Two-Sample Test with Kernel Projected Wasserstein Distance

Jie Wang[1], Rui Gao[2], and Yao Xie[1]

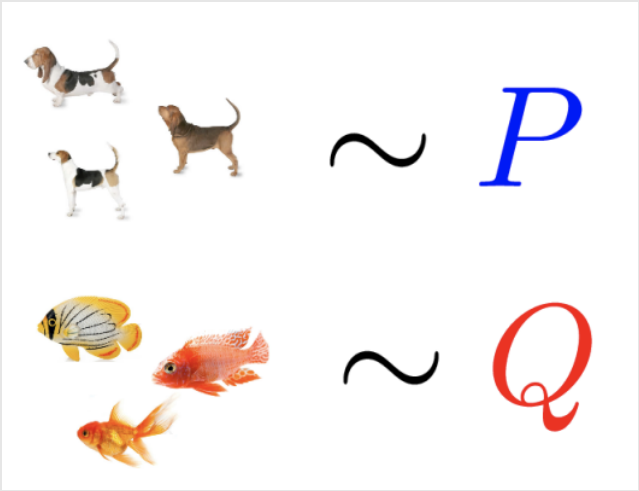
[1]Georgia Institute of Technology

[2]The University of Texas at Austin

Oral Presentation at *Artificial Intelligence and Statistics 2022*

Motivation: Comparing Two Samples

- Given: Samples from unknown distributions P and Q.
- Goal: Do P and Q differ?



Problem Setup

Given two independent sample sets:

$$X = \{x_1, \ldots, x_n\} \sim P, Y = \{y_1, \ldots, y_m\} \sim Q,$$

A test $T: (X, Y) \mapsto \{d_0, d_1\}$ decide:

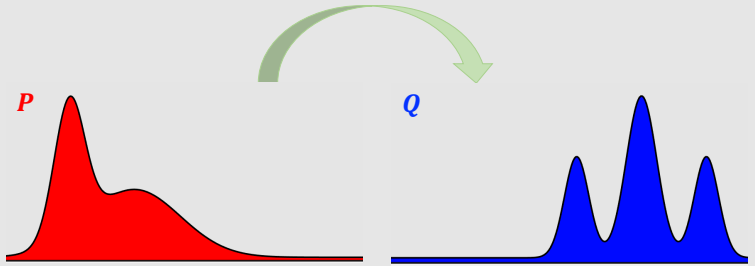
$$\mathscr{H}_0: P = Q, \qquad \mathscr{H}_1: P \neq Q.$$

Risk functions:

$$\text{Type-I Error}: \mathbb{P}_{x^n \sim P, y^m \sim Q}\left(T(x^n, y^m) = d_1\right), \quad \text{under } \mathscr{H}_0,$$

$$\text{Type-II Error}: \mathbb{P}_{x^n \sim P, y^m \sim Q}\left(T(x^n, y^m) = d_0\right), \quad \text{under } \mathscr{H}_1.$$

Wasserstein Distance



$$W(P,Q) = \min_{g \in \mathbb{G}(P,Q)} \left\{ \mathbb{E}_{(w,w') \sim g} [\|w - w'\|] \right\}$$

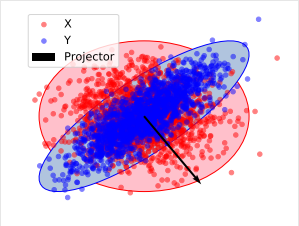
- Cheapest cost of transporting probability mess from one distribution to another.
- Advantages:
  - Flexibility: non-overlapping support, discrete and continuous.
  - Geometric properties.

Projected Wasserstein Distance

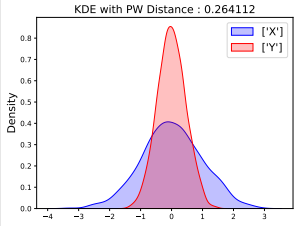
- Design projected Wasserstein distance for testing[1]:

$$\mathscr{P}W(P,Q) = \max_{\mathscr{A}\in\mathbb{V}_d} W(\mathscr{A}\#P, \mathscr{A}\#Q).$$

  - Find linear projector $\mathscr{A} \in \mathbb{V}_d = \{\mathscr{A} : \mathscr{A}(z) = A^\mathsf{T} z, A^\mathsf{T} A = I_d\}$ for which the Wasserstein distance between the projected distributions is as large as possible.



(a) Scatter plots



(b) Projected samples

[1] Jie Wang, Rui Gao, and Yao Xie. "Two-sample Test using Projected Wasserstein Distance". In: *Proceedings of IEEE International Symposium on Information Theory*. July 2021.

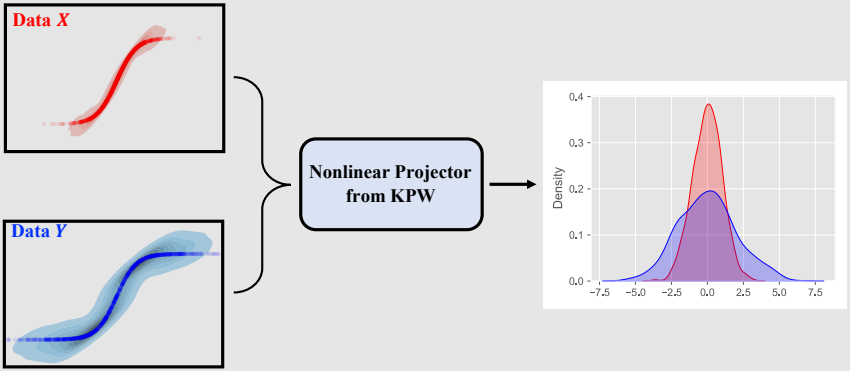
Kernel Projected Wasserstein Distance

- Develop kernel projected Wasserstein distance for testing:

$$KPW(P,Q) = \max_{f \in \mathscr{F}} W(f\#P,\ f\#Q)$$

$$\text{where} \quad \mathscr{F} = \{f \in \mathscr{H} :\ \|f\|_{\mathscr{H}} \leq 1\}.$$

where $\mathscr{H}$ is a $\mathbb{R}^d$-valued RKHS.



Reproducing Kernel Hilbert Space (RKHS)

Scalar-valued RKHS

- $K: \mathbb{R}^D \times \mathbb{R}^D \to \mathbb{R}$ is P.S.D. if

$$\sum_{i=1}^{N}\sum_{j=1}^{N} y_i K(x_i, x_j) y_j \geq 0, \ \forall x_i \in \mathbb{R}^D, y_i \in \mathbb{R}$$

- K induces a scalar-valued RKHS $\mathscr{H}_K$.
- Reproducing Property:

$$f(x) = \langle f, K_x \rangle_{\mathscr{H}_K}, \quad \forall f \in \mathscr{H}_K,$$

where the kernel section

$$K_x(x') \triangleq K(x', x), \quad \forall x' \in \mathbb{R}^D.$$

Vector-valued RKHS[2]

- $K: \mathbb{R}^D \times \mathbb{R}^D \to \mathbb{R}^{d \times d}$ is P.S.D. if

$$\sum_{i=1}^{N}\sum_{j=1}^{N} \langle y_i, K(x_i, x_j) y_j \rangle \geq 0, \ \forall x_i \in \mathbb{R}^D, y_i \in \mathbb{R}^d$$

- K induces a vector-valued RKHS $\mathscr{H}_K$.
- Reproducing Property:

$$\langle f(x), y \rangle = \langle f, K_x y \rangle_{\mathscr{H}_K} \quad \forall f \in \mathscr{H}_K,$$

where the kernel section

$$(K_x y)(x') \triangleq K(x', x) y, \quad \forall x' \in \mathbb{R}^D, y \in \mathbb{R}^d.$$

[2] Charles A. Micchelli and Massimiliano A. Pontil. "On Learning Vector-Valued Functions". In: *Neural Computation* 17.1 (Jan. 2005), pp. 177–204.

KPW Test

- Develop kernel projected Wasserstein distance for testing:

$$KPW(P,Q) = \max_{f \in \mathscr{F}} W(f\#P,\ f\#Q)$$

$$\text{where} \quad \mathscr{F} = \{f \in \mathscr{H} :\ \|f\|_{\mathscr{H}} \leq 1\}.$$

  where $\mathscr{H}$ is a $\mathbb{R}^d$-valued RKHS.

  1. Compute nonlinear projector in training dataset;
  2. Perform permutation test in testing dataset.

- Outline:
  - Computing KPW Distance
  - (Finite-sample Guarantee)
  - Numerical Simulation

Algorithmic Considerations

Computing KPW distance is equivalent to:

$$KPW(P_n, Q_m) = \max_{f \in \mathscr{H}:\ \|f\|_{\mathscr{H}}^2 \le 1} \left\{ \min_{p \in G} \sum_{i,j} p_{i,j} \|f(x_i) - f(y_j)\|^2 \right\}.$$

- Infinite-dimensional optimization – Develop a representer theorem:

  There exists an optimal solution $\hat{f}$ with

  $$\hat{f}(z) = \sum_{i=1}^{n} K(z, x_i) a_{x,i} - \sum_{j=1}^{m} K(z, y_j) a_{y,j}, \quad z \in \mathbb{R}^D,$$

  where $a_{x,i}, a_{y,j} \in \mathbb{R}^d$ are coefficients.

- Non-convex problem – Focus on finding stationary point.

$$KPW(P_n, Q_m) = \max_{f \in \mathscr{H}:\ \|f\|_{\mathscr{H}}^2 \le 1} \left\{ \min_{p \in G} \sum_{i,j} p_{i,j} \|f(x_i) - f(y_j)\|^2 \right\}.$$

- Step 1: Substituting the form of representer theorem:

$$\hat{f}(z) = \sum_{i=1}^{n} K(z, x_i) a_{x,i} - \sum_{j=1}^{m} K(z, y_j) a_{y,j}$$

$$= \underbrace{\begin{bmatrix} K(z,x_1) & K(z,x_2) & \cdots & K(z,x_n) & -K(z,y_1) & -K(z,y_2) & \cdots & -K(z,y_m) \end{bmatrix}}_{G(z; x^n, y^m)} \underbrace{\begin{bmatrix} a_{x,1} \\ \vdots \\ a_{x,n} \\ a_{y,1} \\ \vdots \\ a_{y,m} \end{bmatrix}}_{\omega}$$

- Step 1: Substituting the form of representer theorem:

$$G \triangleq \begin{bmatrix} G(x_1; x^n, y^m) \\ \vdots \\ G(x_n; x^n, y^m) \\ -G(y_1; x^n, y^m) \\ \vdots \\ -G(y_n; x^n, y^m) \end{bmatrix} \qquad \omega \triangleq \begin{bmatrix} a_{x,1} \\ \vdots \\ a_{x,n} \\ a_{y,1} \\ \vdots \\ a_{y,m} \end{bmatrix}$$

$$KPW(P_n, Q_m) = \max_{\omega} \left\{ \min_{\pi \in \Gamma} \sum_{i,j} \pi_{i,j} c_{i,j} : \ \omega^T G \omega \leq 1 \right\}$$

Algorithmic Development

- Step 1: Substituting the form of representer theorem:

$$\max_{w} \left\{ \min_{p \in \mathsf{G}} \mathring{\mathrm{a}}_{i,j} p_{i,j} c_{i,j} : \; w^{\mathsf{T}} G w \leq 1 \right\}.$$

- Step 2: Take $G^{-1} = UU^{\mathsf{T}}$ and $s = U^{-1}w$, we have

$$\max_{s \in \mathbb{S}^{d(n+m)-1}} \left\{ \min_{p \in \mathsf{G}} \mathring{\mathrm{a}}_{i,j} p_{i,j} c_{i,j} \right\}, \quad \text{where } \mathbb{S}^{d(n+m)-1} \text{ is a sphere.}$$

- Step 3: Adding entropic regularization and reformulate by duality:

$$\max_{s \in \mathbb{S}^{d(n+m)-1}} \left\{ \min_{p \in \mathsf{G}} \mathring{\mathrm{a}}_{i,j} p_{i,j} c_{i,j} - hH(p) \right\}$$

$$= \min_{u,v,s} \left\{ F(u,v,s) : \; s \in \mathbb{S}^{d(n+m)-1}, u \in \mathbb{R}^n, v \in \mathbb{R}^m \right\}.$$

Convergence Analysis

We say that $(\hat{u}, \hat{v}, \hat{s})$ is a (e_1, e_2)-stationary point if

$$\|\mathrm{Grad}_s F(\hat{u}, \hat{v}, \hat{s})\| \leq e_1,$$
$$F(\hat{u}, \hat{v}, \hat{s}) - \min_{u,v} F(u, v, \hat{s}) \leq e_2.$$

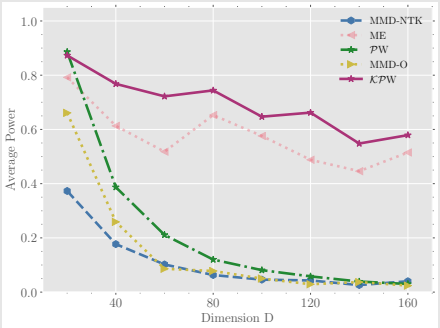
Proposed method returns an (e_1, e_2)-stationary point within

- iteration number:
$$\mathscr{O}\left(\log(mn) \cdot \left[\frac{1}{e_2^3} + \frac{1}{e_1^2 e_2}\right]\right),$$
- Arithmetic operations:
$$\mathscr{O}\left(N^3 d^3 \log(N) \cdot \left[\frac{1}{e_2^3} + \frac{1}{e_1^2 e_2}\right]\right),$$
- Storage requirement: $\mathscr{O}(d^2 N^2)$.

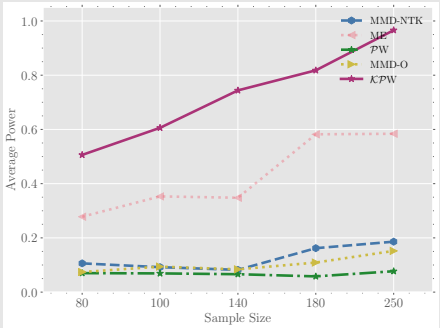
Testing Power for Synthetic Datasets (Gaussian Mixture)

Baseline:

- Projected Wasserstein test (PW) (Wang, Gao, and Xie 2021);
- Gaussian MMD test with optimized bandwidth (MMD-O) (Liu et al. 2020);
- MMD test combining neural network (NTK-MMD) (Cheng and Xie 2021);
- ME test with optimized hyper-parameters (ME) (Jitkrittum et al. 2016).



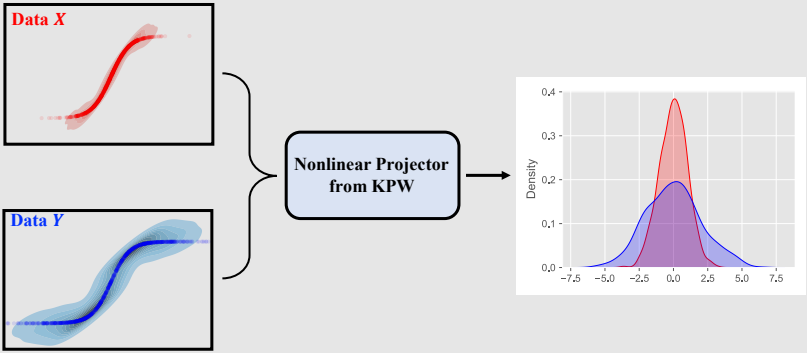
(a) Power v.s. Dimension



(b) Power v.s. Sample Size

Two-Sample Test with Kernel Projected Wasserstein Distance

Online Available: arxiv.org/abs/2102.06449



SCAN ME