Testing for Homogeneity with Kernel Fisher Discriminant Analysis

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

Francis Bach
Willow Project, INRIA-ENS
45, rue d'Ulm, 75230 Paris, France
francis.bach@mines.org

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

Abstract

We propose to investigate test statistics for testing homogeneity based on kernel Fisher discriminant analysis. Asymptotic null distribution under null hypothesis is derived, and consistency against fixed alternatives is assessed. Finally, experimental evidence of the performance of the proposed approach on artificial data and on a speaker verification task is provided.

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 the input space (X, d), with common distributions $\mathbb{P}_1$ and $\mathbb{P}_2$, respectively. The problem consists in testing the null hypothesis $\mathbf{H}_0 : \mathbb{P}_1 = \mathbb{P}_2$ against the alternative $\mathbf{H}_A : \mathbb{P}_1 \neq \mathbb{P}_2$. This problem arises in many applications, ranging from computational anatomy [10] to process monitoring [7]. We shall allow the input space 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 [15]) arising in applications such as bioinformatics (see, *e.g.*, [4]).

Traditional approaches to this problem are based on distribution functions and use a certain distance between the empirical distributions obtained from the two samples. The most popular procedures are the two-sample Kolmogorov-Smirnov tests or the Cramer-Von Mises tests, that have been the standard for addressing these issues (at least when the dimension of the input space is small, and most often when $\mathsf{X} = \mathbb{R}$). Although these tests are popular due to their simplicity, they are known to be insensitive to certain characteristics of the distribution, such as densities containing high-frequency components or local features such as bumps. The low-power of the traditional density based statistics can be improved on using test statistics based on kernel density estimators [2] and [1] and wavelet estimators [6]. Recent work [4] has shown that one could difference in means in RKHSs in order to consistently test for homogeneity. In this paper, we show that taking into account the covariance structure in the RKHS allows to obtain simple limiting distributions.

The paper is organized as follows: in Section 2 and Section 3, we state the main definitions and we construct the test statistics. In Section 4, we give the asymptotic distribution of our test statistic under the null hypothesis, and investigate, the consistency and the power of the test for fixed alternatives. In

Section 5 we provide experimental evidence of the performance of our test statistic on both artificial and real datasets. Detailed proofs are presented in the last sections.

2 Mean and covariance in reproducing kernel Hilbert spaces

We first highlight the main assumptions we make in the paper on the reproducing kernel, then introduce operator-theoretic tools for working with distributions in infinite-dimensional spaces.

2.1 Reproducing kernel Hilbert spaces

Let (X, d) be a separable metric space, and denote by $\mathcal{X}$ the associated σ-algebra. Let X be X-valued random variable, with probability measure $\mathbb{P}$; the corresponding expectation is denoted $\mathbb{E}$. Consider a Hilbert space $(\mathcal{H}, \langle \cdot, \cdot \rangle_{\mathcal{H}})$ of functions from X to $\mathbb{R}$. The Hilbert space $\mathcal{H}$ is an RKHS if at each $x \in \mathsf{X}$, 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 $x \in \mathsf{X}$, there corresponds an element $\Phi(x) \in \mathcal{H}$ (we call Φ the feature map) 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 $k : \mathsf{X} \times \mathsf{X} \to \mathbb{R}$ is a positive definite kernel. We denote by $\|f\|_{\mathcal{H}} = \langle f, f \rangle_{\mathcal{H}}^{1/2}$ the associated norm. It is assumed in the remainder that $\mathcal{H}$ is a separable Hilbert space. Note that this is always the case if X is a separable metric space and if the kernel is continuous (see [16]). Throughout this paper, we make the following two assumptions on the kernel:

- (**A1**) The kernel k is bounded, that is $|k|_\infty = \sup_{(x,y) \in \mathsf{X} \times \mathsf{X}} k(x, y) < \infty$.
- (**A2**) For all probability measures $\mathbb{P}$ on $(\mathsf{X}, \mathcal{X})$, the RKHS associated with $k(\cdot, \cdot)$ is dense in $\mathsf{L}^2(\mathbb{P})$.

The asymptotic normality of our test statistics is valid without assumption (A2), while consistency results against fixed alternatives does need (A2). Assumption (A2) is true for translation-invariant kernels [8], and in particular for the Gaussian kernel on $\mathbb{R}^d$ [16]. Note that we do not require the compactness of X as in [16],

2.2 Mean element and covariance operator

We shall need some operator-theoretic tools to define mean elements and covariance operators in RKHS. 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 infimum 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 quantity $\|T\|_{\mathrm{HS}} = \{\sum_{p=1}^{\infty} \langle Te_p, Te_p \rangle_{\mathcal{H}}\}^{1/2}$ is finite, with $\{e_p\}_{p \geq 1}$ is a complete orthonormal basis of $\mathcal{H}$. The Hilbert-Schmidt norm $\|T\|_{\mathrm{HS}}$ 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 $(u \otimes v) f = \langle v, f \rangle_{\mathcal{H}} u$ (see [9]).

We recall below some basic facts about first and second-order moments of RKHS-valued random variables. If $\int k^{1/2}(x, x) \mathbb{P}(dx) < \infty$, the mean element $\mu_{\mathbb{P}}$ is defined for all functions $f \in \mathcal{H}$ as the unique element in $\mathcal{H}$ satisfying,

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

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 $f, g \in \mathcal{H}$,

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

Note that when assumption (A2) is satisfied, then the map from $\mathbb{P} \mapsto \mu_{\mathbb{P}}$ is injective. 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.

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

$$\hat{\mu} = n^{-1} \sum_{i=1}^{n} k(X_i, \cdot) \ , \quad \hat{\Sigma} = n^{-1} \sum_{i=1}^{n} k(X_i, \cdot) \otimes k(X_i, \cdot) - \hat{\mu} \otimes \hat{\mu} \ . \tag{3}$$

The operator Σ is a self-adjoint nonnegative trace-class operators. Hence, it can de diagonalized in an orthonormal basis, with a spectrum composed of a strictly decreasing sequence $\lambda_p > 0$ tending to zero and potentially a null space $\mathcal{N}(\Sigma)$ composed of functions f in $\mathcal{H}$ such that $\int \{f - \mathbb{P}f\}^2 d\mathbb{P} = 0$ [5], *i.e.,* functions which are constant in the support of $\mathbb{P}$.

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).

3 KFDA-based test statistic

In the feature space, the two-sample homogeneity test procedure can be formulated as follows. Given $\{X_1^{(1)}, \ldots, X_{n_1}^{(1)}\}$ and $\{X_1^{(2)}, \ldots, X_{n_2}^{(2)}\}$ from distributions $\mathbb{P}_1$ and $\mathbb{P}_2$, two independent identically distributed samples respectively from $\mathbb{P}_1$ and $\mathbb{P}_2$, having mean and covariance operators respectively given by (μ_1, Σ_1) and (μ_2, Σ_2), we wish to test the null hypothesis $\mathbf{H}_0$, $\mu_1 = \mu_2$ and $\Sigma_1 = \Sigma_2$, against the alternative hypothesis $\mathbf{H}_A, \mu_1 \neq \mu_2$.

In this paper, we tackle the problem by 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 [12]. 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). 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 Fisher discriminant ratio serves as a basis of our test statistics:

$$n \max_{f \in \mathcal{H}} \frac{\left\langle f, \hat{\Sigma}_B f \right\rangle_{\mathcal{H}}}{\left\langle f, (\hat{\Sigma}_W + \gamma_n \mathrm{I}) f \right\rangle_{\mathcal{H}}} = \frac{n_1 n_2}{n} \ \left\| (\hat{\Sigma}_W + \gamma_n \mathrm{I})^{-\frac{1}{2}} \hat{\delta} \right\|_{\mathcal{H}}^2 \ , \tag{4}$$

where I denotes the identity operator. Note that if the input space is Euclidean, *e.g.* $\mathsf{X} = \mathbb{R}^d$, the kernel is linear $k(x, y) = x^\top y$ and $\gamma_n = 0$, this quantity matches the so-called Hotelling's T^2-statistic in the two-sample case [13]. Moreover, in practice it may be computed thanks to the kernel trick, adapted to the kernel Fisher discriminant analysis and outlined in [15, Chapter 6]. 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, as $\gamma_n \to 0$ at an appropriate rate, 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-calibrated 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 $\lambda_p(S)$, the quantity $d_r(\Sigma, \gamma)$ is defined for all $q \geq 1$ as

$$d_r(\Sigma, \gamma) \stackrel{\text{def}}{=} \left\{ \sum_{p=1}^{\infty} (\lambda_p + \gamma)^{-r} \lambda_p^r \right\}^{1/r} \ . \tag{5}$$

4 Theoretical results

We consider in the sequel the following studentized test statistic:

$$\widehat{T}_n(\gamma_n) = \frac{\frac{n_1 n_2}{n} \left\| (\hat{\Sigma}_W + \gamma_n \mathrm{I})^{-1/2} \hat{\delta} \right\|_{\mathcal{H}}^2 - d_1(\hat{\Sigma}_W, \gamma_n)}{\sqrt{2} d_2(\hat{\Sigma}_W, \gamma_n)} \,. \tag{6}$$

In this paper, we first consider the asymptotic behavior of $\widehat{T}_n$ under the null hypothesis, and then against a fixed alternative. This will establish that our nonparametric test procedure is consistent in power.

4.1 Asymptotic normality under null hypothesis

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, *i.e.* $\mu_1 = \mu_2$ and $\Sigma_1 = \Sigma_2 = \Sigma$. As $\gamma_n \to 0$ tends to zero,

Theorem 1. *Assume (A1) and (B1). If $\mathbb{P}_1 = \mathbb{P}_2 = \mathbb{P}$ and if $\gamma_n + \gamma_n^{-1} n^{-1/2} \to 0$, then*

$$\widehat{T}_n(\gamma_n) \xrightarrow{\mathcal{D}} \mathcal{N}(0,1) \tag{7}$$

The proof is postponed to Section 7. Under the assumptions of Theorem 1, the sequence of tests that rejects the null hypothesis when $\widehat{T}_n(\gamma_n) \geq z_{1-\alpha}$, where $z_{1-\alpha}$ is the $(1-\alpha)$-quantile of the standard normal distribution, is asymptotically level α. Note that the limiting distribution does not depend on the kernel nor on the regularization parameter.

4.2 Power consistency

We study the power of the test based on $\widehat{T}_n(\gamma_n)$ under alternative hypotheses. The minimal requirement is to to prove that this sequence of tests is consistent in power. 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.

The following proposition shows that the limit is finite, strictly positive and independent of the kernel otherwise (see [8] for similar results for canonical correlation analysis). The following result gives some useful insights on $\left\| \Sigma_W^{-1/2} \delta \right\|_{\mathcal{H}}$, *i.e.*the population counterpart of $\left\| (\hat{\Sigma}_W^{-1/2} + \gamma_n \mathrm{I})^{-1/2} \hat{\delta} \right\|_{\mathcal{H}}$ on which our test statistics is based upon.

Proposition 2. *Assume (A1) and (A2). If $\gamma_n + \gamma_n^{-1} n^{-1/2} \to 0$, then for any probability distributions $\mathbb{P}_1$ and $\mathbb{P}_2$,*

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

where ν is any probability measure such that $\mathbb{P}_1$ and $\mathbb{P}_2$ are absolutely continuous w.r.t. ν and p_1 and p_2 are the densities of $\mathbb{P}_1$ and $\mathbb{P}_2$ with respect to ν.

The norm $\left\| \Sigma_W^{-1/2} \delta \right\|_{\mathcal{H}}^2$ is finite when the χ^2-divergence $\int p_1^{-1} (p_2 - p_1)^2 d\rho$ is finite. 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 3. *Assume (A1) and (A2). Let $\mathbb{P}_1$ and $\mathbb{P}_2$ be two distributions over $(\mathsf{X}, \mathcal{X})$, such that $\mathbb{P}_2 \neq \mathbb{P}_1$. If $\gamma_n + \gamma_n^{-1} n^{-1/2} \to 0$, then*

$$\mathbb{P}_{\boldsymbol{H}_A}(\widehat{T}_n(\gamma) > z_{1-\alpha}) \to \infty \,. \tag{8}$$

5 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.

$\gamma =$	10^{-1}	10^{-4}	10^{-7}	10^{-10}
KFDA	0.01±0.0032	0.11±0.0062	0.98±0.0031	0.99±0.0001
MMD	0.01±0.0023	id.	id.	id.

Table 1: Evolution of power of KFDA and MMD respectively, as γ goes to 0.

5.1 Artificial data

We shall focus here on a particularly simple setting, in order analyze the major issues arising in applying our approach in practice. Indeed, we consider the periodic smoothing spline kernel (see [17] for a detailed derivation), for which explicit formulae are available for the eigenvalues of the corresponding covariance operator when the underlying distribution is uniform. This allows us to alleviate the issue of estimating the spectrum of the covariance operator, and weigh up the practical impact of the regularization on the power of our test statistic.

Periodic smoothing spline kernel Consider X as the two-dimensional circle identified with the interval $[0, 1]$ (with periodicity conditions). We consider the strictly positive sequence $K_\nu = (2\pi\nu)^{-2m}$ and the following norm:

$$\|f\|_{\mathcal{H}}^2 = \frac{\langle f, c_0\rangle^2}{K_0} + \sum_{\nu>0} \frac{\langle f, c_\nu\rangle^2 + \langle f, s_\nu\rangle^2}{K_\nu}$$

where $c_\nu(t) = \sqrt{2}\cos 2\pi\nu t$ and $s_\nu(t) = \sqrt{2}\sin 2\pi\nu t$ for $\nu \geq 1$ and $c_0(t) = \mathbf{1}_{\mathsf{X}}$. This is always an RKHS norm associated with the following kernel

$$K(s,t) = \frac{(-1)^{m-1}}{(2m)!} B_{2m}((s-t) - \lfloor s-t \rfloor)$$

where B_{2m} is the $2m$-th Bernoulli polynomial. We have $B_2(x) = x^2 - x + 1/6$.

We consider the following testing problem

$$\begin{aligned} \mathbf{H}_0 : &\quad p_1 = p_2 \\ \mathbf{H}_A : &\quad p_2 \neq p_2 \end{aligned}$$

with p_1 the uniform density (i.e., the density with respect to the Lebesgue measure is equal to c_0), and densities $p_2 = p_1(c_0 + .25 * c_4)$. The covariance operator $\Sigma(p_1)$ has eigenvectors c_0, c_ν, s_ν with eigenvalues 0 for c_0 and K_ν for others.

Comparison with MMD We conducted experimental comparison in terms of power, for $m = 2$ and $n = 10^4$ and $\varepsilon = 0.5$. All quantities involving the eigenvalues of the covariance operator were computed from their counterparts instead of being estimated. The sampling from p_2^n was performed by inverting the cumulative distribution function. The table below displays the results, averaged over 10 Monte-Carlo runs.

5.2 Speaker verification

We conducted experiments in a speaker verification task [3] , on a subset of 8 female speakers using data from the NIST 2004 Speaker Recognition Evaluation. We refer the reader to [14] 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 $\alpha = 0.05$, since the empirical level seemed to match the prescribed for this value of the level as we noticed in previous subsection. 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.

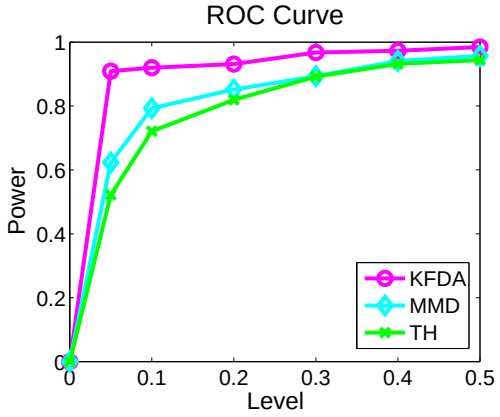

Figure 1: Comparison of ROC curves in a speaker verification task

6 Conclusion

We proposed a well-calibrated test statistic, built on kernel Fisher discriminant analysis, for which we proved that the asymptotic limit distribution under null hypothesis is standard normal distribution. Our test statistic can be readily computed from Gram matrices once a kernel is defined, and allows us to perform nonparametric hypothesis testing for homogeneity for high-dimensional data. The KFDA-test statistic yields competitive performance for speaker identification.

7 Sketch of proof of asymptotic normality under null hypothesis

Outline. The proof of the asymptotic normality of the test statistics under null hypothesis follows four steps. As a first step, we derive an asymptotic approximation of the test statistics as $\gamma_n + \gamma_n^{-1} n^{-1/2} \to 0$, where the only remaining stochastic term is $\hat{\delta}$. The test statistics is then spanned onto the eigenbasis of Σ, and decomposed into two terms B_n and C_n. The second step allows to prove the asymptotic negligibility of B_n, while the third step establishes the asymptotic normality of C_n by a martingale central limit theorem (MCLT). □

Step 1: $\widehat{T}_n(\gamma_n) = \tilde{T}_n(\gamma_n) + o_P(1)$. First, we may prove, using perturbation results of covariance operators, that, as $\gamma_n + \gamma_n^{-1} n^{-1/2} \to 0$, we have

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

□

For ease of notation, in the following, we shall often omit Σ in quantities involving it. Hence, from now on, $\lambda_p, \lambda_q, d_{2,n}$ stand for $\lambda_p(\Sigma), \lambda_q(\Sigma), d_2(\Sigma, \gamma_n)$. Define

$$Y_{n,p,i} \stackrel{\text{def}}{=} \begin{cases} \left(\frac{n_2}{n_1 n}\right)^{1/2} \left(e_p(X_i^{(1)}) - \mathbb{E}[e_p(X_1^{(1)})] \right) & 1 \le i \le n_1 \; , \\ -\left(\frac{n_1}{n_2 n}\right)^{1/2} \left(e_p(X_{i-n_1}^{(2)}) - \mathbb{E}[e_p(X_1^{(2)})] \right) & n_1 + 1 \le i \le n \; . \end{cases} \tag{10}$$

We now give formulas for the moments of $\{Y_{n,p,i}\}_{1 \le i \le n, p \ge 1}$, often used in the proof. Straightforward calculations give

$$\sum_{i=1}^{n} \mathbb{E}[Y_{n,p,i} Y_{n,q,i}] = \lambda_p^{1/2} \lambda_q^{1/2} \delta_{p,q} \; , \tag{11}$$

while the Cauchy-Schwarz inequality and the reproducing property give

$$\mathrm{Cov}(Y_{n,p,i}^2, Y_{n,q,i}^2) \le C n^{-2} |k|_\infty \lambda_p^{1/2} \lambda_q^{1/2} \; . \tag{12}$$

Denote $S_{n,p} \stackrel{\text{def}}{=} \sum_{i=1}^{n} Y_{n,p,i}$. Using Eq. (11), our test statistics now writes as $\tilde{T}_n = (\sqrt{2} d_{2,n})^{-1} A_n$ with

$$A_n \stackrel{\text{def}}{=} \frac{n_1 n_2}{n} \left\| (\Sigma + \gamma_n \mathrm{I})^{-1/2} \hat{\delta} \right\|^2 - d_{1,n} = \sum_{p=1}^{\infty} (\lambda_p + \gamma_n)^{-1} \left\{ S_{n,p}^2 - \mathbb{E} S_{n,p}^2 \right\} = B_n + 2C_n \ . \tag{13}$$

where B_n and C_n are defined as follows

$$B_n \stackrel{\text{def}}{=} \sum_{p=1}^{\infty} \sum_{i=1}^{n} \left\{ Y_{n,p,i}^2 - \mathbb{E} Y_{n,p,i}^2 \right\} \ , \tag{14}$$

$$C_n \stackrel{\text{def}}{=} \sum_{p=1}^{\infty} (\lambda_p + \gamma_n)^{-1} \sum_{i=1}^{n} Y_{n,p,i} \left\{ \sum_{j=1}^{i-1} Y_{n,p,j} \right\} \ . \tag{15}$$

Step 2: $B_n = o_P(1)$. The proof consists in computing 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

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

Using Eq. (12), we get

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

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 3: $d_{2,n}^{-1} C_n \xrightarrow{\mathcal{D}} \mathrm{N}(0, 1/2)$. We use the central limit theorem (MCLT) for triangular arrays of martingale differences (see e.g. [11, Theorem 3.2]). For $= 1, \ldots, n$, denote

$$\xi_{n,i} \stackrel{\text{def}}{=} d_{2,n}^{-1} \sum_{p=1}^{\infty} (\lambda_p + \gamma_n)^{-1} Y_{n,p,i} M_{n,p,i-1} \ , \quad \text{where} \quad M_{n,p,i} \stackrel{\text{def}}{=} \sum_{j=1}^{i} Y_{n,p,j} \ , \tag{16}$$

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, *i.e.* $\mathbb{E}\left[\xi_{n,i} \,|\, \mathcal{F}_{n,i-1} \right] = 0$. The first step in the proof of the CLT is to establish that

$$s_n^2 = \sum_{i=1}^{n} \mathbb{E}\left[\xi_{n,i}^2 \,\middle|\, \mathcal{F}_{n,i-1} \right] \xrightarrow{\mathrm{P}} 1/2 \ . \tag{17}$$

The second step of the proof is to establish the negligibility condition. We use [11, Theorem 3.2], which requires to establish that $\max_{1 \leq i \leq n} |\xi_{n,i}| \xrightarrow{\mathrm{P}} 0$ (smallness) and $\mathbb{E}(\max_{1 \leq i \leq n} \xi_{n,i}^2)$ is bounded in n (tightness), where $\xi_{n,i}$ is defined in (16). We will establish the two conditions simultaneously by checking that

$$\mathbb{E}\left(\max_{1 \leq i \leq n} \xi_{n,i}^2 \right) = o(1) \ . \tag{18}$$

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

$$D_n = d_{2,n}^{-2} \sum_{p=1}^{\infty} (\lambda_p + \gamma_n)^{-2} \sum_{i=1}^{n} M_{n,p,i-1}^2 \mathbb{E}[Y_{n,p,i}^2] \ , \tag{19}$$

$$E_n = d_{2,n}^{-2} \sum_{p \neq q} (\lambda_p + \gamma_n)^{-1} (\lambda_q + \gamma_n)^{-1} \sum_{i=1}^{n} M_{n,p,i-1} M_{n,q,i-1} \mathbb{E}[Y_{n,p,i} Y_{n,q,i}] \ . \tag{20}$$

Consider first the diagonal terms E_n. We first compute its mean. Note that $\mathbb{E}[M_{n,p,i}^2] = \sum_{j=1}^{i} \mathbb{E}[Y_{n,p,j}^2]$. Using Eq. (11) we get

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

Therefore, $\mathbb{E}[D_n] = 1/2 + o(1)$. Next, we may prove that $D_n - \mathbb{E}[D_n] = o_P(1)$ is negligible, by checking that $\mathrm{Var}[D_n] = o(1)$. We finally consider E_n defined in (20), and prove that $E_n = o_P(1)$ using Eq. (11). This concludes the proof of Eq. (17).

We finally show Eq. (18). Since $|Y_{n,p,i}| \leq n^{-1/2}|k|_\infty^{1/2}$ $\mathbb{P}$-a.s we may bound

$$\max_{1\leq i\leq n}|\xi_{n,i}| \leq Cd_{2,n}^{-1}n^{-1/2}\sum_{p=1}^{\infty}(\lambda_p+\gamma_n)^{-1}\max_{1\leq i\leq n}|M_{n,p,i-1}| . \tag{21}$$

Then, the Doob inequality implies that $\mathbb{E}^{1/2}[\max_{1\leq i\leq n}|M_{n,p,i-1}|^2] \leq \mathbb{E}^{1/2}[M_{n,p,n-1}^2] \leq C\lambda_p^{1/2}$. Plugging this bound in (21), 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 $\gamma_n + \gamma_n^{-1}n^{-1/2} \to 0$ and Assumption (B1). □

References

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

[2] 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.

[3] 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.

[4] 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.

[5] H. Brezis. *Analyse Fonctionnelle*. Masson, 1980.

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

[7] E. Carlstein, H. Müller, and D. Siegmund, editors. *Change-point Problems*, number 23 in IMS Monograph. Institute of Mathematical Statistics, Hayward, CA, 1994.

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

[9] I. Gohberg, S. Goldberg, and M. A. Kaashoek. *Classes of Linear Operators Vol. I*. Birkhäuser, 1990.

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

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

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

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

[14] 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.

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

[16] 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, 2006.

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