdot), \ldots, k(X_{n_u}^{(u)}, \cdot) \right] . \tag{7}$$

We denote by $\mathbf{G}_n = \left[\mathbf{G}_n^{(1)} \ \mathbf{G}_n^{(2)} \right]$. We denote by $\mathbf{K}_n^{(u,v)} = [\mathbf{G}_n^{(u)}]^T \mathbf{G}_n^{(v)}$, $u, v \in \{0, 1\}$, the Gram matrix given by $\mathbf{K}_n^{(u,v)}(i, j) \stackrel{\text{def}}{=} k(X_i^{(u)}, X_j^{(v)})$ for $i \in \{1, \ldots, n_u\}$, $j \in \{1, \ldots, n_v\}$. Define, for any integer ℓ, $\mathbf{P}_\ell = \mathrm{I}_\ell - \ell^{-1} \mathbf{1}_\ell \mathbf{1}_\ell^T$ where $\mathbf{1}_\ell$ is the $(\ell \times 1)$ vector whose components are all equal to one and I_ℓ is the $(\ell \times \ell)$ identity matrix and let $\mathbf{N}_n$ be given by

$$\mathbf{N}_n \stackrel{\text{def}}{=} \begin{pmatrix} \mathbf{P}_{n_1} & 0 \\ 0 & \mathbf{P}_{n_2} \end{pmatrix} . \tag{8}$$

Finally, define the vector $\mathbf{m}_n = (\mathbf{m}_{n,i})_{1 \leq i \leq n}$ with $\mathbf{m}_{n,i} = -n_1^{-1}$ for $i = 1, \ldots, n_1$ and $\mathbf{m}_{n,i} = n_2^{-1}$ for $i = n_1 + 1, \ldots, n_1 + n_2$. With the notations introduced above,

$$\hat{\mu}_2 - \hat{\mu}_1 = \mathbf{G}_n \mathbf{m}_n , \quad \hat{\Sigma}_u = n_u^{-1} \mathbf{G}_n^{(u)} \mathbf{P}_{n_u} \mathbf{P}_{n_u}^T (\mathbf{G}_n^{(u)})^T , u = 1, 2 , \quad \hat{\Sigma}_W = n^{-1} \mathbf{G}_n \mathbf{N}_n \mathbf{N}_n^T \mathbf{G}_n^T ,$$

which implies that

$$\left\langle \hat{\mu}_2 - \hat{\mu}_1, (\hat{\Sigma}_W + \gamma I)^{-1}(\hat{\mu}_2 - \hat{\mu}_1) \right\rangle_{\mathcal{H}} = \mathbf{m}_n^T \mathbf{G}_n^T (n^{-1}\mathbf{G}_n\mathbf{N}_n\mathbf{N}_n^T\mathbf{G}_n^T + \gamma \mathrm{I})^{-1}\mathbf{G}_n\mathbf{m}_n \ .$$

Then, using the matrix inversion lemma, we get

$$\begin{aligned}
&\mathbf{m}_n^T \mathbf{G}_n^T (n^{-1}\mathbf{G}_n\mathbf{N}_n\mathbf{N}_n^T\mathbf{G}_n^T + \gamma \mathrm{I})^{-1}\mathbf{G}_n\mathbf{m}_n \\
&\quad = \gamma^{-1}\mathbf{m}_n^T\mathbf{G}_n^T \left\{ \mathrm{I} - n^{-1}\mathbf{G}_n\mathbf{N}_n(\gamma \mathrm{I} + n^{-1}\mathbf{N}_n^T\mathbf{G}_n^T\mathbf{G}_n\mathbf{N}_n)^{-1}\mathbf{N}_n^T\mathbf{G}_n^T \right\} \mathbf{G}_n\mathbf{m}_n \\
&\quad = \gamma^{-1}\left\{ \mathbf{m}_n^T\mathbf{K}_n\mathbf{m}_n - n^{-1}\mathbf{m}_n^T\mathbf{K}_n\mathbf{N}_n(\gamma \mathrm{I} + n^{-1}\mathbf{N}_n\mathbf{K}_n\mathbf{N}_n)^{-1}\mathbf{N}_n\mathbf{K}_n\mathbf{m}_n \right\} \ .
\end{aligned}$$

Hence, the maximum kernel Fisher discriminant ratio may be computed from

$$\begin{aligned}
&n_1 n_2/n \ \left\| (\hat{\Sigma}_W + \gamma_n \mathrm{I})^{-1/2}(\hat{\mu}_2 - \hat{\mu}_1) \right\|_{\mathcal{H}}^2 \\
&\qquad = n_1 n_2/\gamma n \ \left\{ \mathbf{m}_n^T\mathbf{K}_n\mathbf{m}_n - n^{-1}\mathbf{m}_n^T\mathbf{K}_n\mathbf{N}_n(\gamma \mathrm{I} + n^{-1}\mathbf{N}_n\mathbf{K}_n\mathbf{N}_n)^{-1}\mathbf{N}_n\mathbf{K}_n\mathbf{m}_n \right\} \ .
\end{aligned}$$

4. Main results

This discussion yields the following normalized test statistics:

$$\widehat{T}_n(\gamma_n) = \frac{n_1 n_2/n \ \left\| (\hat{\Sigma}_W + \gamma_n \mathrm{I})^{-1/2}\hat{\delta} \right\|_{\mathcal{H}}^2 - d_1(\hat{\Sigma}_W, \gamma_n)}{\sqrt{2}\ d_2(\hat{\Sigma}_W, \gamma_n)} \ . \tag{9}$$

In this paper, we first consider the asymptotic behavior of $\widehat{T}_n$ under the null hypothesis, and against a fixed alternative. This will establish that our nonparametric test procedure is consistent. However, this is not enough, as it can be arbitrarily slow. We thus then consider local alternatives.

For all our results, we consider two situations regarding the regularization parameter γ_n; (a) a situation where γ_n is held fixed, and in which the limiting distribution is somewhat similar to the maximum mean discrepancy test statistics, and (b) a situation where γ_n tends to zero slowly enough, and in which we obtain qualitatively different results.

4.1 Limiting distribution under null hypothesis

Throughout this paper, we assume that the proportions n_1/n and n_2/n converge to strictly positive numbers, that is

$$n_u/n \to \rho_u, \quad \text{as } n = n_1 + n_2 \to \infty \ , \quad \text{with } \rho_u > 0 \text{ for } u = 1, 2 \ .$$

In this section, we derive the distribution of the test statistics under the null hypothesis $\mathbf{H}_0 : \mathbb{P}_1 = \mathbb{P}_2$ of homogeneity, which implies $\mu_1 = \mu_2$ and $\Sigma_1 = \Sigma_2 = \Sigma_W$. We first consider the case where the regularization factor is held constant $\gamma_n \equiv \gamma$. We denote $\xrightarrow{\mathcal{D}}$ the convergence in distribution.

Theorem 1. *Assume (A1-B1). Assume in addition that the probability distributions $\mathbb{P}_1$ and $\mathbb{P}_2$ are equal, i.e. $\mathbb{P}_1 = \mathbb{P}_2 = \mathbb{P}$, and that $\gamma_n \equiv \gamma > 0$. Then,*

$$\widehat{T}_n(\gamma) \xrightarrow{\mathcal{D}} T_\infty(\Sigma_W, \gamma) \stackrel{\text{def}}{=} 2^{-1/2}\, d_2^{-1}(\Sigma_W, \gamma) \sum_{p=1}^{\infty} (\lambda_p(\Sigma_W) + \gamma)^{-1} \lambda_p(\Sigma_W)(Z_p^2 - 1)\ , \qquad (10)$$

where $\{\lambda_p(\Sigma_W)\}_{p\geq 1}$ are the eigenvalues of the covariance operator Σ_W, and $d_2(\Sigma_W, \gamma)$ is defined in (6), and Z_p, $p \geq 1$ are independent standard normal variables.

If the number of non-vanishing eigenvalues is equal to p and if $\gamma = 0$, then the limiting distribution coincides with the limiting distribution of the Hotelling T^2 for comparisons of two p-dimensional vectors (which is a central chi-square with p degrees of freedom (Lehmann and Romano, 2005). The previous result is similar to what is obtained by Gretton et al. (2006) for the Maximum Mean Discrepancy test statistics (MMD), we obtain a weighted sum of chi-squared distributions with *summable* weights. For a given level $\alpha \in [0, 1]$, denote by $t_{1-\alpha}(\Sigma_W, \gamma)$ the $(1-\alpha)$-quantile of the distribution of $T_\infty(\Sigma_W, \gamma)$. Then, the sequence of test $\widehat{T}_n(\gamma) \geq t_{1-\alpha}(\Sigma_W, \gamma)$, is pointwise asymptotically level α to test homogeneity. Because in practice the covariance Σ_W is unknown, it is not possible to compute the quantile $t_{1-\alpha}(\Sigma_W, \gamma)$. Nevertheless, this quantile can still be consistently estimated by $t_{1-\alpha}(\hat{\Sigma}_W, \gamma)$, which can be obtained from the sample covariance matrix (see Proposition 24).

Corollary 2. *The test $\widehat{T}_n(\gamma) \geq t_{1-\alpha}(\hat{\Sigma}_W, \gamma)$ is pointwise asymptotically level α.*

In practice, the quantile $t_{1-\alpha}(\hat{\Sigma}_W, \gamma)$ can be numerically computed by inverse Laplace transform (see Strawderman, 2004; Hughett, 1998).

For all $\gamma > 0$, the weights $\{(\lambda_p + \gamma)^{-1}\lambda_p\}_{p\geq 1}$ are summable. However, if Assumption (B2) is satisfied, both $d_{1,n}(\gamma, \Sigma_W)$ and $d_{1,n}(\gamma, \Sigma_W)$ tend to infinity when $n \to 0$. The following theorem shows that if γ_n tends to zero slowly enough, then our test statistics is asymptotically normal:

Theorem 3. *Assume (A1), (B1-B2). Assume in addition that the probability distributions $\mathbb{P}_1$ and $\mathbb{P}_2$ are equal, i.e. $\mathbb{P}_1 = \mathbb{P}_2 = \mathbb{P}$ and that the sequence $\{\gamma_n\}$ is such that*

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

then

$$\widehat{T}_n(\gamma_n) \xrightarrow{\mathcal{D}} \mathcal{N}(0, 1)\ .$$

The proof of the theorem is postponed to Section 9. Under the assumptions of Theorem 3, the sequence of tests that rejects the null hypothesis when $\hat{T}_n(\gamma_n) \geq z_{1-\alpha}$, where $z_{1-\alpha}$ is the $(1-\alpha)$-quantile of the standard normal distribution, is asymptotically level α.

Contrary to the case where $\gamma_n \equiv \gamma$, the limiting distribution does not depend on the reproducing kernel, nor on the sequence of regularization parameters $\{\gamma\}_{n\geq 1}$. 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 $\{\gamma\}_{n\geq 1}$ goes to zero at a slower rate than $n^{-1/2}$. For instance, if the eigenvalues $\{\lambda_p\}_{p\geq 1}$ decrease at a polynomial rate, that is if there exists $s > 0$ such that we have $\lambda_p = p^{-s}$ for all $p \geq 1$, then, by Lemma 20, we have $d_1(\Sigma_W, \gamma_n) \sim \gamma_n^{-1/s}$ and $d_2(\Sigma_W, \gamma_n) \sim \gamma_n^{-1/2s}$ as $n \to \infty$. Therefore, the condition

$d_2^{-1}(\Sigma_W, \gamma_n) d_1(\Sigma_W, \gamma_n) \gamma_n^{-1} n^{-1/2} \to 0$ entails in this particular case that $\gamma_n^{-1} = o(n^{2s/1+4s})$, where the rate of decay s of the eigenvalues of the covariance operator Σ_W, depends *both on the kernel and the underlying distribution* $\mathbb{P}_1 = \mathbb{P}_2 = \mathbb{P}$. Besides, it may seem surprising that the limiting distribution is normal. This is due to two facts. First, we regularize the sample covariance operator prior to inversion (being of finite rank, the inverse of $\hat{\Sigma}$ is obviously not defined). Second, the problem is here truly infinite dimensional, because we have assumed that the eigenvalues are infinite dimensional $\lambda_p(\Sigma_W) > 0$ for all p (see Lehmann and Romano, 2005, Theorem 14.4.2, for a related result).

4.2 Limiting behavior against fixed alternatives

We study the power of the test based on $\widehat{T}_n(\gamma_n)$ under alternative hypotheses. The minimal requirement is to prove that this sequence of tests is consistent. A sequence of tests of constant level α is said to be *consistent in power* if the probability of accepting the null hypothesis of homogeneity goes to zero as the sample size goes to infinity under a *fixed* alternative. Recall that two probability $\mathbb{P}_1$ and $\mathbb{P}_2$ defined on a measurable space $(\mathcal{X}, \mathfrak{X})$ are called *singular* if there exist two disjoint sets A and B in $\mathfrak{X}$ whose union is $\mathcal{X}$ such that $\mathbb{P}_1$ is zero on all measurable subsets of B while $\mathbb{P}_2$ is zero on all measurable subsets of B. This is denoted by $\mathbb{P}_1 \perp \mathbb{P}_2$.

When $\gamma_n \equiv \gamma$ or when $\gamma_n \to 0$, and $\mathbb{P}_1$ and $\mathbb{P}_2$ are not singular, then the following proposition shows that the limits in both cases are finite, strictly positive and independent of the kernel otherwise (see Fukumizu et al., 2008, for similar results for canonical correlation analysis). The following result gives some useful insights on $\|\Sigma_W^{-1/2}(\mu_2 - \mu_1)\|_{\mathcal{H}}$, the population counterpart of $\|(\hat{\Sigma}_W + \gamma_n \mathrm{I})^{-1/2}(\hat{\mu}_2 - \hat{\mu}_1)\|_{\mathcal{H}}$ on which our test statistics is based upon.

Proposition 4. *Assume (A1-A2). Let ν a measure dominating $\mathbb{P}_1$ and $\mathbb{P}_2$, and let p_1 and p_2 the densities of $\mathbb{P}_1$ and $\mathbb{P}_2$ with respect to ν. The norm $\|\Sigma_W^{-1/2}(\mu_2 - \mu_1)\|_{\mathcal{H}}$ is infinite if and only if $\mathbb{P}_1$ and $\mathbb{P}_2$ are mutually singular. If $\mathbb{P}_1$ and $\mathbb{P}_2$ are nonsingular, $\|\Sigma_W^{-1/2}(\mu_2 - \mu_1)\|_{\mathcal{H}}$ is finite and is given by*

$$\left\|\Sigma_W^{-1/2}(\mu_2 - \mu_1)\right\|_{\mathcal{H}}^2 = \frac{1}{\rho_1 \rho_2}\left(1 - \int \frac{p_1 p_2}{\rho_1 p_1 + \rho_2 p_2} d\nu\right)\left(\int \frac{p_1 p_2}{\rho_1 p_1 + \rho_2 p_2} d\nu\right)^{-1}.$$

It is equal to zero if the χ^2-divergence is null, that is, if and only if $\mathbb{P}_1 = \mathbb{P}_2$.

By combining the two previous propositions, we therefore obtain the following consistency theorem:

Theorem 5. *Assume (A1-A2). Let $\mathbb{P}_1$ and $\mathbb{P}_2$ be two distributions over $(\mathcal{X}, \mathfrak{X})$, such that $\mathbb{P}_2 \neq \mathbb{P}_1$. If either $\gamma_n \equiv \gamma$ or $\gamma_n + d_2^{-1}(\Sigma_1, \gamma_n) d_1(\Sigma_1, \gamma_n) \gamma_n^{-1} n^{-1/2} \to 0$, then for any $t > 0$*

$$\mathbb{P}_{\boldsymbol{H}_A}(\widehat{T}_n(\gamma) > t) \to 1\,. \tag{11}$$

4.3 Limiting distribution against local alternatives

When the alternative is fixed, any sensible test procedure will have a power that tends to one as the sample size n tends to infinity. This property is not suitable for comparing the

limiting power of different test procedures. Several approaches are possible to answer this question. One such approach is to consider sequences of *local alternatives* (Lehmann and Romano, 2005). Such alternatives tend to the null hypothesis as $n \to \infty$ at a rate which is such that the limiting distribution of sequence the test statistics under the sequence of alternatives converge to a non-degenerate random variable. To compare two sequences of tests for a given sequence of alternatives, one may then compute the ratio of the limiting powers, and choose the test which has the largest power.

In our setting, let $\mathbb{P}_1$ denote a fixed probability on $(\mathcal{X}, \mathfrak{X})$ and let $\mathbb{P}_2^n$ be a sequence of probability on $(\mathcal{X}, \mathfrak{X})$. The sequence $\mathbb{P}_2^n$ depends on the sample size n and converge to $\mathbb{P}_1$ as n goes to infinity with respect to a certain distance. In the asymptotic analysis of our test statistics against sequences of local alternatives, the χ^2-divergence $D_{\chi^2}\left(\mathbb{P}_1 \parallel \mathbb{P}_2^n\right)$ is defined for all n as

$$\eta_n \stackrel{\text{def}}{=} \left\| \frac{d\mathbb{P}_2^n}{d\mathbb{P}_1} - 1 \right\|_{L^2(\mathbb{P}_1)} , \tag{12}$$

for $\mathbb{P}_2^n$ absolutely continuous with respect to $\mathbb{P}_1$. Therefore, in the subsequent sections, we shall make the following assumption:

(**C**) For any n, $\mathbb{P}_2^n$ is absolutely continuous with respect to $\mathbb{P}_1$, and $D_{\chi^2}\left(\mathbb{P}_1 \parallel \mathbb{P}_2^n\right) \to 0$ as n tends to infinity.

The following theorem shows that under local alternatives, we get a series of shift in the chi-squared distributions when $\gamma_n \equiv \gamma$:

Theorem 6. *Assume (A1), (B1), and (C). Assume in addition $\gamma_n \equiv \gamma > 0$ and that $n\eta_n^2 = O(1)$, then*

$$\widehat{T}_n(\gamma) \xrightarrow{\mathcal{D}} 2^{-1/2} d_2^{-1}(\Sigma_1, \gamma) \sum_{p=1}^{\infty} (\lambda_p(\Sigma_1) + \gamma)^{-1} \lambda_p(\Sigma_1) \{(Z_p + a_{n,p}(\gamma))^2 - 1\} ,$$

with

$$a_{n,p}(\gamma) = (n_1 n_2/n)^{1/2} \left\langle (\Sigma_1 + \gamma \mathrm{I})^{-1/2} (\mu_2^n - \mu_1), e_p \right\rangle_{\mathcal{H}} , \tag{13}$$

where $\{Z_p\}_{p \geq 1}$ are independent standard normal random variables, defined on a common probability space.

When the sequence of regularization parameters $\{\gamma_n\}_{n \geq 1}$ tends to zero at a slower rate than $n^{-1/2}$, the test statistics is shown to be asymptotically normal, with the same limiting variance as the one under the null hypothesis, but with a non-zero limiting mean, as detailed in the next two results. While the former states the asymptotic normality under general conditions, the latter highlights the fact that the asymptotic mean-shift in the limiting distribution may be conveniently expressed from the limiting χ^2-divergence of $\mathbb{P}_1$ and $\mathbb{P}_2^n$ under additional smoothness assumptions on the spectrum of the covariance operator.

Theorem 7. *Assume (A1), and (B1-2), and (C). Let $\{\gamma_n\}_{n \geq 1}$ be a sequence such that*

$$\gamma_n + d_2^{-1}(\Sigma_1, \gamma_n) d_1(\Sigma_1, \gamma_n) \gamma_n^{-1} n^{-1/2} \to 0 \tag{14}$$

$$d_2^{-1}(\Sigma_1, \gamma_n) n\eta_n^2 = O(1) \quad \textit{and} \quad d_2^{-1}(\Sigma_1, \gamma_n) d_1(\Sigma_1, \gamma_n) \eta_n \to 0 , \tag{15}$$

where $\{\eta_n\}_{n\geq 1}$ *is defined in (12). If the following limit exists,*

$$\Delta \stackrel{\text{def}}{=} \lim_{n\to\infty} \frac{n \left\|(\Sigma_1 + \gamma_n I)^{-1/2}(\mu_2^n - \mu_1)\right\|_{\mathcal{H}}^2}{d_2(\Sigma_1, \gamma_n)} , \tag{16}$$

then,

$$\widehat{T}_n(\gamma_n) \xrightarrow{\mathcal{D}} \mathcal{N}(\rho_1\rho_2\Delta, 1) .$$

Corollary 8. *Under the assumptions of Theorem 7, if there exists* $a > 0$ *such that*

$$\left\langle \mu_2^n - \mu_1, \Sigma_1^{-1-a}(\mu_2^n - \mu_1)\right\rangle_{\mathcal{H}} < \infty ,$$

and if the following limit exists,

$$\Delta = \lim_{n\to\infty} d_2(\Sigma_1, \gamma_n)^{-1} n\eta_n^2 ,$$

then,

$$\widehat{T}_n(\gamma_n) \xrightarrow{\mathcal{D}} \mathcal{N}(\rho_1\rho_2\Delta, 1) .$$

It is worthwhile to note that $\rho_1\rho_2\Delta$, the limiting mean-shift of our test statistics against sequences of local alternatives does not depend on the choice of the reproducing kernel. This means that, at least in the large-sample setting $n \to \infty$, the choice of the kernel is irrelevant, provided that for some $a > 0$ we have $\left\langle \mu_2^n - \mu_1, \Sigma_1^{-1-a}(\mu_2^n - \mu_1)\right\rangle_{\mathcal{H}} < \infty$. Then, we get that the sequences of local alternatives converge to the null at rate $\eta_n = C\, d_2^{1/2}(\Sigma_1, \gamma_n)n^{-1/2}$ for some constant $C > 0$, which is slower than the usual parametric rate $n^{-1/2}$ since $d_2(\Sigma_1, \gamma_n) \to \infty$ as $n \to \infty$ as shown in Lemma 18. Note also that conditions of the form $\left\langle \mu_2^n - \mu_1, \Sigma_1^{-1-\alpha}(\mu_2^n - \mu_1)\right\rangle_{\mathcal{H}} < \infty$ imply that the sequence of local alternatives are limited to *smooth enough* densities p_2^n around p_1.

5. Discussion

We illustrate now the behaviour of the limiting power of our test statistics against two different types of sequences of local alternatives. Then, we compare the power of our test statistics against the power of the Maximum Mean Discrepancy test statistics proposed by Gretton et al. (2006). Finally, we highlights some links between testing for homogeneity and supervised binary classification.

5.1 Limiting power against local alternatives of KFDA

We have seen that our test statistics is consistent in power against fixed alternatives, for both regularization schemes $\gamma_n \equiv \gamma$ and $\gamma_n \to 0$. We shall now examine the behaviour of the power of our test statistics, against different types of sequences of local alternatives: i) directional alternatives, ii) non-directional alternatives. For this purpose, we consider a specific reproducing kernel, the periodic spline kernel, whose derivation is given below. Indeed, when $\mathbb{P}_1$ is the uniform distribution on $[0, 1]$, and $d\mathbb{P}_2/d\mathbb{P}_1 = 1 + \eta c_q$ with c_q is a one-component contamination on the Fourier basis, we may conveniently compute a closed-form equivalent when $n \to \infty$ of the eigenvalues of the covariance operator Σ_1, and therefore the power function of the test statistics.

Periodic spline kernel The periodic spline kernel, described in (Wahba, 1990, Chapter 2), is defined as follows. Any function f in $L^2(\mathcal{X})$, where $\mathcal{X}$ is taken as the torus $\mathbb{R}/2\pi\mathbb{Z}$, may be expressed in the form of a Fourier series expansion $f(t) = \sum_{p=0}^{\infty} a_p c_p(t)$ where $\sum_{p=0}^{\infty} a_p^2$, and for all $\ell \geq 1$

$$c_0(t) = \mathbf{1}_{\mathcal{X}} \tag{17}$$

$$c_{2\ell-1}(t) = \sqrt{2}\ \sin(2\pi(2\ell-1)t) \tag{18}$$

$$c_{2\ell}(t) = \sqrt{2}\ \cos(2\pi(2\ell-1)t)\ . \tag{19}$$

Let us consider the family of RKHS defined by $\mathcal{H}^m = \{f :\ f \in L^2(\mathcal{X}),\ \sum_{p=0}^{\infty} \lambda_p^{-1} a_p^2 < \infty\}$ with $m > 1$, where $\lambda_p = (2\pi p)^{-2m}$ for all $p \geq 1$, whose norm is defined for all $f \in L^2(\mathcal{X})$ as

$$\|f\|_{\mathcal{H}}^2 = 1/2\ \sum_{p=0}^{\infty} (2\pi p)^{-2m} a_p^2\ . \tag{20}$$

Therefore, the associated reproducing kernel $k(x, y)$ writes as

$$k_m(x,y) = 2\sum_{p=0}^{\infty} (2\pi p)^{-2m} c_p(x-y) = \frac{(-1)^{m-1}}{(2m)!} B_{2m}\left((x-y) - \lfloor x-y \rfloor\right)\ ,$$

where B_{2m} is the $2m$-th Bernoulli polynomial.

The set $\{e_p(t), p \geq 1\}$ is actually an orthonormal basis of $\mathcal{H}$, where $e_p(t) \stackrel{\text{def}}{=} \lambda_p^{1/2} c_p(t)$ for all $p \geq 1$. Let us consider $\mathbb{P}_1$ the uniform probability measure on $[0,1]$. We have $e_p - \mathbb{E}_{\mathbb{P}_1}[e_p] \equiv e_p$ and $\mu_1 \equiv 0$, where μ_1 is the mean element associated with $\mathbb{P}_1$. Hence, $\{(\lambda_p, e_p(t)), p \geq 1\}$ is an eigenbasis of Σ_1 the covariance operator associated with $\mathbb{P}_1$, where for all $\ell \geq 1$

$$\lambda_0 = 1 \tag{21}$$

$$\lambda_{2\ell-1} = (4\pi\ell)^{-2m} \tag{22}$$

$$\lambda_{2\ell} = (4\pi\ell)^{-2m}\ . \tag{23}$$

Note that the parameter m characterizes the RKHS $\mathcal{H}^m$ and its associated reproducing kernel $k_m(\cdot,\cdot)$, and therefore controls the rate of decay of the eigenvalues of the covariance operator Σ_1. Indeed, by Lemma 20, we have $d_1(\Sigma_1, \gamma_n) = C_1\ \gamma_n^{-1/2m}$ and $d_2(\Sigma_1, \gamma_n) = C_2\ \gamma_n^{-1/4m}$ for some constants $C_1, C_2 > 0$ as $n \to \infty$.

Directional alternatives Let us consider the limiting power of our test statistics in the following setting:

$$\mathbf{H}_0 :\ \mathbb{P}_1 = \mathbb{P}_2^n \quad \text{against} \quad \mathbf{H}_A^n :\ \mathbb{P}_1 \neq \mathbb{P}_2^n,\ \text{ with } \mathbb{P}_2^n \text{ such that } d\mathbb{P}_2^n/d\mathbb{P}_1 = 1 + An^{-1/2} c_q\ , \tag{24}$$

where $\mathbb{P}_1$ is the uniform probability measure on $[0,1]$, and $c_q(t)$ is defined in (17). In the case $\gamma_n \equiv \gamma$, given a significance level $\alpha \in (0,1)$, the associated critical level $t_{1-\alpha}$ is defined as satisfying

$$\mathbb{P}\left(2^{-1/2} d_2^{-1}(\Sigma_1, \gamma) \sum_{p=1}^{\infty} (\lambda_p(\Sigma_1) + \gamma)^{-1} \lambda_p(\Sigma_1)\{Z_p^2 - 1\} > t_{1-\alpha}\right) = \alpha\ .$$

Note that $a_{n,p}(\gamma) = 0$ for all $p \geq 1$ (from Theorem 6) except for $p = q$ where

$$a_{n,q}(\gamma) = \sqrt{A}\,\sqrt{n_1 n_2/n^2}\,(\lambda_q + \gamma)^{-1/2}\lambda_q^{1/2}\,.$$

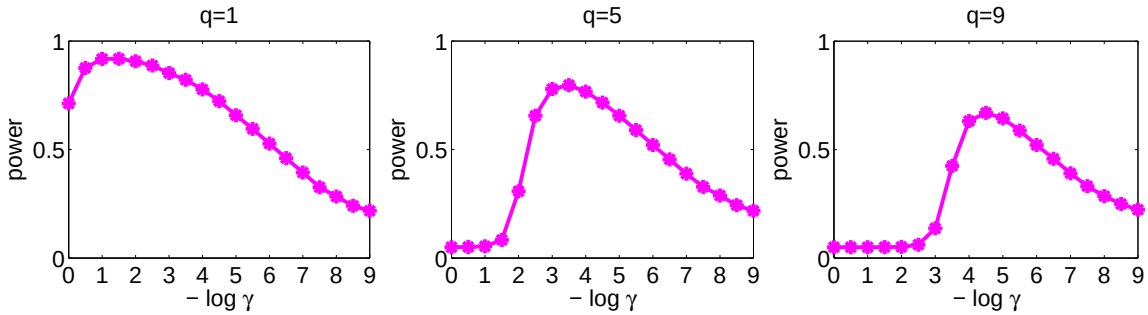

Figure 1: Evolution of power of KFDA as $\gamma = 1, 10^{-1}, \ldots, 10^{-9}$, for q-th component alternatives with (from left to right) with $q = 1, 5, 9$.

In order to analyze the behaviour of the power for varying values of γ and for different values of q, we compute the limiting power, when taking $m = 2$ in the periodic reproducing kernel, and for $q = 1, 5, 9$, and investigate the evolution of the power as a function of the regularization parameter γ. As Figure 1 shows, our test statistics has trivial power, that is equal to α, when $\gamma \gg \lambda_q$, while it reaches stricly nontrivial power as long as $\gamma \leq \lambda_q$. This motivates the study of the decaying regularization scheme $\gamma_n \to 0$ of our test statistics, in order to incorporate the $\gamma \to 0$ into our large-sample framework. In the next paragraph, we shall demonstrate that the version of our test statistics with decaying regularization parameter $\gamma_n \to 0$ reaches high power against a broader class of local alternatives, which we call *non-directional alternatives*, where $q \equiv q_n \to \infty$, as opposed to *directional alternatives* where q was kept constant. Yet, for having nontrivial power with the test statistics $\widehat{T}(\gamma_n)$ against such sequences of local alternatives, the non-directional sequence of local alternatives have to converge to the null at a slower rate than $\sqrt{n}$.

Non-directional alternatives Now, we consider the limiting power of our test statistics in the following setting:

$$\mathbf{H}_0 : \mathbb{P}_1 = \mathbb{P}_2^n \quad \text{against} \quad \mathbf{H}_A^n : \mathbb{P}_1 \neq \mathbb{P}_2^n, \ \text{ with } \mathbb{P}_2^n \text{ such that } d\mathbb{P}_2^n/d\mathbb{P}_1 = 1 + \eta_n c_{q_n}\,, \quad (25)$$

Assume $\mathbb{P}_1$ is the uniform probability measure on $[0,1]$, and consider again the periodic spline kernel of order $2m$. Take $\{q_n\}_{n\geq 1}$ a nonnegative nondecreasing sequence of integers. Now, if the sequence of local alternatives is converging to the null at rate $\eta_n = (2\Delta)^{1/2} q_n^{1/4} n^{-1/2}$ for some $\Delta > 0$, with $q_n = o(n^{1/1+4m})$ for our asymptotic analysis to hold, then as long as $\gamma_n \equiv \lambda_{q_n} = q_n^{-2m}$ we have

$$\begin{aligned}\lim_{n\to\infty} \mathbb{P}_{\mathbf{H}_A^n}\left(\widehat{T}_n(\gamma_n) > z_{1-\alpha}\right) &= \mathbb{P}_{\mathbf{H}_A^n}\left(Z_1 + \rho_1\rho_2\Delta > z_{1-\alpha}\right) \\ &= 1 - \Phi\left[z_{1-\alpha} - \rho_1\rho_2\Delta\right]\,.\end{aligned}$$

	$\lambda_p(\Sigma)$	$d_1(\Sigma,\gamma)$	$d_2(\Sigma,\gamma)$
Normal tails	$O\left(\exp(-cp^{1/d})\right)$	$O\left(\log^d(1/\gamma)\right)$	$O\left(\log^{d/2}(1/\gamma)\right)$
Polynomial tails	$O\left(p^{-\beta}\right)$ for any $\beta > \alpha$	$O\left(\gamma^{-1/\beta}\right)$	$O\left(\gamma^{-1/2\beta}\right)$

Table 1: examples of rate of convergence for the gaussian kernels for $\mathcal{X} = \mathbb{R}^p$

where we used Lemma 20 together with Theorem 7. On the other hand, if $\gamma_n^{-1} q_n^{-2m} = o(1)$, then the limiting power is trivial and equal to α.

Back to the fixed-regularization test statistics $\widehat{T}_n(\gamma)$, we may also compute the limiting power of $\widehat{T}_n(\gamma)$ against the non-directional sequence of local alternatives defined in (25) by taking into account Remark 17 to use Theorem 6. Indeed, as n tends to infinity, since $a_{n,q_n}(\gamma) = (\rho_1\rho_2)^{1/2}(\lambda_{q_n} + \gamma)^{-1/2}\lambda_{q_n}\eta_n$, then the fixed-regularization version $\widehat{T}_n(\gamma)$ of the test statistics has trivial power against non-directional alternatives.

Remark 9. *We analyzed the limiting power of our test statistics in the specific case where $\mathbb{P}_1$ is the uniform distribution on $[0,1]$ and the reproducing kernel belongs to the family of periodic spline kernels. Yet, our findings carry over more general settings as illustrated by Table 1. Indeed, for general distributions with polynomial decay in the tail and (nonperiodic) gaussian kernels, the eigenvalues of the covariance operator still exhibit similar behaviour as in the example treated above.*

We now discuss the links between our procedure with the previously proposed Maximum Mean Discrepancy (MMD) test statistics. We also highlight interesting links with supervised kernel-based classification.

5.2 Comparison with Maximum Mean Discrepancy

Our test statistics share many similarities with the Maximum Mean Discrepancy test statistics of Gretton et al. (2006). In the case $\gamma_n \equiv \gamma$, both have limiting null distribution which may be expressed as an infinite weighted mixture of chi-squared random variables. Yet, while $\widehat{T}_n^{\mathrm{MMD}} \xrightarrow{\mathcal{D}} C\sum_{p=1}^{\infty}\lambda_p(Z_p^2 - 1)$ where $\widehat{T}_n^{\mathrm{MMD}}$ denotes the test statistics used by MMD, we have in our case $\widehat{T}_n^{\mathsf{KFDA}}(\gamma_n) \xrightarrow{\mathcal{D}} C\sum_{p=1}^{\infty}(\lambda_p + \gamma_n)^{-1}\lambda_p(Z_p^2 - 1)$. Roughly speaking, the test statistics based on KFDA uniformly weights the components associated with the first eigenvalues of the covariance operator, and downweights the remaining ones, which allows to gain greater power for testing by focusing on the user-tunable number of components of the covariance operator. On the other hand, the test statistics based on MMD is naturally sensitive to differences lying on the first components, and gets progressively less sensitive to differences in higher components. Thus, our test statistics based on KFDA allows to give equal weights to differences lying in (almost) all components, the effective number of components on the which the test statistics focus on being tuned *via* the regularization parameter γ_n. These differences may be illustrated by considering the behavuour of MMD against sequences of local alternatives respectively with fixed-frequency and non-directional, for periodic kernels.

Directional alternatives Let us consider the setting defined in (24). By a similar reasoning, we may also compute the limiting power of $\widehat{T}_n^{\mathrm{MMD}}$ against directional sequences of

local alternatives, with a periodic spline kernel of order $m = 2$, for different components $q = 1, 5, 9$. Both test statistics KFDA and MMD reach high power when the sequences of local alternatives lies on the first component. However, the power of MMD tumbles down for higher-order alternatives whereas the power of KFDA remains strictly nontrivial for high-order alternatives as long as γ is sufficiently small.

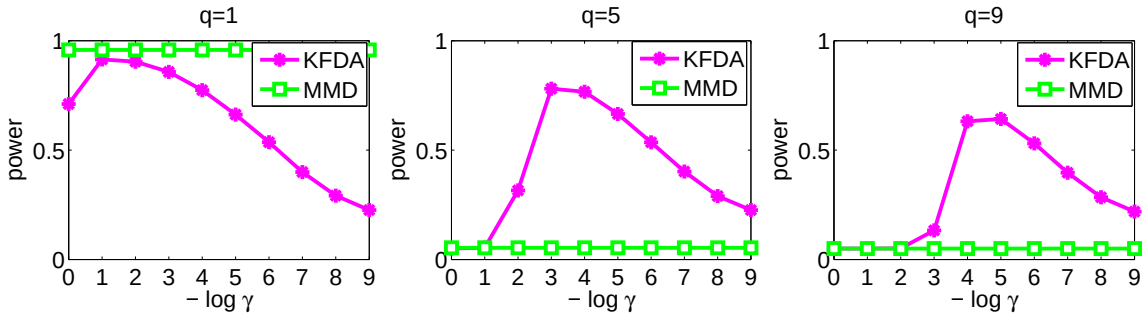

Figure 2: Comparison of the evolution of power of KFDA versus the power of MMD as $\gamma = 1, 10^{-1}, \ldots, 10^{-9}$, for q-th component alternatives with (from left to right) with $q = 1, 5, 9$.

Non-directional alternatives Now, consider sequences of local alternatives as defined in (25). The test statistics MMD does not notice such alternatives. Therefore, MMD has trivial power equal to α against non-directional alternatives.

5.3 Links with supervised classification

When the sample sizes of each sample are equal, that is when $n_1 = n_2$, KFDA is known to be equivalent to Kernel Ridge Regression (KRR), also referred to as smoothing spline regression in statistics. In this case, KRR performs a kernel-based least-square regression fit on the labels, where the samples are respectively labelled -1 and $+1$. The recentering parameter $d_1(\Sigma_1, \gamma_n)$ in our procedure coincides with the so-called *degrees of freedom* in smoothing spline regression, which were often advocated to provide a relevant measure of complexity for model selection (see Efron, 2004). In particular, since the mean-shift in the limiting normal distribution against local alternatives is lower-bounded by $n d_1^{-1}(\Sigma_1, \gamma_n) \langle (\mu_2 - \mu_1), (\Sigma_1 + \gamma_n I)^{-1}(\mu_2 - \mu_1) \rangle$, this suggests an algorithm for selecting γ_n and the kernel. For a fixed degree of freedom $d_1(\Sigma_1, \gamma_n)$, maximizing the asymptotic mean-shift (which corresponds to the class separation) is likely to yield greater power. As future work, we plan to investigate, both theoretically and practically, the use of (single and multiple) kernel learning procedures as developed by Bach et al. (2004) for maximizing the expected power of our test statistics in specific applications.

6. Experiments

In this section, we investigate the experimental performances of our test statistic KFDA, and compare it in terms of power against other nonparametric test statistics.

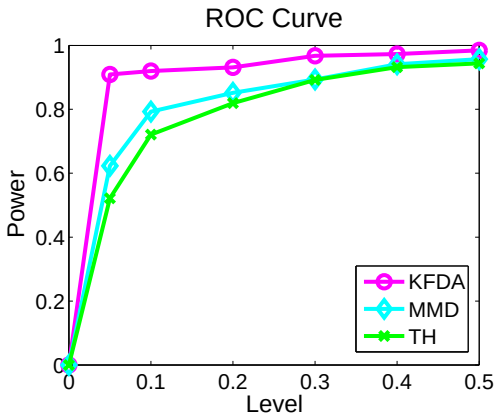

Figure 3: Comparison of ROC curves in a speaker verification task

6.1 Speaker verification

We conducted experiments in a speaker verification task Bimbot et al. (2004), on a subset of 8 female speakers using data from the NIST 2004 Speaker Recognition Evaluation. We refer the reader to (Louradour et al., 2007) for instance for details on the pre-processing of data. The figure shows averaged results over all couples of speakers. For each couple of speaker, at each run we took 3000 samples of each speaker and launched our KFDA-test to decide whether samples come from the same speaker or not, and computed the type II error by comparing the prediction to ground truth. We averaged the results for 100 runs for each couple, and all couples of speaker. The level was set to $\alpha = 0.05$, and the critical values were computed by a bootstrap resampling procedure. Since the observations may be considered dependent within the sequences, and independent between the sequences, we used a fixed-block variant of the boostrap, which consists in using boostrap samples built by piecing together several boostrap samples drawn in each sequence. We performed the same experiments for the Maximum Mean Discrepancy and the Tajvidi-Hall test statistic (TH). We summed up the results by plotting the ROC-curve for all competing methods. Our method reaches good empirical power for a small value of the prescribed level ($1 - \beta = 90\%$ for $\alpha = 0.05\%$). Maximum Mean Discrepancy also yields good empirical performance on this task.

7. Conclusion

We proposed a well-calibrated kernel-based test statistic for testing the homogeneity of two samples, built on the kernel Fisher discriminant analysis algorithm, for which we proved that the asymptotic limit distribution under null hypothesis is standard normal distribution when de regularization parameter decays to zero at a slower rate than $n^{-1/2}$. Besides, our test statistic can be readily computed from Gram matrices once a reproducing kernel is defined, and reaches nontrivial power aqgainst a large class of alternatives under mild conditions on the regularization parameter. Finally, our KFDA-test statistic yields competitive performance for speaker identification purposes.

8. Proof of some preliminary results

We preface the proof by some useful results relating the KFDA statistics to kernel independent quantities.

Proposition 10. *Assume (A1)-(A2). Let $\mathbb{P}_1$ and $\mathbb{P}_2$ be two probability distributions on $(\mathcal{X}, \mathfrak{X})$, and denote by μ_1, μ_2 the associated mean (see (1)). Let $\mathbb{Q}$ be a probability dominating $\mathbb{P}_1$ and $\mathbb{P}_2$, and let Σ be the associated covariance operator (see (2)). Then,*

$$\left\| \frac{d\mathbb{P}_1}{d\mathbb{Q}} - \frac{d\mathbb{P}_2}{d\mathbb{Q}} \right\|_{L^2(\mathbb{Q})} < \infty,$$

if and only if the vector $(\mu_2 - \mu_1) \in \mathcal{H}$ belongs to the range of the square root $\Sigma^{1/2}$. In addition,

$$\left\langle \mu_2 - \mu_1, \Sigma^{-1}(\mu_2 - \mu_1) \right\rangle_{\mathcal{H}} = \left\| \frac{d\mathbb{P}_1}{d\mathbb{Q}} - \frac{d\mathbb{P}_2}{d\mathbb{Q}} \right\|^2_{L^2(\mathbb{Q})} . \tag{26}$$

Proof. Denote by $\{\lambda_k\}_{k\geq 1}$ and $\{e_k\}_{k\geq 1}$ the strictly positive eigenvalues and the corresponding eigenvectors of the covariance operator Σ, respectively. For $k \geq 1$, set

$$f_k = \lambda_k^{-1/2} \{e_k - \mathbb{Q}e_k\} . \tag{27}$$

By construction, for any $k, \ell \geq 1$,

$$\lambda_k \delta_{k,\ell} = \langle e_k, \Sigma e_\ell \rangle_{\mathcal{H}} = \langle e_k - \mathbb{Q}e_k, e_\ell - \mathbb{Q}e_\ell \rangle_{L^2(\mathbb{Q})} = \lambda_k^{1/2} \lambda_\ell^{1/2} \langle f_k, f_\ell \rangle_{L^2(\mathbb{Q})} ,$$

where $\delta_{k,\ell}$ is Kronecker's delta. Hence $\{f_k\}_{k\geq 1}$ is an orthonormal system of $L^2(\mathbb{Q})$. Note that $\mu_2 - \mu_1$ belongs to the range of $\Sigma^{1/2}$ if and only if

(a) $\langle \mu_2 - \mu_1, g \rangle_{\mathcal{H}} = 0$ for all g in the null space of Σ,

(b) $\left\langle \mu_1 - \mu_2, \Sigma^{-1}(\mu_1 - \mu_2) \right\rangle_{\mathcal{H}} \stackrel{\text{def}}{=} \sum_{p=1}^{\infty} \lambda_p^{-1} \langle e_p, (\mu_1 - \mu_2) \rangle^2_{\mathcal{H}} < \infty$.

Consider first condition (a). For any $g \in \mathcal{H}$, il follows from the definitions that

$$\begin{aligned} \langle \mu_2 - \mu_1, g \rangle_{\mathcal{H}} = \int (d\mathbb{P}_1 - d\mathbb{P}_2)\, g = \int (d\mathbb{P}_1 - d\mathbb{P}_2)\, (g - \mathbb{Q}g) \\ = \left\langle \frac{d\mathbb{P}_1}{d\mathbb{Q}} - \frac{d\mathbb{P}_2}{d\mathbb{Q}}, g - \mathbb{Q}g \right\rangle_{L^2(\mathbb{Q})} . \end{aligned}$$

If g belongs to the null space of Σ, then $\|g - \mathbb{Q}g\|_{L^2(\mathbb{Q})} = 0$, and the previous relation implies that $\langle \mu_2 - \mu_1, g \rangle_{\mathcal{H}} = 0$. Consider now (b).

$$\begin{aligned} \sum_{p=1}^{\infty} \lambda_p^{-1} \langle e_p, (\mu_1 - \mu_2) \rangle^2_{\mathcal{H}} &= \sum_{p=1}^{\infty} \lambda_p^{-1} \left(\int \{d\mathbb{P}_1(x) - d\mathbb{P}_2(x)\} e_p(x) \right)^2 \\ &= \sum_{p=1}^{\infty} \left\langle \frac{d\mathbb{P}_1}{d\mathbb{Q}} - \frac{d\mathbb{P}_2}{d\mathbb{Q}}, f_p \right\rangle^2_{L^2(\mathbb{Q})} \leq \left\| \frac{d\mathbb{P}_1}{d\mathbb{Q}} - \frac{d\mathbb{P}_2}{d\mathbb{Q}} \right\|^2_{L^2(\mathbb{Q})} . \end{aligned} \tag{28}$$

In order to prove the equality, we simply notice that because of the density of the RKHS in $L^2(\mathbb{Q})$, then $\{f_k\}_{k\geq 1}$ is a complete orthonormal basis of the space of functions $L_0^2(\mathbb{Q})$, defined as

$$L_0^2(\mathbb{Q}) \stackrel{\text{def}}{=} \left\{ g \in L^2(\mathbb{Q}) \, , \int (g - \mathbb{Q}g)^2 d\mathbb{Q} > 0 \quad \text{and} \quad \mathbb{Q}g = 0 \right\} . \tag{29}$$

□

Lemma 11. *Assume (A1)-(A2). Let $\mathbb{P}_1$ and $\mathbb{P}_2$ two probability distributions on $(\mathcal{X}, \mathfrak{X})$ such that $\mathbb{P}_2 \ll \mathbb{P}_1$.*

Denote by Σ_1 and Σ_2 the associated covariance operators. Then, for any $\gamma > 0$,

$$\left\| \mathrm{I} - \Sigma_1^{-1/2} \Sigma_2^n \Sigma_1^{-1/2} \right\|_{\mathrm{HS}}^2 \leq 4 \left\| \frac{d\mathbb{P}_2}{d\mathbb{P}_1} - 1 \right\|_{L^2(\mathbb{P}_1)}^2 , \tag{30}$$

$$\left| \mathrm{Tr}\{(\Sigma_1 + \gamma \mathrm{I})^{-1} (\Sigma_2 - \Sigma_1)\} \right| \leq 2 d_2(\Sigma_1, \gamma) \left\| \frac{d\mathbb{P}_2}{d\mathbb{P}_1} - 1 \right\|_{L^2(\mathbb{P}_1)} . \tag{31}$$

where $d_2(\Sigma_1, \gamma)$ is defined in (6).

Proof. Denote by $\{\lambda_k\}_{k\geq 1}$ and $\{e_k\}_{k\geq 1}$ the strictly positive eigenvalues and the corresponding eigenvectors of the covariance operator Σ_1. Note that $\langle e_k, \Sigma_1 e_\ell \rangle = \lambda_k \delta_{k,\ell}$ for all k and ℓ. Let us denote $f_k = \lambda_k^{-1/2} \{e_k - \mathbb{P}_1 e_k\}$. Then, we have $\langle f_k, f_\ell \rangle_{L^2(\mathbb{P}_1)} = \delta_{k,\ell}$. Note that

$$\begin{aligned}
&\sum_{k,\ell=1}^{\infty} \left\{ \delta_{k,\ell} - \lambda_k^{-1/2} \lambda_\ell^{-1/2} \langle e_k, \Sigma_2 e_\ell \rangle_{\mathcal{H}} \right\}^2 \\
&= \sum_{k,\ell=1}^{\infty} \left\{ \left\langle f_k, \left(1 - \frac{d\mathbb{P}_2}{d\mathbb{P}_1} \right) f_\ell \right\rangle_{L^2(\mathbb{P}_1)} + \lambda_k^{-1/2} \lambda_\ell^{-1/2} \langle \mu_2 - \mu_1, e_k \rangle_{\mathcal{H}} \langle \mu_2 - \mu_1, e_\ell \rangle_{\mathcal{H}} \right\}^2 .
\end{aligned}$$

Then, using that $(a+b)^2 \leq 2(a^2 + b^2)$, and (28) in Proposition 10 with $\Sigma = \Sigma_1$, we obtain

$$\left\| \mathrm{I} - \Sigma_1^{-1/2} \Sigma_2^n \Sigma_1^{-1/2} \right\|_{\mathrm{HS}}^2 \leq 4 \left\| \frac{d\mathbb{P}_2}{d\mathbb{P}_1} - 1 \right\|_{L^2(\mathbb{P}_1)}^2 . \tag{32}$$

Denote, for all $p, q \geq 1$

$$\varepsilon_{p,q} \stackrel{\text{def}}{=} \left\langle e_p, (\Sigma_1^{-1/2} \Sigma_2 \Sigma_1^{-1/2} - \mathrm{I}) e_q \right\rangle . \tag{33}$$

By applying the Hölder inequality, and using (30), we get

$$\begin{aligned}
\left| \mathrm{Tr}\{(\Sigma_1 + \gamma \mathrm{I})^{-1} (\Sigma_2 - \Sigma_1)\} \right| &= \sum_{p=1}^{\infty} \left| \left\langle e_p, (\Sigma_1 + \gamma \mathrm{I})^{-1} \Sigma_1 e_p \right\rangle \varepsilon_{p,p} \right| \\
&\leq \left(\sum_{p=1}^{\infty} \left\langle e_p, (\Sigma_1 + \gamma \mathrm{I})^{-1} \Sigma_1 e_p \right\rangle^2 \right)^{1/2} \left(\sum_{p=1}^{\infty} \varepsilon_{p,p}^2 \right)^{1/2} \leq 2 d_2(\Sigma_1, \gamma) \left\| \frac{d\mathbb{P}_2}{d\mathbb{P}_1} - 1 \right\|_{L^2(\mathbb{P}_1)} ,
\end{aligned}$$

which completes the proof of (31). □

Proposition 12. *Assume (A1). Let $\{X_1^n, \ldots, X_n^n\}$ 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 $\lambda_p(\Sigma^n)$ of Σ^n are non-negative, and if there exists $C > 0$ such that for all n we have $\sum_{p=1}^{\infty} \lambda_p^{1/2}(\Sigma^n) < C$, then $\sum_{p=1}^{\infty} |\lambda_p(\hat{\Sigma} - \Sigma^n)| = O_P(n^{-1/2})$.*

Proof. Lemma 21 shows that, for any orthonormal basis $\{e_p\}_{p\geq 1}$ in the RKHS $\mathcal{H}$:

$$\sum_{p=1}^{\infty} |\lambda_p(\hat{\Sigma} - \Sigma^n)| \leq \sum_{p=1}^{\infty} \left\| (\hat{\Sigma} - \Sigma^n) e_p \right\|_{\mathcal{H}} .$$

We take the orthonormal family of eigenvectors $\{e_p\}_{p\geq 1}$ of the covariance operator Σ^n (associated to the eigenvalues $\lambda_p(\Sigma^n)$ ranked in decreasing order). Then, it suffices to show that $\sum_{p=1}^{\infty} \left\| (\hat{\Sigma} - \Sigma^n) e_p \right\|_{\mathcal{H}} = O_P(n^{-1/2})$. Note that,

$$\left(\hat{\Sigma} - \Sigma^n\right) e_p = n^{-1} \sum_{i=1}^{n} \zeta_{p,n,i} - \left(n^{-1} \sum_{i=1}^{n} k(X_i, \cdot) \right) \left(n^{-1} \sum_{i=1}^{n} \overline{e}_{p,n}(X_i) \right) ,$$

where $\overline{e}_{p,n} = e_p - \mathbb{E}^n[e_p(X_1)]$ and

$$\zeta_{p,n,i} \stackrel{\text{def}}{=} k(X_i, \cdot)\overline{e}_{p,n}(X_i) - \mathbb{E}^n \left\{ k(X_1, \cdot)\overline{e}_{p,n}(X_1) \right\}$$

By the Minkowski inequality,

$$\left\{ \mathbb{E}^n \left\| \left(\hat{\Sigma} - \Sigma^n\right) e_p \right\|_{\mathcal{H}}^2 \right\}^{1/2} \leq \left\{ \mathbb{E}^n \left\| n^{-1} \sum_{i=1}^{n} \zeta_{p,n,i}^2 \right\| \right\}^{1/2}$$
$$+ \left\{ \mathbb{E}^n \left[\left\| n^{-1} \sum_{i=1}^{n} k(X_i, \cdot) \right\|_{\mathcal{H}}^2 \left| n^{-1} \sum_{i=1}^{n} \overline{e}_{p,n}(X_i) \right|^2 \right] \right\}^{1/2} = A_1 + A_2 .$$

We consider these two terms separately. Consider first A_1. We have

$$A_1^2 = n^{-1}\mathbb{E}^n \left\| \zeta_{p,n,i} \right\|_{\mathcal{H}}^2 \leq n^{-1}\mathbb{E}^n \left\{ \|k(X_1, \cdot)\|_{\mathcal{H}}^2 |\overline{e}_{p,n}(X_1)|^2 \right\} \leq n^{-1}|k|_{\infty}\mathbb{E}^n \left[|\overline{e}_{p,n}(X_1)|^2 \right] .$$

Consider now A_2. Since $\left\| n^{-1} \sum_{i=1}^{n} k(X_i, \cdot) \right\|_{\mathcal{H}}^2 \leq |k|_{\infty}$, we have

$$A_2^2 \leq n^{-1}|k|_{\infty}\mathbb{E}^n \left[|\overline{e}_{p,n}(X_1)|^2 \right] .$$

This shows, using the Minkowski inequality, that

$$\left\{ \mathbb{E}^n \left(\sum_{p=1}^{\infty} \left\| (\hat{\Sigma} - \Sigma^n) e_p \right\|_{\mathcal{H}} \right)^2 \right\}^{1/2} \leq 2|k|_{\infty}^{1/2} n^{-1/2} \sum_{p=1}^{\infty} \left\{ \mathbb{E}^n \left[|\overline{e}_{p,n}(X_1)|^2 \right] \right\}^{1/2} .$$

Since by assumption $\sum_{p=1}^{\infty} \left\{ \mathbb{E}^n \left[|\overline{e}_{p,n}(X_1)|^2 \right] \right\}^{1/2} = \sum_{p=1}^{\infty} \lambda_p^{1/2}(\Sigma^n) < \infty$, the proof is concluded. □

Corollary 13. *Assume (A1). Let* $\{X^{(1)}_{1,n_1}, \ldots, X^{(1)}_{n_1,n_1}\}$ *and* $\{X^{(2)}_{1,n_2}, \ldots, X^{(2)}_{n_2,n_2}\}$ *be two triangular arrays, whose mean elements and covariance operators are respectively* (μ^n_1, Σ^n_1) *and* (μ^n_2, Σ^n_2)*, where* $n_1/n \to \rho_1$ *and* $n_2/n \to \rho_2$ *as* n *tends to infinity. If* $\sup_{n \geqslant 0} \sum_{p=1}^{\infty} \lambda_p^{1/2}(\Sigma^n_a) < \infty$*, then*

$$\sum_{p=1}^{\infty} |\lambda_p(\hat{\Sigma}_W - \Sigma_W)| = O_P(n^{-1/2}) \; . \tag{34}$$

In addition, we also have

$$\left\| \hat{\Sigma}_W - \Sigma_W \right\|_{\mathrm{HS}} = O_P(n^{-1/2}) \; . \tag{35}$$

Proof. Since $\hat{\Sigma}_W - \Sigma_W = n_1 n^{-1}(\hat{\Sigma}_1 - \Sigma_1) + n_2 n^{-1}(\hat{\Sigma}_2 - \Sigma^n_2)$, then

$$\sum_{p=1}^{\infty} \left\| (\hat{\Sigma}_W - \Sigma_W) e_p \right\|_{\mathcal{H}} \leq n_1 n^{-1} \sum_{p=1}^{\infty} \left\| (\hat{\Sigma}_1 - \Sigma_1) e_p \right\|_{\mathcal{H}} + n_2 n^{-1} \sum_{p=1}^{\infty} \left\| (\hat{\Sigma}_2 - \Sigma^n_2) e_p \right\|_{\mathcal{H}} \; ,$$

and applying twice Proposition 12 leads to (34). Now, using that

$$\left\| \hat{\Sigma}_W - \Sigma_W \right\|_{\mathrm{HS}} \leq \sum_{p=1}^{\infty} |\lambda_p(\hat{\Sigma}_W - \Sigma_W)| \; , \tag{36}$$

then (35) follows as a direct consequence of (34). □

9. Asymptotic approximation of the test statistics

The following proposition shows that in the asymptotic study of our test statistics, we can replace most empirical quantities by population quantities. For ease of notation, we shall denote $\mu_2 - \mu_1$ by δ. $\hat{\mu}_2 - \hat{\mu}_1$ by $\hat{\delta}$.

Proposition 14. *Assume (C). If*

$$\begin{aligned} &\gamma_n + d_2^{-1}(\Sigma_1, \gamma_n) d_1(\Sigma_1, \gamma_n) \gamma_n^{-1} n^{-1/2} \to 0 \\ &d_2^{-1}(\Sigma_1, \gamma_n) n \eta_n^2 = O(1) \quad \textit{and} \quad d_2^{-1}(\Sigma_1, \gamma_n) d_1(\Sigma_1, \gamma_n) \eta_n \to 0 \; , \end{aligned}$$

then, $\widehat{T}_n(\gamma_n) = \tilde{T}_n(\gamma_n) + o_P(1)$*, where*

$$\tilde{T}_n(\gamma) \stackrel{\text{def}}{=} \frac{(n_1 n_2/n) \; \left\| (\Sigma_1 + \gamma \mathrm{I})^{-1/2} \, \hat{\delta} \right\|^2_{\mathcal{H}} - d_1(\Sigma_1, \gamma)}{\sqrt{2} d_2(\Sigma_1, \gamma)} \; . \tag{37}$$

Proof. Notice that

$$|d_2(\hat{\Sigma}_W, \gamma_n) - d_2(\Sigma_1, \gamma_n)| \leq |d_2(\hat{\Sigma}_W, \gamma_n) - d_2(\Sigma_W, \gamma_n)| + |d_2(\Sigma_W, \gamma_n) - d_2(\Sigma_1, \gamma_n)| \; .$$

Then, on the one hand, using Eq. (77) for $r = 2$ in Lemma 23 with $S = \Sigma_W$ and $\Delta = \hat{\Sigma}_W - \Sigma_W$ and Eq. (34) in Corollary 13, we get $\left| d_2(\hat{\Sigma}_W, \gamma_n) - d_2(\Sigma_W, \gamma_n) \right| = O_P(\gamma_n^{-1} n^{-1/2})$. On the other hand, using Eq. (79) in Lemma 23 with $S = \Sigma_1$ and $\Delta = n_2 n^{-1}(\Sigma^n_2 -$

$\Sigma_1)$, we get $d_2(\Sigma_W,\gamma_n) - d_2(\Sigma_1,\gamma_n) = O(\eta_n)$. Furthermore, similar reasoning, using Eq. (77) and Eq. (78) again in Lemma 23 allows to prove that $d_2^{-1}(\Sigma_1,\gamma_n)d_1(\hat{\Sigma}_W,\gamma_n) = d_2^{-1}(\Sigma_1,\gamma_n)d_1(\Sigma_1,\gamma_n) + o_P(1)$. Next, we shall prove that

$$\left\|(\hat{\Sigma}_W+\gamma_n\mathrm{I})^{-1/2}\hat{\delta}\right\|_{\mathcal{H}}^2 = \left\|(\Sigma_1+\gamma_n\mathrm{I})^{-1/2}\hat{\delta}\right\|_{\mathcal{H}}^2 + n^{-1}O_P\left\{(d_1(\Sigma_1,\gamma_n)+n\eta_n^2)(\gamma_n^{-1}n^{-1/2}+\eta_n)\right\} . \tag{38}$$

Using straightforward algebra, we may write

$$\left|\left\|(\hat{\Sigma}_W+\gamma_n\mathrm{I})^{-1/2}\hat{\delta}\right\|_{\mathcal{H}}^2 - \left\|(\Sigma_1+\gamma_n\mathrm{I})^{-1/2}\hat{\delta}\right\|_{\mathcal{H}}^2\right| \leq A_1A_2\left\{B_1+B_2\right\} , \tag{39}$$

with

$$A_1 \stackrel{\text{def}}{=} \left\|(\Sigma_1+\gamma_n\mathrm{I})^{-1/2}\hat{\delta}\right\|_{\mathcal{H}} , \qquad B_1 \stackrel{\text{def}}{=} \left\|(\hat{\Sigma}_W+\gamma_n\mathrm{I})^{-1/2}(\hat{\Sigma}_W-\Sigma_W)(\Sigma_1+\gamma_n\mathrm{I})^{-1/2}\right\| ,$$
$$A_2 \stackrel{\text{def}}{=} \left\|(\hat{\Sigma}_W+\gamma_n\mathrm{I})^{-1/2}\hat{\delta}\right\|_{\mathcal{H}} , \qquad B_2 \stackrel{\text{def}}{=} \left\|(\hat{\Sigma}_W+\gamma_n\mathrm{I})^{-1/2}(\Sigma_2^n-\Sigma_1)(\Sigma_1+\gamma_n\mathrm{I})^{-1/2}\right\| .$$

We now prove that

$$A_1^2 = O_P(n^{-1}d_1(\Sigma_1,\gamma_n)+\eta_n^2) , \tag{40}$$
$$A_2^2 = O_P(n^{-1}d_1(\Sigma_1,\gamma_n)+\eta_n^2) . \tag{41}$$

We first consider (40). Note that $\mathbb{E}\left(\hat{\delta}\otimes\hat{\delta}\right) = \delta_n\otimes\delta_n + n_1^{-1}\Sigma_1 + n_2^{-1}\Sigma_2^n$, which yields

$$\mathbb{E}\|(\Sigma_1+\gamma_n\mathrm{I})^{-1/2}\hat{\delta}\|^2 = \mathrm{Tr}\Big\{(\Sigma_1+\gamma_n\mathrm{I})^{-1}\mathbb{E}\left(\hat{\delta}\otimes\hat{\delta}\right)\Big\} = \left\langle\delta_n,(\Sigma_1+\gamma_n\mathrm{I})^{-1}\delta_n\right\rangle_{\mathcal{H}} + \frac{n}{n_1n_2}\mathrm{Tr}\big\{(\Sigma_1+\gamma_n\mathrm{I})^{-1}\Sigma_1\big\} + n_2^{-1}\mathrm{Tr}\big\{(\Sigma_1+\gamma_n\mathrm{I})^{-1}(\Sigma_2^n-\Sigma_1)\big\} . \tag{42}$$

Using Proposition 10 with $\Sigma=\Sigma_1$ together with Assumption (C), we may write

$$\left|\left\langle\delta_n,(\Sigma_1+\gamma_n\mathrm{I})^{-1}\delta_n\right\rangle_{\mathcal{H}}\right| \leq \left|\left\langle\delta_n,\Sigma_1^{-1}\delta_n\right\rangle_{\mathcal{H}}\right| \leq \left\|\frac{d\mathbb{P}_2^n}{d\mathbb{P}_1}-1\right\|_{L^2(\mathbb{P}_1)}^2 = \eta_n^2 .$$

Next, applying Lemma 11, we obtain

$$\left|\mathrm{Tr}\{(\Sigma_1+\gamma_n\mathrm{I})^{-1}(\Sigma_2^n-\Sigma_1)\}\right| = O(d_2(\Sigma_1,\gamma_n)\eta_n) ,$$

which yields

$$\mathbb{E}\|(\Sigma_1+\gamma_n\mathrm{I})^{-1/2}\hat{\delta}\|^2 = (n/n_1n_2)d_1(\Sigma_1,\gamma_n)\left\{1+O(\eta_n)\right\}+O(\eta_n^2) . \tag{43}$$

Finally, we get (40) by the Markov inequality. Now, to prove (41), it suffices to observe that $\left\|(\hat{\Sigma}_W+\gamma_n\mathrm{I})^{-1}(\Sigma_1+\gamma_n\mathrm{I})\right\| = 1+o_P(1)$, and then conclude from (40). Next, using the upper-bound $\left\|(\Sigma+\gamma_n\mathrm{I})^{-1/2}\right\| \leq \gamma_n^{-1/2}$, and Corollary 13 which gives $\left\|\hat{\Sigma}_W-\Sigma_W\right\|_{\mathrm{HS}} = O_P(n^{-1/2})$, we get

$$B_1 = O_P(\gamma_n^{-1}n^{-1/2}) . \tag{44}$$

Finally, under Assumption (C), using Eq. (30) in Lemma 11, we obtain

$$B_2 = O_P(\eta_n) \ . \tag{45}$$

The proof of (38) is concluded by plugging (40-41-44-45) into (39). □

Remark 15. *For the sake of generality, we proved the approximation result under the assumptions* $\gamma_n + d_2^{-1}(\Sigma_1,\gamma_n)d_1(\Sigma_1,\gamma_n)\gamma_n^{-1}n^{-1/2} \to 0$ *on the one hand,* $d_2^{-1}(\Sigma_1,\gamma_n)n\eta_n^2 = O(1)$ *and* $d_2^{-1}(\Sigma_1,\gamma_n)d_1(\Sigma_1,\gamma_n)\eta_n \to 0$ *on the other hand. However, in the case* $\gamma_n \equiv \gamma$*, the approximation is still valid if* $n\eta_n^3 \to 0$*, which allows to use this approximation to derive the limiting power of our test statistics against non-directional sequences of local alternatives as in (25).*

10. Proof of Theorems 6-7

For ease of notation, in the subsequent proofs, we shall often omit Σ_1 in quantities involving it. Hence, from now on, $\lambda_p, \lambda_q, d_2$ stand for $\lambda_p(\Sigma_1), \lambda_q, d_2(\Sigma_1,\gamma)$. Define

$$Y_{n,p,i} \stackrel{\text{def}}{=} \begin{cases} \left(\frac{n_2}{n_1 n}\right)^{1/2} \left(e_p(X_i^{(1)}) - \mathbb{E}[e_p(X_1^{(1)})]\right) & 1 \le i \le n_1 \ , \\ -\left(\frac{n_1}{n_2 n}\right)^{1/2} \left(e_p(X_{i-n_1}^{(2)}) - \mathbb{E}[e_p(X_1^{(2)})]\right) & n_1+1 \le i \le n \ . \end{cases} \tag{46}$$

The following lemma gives formulas for the moments of $Y_{n,p,i}$, used throughout the actual proof of the main results.

Lemma 16. *Consider* $\{Y_{n,p,i}\}_{1\le i\le n, p\ge 1}$ *and as defined respectively in (46) . Then*

$$\sum_{i=1}^{n} \mathbb{E}[Y_{n,p,i}Y_{n,q,i}] = \lambda_p^{1/2}\lambda_q^{1/2}\{\delta_{p,q} + n_1 n^{-1}\varepsilon_{p,q}\} \tag{47}$$

$$\mathrm{Cov}(Y_{n,p,i}^2, Y_{n,q,i}^2) \le Cn^{-2}|k|_\infty \lambda_p^{1/2}\lambda_q^{1/2}(1+\varepsilon_{p,p})^{1/2}(1+\varepsilon_{q,q})^{1/2} \ . \tag{48}$$

Proof. The first expressions are proved by elementary calculations from

$$\begin{aligned} &\mathbb{E}[Y_{n,p,1}Y_{n,q,1}] = \frac{n_2}{n_1 n}\delta_{p,q}\lambda_p(\Sigma_1) \\ &\mathbb{E}[Y_{n,p,1}Y_{n,q,n_1+1}] = 0, \qquad \text{since} \quad X_1^{(1)} \perp X_1^{(2)} \\ &\mathbb{E}[Y_{n,p,n_1+1}Y_{n,q,n_1+1}] = \frac{n_1}{n_2 n}\lambda_p^{1/2}\lambda_q^{1/2}\left\{\delta_{p,q} + \left\langle e_p, (\Sigma_1^{-1/2}\Sigma_2^n\Sigma_1^{-1/2} - \mathrm{I})e_q\right\rangle\right\} \ . \end{aligned}$$

Next, notice that, for all $p \ge 1$, we have by the reproducing property and the the Cauchy-Schwarz inequality

$$|e_p(x)| = \langle e_p, k(x,\cdot)\rangle_{\mathcal{H}} \le \|e_p\|_{\mathcal{H}} \, \|k(x,\cdot)\|_{\mathcal{H}} \le |k|_\infty^{1/2} \ .$$

which yields

$$\begin{aligned} \left|\mathrm{Cov}(Y_{n,p,i}^2, Y_{n,q,i}^2)\right| &\le \mathbb{E}[Y_{n,p,i}^2 Y_{n,q,i}^2] + \mathbb{E}[Y_{n,p,i}^2]\mathbb{E}[Y_{n,q,i}^2] \\ &\le C\mathbb{E}^{1/2}[Y_{n,p,i}^4]\mathbb{E}^{1/2}[Y_{n,q,i}^4] \\ &\le Cn^{-1}|k|_\infty \mathbb{E}^{1/2}[Y_{n,p,i}^2]\mathbb{E}^{1/2}[Y_{n,q,i}^2] \\ &\le Cn^{-2}|k|_\infty \lambda_p^{1/2}\lambda_q^{1/2}(1+\varepsilon_{p,p})^{1/2}(1+\varepsilon_{q,q})^{1/2} \ . \end{aligned}$$

□

10.1 Proof of Theorem 6

Proof. The proof is adapted from (Serfling, 1980, pages 195-199). By Proposition 14,

$$\widehat{T}_n(\gamma) = \frac{\hat{V}_{n,\infty}(\gamma) - d_1(\Sigma_1,\gamma)}{\sqrt{2}d_2(\Sigma_1,\gamma)} + o_P(1) \ ,$$

where

$$\hat{V}_{n,\infty}(\gamma) \stackrel{\text{def}}{=} \sum_{p=1}^{\infty} (\lambda_p+\gamma)^{-1} \left(S_{n,p} + \sqrt{\frac{n_1 n_2}{n}} \langle \delta_n, e_p \rangle \right)^2 \ ,$$

with

$$S_{n,p} \stackrel{\text{def}}{=} \sqrt{\frac{n_1 n_2}{n}} \left\langle \hat{\delta} - \delta_n, e_p \right\rangle = \sum_{i=1}^{n} Y_{n,p,i} \ . \tag{49}$$

Now put

$$\hat{V}_{n,N}(\gamma) \stackrel{\text{def}}{=} \sum_{p=1}^{N} (\lambda_p+\gamma)^{-1} \left(S_{n,p} + \sqrt{\frac{n_1 n_2}{n}} \langle \delta_n, e_p \rangle \right)^2 \ . \tag{50}$$

Because $\{Y_{n,p,i}\}$ are zero mean, independent, Lemma 16-Eq. (47) shows that, as n goes to infinity, $\sum_{i=1}^{n} \mathrm{Cov}(Y_{n,p,i}, Y_{n,q,i}) \to \lambda_p^{1/2}\lambda_q^{1/2}\delta_{p,q}$. In addition, the Lyapunov condition is satisfied, since using (48), $\sum_{i=1}^{n} \mathbb{E}[Y_{n,p,i}^4] \leq C n^{-1}\lambda_p$. We may thus apply the central limit theorem for multivariate triangular arrays, which yields $\mathbf{S}_{n,N} \xrightarrow{\mathcal{D}} \mathcal{N}(0, \mathbf{\Lambda}_N)$ where $\mathbf{S}_{n,N} = (S_{n,1}, \ldots, S_{n,N})$ and $(\mathbf{\Lambda}_N)_{p,q} = \delta_{p,q}\lambda_p$, $1 \leq p,q \leq N$. Fix u and let $\epsilon > 0$ be given. Then, using the version of the continuous mapping theorem stated in (van der Vaart, 1998, Theorem 18.11), with the sequence of quadratic functions $\{g_n\}_{n\geq 1}$ defined as $[\ g_n : \mathbf{T}_N = (T_1, \ldots, T_N) \mapsto (\mathbf{T}_N + \mathbf{a}_n)^T [\mathrm{diag}(\alpha_1, \ldots, \alpha_N)](\mathbf{T}_N + \mathbf{a}_n)\ ]$, we may write

$$|\mathbb{E}[\mathrm{e}^{\mathrm{i}u\hat{V}_{n,N}(\gamma)}] - \mathbb{E}[\mathrm{e}^{\mathrm{i}uV_{n,N}(\gamma)}]| \leq \epsilon \ , \tag{51}$$

with $V_{n,N}(\gamma) \stackrel{\text{def}}{=} \sum_{p=1}^{N} (\lambda_p+\gamma)^{-1}\lambda_p (Z_p + a_{n,p})^2$, where $\{Z_p\}_{p\geq 1}$ are independent standard normal random variables, defined on a common probability space, and $\{a_{n,p}\}_{p\geq 1}$ are defined in (13). Next, we prove that $\lim_{N\to\infty} \limsup_{n\to\infty} \mathbb{E}[(\hat{V}_{n,\infty}(\gamma) - \hat{V}_{n,N}(\gamma))^2] = 0$. By the Rosenthal inequality (see (Petrov, 1995, theorem 2.12), there exists a constant C such that $\mathbb{E}[S_{n,p}^4] \leq C(n^{-1}\lambda_p + \lambda_p^2)$. The Minkowski inequality then leads to

$$\begin{aligned}
&\mathbb{E}^{1/2}[(\hat{V}_{n,\infty}(\gamma) - \hat{V}_{n,N}(\gamma))^2] \\
&\leq \sum_{p=N+1}^{\infty} (\lambda_p+\gamma)^{-1} \, \mathbb{E}^{1/2} \left\{ \left(S_{n,p} + \sqrt{\frac{n_1 n_2}{n}} \langle \delta_n, e_p \rangle \right)^4 \right\} \\
&\leq C \left\{ \gamma^{-1} \sum_{p=N+1}^{\infty} \lambda_p^{1/2}(n^{-1/2} + \lambda_p^{1/2}) + n \sum_{p=N+1}^{\infty} (\lambda_p+\gamma)^{-1} \langle \delta_n, e_p \rangle^2 \right\} \\
&\leq C \left\{ \gamma^{-1} \sum_{p=N+1}^{\infty} \lambda_p^{1/2} + n \sum_{p=N+1}^{\infty} (\lambda_p+\gamma)^{-1} \langle \delta_n, e_p \rangle^2 \right\} + o(1) \ .
\end{aligned}$$

Notice that, using (28) in Proposition 10 with $\Sigma = \Sigma_1$, we have

$$n \sum_{p=N+1}^{\infty} (\lambda_p + \gamma)^{-1} \langle \delta_n, e_p \rangle^2 \leq n\gamma^{-1}\lambda_{N+1} \sum_{p=1}^{\infty} \lambda_p^{-1} \langle \delta_n, e_p \rangle^2 \leq \gamma^{-1}\lambda_{N+1}\, n\eta_n^2 \,, \tag{52}$$

which goes to zero uniformly in n as $N \to \infty$. Therefore, under Assumptions (B1) and (C), we may choose N large enough so that

$$|\mathbb{E}[\mathrm{e}^{\mathrm{i}u\hat{V}_{n,\infty}(\gamma)}] - \mathbb{E}[\mathrm{e}^{\mathrm{i}u\hat{V}_{n,N}(\gamma)}]| < \epsilon \,. \tag{53}$$

Similar calculations allow to prove that $\mathbb{E}[(V_{n,\infty}(\gamma) - V_{n,N}(\gamma))^2] = o(1)$, which yields that for all $\epsilon > 0$, for a sufficiently large N, we have

$$|\mathbb{E}[\mathrm{e}^{\mathrm{i}uV_{n,\infty}(\gamma)}] - \mathbb{E}[\mathrm{e}^{\mathrm{i}uV_{n,N}(\gamma)}]| < \epsilon \,. \tag{54}$$

Finally, combining (51) and (53) (54), by the triangular inequality, we have proved that, for $\epsilon > 0$, we may choose a sufficiently large N, such that

$$|\mathbb{E}[\mathrm{e}^{\mathrm{i}u\hat{V}_{n,\infty}(\gamma)}] - \mathbb{E}[\mathrm{e}^{\mathrm{i}uV_{n,\infty}(\gamma)}]| < \epsilon \,, \tag{55}$$

and the proof is concluded by invoking Lévy's continuity theorem (Billingsley, 1995, Theorem 26.3). □

Remark 17. *For the sake of generality, we proved the result under the assumption that $n\eta_n^2 = O(1)$. However, if there exists a nonnegative nondecreasing sequence of integers $\{q_n\}_{n\geq 1}$ such that for all n we have $\sum_{p=1}^{\infty}(\lambda_p + \gamma)^{-1} \langle \delta_n, e_p \rangle^2 = (\lambda_{q_n} + \gamma)^{-1} \langle \delta_n, e_{q_n} \rangle^2$, then the truncation argument used in (52) is valid under a weaker assumption. In particular, when considering non-directional sequences of local alternatives as in (25), it suffices to take $N \to \infty$ such that $N^{-1}q_n = o(1)$, which for n sufficiently large allows to get $n \sum_{p=N+1}^{\infty}(\lambda_p + \gamma)^{-1} \langle \delta_n, e_p \rangle^2 = 0$ in place of (52) in the proof. The rest of the proof follows similarly.*

The following lemma highlights the main difference between the asymptotics respectively when $\gamma_n \equiv \gamma$ and $\gamma_n \to 0$, which is that $d_1(\Sigma_1, \gamma_n) \to \infty$ and $d_2(\Sigma_1, \gamma_n) \to \infty$ in the case $\gamma_n \to 0$, whereas they acted as irrelevant constants in the case $\gamma_n \equiv \gamma$.

Lemma 18. *If $\gamma_n = o(1)$, then, $d_1(\Sigma_1, \gamma_n) \to \infty$, and $d_2(\Sigma_1, \gamma_n) \to \infty$, as n tends to infinity.*

Proof. Since the function $x \mapsto x/(x + \gamma_n)$ is monotone increasing, for any $\lambda \geq \gamma_n$, $\lambda/(\lambda + \gamma_n) \geq 1/2$. Therefore,

$$\sum_{p=1}^{n} \frac{\lambda_p(\Sigma_1)}{\lambda_p(\Sigma_1) + \gamma_n} \geq \frac{1}{2} \# \{k \leq n : \lambda_p(\Sigma_1) \geq \gamma_n\} \,,$$

and the proof is concluded by noting that since $\gamma_n \to 0$, $\# \{k : \lambda_p(\Sigma_1) \geq \gamma_n\} \to \infty$, as n tends to infinity. □

The quantities $\lambda_p(\Sigma_1), \lambda_q(\Sigma_1), d_1(\Sigma_1, \gamma_n), d_2(\Sigma_1, \gamma_n)$ being pervasive in the subsequent proofs, they shall be respectively be abbreviated as $\lambda_p, \lambda_q, d_{1,n}, d_{2,n}$. Our test statistics writes as $\tilde{T}_n = (\sqrt{2} d_{2,n})^{-1} A_n$ with

$$A_n \stackrel{\text{def}}{=} \frac{n_1 n_2}{n} \left\| (\Sigma_1 + \gamma_n \mathrm{I})^{-1/2} \hat{\delta} \right\|^2 - d_{1,n} \, . \tag{56}$$

Using the quantities $S_{n,p}$ and $Y_{n,p,i}$ defined respectively in (49) and (46), A_n may be expressed as

$$\begin{aligned} A_n &= \sum_{p=1}^{\infty} (\lambda_p + \gamma_n)^{-1} \left(S_{n,p} + \sqrt{\frac{n_1 n_2}{n}} \langle \delta_n, e_p \rangle \right)^2 - d_{1,n} \\ &= \sum_{p=1}^{\infty} (\lambda_p + \gamma_n)^{-1} \left\{ S_{n,p}^2 - \mathbb{E} S_{n,p}^2 + 2 \sqrt{\frac{n_1 n_2}{n}} S_{n,p} \langle \delta_n, e_p \rangle \right\} \\ &\quad + \frac{n_1 n_2}{n} \left\langle \delta_n, (\Sigma_1 + \gamma_n \mathrm{I})^{-1} \delta_n \right\rangle + \sum_{p=1}^{\infty} (\lambda_p + \gamma_n)^{-1} \left\{ \mathbb{E} S_{n,p}^2 - \lambda_p \right\} \, . \end{aligned}$$

Since, by Lemma 16 Eq. (47), $\mathbb{E} S_{n,p}^2 - \lambda_p = (n_1/n) \lambda_p \varepsilon_{p,p}$, where $\varepsilon_{p,p}$ is defined in (33), then, by Hölder inequality, we obtain

$$\left| \sum_{p=1}^{\infty} (\lambda_p + \gamma_n)^{-1} \left\{ \mathbb{E} S_{n,p}^2 - \lambda_p \right\} \right| \leq \left(\sum_{p=1}^{\infty} (\lambda_p + \gamma_n)^{-2} \lambda_p^2 \right)^{1/2} \left(\sum_{p=1}^{\infty} \varepsilon_{p,p}^2 \right)^{1/2} = O(d_{2,n} \eta_n) \, .$$

We now decompose

$$\sum_{p=1}^{\infty} (\lambda_p + \gamma_n)^{-1} \left\{ S_{n,p}^2 - \mathbb{E} S_{n,p}^2 + 2 \sqrt{\frac{n_1 n_2}{n}} S_{n,p} \langle \delta_n, e_p \rangle \right\} = B_n + 2C_n + 2D_n \, ,$$

where B_n and C_n and D_n are defined as follows

$$B_n \stackrel{\text{def}}{=} \sum_{p=1}^{\infty} \sum_{i=1}^{n} \left\{ Y_{n,p,i}^2 - \mathbb{E} Y_{n,p,i}^2 \right\} \, , \tag{57}$$

$$C_n \stackrel{\text{def}}{=} \sum_{p=1}^{\infty} (\lambda_p + \gamma_n)^{-1} \sum_{i=1}^{n} Y_{n,p,i} \sqrt{\frac{n_1 n_2}{n}} \langle \delta_n, e_p \rangle \, , \tag{58}$$

$$D_n \stackrel{\text{def}}{=} \sum_{p=1}^{\infty} (\lambda_p + \gamma_n)^{-1} \sum_{i=1}^{n} Y_{n,p,i} \left\{ \sum_{j=1}^{i-1} Y_{n,p,j} \right\} \, . \tag{59}$$

The proof is in three steps. We will first show that B_n is negligible, then that C_n is negligible, and finally establish a central limit theorem for D_n.

Step 1: $B_n = o_P(1)$. The proof amounts to compute the variance of this term. Since the variables $Y_{n,p,i}$ and $Y_{n,q,j}$ are independent if $i \neq j$, then $\mathrm{Var}(B_n) = \sum_{i=1}^n v_{n,i}$, where

$$v_{n,i} \stackrel{\text{def}}{=} \mathrm{Var}\left(\sum_{p=1}^{\infty}(\lambda_p+\gamma_n)^{-1}\{Y_{n,p,i}^2 - \mathbb{E}[Y_{n,p,i}^2]\}\right)$$
$$= \sum_{p,q=1}^{\infty}(\lambda_p+\gamma_n)^{-1}(\lambda_q+\gamma_n)^{-1}\mathrm{Cov}(Y_{n,p,i}^2, Y_{n,q,i}^2)\ .$$

Using Lemma 16, Eq. (48), we get

$$\sum_{i=1}^n v_{n,i} \leq Cn^{-1}\left(\sum_{p=1}^{\infty}(\lambda_p+\gamma_n)^{-1}\lambda_p^{1/2}(1+\varepsilon_{p,p})^{1/2}\right)^2 \leq Cn^{-1}\gamma_n^{-2}\left(\sum_{p=1}^{\infty}\lambda_p^{1/2}\right)^2\{1+O(\eta_n)\}$$

where the RHS above is indeed negligible, since by assumption we have $\gamma_n^{-1}n^{-1/2} \to 0$ and $\sum_{p=1}^{\infty}\lambda_p^{1/2} < \infty$. □

Step 2: $C_n = o_P(d_{2,n}^2)$. Again, the proof essentially consists in computing the variance of this term, and then conclude by the Markov inequality. As previously, since the variables $Y_{n,p,i}$ and $Y_{n,q,j}$ are independent if $i \neq j$, then $\mathrm{Var}(C_n) = \sum_{i=1}^n u_{n,i}$, where

$$u_{n,i} \stackrel{\text{def}}{=} \sum_{p=1}^{\infty}(\lambda_p+\gamma_n)^{-2}\mathbb{E}[Y_{n,p,i}^2]\frac{n_1n_2}{n}\langle\delta_n, e_p\rangle^2$$
$$+ \sum_{p,q=1}^{\infty}(\lambda_q+\gamma_n)^{-1}(\lambda_q+\gamma_n)^{-1}\mathbb{E}[Y_{n,p,i}Y_{n,q,i}]\frac{n_1n_2}{n}\langle\delta_n, e_p\rangle\langle\delta_n, e_q\rangle\ .$$

Moreover, note that $\mathbb{E}[Y_{n,p,i}^2] \leq Cn^{-1}\lambda_p$, and under Assumption (C1)

$$\frac{n_1n_2}{n}d_{2,n}^{-2}\sum_{p=1}^{\infty}(\lambda_p+\gamma_n)^{-2}\lambda_p\langle\delta_n, e_p\rangle^2$$
$$= \frac{n_1n_2}{n}d_{2,n}^{-2}\sum_{p=1}^{\infty}(\lambda_p+\gamma_n)^{-1}\langle\delta_n, e_p\rangle^2 \leq d_{2,n}^{-1}\frac{n\left\langle\delta_n, (\Sigma_1+\gamma_n)^{-1}\delta_n\right\rangle}{d_{2,n}} = o(1)\ .$$

Similarly, for $p \neq q$ we have $|\mathbb{E}[Y_{n,p,i}Y_{n,q,i}]| \leq Cn^{-1}\lambda_p^{1/2}\lambda_q^{1/2}|\varepsilon_{p,q}|$, which implies that

$$\frac{n_1n_2}{n}d_{2,n}^{-2}\sum_{p\neq q}(\lambda_p+\gamma_n)^{-1}(\lambda_q+\gamma_n)^{-1}\lambda_p^{1/2}\lambda_q^{1/2}|\langle\delta_n, e_p\rangle||\langle\delta_n, e_q\rangle||\varepsilon_{p,q}|$$
$$\leq \frac{n_1n_2}{n}d_{2,n}^{-2}\left(\sum_{p=1}^{\infty}(\lambda_p+\gamma_n)^{-2}\lambda_p\langle\delta_n, e_p\rangle^2\right)\left(\sum_{p\neq q}\varepsilon_{p,q}^2\right)^{1/2} = o(1)\ .$$

□

Step 3: $d_{2,n}^{-1} D_n \xrightarrow{\mathcal{D}} \mathcal{N}(0, 1/2)$. We use the central limit theorem (CLT) for triangular array of martingale difference (Hall and Heyde, 1980, Theorem 3.2). For $= 1, \ldots, n$, denote

$$\xi_{n,i} \stackrel{\text{def}}{=} d_{2,n}^{-1} \sum_{p=1}^{\infty} (\lambda_p + \gamma_n)^{-1} Y_{n,p,i} M_{n,p,i-1} \;, \quad \text{where} \quad M_{n,p,i} \stackrel{\text{def}}{=} \sum_{j=1}^{i} Y_{n,p,j} \;, \tag{60}$$

and let $\mathcal{F}_{n,i} = \sigma\left(Y_{n,p,j}, p \in \{1, \ldots, n\}, j \in \{0, \ldots, i\}\right)$. Note that, by construction, $\xi_{n,i}$ is a martingale increment, that is $\mathbb{E}\left[\xi_{n,i} \,|\, \mathcal{F}_{n,i-1}\right] = 0$. The first step in the proof of the CLT is to establish that

$$s_n^2 = \sum_{i=1}^{n} \mathbb{E}\left[\xi_{n,i}^2 \,\middle|\, \mathcal{F}_{n,i-1}\right] \xrightarrow{\text{P}} 1/2 \;. \tag{61}$$

The second step of the proof is to establish the negligibility condition. We invoke (Hall and Heyde, 1980, Theorem 3.2), which requires to establish that $\max_{1 \le i \le n} |\xi_{n,i}| \xrightarrow{\text{P}} 0$ (smallness) and $\mathbb{E}(\max_{1 \le i \le n} \xi_{n,i}^2)$ is bounded in n (tightness), where $\xi_{n,i}$ is defined in (60). We will establish the two conditions simultaneously by checking that

$$\mathbb{E}\left(\max_{1 \le i \le n} \xi_{n,i}^2\right) = o(1) \;. \tag{62}$$

Splitting the sum s_n^2, between diagonal terms E_n, and off-diagonal terms F_n, we have

$$E_n = d_{2,n}^{-2} \sum_{p=1}^{\infty} (\lambda_p + \gamma_n)^{-2} \sum_{i=1}^{n} M_{n,p,i-1}^2 \mathbb{E}[Y_{n,p,i}^2] \;, \tag{63}$$

$$F_n = d_{2,n}^{-2} \sum_{p \neq q} (\lambda_p + \gamma_n)^{-1} (\lambda_q + \gamma_n)^{-1} \sum_{i=1}^{n} M_{n,p,i-1} N_{n,q,i-1} \mathbb{E}[Y_{n,p,i} Y_{n,q,i}] \;. \tag{64}$$

Consider first the diagonal terms E_n. We first compute its mean. Note that $\mathbb{E}[N_{n,p,i}^2] = \sum_{j=1}^{i} \mathbb{E}[Y_{n,p,j}^2]$. Using Lemma 16, we get

$$\begin{aligned}
&\sum_{p=1}^{\infty} (\lambda_p + \gamma_n)^{-2} \sum_{i=1}^{n} \sum_{j=1}^{i-1} \mathbb{E}[Y_{n,p,j}^2] \mathbb{E}[Y_{n,p,i}^2] \\
&= \frac{1}{2} \sum_{p=1}^{\infty} (\lambda_p + \gamma_n)^{-2} \left\{ \left[\sum_{i=1}^{n} \mathbb{E}[Y_{n,p,i}^2] \right]^2 - \sum_{i=1}^{n} \mathbb{E}^2[Y_{n,p,i}^2] \right\} = \frac{1}{2} d_{2,n}^2 \left\{ 1 + O(d_{2,n}^{-1} \eta_n) + O(n^{-1}) \right\} \;.
\end{aligned}$$

Therefore, $\mathbb{E}[E_n] = 1/2 + o(1)$. Next, we check that $E_n - \mathbb{E}[E_n] = o_P(1)$ is negligible. We write $E_n - \mathbb{E}[E_n] = d_{2,n}^{-2} \sum_{p=1}^{n} (\lambda_p + \gamma_n)^{-2} Q_{n,p}$, with

$$Q_{n,p} \stackrel{\text{def}}{=} \sum_{i=1}^{n} \mathbb{E}[Y_{n,p,i+1}^2] \left\{ N_{n,p,i}^2 - \mathbb{E}\left[N_{n,p,i}^2\right] \right\} \;. \tag{65}$$

Using this notation,

$$\mathrm{Var}[E_n] = d_{2,n}^{-4} \sum_{p=1}^{n} (\lambda_p + \gamma_n)^{-4} \mathbb{E}[Q_{n,p}^2] + 2 d_{2,n}^{-4} \sum_{1 \le p < q \le n} (\lambda_p + \gamma_n)^{-2} (\lambda_q + \gamma_n)^{-2} \mathbb{E}[Q_{n,p} Q_{n,q}] \; . \quad (66)$$

We will establish that

$$|\mathbb{E}[Q_{n,p} Q_{n,q}]| \le C \left\{ \lambda_p^2 \lambda_q^2 (\delta_{p,q} + |\varepsilon_{p,q}|)^2 + n^{-1} \lambda_p^{3/2} \lambda_q^{3/2} \right\} \; . \quad (67)$$

Plugging this bound into (66) and using that $\lambda_p/(\lambda_p + \gamma_n) \le 1$ and $d_{2,n} \to \infty$ as n tends to infinity, yields under Assumption (B1)

$$\mathrm{Var}[E_n] \le \left\{ d_{2,n}^{-2} + n^{-1} \gamma_n^{-1} d_{2,n}^{-2} \right\} + C \left\{ d_{2,n}^{-2} \eta_n + n^{-1} d_{2,n}^{-4} \left(\sum_{p=1}^{\infty} \lambda_p \right)^2 \right\} \; ,$$

showing that $\mathrm{Var}[E_n] = o(1)$, and hence that $E_n - \mathbb{E}[E_n] = o_P(1)$. To show (67), note first that $\{M_{n,p,i}^2 - \mathbb{E}[M_{n,p,i}^2]\}_{1 \le i \le n}$ is a $\mathcal{F}_n$-adapted martingale. Denote by $\nu_{n,p,i}$ its increment defined recursively as follows: $\nu_{n,p,1} = N_{n,p,1}^2 - \mathbb{E}[N_{n,p,1}^2]$ and for $i \ge 1$ as

$$\nu_{n,p,i} = M_{n,p,i}^2 - \mathbb{E}[M_{n,p,i}^2] - \left\{ N_{n,p,i-1}^2 - \mathbb{E}[N_{n,p,i-1}^2] \right\} = Y_{n,p,i}^2 - \mathbb{E}[Y_{n,p,i}^2] + 2 Y_{n,p,i} M_{n,p,i-1} \; .$$

Using the summation by part formula, $Q_{n,p}$ may be expressed as

$$Q_{n,p} = \sum_{i=1}^{n-1} \nu_{n,p,i} \left[\sum_{j=i+1}^{n} \mathbb{E}[Y_{n,p,j}^2] \right] \; .$$

Using Lemma 16, Eq. (47), we obtain for any $1 \le p \le q \le n$,

$$\begin{aligned} |\mathbb{E}[Q_{n,p} Q_{n,q}]| &\le \left(\sum_{j=1}^{n} \mathbb{E}[Y_{n,p,j}^2] \right) \left(\sum_{j=1}^{n} \mathbb{E}[Y_{n,q,j}^2] \right) \left| \sum_{i=1}^{n-1} \mathbb{E}[\nu_{n,p,i} \nu_{n,q,i}] \right| \\ &\le C \lambda_p \lambda_q (1 + O(\eta_n)) \left| \sum_{i=1}^{n-1} \mathbb{E}[\nu_{n,p,i} \nu_{n,q,i}] \right| \; . \end{aligned} \quad (68)$$

We get

$$\mathbb{E}[\nu_{n,p,i} \nu_{n,q,i}] = \mathrm{Cov}(Y_{n,p,i}^2, Y_{n,q,i}^2) + 4 \mathbb{E}\left\{ Y_{n,p,i} Y_{n,q,i} \right\} \mathbb{E}\left\{ M_{n,p,i-1} N_{n,q,i-1} \right\} \; .$$

First, applying Eq. (48) in Lemma 16 gives

$$\sum_{i=1}^{n-1} \mathrm{Cov}(Y_{n,p,i}^2, Y_{n,q,i}^2) \le C n^{-1} \lambda_p^{1/2} \lambda_q^{1/2} \; . \quad (69)$$

Since $\mathbb{E}[M_{n,p,i-1}N_{n,q,i-1}] = \sum_{j=1}^{i-1}\mathbb{E}[Y_{n,p,j}Y_{n,q,j}]$, Lemma 16, Eq. (47) shows that

$$\left|\sum_{i=1}^{n}\mathbb{E}[Y_{n,p,i}Y_{n,q,i}]\mathbb{E}[M_{n,p,i-1}N_{n,q,i-1}]\right| = \frac{1}{2}\left|\left\{\left(\sum_{i=1}^{n}\mathbb{E}[Y_{n,p,i}Y_{n,q,i}]\right)^2 - \sum_{i=1}^{n}\mathbb{E}^2[Y_{n,p,i}Y_{n,q,i}]\right\}\right|$$
$$\leq C\lambda_p\lambda_q(\delta_{p,q}+|\varepsilon_{p,q}|)^2 \,. \tag{70}$$

Eq. 67 follows by plugging (69) and (70) into (68). We finally consider F_n defined in (64). We will establish that $F_n = o_P(1)$. Using Lemma 16-Eq. (47),

$$\mathbb{E}^{1/2}[M_{n,p,i-1}^2]\mathbb{E}^{1/2}[N_{n,q,i-1}^2] \leq C\lambda_p^{1/2}\lambda_q^{1/2} \,,$$

and $|\mathbb{E}[Y_{n,p,i}Y_{n,q,i}]| \leq Cn^{-1}\lambda_p^{1/2}\lambda_q^{1/2}\varepsilon_{p,q}$, the Minskovski inequality implies that

$$\{\mathbb{E}|F_n|^2\}^{1/2} \leq Cd_{2,n}^{-2}\sum_{p\neq q}(\lambda_p+\gamma_n)^{-1}(\lambda_q+\gamma_n)^{-1}\lambda_p\lambda_q\varepsilon_{p,q} \leq C\eta_n \,,$$

showing that $F_n = o(1)$. This concludes the proof of Eq. (61).

We finally show Eq. (62). Since $|Y_{n,p,i}| \leq n^{-1/2}|k|_\infty^{1/2}$ $\mathbb{P}$-a.s we may bound

$$\max_{1\leq i\leq n}|\xi_{n,i}| \leq Cd_{2,n}^{-1}n^{-1/2}\sum_{p=1}^{\infty}(\lambda_p+\gamma_n)^{-1}\max_{1\leq i\leq n}|M_{n,p,i-1}| \,. \tag{71}$$

Then, the Doob inequality implies that $\mathbb{E}^{1/2}[\max_{1\leq i\leq n}|M_{n,p,i-1}|^2] \leq \mathbb{E}^{1/2}[N_{n,p,n-1}^2] \leq C\lambda_p^{1/2}$. Plugging this bound in (71), the Minkowski inequality

$$\mathbb{E}^{1/2}\left(\max_{1\leq i\leq n}\xi_{n,i}^2\right) \leq C\left\{d_{2,n}^{-1}\gamma_n^{-1}n^{-1/2}\sum_{p=1}^{\infty}\lambda_p^{1/2}\right\} \,,$$

and the proof is concluded using the fact that $\gamma_n + d_2^{-1}(\Sigma_1,\gamma_n)d_1(\Sigma_1,\gamma_n)\gamma_n^{-1}n^{-1/2} \to 0$ and Assumption (B1). □

11. Proof of Theorem 5

Proof of Proposition 4. We denote by $\Sigma = \rho_1\Sigma_1 + \rho_2\Sigma_2 + \rho_1\rho_2\delta\otimes\delta$ the covariance operator associated with the probability density $p = \rho_1 p_1 + \rho_2 p_2$, and $\delta = \mu_2 - \mu_1$. Then, Proposition 10 applied to the probability densities p_1, p_2 and $p = \rho_1 p_1 + \rho_2 p_2$ shows that $\left\langle \delta, \Sigma^{-1}\delta\right\rangle_{\mathcal{H}} = \int \frac{(p_1-p_2)^2}{\rho_1 p_1+\rho_2 p_2}d\rho$. Thus

$$\begin{aligned}
\rho_1\rho_2\left\langle \delta, \Sigma^{-1}\delta\right\rangle_{\mathcal{H}} &= \frac{1}{2}\int \frac{\frac{\rho_1}{\rho_2}(p_1-p)^2 + \frac{\rho_2}{\rho_1}(p_2-p)^2}{p}d\rho \\
&= \frac{1}{2\rho_1\rho_2}\int\frac{\rho_1^2p_1^2+\rho_2^2p_2^2}{p}d\rho - \frac{1}{2}\frac{\rho_2}{\rho_1} - \frac{1}{2}\frac{\rho_1}{\rho_2} \\
&= \frac{1}{2\rho_1\rho_2} - \frac{1}{2}\frac{\rho_2}{\rho_1} - \frac{1}{2}\frac{\rho_1}{\rho_2} - \int\frac{p_1p_2}{p}d\rho = 1 - \int\frac{p_1p_2}{p}d\rho \,.
\end{aligned}$$

The previous inequality shows that $\rho_1\rho_2 \left\langle \delta, \Sigma^{-1}\delta \right\rangle_{\mathcal{H}} < 1$ is satisfied when $\int p_1 p_2 / p d\rho \neq 0$. Therefore, in this situation,

$$\begin{aligned}\left\langle \delta, (\rho_1\Sigma_1 + \rho_2\Sigma_2)^{-1}\delta \right\rangle_{\mathcal{H}} &= \left\langle \delta, (\Sigma - \rho_1\rho_2\delta\otimes\delta)^{-1}\delta \right\rangle_{\mathcal{H}} \\ &= \left\langle \delta, \Sigma^{-1}\delta \right\rangle_{\mathcal{H}} \left(1 - \rho_1\rho_2 \left\langle \delta, \Sigma^{-1}\delta \right\rangle_{\mathcal{H}}\right)^{-1} ,\end{aligned}$$

and the proof follows by combining the two latter equations.

Consider now the case where $\int p_1 p_2 / p d\rho = 0$, that is when the probability distribution $\mathbb{P}_1$ and $\mathbb{P}_2$ are singular (for any set $A \in \mathfrak{X}$ such as $\mathbb{P}_1(A) \neq 0$, $\mathbb{P}_2(A) = 0$ and vice-versa). In that case, $\left\langle \delta, (\rho_1\Sigma_1 + \rho_2\Sigma_2)^{-1}\delta \right\rangle_{\mathcal{H}}$ is infinite. □

Proof. We first prove that

$$\left\| (\hat{\Sigma}_W + \gamma_n \mathrm{I})^{-1/2}\hat{\delta} \right\|_{\mathcal{H}}^2 \xrightarrow{\mathrm{P}} \left\| (\Sigma_W + \gamma_\infty I)^{-1/2}\delta \right\|_{\mathcal{H}}^2 , \tag{72}$$

where $\gamma_\infty \stackrel{\text{def}}{=} \lim_{n\to\infty}\gamma_n$. Using straightforward algebra, we may write

$$\left| \left\| (\hat{\Sigma}_W + \gamma_n \mathrm{I})^{-1/2}\hat{\delta} \right\|_{\mathcal{H}}^2 - \left\| (\Sigma_W + \gamma_\infty I)^{-1/2}\delta \right\|_{\mathcal{H}}^2 \right| \leq C_1 + C_2 + C_3 , \tag{73}$$

where

$$\begin{aligned}
C_1 &\stackrel{\text{def}}{=} \left\| (\Sigma_W + \gamma_n \mathrm{I})^{-1/2}\hat{\delta} \right\|_{\mathcal{H}} \left\| (\hat{\Sigma}_W + \gamma_n \mathrm{I})^{-1/2}\hat{\delta} \right\|_{\mathcal{H}} \left\| (\hat{\Sigma}_W + \gamma_n \mathrm{I})^{-1/2}(\hat{\Sigma}_W - \Sigma_W)\,(\Sigma_W + \gamma_n \mathrm{I})^{-1/2} \right\|_{\mathrm{HS}} , \\
C_2 &\stackrel{\text{def}}{=} \left| \left(\left\| (\Sigma_W + \gamma_n \mathrm{I})^{-1/2}\hat{\delta} \right\|_{\mathcal{H}}^2 - \left\| (\Sigma_W + \gamma_n \mathrm{I})^{-1/2}\delta \right\|_{\mathcal{H}}^2 \right) \right| , \\
C_3 &\stackrel{\text{def}}{=} \left| \left(\left\| (\Sigma_W + \gamma_n \mathrm{I})^{-1/2}\delta \right\|_{\mathcal{H}}^2 - \left\| (\Sigma_W + \gamma_\infty I)^{-1/2}\delta \right\|_{\mathcal{H}}^2 \right) \right| .
\end{aligned}$$

First, prove that $C_1 = o_P(1)$. Write $C_1 = A_1 A_2 B_1$. Using (with obvious changes) the relation (42), the monotone convergence theorem yields

$$\lim_{n\to\infty} \mathbb{E} \left\| (\Sigma_W + \gamma_n \mathrm{I})^{-1/2}\hat{\delta} \right\|_{\mathcal{H}}^2 = \left\langle \delta, (\Sigma_W + \gamma_\infty I)^{-1}\delta \right\rangle_{\mathcal{H}} .$$

which gives $A_1 = O_P(1)$. As for proving $A_2 = O_P(1)$, using an argument similar to the one used to derive Eq. (41), it suffices to observe that $A_2 = A_1 + o_P(1)$. Then, Eq. (35) in Corollary 13 gives $B_1 = O_P(\gamma_n^{-1} n^{-1/2})$, which shows that $C_1 = A_1 A_2 B_1 = o_P(1)$. Next, prove that $C_2 = o_P(1)$. We may write

$$C_2 = 2 \left\langle \hat{\delta} - \delta, (\Sigma_W + \gamma_n \mathrm{I})^{-1}\delta \right\rangle_{\mathcal{H}} + \left\| (\Sigma_W + \gamma_n \mathrm{I})^{-1/2}(\hat{\delta} - \delta) \right\|_{\mathcal{H}}^2 .$$

Since $\left\| (\Sigma_W + \gamma_n \mathrm{I})^{-1/2} \right\|_{\mathcal{H}} \leq \gamma_n^{-1/2}$, and $\left\| (\Sigma_W + \gamma_n \mathrm{I})^{-1/2}\delta \right\|_{\mathcal{H}} < \infty$, and moreover $\|\hat{\delta} - \delta\|_{\mathcal{H}} = O_P(n^{-1/2})$, then we get $C_2 = O_P(\gamma_n^{-1/2} n^{-1/2}) = o_P(1)$ Finally, prove that $C_3 = o(1)$ Note that $C_3 = -\sum_{p=1}^{\infty} \gamma_n^{-1}(\lambda_p + \gamma_n)^{-1}\lambda_p \langle \delta, e_p \rangle_{\mathcal{H}}^2$, where $\{\lambda_p\}$ and $\{e_p\}$ denote respectively the eigenvalues and eigenvectors of Σ_W. Since $[\gamma \mapsto (\lambda_p + \gamma)^{-1}\gamma]$ is monotone, the monotone convergence theorem shows that $C_3 = o(1)$.

Now, when $\mathbb{P}_1 \neq \mathbb{P}_2$, Proposition 4 with $\mathbb{P} = \rho_1\mathbb{P}_1 + \rho_2\mathbb{P}_2$ ensures that $\delta \in \mathcal{R}(\Sigma_W^{1/2})$ as long as $\left\| \frac{d\mathbb{P}_2}{d\mathbb{P}_1} - 1 \right\|_{L^2(\mathbb{P}_1)} < \infty$. Then, under assumption (A2), by injectivity of Σ_W we have $\delta \neq 0$. Hence, since Σ_W is trace-class, we may apply Lemma 19 with $\alpha = 1$, which yields $d^{-1}(\Sigma_W, \gamma_n)\ n \to \infty$. Therefore, $\widehat{T}_n(\gamma_n) \xrightarrow{\mathrm{P}} \infty$, and the proof ois concluded. Otherwise, that is when $\left\| \frac{d\mathbb{P}_2}{d\mathbb{P}_1} - 1 \right\|_{L^2(\mathbb{P}_1)} = \infty$, we have $\widehat{T}_n(\gamma_n) \xrightarrow{\mathrm{P}} \infty$. □

Appendix A. Technical Lemmas

Lemma 19. *Let $\{\lambda_p\}_{p\geq 1}$ be a non-increasing sequence of non-negative numbers. Let $\alpha > 0$. Assume that $\sum_{p\geq 1} \lambda_p^\alpha < \infty$. Then, for any $\beta \geq \alpha$,*

$$\sup_{\gamma>0} \gamma^\alpha \sum_{p=1}^{\infty} \lambda_p^\beta (\lambda_p + \gamma)^{-\beta} \leq 2 \sum_{p=1}^{\infty} \lambda_p^\alpha \,. \tag{74}$$

In addition, if $\lim_{p\to\infty} p\lambda_p^\alpha = \infty$, then for any $\beta > 0$,

$$\lim_{\gamma\to 0} \gamma^\alpha \sum_{p=1}^{\infty} \lambda_p^\beta (\lambda_p + \gamma)^{-\beta} = \infty \,. \tag{75}$$

Proof. For $\gamma > 0$, denote by $q_\gamma = \sup_{p\geq 1}\{p : \lambda_p > \gamma\}$. Then,

$$\gamma^\alpha \sum_{p=1}^{\infty} \lambda_p^\beta (\lambda_p + \gamma)^{-\beta} \leq \gamma^\alpha \sum_{p=1}^{\infty} \lambda_p^\alpha (\lambda_p + \gamma)^{-\alpha} \leq \gamma^\alpha q_\gamma + \sum_{p>q_\gamma}^{\infty} \lambda_p^\alpha \,. \tag{76}$$

Since the sequence $\{\lambda_p\}$ is non-increasing, the condition $C \stackrel{\mathrm{def}}{=} \sum_{p\geq 1} \lambda_p^\alpha < \infty < \infty$ implies that $p\lambda_p^\alpha \leq C$. Therefore, $\lambda_p \leq C^{1/\alpha} p^{-1/\alpha}$, which implies that for any p satisfying $C\gamma^{-\alpha} \leq p$, $\lambda_p \leq \gamma$, showing that $q_\gamma \leq C\gamma^{-\alpha}$. This establishes (74).

Since $\lambda \mapsto \lambda(\lambda + \gamma)^{-1}$ is non-decreasing, for $p \leq q_\gamma$, $\lambda_p(\lambda_p + \gamma)^{-1} \geq (1/2)$. Therefore, $\gamma^\alpha \sum_{p=1}^{\infty} \lambda_p^\beta (\lambda_p + \gamma)^{-\beta} \geq (2)^{-\beta}\gamma^\alpha q_\gamma$. Since $\lim_{p\to\infty} p\lambda_p^\alpha = \infty$, this means that $\lambda_p > 0$ for any p, which implies that $\lim_{\gamma\to 0^+} q_\gamma = \infty$. Therefore, $\lim_{\gamma\to 0^+} q_\gamma \lambda_{q_\gamma}^\alpha = \lim_{\gamma\to 0^+} q_\gamma \gamma^\alpha = \infty$. The proof follows. □

Lemma 20. *Let $\{\lambda_p\}_{p\geq 1}$ be a non-increasing sequence of non-negative numbers. Assuse there exists $s > 0$ such that $\lambda_p = p^{-s}$ for all $p \geq 1$. Then,*

$$\left[\sum_{p=1}^{\infty} (\lambda_p + \gamma_n)^{-r} \lambda_p^r \right]^{1/r} = \gamma^{-1/sr} \left\{ \int_0^\infty (1 + v^s)^{-r} dv \right\}^{1/r} (1 + o(1)) \,, \quad \textit{as } \gamma \to 0 \,.$$

Proof. First note that

$$\sum_{p=1}^{\infty} (\lambda_p + \gamma_n)^{-r} \lambda_p^r = \sum_{p=1}^{\infty} (1 + \gamma_n \lambda_p^{-1})^{-r} = \sum_{p=1}^{\infty} (1 + (\gamma_n^{1/s} p)^s)^{-r} \,.$$

For all $\gamma > 0$, the function $[u \mapsto (1 + (\gamma^{1/s}u)^s)^{-r}]$ is increasing and nonnegative. Therefore, for all $p \geq 1$ we may write

$$\begin{aligned} \int_p^{p+1} (1 + (\gamma^{1/s}u)^s)^{-r} du &\leq (1 + (\gamma^{1/s}p)^s)^{-r} \leq \int_{p-1}^{p} (1 + (\gamma^{1/s}u)^s)^{-r} du \ , \\ \gamma^{-1/s} \int_{\gamma^{1/s}p}^{\gamma^{1/s}(p+1)} (1 + v^s)^{-r} dv &\leq (1 + (\gamma^{1/s}p)^s)^{-r} \leq \gamma^{-1/s} \int_{\gamma^{1/s}(p-1)}^{\gamma^{1/s}p} (1 + v^s)^{-r} dv \ . \end{aligned}$$

Hence, sussing on p over $1, \ldots, N-1$, we obtain

$$\gamma^{-1/s} \int_{\gamma^{1/s}}^{\gamma^{1/s}N} (1 + v^s)^{-r} dv \leq \sum_{p=1}^{N} (1 + (\gamma^{1/s}p)^s)^{-r} \leq \gamma^{-1/s} \int_{0}^{\gamma^{1/s}N} (1 + v^s)^{-r} dv \ .$$

Therefore, taking $N \to \infty$ in such a way that $\gamma^{1/s}N \to \infty$ as $\gamma \to 0$, we finally get

$$\sum_{p=1}^{\infty} (1 + (\gamma^{1/s}p)^s)^{-r} = \gamma^{-1/s} \left\{ \int_0^{\infty} (1 + v^s)^{-r} dv \right\} (1 + o(1)) \ . \qquad \square$$

Lemma 21. *Let A be a self-adjoint compact operator on $\mathcal{H}$. Then, for any orthonormal basis $\{\varphi_p\}_{p\geq 1}$ of $\mathcal{H}$,*

$$\sum_{p=1}^{\infty} |\lambda_p(A)| \leq \sum_{p=1}^{\infty} \|A\varphi_p\|_{\mathcal{H}} \ .$$

Proof. Let $\{\psi_p\}_{p\geq 1}$ be an orthonormal basis of $\mathcal{H}$ consisting of a sequence of eigenvectors of A corresponding to the eigenvalues $\{\lambda_p(A)\}$ of this latter operator, so that $\langle \psi_p, A\psi_p \rangle_{\mathcal{H}} = \lambda_p(A)$. Then,

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

$\square$

Appendix B. Perturbation results on covariance operators

Lemma 22. *Let A be a compact self-adjoint operator, with $\{\lambda_p\}_{p\geq 1}$ the eigenvalues of A, and $\{e_p\}_{p\geq 1}$ an orthonormal system of eigenvectors of A. Then, for all integer $k > 1$, using the convention $p_{k+1} = p_1$,*

$$\sum_{p=1}^{\infty} \left\langle e_p, (AB)^k e_p \right\rangle = \sum_{p_1=1}^{\infty} \sum_{p_2=1}^{\infty} \cdots \sum_{p_k=1}^{\infty} \left\{ \left(\prod_{j=1}^{k} \lambda_{p_j} \right) \left(\prod_{j=1}^{k} \langle e_{p_j}, B e_{p_{j+1}} \rangle \right) \right\} \ .$$

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

$$\sum_{p=1}^{\infty} \left\langle e_p, (AB)^k e_p \right\rangle$$
$$= \sum_{p_1=1}^{\infty} \sum_{p_2=1}^{\infty} \cdots \sum_{p_\ell=1}^{\infty} \left\{ \left(\prod_{j=1}^{\ell-1} \lambda_{p_j} \right) \left(\prod_{j=1}^{\ell-1} \langle e_{p_j}, B e_{p_{j+1}} \rangle \right) \left\langle e_{p_\ell}, (AB)^{k-\ell+1} e_{p_1} \right\rangle \right\}, \quad \mathcal{P}(\ell) \ .$$

First, for $\ell = 2$, using that $A^* e_{p_1} = A e_{p_1} = \lambda_{p_1} e_{p_1}$, and $B^* e_{p_1} = \sum_{p_2}^{\infty} \langle e_{p_1}, B e_{p_2} \rangle \, e_{p_2}$, we indeed have

$$\begin{aligned} \sum_{p_1=1}^{\infty} \left\langle e_{p_1}, AB(AB)^{k-1} e_{p_1} \right\rangle &= \sum_{p_1=1}^{\infty} \lambda_{p_1} \left\langle B^* e_{p_1}, (AB)^{k-1} e_{p_1} \right\rangle \\ &= \sum_{p_1=1}^{\infty} \lambda_{p_1} \left\langle \sum_{p_2}^{\infty} \langle e_{p_1}, B e_{p_2} \rangle \, e_{p_2}, (AB)^{k-1} e_{p_1} \right\rangle \\ &= \sum_{p_1=1}^{\infty} \sum_{p_2=1}^{\infty} \lambda_{p_1} \langle e_{p_1}, B e_{p_2} \rangle \left\langle e_{p_2}, (AB)^{k-1} e_{p_1} \right\rangle, \quad \mathcal{P}(2) \ . \end{aligned}$$

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

$$\begin{aligned} & \sum_{p=1}^{\infty} \left\langle e_p, (AB)^k e_p \right\rangle \\ = & \sum_{p_1=1}^{\infty} \cdots \sum_{p_\ell=1}^{\infty} \left\{ \left(\prod_{j=1}^{\ell-1} \lambda_{p_j} \right) \left(\prod_{j=1}^{\ell-1} \langle e_{p_j}, B e_{p_{j+1}} \rangle \right) \left\langle e_{p_\ell}, AB(AB)^{k-\ell} e_{p_1} \right\rangle \right\} \\ = & \sum_{p_1=1}^{\infty} \cdots \sum_{p_\ell=1}^{\infty} \left\{ \left(\prod_{j=1}^{\ell-1} \lambda_{p_j} \right) \lambda_{p_\ell} \left(\prod_{j=1}^{\ell-1} \langle e_{p_j}, B e_{p_{j+1}} \rangle \right) \left\langle B^* e_{p_\ell}, (AB)^{k-\ell} e_{p_1} \right\rangle \right\} \\ = & \sum_{p_1=1}^{\infty} \cdots \sum_{p_\ell=1}^{\infty} \sum_{p_{\ell+1}=1}^{\infty} \left\{ \left(\prod_{j=1}^{\ell} \lambda_{p_j} \right) \left(\prod_{j=1}^{\ell-1} \langle e_{p_j}, B e_{p_{j+1}} \rangle \right) \langle e_{p_\ell}, B e_{p_{\ell+1}} \rangle \left\langle e_{p_{\ell+1}}, (AB)^{k-\ell} e_{p_1} \right\rangle \right\} \ , \end{aligned}$$

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

The proof is concluded by a k-step induction, that is once A in $(AB)^k$ is eventually marginalized out k-times and only the last term $\langle e_{p_k}, B e_{p_1} \rangle$ remains. □

Lemma 23. *Let $\gamma > 0$, and S a trace-class operator. Denote $\{\lambda_p\}_{p \geq 1}$ and $\{e_p\}_{p \geq 1}$ respectively the positive eigenvalues and the corresponding eigenvectors of S. Consider $d_r(T, \gamma)$*

for $r = 1, 2$, with T a compact operator, as defined in (6). If Δ is a trace-class perturbation operator such that $\left\|(S+\gamma \mathrm{I})^{-1}\Delta\right\| < 1$, and $\|\Delta\|_{\mathcal{C}_1} = \sum_{p=1}^{\infty} \|\Delta e_p\| < \gamma$, then

$$|d_r(S+\Delta,\gamma) - d_r(S,\gamma)| \leq \frac{\gamma^{-1}\|\Delta\|_{\mathcal{C}_1}}{1-\gamma^{-1}\|\Delta\|_{\mathcal{C}_1}}\,, \qquad \textit{for} \quad r = 1, 2\,. \tag{77}$$

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

$$|d_1(S+\Delta,\gamma) - d_1(S,\gamma)| \leq \frac{d_2(S,\gamma)\left\|S^{-1/2}\Delta S^{-1/2}\right\|_{\mathrm{HS}}}{1-d_2(S,\gamma)\left\|S^{-1/2}\Delta S^{-1/2}\right\|_{\mathrm{HS}}}\,, \tag{78}$$

$$|d_2(S+\Delta,\gamma) - d_2(S,\gamma)| \leq \frac{\left\|S^{-1/2}\Delta S^{-1/2}\right\|_{\mathrm{HS}}}{1-\left\|S^{-1/2}\Delta S^{-1/2}\right\|_{\mathrm{HS}}}\,. \tag{79}$$

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

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

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

$$\begin{aligned}|d_1(S+\Delta,\gamma) - d_1(S,\gamma)| &= \left|\mathrm{Tr}\left\{(S+\Delta+\gamma \mathrm{I})^{-1}(S+\Delta)\right\} - \mathrm{Tr}\left\{(S+\gamma \mathrm{I})^{-1}S\right\}\right|\\ &= \sum_{k=1}^{\infty}\left|\mathrm{Tr}\left\{\left\{(S+\gamma \mathrm{I})^{-1}\Delta\right\}^k\left\{(S+\gamma \mathrm{I})^{-1}S-\mathrm{I}\right\}\right\}\right| \leq \sum_{k=1}^{\infty}\left|\mathrm{Tr}\left\{\left((S+\gamma \mathrm{I})^{-1}\Delta\right)^k\right\}\right|\,.\end{aligned} \tag{80}$$

Applying Lemma 22 with $B = \Delta$, and $A = (S+\gamma \mathrm{I})^{-1}$, we obtain

$$\begin{aligned}\mathrm{Tr}\left\{\left((S+\gamma \mathrm{I})^{-1}\Delta\right)^k\right\} &= \sum_{p=1}^{\infty}\left\langle e_p, \left((S+\gamma \mathrm{I})^{-1}\Delta\right)^k e_p\right\rangle\\ &= \sum_{p_1=1}^{\infty}\cdots\sum_{p_k=1}^{\infty}\left\{\left(\prod_{j=1}^{k}(\lambda_{p_j}+\gamma)^{-1}\right)\left(\prod_{j=1}^{k}\left\langle e_{p_j}, \Delta e_{p_{j+1}}\right\rangle\right)\right\}\,.\end{aligned}$$

Since, for all $1 \leq j \leq k$, we have $\left|\left\langle e_{p_j}, \Delta e_{p_{j+1}}\right\rangle\right| \leq \left\|\Delta e_{p_j}\right\|$ and $(\lambda_{p_j}+\gamma)^{-1} \leq \gamma^{-1}$, the upper-bound in (80) is actually the sum of a geometric series whose ratio is $\gamma^{-1}\sum_{p=1}^{\infty}\|\Delta e_p\| = \gamma^{-1}\|\Delta\|_{\mathcal{C}_1}$, where $\gamma^{-1}\|\Delta\|_{\mathcal{C}_1} < 1$ by assumption, which completes the proof of (77) when $r = 1$. A similar reasoning as above allows to prove (77) when $r = 2$.

We now prove the second upper-bound (78). Using that

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

we may apply Lemma 22 again, but with $B = S^{-1/2}\Delta S^{-1/2}$, and $A = S^{1/2}(S+\gamma \mathrm{I})^{-1}S^{1/2}$, yielding

$$\begin{aligned}\mathrm{Tr}\Big\{\big((S+\gamma\mathrm{I})^{-1}\Delta\big)^k\Big\} &= \sum_{p=1}^{\infty}\Big\langle e_p, \big((S+\gamma\mathrm{I})^{-1}\Delta\big)^k e_p\Big\rangle \\ &= \sum_{p_1=1}^{\infty}\cdots\sum_{p_k=1}^{\infty}\left\{\left(\prod_{j=1}^{k}(\lambda_{p_j}+\gamma)^{-1}\lambda_{p_j}\right)\left(\prod_{j=1}^{k}\Big\langle e_{p_j}, \Big(S^{-1/2}\Delta S^{-1/2}\Big)\, e_{p_{j+1}}\Big\rangle\right)\right\}\,.\end{aligned}$$

Then, using that $\big|\big\langle e_{p_j}, \big(S^{-1/2}\Delta S^{-1/2}\big)\, e_{p_{j+1}}\big\rangle\big| \le \big\|\big(S^{-1/2}\Delta S^{-1/2}\big)\, e_{p_j}\big\|$, and applying Hölder inequality, we obtain

$$\begin{aligned}&\Big|\mathrm{Tr}\Big\{\big((S+\gamma\mathrm{I})^{-1}\Delta\big)^k\Big\}\Big| \\ \le\;& \left\{\sum_{p=1}^{\infty}(\lambda_p+\gamma)^{-2}\lambda_p^2\right\}^{k/2}\left\{\sum_{p_1=1}^{\infty}\cdots\sum_{p_k=1}^{\infty}\left(\prod_{j=1}^{k}\Big\langle e_{p_j}, \Big(S^{-1/2}\Delta S^{-1/2}\Big)\, e_{p_{j+1}}\Big\rangle^2\right)\right\}^{1/2} \\ \le\;& d^k(S)\,\Big\|S^{-1/2}\Delta S^{-1/2}\Big\|_{\mathrm{HS}}^k\,.\end{aligned}$$

Finally, going back to (80), the upper-bound is actually the sum of a geometric series whose ratio is $d(S)\,\big\|S^{-1/2}\Delta S^{-1/2}\big\|_{\mathrm{HS}}$, where $d(S)\,\big\|S^{-1/2}\Delta S^{-1/2}\big\|_{\mathrm{HS}} < 1$ by assumption, which completes the proof of (78). As for (79), observe that

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

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

Appendix C. Miscellaneous proofs

Proposition 24. *Assume (A1) and (B1). Assume in addition that $\mathbb{P}_1 = \mathbb{P}_2 = \mathbb{P}$. If $\gamma_n \equiv \gamma$, then*

$$Sup_x\Big|\mathbb{P}(T_\infty(\hat{\Sigma}_W,\gamma)\le x) - \mathbb{P}(T_\infty(\Sigma_W,\gamma)\le x)\Big| \to 0\,, \tag{81}$$

where $T_\infty(S,\gamma)$ for a trace-class operator S is defined in (10).

Proof. First, define the random variables $\{Y_n\}$ and $\{Y\}$ as follows

$$Y_n \stackrel{\text{def}}{=} \sum_{p=1}^{\infty} (\hat{\lambda}_p + \gamma)^{-1} \hat{\lambda}_p (Z_p^2 - 1) \ , \quad Y \stackrel{\text{def}}{=} \sum_{p=1}^{\infty} (\lambda_p + \gamma)^{-1} \lambda_p (Z_p^2 - 1) \ ,$$

where $\{Z_p\}_{p \geq 1}$ are independent standard normal variables. Considering the random element $h \in \mathcal{H}$, such that $\langle h, e_p \rangle_{\mathcal{H}} = Z_p$ for all $p \geq 1$, we may write

$$\begin{aligned} Y_n &= \left\| (\hat{\Sigma}_W + \gamma I)^{-1/2} \hat{\Sigma}_W^{-1/2} h \right\|_{\mathcal{H}}^2 - d_{1,n}(\hat{\Sigma}_W, \gamma) \ , \\ Y &= \left\| (\Sigma_W + \gamma I)^{-1/2} \Sigma_W^{-1/2} h \right\|_{\mathcal{H}}^2 - d_{1,n}(\Sigma_W, \gamma) \ . \end{aligned}$$

Then, using Eq. (77) for $r = 1$ in Lemma 23 with $S = \Sigma_W$, and Corollary 13 which gives $\left\| \hat{\Sigma}_W - \Sigma_W \right\|_{\text{HS}} = O_P(n^{-1/2})$, we get $|Y_n - Y| = O_P(n^{-1/2})$, and hence that $Y_n \xrightarrow{\text{P}} Y$ in case $\gamma_n \equiv \gamma$. Next, applying the Polya theorem (Lehmann and Romano, 2005, Theorem 11.2.9) gives the result

$$\text{Sup}_x \left| \mathbb{P}(Y_n \leq x) - \mathbb{P}(Y \leq x) \right| \to 0 \ .$$

□

Appendix D. Eigenvalues of covariance operators

In this section, we give new general results regarding the decay of eigenvalues of covariance operators. We assume that we have a bounded density $p(x)$ on $\mathbb{R}^p$ with respect to the Lebesgue measure, and a translation invariant kernel $k(x - y)$ with positive integrable Fourier transform. In this section, we consider eigenvalues of the second order moment operator, which dominates the covariance operator. From the proof of Proposition 10, the eigenvalues of the second order moment operator are the eigenvalues of the following operator from $L^2(\mathbb{R}^p)$ to $L^2(\mathbb{R}^p)$, defined as

$$Qf(x) = \int_{\mathbb{R}^p} p(x)^{1/2} k(x - y) f(y) p(y)^{1/2} dy$$

We let denote $\lambda_n(p, K)$ the eigenvalues of this operator ranked in decreasing order.

We let denote $T(p)$ the pointwise multiplication by p, defined from $L^2(\mathbb{R}^p)$ to $L^2(\mathbb{R}^p)$. We also denote $C(k)$ the convolution operator by k. We thus get $Q = T(p)^{1/2} C(k) T(p)^{1/2}$. Note that by taking Fourier transforms (P of p, and K of k), the eigenvalues are the same as the one of $T(K)^{1/2} C(P) T(K)^{1/2}$ and thus p and K plays equivalent roles (Widom, 1963).

The following lemma, taken from Widom (1964), gives an upperbound of the eigenvalues in the situation where p and K are indicator functions:

Lemma 25. *Let $\varepsilon > 0$. Then there exists $\delta > 0$ such that, if $p(x)$ is the indicator function of $[-1, 1]$ and K is the indicator function of $[-\gamma, \gamma]$, with $\gamma \leqslant (1 - \varepsilon) n\pi/2$, then $\lambda_n(p, K) \leqslant e^{-n\delta}$.*

This result is very useful because it is uniform in γ, as long as $\gamma \leqslant (1 - \varepsilon) n\pi/2$. We now take $\varepsilon = \frac{1}{2}$, and we thus get $\lambda_n(1_{[-1,1]}, 1_{[-n\pi/4, n\pi/4]}) \leqslant e^{-n\delta}$ for some $\delta > 0$.

We consider the tail behavior of $p(x)$ and of the Fourier transform $K(\omega)$ of k, through $M(p,u) = \max_{\|x\|_\infty \geqslant u} p(x)$ and $M(K,v) = \max_{\|\omega\|_\infty \geqslant v} K(\omega)$, where, for $x = (x_1, \ldots, x_p)$, $\|x\|_\infty = \max_{1 \leq i \leq p} |x_i|$. We also let denote $M_0(K)$ and $M_0(p)$ the supremum of K and p over $\mathbb{R}^d$.

Proposition 26. *For all (u,v) such that $uv = n\pi/4$, then*

$$\lambda_n(p,K) \leqslant M(p,u)M_0(K) + M(K,v)M_0(p) + M_0(K)M_0(p)e^{-\delta n^{1/p}}$$

Proof. We divide twice $\mathbb{R}^p$ in two parts, the spatial version $\mathbb{R}^p = \{x,\ \|x\|_\infty \leqslant u\} \cup \{x,\ \|x\|_\infty > u\} = A_u \cup B_u$ and the Fourier version $\mathbb{R}^p = \{\omega,\ \|\omega\|_\infty \leqslant v\} \cup \{\omega,\ \|\omega\|_\infty > v\} = A_v \cup B_v$. We have for all p and K,

$$\lambda_n(p,K) \leqslant \lambda_n(p1_{A_u}, K) + \lambda_1(p1_{B_u}, K)$$

which is classical results for perturbation of eigenvalues By definition of $M(p,u)$, we have $T(p1_{B_u}) \preccurlyeq M(p,u)\mathrm{I}$, and moreover $C(k) \preccurlyeq M_0(K)\mathrm{I}$, which implies that $\lambda_1(p1_{B_u}, K) \leqslant M(p,u)M_0(K)$. We thus get

$$\lambda_n(p,K) \leqslant \lambda_n(p1_{A_u}, K) + M(p,u)M_0(K).$$

Similarly, we get

$$\lambda_n(p,K) \leqslant \lambda_n(p1_{A_u}, K1_{A_v}) + M(p,u)M_0(K) + M(K,v)M_0(p).$$

We know that if two operators satisties $A \preccurlyeq B$, then $\lambda_n(A) \leqslant \lambda_n(B)$, thus since $T(p1_{A_u}) \preccurlyeq T(M_0(p)1_{Au})$ and similarly for K, we get

$$\lambda_n(p,K) \leqslant M_0(K)M_0(p)\lambda_n(1_{A_u}, 1_{A_v}) + M(p,u)M_0(K) + M(K,v)M_0(p)$$

By a simple change of variable, it easy to show that $\lambda_n(1_{A_u}, 1_{A_v}) = \lambda_n(1_{A_1}, 1_{A_{vu}})$ When $p = 1$, we immediately have $\lambda_n(1_{A_1}, 1_{A_{vu}}) \leqslant e^{-\delta n}$. When $p > 1$, then we notice that the eigenfunctions and eigenvalue of the operators will be product of eigenfunctions and eigenvectors of the univariate operators. That is, the eigenvalues are of the form $\mu_{i_1} \cdots \mu_{i_p}$ where $(i_1, \ldots, i_p)$ are positive integer and $\mu_i \leqslant e^{-\delta i}$ are eigenvalues of the univariate operator. From the product formulation, we get that if n is equal to the number of partitions of a certain integer k into p strictly positive integers, then $\lambda_n \leqslant e^{-\delta k}$. This number of partitions is exactly equal to $[(p-1)!(p-k)!]^{-1}(k-1)! \leqslant (k-1)^p$.

Thus, given any n, we can find an integer k such that $(k-1)^p \leqslant n$, and we have $\lambda_n(1_{A_1}, 1_{A_{vu}}) \leqslant e^{-\delta k}$. This leads to $\lambda_n(1_{A_1}, 1_{A_{vu}}) \leqslant e^{-\delta n^{1/p}}$. The proposition follows. □

We can now derive a number of corollaries:

Corollary 27. *If $p(x)$ is upper bounded by a constant times $e^{-\alpha\|x\|^2}$ and $K(\omega)$ is upper bounded by a constant times $e^{-\beta\|\omega\|^2}$, then there exists $\eta > 0$ such that $\lambda_n(p,K) = O(e^{-\eta n^{1/p}})$.*

Proof. Take $u = v = \sqrt{n\pi/4}$. □

Corollary 28. *If $p(x)$ is upper bounded by a constant times $(1+\|x\|)^{-\alpha}$ (with $\alpha > p$ such that we have integrability) and $K(\omega)$ is upper bounded by a constant times $e^{-\beta\|\omega\|^2}$, then $\lambda_n(p,K) = O(\frac{1}{n^{\alpha-\eta}})$ for any $\eta > 0$.*

Proof. Take v proportional to $n^{\eta/\alpha}$. □

Corollary 29. *If $p(x)$ is upper bounded by a constant times $(1+\|x\|)^{-\alpha}$ (with $\alpha > p$ such that we have integrability) and $K(\omega)$ is upper bounded by a constant times $(1+\|x\|)^{-\beta}$ (with $\beta > p$ such that we have integrability) , then $\lambda_n(p,K) = O(n^{-\alpha\beta/(\alpha+\beta)})$.*

Proof. Take v proportional to $n^{\alpha/(\alpha+\beta)}$. □

References

D. L. Allen. Hypothesis testing using an L_1-distance bootstrap. *The American Statistician*, 51(2):145–150, 1997.

N. H. Anderson, P. Hall, and D. M. Titterington. Two-sample test statistics for measuring discrepancies between two multivariate probability density functions using kernel-based density estimates. *Journal of Multivariate Analysis*, 50(1):41–54, 1994.

N. Aronszajn. Theory of reproducing kernels. *Trans. Am. Math. Soc.*, 68:337 – 404, 1950.

J.-P. Aubin. *Applied Functional Analysis (2nd ed.)*. Wiley-Interscience, 2000.

F. R. Bach, G. R. G. Lanckriet, and M. I. Jordan. Multiple kernel learning, conic duality, and the smo algorithm. In *Proc. ICML*, 2004.

P. Billingsley. *Probability and Measure (3rd ed.)*. Wiley Series in Probability and Mathematical Statistics. Wiley Interscience, 1995.

F. Bimbot, J.-F. Bonastre, C. Fredouille, G. Gravier, I. Magrin-Chagnolleau, S. Meignier, T. Merlin, J. Ortega-Garcia, D. Petrovska-Delacretaz, and D. A. Reynolds. A tutorial on text-independent speaker verification. *EURASIP*, 4:430–51, 2004.

G. Blanchard, O. Bousquet, and L. Zwald. Statistical properties of kernel principal component analysis. *Machine Learning*, 66(2-3), 2007.

K. Borgwardt, A. Gretton, M. Rasch, H.-P. Kriegel, Schölkopf, and A. J. Smola. Integrating structured biological data by kernel maximum mean discrepancy. *Bioinformatics*, 22(14): 49–57, 2006.

C. Butucea and K. Tribouley. Nonparametric homogeneity tests. *Journal of Statistical Planning and Inference*, 136(3):597–639, 2006.

B. Efron. The estimation of prediction error: Covariance penalties and cross-validation. *Journal of American Statistical Association*, 99(467):619–632, 2004.

K. Fukumizu, A. Gretton, X. Sunn, and B. Schölkopf. Kernel measures of conditional dependence. In *Adv. NIPS*, 2008.

U. Grenander and M. Miller. *Pattern Theory: from representation to inference*. Oxford Univ. Press, 2007.

A. Gretton, K. Borgwardt, M. Rasch, B. Schoelkopf, and A. Smola. A kernel method for the two-sample problem. In *Adv. NIPS*, 2006.

P. Hall and C. Heyde. *Martingale Limit Theory and Its Application*. Academic Press, 1980.

T. Hastie, R. Tibshirani, and J. Friedman. *The Elements of Statistical Learning*. Springer Series in Statistics. Springer, 2001.

P. Hughett. Error bounds for numerical inversion of a probability characteristic function. *SIAM Journal on Numerical Analysis*, 35:1368–1392, 1998.

E. Lehmann and J. Romano. *Testing Statistical Hypotheses (3rd ed.)*. Springer, 2005.

J. Louradour, K. Daoudi, and F. Bach. Feature space mahalanobis sequence kernels: Application to svm speaker verification. *IEEE Transactions on Audio, Speech and Language Processing*, 2007. To appear.

V. V. Petrov. *Limit Theorems of Probability Theory: sequences of independent random variables*. Oxford Studies in Probability. Oxford University Press, 1995.

R. Serfling. *Approximation Theorems in Statistics*. Wiley, 1980.

J. Shawe-Taylor and N. Cristianini. *Kernel Methods for Pattern Analysis*. Cambridge Univ. Press, 2004.

I. Steinwart, D. Hush, and C. Scovel. An explicit description of the reproducing kernel hilbert spaces of gaussian RBF kernels. *IEEE Transactions on Information Theory*, 52: 4635–4643, 2006a.

I. Steinwart, D. Hush, and C. Scovel. Function classes that approximate the bayes risk. In *Proceedings of the 19th Conference on Learning Theory (COLT 2006)*, 2006b.

R. L. Strawderman. Computing tail probabilities by numerical fourier inversion: the absolutely continuous case. *Statistica Sinica*, 14:175–201, 2004.

A. van der Vaart. *Asymptotic Statistics*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, 1998.

G. Wahba. *Spline Models for Observational Data*. SIAM, 1990.

H. Widom. Asymptotic behavior of the eigenvalues of certain integral equations I. *Transactions of the American Mathematical Society*, 109:278–295, 1963.

H. Widom. Asymptotic behavior of the eigenvalues of certain integral equations II. *Archive for Rational Mechanics and Analysis*, 17:215–229, 1964.