Comparing distributions: ℓ_1 geometry improves kernel two-sample testing

Meyer Scetbon, Gaël Varoquaux
Inria, Université Paris-Saclay

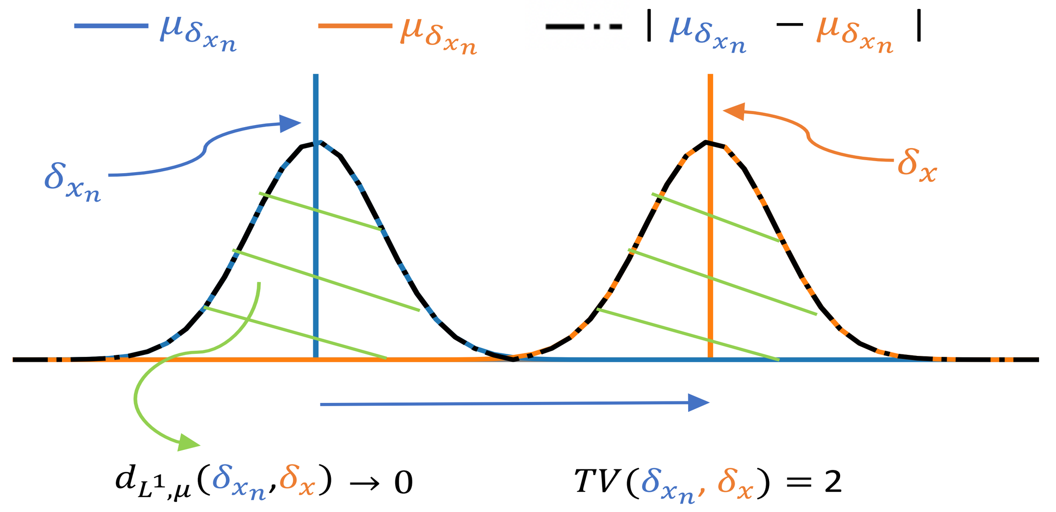
Overview

Problem: Are two sets of observations drawn from the same distribution?

Contributions:

- We exhibit a family of L^p-based metrics which metrize the weak convergence.
- We derive linear-time, nonparametric, a.s - consistent L^1-based two sample tests.
- We show L^1geometry provides better power than its L^2 counterpart.
- We maximize a lower bound on the test power and learn distinguishing features between distributions.

Weak Convergence



Theorem: Let k a characteric and bounded kernel. For all $p \geq 1$,

$$d_{L^p,\mu}(P,Q) := \left(\int_t |\mu_P(t) - \mu_Q(t)|^p \, d\Gamma(t)\right)^{1/p}$$

where $\mu_P(t) := \int_t k(x,t)dP(x)$ is a metric which metrize the weak convergence.

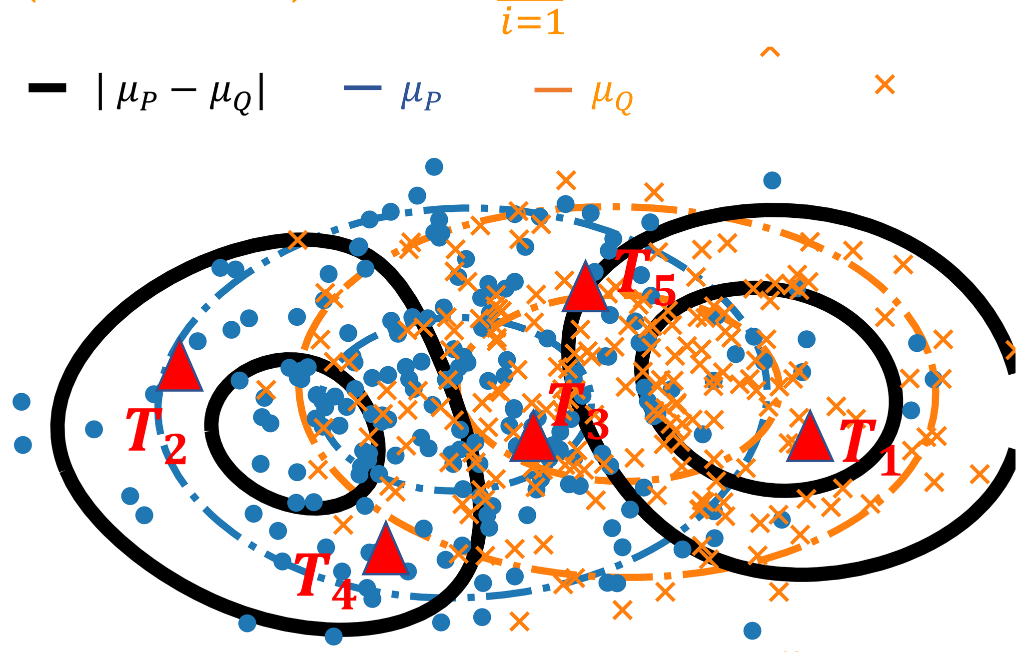
Sketch of proof:

- Integral Operator: $T_k: f \in L_2^{d\Gamma}(\mathbb{R}^d) \longrightarrow \int_{t \in \mathbb{R}^d} k(t,.)f(t)d\Gamma$
- Unit Ball of $L_\infty^{d\Gamma}(\mathbb{R}^d)$: $B_\infty^{d\Gamma} := \{f: \sup|f(x)| \leq 1 \; \Gamma\text{ - a.s}\}$
- IPM formulation: $d_{L^1,\mu}(P,Q) = \sup_{f \in T_k(B_\infty^{d\Gamma})} \{E_P(f(X)) - E_Q(f(Y))\}$

Mean Embedding test

- Test $H_0: P = Q$ vs $H_1: P \neq Q$:
- Samples: $X := \{x_i\}_{i=1}^n \sim P$ and $Y := \{y_i\}_{i=1}^n \sim Q$
- Empirical ME: $\mu_X(T) := \frac{1}{n}\sum_{i=1}^n k_\sigma(x_i, T)$
- k_σ the Gaussian kernel of width σ
- Test locations: $\{T_i\}_{i=1}^J \sim \Gamma$
- Test statistic:

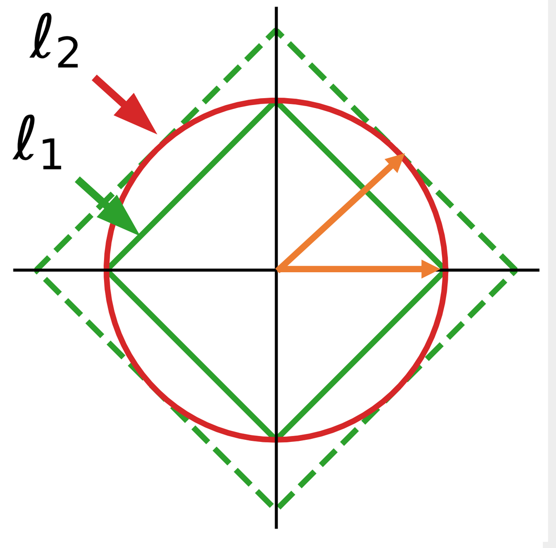
$$\left(\hat{d}_{\ell_p,\mu}(X,Y)\right)^p := n^{\frac{p}{2}} \sum_{i=1}^{J} |\mu_X(T_i) - \mu_Y(T_i)|^p$$



Test of level α : Compute $\left(\hat{d}_{\ell_p,\mu}(X,Y)\right)^p$ and reject H_0 if $\left(\hat{d}_{\ell_p,\mu}(X,Y)\right)^p > \zeta_{\alpha,p} = 1 - \alpha$ quantile of the null distribution.

Why $\ell_1 \gg \ell_2$?

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



Proposition: Let. $\delta > 0$. Under the alternative hypothesis H_1, almost surely there exists $N \geq 1$, such that for all $n \geq N$, with a probability of $1 - \delta$:

$$\left(\hat{d}_{\ell_2,\mu}(X,Y)\right)^2 > \zeta_{\alpha,2} \implies \hat{d}_{\ell_1,\mu}(X,Y) > \zeta_{\alpha,1}$$

Normalized Tests

Remark: Under H_0, $\hat{d}_{\ell_1,\mu}(X,Y)$ converge to a sum of correlated Nakagami variables.

Normalized Mean Embedding (ME) Test:

$$\text{L1-ME}[X,Y] := ||\sqrt{n}\, \Sigma_n^{-\frac{1}{2}} S_n||_1$$

- $S_n := \frac{1}{n}\sum_{i=1}^n Z_X^i - Z_Y^i$
- $\Sigma_n := \widehat{cov}(Z_X) + \widehat{cov}(Z_Y)$
- $Z_X^i := (k_\sigma(x_i, T_1), \dots, k_\sigma(x_i, T_J))$

Proposition: Under H_0, L1-ME$[X,Y]$ is a.s asymptotically distributed as a sum of J i.i.d Nakagami variables of parameter $m = \frac{1}{2}$ and $\varpi = \frac{1}{2}$.

Normalized Smooth Characteristic Function (SCF) Test:

$$\text{L1-SCF}[X,Y] := ||\sqrt{n}\, \Sigma_n^{-\frac{1}{2}} S_n||_1$$

- $Z_X^i := (\cos(x_i^T T_1)f(x_i), \sin(x_i^T T_1)f(x_i), \dots, \sin(x_i^T T_J)f(x_i))$
- f is the inverse Fourier transform of k_σ.

Optimization Procedure

Regularized Statistic:

$$\text{L1-ME}[X,Y] := ||\sqrt{n}\,(\Sigma_n + \gamma_n)^{-1/2} S_n||_1$$

- $\gamma_n \longrightarrow 0$

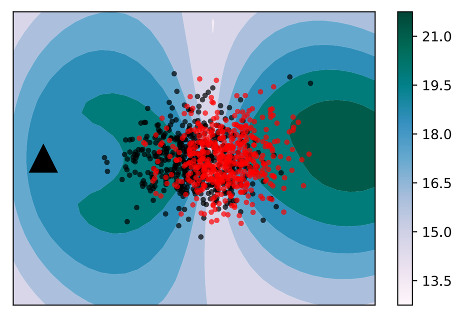
Proposition: The test power $P(\text{L1-ME}[X,Y] > \epsilon)$ of the the L1-ME test satisifes $P(\text{L1-ME}[X,Y] > \epsilon) \geq L(\lambda_n)$ where $L(\lambda_n)$ is an increasing function of λ_n and goes to 1 when n goes to infinity.

- $\lambda_n := ||\sqrt{n}\, \Sigma^{-\frac{1}{2}} S||_1$ is the population counterpart of L1-ME$[X,Y]$.

Optimization Procedure:

- Optimize $\{T_i\}_{i=1}^J, \sigma = argmax\, L(\lambda_n)$
- Estimation of λ_n on a separate training set.

Informative Features



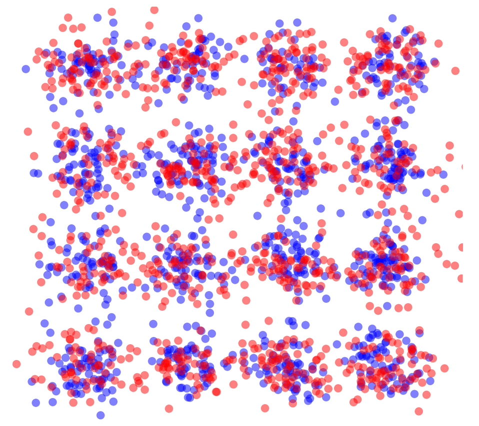
Contour plot of L1$-$ME$[X,Y]$ as a function of T_2 with $J = 2$ and T_1 fixed.

- $P \sim N([0,0], I_2)$
- $Q \sim N([0,1], I_2)$
- L1$-$ME$[X,Y]$ detects the differences.

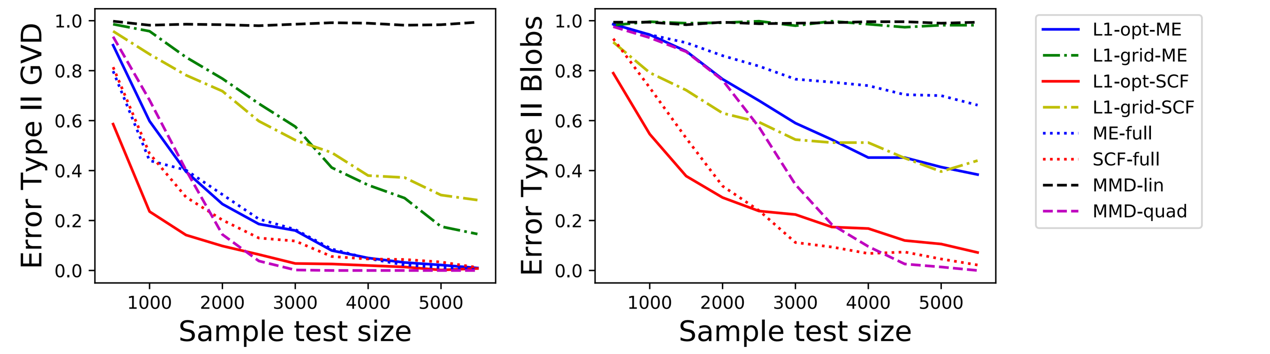
Test Power: Synthetic Problems

- **Test Power vs. Sample Size**

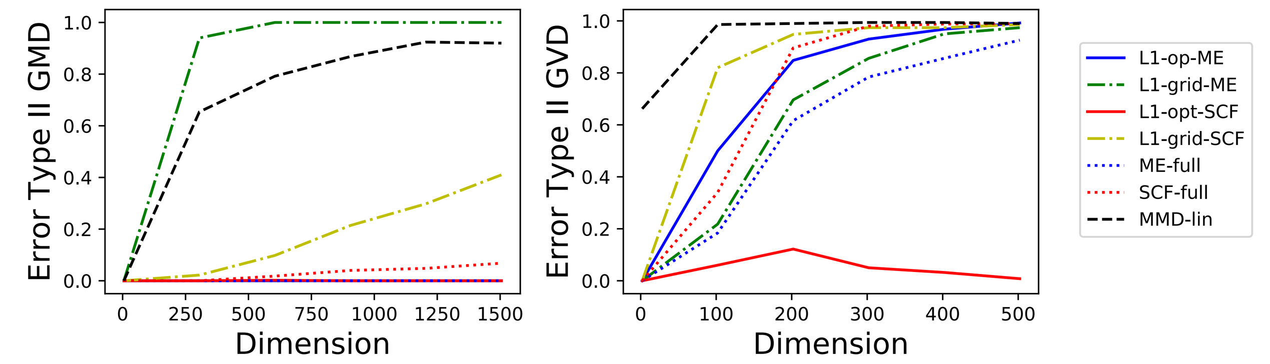
Data	P	Q
GMD	$\mathcal{N}(0, I_d)$	$\mathcal{N}((1,0,..,0)^T, I_d)$
GVD	$\mathcal{N}(0, I_d)$	$\mathcal{N}(0, diag(2,1,..,1))$
Blobs	Mixture of 16 Gaussians in $\mathbb{R}^2$	



- L1-opt-ME, L1-opt-SCF: Proposed Methods
- L1-grid-ME, L1-grid-SCF: Random settings
- ME-full, SCF-full: Optimized ℓ_2-based methods
- MMD-quad, MMD-lin: Quadratic and linear-time MMD tests



- **Test Power vs. Dimension**



Higgs Dataset

Higgs Dataset: $d = 4$, J= 3. Plot of Type-II error for ℓ_1 and ℓ_2 based test .

- Optimized tests outperform their random versions.
- ℓ_1 norm provides better power.