Comparing distributions: L1 geometry improves kernel two-sample testing

Meyer Scetbon, Gaël Varoquaux

Comparing distributions: ℓ_1 geometry improves kernel two-sample testing

Meyer Scetbon
CREST, ENSAE & Inria, Université Paris-Saclay

Gaël Varoquaux
Inria, Université Paris-Saclay

Abstract

Are two sets of observations drawn from the same distribution? This problem is a two-sample test. Kernel methods lead to many appealing properties. Indeed state-of-the-art approaches use the L^2 distance between kernel-based distribution representatives to derive their test statistics. Here, we show that L^p distances (with $p \geq 1$) between these distribution representatives give metrics on the space of distributions that are well-behaved to detect differences between distributions as they metrize the weak convergence. Moreover, for analytic kernels, we show that the L^1 geometry gives improved testing power for scalable computational procedures. Specifically, we derive a finite dimensional approximation of the metric given as the ℓ_1 norm of a vector which captures differences of expectations of analytic functions evaluated at spatial locations or frequencies (i.e, features). The features can be chosen to maximize the differences of the distributions and give interpretable indications of how they differs. Using an ℓ_1 norm gives better detection because differences between representatives are dense as we use analytic kernels (non-zero almost everywhere). The tests are consistent, while much faster than state-of-the-art quadratic-time kernel-based tests. Experiments on artificial and real-world problems demonstrate improved power/time tradeoff than the state of the art, based on ℓ_2 norms, and in some cases, better outright power than even the most expensive quadratic-time tests.

We consider two sample tests: testing whether two random variables are identically distributed without assumption on their distributions. This problem has many applications such as data integration [4] or automated model checking [21]. Distances between distributions underlie progress in unsupervised learning with generative adversarial networks [19, 1]. A kernel on the sample space can be used to build the Maximum Mean Discrepancy (MMD) [11, 13, 12, 25], a metric on distribution which has the strong propriety of metrizing the weak convergence of probability measures. It leads to non-parametric two-sample tests using the reproducing kernel Hilbert space (RKHS) distance [15, 9], or energy distance [31, 3]. The MMD has a quadratic computational cost, which may force to use of subsampled estimates [32, 14]. [5] approximate the L^2 distance between distribution representatives in the RKHS, to compute in linear time a pseudo metric over the space of distributions. Such approximations are related to random (Fourier) features, used in kernels algorithms [23, 18]. Distribution representatives can be mean embeddings [29, 28] or smooth characteristic functions [5, 16].

We first introduce the state of the art on Kernel-based two-sample testing built from the L^2 distance between mean embeddings in the RKHS. In fact, a wider family of distance is well suited for the two-sample problem: we show that for any $p \geq 1$, the L^p distance between these distribution representatives is a metric on the space of Borel probability measures that metrizes their weak convergence. We then define our ℓ_1-based statistic derived from the L^1 geometry and study its asymptotic behavior. We consider the general case where the number of samples of the two distributions may differ. We show that using the ℓ_1 norm provides a better testing power. Indeed, test statistics approximate such metrics and are defined as the norm of a J-dimensional vector which is the difference between the two distribution representatives at J locations. Under the alternative hypothesis H_1: $P \neq Q$, the

33rd Conference on Neural Information Processing Systems (NeurIPS 2019), Vancouver, Canada.

analyticity of the kernel ensures that all the features of this vector are non zero almost surely. We show that the ℓ_1 norm captures this dense difference better than the ℓ_2 norm and leads to better tests. We show also that improvements of Kernel two-sample tests established with the ℓ_2 norm [16] hold in the ℓ_1 case: optimizing features and the choice of kernel. We adapt the construction in the frequency domain as in [5]. Finally, we show that on 4 synthetic and 3 real-life problems, our new ℓ_1-based tests outperform the state of the art.

1 Prior art: kernel embeddings for two-sample tests

Given two samples $X := \{x_i\}_{i=1}^n$, $Y := \{y_i\}_{i=1}^n \subset \mathcal{X}$ independently and identically distributed (i.i.d.) according to two probability measures P and Q on a metric space $(\mathcal{X}, d)$ respectively, the goal of a two-sample test is to decide whether P is different from Q on the basis of the samples. Kernel methods arise naturally in two-sample testing as they provide Euclidean norms over the space of probability measures that metrize the convergence in law. To define such a metric, we need first to introduce the notion of Integral Probability Metric (IPM):

$$\text{IPM}[F, P, Q] := \sup_{f \in F} \Big(\mathbb{E}_{x \sim P} \left[f(x) \right] - \mathbb{E}_{y \sim Q} \left[f(y) \right] \Big) \tag{1}$$

where F is an arbitrary class of functions. When F is the unit ball B_k in the RKHS H_k associated with a positive definite bounded kernel $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$, the IPM is known as the Maximum Mean Discrepancy (MMD) [11], and it can be shown that the MMD is equal to the RKHS distance between so called mean embeddings [12],

$$\text{MMD}[P, Q] = \|\mu_P - \mu_Q\|_{H_k} \tag{2}$$

where μ_P is an embedding of the probability measure P to H_k,

$$\mu_P(t) := \int_{\mathbb{R}^d} k(x, t) dP(x) \tag{3}$$

and $\|.\|_{H_k}$ denotes the norm in the RKHS H_k. Moreover for kernels said to be *characteristic* [10], *eg* Gaussian kernels, $\text{MMD}[P, Q] = 0$ if and only if $P = Q$ [11]. In addition, when the kernel is bounded, and $\mathcal{X}$ is a compact Hausdorff space, [27] show that the MMD metrizes the weak convergence. Tests between distributions can be designed using an empirical estimation of the MMD.

A drawback of the MMD is the computation cost of empirical estimates, these being the sum of two U-statistics and an empirical average, with a quadratic cost in the sample size.

[5] study a related expression defined as the L^2 distance between mean embeddings of Borel probability measures:

$$d^2_{L^2,\mu}(P, Q) := \int_{t \in \mathbb{R}^d} \Big| \mu_P(t) - \mu_Q(t) \Big|^2 d\Gamma(t) \tag{4}$$

where Γ is a Borel probability measure. They estimate the integral (4) with the random variable,

$$d^2_{\ell_2,\mu,J}(P, Q) := \frac{1}{J} \sum_{j=1}^{J} \Big| \mu_P(T_j) - \mu_Q(T_j) \Big|^2 \tag{5}$$

where $\{T_j\}_{j=1}^J$ are sampled i.i.d. from the distribution Γ. This expression still has desirable metric-like properties, provided that the kernel is *analytic*:

Definition 1.1 (Analytic kernel)**.** *A positive definite kernel $k : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$ is* analytic *on its domain if for all $x \in \mathbb{R}^d$, the feature map $k(x, .)$ is an analytic function on $\mathbb{R}^d$.*

Indeed, for k a definite positive, characteristic, analytic, and bounded kernel on $\mathbb{R}^d$, [5] show that $d_{\ell_2,\mu,J}$ is a random metric[1] from which consistent two-sample test can be derived. By denoting μ_X and μ_Y respectively the empirical mean embeddings of P and Q,

$$\mu_X(T) := \frac{1}{n} \sum_{i=1}^{n} k(x_i, T), \qquad \mu_Y(T) := \frac{1}{n} \sum_{i=1}^{n} k(y_i, T)$$

[1] A random metric is a random process which satisfies all the conditions for a metric 'almost surely' [5].

[5] show that for $\{T_j\}_{j=1}^J$ sampled from the distribution Γ, under the null hypothesis $H_0 : P = Q$, as $n \to \infty$, the following test statistic:

$$\widehat{d}^2_{\ell_2,\mu,J}[X,Y] := n \sum_{j=1}^{J} \Big| \mu_X(T_j) - \mu_Y(T_j) \Big|^2 \tag{6}$$

converges in distribution to a sum of correlated chi-squared variables. Moreover, under the alternative hypothesis $H_1 : P \neq Q$, $\widehat{d}^2_{\ell_2,\mu,J}[X,Y]$ can be arbitrarly large as $n \to \infty$, allowing the test to correctly reject H_0. For a fixed level α, the test rejects H_0 if $\widehat{d}^2_{\ell_2,\mu,J}[X,Y]$ exceeds a predetermined test threshold, which is given by the $(1-\alpha)$-quantile of its asymptotic null distribution. As it is very computationally costly to obtain quantiles of this distribution, [5] normalize the differences between mean embeddings, and consider instead the test statistic ME[X,Y]:=$\|\sqrt{n}\mathbf{\Sigma}_n^{-1/2}\mathbf{S}_n\|_2^2$ where $\mathbf{S}_n := \frac{1}{n}\sum_{i=1}^n \mathbf{z}_i$, $\mathbf{\Sigma}_n := \frac{1}{n-1}\sum_{i=1}^n (\mathbf{z}_i - \mathbf{S}_n)(\mathbf{z}_i - \mathbf{S}_n)^T$, and $\mathbf{z}_i := (k(x_i, T_j) - k(y_i, T_j))_{j=1}^J \in \mathbb{R}^J$. Under the null hypothesis H_0, asymptotically the ME statistic follows $\chi^2(J)$, a chi-squared distribution with J degrees of freedom. Moreover, for k a translation-invariant kernel, [5] derive another statistical test, called the SCF test (for Smooth Characteristic Function), where its statistic SCF$[X,Y]$ is of the same form as the ME test statistic with a modified $\mathbf{z}_i := [f(x_i)\sin(x_i^T T_j) - f(y_i)\sin(y_i^T T_j), f(x_i)\cos(x_i^T T_j) - f(y_i)\cos(y_i^T T_j)]_{j=1}^J \in \mathbb{R}^{2J}$ where f is the inverse Fourier transform of k, and show that under H_0, SCF$[X,Y]$ follows asymptotically $\chi^2(2J)$.

2 A family of metrics that metrize of the weak convergence

[5] build their ME statistic by estimating the L^2 distance between mean embeddings. This metric can be generalized using any L^p distance with $p \geq 1$. These metrics are well suited for the two-sample problem as they metrize the weak convergence (see proof in supp. mat. A.1):

Theorem 2.1. *Given $p \geq 1$, k a definite positive, characteristic, continuous, and bounded kernel on $\mathbb{R}^d$, μ_P and μ_Q the mean embeddings of the Borel probability measures P and Q respectively, the function defined on $\mathcal{M}_+^1(\mathbb{R}^d) \times \mathcal{M}_+^1(\mathbb{R}^d)$:*

$$d_{L^p,\mu}(P,Q) := \left(\int_{t \in \mathbb{R}^d} \Big| \mu_P(t) - \mu_Q(t) \Big|^p d\Gamma(t) \right)^{1/p} \tag{7}$$

is a metric on the space of Borel probability measures, for Γ a Borel probability measure absolutely continuous with respect to Lebesgue measure. Moreover a sequence $(\alpha_n)_{n\geq 0}$ of Borel probability measures converges weakly towards α if and only if $d_{L^p,\mu}(\alpha_n, \alpha) \to 0$.

Therefore, as the MMD, these metrics take into account the geometry of the underlying space and metrize the convergence in law. If we assume in addition that the kernel is analytic, we will show that deriving test statistics from the L^1 distance instead of the L^2 distance improves the test power for two-sample testing.

3 Two-sample testing using the ℓ_1 norm

3.1 A test statistic with simple asymptotic distribution

From now, we assume that k is a positive definite, characteristic, analytic, and bounded kernel.

The statistic presented in eq. 6 is based on the ℓ_2 norm of a vector that capture differences between distributions in the RKHS at J locations. We will show that using an ℓ_1 norm instead of an ℓ_2 norm improves the test power (Proposition 3.1). It captures better the geometry of the problem. Indeed, when $P \neq Q$, the differences between distributions are dense which allow the ℓ_1 norm to reject better the null hypothesis H_0: $P = Q$.

We now build a consistent statistical test based on an empirical estimation of the L^1 metric introduced in eq. 7:

$$\widehat{d}_{\ell_1,\mu,J}[X,Y] := \sqrt{n} \sum_{j=1}^{J} \Big| \mu_X(T_j) - \mu_Y(T_j) \Big| \tag{8}$$

where $\{T_j\}_{j=1}^J$ are sampled from the distribution Γ. We show that under H_0, $\widehat{d}_{\ell_1,\mu,J}[X,Y]$ converges in distribution to a sum of correlated Nakagami variables[2] and under H_1, $\widehat{d}_{\ell_1,\mu,J}[X,Y]$ can be arbitrary large as $n \to \infty$ (see supp. mat. C.1). For a fixed level α, the test rejects H_0 if $\widehat{d}_{\ell_1,\mu,J}[X,Y]$ exceeds the $(1-\alpha)$-quantile of its asymptotic null distribution. We now compare the power of the statistics based respectively on the ℓ_2 norm (eq. 6) and the ℓ_1 norm (eq. 8) at the same level $\alpha > 0$ and we show that the power of the test using the ℓ_1 norm is better with high probability (see supp. mat. C.2):

Proposition 3.1. *Let $\alpha \in]0,1[$, $\gamma > 0$ and $J \geq 2$. Let $\{T_j\}_{j=1}^J$ sampled i.i.d. from the distribution Γ and let $X := \{x_i\}_{i=1}^n$ and $Y := \{y_i\}_{i=1}^n$ i.i.d. samples from P and Q respectively. Let us denote δ the $(1-\alpha)$-quantile of the asymptotic null distribution of $\widehat{d}_{\ell_1,\mu,J}[X,Y]$ and β the $(1-\alpha)$-quantile of the asymptotic null distribution of $\widehat{d}^2_{\ell_2,\mu,J}[X,Y]$. Under the alternative hypothesis, almost surely, there exists $N \geq 1$ such that for all $n \geq N$, with a probability of at least $1-\gamma$ we have:*

$$\widehat{d}^2_{\ell_2,\mu,J}[X,Y] > \beta \Rightarrow \widehat{d}_{\ell_1,\mu,J}[X,Y] > \delta \tag{9}$$

Therefore, for a fixed level α, under the alternative hypothesis, when the number of samples is large enough, with high probability, the ℓ_1-based test rejects better the null hypothesis. However, even for fixed $\{T_j\}_{j=1}^J$, computing the quantiles of these distributions requires a computationally-costly bootstrap or permutation procedure. Thus we follow a different approach where we allow the number of samples to differ. Let $X := \{x_i\}_{i=1}^{N_1}$ and $Y := \{y_i\}_{i=1}^{N_2}$ i.i.d according to respectively P and Q. We define for any sequence of $\{T_j\}_{j=1}^J$ in $\mathbb{R}^d$:

$$\mathbf{S}_{N_1,N_2} := \Big(\mu_X(T_1) - \mu_Y(T_1), ..., \mu_X(T_J) - \mu_Y(T_J)\Big) \tag{10}$$

$$\mathbf{Z}_X^i := (k(x_i,T_1), ..., k(x_i,T_J)) \in \mathbb{R}^J \qquad \mathbf{Z}_Y^j := (k(y_j,T_1), ..., k(y_j,T_J)) \in \mathbb{R}^J$$

And by denoting:

$$\mathbf{\Sigma}_{N_1} := \frac{1}{N_1-1}\sum_{i=1}^{N_1}(\mathbf{Z}_X^i - \overline{\mathbf{Z}}_X)(\mathbf{Z}_X^i - \overline{\mathbf{Z}}_X)^T \qquad \mathbf{\Sigma}_{N_2} := \frac{1}{N_2-1}\sum_{j=1}^{N_2}(\mathbf{Z}_Y^j - \overline{\mathbf{Z}}_Y)(\mathbf{Z}_Y^j - \overline{\mathbf{Z}}_Y)^T$$

$$\mathbf{\Sigma}_{N_1,N_2} := \frac{\mathbf{\Sigma}_{N_1}}{\rho} + \frac{\mathbf{\Sigma}_{N_2}}{1-\rho}$$

We can define our new statistic as:

$$\text{L1-ME}[X,Y] := \left\| \sqrt{t}\mathbf{\Sigma}_{N_1,N_2}^{-\frac{1}{2}}\mathbf{S}_{N_1,N_2} \right\|_1 \tag{11}$$

We assume that the number of samples of the distributions P and Q are of the same order, i.e: let $t = N_1 + N_2$, we have: $\frac{N_1}{t} \to \rho$ and therefore $\frac{N_2}{t} \to 1-\rho$ with $\rho \in]0,1[$. The computation of the statistic requires inverting a $J \times J$ matrix $\mathbf{\Sigma}_{N_1,N_2}$, but this is fast and numerically stable: J is typically be small, *eg* less than 10. The next proposition demonstrates the use of this statistic as a consistent two-sample test (see supp. mat. C.3 for the proof).

Proposition 3.2. *Let $\{T_j\}_{j=1}^J$ sampled i.i.d. from the distribution Γ and $X := \{x_i\}_{i=1}^{N_1}$ and $Y := \{y_i\}_{i=1}^{N_2}$ be i.i.d. samples from P and Q respectively. Under H_0, the statistic L1-ME$[X,Y]$ is almost surely asymptotically distributed as Naka$(\frac{1}{2},1,J)$, a sum of J random variables i.i.d which follow a Nakagami distribution of parameter $m = \frac{1}{2}$ and $\omega = 1$. Finally under H_1, almost surely the statistic can be arbitrarily large as $t \to \infty$, enabling the test to correctly reject H_0.*

Statistical test of level α: Compute $\|\sqrt{t}\mathbf{\Sigma}_{N_1,N_2}^{-\frac{1}{2}}\mathbf{S}_{N_1,N_2}\|_1$, choose the threshold δ corresponding to the $(1-\alpha)$-quantile of Naka$(\frac{1}{2},1,J)$, and reject the null hypothesis whenever $\|\sqrt{t}\mathbf{\Sigma}_{N_1,N_2}^{-\frac{1}{2}}\mathbf{S}_{N_1,N_2}\|_1$ is larger than δ.

[2] the pdf of the Nakagami distribution of parameters $m \geq \frac{1}{2}$ and $\omega > 0$ is $\forall x \geq 0$, $f(x,m,\omega) = \frac{2m^m}{\Gamma(m)\omega^m}x^{2m-1}\exp(\frac{-m}{\omega}x^2)$ where Γ is the Gamma function.

3.2 Optimizing test locations to improve power

As in [16], we can optimize the test locations $\mathcal{V}$ and kernel parameters (jointly referred to as θ) by maximizing a lower bound on the test power which offers a simple objective function for fast parameter tuning. We make the same regularization as in [16] of the test statistic for stability of the matrix inverse, by adding a regularization parameter $\gamma_{N_1,N_2} > 0$ which goes to 0 as t goes to infinity, giving L1-ME$[X, Y] := \|\sqrt{t}(\mathbf{\Sigma}_{N_1,N_2} + \gamma_{N_1,N_2}\mathbf{I})^{-\frac{1}{2}}\mathbf{S}_{N_1,N_2}\|_1$ (see proof in supp. mat. D.1).

Proposition 3.3. *Let $\mathcal{K}$ be a uniformly bounded family of $k : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$ measurable kernels (i.e., $\exists\, K < \infty$ such that $\sup_{k\in\mathcal{K}} \sup_{(x,y)\in(\mathbb{R}^d)^2} |k(x, y)| \leq K$). Let $\mathcal{V}$ be a collection in which each element is a set of J test locations. Assume that $c := \sup_{V\in\mathcal{V},k\in\mathcal{K}} \|\mathbf{\Sigma}^{-1/2}\| < \infty$. Then the test power $\mathbb{P}\left(\widehat{\lambda}_t \geq \delta\right)$ of the L1-ME test satisfies $\mathbb{P}\left(\widehat{\lambda}_t \geq \delta\right) \geq L(\lambda_t)$ where:*

$$L(\lambda_t) = 1 - 2\sum_{k=1}^{J} \exp\left(-\left(\frac{\lambda_t - \delta}{J^2 + J}\right)^2 \frac{\gamma_{N_1,N_2} N_1 N_2}{(N_1 + N_2)^2}\right) - 2\sum_{k,q=1}^{J} \exp\left(-2\frac{\left(\frac{\gamma_{N_1,N_2}}{K_3 J^2}\frac{\lambda_t - \delta}{(J^2+J)\sqrt{t}} - \frac{J^3 K_2}{\sqrt{\gamma_{N_1,N_2}}} - J^4 K_1\right)^2}{K_\lambda^2 (N_1 + N_2)\max\left(\frac{8}{\rho N_1}, \frac{8}{(1-\rho)N_2}\right)^2}\right)$$

and K_1, K_2, K_3 and K_λ, are positive constants depending on only K, J and c. The parameter $\lambda_t := \|\sqrt{t}\mathbf{\Sigma}^{-\frac{1}{2}}\mathbf{S}\|_1$ is the population counterpart of $\widehat{\lambda}_t := \|\sqrt{t}(\Sigma_{N_1,N_2} + \gamma_{N_1,N_2}\mathbf{I})^{-\frac{1}{2}}\mathbf{S}_{N_1,N_2}\|_1$ where $\mathbf{S} = \mathbb{E}_{x,y}(S_{N_1,N_2})$ and $\mathbf{\Sigma} = \mathbb{E}_{x,y}(\mathbf{\Sigma}_{N_1,N_2})$. Moreover for large t, $L(\lambda_t)$ is increasing in λ_t.

Proposition 3.3 suggests that it is sufficient to maximize λ_t to maximize a lower bound on the L1-ME test power. The statistic λ_t for this test depends on a set of test locations $\mathcal{V}$ and a kernel parameter σ. We set $\theta^* := \{\mathcal{V}, \sigma\} = \arg\max_\theta \lambda_t = \arg\max_\theta \|\sqrt{t}\,\mathbf{\Sigma}^{-\frac{1}{2}}\mathbf{S}\|_1$. As proposed in [14], we can maximize a proxy test power to optimize θ: it does not affect H_0 and H_1 as long as the data used for parameter tuning and for testing are disjoint.

3.3 Using smooth characteristic functions (SCF)

As the ME statistic, the SCF statistic estimates the L^2 distance between well chosen distribution representatives. Here, the representatives of the distributions are the convolution of their characteristic functions and the kernel k, assumed translation-invariant. [5] use them to detect differences between distributions in the frequency domain. We show that the L^1 version (denoted $d_{L^1,\Phi}$) is a metric on the space of Borel probability measures with integrable characteristic functions such that if α_n converge weakly towards α, then $d_{L^1,\Phi}(\alpha_n, \alpha) \to 0$ (see supp. mat. A.2). Let us introduce the test statistics in the frequency domain respectively based on the ℓ_2 norm and on the ℓ_1 norm which lead to consistent tests:

$$\widehat{d}^2_{\ell_2,\Phi,J}[X, Y] := \|\sqrt{n}\mathbf{S}_n\|_2^2 \quad \text{and} \quad \widehat{d}_{\ell_1,\Phi,J}[X, Y] := \|\sqrt{n}\mathbf{S}_n\|_1 \tag{12}$$

where $\mathbf{S}_n := \frac{1}{n}\sum_{i=1}^n \mathbf{z}_i$, $\mathbf{z}_i := [f(x_i)\sin(x_i^T T_j) - f(y_i)\sin(y_i^T T_j), f(x_i)\cos(x_i^T T_j) - f(y_i)\cos(y_i^T T_j)]_{j=1}^J \in \mathbb{R}^{2J}$ and f is the inverse Fourier transform of k. We show that, at the same level α, using the ℓ_1 norm in the frequency domain provides a better power with high probability (see supp. mat. E.1):

Proposition 3.4. *Let $\alpha \in]0, 1[$, $\gamma > 0$ and $J \geq 2$. Let $\{T_j\}_{j=1}^J$ sampled i.i.d. from the distribution Γ and let $X := \{x_i\}_{i=1}^n$ and $Y := \{y_i\}_{i=1}^n$ i.i.d. samples from P and Q respectively. Let us denote δ the $(1 - \alpha)$-quantile of the asymptotic null distribution of $\widehat{d}_{\ell_1,\Phi,J}[X, Y]$ and β the $(1 - \alpha)$-quantile of the asymptotic null distribution of $\widehat{d}^2_{\ell_2,\Phi,J}[X, Y]$. Under the alternative hypothesis, almost surely, there exists $N \geq 1$ such that for all $n \geq N$, with a probability of at least $1 - \gamma$ we have:*

$$\widehat{d}^2_{\ell_2,\Phi,J}[X, Y] > \beta \Rightarrow \widehat{d}_{\ell_1,\Phi,J}[X, Y] > \delta \tag{13}$$

We now adapt the construction of the L1-ME test to the frequency domain to avoid computational issues of the quantiles of the asymptotic null distribution:

$$\text{L1-SCF}[X, Y] := \|\sqrt{t}\, \mathbf{\Sigma}_{N_1,N_2}^{-\frac{1}{2}} \mathbf{S}_{N_1,N_2}\|_1 \tag{14}$$

with $\mathbf{\Sigma}_{N_1,N_2}$, and $\mathbf{S}_{N_1,N_2}$ defined as in the L1-ME statistic with new expression for $\mathbf{Z}_X^i$ (and $\mathbf{Z}_Y^j$):

$$\mathbf{Z}_X^i = \left(cos\left(T_1^T x_i\right) f(x_i), ..., sin\left(T_J^T x_i\right) f(x_i)\right) \in \mathbb{R}^{2J}$$

From this statistic, we build a consistent test. Indeed, an analogue proof of the Proposition 3.2 gives that under H_0, L1-SCF$[X, Y]$ is a.s. asymptotically distributed as Naka($\frac{1}{2}, 1, 2J$), and under H_1, the test statistic can be arbitrarily large as t goes to infinity. Finally an analogue proof of Proposition 3.3 shows that we can optimize the test locations and the kernel parameter to improve the power as well.

4 Experimental study

We now run empirical comparisons of our ℓ_1-based tests to their ℓ_2 counterparts, state-of-the-art Kernel-based two-sample tests. We study both toy and real problems. We use the isotropic Gaussian kernel class $\mathcal{K}_g$. We call **L1-opt-ME** and **L1-opt-SCF** the tests based respectively on mean embeddings and smooth characteristic functions proposed in this paper when optimizing test locations and the Gaussian width σ on a separate training set of the same size as the test set. We denote also **L1-grid-ME** and **L1-grid-SCF** where only the Gaussian width is optimized by a grid search, and locations are randomly drawn from a multivariate normal distribution. We write **ME-full** and **SCF-full** for the tests of [16], also fully optimized according to their criteria. **MMD-quad** (quadratic-time) and **MMD-lin** (linear-time) refer to the MMD-based tests of [11], where, to ensure a fair comparison, the kernel width is also set to maximize the test power following [14]. For **MMD-quad**, as its null distribution is an infinite sum of weighted chi-squared variables (no closed-form quantiles), we approximate the null distribution with 200 random permutations in each trial.

In all the following experiments, we repeat each problem 500 times. For synthetic problems, we generate new samples from the specified P, Q distributions in each trial. For the first real problem (Higgs dataset), as the dataset is big enough we use new samples from the two distributions for each trial. For the second and third real problem (Fast food and text datasets), samples are split randomly into train and test sets in each trial. In all the simulations we report an empirical estimate of the Type-I error when H_0 hold and of the Type-II error when H_1 hold. We set $\alpha = 0.01$. The code is available at https://github.com/meyerscetbon/l1_two_sample_test.

How to realize ℓ_1-based tests ? The asymptotic distributions of the statistics is a sum of i.i.d. Nakagami distribution. [8] give a closed form for the probability density function. As the formula is not simple, we can also derive an estimate of the CDF (see supp. mat. F.1).

Optimization For a fair comparison between our tests and those of [16], we use the same initialization of the test locations[3]. For the ME-based tests, we initialize the test locations with realizations from two multivariate normal distributions fitted to samples from P and Q and for the for initialization of the SCF-based tests, we use the standard normal distribution. The regularization parameter is set to $\gamma_{N_1,N_2} = 10^{-5}$. The computation costs for our proposed tests are the same as that of [16]: with t samples, optimization is $\mathcal{O}(J^3 + dJt)$ per gradient ascent iteration and testing $\mathcal{O}(J^3 + Jt + dJt)$ (see supp. mat. Table 3).

The experiments on synthetic problems mirror those of [16] to make a fair comparison between the prior art and the proposed methods.

Test power vs. sample size We consider four synthetic problems: Same Gaussian (SG, dim$= 50$), Gaussian mean difference (GMD, dim$= 100$), Gaussian variance difference (GVD, dim$= 30$), and Blobs. Table 1 summarizes the specifications of P and Q. In the Blobs problem, P and Q are a mixture of Gaussian distributions on a 4×4 grid in $\mathbb{R}^2$. This problem is challenging as the difference of P and Q is encoded at a much smaller length scale compared to the global structure as explained in [14]. We set $J = 5$ in this experiment.

Data	P	Q
SG	$\mathcal{N}(0, I_d)$	$\mathcal{N}(0, I_d)$
GMD	$\mathcal{N}(0, I_d)$	$\mathcal{N}((1, 0, .., 0)^T, I_d)$
GVD	$\mathcal{N}(0, I_d)$	$\mathcal{N}(0, diag(2, 1, .., 1))$
Blobs	Gaussian mixtures in $\mathbb{R}^2$ as [16]	

Table 1: Synthetic problems. H_0 holds only in SG.

[3] [16]: github.com/wittawatj/interpretable-test

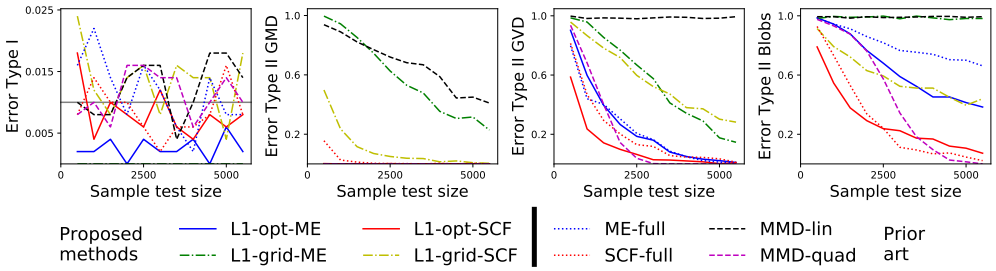


Figure 1: Plots of type-I/type-II errors against the test sample size n^{te} in the four synthetic problems.

Figure 2: Plots of type-I/type-II error against the dimension in three synthetic problems: SG (Same Gaussian), GMD (Gaussian Mean Difference), and GVD (Gaussian Variance Difference).

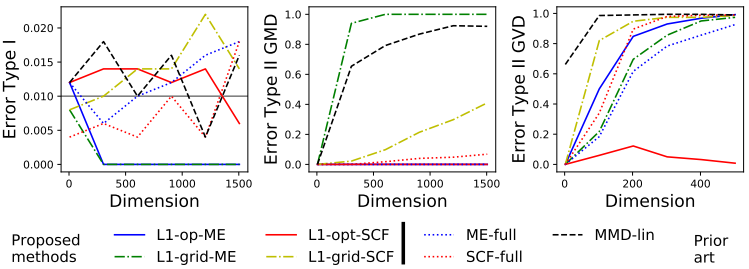


Figure 1 shows type-I error (for SG problem), and test power (for GMD, GVD and Blobs problem) as a function of test sample size. In the SG problem, the type-I error roughly stays at the specified level α for all tests except the L1-ME tests, which reject the null at a rate below the specified level α. Therefore, here these tests are more conservative.

GMD with 100 dimensions is an easy problem for **L1-opt-ME**, **L1-opt-SCF**, **ME-full MMD-quad**, while the **SCF-full** test requires many samples to achieve optimal test power. In the GMD, GVD and Blobs cases, **L1-opt-ME** and **L1-opt-SCF** achieve substantially higher test power than **L1-grid-ME** and **L1-grid-SCF**, respectively: optimizing the test locations brings a clear benefit. Remarkably **L1-opt-SCF** consistently outperforms the quadratic-time **MMD-quad** up to 2 500 samples in the GVD case. SCF variants perform significantly better than ME variants on the Blobs problem, as the difference in P and Q is localized in the frequency domain. For the same reason, **L1-opt-SCF** does much better than the quadratic-time MMD up to 3 000 samples, as the latter represents a weighted distance between characteristic functions integrated across the frequency domain as explained in [29].

We also perform a more difficult GMD problem to distinguish the power of the proposed tests with the **ME-full** as all reach maximal power. **L1-opt-ME** then performs better than **ME-full**, its ℓ_2 counterpart, as it needs less data to achieve good control (see mat. supp. F.2).

Test power vs. dimension d On fig 2, we study how the dimension of the problem affects type-I error and test power of our tests. We consider the same synthetic problems: SG, GMD and GVD, we fix the test sample size to 10000, set $J = 5$, and vary the dimension. Given that these experiments explore large dimensions and a large number of samples, computing the **MMD-quad** was too expensive.

In the SG problem, we observe the **L1-ME** tests are more conservative as dimension increases, and the others tests can maintain type-I error at roughly the specified significance level $\alpha = 0.01$. In the GMD problem, we note that the tests proposed achieve the maximum test power without making error of type-II whatever the dimension is, while the **SCF-full** loses power as dimension increases. However, this is true only with optimization of the test locations as it is shown by the test power of **L1-grid-ME** and **L1-grid-SCF** which drops as dimension increases. Moreover the performance of **MMD-lin** degrades quickly with increasing dimension, as expected from [24]. Finally in the GVD problem, all tests failed to keep a good test power as the dimension increases, except the **L1-opt-SCF**, which has a very low type-II for all dimensions. These results echo those obtained by [33]. Indeed [33] study a class of two sample test statistics based on inter-point distances and they show benefits of using the ℓ_1 norm over the Euclidean distance and the Maximum Mean Discrepancy (MMD) when the dimensionality goes to infinity. For this class of test statistics, they characterize asymptotic power

loss w.r.t the dimension and show that the ℓ_1 norm is beneficial compared to the ℓ_2 norm provided that the summation of discrepancies between marginal univariate distributions is large enough.

Informative features Figure 3 we replicate the experiment of [16], showing that the selected locations capture multiple modes in the ℓ_1 case, as in the ℓ_2 case. (details in supp. mat. F.3). The figure shows that the objective function $\widehat{\lambda}_t^{tr}(T_1, T_2)$ used to position the second test location T_2 has a maximum far from the chosen position for the first test location T_1.

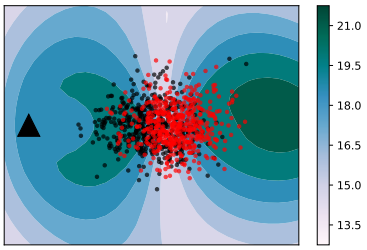


Figure 3: **Illustrating interpretable features**, replicating in the ℓ_1 case the figure of [16]. A contour plot of $\widehat{\lambda}_t^{tr}(T_1, T_2)$ as a function of T_2, when $J = 2$, and T_1 is fixed. The red and black dots represent the samples from the P and Q distributions, and the big black triangle the position of T_1 –complete figure in supp. mat. F.3.

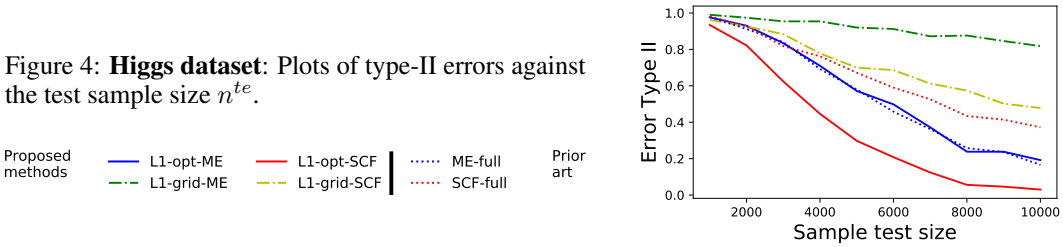
Real Data 1, Higgs: The first real problem is the Higgs dataset [20] described in [2]: distinguishing signatures of Higgs bosons from the background. We use a two-sample test on 4 derived features as in [5]. We compare for various sample sizes the performance of the proposed tests with those of [16]. We do not study the **MMD-quad** test as its computation is too expensive with 10 000 samples. To make the problem harder, we only consider $J = 3$ locations. Fig. 4 shows a clear benefit of the optimized ℓ_1-based tests, in particular for SCF (**L1-opt-SCF**) compared to its ℓ_2 counterpart (**SCF-full**). Optimizing the location is important, as **L1-opt-SCF** and **L1-opt-ME** perform much better than their grid versions (which are comparable to the tests of [5]).

Real Data 2, Fastfood: We use a Kaggle dataset listing locations of over 10,000 fast food restaurants across America[4]. We consider the 6 most frequent brands in mainland USA: Mc Donald's, Burger King, Taco Bell, Wendy's, Arby's and KFC. We benchmark the various two-sample tests to test whether the spatial distribution (in $\mathbb{R}^2$) of restaurants differs across brand. This is a non trivial question, as it depends on marketing strategy of the brand. We compare the distribution of Mc Donald's restaurants with others. We also compare the distribution of Mc Donald's restaurants with itself to evaluate the level of the tests (see supp. mat. Table 5). The number of samples differ across the distributions; hence to perform the tests from [16], we randomly subsample the largest distribution. We use $J = 3$ as the number of locations.

Table 2 summarizes type-II errors of the tests. Note that it is not clear that distributions must differ, as two brands sometimes compete directly, and target similar locations. We consider the **MMD-quad**

[4] www.kaggle.com/datafiniti/fast-food-restaurants

Figure 4: **Higgs dataset**: Plots of type-II errors against the test sample size n^{te}.



Problem	L1-opt-ME	L1-grid-ME	L1-opt-SCF	L1-grid-SCF	ME-full	SCF-full	MMD-quad
McDo vs Burger King (1141)	0.112	0.426	0.428	0.960	0.170	0.094	0.184
McDo vs Taco Bell (877)	0.554	0.624	0.710	0.834	0.684	0.638	0.666
McDo vs Wendy's (733)	0.156	0.246	0.752	0.942	0.416	0.624	0.208
McDo vs Arby's (517)	0.000	0.004	0.006	0.468	0.004	0.012	0.004
McDo vs KFC (429)	0.912	0.990	1.00	0.998	0.996	0.856	0.980

Table 2: **Fast food dataset:** Type-II errors for distinguishing the distribution of fast food restaurants. $\alpha = 0.01$. $J = 3$. The number in brackets denotes the sample size of the distribution on the right. We consider MMD-quad as the gold standard.

as the gold standard to decide whether distributions differ or not. The three cases for which there seems to be a difference are Mc Donald's vs Burger King, Mc Donald's vs Wendy's, and Mc Donalds vs Arby's. Overall, we find that the optimized **L1-opt-ME** agrees best with this gold standard. The Mc Donald's vs Arby's problem seems to be an easy problem, as all tests reach a maximal power, except for the **L1-grid-SCF** test which shows the gain of power brought by the optimization. In the Mc Donald's vs Wendy's problem the **L1-opt-ME** test outperforms the ℓ_2 tests and even the quadratic-time MMD. Finally, all the tests fail to discriminate Mc Donald's vs KFC. The data provide no evidence that these brands pursue different strategies to chose locations.

In the Mc Donald's vs Burger King and Mc Donald's vs Wendy's problems, the optimized version of the test proposed based on mean embedding outperform the grid version. This success implies that the locations learned are each informative, and we plot them (see supp. mat. Figure 7), to investigate the interpretability of the **L1-opt-ME** test. The figure shows that the procedure narrows on specific regions of the USA to find differences between distributions of restaurants.

Real Data 3, text: For a high-dimension problem, we consider the problem of distinguishing the newsgroups text dataset [17] (details in supp. Mat. F.4). Compared to their ℓ_2 counterpart, ℓ_1-optimized tests bring clear benefits and separate all topics of articles based on their word distribution.

Discussion: Our theoretical results suggest it is always beneficial for statistical power to build tests on ℓ_1 norms rather than ℓ_2 norm of differences between kernel distribution representatives (Propositions 3.1, 3.4). In practice, however, optimizing test locations with ℓ_1-norm tests leads to non-smooth objective functions that are harder to optimize. Our experiments confirm the theoretical benefit of the ℓ_1-based framework. The benefit is particularly pronounced for a large number J of test locations –as the difference between ℓ_1 and ℓ_2 norms increases with dimension (see in supp. mat. Lemmas 8, 12)– as well as for large dimension of the native space (Figure 2). The benefit of ℓ_1 distances for two-sample testing in high dimension has also been reported by [33], though their framework does not link to kernel embeddings or to the convergence of probability measures.

5 Conclusion

In this paper, we show that statistics derived from the L^p distances between well-chosen distribution representatives are well suited for the two-sample problem as these distances metrize the weak convergence (Theorem 2.1). We then compare the power of tests introduced in [5] and their ℓ_1 counterparts and we show that ℓ_1-based statistics have better power with high probability (Propositions 3.1, 3.4). As with state-of-the-art Euclidean approaches, the framework leads to tractable computations and learns interpretable locations of where the distributions differ. Empirically, on all 4 synthetic and 3 real problems investigated, the ℓ_1 geometry gives clear benefits compared to the Euclidean geometry. The L^1 distance is known to be well suited for densities, to control differences or estimation [7]. It is also beneficial for kernel embeddings of distributions.

Acknowledgments This work was funded by the DirtyDATA ANR grant (ANR-17-CE23-0018). We also would like to thank Zoltán Szabó from École Polytechnique for crucial suggestions.

References

[1] M. Arjovsky, S. Chintala, and L. Bottou. Wasserstein generative adversarial networks. In *International Conference on Machine Learning*, pages 214–223, 2017.

[2] P. Baldi, P. Sadowski, and D. Whiteson. Searching for exotic particles in high-energy physics with deep learning. *Nature communications*, 5:4308, 2014.

[3] L. Baringhaus and C. Franz. On a new multivariate two-sample test. *Journal of multivariate analysis*, 88(1):190–206, 2004.

[4] K. M. Borgwardt, A. Gretton, M. J. Rasch, H.-P. Kriegel, B. Schölkopf, and A. J. Smola. Integrating structured biological data by kernel maximum mean discrepancy. *Bioinformatics*, 22(14):e49–e57, 2006.

[5] K. P. Chwialkowski, A. Ramdas, D. Sejdinovic, and A. Gretton. Fast two-sample testing with analytic representations of probability measures. In *Advances in Neural Information Processing Systems*, pages 1981–1989, 2015.

[6] F. Cucker and S. Smale. On the mathematical foundations of learning. *Bulletin of the American mathematical society*, 39(1):1–49, 2002.

[7] L. Devroye and L. Györfi. Nonparametric density estimation: The l1 view, 1985.

[8] P. Dharmawansa, N. Rajatheva, and K. Ahmed. On the distribution of the sum of nakagami-m random variables. *IEEE transactions on communications*, 55(7):1407–1416, 2007.

[9] M. Fromont, M. Lerasle, P. Reynaud-Bouret, et al. Kernels based tests with non-asymptotic bootstrap approaches for two-sample problems. In *Conference on Learning Theory*, page 23, 2012.

[10] K. Fukumizu, A. Gretton, X. Sun, and B. Schölkopf. Kernel measures of conditional dependence. In *Advances in neural information processing systems*, pages 489–496, 2008.

[11] A. Gretton, K. M. Borgwardt, M. Rasch, B. Schölkopf, and A. J. Smola. A kernel method for the two-sample-problem. In *Advances in neural information processing systems*, pages 513–520, 2007.

[12] A. Gretton, K. M. Borgwardt, M. J. Rasch, B. Schölkopf, and A. Smola. A kernel two-sample test. *Journal of Machine Learning Research*, 13(Mar):723–773, 2012.

[13] A. Gretton, K. Fukumizu, Z. Harchaoui, and B. K. Sriperumbudur. A fast, consistent kernel two-sample test. In *Advances in neural information processing systems*, pages 673–681, 2009.

[14] 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*, page 1205, 2012.

[15] Z. Harchaoui, E. Moulines, and F. R. Bach. Testing for homogeneity with kernel fisher discriminant analysis. In *Advances in Neural Information Processing Systems*, page 609, 2008.

[16] W. Jitkrittum, Z. Szabó, K. P. Chwialkowski, and A. Gretton. Interpretable distribution features with maximum testing power. In *Advances in Neural Information Processing Systems*, pages 181–189, 2016.

[17] K. Lang. Newsweeder: Learning to filter netnews. In *Proceedings of the Twelfth International Conference on Machine Learning*, pages 331–339, 1995.

[18] Q. Le, T. Sarlós, and A. Smola. Fastfood-computing hilbert space expansions in loglinear time. In *International Conference on Machine Learning*, pages 244–252, 2013.

[19] C.-L. Li, W.-C. Chang, Y. Cheng, Y. Yang, and B. Póczos. Mmd gan: Towards deeper understanding of moment matching network. In *Advances in Neural Information Processing Systems*, pages 2203–2213, 2017.

[20] M. Lichman et al. UCI machine learning repository, 2013.

[21] J. R. Lloyd and Z. Ghahramani. Statistical model criticism using kernel two sample tests. In *Advances in Neural Information Processing Systems*, pages 829–837, 2015.

[22] A. Paszke, S. Gross, S. Chintala, G. Chanan, E. Yang, Z. DeVito, Z. Lin, A. Desmaison, L. Antiga, and A. Lerer. Automatic differentiation in pytorch. 2017.

[23] A. Rahimi and B. Recht. Random features for large-scale kernel machines. In *Advances in neural information processing systems*, pages 1177–1184, 2008.

[24] A. Ramdas, S. J. Reddi, B. Póczos, A. Singh, and L. A. Wasserman. On the decreasing power of kernel and distance based nonparametric hypothesis tests in high dimensions. In *AAAI*, pages 3571–3577, 2015.

[25] D. Sejdinovic, B. Sriperumbudur, A. Gretton, and K. Fukumizu. Equivalence of distance-based and rkhs-based statistics in hypothesis testing. *The Annals of Statistics*, pages 2263–2291, 2013.

[26] B. Simon. *Trace ideals and their applications*. Number 120. Am. Math. Soc., 2005.

[27] C.-J. Simon-Gabriel and B. Schölkopf. Kernel distribution embeddings: Universal kernels, characteristic kernels and kernel metrics on distributions. *arXiv preprint arXiv:1604.05251*, 2016.

[28] B. K. Sriperumbudur, K. Fukumizu, and G. R. Lanckriet. Universality, characteristic kernels and rkhs embedding of measures. *Journal of Machine Learning Research*, 12:2389, 2011.

[29] B. K. Sriperumbudur, A. Gretton, K. Fukumizu, B. Schölkopf, and G. R. Lanckriet. Hilbert space embeddings and metrics on probability measures. *Journal of Machine Learning Research*, 11:1517, 2010.

[30] I. Steinwart. On the influence of the kernel on the consistency of support vector machines. *Journal of machine learning research*, 2(Nov):67–93, 2001.

[31] G. J. Székely and M. L. Rizzo. Testing for equal distributions in high dimension. *InterStat*, 5(16.10):1249–1272, 2004.

[32] W. Zaremba, A. Gretton, and M. Blaschko. B-test: A non-parametric, low variance kernel two-sample test. In *Advances in neural information processing systems*, pages 755–763, 2013.

[33] C. Zhu and X. Shao. Interpoint distance based two sample tests in high dimension. *arXiv preprint arXiv:1902.07279*, 2019.

Supplementary materials

Table of Contents

A A family of metrics that metrize of the weak convergence

A.1 Distances between Mean Embeddings

Theorem 1. *Given $p \geq 1$, k a definite positive, characteristic, continuous, and bounded kernel on $\mathbb{R}^d$, μ_P and μ_Q the mean embeddings of the Borel probability measures P and Q respectively, the function defined on $\mathcal{M}_+^1(\mathbb{R}^d) \times \mathcal{M}_+^1(\mathbb{R}^d)$:*

$$d_{L^p,\mu}(P,Q) := \left(\int_{t\in\mathbb{R}^d} \left| \mu_P(t) - \mu_Q(t) \right|^p d\Gamma(t) \right)^{1/p} \tag{15}$$

is a metric on the space of Borel probability measures, for Γ a Borel probability measure absolutely continuous with respect to Lebesgue measure. Moreover a sequence $(\alpha_n)_{n\geq 0}$ of Borel probability measures converges weakly towards α if and only if $d_{L^p,\mu}(\alpha_n, \alpha) \to 0$.

Proof. *First, let us prove that for any $p \geq 1$ $d_{L^p,\mu}$ is metric of on the space of Borel probability measures. Let $p \geq 1$, we have:*

$$|\mu_P(t) - \mu_Q(t)|^p = |\langle \mu_P - \mu_Q, k_t \rangle|^p$$

Therefore:

$$|\mu_P(t) - \mu_Q(t)|^p \leq ||\mu_P - \mu_Q||_H^p (k(t,t))^{p/2}$$

But as k is bounded, and Γ is finite, $d_{L^p,\mu}$ is well defined on $\mathcal{M}_+^1(\mathcal{X})^2$. Let us prove now that if $P \neq Q$ then $d_{L^p,\mu}(P,Q) > 0$.

Definition 1. *[10] A kernel is characteristic if the mapping $P \in \mathcal{M}_+^1(\mathcal{X}) \to \mu_P \in H_k$ is injective, where H_k is the RKHS associated with k.*

Lemma 1. *[30] If k is a continuous kernel on a metric space then every feature maps associated with the kernel are continuous.*

Let P and Q two Borel distributions such that $P \neq Q$. Since the mapping $p \to \mu_P$ is injective, there must exists at least one point o where $\mu_P - \mu_Q$ is non-zero. By continuity of $\mu_P - \mu_Q$, there exists a ball around o in which $\mu_P - \mu_Q$ is non-zero. Then $d_{L^p,\mu}(P,Q) > 0$. Finally all the other proprieties of a metric are clearly verified by this function.

Let us now show that $d_{L^1,\mu}$ metrize the weak convergence. For that purpose, we first show that this metric has an IPM formulation:

Lemma 2. *l We denote by $\mathcal{T}_k$ the integral operator on $L_2^{d\Gamma}(\mathbb{R}^d)$ associated to the positive definite, characteristic, continuous, and bounded kernel k defined as:*

$$\begin{array}{rccc} \mathcal{T}_k & : & L_2^{d\Gamma}(\mathbb{R}^d) & \to & L_2^{d\Gamma}(\mathbb{R}^d) \\ & & f & \to & \int_{\mathbb{R}^d} k(x,.) f(x) d\Gamma(x) \end{array}$$

By denoting $B_\infty^{d\Gamma}$ the unit ball of $L_\infty^{d\Gamma}(\mathbb{R}^d)$, we have that:

$$d_{L^1,\mu}(P,Q) = \sup_{f \in \mathcal{T}_k(B_\infty^{d\Gamma})} \Big(\mathbb{E}_P[f(X)] - \mathbb{E}_Q[f(Y)] \Big)$$

Proof. *We have:*

$$\begin{aligned} d_{L^1,\mu}(P,Q) &= \int_{x \in \mathbb{R}^d} |\mu_P(x) - \mu_Q(x)| d\Gamma(x) \\ &= \int_{x \in \mathbb{R}^d} |\langle \mu_P - \mu_Q, k_x \rangle_H| d\Gamma(x) \\ &= \int_{x \in \{v : \mu_P(v) \geq \mu_Q(v)\}} \langle \mu_P - \mu_Q, k_x \rangle_H d\Gamma(x) - \int_{x \in \{v : \mu_P(v) < \mu_Q(v)\}} \langle \mu_P - \mu_Q, k_x \rangle_H d\Gamma(x) \\ &= \langle \mu_P - \mu_Q, \int_{x \in \{v : \mu_P(v) \geq \mu_Q(v)\}} k_x d\Gamma(x) - \int_{x \in \{v : \mu_P(v) < \mu_Q(v)\}} k_x d\Gamma(x) \rangle_H \end{aligned}$$

Then:

$$d_{L^1,\mu}(P,Q) = \langle \mu_P - \mu_Q, f \rangle_H$$

with

$$f = \int_{t \in \mathbb{R}^d} g_t d\Gamma(t) \quad \textit{where} \quad g_t = \begin{cases} k_t \textit{ if } t \in \{x : \mu_P(x) \geq \mu_Q(x)\} \\ -k_t \textit{ otherwise.} \end{cases} \tag{16}$$

Therefore, $f \in T_k(B_\infty^{d\Gamma}) \subset H_k$ and we have:

$$d_{L^1,\mu}(P,Q) = \mathbb{E}_P(f(X)) - \mathbb{E}_Q(f(Y))$$

Now, let f be an element of $\mathcal{T}_k(B_\infty^{d\Gamma}) \subset H_k$. Therefore there exists $g \in B_\infty^{d\Gamma}$ such that $f = \mathcal{T}_k(g)$ and we have then:

$$\begin{aligned} \mathbb{E}_P(f(X)) - \mathbb{E}_Q(f(Y)) &= \langle \mu_P - \mu_Q, f \rangle \\ &= \langle \mu_P - \mu_Q, \int_{t \in \mathbb{R}^d} g(t) k_t d\Gamma(t) \rangle \\ &= \int_{t \in \mathbb{R}^d} g(t) \langle \mu_P - \mu_Q, k_t \rangle d\Gamma(t) \\ &= \int_{t \in \mathbb{R}^d} g(t) (\mu_P(t) - \mu_Q(t)) d\Gamma(t) \\ &\leq \int_{t \in \mathbb{R}^d} |\mu_P(t) - \mu_Q(t)| d\Gamma(t) \end{aligned}$$

Therefore we have:

$$d_{L^1,\mu}(P,Q) = \sup_{f \in \mathcal{T}_k(B_\infty^{d\Gamma})} \Big(\mathbb{E}_P[f(X)] - \mathbb{E}_Q[f(Y)] \Big)$$

From this IPM formulation we now show that $d_{L^1,\mu}$ metrize the weak convergence. First, as the kernel k is assumed to be continuous, then $\mathcal{T}_k(B_\infty^{d\Gamma}) \subset H_k \subset C^0(\mathbb{R}^d)$, the set of continuous functions. Therefore, thanks to the IPM formulation of the metric, the weak convergence of a sequence of distributions $(\alpha_n)_{n\geq 0}$ towards a distribution α implies the convergence according to the $d_{L^1,\mu}$-distance. Conversely let $\alpha \in \mathcal{M}_+^1(\mathcal{X})$ and let us assume that $(\alpha_n)_{n\geq 0}$ is a sequence of Borel probability measures such that $d_{L^1,\mu}(\alpha_n, \alpha) \to 0$. Since $\int\limits_{x\in\mathbb{R}^d} k(x,x)d\Gamma(x)$ is finite, T_k is self-adjoint, positive semi-definite and trace-class [26]. It has at most countably many positive eigenvalues $(\lambda_m)_{m\geq 0}$ and corresponding orthonormal eigenfunctions $(e_m)_{m\geq 0}$. Then the Mercer theorem [6] gives that $(\lambda_m^{1/2} e_m)_{m\geq 0}$ is an orthonormal basis of H_k. Let us denote $C = \sup\limits_{x\in\mathbb{R}^d} \sqrt{k(x,x)}$ And $V_m = \frac{\lambda_m^{1/2} e_m}{C}$. Therefore we have:

$$\|V_m\|_{\infty,d\Gamma} \leq \frac{\|\lambda_m^{1/2} e_m\|_{H_k}}{C} C \leq 1$$

Therefore, thanks to Lemma 2, for all $m \geq 0$, we have:

$$\langle \mu_{\alpha_n} - \mu_\alpha, T_k(V_m)\rangle_{H_k} \to 0$$

Now, we want to show that for every $f \in H_k$, $\langle \mu_{\alpha_n} - \mu_\alpha, f\rangle_{H_k} \to 0$. Let us consider $f \in H_k$. As $(\lambda_m^{1/2} e_m)_{m\geq 0}$ is an orthonormal basis of H_k, we have:

$$f = \sum_{m\geq 0} \langle f, \lambda_m^{1/2} e_m\rangle_{H_k} \lambda_m^{1/2} e_m$$

Therefore if we define for every $m \geq 0$:

$$f_m := \sum_{i=0}^{m} \langle f, \lambda_i^{1/2} e_i\rangle_{H_k} \lambda_i^{1/2} e_i$$

We have that:

$$\|f_m - f\|_{H_k} \to 0$$

Therefore let $\epsilon > 0$, and K such that:

$$\|f_K - f\|_{H_k} \leq \epsilon$$

First we remarks that:

$$\begin{aligned}
\langle \mu_{\alpha_n} - \mu_\alpha, f_K\rangle &= \sum_{i=0}^{K} \langle f, \lambda_i^{1/2} e_i\rangle \langle \mu_{\alpha_n} - \mu_\alpha, \lambda_i^{1/2} e_i\rangle \\
&= \sum_{i=0}^{K} \langle f, \lambda_i^{1/2} e_i\rangle C \langle \mu_{\alpha_n} - \mu_\alpha, V_i\rangle \\
&= \sum_{i=0}^{K} \langle f, \lambda_i^{1/2} e_i\rangle \frac{C}{\lambda_i} \langle \mu_{\alpha_n} - \mu_\alpha, T_k(V_i)\rangle
\end{aligned}$$

Indeed the last equality hold as all the eigenvalues are positives. Finally we have that:

$$\langle \mu_{\alpha_n} - \mu_\alpha, f_K\rangle_{H_k} \to 0 \text{ as } n \text{ goes to infinity.}$$

Let N, such that for $n \geq N$:

$$\langle \mu_{\alpha_n} - \mu_\alpha, f_K\rangle_{H_k} \leq \epsilon$$

Therefore we have for all $n \geq N$:

$$\begin{aligned}
\langle \mu_{\alpha_n} - \mu_\alpha, f\rangle &= \langle \mu_{\alpha_n} - \mu_\alpha, f_K\rangle + \langle \mu_{\alpha_n} - \mu_\alpha, f - f_K\rangle \\
&\leq \epsilon + \|\mu_{\alpha_n} - \mu_\alpha\|_{H_k} \|f - f_K\|_{H_k} \\
&\leq \epsilon + \|\mu_{\alpha_n} - \mu_\alpha\|_{H_k} \epsilon
\end{aligned}$$

Finally as k is bounded, we have that:

$$\|\mu_{\alpha_n} - \mu_\alpha\|_{H_k} \leq 2 \sup_{x,t} \sqrt{k(x,t)}$$

Finally we have that for every $f \in H_k$:

$$\langle \mu_{\alpha_n} - \mu_\alpha, f \rangle \to 0$$

Therefore for any $f \in B_{H_k}$, the unit ball of the RKHS, we have:

$$\langle \mu_{\alpha_n} - \mu_\alpha, f \rangle \to 0$$

And then:

$$\textit{MMD}[\alpha_n, \alpha] \to 0$$

Moreover we have the following theorem:

Theorem 2. *([27]) A bounded kernel over a locally compact Hausdorff space $\mathcal{X}$ metrizes the weak convergence of probability measures iff it is continuous and characteristic.*

Therefore α_n converge weakly towards α and $d_{L^1,\mu}$ metrize the weak convergence. Moreover thanks to Hölder's inequality we have that for any $p \geq 1$:

$$d_{L^1,\mu}(P,Q) \leq d_{L^p,\mu}(P,Q) \tag{17}$$

Moreover as the kernel k is bounded we have also:

$$d_{L^p,\mu}(P,Q)^p \leq \|\mu_P - \mu_Q\|_\infty^{p-1} d_{L^1,\mu}(P,Q) \tag{18}$$

$$\leq (2C^2)^{p-1} d_{L^1,\mu}(P,Q) \tag{19}$$

Therefore for any $p \geq 1$ $d_{L^p,\mu}$ metrizes the weak convergence.

A.2 Distances between Smooth Characteristic Functions

Definition 2. *[5] Let $k : \mathbb{R}^d \to \mathbb{R}$ be a translation-invariant kernel i.e., $k(x-y)$ defines a positive definite kernel for x and y, P a Borel probability measure and $\psi_P(t) := \mathbb{E}_x\left(\exp(ix^T t)\right)$ be the characteristic function of P. A smooth characteristic function Φ_P is defined as:*

$$\Phi_P(v) := \int_{\mathbb{R}^d} \psi_P(t) k(v-t) dt \tag{20}$$

Lemma 3. *[5] If k is a continuous, integrable and translation invariant kernel with an inverse Fourier transform strictly greater then zero an and P has integrable characteristic function, then the mapping:*

$$\Gamma : P \to \Phi_P \tag{21}$$

is injective and Φ_P is an element of the RKHS H_k associated with k.

Theorem 3. *Given $p \geq 1$, k a translation invariant with an inverse Fourier transform strictly greater then zero, continuous, and integrable kernel on $\mathbb{R}^d$, Φ_P and Φ_Q the smooth characteristic functions of the Borel probability measures with integrable characteristic functions P and Q respectively, the following function:*

$$d_{L^p,\Phi}(P,Q) := \left(\int_{t \in \mathbb{R}^d} \left| \Phi_P(t) - \Phi_Q(t) \right|^p d\Gamma(t) \right)^{1/p} \tag{22}$$

where Γ a Borel probability measure absolutely continuous with respect to Lebesgue measure, is a metric on the space of Borel probability measures with integrable characteristic functions. Moreover if a sequence $(\alpha_n)_{n \geq 0}$ of Borel probability measures with integrable characteristic functions converges weakly towards α then $d_{L^1,\mu}(\alpha_n, \alpha) \to 0$.

Proof. *Let $p \geq 1$. First, as ψ_P and ψ_Q live in H_k, the RKHS associated with k, we have:*

$$|\Phi_P(t) - \Phi_Q(t)|^p \leq ||\Phi_P - \Phi_Q||_H^p \, k(0)^{p/2}$$

Let us prove now that if $P \neq Q$ then $d(P,Q) > 0$. Thanks to Lemma 1, Φ_P and Φ_Q are continuous. Since the mapping $P \to \Phi_P$ is injective, there must exists at least one point o where $\Phi_P - \Phi_Q$ is non-zero. By continuity of $\Phi_P - \Phi_Q$, there exists a ball around o in which $\Phi_P - \Phi_Q$ is non-zero. Then $d_{L^p,\Phi}(P,Q) > 0$. Moreover, all the other proprieties of a metric are clearly verified by this function. Let us now show that $d_{L^1,\Phi}$ admits a IPM formulation:

Lemma 4. *Let $\mathcal{T}_k$ be the integral operator on $L_2^{d\Gamma}(\mathbb{R}^d)$ associated with the kernel k. By denoting $B_\infty^{d\Gamma}$ the unit ball of $L_\infty^{d\Gamma}(\mathbb{R}^d)$, we have that:*

$$d_{L^1,\Phi}(P,Q) = \sup_{f\in\mathcal{L}(\mathcal{T}_k(B_\infty^{d\Gamma}))} \Big(\mathbb{E}_P[f(X)] - \mathbb{E}_Q[f(Y)]\Big)$$

where:

$$\mathcal{L}(f)(x) := \int_{t\in\mathbb{R}^d} \exp(it^T x) f(t) dt \tag{23}$$

Proof. *Let P and Q be Borel probability measures with integrable characteristic functions. As Φ_P and Φ_Q live in the RKHS associated with k, we obtain, as in the proof of Theorem 2.1, that:*

$$d_{L^1,\Phi}(P,Q) = \langle \Phi_P - \Phi_Q, f\rangle$$

with

$$f = \int_{t\in\mathbb{R}^d} g_t d\Gamma(t) \quad \textit{where} \quad g_t = \begin{cases} k_t \textit{ if } t\in\{x:\Phi_P(x)\geq\Phi_Q(x)\} \\ -k_t \textit{ otherwise.}\end{cases}$$

Therefore $f\in\mathcal{T}_k(B_\infty^{d\Gamma})$ and we have:

$$\begin{aligned} d_{L^1,\Phi}(P,Q) &= \int_{\mathbb{R}^d} \psi_P(t) f(t) dt - \int_{\mathbb{R}^d} \psi_Q(t) f(t) dt \\ &= \int_{t\in\mathbb{R}^d}\int_{\epsilon\in\mathbb{R}^d} \exp(i\epsilon^T t) f(t) dP(\epsilon) dt - \int_{t\in\mathbb{R}^d}\int_{\epsilon\in\mathbb{R}^d} \exp(i\epsilon^T t) f(t) dQ(\epsilon) dt \end{aligned}$$

Let us now show that for any $g\in B_\infty^{d\Gamma}$, $\mathcal{T}_k(g)$ is integrable (w.r.t the Lebesgue measure):

$$\int_{x\in\mathbb{R}^d} |\mathcal{T}_k(g)(x)| dx \leq \int_{x\in\mathbb{R}^d}\int_{t\in\mathbb{R}^d} |k(x,t)g(t)| d\Gamma(t) dx$$

But as k is translation-invariant we have:

$$\begin{aligned} \int_{t\in\mathbb{R}^d}\int_{x\in\mathbb{R}^d} |k(x,t)g(t)| d\Gamma(t) dx &= \int_{x\in\mathbb{R}^d} \left(\int_{u\in\mathbb{R}^d} |k(u)| du\right) |g(t)| d\Gamma(t) \\ &= \int_{u\in\mathbb{R}^d} |k(u)| du \int_{x\in\mathbb{R}^d} |g(t)| d\Gamma(t) \end{aligned}$$

And as k is integrable, and $g\in B_\infty^{d\Gamma}$, we can apply the Fubini–Tonelli theorem, and $\mathcal{T}_k(g)$ is integrable.

Therefore for any Borel probability measure P with integrable characteristic function, $\int_{x\in\mathbb{R}^d}\int_{\epsilon\in\mathbb{R}^d} |f(t)| dP(\epsilon) dt < \infty$ and by Fubini–Tonelli theorem, we can rewrite $d_{L^1,\Phi}(P,Q)$ as:

$$d_{L^1,\Phi}(P,Q) = \int_{\epsilon\in\mathbb{R}^d} \left(\int_{t\in\mathbb{R}^d} \exp(i\epsilon^T t) f(t) dt\right) dP(\epsilon) - \int_{\epsilon\in\mathbb{R}^d} \left(\int_{t\in\mathbb{R}^d} \exp(i\epsilon^T t) f(t) dt\right) dQ(\epsilon)$$

Therefore we have:

$$\begin{aligned} d_{L^1,\Phi}(P,Q) &= \int_{\epsilon\in\mathbb{R}^d} \mathcal{L}(f)(\epsilon) dP(\epsilon) - \int_{\epsilon\in\mathbb{R}^d} \mathcal{L}(f)(\epsilon) dQ(\epsilon) \\ &= \mathbb{E}_P(\mathcal{L}(f)(X)) - \mathbb{E}_Q(\mathcal{L}(f)(Y)) \end{aligned}$$

Let now g be an abritary function in $B_\infty^{d\Gamma}$. Then we have:

$$\mathbb{E}_P(L(\mathcal{T}_k(g))(X)) - \mathbb{E}_Q(L(\mathcal{T}_k(g))(Y)) = \int_{\epsilon\in\mathbb{R}^d} \mathcal{L}(\mathcal{T}_k(g))(\epsilon) dP(\epsilon) - \int_{\epsilon\in\mathbb{R}^d} \mathcal{L}(\mathcal{T}_k(g))(\epsilon) dQ(\epsilon)$$

But we have that:

$$\begin{aligned} \int_{\epsilon\in\mathbb{R}^d} \mathcal{L}(\mathcal{T}_k(g))(\epsilon) dP(\epsilon) &= \int_{\epsilon\in\mathbb{R}^d} \left(\int_{t\in\mathbb{R}^d} \exp(i\epsilon^T t) \mathcal{T}_k(g)(t) dt\right) dP(\epsilon) \\ &= \int_{t\in\mathbb{R}^d} \left(\int_{\epsilon\in\mathbb{R}^d} \exp(i\epsilon^T t) dP(\epsilon)\right) \mathcal{T}_k(g)(t) dt \\ &= \int_{\mathbb{R}^d} \psi_P(t) \mathcal{T}_k(g)(t) dt \\ &= \langle \Phi_P, \mathcal{T}_k(g)\rangle \end{aligned}$$

Finally we have:

$$\begin{aligned}\mathbb{E}_P(\mathcal{L}(\mathcal{T}_k(g))(X)) - \mathbb{E}_Q(\mathcal{L}(\mathcal{T}_k(g))(Y)) &= \langle \Phi_P - \Phi_Q, \mathcal{T}_k(g)\rangle \\ &= \int_{\mathbb{R}^d} g(t)(\Phi_P(t) - \Phi_Q(t))d\Gamma(t) \\ &\leq \int_{\mathbb{R}^d} |\Phi_P(t) - \Phi_Q(t)|d\Gamma(t)\end{aligned}$$

The results follows.

Therefore thanks to the IPM formulation of the $d_{L^1,\Phi}$-distance, we deduce that for all $p \geq 1$, if α_n converge weakly towards α, then $d_{L^1,\Phi}(\alpha_n, \alpha) \to 0$. Indeed, we have shown that $\mathcal{T}_k(B_\infty^{d\Gamma}) \subset L^1(\mathbb{R}^d)$, therefore $\mathcal{L}(\mathcal{T}_k(B_\infty^{d\Gamma})) \subset C^0(\mathbb{R}^d)$, and the result follows.

B Two-sample testing using the ℓ_1 norm

B.1 ℓ_1-based random metric with mean embeddings

Definition 3. *Let k be a kernel. For any $J > 0$, we define:*

$$d_{\ell_1,\mu,J} := \left\{ d_{\ell_1,\mu,J}\left[P, Q\right] = \frac{1}{J}\sum_{j=1}^{J} |\mu_P\left(T_j\right) - \mu_Q\left(T_j\right)| : \quad P, Q \in \mathcal{M}_+^1\left(\mathbb{R}^d\right) \right\}$$

with $\{T_j\}_{j=1}^J$ sampled independently from the distribution Γ.

Theorem 4. *Let k be a bounded, analytic, and characteristic kernel. Then for any $J > 0$, $d_{\ell_1,\mu,J}$ is a random metric on the space of Borel probability measures.*

Proof. *To prove this theorem we have first to introduce the fact that analytic functions are 'well behaved'.*

Lemma 5. *Let μ be absolutely continuous measure on $\mathbb{R}^d$ (wrt. the Lebesgue measure). Non-zero, analytic function f can be zero at most at the set of measure 0, with respect to the measure μ.*

Proof. *If f is zero at the set with a limit point then it is zero everywhere. Therefore f can be zero at most at a set A without a limit point, which by definition is a discrete set (distance between any two points in A is greater then some $\epsilon > 0$). Discrete sets have zero Lebesgue measure (as a countable union of points with zero measure). Since μ is absolutely continuous then $\mu(A)$ is zero as well.*

Let us now show how to build a random metric based on the ℓ_1 norm.

Lemma 6. *Let Λ be an injective mapping from the space of the Borel probability measures into a space of analytic functions on $\mathbb{R}^d$. Define*

$$d_{\Lambda,J}\left[P, Q\right] := \frac{1}{J}\sum_{j=1}^{J} |\Lambda P\left(T_j\right) - \Lambda Q\left(T_j\right)|$$

with $\{T_j\}_{j=1}^J$ sampled independently from the distribution Γ.

Then $d_{\Lambda,J}$ is a random metric.

Proof. *Let ΛP and ΛQ be images of measures P and Q respectively. We want to apply Lemma 5 to the analytic function $f = \Lambda P - \Lambda Q$, with the measure Γ, to see that if $P \neq Q$ then $f \neq 0$ a.s. To do so, we need to show that $P \neq Q$ implies that f is non-zero. Since mapping to Γ is injective, there must exists at least one point o where f is non-zero. By continuity of f, there exists a ball around o in which f is non-zero.*

We have shown that $P \neq Q$ implies f is almost everywhere non zero which in turn implies that $d_{\Lambda,J}\left(P, Q\right) > 0$ a.s. If $P = Q$ then $f = 0$ and $d_{\Lambda,J}\left(P, Q\right) = 0$.

By the construction $d_{\Lambda,J}$ is clearly symmetric and satisfies the triangle inequality.

Before proving the theorem we need to introduce a Lemma:

Lemma 7. *[5] If k is a bounded, analytic kernel on $\mathbb{R}^d \times \mathbb{R}^d$, then all functions in the RKHS H associated with this kernel are analytic.*

Since k is characteristic the mapping $\Lambda : P \to \mu_P$ is injective. Since k is a bounded, analytic kernel on $\mathbb{R}^d \times \mathbb{R}^d$,the Lemma 7 guarantees that μ_P is analytic, hence the image of Λ is a subset of analytic functions. Therefore, we can use Lemma 6 to see that $d_{\Lambda,J}[P,Q] = d_{\ell_1,\mu,J}[P,Q]$ is a random metric and this concludes the proof.

B.2 A first test with finite-sample control

Let us now build a statistic based on an estimation of the random metric introduced in eq.7. Let $X = \{x_1, ..., x_{N_1}\}$ and $Y = \{y_1, ..., y_{N_2}\} \subset \mathbb{R}^d$ i.i.d. two samples drawn respectively from the Borel probability measures P and Q. From these samples we define their empirical mean embeddings μ_X and μ_Y:

$$\mu_X(T) := \frac{1}{N_1}\sum_{i=1}^{N_1} k(x_i, T), \qquad \mu_Y(T) := \frac{1}{N_2}\sum_{i=1}^{N_2} k(y_i, T)$$

And we define:

$$\mathbf{S}_{N_1,N_2} := \Big(\mu_X(T_1) - \mu_Y(T_1), ..., \mu_X(T_J) - \mu_Y(T_J)\Big) \tag{24}$$

with $\{T_j\}_{j=1}^J$ sampled independently from the distribution Γ. Finally we define a first statistic:

$$d_{\ell_1,\mu,J}[X,Y] := \frac{1}{J}\|S_{N_1,N_2}\|_1 \tag{25}$$

We now derive a control of the statistic:

Proposition 1. *With K such that* $\sup_{x,y\in\mathbb{R}^d} |k(x,y)| \leq \frac{K}{2}$,

$$\mathbb{P}_{X,Y}\left(\Big|d_{\ell_1,\mu,J}[X,Y] - d_{\ell_1,\mu,J}[P,Q]\Big| > t\right) \leq 2J\exp\left(\frac{-t^2 N_1 N_2}{2K^2(N_1+N_2)}\right)$$

Proof. *We have:*

$$|d_{\ell_1,\mu,J}[X,Y] - d_{\ell_1,\mu,J}[P,Q]| \leq \frac{1}{J}\sum_{j=1}^{J}\Big||\mu_X(T_j) - \mu_Y(T_j)| - |\mu_p(T_j) - \mu_p(T_j)|\Big|$$

Then:

$$|d_{\ell_1,\mu,J}[X,Y] - d_{\ell_1,\mu,J}[P,Q]| \leq \frac{1}{J}\sum_{j=1}^{J}\Big|\left(\mu_X(T_j) - \mu_Y(T_j)\right) - \mathbb{E}_{X,Y\sim p,q}\left(\mu_X(T_j) - \mu_Y(T_j)\right)\Big|$$

Let us now consider the upper bound of the difference. By applying a union bound we have:

$$\mathbb{P}\Big(\frac{1}{J}\sum_{j=1}^{J}\Big|\left(\mu_X(T_j) - \mu_Y(T_j)\right) - \mathbb{E}_{X,Y}\left(\mu_X(T_j) - \mu_Y(T_j)\right)\Big| \geq t\Big)$$
$$\leq \sum_{j=1}^{J}\mathbb{P}_{X,Y}\left(\frac{1}{J}\Big|\left(\mu_X(T_j) - \mu_Y(T_j)\right) - \mathbb{E}_{X,Y}(..)\Big| \geq \frac{t}{J}\right)$$

Then by applying Hoeffding's inequality on each term of the sum of the right term of the inequality, we have:

$$\mathbb{P}_{X,Y}\left(\frac{1}{J}\Big|\left(\mu_X(T_j) - \mu_Y(T_j)\right) - \mathbb{E}_{X,Y}(..)\Big| \geq \frac{t}{J}\right) \leq 2\exp\left(-\frac{t^2 N_1 N_2}{2K^2(N_1+N_2)}\right)$$

Finally we have:

$$\mathbb{P}_{X,Y}\left(\Big|d_{\ell_1,\mu,J}[X,Y] - d_{\ell_1,\mu,J}[P,Q]\Big| \geq t\right) \leq 2J\exp\left(-\frac{t^2 N_1 N_2}{2K^2(N_1+N_2)}\right)$$

Corollary 1. *The hypothesis test associated with the statistic $d_{\ell_1,\mu,J}[X,Y]$ of level α for the null hypothesis $P = Q$, that is for $d_{\ell_1,\mu,J}[P,Q] = 0$ almost surely, has almost surely the acceptance region:*

$$d_{\ell_1,\mu,J}[X,Y] < K\sqrt{\frac{N_1+N_2}{N_1N_2}}\sqrt{2\log\left(\frac{J}{\alpha}\right)}$$

Moreover, the test is consistent almost surely.

Proof. Let us note the probability space of random variables $\{T_j\}_{j=1}^J$ as $(\Omega, \mathcal{F}, P)$.

Let $\omega \in \Omega$ such that $d^{\omega}_{\ell_1,\mu,J}[P,Q] = 0$. Then we have thanks to Proposition **??** that:

$$d^{\omega}_{\ell_1,\mu,J}[X,Y] < K\sqrt{\frac{N_1+N_2}{N_1N_2}}\sqrt{2\log\left(\frac{J}{\alpha}\right)}$$

with a probability at last of $1-\alpha$.

By assuming the null hypothesis $P = Q$, we have thanks to Theorem 4 that $d_{\ell_1,\mu,J}[P,Q] = 0$ a.s., then the result above hold a.s.

Moreover the statistic converges in probability to its population value a.s which give us the consistency of the test a.s. □

We now show that, under the alternative hypothesis, the statistic captures dense differences between distributions with high probability:

Corollary 2. *Let $\gamma > 0$, then under the alternative hypothesis, almost surely there exist $\Delta > 0$ such that for all $N_1, N_2 \geq 1$:*

$$\mathbb{P}_{X,Y}\left(\forall j \in [\![1,J]\!], \frac{|\mu_X(T_j) - \mu_Y(T_j)|}{J} \geq \frac{\Delta}{J} - \omega_{N_1,N_2}\right) \geq 1-\gamma$$

$$\textit{where } \omega_{N_1,N_2} = \frac{1}{J}\sqrt{\log(\frac{J^2}{\gamma})\frac{2K^2(N_1+N_2)}{N_1N_2}}$$

Proof. Let Δ be the minimum of $\mu_p - \mu_q$ over the set of locations $\{T_j\}_{j=1}^J$. Thanks to the analycity of the kernel we have that under the alternative hypothesis, $\mu_p - \mu_q$ is non zero everywhere almost surely. Therefore $\Delta > 0$ almost surely. Moreover by applying Proposition 1 for each T_j we obtain that for all $N_1, N_2 \geq 0$:

$$\mathbb{P}_{X,Y}\left(\frac{|\mu_X(T_j) - \mu_Y(T_j)|}{J} \geq \frac{\Delta}{J} - \omega_{N_1,N_2}\right) \geq 1 - \frac{\gamma}{J}$$

$$\text{where } \omega_{N_1,N_2} = \frac{1}{J}\sqrt{\log(\frac{J^2}{\gamma})\frac{2K^2(N_1+N_2)}{N_1N_2}}$$

Finally by applying an union bound, the result follows. □

C A test statistic with simple asymptotic distribution

C.1 Asymptotic distribution of $\widehat{d}_{\ell_1,\mu,J}[X,Y]$

Proposition 2. *Let $\{T_j\}_{j=1}^J$ sampled independently from the distribution Γ and $X := \{x_i\}_{i=1}^n$ and $Y := \{y_i\}_{i=1}^n$ be i.i.d. samples from P and Q respectively. Under H_0, the statistic $\widehat{d}_{\ell_1,\mu,J}[X,Y]$ is almost surely asymptotically distributed as a sum of J correlated Nakagami variables. Finally under H_1, almost surely the statistic can be arbitrarily large as $n \to \infty$, allowing the test to correctly reject H_0.*

Proof. *Let us note the probability space of random variables $\{T_j\}_{j=1}^J$ as $(\Omega, \mathcal{F}, P)$. Let $\omega \in \Omega$ such that $d^{\omega}_{\ell_1,\mu,J}[P,Q] = 0$ (see Definition 3) and let us define:*

$$\mathbf{z}_i^{\omega} := \left(k\left(x_i, T_1\left(\omega\right)\right) - k\left(y_j, T_1\left(\omega\right)\right), ..., k\left(x_i, T_J\left(\omega\right)\right) - k\left(y_j, T_J\left(\omega\right)\right)\right) \in \mathbb{R}^J$$

Therefore we can define:

$$\mathbf{S}_n := \frac{1}{n}\sum_{i=1}^{n} \mathbf{z}_i^{\omega}$$

By applying the Central-Limit Thoerem, we have:

$$\sqrt{n}\mathbf{S}_n \to \mathcal{N}\left(0, \mathbf{\Sigma}^{\omega}\right) \qquad \textit{with} \qquad \mathbf{\Sigma}^{\omega} := \textit{Cov}(\mathbf{z}^{\omega})$$

Therefore $\widehat{d}^{\omega}_{\ell_1,\mu,J}[X,Y] = \|\sqrt{n}\,\mathbf{S}_n^{\omega}\|_1$ converges to a sum of correlated Nakagami variables. But under, the null hypothesis $P = Q$, we have thanks to Theorem 4 that $d_{\ell_1,\mu,J}[P,Q] = 0$ a.s., then a.s. $\widehat{d}_{\ell_1,\mu,J}$ converges to a sum of correlated Nakagami variables. Let's now consider an ω such that $d^{\omega}_{\ell_1,\mu,J}[P,Q] > 0$. Since $\mathbf{S}_n^{\omega}$ converges in probability to the vector $\mathbf{S}^{\omega} = \mathbb{E}\left(\mathbf{z}^{\omega}\right) \neq 0$, then we have:

$$\mathbb{P}\left(\left\|\sqrt{n}\mathbf{S}_n^{\omega}\right\|_1 > r\right) = \mathbb{P}\left(\|\mathbf{S}_n^{\omega}\|_1 - \frac{r}{\sqrt{n}} > 0\right)$$

And as $\frac{r}{\sqrt{t}} \to 0$ as $t \to \infty$, we have finally:

$$\mathbb{P}\left(\left\|\sqrt{n}\mathbf{S}_n^{\omega}\right\|_1 > r\right) \to 1 \quad \textit{as} \quad t \to \infty.$$

Finally, under H_1, $d_{\ell_1,\mu,J}[P,Q] > 0$ almost surely and the statistic can be arbitrarily large as $n \to \infty$ almost surely.

C.2 Proof of Proposition 3.1

Proposition 3. *Let $\alpha \in]0,1[$, $\gamma > 0$ and $J \geq 2$. Let $\{T_j\}_{j=1}^J$ sampled i.i.d. from the distribution Γ and let $X := \{x_i\}_{i=1}^n$ and $Y := \{y_i\}_{i=1}^n$ i.i.d. samples from P and Q respectively. Let us denote δ the $(1-\alpha)$-quantile of the asymptotic null distribution of $\widehat{d}_{\ell_1,\mu,J}[X,Y]$ and β the $(1-\alpha)$-quantile of the asymptotic null distribution of $\widehat{d}^2_{\ell_2,\mu,J}[X,Y]$. Under the alternative hypothesis, almost surely, there exists $N \geq 1$ such that for all $n \geq N$, with a probability of at least $1-\gamma$ we have:*

$$\widehat{d}^2_{\ell_2,\mu,J}[X,Y] > \beta \Rightarrow \widehat{d}_{\ell_1,\mu,J}[X,Y] > \delta \tag{26}$$

Proof. *First we remarks that:*

$$\widehat{d}^2_{\ell_2,\mu,J}[X,Y] = \|\sqrt{n}\mathbf{S}_n\|_2^2$$

and

$$\widehat{d}_{\ell_1,\mu,J}[X,Y] = \|\sqrt{n}\mathbf{S}_n\|_1$$

where $\mathbf{S}_n := \frac{1}{n}\sum_{i=1}^n \mathbf{z}_i^{\omega}$ and $\mathbf{z}_i := \left(k\left(x_i, T_1\left(\omega\right)\right) - k\left(y_j, T_1\left(\omega\right)\right), ..., k\left(x_i, T_J\left(\omega\right)\right) - k\left(y_j, T_J\left(\omega\right)\right)\right)$. Let us now introduce the following Lemma:

Lemma 8. *Let $\mathbf{x}$ a random vector $\in \mathbb{R}^J$ with $J \geq 2$, $\mathbf{z} := \min_{j\in[|1,J|]} |x_j|$, $\epsilon > 0$ and $\gamma > 0$. If*

$$\mathbb{P}(\mathbf{z} \geq \epsilon) \geq 1 - \gamma$$

we have with a probability of at least $1-\gamma$ that, $\forall t_1 \geq t_2 \geq 0$, if $\epsilon \geq \sqrt{\frac{t_1^2 - t_2^2}{J(J-1)}}$, then

$$\|\mathbf{x}\|_2 > t_2 \Rightarrow \|\mathbf{x}\|_1 > t_1.$$

Proof. *First we remarks that:*

$$\begin{aligned}\epsilon > \sqrt{\frac{t_1^2 - t_2^2}{J\left(J-1\right)}} \Rightarrow & J\left(J-1\right)\epsilon > t_1^2 - t_2^2 \\ \Rightarrow & t_2^2 > t_1^2 - J\left(J-1\right)\epsilon^2\end{aligned}$$

Therefore, we have:

$$\begin{aligned}\|\mathbf{x}\|_2 \geq t_2 \Rightarrow & \|\mathbf{x}\|_2^2 + J(J-1)\epsilon^2 \geq t_1^2 \\ \Rightarrow & \sqrt{\|\mathbf{x}\|_2^2 + J(J-1)\epsilon^2} \geq t_1\end{aligned}$$

But we have that:

$$\|\mathbf{x}\|_1^2 = \sum_{i=1}^{J} |\mathbf{x}_i|^2 + \sum_{i \neq j} |\mathbf{x}_i||\mathbf{x}_j|$$

Therefore we have with a probability of 1-γ that:

$$\|\mathbf{x}\|_1^2 \geq \|\mathbf{x}\|_2^2 + J(J-1)\epsilon^2$$

And:

$$\|\mathbf{x}\|_2 \geq t_2 \Rightarrow \|\mathbf{x}\|_1 \geq t_1$$

Moreover by denoting δ the $(1-\alpha)$-quantile of the asymptotic null distribution of $\widehat{d}_{\ell_1,\mu,J}[X,Y]$ and β the $(1-\alpha)$-quantile of the asymptotic null distribution of $\widehat{d}^2_{\ell_2,\mu,J}[X,Y]$ we have that $\delta \geq \sqrt{\beta}$:

Lemma 9. *Let $\mathbf{x}$ be a random vector in $\mathbb{R}^J$, δ the $(1-\alpha)$-quantile of $\|\mathbf{x}\|_1$ and β the $(1-\alpha)$-quantile of $\|\mathbf{x}\|_2$. We have then:*

$$\delta \geq \beta \geq 0. \tag{27}$$

Proof. *The results is a direct consequence of the domination of the ℓ_1 norm:*

$$\|\mathbf{x}\|_1 \geq \|\mathbf{x}\|_2$$

Indeed, under H_0, we have shown that (see proof Proposition 2):

$$\sqrt{n}\mathbf{S}_n \to \mathcal{N}(0, \mathbf{\Sigma}^\omega) \qquad \text{with} \qquad \mathbf{\Sigma} := Cov(\mathbf{z})$$

Therefore by applying the Lemma 9 to $\mathbf{x}$ which follows $\mathcal{N}(0, \mathbf{\Sigma}^\omega)$, we obtain that $\delta \geq \sqrt{\beta}$. Now, To show the result we only need to show that the assumption of the Lemma 8 is sastified for the random vector $\mathbf{x} := \sqrt{n}\mathbf{S}_n$, $t_1 = \delta$ and $t_2 = \sqrt{\beta}$, i.e. for $\epsilon = \sqrt{\frac{\delta^2 - \beta}{J(J-1)}}$ under the alternative hypothesis. Under $H_1 : P \neq Q$, we have that $\mathbf{S}_n$ converge in probability to $\mathbf{S} := \mathbb{E}_{(x,y)\sim(P,Q)}(\mathbf{S}_n)$. Then by continuity of the application:

$$\phi_j : x := (x_j)_{j=1}^J \mathbb{R}^J \to |x_j|$$

, we have that for all $j \in [|1, J|]$, $|(\mathbf{S}_n)_j|$ converges in probability towards $\mathbf{S}_j$, the j-th coordinate of $\mathbf{S}$. Since $\mathbf{S} = (\mu_P(T_j) - \mu_Q(T_j))_{j=1}^J$, thanks to the analycity of the kernel k, the Lemma 7 guarantees the analycity of $\mu_P - \mu_Q$. And thanks to the injectivity of the mean embedding function, $\mu_P - \mu_Q$ is a non-zero function, therefore thanks to Lemma 5 $\mu_P - \mu_Q$ is non zero almost everywhere. Moreover the $(T_j)_{j=1}^J$ are independent, therefore the coordinates of $\mathbf{S}$ are almost surely all nonzero. Then we have then for all $j \in [\![1, J]\!]$:

$$\mathbb{P}\left(\left|(\sqrt{n}\mathbf{S}_n)_j\right| > \epsilon\right) = \mathbb{P}\left(\left|(\mathbf{S}_n)_j\right| - \frac{\epsilon}{\sqrt{n}} > 0\right)$$

And as $\frac{\epsilon}{\sqrt{n}} \to 0$ as $n \to \infty$, we have finally almost surely for all $j \in [\![1, J]\!]$:

$$\mathbb{P}_{X,Y}\left(\left|(\sqrt{n}\mathbf{S}_n)_j\right| \geq \epsilon\right) \to 1 \text{ as } n \to \infty$$

Therefore almost surely there exist $N \geq 1$ such that for all $n \geq N$ and for all $j \in [\![1, J]\!]$:

$$\mathbb{P}_{X,Y}\left(\left|(\sqrt{n}\mathbf{S}_n)_j\right| \geq \epsilon\right) \geq 1 - \frac{\gamma}{J}$$

Finally by applying a union bound we obtain that almost surely, for all $n \geq N$:

$$\mathbb{P}_{X,Y}\left(\forall j \in [|1, J|], \left|(\sqrt{n}\mathbf{S}_n)_j\right| \geq \epsilon\right) \geq 1 - \gamma$$

Therefore by applying Lemma 8, we obtain that, almost surely, for all $n \geq N$, with a probability of at least $1 - \gamma$:

$$\|\sqrt{n}\mathbf{S}_n\|_2 > \sqrt{\beta} \Rightarrow \|\sqrt{n}\mathbf{S}_n\|_1 > \delta.$$

C.3 Proof of the Proposition 3.2

Proposition 4. *Let $\{T_j\}_{j=1}^J$ sampled independently from the distribution Γ and $X := \{x_i\}_{i=1}^{N_1}$ and $Y := \{y_i\}_{i=1}^{N_2}$ be i.i.d. samples from P and Q respectively. Under H_0, the statistic L1-ME$[X, Y]$ is almost surely asymptotically distributed as Naka$(\frac{1}{2}, 1, J)$, a sum of J random variables i.i.d which follow a Nakagami distribution of parameter $m = \frac{1}{2}$ and $\omega = 1$. Finally under H_1, almost surely the statistic can be arbitrarily large as $t \to \infty$, allowing the test to correctly reject H_0.*

Proof. *Let us note the probability space of random variables $\{T_j\}_{j=1}^J$ as $(\Omega, \mathcal{F}, P)$. Let $\omega \in \Omega$ such that $d^{\omega}_{\ell_1,\mu,J}[P, Q] = 0$ (see Definition 3). Let us denote:*

$$\mathbf{Z}_X^{i,\omega} := \left(k\left(x_i, T_1\left(\omega\right)\right), ..., k\left(x_i, T_J\left(\omega\right)\right)\right) \qquad \mathbf{Z}_Y^{j,\omega} := \left(k\left(y_j, T_1\left(\omega\right)\right), ..., k\left(y_j, T_J\left(\omega\right)\right)\right),$$

$$\mathbf{S}^{\omega}_{N_1,N_2} := \frac{1}{N_1}\sum_{i=1}^{N_1} \mathbf{Z}_X^{i,\omega} - \frac{1}{N_2}\sum_{j=1}^{N_2} \mathbf{Z}_Y^{j,\omega}.$$

As $d^{\omega}_{\ell_1,\mu,J}[P, Q] = 0$ then for all j, $\mu_p\left(T_j\left(\omega\right)\right) = \mu_q\left(T_j\left(\omega\right)\right)$, which implies that $\mathbb{E}\left(\mathbf{Z}_X^{i,\omega}\right) = \mathbb{E}\left(\mathbf{Z}_Y^{j,\omega}\right)$. Therefore, by applying the Central-Limit Theorem, we have:

$$\sqrt{t}\,\mathbf{S}^{\omega}_{N_1,N_2} \longrightarrow \mathcal{N}\left(0, \frac{\mathbf{\Sigma}_1^{\omega}}{\rho}\right) - \mathcal{N}\left(0, \frac{\mathbf{\Sigma}_2^{\omega}}{1-\rho}\right) \qquad \textit{with} \quad \mathbf{\Sigma}_1 = \textit{Cov}\left(\mathbf{Z}_X^{\omega}\right) \quad \textit{and} \quad \mathbf{\Sigma}_2 = \textit{Cov}\left(\mathbf{Z}_Y^{\omega}\right)$$

As $\mathbf{Z}_X^{\omega}$ and $\mathbf{Z}_Y^{\omega}$ are independent, we have then that:

$$\sqrt{t}\,\mathbf{S}^{\omega}_{N_1,N_2} \longrightarrow \mathcal{N}\left(0, \mathbf{\Sigma}^{\omega}\right) \qquad \textit{with} \quad \mathbf{\Sigma}^{\omega} = \frac{\mathbf{\Sigma}_1^{\omega}}{\rho} + \frac{\mathbf{\Sigma}_2^{\omega}}{1-\rho}$$

And by Slutsky's theorem we deduce that:

$$\sqrt{t}\left(\mathbf{\Sigma}^{\omega}_{N_1,N_2}\right)^{-\frac{1}{2}} \mathbf{S}^{\omega}_{N_1,N_2} \longrightarrow \mathcal{N}\left(0, \mathbf{I}\right)$$

So by noting, $\sqrt{t}\left(\mathbf{\Sigma}^{\omega}_{N_1,N_2}\right)^{-\frac{1}{2}} \mathbf{S}^{\omega}_{N_1,N_2} = \left(W^{1,\omega}_{N_1,N_2}, ..., W^{J,\omega}_{N_1,N_2}\right)$, we have that for each coordinate:

$$\left(W^{j,\omega}_{N_1,N_2}\right) \longrightarrow \mathbf{S}_j^{\omega}$$

where $\left(\mathbf{S}_j^{\omega}\right)$ are i.i.d and follow a standard normal distribution. Therefore by considering the ℓ_1 norm of the statistic we have that:

$$||\sqrt{t}\left(\mathbf{\Sigma}^{\omega}_{N_1,N_2}\right)^{-\frac{1}{2}} \mathbf{S}^{\omega}_{N_1,N_2}||_1 \longrightarrow \sum_{j=1}^{J} |\mathbf{S}_j^{\omega}|$$

where $\left(\mathbf{S}_j^{\omega}\right)$ are independent and $\mathbf{S}_j^{\omega} \sim Naka\left(\frac{1}{2}, 1\right)$. And by assuming the null hypothesis $P = Q$, we have thanks to Theorem 4 that $d_{\ell_1,\mu,J}[P, Q] = 0$ a.s., then the result above hold a.s. Moreover, let's consider an ω such that $d^{\omega}_{\ell_1,\mu,J}[P, Q] > 0$. First we need show that $\left(\mathbf{\Sigma}^{\omega}_{N_1,N_2}\right)^{-\frac{1}{2}}$ converges in probability to the positive definite matrix $\left(\mathbf{\Sigma}^{\omega}\right)^{-\frac{1}{2}}$. For that we need to prove the following:

Lemma 10. *The function $h\left(\mathbf{X}\right) = \mathbf{X}^{-\frac{1}{2}}$ is well defined on $\mathcal{S}_J^{++}\left(R\right)$ and is continuous.*

Proof. *First we observe that h is the composition of two function which are:*

- *$h_1\left(\mathbf{X}\right) = \mathbf{X}^{-1}$ which is well defined and continuous on $\mathcal{S}_J^{++}\left(R\right)$*
- *$h_2\left(\mathbf{X}\right) = \mathbf{X}^{\frac{1}{2}}$ which is well defined on $\mathcal{S}_J^{+}\left(R\right)$ because each matrix of $\mathcal{S}_J^{+}\left(R\right)$ admits a unique square root matrix on $\mathcal{S}_J^{+}\left(R\right)$, so the result hold on $\mathcal{S}_J^{++}\left(R\right)$.*

Let us prove now the continuity of h_2. Let $\left(\mathbf{U}_n\right)$ a sequence in $\mathcal{S}_n^{++}\left(R\right)$ such that $\mathbf{U}_n \to \mathbf{U}$ and let us prove that $h_2\left(\mathbf{U}_n\right) \to h_2\left(\mathbf{U}\right)$ to prove the continuity of h_2. As $\left(\mathbf{U}_n\right)$ converges, then $\left(\mathbf{U}_n\right)$ is bounded, and we have:

$$|||\mathbf{U}_n|||\leq K \Longrightarrow |||h_2\left(\mathbf{U}_n\right)|||= \sqrt{|||\mathbf{U}_n|||} \leq \sqrt{K}$$

Then $(h_2(\mathbf{U}_n))$ is bounded. Let us show now that: $\forall \mathbf{A}$ s.t $\exists\ \phi$ strictly increasing and $h_2\left(\mathbf{U}_{\phi(n)}\right) \to \mathbf{A}$ we have $\mathbf{A} = h_2(\mathbf{U})$. Let $\mathbf{A}$ defined as above. Then $\exists\ \phi$ strictly increasing such that $h_2\left(\mathbf{U}_{\phi(n)}\right) \to \mathbf{A}$. As $\mathcal{S}_n^+(R)$ is closed, $\mathbf{A} \in \mathcal{S}_n^+(R)$, and by continuity of $\mathbf{M} \to \mathbf{M}^2$ we have also that $\mathbf{U}_{\phi(n)} \to \mathbf{A}^2$. And as $\mathbf{U}_n \to \mathbf{U}$, we have $\mathbf{A}^2 = \mathbf{U}$. And by uniqueness, we have finally:

$$h_2(\mathbf{U}) = \mathbf{A}.$$

So h_2 est continuous, and that conclude the proof.

Then each entry of the matrix $\mathbf{\Sigma}^{\omega}_{N_1,N_2}$ converges to the matrix $\mathbf{\Sigma}^{\omega}$, hence entires of the matrix $(\mathbf{\Sigma}^{\omega})^{-\frac{1}{2}}$, given by a continuous function of the entries of $\mathbf{\Sigma}^{\omega}$, are limit of the sequence $\left(\mathbf{\Sigma}^{\omega}_{N_1,N_2}\right)^{-\frac{1}{2}}$.

Similarly $\mathbf{S}^{\omega}_{N_1,N_2}$ converges in probability to the vector $\mathbf{S}^{\omega} = \mathbb{E}\left(\mathbf{Z}^{1,\omega}\right) - \mathbb{E}\left(\mathbf{Z}^{2,\omega}\right) \neq 0$. Since $\|(\mathbf{\Sigma}^{\omega})^{-\frac{1}{2}}\mathbf{S}^{\omega}\|_1 = \mathbf{A}_{\omega} > 0$ (indeed $(\mathbf{\Sigma}^{\omega})^{-\frac{1}{2}}$ is positive definite), then $\|\left(\mathbf{\Sigma}^{\omega}_{N_1,N_2}\right)^{-\frac{1}{2}}\mathbf{S}^{\omega}_{N_1,N_2}\|_1$, being a continuous function of the entries of $\mathbf{S}^{\omega}_{N_1,N_2}$ and $\left(\mathbf{\Sigma}^{\omega}_{N_1,N_2}\right)^{-\frac{1}{2}}$, converges to $\mathbf{A}_{\omega}$. Then

$$\mathbb{P}\left(\left\|\sqrt{t}(\mathbf{\Sigma}^{\omega}_{N_1,N_2})^{-\frac{1}{2}}\mathbf{S}^{\omega}_{N_1,N_2}\right\|_1 > r\right) = \mathbb{P}\left(\left\|\left(\mathbf{\Sigma}^{\omega}_{N_1,N_2}\right)^{-\frac{1}{2}}\mathbf{S}^{\omega}_{N_1,N_2}\right\|_1 - \frac{r}{\sqrt{t}} > 0\right)$$

And as $\frac{r}{\sqrt{t}} \to 0$ as $t \to \infty$, we have finally:

$$\mathbb{P}\left(\left\|\sqrt{t}\left(\mathbf{\Sigma}^{\omega}_{N_1,N_2}\right)^{-\frac{1}{2}}\mathbf{S}^{\omega}_{N_1,N_2}\right\|_1 > r\right) \to 1 \quad as \quad t \to \infty.$$

Finally, since $d_{\ell_1,\mu,J}[P,Q] > 0$ almost surely then $\mathbb{E}\left(\mathbf{Z}^{1,\omega}\right) - \mathbb{E}\left(\mathbf{Z}^{2,\omega}\right) \neq 0$ for almost all $\omega \in \Omega_1$, therefore under H_1, the statistic can be arbitrarily large as $t \to \infty$ almost surely.

D Optimizing test locations to improve power

D.1 Proof of Proposition 3.3

Proposition D.1. *Let $\mathcal{K}$ be a uniformly bounded family of $k : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$ measurable kernels (i.e., $\exists\ K < \infty$ such that $\sup_{k\in\mathcal{K}} \sup_{(x,y)\in(\mathbb{R}^d)^2} |k(x,y)| \leq K$). Let $\mathcal{V}$ be a collection in which each element is a set of J test locations. Assume that $c := \sup_{V\in\mathcal{V},k\in\mathcal{K}} \|\mathbf{\Sigma}^{-1/2}\| < \infty$. Then the test power $\mathbb{P}\left(\widehat{\lambda}_t \geq \delta\right)$ of the L1-ME test satisfies $\mathbb{P}\left(\widehat{\lambda}_t \geq \delta\right) \geq L(\lambda_t)$ where:*

$$L(\lambda_t) = 1 - 2\sum_{k=1}^{J} \exp\left(-\left(\frac{\lambda_t - \delta}{J^2 + J}\right)^2 \frac{\gamma_{N_1,N_2} N_1 N_2}{(N_1+N_2)^2}\right)$$
$$- 2\sum_{k,q=1}^{J} \exp\left(-2\frac{\left(\frac{\gamma_{N_1,N_2}}{K_3 J^2}\frac{\lambda_t - \delta}{(J^2+J)\sqrt{t}} - \frac{J^3 K_2}{\sqrt{\gamma_{N_1,N_2}}} - J^4 K_1\right)^2}{K_\lambda^2 (N_1+N_2)\max\left(\frac{8}{\rho N_1}, \frac{8}{(1-\rho)N_2}\right)^2}\right)$$

and K_1, K_2, K_3 and K_λ, are positive constants depending on only K, J and c. The parameter $\lambda_t := \|\sqrt{t}\mathbf{\Sigma}^{-\frac{1}{2}}\mathbf{S}\|_1$ is the population counterpart of $\widehat{\lambda}_t := \|\sqrt{t}(\Sigma_{N_1,N_2} + \gamma_{N_1,N_2}\mathbf{I})^{-\frac{1}{2}}\mathbf{S}_{N_1,N_2}\|_1$ where $\mathbf{S} = \mathbb{E}_{x,y}(S_{N_1,N_2})$ and $\mathbf{\Sigma} = \mathbb{E}_{x,y}(\mathbf{\Sigma}_{N_1,N_2})$. Moreover for large t, $L(\lambda_t)$ is increasing in λ_t.

Proof. *We will first find an upper bound of $|\widehat{\lambda}_t - \lambda_t|$, then we will compute a lower bound of $\mathbb{P}\left(\widehat{\lambda}_t > \delta\right)$. To simplify the notation In the following, we denote:*

$$\mathbf{\Sigma}_{N_1,N_2} := \frac{\mathbf{\Sigma}_{N_1}}{\rho} + \frac{\mathbf{\Sigma}_{N_2}}{1-\rho} + \gamma_{N_1,N_2}\mathbf{I} \tag{28}$$

such that $\widehat{\lambda}_t := \|\sqrt{t}(\Sigma_{N_1,N_2})^{-\frac{1}{2}}\mathbf{S}_{N_1,N_2}\|_1$. We have:

$$|\widehat{\lambda}_{N_1,N_2} - \lambda_t| = \left|\sqrt{t}\left(\|\mathbf{\Sigma}^{-\frac{1}{2}}_{N_1,N_2}\mathbf{S}_{N_1,N_2}\|_1 - \|\mathbf{\Sigma}^{-\frac{1}{2}}\mathbf{S}\|_1\right)\right|$$

Then we have:

$$\begin{aligned}\left|\|\mathbf{\Sigma}_{N_1,N_2}^{-\frac{1}{2}}\mathbf{S}_{N_1,N_2}\|_1-\|\mathbf{\Sigma}^{-\frac{1}{2}}\mathbf{S}\|_1\right| &\leq \|\mathbf{\Sigma}_{N_1,N_2}^{-\frac{1}{2}}\mathbf{S}_{N_1,N_2}-\mathbf{\Sigma}^{-\frac{1}{2}}\mathbf{S}\|_1 \\ &\leq \|\mathbf{\Sigma}_{N_1,N_2}^{-\frac{1}{2}}\mathbf{S}_{N_1,N_2}-\mathbf{\Sigma}_{N_1,N_2}^{-\frac{1}{2}}\mathbf{S}+\mathbf{\Sigma}_{N_1,N_2}^{-\frac{1}{2}}\mathbf{S}-\mathbf{\Sigma}^{-\frac{1}{2}}\mathbf{S}\|_1 \\ &\leq \|\mathbf{\Sigma}_{N_1,N_2}^{-\frac{1}{2}}\left(\mathbf{S}_{N_1,N_2}-\mathbf{S}\right)\|_1+\|\left(\mathbf{\Sigma}_{N_1,N_2}^{-\frac{1}{2}}-\mathbf{\Sigma}^{-\frac{1}{2}}\right)\mathbf{S}\|_1\end{aligned}$$

Let us now consider the first term on the right side of the inequality:

$$\|\mathbf{\Sigma}_{N_1,N_2}^{-\frac{1}{2}}\left(\mathbf{S}_{N_1,N_2}-\mathbf{S}\right)\|_1=\sum_{j=1}^{J}|\mathbf{\Sigma}_{N_1,N_2}^{-\frac{1}{2}}\left(\mathbf{S}_{N_1,N_2}-\mathbf{S}\right)|_j$$

But since $\mathbf{\Sigma}_{N_1,N_2}$ is symmetric definite positive, we can write:

$$\mathbf{\Sigma}_{N_1,N_2}=\mathbf{U}\mathbf{D}\mathbf{U}^T$$

where $\mathbf{U}$ is orthogonal and $\mathbf{D}=diag\left(\lambda_i\right)$ with $\lambda_i>0$. So:

$$\mathbf{\Sigma}_{N_1,N_2}^{-\frac{1}{2}}=\mathbf{U}\mathbf{D}^{-\frac{1}{2}}\mathbf{U}^T$$

But the regularization of $\mathbf{\Sigma}_{N_1,N_2}=(\frac{\mathbf{\Sigma}_{N_1}}{\rho}+\frac{\mathbf{\Sigma}_{N_2}}{1-\rho}+\gamma_{N_1,N_2}\mathbf{I})$ ensure that $\lambda_i\geq\gamma_{N_1,N_2}$. Thus $\lambda_i^{-\frac{1}{2}}\leq\gamma_{N_1,N_2}^{-\frac{1}{2}}$, and we have now:

$$\left|[\mathbf{\Sigma}_{N_1,N_2}^{-\frac{1}{2}}]_{i,j}\right|=\left|\sum_{j=1}^{J}\lambda_j^{-\frac{1}{2}}\left(\mathbf{U}_k\right)_i\left(\mathbf{U}_k\right)_j\right|$$

where $\mathbf{U}=[\mathbf{U}_1,...,\mathbf{U}_J]$ and $\|\mathbf{U}_k\|_2=1$. And finally:

$$\left|[\mathbf{\Sigma}_{N_1,N_2}^{-\frac{1}{2}}]_{i,j}\right|\leq\frac{J}{\sqrt{\gamma_{N_1,N_2}}}$$

Now we have:

$$\begin{aligned}\left\|\mathbf{\Sigma}_{N_1,N_2}^{-\frac{1}{2}}\left(\mathbf{S}_{N_1,N_2}-\mathbf{S}\right)\right\|_1 &\leq \sum_{j=1}^{J}\left|\sum_{k=1}^{J}[\mathbf{\Sigma}_{N_1,N_2}^{-\frac{1}{2}}]_{j,k}\left(\mathbf{S}_{N_1,N_2}-\mathbf{S}\right)_k\right| \\ &\leq \frac{J^2}{\sqrt{\gamma_{N_1,N_2}}}\sum_{k=1}^{J}|\left(\mathbf{S}_{N_1,N_2}-\mathbf{S}\right)_k| \\ &\leq \frac{J^2}{\sqrt{\gamma_{N_1,N_2}}}\sum_{k=1}^{J}|\mu_X\left(T_k\right)-\mu_Y\left(T_k\right)-\mathbb{E}\left(\mu_X\left(T_k\right)-\mu_Y\left(T_k\right)\right)|\end{aligned}$$

Let us note $\frac{\mathbf{\Sigma}_{N_1}}{\rho}+\frac{\mathbf{\Sigma}_{N_2}}{1-\rho}=\mathbf{M}_{N_1,N_2}$ and consider the second term of the inequality:

$$\begin{aligned}\mathbf{\Sigma}_{N_1,N_2}^{-\frac{1}{2}}-\mathbf{\Sigma}^{-\frac{1}{2}} &= \left(\mathbf{M}_{N_1,N_2}+\gamma_{N_1,N_2}\mathbf{I}\right)^{-\frac{1}{2}}-\mathbf{\Sigma}^{-\frac{1}{2}} \\ &= \left[\left(\mathbf{M}_{N_1,N_2}+\gamma_{N_1,N_2}\mathbf{I}\right)^{-\frac{1}{2}}-\left(\mathbf{\Sigma}+\gamma_{N_1,N_2}\mathbf{I}\right)^{-\frac{1}{2}}\right]+\left[\left(\mathbf{\Sigma}+\gamma_{N_1,N_2}\mathbf{I}\right)^{-\frac{1}{2}}-\mathbf{\Sigma}\right] \\ &= (1)+(2)\end{aligned}$$

Let us first consider (1):

$$\begin{aligned}(1) =&\mathbf{\Sigma}_{N_1,N_2}^{-\frac{1}{2}}\left(\left(\mathbf{\Sigma}+\gamma_{N_1,N_2}\mathbf{I}\right)^{\frac{1}{2}}-\left(\mathbf{M}_{N_1,N_2}+\gamma_{N_1,N_2}\mathbf{I}\right)^{\frac{1}{2}}\right)\left(\mathbf{\Sigma}+\gamma_{N_1,N_2}\mathbf{I}\right)^{-\frac{1}{2}} \\ =&\mathbf{\Sigma}_{N_1,N_2}^{-\frac{1}{2}}\left[\left(\mathbb{E}\left(\mathbf{M}_{N_1,N_2}+\gamma_{N_1,N_2}\mathbf{I}\right)\right)^{\frac{1}{2}}-\left(\mathbf{M}_{N_1,N_2}+\gamma_{N_1,N_2}\mathbf{I}\right)^{\frac{1}{2}}\right]\left(\mathbb{E}\left(\mathbf{M}_{N_1,N_2}+\gamma_{N_1,N_2}\mathbf{I}\right)\right)^{\frac{1}{2}} \\ =&\mathbf{\Sigma}_{N_1,N_2}^{-\frac{1}{2}}\left[\left(\mathbb{E}\left(\mathbf{\Sigma}_{N_1,N_2}\right)\right)^{\frac{1}{2}}-\mathbf{\Sigma}_{N_1,N_2}^{\frac{1}{2}}\right]\left(\mathbb{E}\left(\mathbf{\Sigma}_{N_1,N_2}\right)\right)^{-\frac{1}{2}} \\ =&\mathbf{\Sigma}_{N_1,N_2}^{-\frac{1}{2}}\left[\left(\mathbb{E}\left(\mathbf{\Sigma}_{N_1,N_2}^{\frac{1}{2}}\right)\right)-\mathbf{\Sigma}_{N_1,N_2}^{\frac{1}{2}}\right]\left(\mathbb{E}\left(\mathbf{\Sigma}_{N_1,N_2}\right)\right)^{\frac{1}{2}}+\mathbf{\Sigma}_{N_1,N_2}^{-\frac{1}{2}}\left[\left(\mathbb{E}\left(\mathbf{\Sigma}_{N_1,N_2}\right)\right)^{\frac{1}{2}}-\mathbb{E}\left(\mathbf{\Sigma}_{N_1,N_2}^{\frac{1}{2}}\right)\right]\left(\mathbb{E}\left(\mathbf{\Sigma}_{N_1,N_2}\right)\right)^{-\frac{1}{2}}\end{aligned}$$

And we have for (2)*:*

$$(2) = \left(\mathbf{\Sigma} + \gamma_{N_1,N_2}\mathbf{I}\right)^{-\frac{1}{2}} \left(\mathbf{\Sigma}^{\frac{1}{2}} - \left(\mathbf{\Sigma} + \gamma_{N_1,N_2}\mathbf{I}\right)^{\frac{1}{2}}\right) \mathbf{\Sigma}^{-\frac{1}{2}}$$

Thus we have:

$$\begin{aligned}\left\|\left(\mathbf{\Sigma}_{N_1,N_2}^{-\frac{1}{2}} - \mathbf{\Sigma}^{-\frac{1}{2}}\right)\mathbf{S}\right\|_1 \leq & \left\|\mathbf{\Sigma}_{N_1,N_2}^{-\frac{1}{2}}\left[\left(\mathbb{E}\left(\mathbf{\Sigma}_{N_1,N_2}^{\frac{1}{2}}\right)\right) - \mathbf{\Sigma}_{N_1,N_2}^{\frac{1}{2}}\right]\left(\mathbb{E}\left(\mathbf{\Sigma}_{N_1,N_2}\right)\right)^{-\frac{1}{2}}\mathbf{S}\right\|_1 \\ & + \left\|\mathbf{\Sigma}_{N_1,N_2}^{-\frac{1}{2}}\left[\left(\mathbb{E}\left(\mathbf{\Sigma}_{N_1,N_2}\right)\right)^{\frac{1}{2}} - \mathbb{E}\left(\mathbf{\Sigma}_{N_1,N_2}^{\frac{1}{2}}\right)\right]\left(\mathbb{E}\left(\mathbf{\Sigma}_{N_1,N_2}\right)\right)^{-\frac{1}{2}}\mathbf{S}\right\|_1 \\ & + \left\|\left(\mathbf{\Sigma} + \gamma_{N_1,N_2}\mathbf{I}\right)^{-\frac{1}{2}}\left(\mathbf{\Sigma}^{\frac{1}{2}} - \left(\mathbf{\Sigma} + \gamma_{N_1,N_2}\mathbf{I}\right)^{\frac{1}{2}}\right)\mathbf{\Sigma}^{-\frac{1}{2}}\mathbf{S}\right\|_1\end{aligned}$$

But we know that $|\mathbf{\Sigma}_{N_1,N_2}^{-\frac{1}{2}}|_{i,j} \leq \frac{J}{\sqrt{\gamma_{N_1,N_2}}}$ *and by the same reasoning we have also that* $|(\mathbf{\Sigma} + \gamma_{N_1,N_2}\mathbf{I})_{i,j}^{-\frac{1}{2}}| \leq \frac{J}{\sqrt{\gamma_{N_1,N_2}}}$*. By noting:*

$$\begin{aligned}\mathbf{K}_1 &= \sup_{k\in[\![1,J]\!]} |[\mathbf{\Sigma}^{-\frac{1}{2}}\mathbf{S}]_k| \\ \mathbf{K}_2 &= \sup_{k\in[\![1,J]\!]} \left|\left[\left(\mathbb{E}\left(\mathbf{\Sigma}_{N_1,N_2}\right)\right)^{\frac{1}{2}} - \mathbb{E}\left(\mathbf{\Sigma}_{N_1,N_2}^{\frac{1}{2}}\right)\left(\mathbb{E}\left(\mathbf{\Sigma}_{N_1,N_2}\right)\right)^{-\frac{1}{2}}\mathbf{S}\right]_k\right| \\ \mathbf{K}_3 &= \sup_{k\in[\![1,J]\!]} |[\left(\mathbb{E}\left(\mathbf{\Sigma}_{N_1,N_2}\right)\right)^{-\frac{1}{2}}\mathbf{S}]_k|\end{aligned}$$

All these constants are independent from $N_1, N_2, (x_i)$ *and* (y_j)*. We have finally:*

$$\begin{aligned}\left\|(\mathbf{\Sigma}_{N_1,N_2}^{-\frac{1}{2}} - \mathbf{\Sigma}^{-\frac{1}{2}})\mathbf{S}\right\|_1 &\leq \sum_{j=1}^{J}\sum_{q=1}^{J}\sum_{k=1}^{J}\left|\left(\mathbb{E}\left(\mathbf{\Sigma}_{N_1,N_2}^{\frac{1}{2}}\right)\right)_{q,k} - \left(\mathbf{\Sigma}_{N_1,N_2}^{\frac{1}{2}}\right)_{q,k}\right|\frac{\mathbf{K}_3 J}{\sqrt{\gamma_{N_1,N_2}}} + J^4\mathbf{K}_1 + \frac{J^3\mathbf{K}_2}{\sqrt{\gamma_{N_1,N_2}}} \\ &\leq \left[\sum_{q,k=1}^{J}\left|\left(\mathbb{E}\left(\mathbf{\Sigma}_{N_1,N_2}^{\frac{1}{2}}\right)\right)_{q,k} - \left(\mathbf{\Sigma}_{N_1,N_2}^{\frac{1}{2}}\right)_{q,k}\right|\right]\frac{\mathbf{K}_3 J^2}{\sqrt{\gamma_{N_1,N_2}}} + J^4\mathbf{K}_1 + \frac{J^3\mathbf{K}_2}{\sqrt{\gamma_{N_1,N_2}}}\end{aligned}$$

And by applying a union bound on all the terms that compose the upper bound of $|\widehat{\lambda}_{N_1,N_2} - \lambda_t|$ *we have thus:*

$$\begin{aligned}\mathbb{P}\left(\left|\widehat{\lambda}_t - \lambda_t\right| \leq \alpha\right) \geq & \sum_{k=1}^{J}\mathbb{P}\left(\sqrt{t}\frac{J^2}{\sqrt{\gamma_{N_1,N_2}}}|\mu_X(T_k) - \mu_Y(T_k) - \mathbb{E}\left(\mu_X(T_k) - \mu_Y(T_k)\right)| \leq \frac{\alpha}{J+J^2}\right) \\ & + \sum_{q,k=1}^{J}\mathbb{P}\left(\sqrt{t}\left(\left(\left|\left(\mathbb{E}\left(\mathbf{\Sigma}_{N_1,N_2}^{\frac{1}{2}}\right)\right)_{q,k} - \left(\mathbf{\Sigma}_{N_1,N_2}^{\frac{1}{2}}\right)_{q,k}\right|\right)\frac{\mathbf{K}_3 J^2}{\sqrt{\gamma_{N_1,N_2}}} + J^4\mathbf{K}_1 + \frac{J^3\mathbf{K}_2}{\sqrt{\gamma_{N_1,N_2}}}\right) \leq \frac{\alpha}{J^2+J}\right) \\ & - \left(J^2 + J - 1\right)\end{aligned}$$

As $\mu_X(T) - \mu_Y(T) = \sum_{k=1}^{t} Z_i$ *where* Z_k *are independent and:*

- $\forall i \leq N_1, Z_i = \frac{k(x_i,T)}{N_1}$*, so* $|Z_i| \leq \frac{K}{N_1}$
- $\forall N_1 < i \leq N_2, Z_i = -\frac{k(y_i,T)}{N_1}$ *so* $|Z_i| \leq \frac{K}{N_2}$

We have thanks to Hoeffding's inequality that $\forall k \in [\![1, J]\!]$:

$$\begin{aligned}\mathbb{P}\left(\sqrt{t}\frac{J^2}{\sqrt{\gamma_{N_1,N_2}}}\left|\mu_X(T_k) - \mu_Y(T_k) - \mathbb{E}\left(\mu_X(T_k) - \mu_Y(T_k)\right)\right| \leq \frac{\alpha}{J+J^2}\right) & \\ \geq 1 - 2\exp\left(-\left(\frac{\alpha}{J^2+J}\right)^2\frac{\gamma_{N_1,N_2}N_1N_2}{K^2\left(N_1+N_2\right)^2}\right) &\end{aligned}$$

Moreover $\forall k,q \in [\![1,J]\!]$:

$$\mathbb{P}\left[\sqrt{t}\left(\left|\left(\mathbb{E}\left(\mathbf{\Sigma}^{\frac{1}{2}}_{N_1,N_2}\right)\right)_{q,k} - \left(\mathbf{\Sigma}^{\frac{1}{2}}_{N_1,N_2}\right)_{q,k}\right| \frac{\mathbf{K}_3 J^2}{\sqrt{\gamma_{N_1,N_2}}} + \frac{J^4}{\mathbf{K}_1} + \frac{J^3 K_2}{\sqrt{\gamma_{N_1,N_2}}}\right) \leq \frac{\alpha}{J^2+J}\right]$$
$$= \mathbb{P}\left[\left|\left(\mathbf{\Sigma}^{\frac{1}{2}}_{N_1,N_2}\right)_{k,q} - \mathbb{E}\left(\mathbf{\Sigma}^{\frac{1}{2}}_{N_1,N_2}\right)_{k,q}\right| \leq \frac{\gamma_{N_1,N_2}}{\mathbf{K}_3 J^2}\left[\frac{\alpha}{(J^2+J)\sqrt{t}} - \left(\frac{J^3\mathbf{K}_2}{\sqrt{\gamma_{N_1,N_2}}} + J^4\mathbf{K}_1\right)\right]\right]$$

Let define $F(x_1,...,x_{N_1},y_1,...,y_{N_2}) := \mathbf{\Sigma}_{N_1,N_2}$ *and* $F_{k,q}(x_1,...,x_{N_1},y_1,...,y_{N_2}) := (\mathbf{\Sigma}_{N_1,N_2})_{k,q}$ *We can see easily that* $\forall (x_i),(y_i),x,x',y,y'$*:*

$$\left|F_{k,q}(x_1,..,x,..,x_{N_1},y_1,..,y_{N_2}) - F_{k,q}(x_1,..,x',..,x_{N_1},y_1,..,y_{N_2})\right| \leq \frac{8}{\rho N_1}$$

and

$$\left|F_{k,q}(x_1,..,x_{N_1},y_1,..y,..,y_{N_2}) - F_{k,q}(x_1,..,x_{N_1},y_1,..,y',..,y_{N_2})\right| \leq \frac{8}{(1-\rho)N_2}$$

Let $g(\mathbf{X}) = \mathbf{X}^{\frac{1}{2}}$ *defined on* $\mathbf{S}^{++}_J(R)$ *and takes values in* $\mathbf{S}^{++}_J(R)$*. This fuction is well defined because each matrix of* $\mathbf{S}^{++}_J(R)$ *admits a unique square root matrix on* $\mathbf{S}^{++}_J(R)$*. Moreover The result hold on* $\mathbf{S}^{+}_J(R)$*.*

Lemma 11. *g is locally Lipschitz continuous on* $\mathcal{S}^{++}_J(R)$ *which means that:*

$$\forall N>0,\ \forall \mathbf{X},\mathbf{Y} \in B(0,N) \subset \mathcal{S}^{++}_J(\mathbb{R}), \quad \exists K_N / \|g(\mathbf{X}) - g(\mathbf{Y})\| \leq K_N \|\mathbf{X}-\mathbf{Y}\|$$

Proof. *Let us first prove that g is* C^∞*. First thanks to Lemma 10 g is continuous on* $\mathbf{S}^{++}_J(R)$*. Let us show now that g is* C^∞ *on this space. We know that* $\mathbf{M} \to \mathbf{M}^2$ *induces a bijection from* $\mathcal{S}^{++}_n(R)$ *on itself where the inverse is g. To prove then that g is* C^∞*, thanks to the inverse function theorem, we just have to show that* $\mathbf{D}_{\mathbf{U}_0}\left(\mathbf{M} \to \mathbf{M}^2\right)$ *is invertible for every* $\mathbf{U}_0 \in \mathcal{S}^{++}_n(R)$*. Let* $\mathbf{U}_0 \in \mathcal{S}^{++}_n(R)$*. And let's consider the differential defined on* $\mathcal{S}_n(R)$ *in* $\mathcal{S}_n(R)$ *which is a linear application and which associates* $\mathbf{H}$ *to* $\mathbf{U}_0\mathbf{H} + \mathbf{H}\mathbf{U}_0$*. If we prove the injectivity of this function we will have its invertibility as* $\mathcal{S}_n(R)$ *is a finite dimensional space. Let* $\mathbf{H} \in \mathcal{S}_n(R)$ *such that* $\mathbf{U}_0\mathbf{H} + \mathbf{H}\mathbf{U}_0 = 0$ *and* $\mathbf{x}$ *an eigenvector of* $\mathbf{U}_0$ *associated with the eigenvalue* λ *which is strictly positive as* $\mathbf{U}_0$ *is definite positive. We have:*

$$\mathbf{U}_0\mathbf{H}\mathbf{x} = -\mathbf{H}\mathbf{U}_0\mathbf{x} = -\lambda\mathbf{H}\mathbf{x}$$

As $-\lambda < 0$ *it is not an eigenvalue of* $\mathbf{U}_0$ *and then* $\mathbf{H}\mathbf{x} = 0$*. This is true for all the eigenvectors of* $\mathbf{U}_0$*, then* $\mathbf{H} = 0$ *and the differential is injective, so g is* C^∞ *on* $\mathcal{S}^{++}_n(R)$*. Finally by applying the Mean value theorem, we have that g is locally Lipschitz continuous.*

We also remark that $||F(x_i,y_j)|| = \max_{i,j\in[\![1,J]\!]} |(\mathbf{\Sigma}_{N_1,N_2})_{i,j}| \leq \lambda$ *(because the Gaussian kernel is bounded) with* λ *independent from* $N_1,N_2,(x_i)$ *and* (y_j)*. Then by taking the following norm* $\|\mathbf{M}\| = \max_{i,j\in[\![1,J]\!]} \mathbf{M}_{i,j}$ *we have:*

$$\left\|g(F(x_1,..,x,..,x_{N_1},y_1,..,y_{N_2})) - g(F(x_1,..,x',..,x_{N_1},y_1,..,y_{N_2}))\right\|$$
$$\leq K_\lambda \left\|F(x_1,..,x,..,x_{N_1},y_1,..,y_{N_2}) - F(x_1,..,x',..,x_{N_1},y_1,..,y_{N_2})\right\|$$

And:

$$\left\|F(x_1,..,x,..,x_{N_1},y_1,..,y_{N_2}) - F(x_1,..,x',..,x_{N_1},y_1,..,y_{N_2})\right\| \leq \max\left(\frac{8}{\rho N_1}, \frac{8}{(1-\rho)N_2}\right)$$

Then $\forall k,q \in [\![1,J]\!]$:

$$\left|\mathbf{\Sigma}^{\frac{1}{2}}_{N_1,N_2}(x) - \mathbf{\Sigma}^{\frac{1}{2}}_{N_1,N_2}(x')\right| \leq K_\lambda \max\left(\frac{8}{\rho N_1}, \frac{8}{(1-\rho)N_2}\right)$$

And thanks to the McDiarmid inequality we have:

$$\mathbb{P}\left(\left|\left(\mathbf{\Sigma}^{\frac{1}{2}}_{N_1,N_2}\right)_{k,q} - \mathbb{E}\left(\mathbf{\Sigma}^{\frac{1}{2}}_{N_1,N_2}\right)_{k,q}\right| \leq \frac{\gamma_{N_1,N_2}}{K_3 J^2}\left[\frac{\alpha}{(J^2+J)\sqrt{t}} - \left(\frac{J^3 K_2}{\sqrt{\gamma_{N_1,N_2}}} + J^4 K_1\right)\right]\right)$$
$$\geq 1 - 2\exp\left(-2\frac{\left(\frac{\gamma_{N_1,N_2}}{K_3 J^2}\left(\frac{\alpha}{(J^2+J)\sqrt{t}} - \frac{J^3 K_2}{\sqrt{\gamma_{N_1,N_2}}} - J^4 K_1\right)\right)^2}{K_\lambda^2 (N_1+N_2)\max\left(\frac{8}{\rho N_1}, \frac{8}{(1-\rho)N_2}\right)^2}\right)$$

Then we have:

$$\mathbb{P}\left(|\widehat{\lambda}_{N_1,N_2} - \lambda_t| \leq \alpha\right) \geq 1 - 2\sum_{k=1}^{J}\exp\left(-\left(\frac{\alpha}{J^2+J}\right)^2 \frac{\gamma_{N_1,N_2} N_1 N_2}{(N_1+N_2)^2}\right)$$
$$- 2\sum_{k,q=1}^{J}\exp\left(-2\frac{\left(\frac{\gamma_{N_1,N_2}}{K_3 J^2}\left(\frac{\alpha}{(J^2+J)\sqrt{t}} - \frac{J^3 K_2}{\sqrt{\gamma_{N_1,N_2}}} - J^4 K_1\right)\right)^2}{K_\lambda^2 (N_1+N_2)\max\left(\frac{8}{\rho N_1}, \frac{8}{(1-\rho)N_2}\right)^2}\right)$$

And finally, by taking $\alpha = \lambda_t - \delta$ we have the result.

E Using smooth characteristic functions (SCF)

Theorem 5. *Let k be an analytic, integrable kernel with an inverse Fourier transform strictly greater than zero. For any $J > 0$, we define:*

$$d_{\Phi,J} = \left\{ d_{\Phi,J}[p,q] = \frac{1}{J}\sum_{j=1}^{J} |\Phi_p(T_j) - \Phi_q(T_j)| : \quad p, q \in \mathcal{M}_+^1\left(\mathbb{R}^d\right), \Phi_p, \Phi_q \in L_1\left(\mathbb{R}^d\right) \right\}$$

Then for any $J > 0$, $d_{\Phi,J}$ is a random metric on the space of Borel probability measures with integrable characteristic functions.

Proof. *Since k is an analytic, integrable kernel with an inverse Fourier transform strictly greater then zero then by the Lemma 3 the mapping $\Lambda : P \to \Phi_P$ is injective and $\Lambda(P)$ is an element of the RKHS associated with k. The Lemma 7 shows that Φ_P is analytic. Therefore we can use Lemma 6 to see that $d_{\Lambda,J}(P,Q) = d_{\Phi,J}(P,Q)$ is a random metric. This concludes the proof of the Theorem.*

E.1 Proof of Proposition 3.4

Proposition 5. *Let $\alpha \in]0,1[$, $\gamma > 0$ and $J \geq 2$. Let $\{T_j\}_{j=1}^J$ sampled i.i.d. from the distribution Γ and let $X := \{x_i\}_{i=1}^n$ and $Y := \{y_i\}_{i=1}^n$ i.i.d. samples from P and Q respectively. Let us denote δ the $(1-\alpha)$-quantile of the asymptotic null distribution of $\widehat{d}_{\ell_1,\Phi,J}[X,Y]$ and β the $(1-\alpha)$-quantile of the asymptotic null distribution of $\widehat{d}^2_{\ell_2,\Phi,J}[X,Y]$. Under the alternative hypothesis, almost surely, there exists $N \geq 1$ such that for all $n \geq N$, with a probability of at least $1-\gamma$ we have:*

$$\widehat{d}^2_{\ell_2,\Phi,J}[X,Y] > \beta \Rightarrow \widehat{d}_{\ell_1,\Phi,J}[X,Y] > \delta \tag{29}$$

Proof. *Let us first introduce the following Lemma:*

Lemma 12. *Let $\mathbf{x}$ a random vector $\in \mathbb{C}^J$ with $J \geq 2$, $\epsilon > 0$, $\gamma > 0$ and $\mathbf{z} := \min_{j\in[|1,J|]} |Re(x_j)| + |Im(x_j)|$ where Im and Re are respectively the imaginary and real part functions. Moreover let denote $\mathbf{X} := (Im(x_j), Re(x_j))_{j=1}^J \in \mathbb{R}^{2J}$. If*

$$\mathbb{P}(\mathbf{z} \geq \epsilon) \geq 1 - \gamma$$

we have with a probability of at least $1-\gamma$ that, $\forall t_1 \geq t_2 \geq 0$, if $\epsilon \geq \sqrt{\frac{t_1^2 - t_2^2}{J(J-1)}}$, then

$$\|\mathbf{X}\|_2 > t_2 \Rightarrow \|\mathbf{X}\|_1 \geq t_1.$$

Proof. *First we remarks that:*

$$\epsilon > \sqrt{\frac{t_1^2 - t_2^2}{J(J-1)}} \Rightarrow J(J-1)\epsilon > t_1^2 - t_2^2$$
$$\Rightarrow t_2^2 > t_1^2 - J(J-1)\epsilon^2$$

Therefore, we have:

$$\|\mathbf{X}\|_2 \geq t_2 \Rightarrow \|\mathbf{X}\|_2^2 + J(J-1)\epsilon^2 \geq t_1^2$$
$$\Rightarrow \sqrt{\|\mathbf{X}\|_2^2 + J(J-1)\epsilon^2} \geq t_1$$

But we have that:

$$\|\mathbf{X}\|_1^2 = \|\mathbf{X}\|_2^2 + \sum_{i \neq j} (|Im(x_i)| + |Re(x_i)|)(|Im(x_j)| + |Re(x_j)|)$$

Therefore we have with a probability of 1-γ that:

$$\|\mathbf{X}\|_1^2 \geq \|\mathbf{X}\|_2^2 + J(J-1)\epsilon^2$$

And:

$$\|\mathbf{X}\|_2 \geq t_2 \Rightarrow \|\mathbf{X}\|_1 \geq t_1$$

Moreover by denoting δ the $(1-\alpha)$-quantile of the asymptotic null distribution of $\widehat{d}_{\ell_1,\Phi,J}[X,Y]$ and β the $(1-\alpha)$-quantile of the asymptotic null distribution of $\widehat{d}^2_{\ell_2,\Phi,J}[X,Y]$ thanks to Lemma 9, we have that $\delta \geq \sqrt{\beta}$. Therefore to show the result we only need to show that the assumption of the Lemma 12 is sastified for the random vector $\mathbf{X} := \sqrt{n}\mathbf{S}_n \in \mathbb{R}^{2J}$, $t_1 = \delta$ and $t_2 = \sqrt{\beta}$, i.e. for $\epsilon = \sqrt{\frac{\delta^2 - \beta}{J(J-1)}}$ under the alternative hypothesis. Under $H_1 : P \neq Q$, we have $\mathbf{S}_n$ converges in probability to the vector $\mathbf{S}$ where $\mathbf{\Sigma} := \mathbb{E}_{(x,y)\sim(p,q)}(\Sigma_n)$ and $\mathbf{S} := \mathbb{E}_{(x,y)\sim(p,q)}(\mathbf{S}_n)$. Moreover we have $\mathbf{S} = (Im(\Phi_P(T_j) - \Phi_Q(T_j)), Re(\Phi_P(T_j) - \Phi_Q(T_j)))_{j=1}^J \in \mathbb{R}^{2J}$. Indeed, according to the Definition 2, we have for all $j \in [|1, J|]$:

$$\begin{aligned}
\phi_P(T_j) &:= \int_{\epsilon \in \mathbb{R}^d} \psi_P(\epsilon) k(T_j - \epsilon) d\epsilon \\
&= \int_{\epsilon \in \mathbb{R}^d} \int_{x \in \mathbb{R}^d} \exp(ix^T\epsilon) k(T_j - \epsilon) dP(x) d\epsilon \\
&= \int_{x \in \mathbb{R}^d} \left(\int_{\epsilon \in \mathbb{R}^d} \exp(ix^T(\epsilon - T_j)) k(\epsilon - T_j) \right) \exp(ix^T T_j) d\epsilon dP(x) \\
&= \int_{x \in \mathbb{R}^d} f(x) \exp(ix^T T_j) dP(x)
\end{aligned}$$

and all these equalities hold as k is integrable. Lemma 3 guarantees the injectivity of the function $\Gamma : P \to \Phi_P$, and as $P \neq Q$, therefore $\Phi_P - \Phi_Q$ is a non-zero function. Moreover Φ_P and Φ_Q live in the RKHS H_k associated with k. Therefore thanks to Lemma 7, $\Phi_P - \Phi_Q$ is analytic. Therefore thanks to Lemma 5, $\Phi_P - \Phi_Q$ is almost surely non zero. Moreover the $(T_j)_{j=1}^J$ are independent, therefore almost surely $(|\Phi_P(T_j) - \Phi_Q(T_j)|)_j$ are all non zero, and then $(|Im(\Phi_P(T_j) - \Phi_Q(T_j)))| + |Re(\Phi_P(T_j) - \Phi_Q(T_j))|)_j$ are all non zero. Then by continuity of the functions defined for all $k \in [|1, J|]$ by:

$$\phi_k : x := (x_j^1, x_j^2)_{j=1}^J \in \mathbb{R}^{2J} \to |x_k^1| + |x_k^1| \tag{30}$$

We have that for all $k \in [|1, J|]$, $\phi_k(\mathbf{S}_n)$ converge in probability towards $\phi_k(\mathbf{S})$, which are almost surely all non zeros. Then for all $k \in [\![1, J]\!]$ we have:

$$\mathbb{P}\left(\left| (\sqrt{n}\phi_k(\mathbf{S}_n) \right| > \epsilon \right) = \mathbb{P}_{X,Y}\left(\left| (\phi_k(\mathbf{S}_n) \right| - \frac{\epsilon}{\sqrt{n}} > 0 \right)$$

And as $\frac{\epsilon}{\sqrt{n}} \to 0$ as $n \to \infty$, we have finally almost surely for all $k \in [\![1, J]\!]$:

$$\mathbb{P}_{X,Y}\left(\left| (\sqrt{n}\phi_k(\mathbf{S}_n) \right| \geq \epsilon \right) \to 1 \text{ as } n \to \infty$$

Therefore almost surely there exist $N \geq 1$ such that for all $n \geq N$ and for all $k \in [\![1, J]\!]$:

$$\mathbb{P}_{X,Y}\left(\left|(\sqrt{n}\phi_k(\mathbf{S}_n)\right| \geq \epsilon\right) \geq 1 - \frac{\gamma}{J}$$

Finally by applying a union bound we obtain that almost surely, for all $n \geq N$:

$$\mathbb{P}_{X,Y}\left(\forall k \in [|1, J|],\ \left|(\sqrt{n}\phi_k(\mathbf{S}_n)\right| \geq \epsilon\right) \geq 1 - \gamma$$

Therefore by applying Lemma 8, we obtain that, almost surely, for all $n \geq N$, with a probability of at least $1 - \gamma$:

$$\widehat{d}_{\ell_2,\Phi,J}[X, Y] > \sqrt{\beta} \Rightarrow \widehat{d}_{\ell_1,\Phi,J}[X, Y] > \delta$$

F Experiments

F.1 Realization of the ℓ_1-based tests

Indeed to realize these tests, we need to compute the $1-\alpha$ quantile of the $Nake\left(\frac{1}{2}, 1, J\right)$. To do so we need to obtain the cumulative distribution function (CDF) of the sum of J Nakagami i.i.d. But as we do not have a closed form of this distribution, we need to estimate this CDF by considering the empirical distribution function. Indeed to generate samples from $Nake\left(\frac{1}{2}, 1, J\right)$, it is sufficient to generate samples from multivariate normal distribution $\mathcal{N}\left(0, I_{d_J}\right)$, and to sum the absolute values of the J coordinates of theses vectors.

Moreover, we have the following result:

Theorem 6. ***(Dvoretzky–Kiefer–Wolfowitz inequality)*** *Let x_1,...,x_n be real-valued independent and identically distributed random variables with cumulative distribution function $F(.)$ Let F_n denote the associated empirical distribution function defined by:*

$$F_n(t) = \frac{1}{n}\sum_{i=1}^{n} \mathbf{1}_{x_i \leq t}$$

Then we have $\forall \epsilon > 0$:

$$\mathbb{P}\left(||F_n - F||_\infty > \epsilon\right) \leq 2e^{-2n\epsilon^2}$$

Finally we have, $F(x) - \epsilon \leq F_n(x) \leq F(x) + \epsilon$ with a probability of $1 - \delta$ where $\epsilon = \sqrt{\frac{ln\left(\frac{2}{\delta}\right)}{2n}}$.

Then with a probability of 99%, and by taking $n = 100\,000$ samples i.i.d of the $Naka\left(\frac{1}{2}, 1, J\right)$, we can estimate the CDF with an error of $\epsilon \leq 0.0051$, which is less than $\alpha = 0.01$.

Optimization: The lower bounds that we optimize to perform **L1-opt-ME** and **L1-opt-SCF** are non-convex, as in the prior art [16]. However, the use of the ℓ_1-norm makes optimization even harder, as it is no longer a smooth. Moreover we need to differentiate through the inverse square root matrix operation which can lead is some cases to degenerate matrices during the gradient ascent. Therefore to avoid this, we decide to check at each step the convergence of the inverse square root matrix operation. Further work should consider dedicated optimization algorithms.

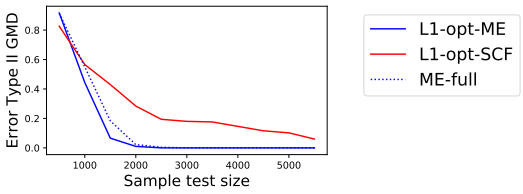
Table 3 gives the run times of the different optimized tests on the Blobs problem when the test sample size is $n^{te} = 1e6$.

	L1-opt-ME	**ME-full**	**L1-opt-SCF**	**SCF-full**
Run Time (s)	164.23	157.97	599.77	579.42

Table 3: Run times of the optimized tests when $n^{te} = 1e6$ and $J = 2$ for the blobs problem.

Software implementation: as the expression of the optimization objective is rather complicated, we use the automatic differentiation of pytorch [22], to compute its gradient, and then proceed with a gradient ascent where the step size after t iterations is the inverse of the euclidean norm of the gradient times $\sqrt{t}$. The specific code can be found at `https://github.com/meyerscetbon/l1_two_sample_test`.

Figure 5: Plot of type-II error against the test sample size n^{te} in the following toy problem: $P = \mathcal{N}(0, I_d)$ and $Q = \mathcal{N}\left((0.3, 0, .., 0)^T, I_d\right)$ with $d = 100$



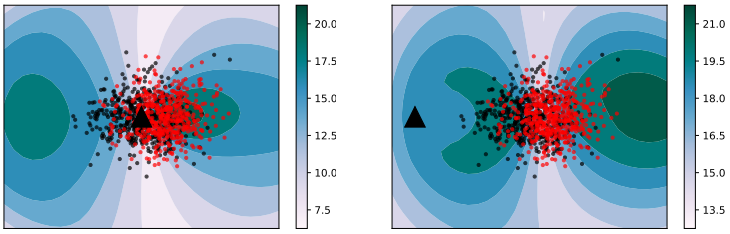
F.2 Experiments on a more difficult problem

In Figure 5, we consider the following GMD problem: $P \sim \mathcal{N}(0, I_d)$, $Q \sim \mathcal{N}\left((0.3, 0, .., 0)^T, I_d\right)$ with $d = 100$. The figure shows that when the problem of GMD is more difficult, we can see that **L1-opt-ME** performs the best.

F.3 Informative features

We show that the optimization of the proxy $\widehat{\lambda}_t^{tr}(\theta)$ for the test power in the ℓ_1 case is informative for revealing the difference of the two samples in the ME test as in [16] with the ℓ_2 version. We consider the Gaussian Mean Difference (GMD) problem (see Table 1), where both P and Q are two-dimensional normal distributions with different means. We use $J = 2$ test locations T_1 and T_2 , where T_1 is fixed to the location indicated by the black triangle in Figure 6. The contour plot shows $T_2 \to \widehat{\lambda}_t^{tr}(T_1, T_2)$.

Figure 6: **Illustrating interpretable features**, replicating in the ℓ_1 case the figure of [16]. A contour plot of $\widehat{\lambda}_t^{tr}(T_1, T_2)$ as a function of T_2, when $J = 2$, and T_1 is fixed. The red and black dots represent the samples from the P and Q distributions, and the big black triangle the position of T_1.



(a) T_1 is centered (b) T_1 lives in the left region

Figure 6a suggests that $\widehat{\lambda}_t^{tr}(T_1, T_2)$ is maximized when T_2 is placed in either of the two regions that captures the difference of the two samples i.e., the region in which the probability masses of P and Q have less overlap. In Figure 6b, we consider placing T_1 in one of the two key regions. In this case, the contour plot shows that T_2 should be placed in the other region to maximize $\widehat{\lambda}_t^{tr}(T_1, T_2)$, implying that placing multiple test locations in the same neighborhood does not increase the discriminability. The two modes on the left and right suggest two ways to place the test location in a region that reveals the difference. The non-convexity of the $\widehat{\lambda}_t^{tr}(T_1, T_2)$ is an indication of many informative ways to detect differences of P and Q, rather than a drawback. A convex objective would not capture this multimodality.

F.4 Real problem: 20 newsgroups text dataset

In this experiment we use the 20 newsgroups text dataset from [17] which comprises around 18000 newsgroups posts on 20 topics. We consider 3 categories which are: "comp", "sci", and "alt" . The first category is about components in hardware systems, the second is about sciences and spaces, and the last is about religion. To perform the tests we need to embed these documents in a metric space. For this, we use the TF-IDF matrix by group of two categories with a $df \geq 30$, which lead to embed the documents in spaces of 3 000 dimensions approximately. Then we perform the two-sample tests on the embedded documents. We compare the distribution of "sci" documents with others, as well as with itself to evaluate the level of the tests. The number of samples of each category is not the same,

P	Q	L1-opt-ME	L1-grid-ME	L1-opt-SCF	L1-grid-SCF	ME-full	SCF-full
sci (1187)	sci (1187)	**0.00**	**0.00**	0.004	**0.00**	1	0.002
sci (1187)	comp (292)	**0.00**	0.496	**0.00**	0.170	**0.00**	0.634
sci (1187)	alt (240)	**0.00**	0.370	**0.00**	0.064	**0.00**	0.510

Table 4: Type-I errors and Type-II errors of various the L1-tests in the problem of distinguishing the newsgroups text dataset. $\alpha = 0.01$. $J = 2$. The number in brackets denotes the test sample size of each samples.

hence to perform the tests from [16], we take randomly n_{min} samples for both distributions without replacement (where n_{min} is in fact the number of samples of the distributions compared to the sci distribution). We set the number of location $J = 2$.

Type-I errors and type-II errors are summarized in Table 4 The two first columns indicates the categories of the papers in the two samples. This task represents a case in which H_0 holds. In this case all the tests are conservative except the **ME-full** test which totally rejecting the null hypothesis. In the other problems, we show the Type-II errors of our tests. The ℓ_1 optimized tests perform very well, which shows that the locations learned are indeed discriminant. The ℓ_1 approaches bring a clear gain in statistical control compared to their ℓ_2 counterparts.

F.5 Real problem: fast-food distribution

Problem	L1-opt-ME	L1-grid-ME	L1-opt-SCF	L1-grid-SCF	ME-full	SCF-full	MMD-quad
McDo vs McDo (2002)	0.010	0.000	0.000	0.000	0.012	0.000	0.000

Table 5: **Fast food dataset:** Type-I errors for distinguishing the distribution of fast food restaurants. $\alpha = 0.01$. $J = 3$. The number in brackets denotes the sample size of the distribution on the right. We consider MMD-quad as the gold standard.

Table 5 summarizes Type-I errors observed on the Mac Donald's vs Mac Donald's problem. It shows that the optimized tests based on mean embeddings stay roughly at the specified level $\alpha = 0.01$ when H_0 hold, and others are more conservative.

Figures 8, 9, 10, 11, 12 give the distributions of the data (restaurant locations) and of the T_J for each of the problems that we consider.

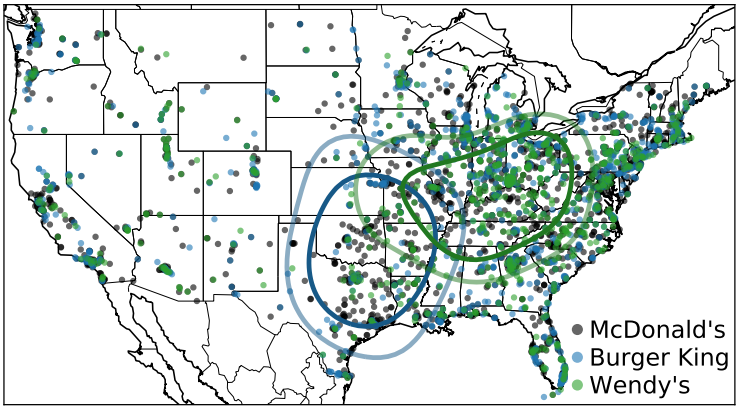


Figure 7: **Fast food data:** Visualizing interpretable locations for differences in Mc Donald's vs Burger King and Mc Donald's vs Wendy's. The lines correspond to the distribution of the locations chosen for the T_J features by the L1-opt-ME procedure. The distributions are estimated with a kernel density estimate. The lines represent the contours probabilities 80% and 90%.

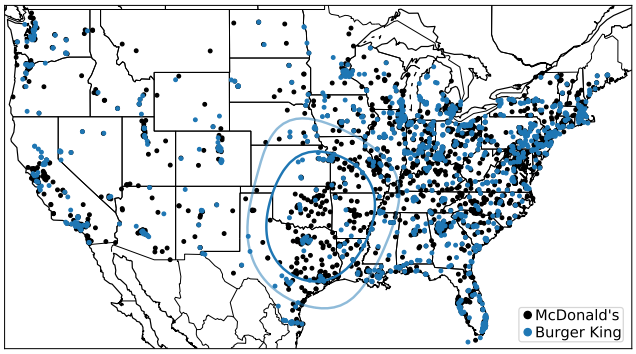


Figure 8: **Mc Donald's vs Burger King**

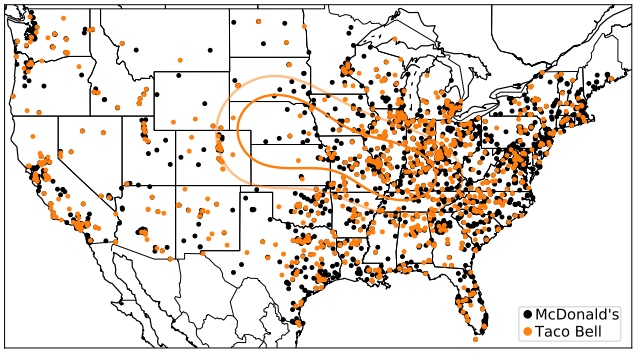


Figure 9: **Mc Donald's vs Taco Bell**

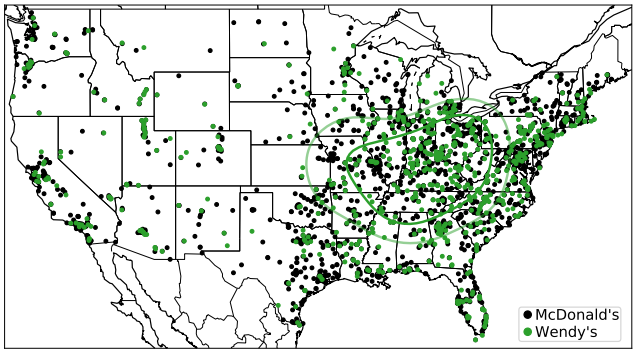


Figure 10: **Mc Donald's vs Wendy's**

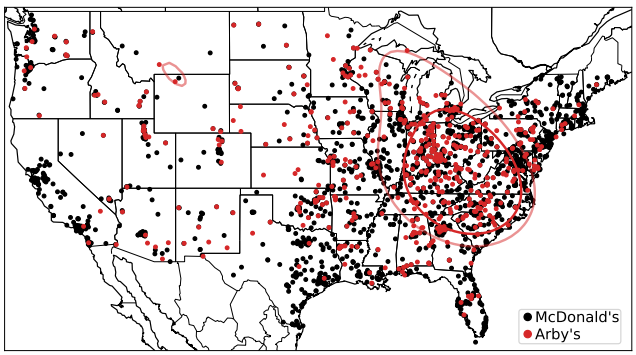


Figure 11: **Mc Donald's vs Arby's**

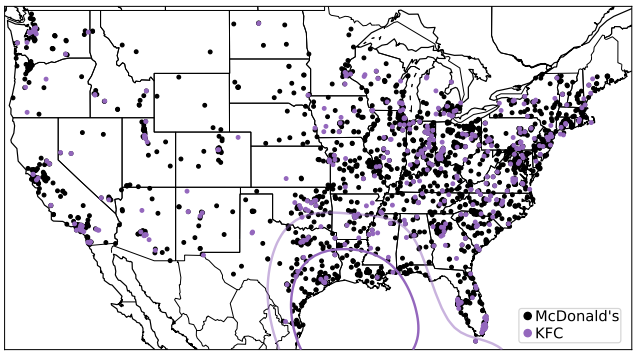


Figure 12: **Mc Donald's vs KFC**