Submitted to Bernoulli

A One-Sample Test for Normality with Kernel Methods

JÉRÉMIE KELLNER and ALAIN CELISSE

Laboratoire de Mathématiques
UMR 8524 CNRS - Université Lille 1 - MODAL team-project Inria
59655, Villeneuve d'Ascq Cedex
E-mail: jeremie.kellner@ed.univ-lille1.fr; alain.celisse@math.univ-lille1.fr

We propose a new one-sample test for normality in a Reproducing Kernel Hilbert Space (RKHS). Namely, we test the null-hypothesis of belonging to a given family of Gaussian distributions. Hence our procedure may be applied either to test data for normality or to test parameters (mean and covariance) if data are assumed Gaussian. Our test is based on the same principle as the MMD (Maximum Mean Discrepancy) which is usually used for two-sample tests such as homogeneity or independence testing. Our method makes use of a special kind of parametric bootstrap (typical of goodness-of-fit tests) which is computationally more efficient than standard parametric bootstrap. Moreover, an upper bound for the Type-II error highlights the dependence on influential quantities. Experiments illustrate the practical improvement allowed by our test in high-dimensional settings where common normality tests are known to fail. We also consider an application to covariance rank selection through a sequential procedure.

MSC 2010 subject classifications: Primary 60K35, 60K35; secondary 60K35.

1. Introduction

Non-vectorial data such as DNA sequences or pictures often require a positive semi-definite kernel [1] which plays the role of a similarity function. For instance, two strings can be compared by counting the number of common substrings. Further analysis is then carried out in the associated reproducing kernel Hilbert space (RKHS), that is the Hilbert space spanned by the evaluation functions $k(x, .)$ for every x in the input space. Thus embedding data into this RKHS through the feature map $x \mapsto k(x, .)$ allows to apply linear algorithms to initially non-vectorial inputs.

Embedded data are often assumed to have a Gaussian distribution. For instance supervised and unsupervised classification are performed in [4] by modeling each class as a Gaussian process. In [31], outliers are detected by modelling embedded data as a Gaussian random variable and by removing points lying in the tails of that Gaussian distribution. This key assumption is also made in [36] where a mean equality test is used in high-dimensional setting. Moreover, Principal Component Analysis (PCA) and its kernelized version Kernel PCA [33] are known to be optimal for Gaussian data as these methods rely on second-order statistics (covariance). Besides, a Gaussian assumption allows to use Expectation-Minimization (EM) techniques to speed up PCA [32].

Depending on the (finite or infinite dimensional) structure of the RKHS, Cramer-von-Mises-type normality tests can be applied, such as Mardia's skewness test [27], the Henze-Zirkler test [21] and the Energy-distance test [39]. However these tests become less powerful as dimension increases (see Table 3 in [39]). An alternative approach consists in randomly projecting high-dimensional objects on one-dimensional directions and then applying univariate test on a few randomly chosen marginals [10]. This projection pursuit method has the advantage of being suited to high-dimensional settings. On the other hand, such approaches also suffer a lack of power because of the limited number of considered directions (see Section 4.2 in [10]).

In the RKHS setting, [16] introduced the Maximum Mean Discrepancy (MMD) which quantifies the gap between two distributions through distances between two elements of an RKHS. The MMD approach has been used for two-sample testing [16] and for independence testing (Hilbert Space Independence Criterion, [19]). However to the best of our knowledge, MMD has not been applied in a one-sample goodness-of-fit testing framework.

The main contribution of the present paper is to provide a one-sample statistical test of normality for data in a general Hilbert space (which can be an RKHS), by means of the MMD principle. This test features two possible applications: testing the normality of the data but also testing parameters (mean and covariance) if data are assumed Gaussian. The latter application encompasses many current methods that assume normality to make inferences on parameters, for instance to test the nullity of the mean [36] or to assess the sparse structure of the covariance [38, 2].

Once the test statistic is defined, a critical value is needed to decide whether to accept or reject the Gaussian hypothesis. In goodness-of-fit testing, this critical value is typically estimated by parametric bootstrap. Unfortunately, parametric bootstrap requires parameters to be computed several times, hence heavy computational costs (*i.e.* diagonalization of covariance matrices). Our test bypasses the recomputation of parameters by implementing a faster version of parametric bootstrap. Following the idea of [25], this fast bootstrap method "linearizes" the test statistic through a Fréchet derivative approximation and thus can estimate the critical value by a *weighted* bootstrap (in the sense of [6]) which is computationally more efficient. Furthermore our version of this bootstrap method allows parameters estimators that are not explicitly "linear" (*i.e.* that consist of a sum of independent terms) and that take values in possible infinite-dimensional Hilbert spaces.

Finally, we illustrate our test and present a sequential procedure that assesses the rank of a covariance operator. The problem of covariance rank estimation is adressed in several domains: functional regression [7, 5], classification [40] and dimension reduction methods such as PCA, Kernel PCA and Non-Gaussian Component Analysis [3, 11, 12] where the dimension of the kept subspace is a crucial problem.

Here is the outline of the paper. Section 2 sets our framework and Section 3 introduces the MMD and how it is used for our one-sample test. The new normality test is described in Section 4, while both its theoretical and empirical performances are detailed in Section 5 in terms of control of Type-I and Type-II errors. A sequential procedure to select covariance rank is presented in Section 6.

2. Framework

Let $(\mathcal{H}, \mathcal{A})$ be a measurable space, and $Y_1, \ldots, Y_n \in \mathcal{H}$ denote a sample of *independent and identically distributed* (*i.i.d.*) random variables drawn from an unknown distribution $P \in \mathcal{P}$, where $\mathcal{P}$ is a set of distributions defined on $\mathcal{A}$.

In our framework, $\mathcal{H}$ is a separable Hilbert space endowed with a dot product $\langle \, . \, , \, . \, \rangle_{\mathcal{H}}$ and the associated norm $\|.\|_{\mathcal{H}}$ (defined by $\|h\|_{\mathcal{H}} = \langle h, h \rangle_{\mathcal{H}}^{1/2}$ for any $h \in \mathcal{H}$). Our goal is to test whether Y_i is a *Gaussian random variable (r.v.)* of $\mathcal{H}$, which is defined as follows.

***Definition* 2.1.** (Gaussian random variable in a Hilbert space)
Let $(\Omega, \mathcal{F}, \mathbb{P})$ a measure space, $(\mathcal{H}, \mathcal{F}')$ a measurable space where $\mathcal{H}$ is a Hilbert space, and $Y : \Omega \to \mathcal{H}$ a measurable map.
Y is a Gaussian r.v. of $\mathcal{H}$ if $\langle Y, h \rangle_{\mathcal{H}}$ is a univariate Gaussian r.v. for any $h \in \mathcal{H}$.
Assuming that $\mathbb{E}_Y ||Y||_{\mathcal{H}} < +\infty$, there exists $m \in \mathcal{H}$ such that:

$$\forall h \in \mathcal{H}, \;\; \langle m, h \rangle_{\mathcal{H}} = \mathbb{E}_Y \langle Y, h \rangle_{\mathcal{H}} \;\; ,$$

and a (finite trace) operator $\Sigma : \mathcal{H} \to \mathcal{H}$ satisfying:

$$\forall h, h' \in \mathcal{H}, \;\; \langle \Sigma h, h' \rangle_{\mathcal{H}} = \mathrm{cov}(\langle Y, h \rangle_{\mathcal{H}}, \langle Y, h' \rangle_{\mathcal{H}}) \;\; .$$

m and Σ are respectively the mean and the covariance operator of Y. The distribution of Y is denoted $\mathcal{N}(m, \Sigma)$.

More precisely, the tested hypothesis is that Y_i follows a Gaussian distribution $\mathcal{N}(m_0, \Sigma_0)$, where $(m_0, \Sigma_0) \in \Theta_0$ and Θ_0 is a subset of the parameter space Θ. [1] Following [26], let us define the null hypothesis $\mathbf{H}_0 : \; P \in \mathcal{P}_0$, and the alternative hypothesis $\mathbf{H}_1 : \; P \notin \mathcal{P} \setminus \mathcal{P}_0$ where the subset of null-hypotheses $\mathcal{P}_0 \subseteq \mathcal{P}$ is

$$\mathcal{P}_0 = \{ \mathcal{N}(m_0, \Sigma_0) \mid (m_0, \Sigma_0) \in \Theta_0 \} \;\; .$$

The purpose of a statistical test $T(Y_1, \ldots, Y_n)$ of $\mathbf{H}_0$ against $\mathbf{H}_1$ is to distinguish between the null ($\mathbf{H}_0$) and the alternative ($\mathbf{H}_1$) hypotheses. It requires two elements: a statistic $n\hat{\Delta}^2$ (which we define in Section 4.1) that measures the gap between the empirical distribution of the data and the considered family of normal distributions $\mathcal{P}_0$, and a rejection region $\mathcal{R}_\alpha$ (at a level of confidence $0 < \alpha < 1$). $\mathbf{H}_0$ is accepted if and only if $n\hat{\Delta}^2 \notin \mathcal{R}_\alpha$. The rejection region is determined by the distribution of $n\hat{\Delta}^2$ under the null-hypothesis such that the probability of wrongly rejecting $\mathbf{H}_0$ (Type-I error) is controlled by α.

[1] The parameter space Θ is endowed with the dot product $\langle (m, \Sigma), (m', \Sigma') \rangle_\Theta = \langle m, m' \rangle_{\mathcal{H}} + \langle \Sigma, \Sigma' \rangle_{HS(\mathcal{H})}$, where $HS(\mathcal{H})$ is the space of Hilbert-Schmidt (finite trace) operators $\mathcal{H} \to \mathcal{H}$ and $\langle \Sigma, \Sigma' \rangle_{HS(\mathcal{H})} = \sum_{i \geq 1} \langle \Sigma e_i, \Sigma' e_i \rangle_{\mathcal{H}}$ for any complete orthonormal basis $(e_i)_{i \geq 1}$ of $\mathcal{H}$. Therefore, for any $\theta \in \Theta$, the tensor product $\theta^{\otimes 2}$ is defined as the operator $\Theta \to \Theta, \theta' \mapsto \langle \theta, \theta' \rangle_\Theta \theta$. For any $\theta \in \Theta$ and $\bar{h} \in H(K)$, the tensor product $\bar{h} \otimes \theta$ is the operator $\Theta \to H(K), \theta' \mapsto \langle \theta, \theta' \rangle_\Theta \bar{h}$.

3. The Maximum Mean Discrepancy (MMD)

Following [16] the gap between two distributions P and Q on $\mathcal{H}$ is measured by

$$\Delta(P,Q) = \sup_{f\in\mathcal{F}} |\mathbb{E}_{Y\sim P} f(Y) - \mathbb{E}_{Z\sim Q} f(Z)|, \tag{3.1}$$

where $\mathcal{F}$ is a class of real valued functions on $\mathcal{H}$. Regardless of $\mathcal{F}$, (3.1) always defines a pseudo-metric [2] on probability distributions [35].

The choice of $\mathcal{F}$ is subject to two requirements: *(i)* (3.1) must define a metric between distributions, that is

$$\forall P,Q,\ \ \Delta(P,Q) = 0 \Rightarrow P = Q\ \ , \tag{3.2}$$

and *(ii)* (3.1) must be expressed in an easy-to-compute form (without the supremum term).

To solve those two issues, several papers [17, 18, 19] have considered the case when $\mathcal{F}$ is the unit ball of a reproducing kernel Hilbert space (RKHS) $H(K)$ associated with a positive semi-definite kernel $K : \mathcal{H}\times\mathcal{H}\to\mathbb{R}$.

***Definition* 3.1.** (Reproducing Kernel Hilbert space, [1]) Let K be a symmetric, positive semi-definite kernel, i.e.

$$\forall x_1,\ldots,x_n\in\mathcal{H}, \forall\alpha_1,\ldots,\alpha_n, \sum_{i,j=1}^{n}\alpha_i\alpha_j K(x_i,x_j)\geq 0\ \ .$$

There exists a unique Hilbert space $H(K)$ of real-valued functions on $\mathcal{H}$ which satisfies:

- $\forall x\in\mathcal{H},\ \ K(x,.)\in H(K)$,
- $\forall f\in\mathcal{H},\ \ \forall x\in\mathcal{X},\ \ \langle f, K(x,.)\rangle_{H(K)} = f(x)$.

$H(K)$ is the reproducing kernel Hilbert space (RKHS) of K.

Let $\|.\|_{H(K)} = \langle .,.\rangle^{1/2}$ be the norm of $H(K)$ and $\mathcal{B}_1(K) = \{f\in H(K)\mid ||f||_{H(K)}\leq 1\}$ denote the unit ball of $H(K)$. When $\mathcal{F} = \mathcal{B}_1(K)$, $\Delta(\cdot,\cdot)$ becomes a metric only for a class of kernels K that are called *characteristic*.

***Definition* 3.2.** (Characteristic kernel, [14])
Let $\mathcal{F} = \mathcal{B}_1(K)$ in (3.1) for some kernel K. Then K is a characteristic kernel if $\Delta(P,Q) = 0$ implies $P = Q$.

Most common kernels are characteristic: Gaussian kernels $K(x,y) = \exp(-\sigma\|x-y\|^2_{\mathcal{H}})$ where $\sigma>0$, the exponential kernel $K(x,y) = \exp(\langle x,y\rangle_{\mathcal{H}})$ and Student kernels $K(x,y) = (1+\sigma||x-y||^2_{\mathcal{H}})^{-\alpha}$ where $\alpha,\sigma>0$, to name a few. Several criteria for a kernel to be characteristic exist (see [35, 15, 9]).

Moreover taking $\mathcal{F} = \mathcal{B}_1(K)$ enables to cast $\Delta(P,Q)$ as an easy to compute quantity. This is done by embedding any distribution P in the RKHS $H(K)$ as follows.

[2] A pseudo-metric $\Delta(.,.)$ satisfies for any P, Q,R : (i) $\Delta(P,P) = 0$, (ii) $\Delta(P,Q) = \Delta(Q,P)$, and (iii) $\Delta(P,R)\leq\Delta(P,Q)+\Delta(Q,R)$.

***Definition* 3.3.** (Hilbert space embedding, Lemma 3 from [17]) Let P be a distribution such that $\mathbb{E}_{Y\sim P}K^{1/2}(Y,Y) < +\infty$.
Then there exists $\mu_P \in H(K)$ such that for every $f \in H(K)$,

$$\langle \mu_P, f\rangle_{H(K)} = \mathbb{E}f(Y) \ .$$

μ_P is called the Hilbert space embedding of P in $H(K)$.

The existence of the embedding μ_P in Definition 3.3 is guaranteed by Riesz's representation theorem applied to the linear form $f \mapsto \mathbb{E}f(Y)$ where $f \in H(K)$, which is bounded due to the assumption $\mathbb{E}_{Y\sim P}K^{1/2}(Y,Y) < +\infty$.

Thus $\Delta(P,Q)$ can be expressed as the gap between the Hilbert space embeddings of P and Q (Lemma 4 in [17]):

$$\begin{aligned}\Delta(P,Q) &= \sup_{f\in H(K),||f||_{H(K)}\leq 1} |\mathbb{E}_P f(Y) - \mathbb{E}_Q f(Z)| \\ &= \sup_{f\in H(K),||f||_{H(K)}\leq 1} |\langle \mu_P - \mu_Q, f\rangle_{H(K)}| \\ &= ||\mu_P - \mu_Q||_{H(K)} \ . \end{aligned} \tag{3.3}$$

(3.3) is called the Maximum Mean Discrepancy (MMD) between P and Q.

Within our framework the goal is to compare P the true distribution of the data with a Gaussian distribution $P_0 = \mathcal{N}(m_0, \Sigma_0)$ for some $(m_0, \Sigma_0) \in \Theta_0$. Hence the quantity of interest is

$$\Delta^2 = \|\mu_P - \mu_{P_0}\|^2_{H(K)} \ . \tag{3.4}$$

For the sake of simplicity, we use the notation

$$\mu_{\mathcal{N}(m,\Sigma)} = N[m,\Sigma]$$

to denote the Hilbert space embedding of a Gaussian distribution.

4. Kernel normality test

This section introduces our one-sample test for normality based on the quantity (3.4). As said in Section 2, we test the null-hypothesis $\mathbf{H}_0 : P \in \{\mathcal{N}(m_0,\Sigma_0) \mid (m_0,\Sigma_0) \in \Theta_0\}$ where Θ_0 is a subset of the parameter space. Therefore our procedure may be used as a test for normality or a test on parameter if data are assumed Gaussian. The test procedure is summed up in Algorithm 1.

4.1. Test statistic

As in [16], Δ^2 can be estimated by replacing μ_P with the sample mean

$$\hat{\mu}_P = \mu_{\hat{P}} = (1/n)\sum_{i=1}^{n} K(Y_i, .) \ ,$$

Algorithm 1 Kernel Normality Test procedure

Input: $Y_1, \ldots, Y_n \in \mathcal{H}$, $K : \mathcal{H} \times \mathcal{H} \to \mathbb{R}$ (kernel) and $0 < \alpha < 1$ (test level).

1. Compute $\mathbf{K} = [\, \langle Y_i, Y_j \rangle \,]_{i,j}$ (Gram matrix).
2. Compute $n\hat{\Delta}^2$ (test statistic) from (4.1) that depends on $\mathbf{K}$ and K (Section 4.1)
3. (a) Draw B (approximate) independent copies of $n\hat{\Delta}^2$ under $\mathbf{H}_0$ by fast parametric bootstrap (Section 4.2).
   (b) Compute $\hat{q}_{\alpha,n}$ ($1-\alpha$ quantile of $n\hat{\Delta}^2$ under $\mathbf{H}_0$) from these replications.

Output: Reject $\mathbf{H}_0$ if $n\hat{\Delta}^2 > \hat{q}_{\alpha,n}$, and accept otherwise.

where $\hat{P} = (1/n)\sum_{i=1}^n \delta_{Y_i}$ is the empirical distribution. The null-distribution embedding $N[m_0, \Sigma_0]$ is estimated by $N[\tilde{m}, \tilde{\Sigma}]$ where $\tilde{m}$ and $\tilde{\Sigma}$ are appropriate and consistent (under $\mathbf{H}_0$) estimators of m_0 and Σ_0. This yields the estimator

$$\hat{\Delta}^2 = \|\hat{\mu}_P - N[\tilde{m}, \tilde{\Sigma}]\|^2_{H(K)} \ ,$$

which can be explicited by expanding the square of the norm and using the reproducing property of $H(K)$ as follows

$$\hat{\Delta}^2 = \frac{1}{n^2}\sum_{i,j=1}^n K(Y_i, Y_j) - \frac{2}{n}\sum_{i=1}^n N[\tilde{m}, \tilde{\Sigma}](Y_i) + \|N[\tilde{m}, \tilde{\Sigma}]\|^2_{H(K)} \ . \tag{4.1}$$

Proposition 4.1 ensures the consistency of the statistic (4.1).

***Proposition* 4.1.** Assume that P is Gaussian $\mathcal{N}(m_0, \Sigma_0)$ where $(m_0, \Sigma_0) \in \Theta_0$ and $(\tilde{m}, \tilde{\Sigma})$ are consistent estimators of (m_0, Σ_0). Also assume that K is continuous, $\mathbb{E}_P K(Y, Y) < +\infty$ and $N[m, \Sigma]$ is a continuous function of (m, Σ) on Θ_0. Then $\hat{\Delta}^2$ is a consistent estimator of Δ^2.

Proof. First, note that μ_P exists since $\mathbb{E}K(Y, Y) < +\infty$ implies $\mathbb{E}K^{1/2}(Y, Y) < +\infty$. Also by the continuity of K, the mapping $y \mapsto K(y, \cdot)$ is continuous and $\hat{\mu}_P$ as a function of $(Y_1, \ldots, Y_n) \in \mathcal{H}^n$ is also continuous hence measurable (with respect to Borel sets of $\mathcal{H}^n$ and $H(K)$), so that $\hat{\mu}_P$ is a proper $H(K)$-valued random variable. Therefore the Law of Large Numbers in Hilbert Spaces [22] can be applied and entails $\hat{\mu}_P \underset{n\to\infty}{\longrightarrow} \mu_P$ P-almost surely since $\mathbb{E}\|K(Y, \cdot) - \mu_P\|^2_{H(K)} = \mathbb{E}K(Y, Y) - \mathbb{E}K(Y, Y') \leq \mathbb{E}K(Y, Y) + \mathbb{E}^2 K(Y, Y') < +\infty$. The continuity of $N[m, \Sigma]$ (with respect to (m, Σ)) and the consistency of $(\tilde{m}, \tilde{\Sigma})$ yield $N[\tilde{m}, \tilde{\Sigma}] \overset{P-a.s.}{\underset{n\to\infty}{\longrightarrow}} N[m_0, \Sigma_0]$ P-a.s.. Finally, the continuity of $\|\cdot\|^2_{\mathcal{H}}$ leads to $\hat{\Delta}^2 \overset{P-a.s.}{\underset{n\to\infty}{\longrightarrow}} \Delta^2$. □

The expressions for $N[\tilde{m}, \tilde{\Sigma}](Y_i)$ and $\|N[\tilde{m}, \tilde{\Sigma}]\|^2_{H(K)}$ in (4.1) depend on the choice of K. Those are given by Propositions 4.2 and 4.3 when K is Gaussian and exponential. Note

that in these cases, the continuity assumption of $N[m, \Sigma]$ required by Proposition 4.1 is satisfied.

Before stating Propositions 4.2 and 4.3, the following notation is introduced. For a symmetric operator $L : \mathcal{H} \to \mathcal{H}$ with eigenexpansion $L = \sum_{r\geq 1} \lambda_r \Psi_r^{\otimes 2}$, its determinant is denoted $|L| = \prod_{r\geq 1} \lambda_r$. For any $q \in \mathbb{R}$, the operator L^q is defined as $L^q = \sum_{r\geq 1} \lambda_r^q \mathbb{1}_{\{\lambda_r > 0\}} \Psi_r^{\otimes 2}$.

***Proposition* 4.2.** (Gaussian kernel case) Let $K(.,.) = \exp(-\sigma\|. - .\|_{\mathcal{H}}^2)$ where $\sigma > 0$. Then,

$$N[\tilde{m}, \tilde{\Sigma}](.) = |I + 2\sigma\tilde{\Sigma}|^{-1/2} \exp\left(-\sigma||(I + 2\sigma\tilde{\Sigma})^{-1/2}(. - \tilde{m})||_{\mathcal{H}}^2\right) ,$$
$$\|N[\tilde{m}, \tilde{\Sigma}]\|_{H(K)}^2 = |I + 4\sigma\tilde{\Sigma}|^{-1/2} .$$

***Proposition* 4.3.** (Exponential kernel case) Let $K(.,.) = \exp(\langle ., . \rangle_{\mathcal{H}})$. Assume that the largest eigenvalue of $\tilde{\Sigma}$ is smaller than 1. Then,

$$N[\tilde{m}, \tilde{\Sigma}](.) = \exp\left(\langle \tilde{m}, . \rangle_{\mathcal{H}} + \frac{1}{2}\langle \tilde{\Sigma}., . \rangle_{\mathcal{H}}\right) ,$$
$$||N[\tilde{m}, \tilde{\Sigma}]||^2 = |I - \tilde{\Sigma}^2|^{-1/2} \exp\left(\|(I - \tilde{\Sigma}^2)^{-1/2}\tilde{m}\|_{\mathcal{H}}^2\right) .$$

The proofs of Propositions 4.2 and 4.3 are provided in Appendix B.1 in the supplemental article [24].

For most estimators $(\tilde{m}, \tilde{\Sigma})$, the quantities provided in Propositions 4.2 and 4.3 are computable via the Gram matrix $K = [\langle Y_i, Y_j \rangle_{\mathcal{H}}]_{1\leq i,j\leq n}$. For instance, assume that $(\tilde{m}, \tilde{\Sigma})$ are the classical estimators $(\hat{m}, \hat{\Sigma})$ where $\hat{m} = (1/n)\sum_{i=1}^n Y_i$ and $\hat{\Sigma} = (1/n)\sum_{i=1}^n (Y_i - \hat{m})^{\otimes 2}$. Let I_n and J_n be respectively the $n \times n$ identity matrix and the $n \times n$ matrix whose all entries equal 1, $H = I_n - (1/n)J_n$, and $K_c = HKH$ be the centered Gram matrix. Then for any $\square \in \mathbb{R}$,

$$\left|I + \square\hat{\Sigma}\right| = \det\left(I_n + \frac{\square}{n}K_c\right) ,$$

where det(.) denotes the (matrix) determinant function and

$$\left\|(I + \square\hat{\Sigma})^{-1/2} Y_i\right\|_{\mathcal{H}}^2 = \left[(I_n + \frac{\square}{n}K_c)^{-1}\right]_{i,i} ,$$

where $[.]_{ii}$ denotes the entry in the i-th row and the i-th column of a matrix.

4.2. Estimation of the critical value

Designing a test with confidence level $0 < \alpha < 1$ requires to compute the $1 - \alpha$ quantile of the $n\hat{\Delta}^2$ distribution under $\mathbf{H}_0$ denoted by $q_{\alpha,n}$. Thus $q_{\alpha,n}$ serves as a critical value to decide whether the test statistic $\hat{\Delta}^2$ is significantly close to 0 or not, so that the probability of wrongly rejecting $\mathbf{H}_0$ (Type-I error) is at most α.

4.2.1. Classical parametric bootstrap

In the case of a goodness-of-fit test, a usual way of estimating $q_{\alpha,n}$ is to perform a parametric bootstrap. Parametric bootstrap consists in generating B samples of n *i.i.d.* random variables $Y_1^{(b)}, \dots, Y_n^{(b)} \sim \mathcal{N}(\tilde{m}, \tilde{\Sigma})$ $(b = 1, \dots, B)$. Each of these B samples is used to compute a bootstrap replication

$$[n\hat{\Delta}^2]^b = n\|\hat{\mu}_P^b - N[\tilde{m}^b, \tilde{\Sigma}^b]\|^2_{H(K)} \ , \tag{4.2}$$

where $\hat{\mu}_P^b$, $\tilde{m}^b$ and $\tilde{\Sigma}^b$ are the estimators of μ_P, m and Σ based on $Y_1^b, \dots, Y_n^b$.

It is known that parametric bootstrap is asymptotically valid. Namely according to [37], under $\mathbf{H}_0$,

$$\forall b = 1, \dots, B, \quad \left(n\hat{\Delta}^2, \ [n\hat{\Delta}^2]^b\right) \xrightarrow[n\to+\infty]{\mathcal{L}} (U, U') \ ,$$

where U and U' are *i.i.d.* random variables. In a nutshell, (4.2) is approximately an independent copy of the test statistic $n\hat{\Delta}^2$ (under $\mathbf{H}_0$). Therefore B replications $[n\hat{\Delta}^2]^b$ can be used to estimate the $1-\alpha$ quantile $q_{\alpha,n}$ of $n\hat{\Delta}^2$ under the null-hypothesis.

However, this approach suffers heavy computational costs. In particular, each bootstrap replication involves the estimators $(\tilde{m}^b, \tilde{\Sigma}^b)$. In our case, this leads to compute the eigendecomposition of the B Gram matrices $K^b = [\langle Y_i^b, Y_j^b \rangle]_{i,j}$ of size $n \times n$ hence a complexity of order $\mathcal{O}(Bn^3)$.

4.2.2. Fast parametric bootstrap

This computational limitation is alleviated by means of another strategy described in [25]. Let us consider in a first time the case when the estimators of m and Σ are the classical empirical mean and covariance $\hat{m} = (1/n)\sum_{i=1}^n Y_i$ and $\hat{\Sigma} = (1/n)\sum_{i=1}^n (Y_i - \hat{m})^{\otimes 2}$. Introducing the Fréchet derivative [13] $D_{(m,\Sigma)}N$ at (m, Σ) of the function

$$N : \Theta \to H(K), \ \ (m, \Sigma) \mapsto N[m, \Sigma] \ ,$$

our bootstrap method relies on the following approximation under $\mathbf{H}_0$

$$\begin{aligned}
\sqrt{n}\left(\hat{\mu}_P - N[\hat{m}, \hat{\Sigma}]\right) &\simeq \sqrt{n}\left(\hat{\mu}_P - \underbrace{N[m_0, \Sigma_0]}_{=\mu_P \text{ under } \mathcal{H}_0} - D_{(m_0,\Sigma_0)}N[\hat{m} - m_0, \hat{\Sigma} - \Sigma_0]\right) \\
&= \frac{1}{\sqrt{n}}\sum_{i=1}^n [K(Y_i, .) - \mu_P] \\
&\quad - \frac{1}{\sqrt{n}}\sum_{i=1}^n D_{(m_0,\Sigma_0)}N[Y_i - m_0, (Y_i - m_0)^{\otimes 2} - \Sigma_0] \ . \qquad (4.3)
\end{aligned}$$

Since (4.3) consists of a sum of centered independent terms (under $\mathbf{H}_0$), it is possible to generate approximate independent copies of this sum via *weighted* bootstrap [6]. Given

$Z_1^b, \dots, Z_n^b$ *i.i.d.* real random variables of mean zero and unit variance and $\bar{Z}^b$ their empirical mean, a bootstrap replication of (4.3) is given by

$$\frac{1}{\sqrt{n}} \sum_{i=1}^{n} (Z_i^b - \bar{Z}^b) \left\{ K(Y_i, .) - D_{(m_0, \Sigma_0)} N[Y_i, (Y_i - m_0)^{\otimes 2}] \right\} \quad . \tag{4.4}$$

Taking the square of the norm of (4.4) in $H(K)$ and replacing the unknown true parameters m_0 and Σ_0 by their estimators $\hat{m}$ and $\hat{\Sigma}$ yields the bootstrap replication $[n\hat{\Delta}^2]_{fast}^b$ of $n\hat{\Delta}^2$

$$[n\hat{\Delta}^2]_{fast}^b \triangleq \left\| \sqrt{n} \left(\hat{\mu}_P^b - D_{(\hat{m}, \hat{\Sigma})} N[\hat{m}^b, \hat{\Sigma}^b] \right) \right\|_{H(K)}^2 \quad , \tag{4.5}$$

where

$$\begin{aligned} \hat{\mu}_P^b &= (1/n) \sum_{i=1}^{n} (Z_i^b - \bar{Z}^b) K(Y_i, .) \ , \\ \hat{m}^b &= (1/n) \sum_{i=1}^{n} (Z_i^b - \bar{Z}^b) Y_i \ , \\ \hat{\Sigma}^b &= (1/n) \sum_{i=1}^{n} (Z_i^b - \bar{Z}^b) (Y_i - \hat{m}^b)^{\otimes 2}. \end{aligned}$$

Therefore this approach avoids the recomputation of parameters for each bootstrap replication, hence a computational cost of order $\mathcal{O}(Bn^2)$ instead of $\mathcal{O}(Bn^3)$. This is illustrated empirically in the right half of Figure 1.

4.2.3. Fast parametric bootstrap for general parameter estimators

The bootstrap method proposed by [25] used in Section 4.2.2 requires that the estimators $(\tilde{m}, \tilde{\Sigma})$ can be written as a sum of independent terms with an additive term which converges to 0 in probability. Formally, $(\tilde{m}, \tilde{\Sigma}) = (m_0, \Sigma_0) + (1/n) \sum_{i=1}^{n} \psi(Y_i) + \epsilon_n$ where $\mathbb{E}\psi(Y) = 0$, $\mathrm{Var}(\psi(Y)) < +\infty$ and $\epsilon_n \xrightarrow[n \to +\infty]{\mathbb{P}} 0$. However there are some estimators which cannot be written in this form straightforwardly. This is the case for instance if we test whether data follow a Gaussian with covariance of fixed rank r (as in Section 6). In this example, the associated estimators are $\tilde{m} = \hat{m} = (1/n) \sum_{i=1}^{n} Y_i$ (empirical mean) and $\tilde{\Sigma} = \hat{\Sigma}_r = \sum_{s=1}^{r} \hat{\lambda}_s \hat{\Psi}_s^{\otimes 2}$ where $(\hat{\lambda}_s)_s$ and $(\hat{\Psi}_s)_s$ are the eigenvalues and eigenvectors of the empirical covariance operator $\hat{\Sigma} = (1/n) \sum_{i=1}^{n} (Y_i - \hat{\mu})^{\otimes 2}$.

We extend (4.5) to the general case when $\Theta_0 \neq \Theta$ and the estimators $(\tilde{m}, \tilde{\Sigma})$ are not the classical $(\hat{m}, \hat{\Sigma})$. We assume that the estimators $(\tilde{m}, \tilde{\Sigma})$ are functions of the empirical estimators $\hat{m}$ and $\hat{\Sigma}$, namely there exists a continuous mapping $\mathcal{T}$ such that

$$(\tilde{m}, \tilde{\Sigma}) = \mathcal{T}(\hat{m}, \hat{\Sigma}), \text{ where } \mathcal{T}(\Theta) \subseteq \Theta_0 \text{ and } \mathcal{T}|_{\Theta_0} = \mathrm{Id}_{\Theta_0}.$$

Under this definition, $(\tilde{m}, \tilde{\Sigma})$ are consistent estimators of (m, Σ) when $(m, \Sigma) \in \Theta_0$. This kind of estimators are met for various choices of the null-hypothesis:

- **Unknown mean and covariance:** $(\tilde{m}, \tilde{\Sigma}) = (\hat{m}, \hat{\Sigma})$ and $\mathcal{T}$ is the identity map Id_Θ,
- **Known mean and covariance:** $(\tilde{m}, \tilde{\Sigma}) = (m_0, \Sigma_0)$ and $\mathcal{T}$ is the constant map $\mathcal{T}(m, \Sigma) = (m_0, \Sigma_0)$,
- **Known mean and unknown covariance:** $(\tilde{m}, \tilde{\Sigma}) = (m_0, \hat{\Sigma})$ and $\mathcal{T}(m, \Sigma) = (m_0, \Sigma)$,
- **Unknown mean and covariance of known rank r:** $(\tilde{m}, \tilde{\Sigma}) = (\hat{m}, \hat{\Sigma}_r)$ and $\mathcal{T}(m, \Sigma) = (m, \Sigma_r)$ where Σ_r is the rank r truncation of Σ.

By introducing $\mathcal{T}$, we get a similar approximation to that in (4.3) by replacing the mapping $N : \Theta_0 \to H(K)$ with $No\mathcal{T} : \Theta_0 \to H(K)$. This leads to the bootstrap replication

$$[n\hat{\Delta}^2]^b_{fast} := \left\| \sqrt{n} \left(\hat{\mu}^b_P - D_{(\hat{m},\hat{\Sigma})}(No\mathcal{T})[\hat{m}^b, \hat{\Sigma}^b] \right) \right\|^2_{H(K)} \quad . \tag{4.6}$$

The validity of this bootstrap method is justified in Section 4.2.4.

Finally we define an estimator $\hat{q}_{\alpha,n}$ of $q_{\alpha,n}$ from the generated B bootstrap replications $[n\hat{\Delta}^2]^1_{fast} < \ldots < [n\hat{\Delta}^2]^B_{fast}$ (assuming they are sorted)

$$\hat{q}_{\alpha,n} = [n\hat{\Delta}^2]^{(\lfloor (1-\alpha)B \rfloor)} \quad ,$$

where $\lfloor . \rfloor$ stands for the integer part. The rejection region is defined by

$$\mathcal{R}_\alpha = \{ n\hat{\Delta}^2 > \hat{q}_{\alpha,n} \} \quad .$$

4.2.4. Validity of the fast parametric bootstrap

Proposition 4.4 hereafter shows the validity of the fast parametric bootstrap as presented in Section 4.2.3. The proof of Proposition 4.4 is provided in Section B.2.

***Proposition* 4.4.** Assume $\mathbb{E}_P K(Y,Y)$, $\mathrm{Tr}(\Sigma)$ and $\mathbb{E}_P ||Y - m_0||^4$ are finite. Also assume that $No\mathcal{T}$ is continuously differentiable on Θ_0 and that $D_\theta(No\mathcal{T})$ is bounded (for the operator norm) for every $\theta \in \Theta_0$.

If $\mathbf{H}_0$ is true, then for each $b = 1, \ldots, B$,

(i) $\sqrt{n} \left(\hat{\mu}_P - N[\tilde{m}, \tilde{\Sigma}] \right) \xrightarrow[n \to +\infty]{\mathcal{L}} G_P - D_{(m_0,\Sigma_0)}(No\mathcal{T})[U_P]$

(ii) $\sqrt{n} \left(\hat{\mu}^b_P - D_{(\hat{m},\hat{\Sigma})}(No\mathcal{T})[\hat{m}^b, \hat{\Sigma}^b] \right) \xrightarrow[n \to +\infty]{\mathcal{L}} G'_P - D_{(m_0,\Sigma_0)}(No\mathcal{T})[U'_P]$

where (G_P, U_P) and (G'_P, U'_P) are *i.i.d.* random variables in $H(K) \times \Theta$.
If otherwise $\mathbf{H}_0$ is false, (ii) is still true.
Furthermore, G_P and U_P are zero-mean Gaussian r.v. with covariances

$$\begin{aligned} \mathrm{Var}\,(G_P) &= \mathbb{E}_{Y \sim P}(K(Y,.) - \mu_P)^{\otimes 2} \\ \mathrm{Var}\,(U_P) &= \mathbb{E}_{Y \sim P} \left[Y - m_0, (Y - m_0)^{\otimes 2} - \Sigma \right]^{\otimes 2} \\ \mathrm{cov}\,(G_P, U_P) &= \mathbb{E}_{Y \sim P}(K(Y,.) - \mu_P) \otimes \left[Y - m_0, (Y - m_0)^{\otimes} - \Sigma_0 \right] \quad . \end{aligned}$$

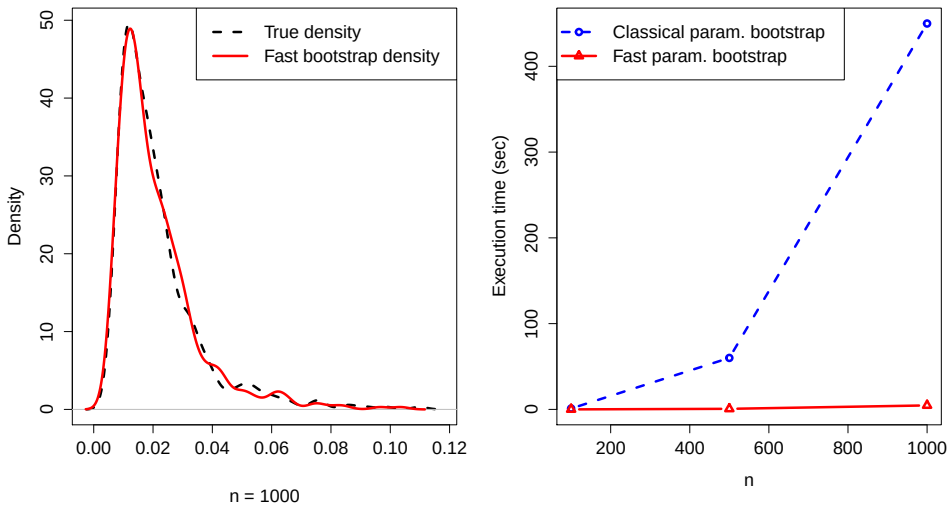

Figure 1. **Left:** Comparison of the distributions of $n\hat{\Delta}^2$ (test statistic) and $[n\hat{\Delta}^2]^b_{fast}$ (fast bootstrap replication) when $n = 1000$. A Kolmogorov-Smirnov two-sample test applied to our simulations returns a p-value of 0.978 which confirms the apparent similarity between the two distributions. **Right:** Comparison of the execution time (in seconds) of both classical and fast bootstrap methods.

By the Continous Mapping Theorem and the continuity of $\|.\|^2_{H(K)}$, Proposition 4.4 guarantees that the estimated quantile converges almost surely to the true one as $n, B \to +\infty$, so that the type-I error equals α asymptotically.

Note that in [25] the parameter subspace Θ_0 must be a subset of $\mathbb{R}^p$ for some integer $p \geq 1$. Proposition 4.4 allows Θ_0 to be a subset of a possibly infinite-dimensional Hilbert space (m belongs to $\mathcal{H}$ and Σ belongs to the space of finite trace operators $\mathcal{H} \to \mathcal{H}$).

Figure 1 (left plot) compares empirically the bootstrap distribution of $[n\hat{\Delta}^2]^b_{fast}$ and the distribution of $n\hat{\Delta}^2$. When $n = 1000$, the two corresponding densities are superimposed and a two-sample Kolmogorov-Smirnov test returns a p-value of 0.978 which confirms the strong similarity between the two distributions. Therefore the fast bootstrap method seems to provide a very good approximation of the distribution of $n\hat{\Delta}^2$ even for a moderate sample size n.

5. Test performances

5.1. An upper bound for the Type-II error

Let us assume the null-hypothesis is false, that is $P \neq \mathcal{N}(m_0, \Sigma_0)$ or $(m_0, \Sigma_0) \notin \Theta_0$. Theorem 5.1 gives the magnitude of the Type-II error, that is the probability of wrongly accepting $\mathbf{H}_0$. The proof can be found in Appendix B.3 in the supplemental article [24].

Before stating Theorem 5.1, let us introduce or recall useful notation :

- $\Delta = \|\mu_P - (NoT)[m_0, \Sigma_0]]\|_{H(K)}$,
- $q_{\alpha,n} = \mathbb{E}\hat{q}_{\alpha,n}$,
- $V_P^2 = \mathbb{E}_P||D_{(m_0,\Sigma_0)}(NoT)[\Psi(Y)] - K(Y,.) + \mu_P||^2_{H(K)}$,

where $\Psi(Y) = (Y - m_0, [Y - m_0]^{\otimes 2} - \Sigma_0)$ and $D_{(m_0,\Sigma_0)}(NoT)$ denotes the Fréchet derivative of NoT at (m_0, Σ_0). According to Proposition 4.4 and the continuous mapping theorem, $\hat{q}_{\alpha,n}$ corresponds to an order statistic of a random variable which converges weakly to $\left\|G'_P - D_{(m_0,\Sigma_0)}(NoT)[U'_P]\right\|^2$ (as defined in Proposition 4.4). Therefore, its mean $q_{\alpha,n}$ tends to a finite quantity as $n \to +\infty$. Δ and V_P^2 do not depend on n as well.

***Theorem* 5.1.** (Type II error) Assume $\sup_{x,y\in\mathcal{H}_0} |K(x,y)| = M < +\infty$ where $\mathcal{H}_0 \subseteq \mathcal{H}$ is the support of the alternative P and $\hat{q}_{\alpha,n}$ is independent of $n\hat{\Delta}^2$.
Then, for any $n > q_{\alpha,n}\Delta^{-2}$

$$\mathbb{P}\left(n\hat{\Delta}^2 \leq \hat{q}_{\alpha,n}\right) \leq \exp\left(-\frac{n\left(\Delta - \frac{q_{\alpha,n}}{n\Delta}\right)^2}{2V_P^2 + CV_P M^{1/2}(\Delta^2 - q_{\alpha,n}/n)}\right) f(\alpha, B, M, \Delta) \ , \qquad (5.1)$$

where

$$f(\alpha, B, M, \Delta) = (1 + o_n(1))\left(1 + \frac{C_{P^b}}{C'\Delta^2 M^{1/2} V_P \sqrt{\alpha B}} + \frac{o_B(B^{-1/2})}{C''\Delta^4 M V_P^2}\right) \ ,$$

and C, C', C'' are absolute constants and C_{P^b} only depends on the distribution of $[n\hat{\Delta}^2]^b_{fast}$.

The first implication of Proposition (5.1) is that our test is consistent, that is

$$\mathbb{P}(n\hat{\Delta}^2 \leq \hat{q}_{\alpha,n} \mid \mathbf{H}_0 \text{ false}) \underset{n\to+\infty}{\longrightarrow} 0 \ .$$

Furthermore, the upper bound in (5.1) reflects the expected behaviour of the Type-II error with respect to meaningful quantities. When Δ decreases, the bound increases (alternative more difficult to detect). When α (Type-I error) decreases, $q_{\alpha,n}$ gets larger and n has to be larger to get the bound. The variance term V_P^2 encompasses the difficulty of estimating μ_P and of estimating the parameters as well. In the special

case when m and Σ are known, $\mathcal{T} = Id$ and the chain rule yields $D_{(m_0,\Sigma_0)}(No\mathcal{T}) = (D_{\mathcal{T}(m_0,\Sigma_0)}N)o(D_{(m_0,\Sigma_0)}\mathcal{T}) = 0$ so that $V_P^2 = \mathbb{E}||\bar{\phi}(Y) - \mu_P||^2$ reduces to the variance of $\hat{\mu}_P$. As expected, a large V_P^2 makes the bound larger. Note that the estimation of the critical value which is related to the term $f(\alpha, B, M, \Delta)$ in (5.1) does not alter the asymptotic rate of convergence of the bound.

Remark that assuming that $\hat{q}_{\alpha,n}$ is independent of $n\hat{\Delta}^2$ is reasonable for a large n, since $n\hat{\Delta}^2$ and $\hat{q}_{\alpha,n}$ are asymptotically independent according to Proposition 4.4.

5.2. Empirical study of type-I/II errors

Empirical performances of our test are inferred on the basis of synthetic data. For the sake of brevity, our test is referred to as KNT (Kernel Normality Test) in the following.

One main concern of goodness-of-fit tests is their drastic loss of power as dimensionality increases. Empirical evidences (see Table 3 in [39]) prove ongoing multivariate normality tests suffer such deficiencies. The purpose of the present section is to check if KNT exhibits a good behavior in high or infinite dimension.

Throughout the following section, we have used a Gaussian kernel $K(x, y) = \exp(-\sigma\|x - y\|^2)$ where the parameter σ is set at the arbitrary value $\sigma = 1$. Kernel parameter choice for MMD-based tests is beyond the scope of this article and is actually a difficult problem which has been scarcely studied in the literature. See for instance [20] or [29] which both have to resort to a "linearized" version of the MMD statistic to tackle the problem, which yields a less powerful test.

5.2.1. Finite-dimensional case (Synthetic data)

Reference tests. The power of our test is compared with that of two multivariate normality tests: the Henze-Zirkler test (HZ) [21] and the energy distance (ED) test [39]. The main idea of these tests is briefly recalled in Appendix A.1 and A.2 in the supplemental article [24].

Null and alternative distributions. Two alternatives are considered: a mixture of two Gaussians with different means ($\mu_1 = 0$ and $\mu_2 = 1.5\ (1, 1/2, \ldots, 1/d)$) and same covariance $\Sigma = 0.5\ \mathrm{diag}(1, 1/4, \ldots, 1/d^2)$, whose mixture proportions equals either $(0.5, 0.5)$ (alternative HA1) or $(0.8, 0.2)$ (alternative HA2). Furthermore, two different cases for d are considered: $d = 2$ (small dimension) and $d = 100$ (large dimension).

Simulation design. 200 simulations are performed for each test, each alternative and each n (ranging from 100 to 500). B is set at $B = 250$ for KNT. The test level is set at $\alpha = 0.05$ for all tests.

Results. In the small dimension case (Figure 2, left column), the actual Type-I error of all tests remain more or less around α (± 0.02). Their Type-II errors are superimposed and quickly decrease down to 0 when $n \geq 200$. On the other hand, experimental results reveal different behaviors as d increases (Figure 2, right column). Whereas ED test lose power, KNT and HZ still exhibits small Type-II error values. Besides, ED and KNT Type-I errors remain around the prescribed level α while that of HZ is close to 1, which

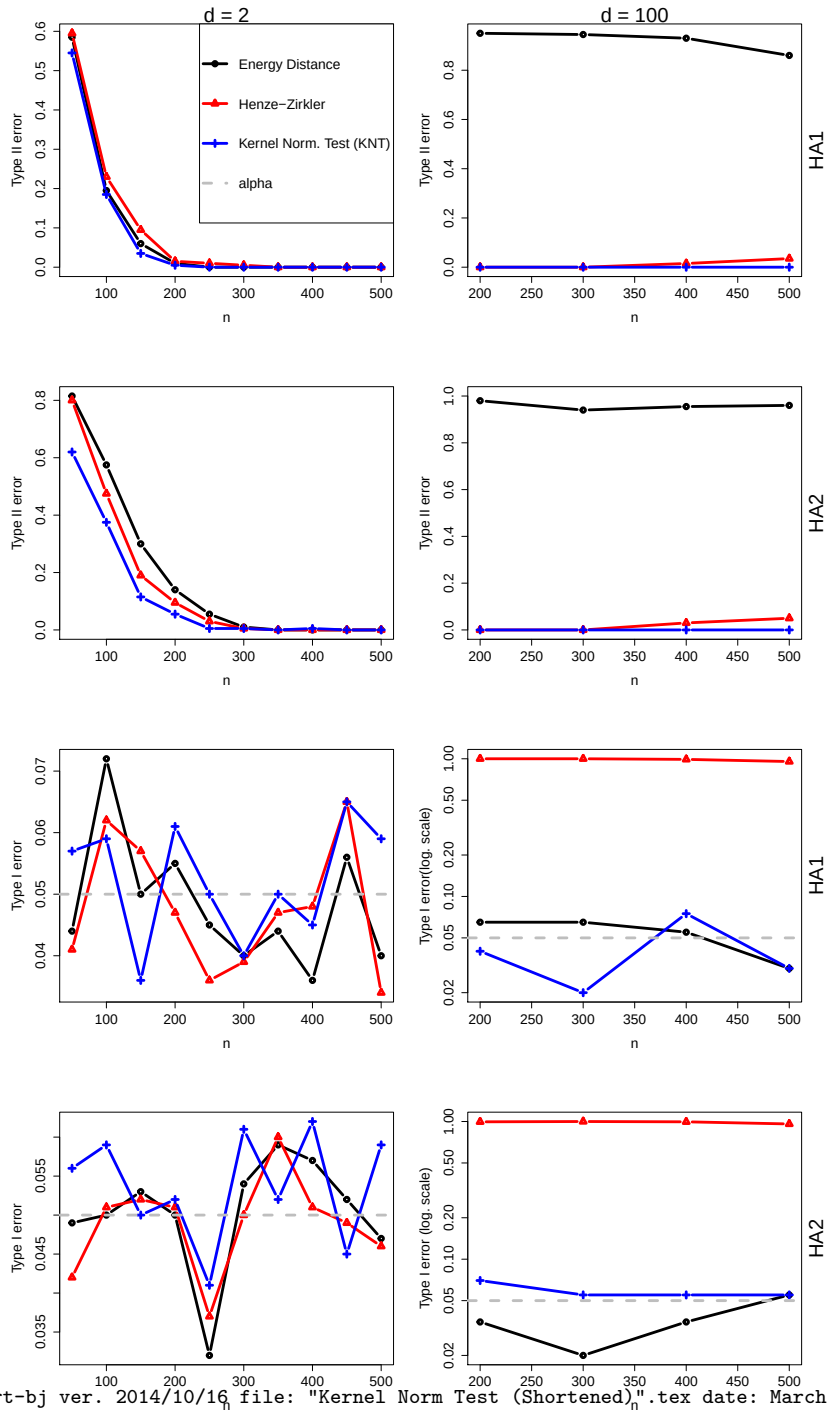

Figure 2. Type-I and type-II errors of KNT (+ blue), Energy Distance (∘ black), and Henze-Zirkler (△ red). Two alternative distributions are considered: HA1 (rows 1 and 3) and HA2 (rows 2 and 4). Two settings are considered: $d = 2$ (left) and $d = 100$ (right).

shows that its small Type-II error is artificial. This seems to confirm that HZ and ED tests are not suited to high-dimensional settings unlike KNT.

Interpretation. The results above may seem surprising as both KZ and ED can be seen as RKHS-based tests (see [34] for Energy Distance). The key difference between our test and HZ/ED tests is that the latter whiten (centering and normalizing) the data beforehand so that they reduce to a test for $\mathcal{N}(0, I_d)$ normality. The advantage of the whitening is that there is no need to compute the parameters of the Gaussian distribution for each bootstrap replication (since the tested Gaussian distribution has fixed parameters) hence a reduced computational cost. On the other hand testing normalized observations seems to induce more variance for the test statistic and the estimation of the critical value. To illustrate this, we have assessed the Type-I errors of KNT in additional experiments (whose results can be found in Appendix C.1 in the supplemental article [24]). These experiments show that in high-dimensional settings, the type-I error of our test is no longer controlled when either the trace of the covariance matrix is too large or when the eigenvalues of the covariance matrix decays too slowly, which may entail a loss of power. On the other hand the type-I error of KNT remains controlled when the number of dimensions increases while both the covariance trace and the decay of the eigenspectrum are fixed. Therefore in high-dimensional settings, our test is less sensitive to loss of power than existing tests by bypassing any whitening pre-process, and in the same time manages to avoid any additional computational costs by means of the fast parametric bootstrap.

5.2.2. Infinite-dimensional case (real data)

Dataset and chosen kernel. Let us consider the USPS dataset which consists of handwritten digits represented by a vectorized 8×8 greyscale matrix ($\mathcal{X} = \mathbb{R}^{64}$). A Gaussian kernel $k_G(\cdot, \cdot) = \exp(-\nu||\cdot - \cdot||^2)$ is used with $\nu = 10^{-4}$ to obtain a new dataset in the infinite-dimensional Hilbert space $\mathcal{H} = H(k_G)$. Comparing sub-datasets "Usps236" (keeping the three classes "2", "3" and "6", 541 observations) and "Usps358" (classes "3", "5" and "8", 539 observations), the 3D-visualization (Figure 3, top panels) suggests three well-separated Gaussian components for "Usps236" (left panel), and more overlapping classes for "Usps358" (right panel).

References tests. KNT is compared with Random Projection (RP) test, specially designed for infinite-dimensional settings. RP is presented in Appendix A.3 in the supplemental article [24]. Several numbers of projections p are considered for the RP test : $p = 1, 5$ and 15.

Simulation design. We set $\alpha = 0.05$ and 200 repetitions have been done for each sample size.

Results. (Figure 3, bottom plots) RP is by far less powerful KNT in both cases, no matter how many random projections p are considered. Indeed, KNT exhibits a Type-II error near 0 when n is barely equal to 100, whereas RP still has a relatively large Type-II error when $n = 400$. On the other hand, RP becomes more powerful as p gets larger as expected. A large enough number of random projections may allow RP to catch up KNT in terms of power. But RP has a computational advantage over KNT only when $p = 1$

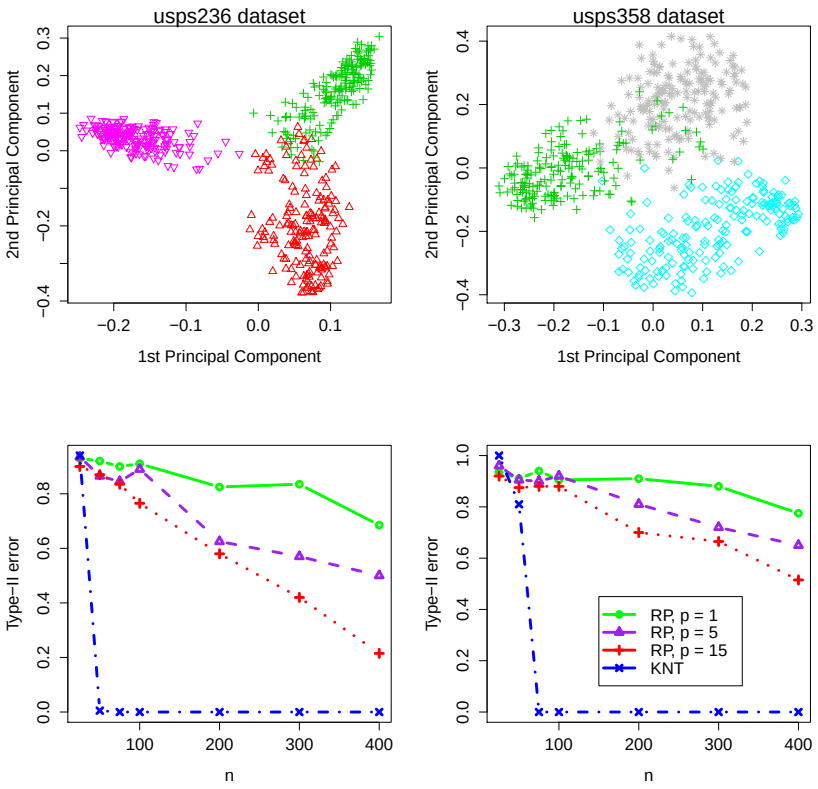

Figure 3. 3D-Visualization (Kernel PCA) of the "Usps236" (top row, left) and "Usps358" (top row, right) datasets; comparison of Type-II error (bottom row, left: "Usps236", right: "Usps358") for: KNT ($\times$ blue) and Random Projection with $p = 1$ ($\bullet$ green), $p = 5$ ($\triangle$ purple) and $p = 15$ ($+$ red) random projections.

where the RP test statistic is distribution-free. This is no longer the case when $p \geq 2$ and the critical value for the RP test is only available through Monte-Carlo methods.

6. Application to covariance rank selection

6.1. Covariance rank selection through sequential testing

Under the Gaussian assumption, the null hypothesis becomes

$$\mathbf{H}_0 : (m_0, \Sigma_0) \in \Theta_0 \ ,$$

and our test reduces to a test on parameters.

We focus on the estimation of the rank of the covariance operator Σ. Namely, we consider a collection of models $(\mathcal{M}_r)_{1 \leq r \leq r_{max}}$ such that, for each $r = 1, \ldots, r_{max}$,

$$\mathcal{M}_r = \{P = \mathcal{N}(m, \Sigma_r) \mid m \in H(k) \text{ and } \mathrm{rk}(\Sigma_r) = r\} \ .$$

Each of these models correspond respectively to the following null hypotheses

$$H_{0,r} : \mathrm{rank}(\Sigma) = r, \ \ r = 1, \ldots, r_{max} \ ,$$

and the corresponding tests can be used to select the most reliable model. These tests are performed in a sequential procedure summarized in Algorithm 2. This sequential procedure yields an estimator $\hat{r}$ defined as

$$\hat{r} \stackrel{\Delta}{=} \min_{\tilde{r}} \{H_{0,r} \text{ rejected for } r = 1, \ldots, \tilde{r} - 1 \text{ and } H_{0,\tilde{r}} \text{ accepted}\} \ .$$

or $\hat{r} \stackrel{\Delta}{=} r_{max}$ if all of the hypotheses are rejected.

Sequential testing to estimate the rank of a covariance matrix (or more generally a noisy matrix) is mentionned in [28] and [30]. Both of these papers focus on the probability to select a wrong rank, that is $\mathbb{P}(\hat{r} \neq r^*)$ where r^* denotes the true rank. The goal is to choose a level of confidence α such that this probability of error converges almost surely to 0 when $n \to +\infty$.

There are two ways of guessing a wrong rank : either by overestimation or by underestimation. Getting $\hat{r}$ greater than r^* implies that the null-hypothesis H_{0,r^*} was tested and wrongly rejected, hence a probability of overestimating r^* at most equal to α. Underestimating means that at least one of the false null-hypothesis $H_{0,1}, \ldots, H_{0,r^*-1}$ was wrongly accepted (Type-II error). Let $\beta_r(\alpha)$ denote the Type-II error of testing $H_{0,r}$ with confidence level α for each $r < r^*$. Thus by a union bound argument,

$$\mathbb{P}(\hat{r} \neq r^*) \leq \sum_{r=1}^{r^*-1} \beta_r(\alpha) + \alpha \ . \tag{6.1}$$

The bound in (6.1) decreases to 0 only if α converges to 0 but at a slow rate. Indeed, the Type-II errors $\beta_r(\alpha)$ grow with decreasing α but converge to zero when $n \to +\infty$. For instance in the case of the sequential tests mentionned in [28] and [30], the correct rate of decrease for α must satisfy $(1/n)\log(1/\alpha) = o_n(1)$.

Algorithm 2 Sequential selection of covariance rank

Input: Gram matrix $K = [K(Y_i, Y_j)]_{i,j}$, confidence level $0 < \alpha < 1$

1. Set $r = 1$ and test $H_{0,r}$
2. If $H_{0,r}$ is rejected and $r < r_{max}$, set $r = r + 1$ and return to 1.
3. Otherwise, set the estimator of the rank $\hat{r} = r$.

Output: estimated rank $\hat{r}$

6.2. Empirical performances

In this section, the sequential procedure to select covariance rank (as presented in Section 6.1) is tested empirically on synthetic data.

Dataset A sample of n zero-mean Gaussian with covariance Σ_{r^*} are generated, where n ranges from 100 to 5000. Σ_{r^*} is of rank $r^* = 10$ and its eigenvalues decrease either polynomially ($\lambda_r = r^{-1}$ for all $r \leq r^*$) or exponentially ($\lambda_r = \exp(-0.2r)$ for all $r \leq r^*$).

Benchmark To illustrate the level of difficulty, we compare our procedure with an oracle procedure which uses the knowledge of the true rank. Namely, the oracle procedure follows our sequential procedure at a level α_{oracle} defined as follows

$$\alpha_{oracle} = \max_{1 \leq r \leq r^*-1} \mathbb{P}_Z(n\hat{\Delta}_r^2 \leq Z_r) \ ,$$

where $n\hat{\Delta}_r^2$ is the observed statistic for the r-th test and Z_r follows the distribution of this statistic under $H_{0,r}$. Hence α_{oracle} is chosen such that the true rank r^* is selected whenever it is possible.

Simulation design To get a consistent estimation of r^*, the confidence level α must decrease with n and is set at $\alpha = \alpha_n = \exp(-0.125n^{0.45})$. Each time, 200 simulations are performed.

Results The top panels of Figure 4 display the proportion of cases when the target rank is found, either for our sequential procedure or the oracle one. When the eigenvalues decay polynomially, the oracle shows that the target rank cannot be almost surely guessed until $n = 1500$. When $n \leq 1500$, our procedure finds the true rank with probability at least 0.8 and quickly catches up to the oracle as n grows. In the exponential decay case, a similar observation is made. This case seems to be easier, as our procedure performs almost as well as the oracle when $n \geq 600$. In all cases, the consistency of our procedure is confirmed by the simulations.

The bottom panels of Figure 4 compare α with the probability of overestimating r^* (denoted by p_+). As noticed in Section 6.1, the former is an upper bound of the latter. But we must check empirically whether the gap between those two quantities is not too large, otherwise the sequential procedure would be too conservative and lead to excessive underestimation of r^*. In the polynomial decay case, the difference between α and p_+ is small, even when $n = 100$. The gap is larger in the exponential case but gets smaller when $n \geq 1500$.

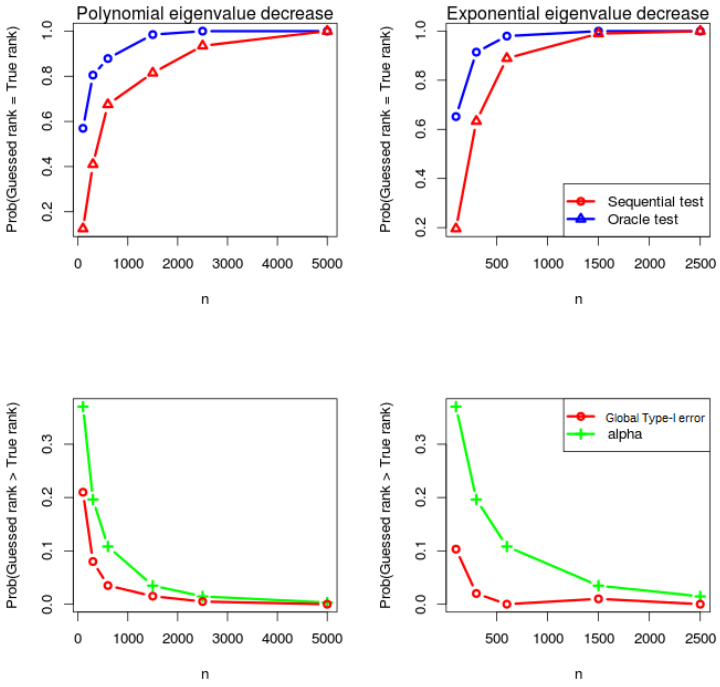

Figure 4. **Top half:** Probabilities of finding the right rank with respect to n for our sequential test ($\bullet$ red) and the oracle procedure ($\triangle$ blue); **bottom half:** probabilities of overestimating the true rank with the sequential procedure compared with fixed alpha ($+$ green). In each case, two decreasing rate for covariance eigenvalues are considered : polynomial (left column) and exponential (right column).

6.3. Robustness analysis

In practice, none of the models $\mathcal{M}_r$ is true. An additive full-rank noise term is often considered in the literature [8, 23]. Namely, we set in our case

$$Y = Z + \epsilon \tag{6.2}$$

where $Z \sim \mathcal{N}(m, \Sigma_{r^*})$ with $\mathrm{rk}(\Sigma_{r^*}) = r^*$ and ϵ is the error term independent of Z. Note that the Gaussian assumption concerns the main signal Z and not the error term whereas usual models assume the converse [8, 23].

Figure 5 illustrates the performance of our sequential procedure under the noisy model (6.2). We set $\mathcal{H} = \mathbb{R}^{100}$, $n = 600$, $r^* = 3$ and $\Sigma_{r^*} = \Sigma_3 = \mathrm{diag}(\lambda_1, \ldots, \lambda_3, 0, \ldots, 0)$ where $\lambda_r = \exp(-0.2r)$ for $r \leq 3$. The noise term is $\epsilon = (\lambda_3 \rho^{-1} \eta_i)_{1 \leq i \leq 100}$ where $\eta_1, \ldots, \eta_{100}$ are

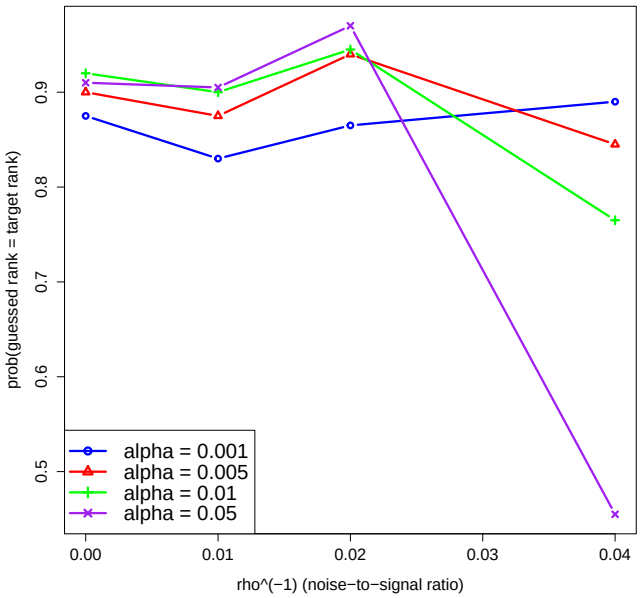

Figure 5. Illustration of the robustness of our sequential procedure under a noisy model.

i.i.d. Student random variables with 10 degrees of freedom and $\rho > 0$ is the *signal-to-noise ratio*.

As expected, the probability of guessing the target rank r^* decreases down to 0 as the signal-to-noise ratio ρ diminishes. However, choosing a smaller level of confidence α allows to improve the probability of right guesses for a fixed ρ. without sacrificing much for smaller signal-to-noise ratios. This is due to the fact that each null-hypothesis $H_{0,r}$ is false, hence the need for a smaller α (smaller Type-I error) which yields greater Type-II errors and avoids the rejection of all of the null-hypotheses.

7. Conclusion

We introduced a new normality test suited to high-dimensional Hilbert spaces. It turns out to be more powerful than ongoing high- or infinite-dimensional tests (such as random projection). In particular, empirical studies showed a mild sensibility to high-dimensionality. Therefore our test can be used as a multivariate normality (MVN) test without strongly suffering a loss of power when d gets larger unlike other MVN tests (Henze-Zirkler, Energy-distance).

If the Gaussian assumption is validated beforehand, our test becomes a general test on parameters. It is illustrated with an application to covariance rank selection that plugs our test into a sequential procedure. Empirical evidences show the good performances and the robustness of this method.

As for future improvements, investigating the influence of the kernel K on the performance of the test would be of interest. In the case of the Gaussian kernel for instance, a method to optimize the Type-II error with respect to the hyperparameter σ would be welcomed (see Appendix C.2 in the supplemental article [24] for simulations showing the influence of this hyperparameter). This aspect has just began to be studied in [20] when performing homogeneity testing with a convex combination of kernels.

Finally, the choice of the level α for the sequential procedure (covariance rank selection) is another subject for future research. Indeed, an asymptotic regime for α has been exhibited to get consistency, but setting the value of α when n is fixed remains an open question.

8. Supplement

The supplemental article [24] to this article features appendix sections. In Appendix A, normality tests mentionned throughout this article (such as Henze-Zirkler or Energy distance) are briefly introduced. In Appendix B, the proofs of the theorems presented in this article are detailed. Appendix C shows additional experiments. Finally, Appendix D explicitly shows closed-forms expressions for the Fréchet derivative of $N[\theta]$ for practitioners.

References

[1] N. Aronszajn. Theory of reproducing kernels. *Transactions of the American Mathematical Society*, May 1950.

[2] J. Bien and R. J. Tibshirani. Sparse estimation of a covariance matrix. *Biometrika*, 98(4):807–820, 2011.

[3] G. Blanchard, M. Sugiyama, M. Kawanabe, V. Spokoiny, and K.-R. Muller. Non-gaussian component analysis: a semi-parametric framework for linear dimension reduction. *NIPS*, 2006.

[4] C. Bouveyron, M. Fauvel, and S. Girard. Kernel discriminant analysis and clustering with parsimonious gaussian process models. *arXiv:1204.4021*, 2012.

[5] E. Brunel, A. Mas, and A. Roche. Non-asymptotic Adaptive Prediction in Functional Linear Models. *arXiv preprint arXiv:1301.3017*, 2013.

[6] M. D. Burke. Multivariate tests-of-fit and uniform confidence bands using a weighted bootstrap. *Statistics & Probability Letters*, 46(1):13–20, January 2000.

[7] H. Cardot and J. Johannes. Thresholding projection estimators in functional linear models. *Journal of Multivariate Analysis*, 101(2):395–408, 2010.

[8] Y. Choi, J. Taylor, and R. Tibshirani. Selecting the number of principal components: estimation of the true rank of a noisy matrix. *arXiv preprint arXiv:1410.8260*, 2014.

[9] A. Christmann and I. Steinwart. Universal kernels on non-standard input spaces. In *Advances in neural information processing systems*, pages 406–414, 2010.

[10] J. A. Cuesta-Albertos, E. del Barrio, R. Fraiman, and C. Matran. The random projection method in goodness of fit for functional data. *Computational Statistics & Data Analysis*, 51(10):4814–4831, June 2006.

[11] E. Diederichs, A. Juditsky, V. Spokoiny, and C. Schutte. Sparse non-Gaussian component analysis. *Information Theory, IEEE Transactions on*, 56(6):3033–3047, 2010.

[12] E. Diederichs, A. Juditsky, A.and Nemirovski, and V. Spokoiny. Sparse non Gaussian component analysis by semidefinite programming. *Machine learning*, 91(2):211–238, 2013.

[13] B. A. Frigyik, S. Srivastava, and M. R. Gupta. An introduction to functional derivatives. *Dept. Electr. Eng., Univ. Washington, Seattle, WA, Tech. Rep*, 1, 2008.

[14] K. Fukumizu, A. Gretton, X. Sun, and B. Schlkopf. Kernel Measures of Conditional Dependence. In *NIPS*, volume 20, pages 489–496, 2007.

[15] K. Fukumizu, B. Sriperumbudur, A. Gretton, and B. Schlkopf. Characteristic kernels on groups and semigroups. *NIPS*, 2009.

[16] A. Gretton, K. Borgwardt, M. Rasch, B. Schoelkopf, and A. Smola. A kernel method for the two-sample-problem. In B. Schoelkopf, J. Platt, and T. Hoffman, editors, *Advances in Neural Information Processing Systems*, volume 19 of *MIT Press, Cambridge*, pages 513–520, 2007.

[17] A. Gretton, K.M. Borgwardt, M.J. Rasch, B. Schlkopf, and A. Smola. A kernel two-sample test. *Journal of Machine Learning Research*, March 2012.

[18] A. Gretton, K. Fukumizu, Z. Harchaoui, and B. K. Sriperumbudur. A fast, consistent kernel two-sample test. *NIPS*, 2009.

[19] A. Gretton, K. Fukumizu, C. H. Teo, L. Song, B Schölkopf, and A. J. Smola. A kernel statistical test of independence. *NIPS*, 21, 2007.

[20] A. Gretton, D. Sejdinovic, H. Strathmann, S. Balakrishnan, M. Pontil, K. Fukumizu, and B. K. Sriperumbudur. Optimal kernel choice for large-scale two-sample tests. In *Advances in Neural Information Processing Systems*, pages 1205–1213, 2012.

[21] N. Henze and B. Zirkler. A class of invariant and consistent tests for multivariate normality. *Comm. Statist. Theory Methods*, 19:3595–3617, 1990.

[22] J. Hoffmann-Jorgensen and G. Pisier. The law of large numbers and the central limit theorem in banach spaces. *The Annals of Probability*, 4:587–599, 1976.

[23] J. Josse and F. Husson. Selecting the number of components in principal component analysis using cross-validation approximations. *Computational Statistics & Data Analysis*, 56(6):1869–1879, 2012.

[24] J. Kellner and A. Celisse. Supplement to "a one-sample test for normality with kernel methods". *Bernoulli*, 2018.

[25] I. Kojadinovic and J. Yan. Goodness-of-fit testing based on a weighted bootstrap: A fast large-sample alternative to the parametric bootstrap. *Canadian Journal of Statistics*, 40:480–500, 2012.

[26] E.L. Lehmann and J. P. Romano. *Testing Statistical hypotheses*. Springer, 2005.

[27] K.V. Mardia. Measures of multivariate skewness and kurtosis with applications. *Biometrika*, 57:519–530, 1970.

[28] Z. Ratsimalahelo. Strongly consistent determination of the rank of matrix. *Econometrics*, 2003.

[29] S. Reddi, A. Ramdas, B. Poczos, A. Singh, and L. Wasserman. On the high dimensional power of a linear-time two sample test under mean-shift alternatives. *Journal of Machine Learning Research*, pages 772–780, 2015.

[30] J.-M. Robin and R. J. Smith. Tests of rank. *Econometric Theory*, 16:151–175, 2000.

[31] V. Roth. Kernel Fisher Discriminants for Outlier Detection. *Neural Computation*, 18(4):942–960, 2006.

[32] S. Roweis. EM Algorithms for PCA and SPCA. In *in Advances in Neural Information Processing Systems*, pages 626–632. MIT Press, 1998.

[33] B. Schölkopf, A. Smola, and K.-R. Müller. Nonlinear Component Analysis as a Kernel Eigenvalue Problem. *Neural Computation*, 10:1299–1319, 1997.

[34] D. Sejdinovic, B. Sriperumbudur, A. Gretton, and K. Fukumizu. Equivalence of distance-based and rkhs-based statistics in hypothesis testing. *The Annals Of Statistics*, 41(5):2263–2291, 2013.

[35] B. K. Sriperumbudur, A. Gretton, K. Fukumiju, B. Schlkopf, and G. R. G. Lanckriet. Hilbert space embeddings and metrics on probability measures. *Journal of Machine Learning Research*, pages 1517–1561, 2010.

[36] M.S. Srivastava, S. Katayama, and Y. Kano. A two-sample test in high dimensional data. *Journal of Multivariate Analysis*, pages 349–358, 2013.

[37] W. Stute, W. G. Manteiga, and M. P. Quindimil. Bootstrap based goodness-of-fit-tests. *Metrika*, 40(1):243–256, 1993.

[38] T. Svantesson and J. W. Wallace. Tests for assessing multivariate normality and the covariance structure of MIMO data. In *Acoustics, Speech, and Signal Processing, 2003. Proceedings.(ICASSP'03). 2003 IEEE International Conference on*, volume 4, pages IV–656. IEEE, 2003.

[39] G.J. Szekely and R.L. Rizzo. A new test for multivariate normality. *Journal of Multivariate Analysis*, 93:58–80, 2005.

[40] L. Zwald. *Performances d'Algorithmes Statistiques d'Apprentissage: "Kernel Projection Machine" et Analyse en Composantes Principales Noyaux*. PhD thesis, Universit Paris XI U.F.R. Scientifique d'Orsay, 2005.