Multivariate Distribution-free Testing using Optimal Transport

Bodhisattva Sen[1]
Department of Statistics
Columbia University, New York

EMFCSC workshop: Robustness and Resilience in Stochastic Optimization and Statistical Learning: Mathematical Foundations

Erice, Italy
21 May, 2022

[1]Supported by NSF grant DMS-2015376

Collaborators

Nabarun Deb
PhD student
Columbia (Stats)

Promit Ghosal
Post-doc
MIT (Math)

Bhaswar Bhattacharya
Assistant Professor
UPenn (Stats)

Multivariate two-sample testing

- **Data**: $\{X_i\}_{i=1}^m$ iid P_1 on $\mathbb{R}^d$; $\{Y_j\}_{j=1}^n$ iid P_2 on $\mathbb{R}^d$, $d \geq 1$

- Test if the two samples came from the same distribution, i.e.,

$$\mathrm{H}_0 : P_1 = P_2 \qquad \text{versus} \qquad \mathrm{H}_1 : P_1 \neq P_2$$

Multivariate two-sample testing

- **Data**: $\{X_i\}_{i=1}^m$ iid P_1 on $\mathbb{R}^d$; $\{Y_j\}_{j=1}^n$ iid P_2 on $\mathbb{R}^d$, $d \geq 1$

- Test if the two samples came from the same distribution, i.e.,

$$\mathrm{H}_0 : P_1 = P_2 \qquad \text{versus} \qquad \mathrm{H}_1 : P_1 \neq P_2$$

- When $d = 1$: Student (1908), Wilcoxon (1945), Cramér von-Mises (1928), Smirnov (1939), Wald and Wolfowitz (1940), Mann and Whitney (1947), Anderson (1962), ...

- When $d > 1$: Hotelling (1931), Weiss (1960), Bickel (1969), Friedman and Rafsky (1979), Schilling (1986), Henze (1988), Liu and Singh (1993), Székely (2003), Rosenbaum (2005), Gretton et al. (2012), Biswas et al. (2014), Chen and Friedman (2017), ...

When $d = 1$

- **Two-sample** t**-test**: Compares $\bar{X}_m = \frac{1}{m}\sum_{i=1}^{m} X_i$ & $\bar{Y}_n = \frac{1}{n}\sum_{j=1}^{n} Y_j$

When $d = 1$

- **Two-sample t-test**: Compares $\bar{X}_m = \frac{1}{m}\sum_{i=1}^{m} X_i$ & $\bar{Y}_n = \frac{1}{n}\sum_{j=1}^{n} Y_j$
- Reject H_0 if test statistic is larger than $(1-\alpha)$-th quantile of t_{m+n-2}
- Approximate (not valid for small sample sizes) level α test; requires additional moment assumptions; not robust to outliers

When $d = 1$

- **Two-sample *t*-test**: Compares $\bar{X}_m = \frac{1}{m}\sum_{i=1}^m X_i$ & $\bar{Y}_n = \frac{1}{n}\sum_{j=1}^n Y_j$
- Reject H_0 if test statistic is larger than $(1-\alpha)$-th quantile of t_{m+n-2}
- Approximate (not valid for small sample sizes) level α test; requires additional moment assumptions; not robust to outliers

Question: Can we find a distribution-free test which is also efficient, and robust to outliers and contamination?

When $d = 1$

- **Two-sample t-test**: Compares $\bar{X}_m = \frac{1}{m}\sum_{i=1}^m X_i$ & $\bar{Y}_n = \frac{1}{n}\sum_{j=1}^n Y_j$
- Reject H_0 if test statistic is larger than $(1-\alpha)$-th quantile of t_{m+n-2}
- Approximate (not valid for small sample sizes) level α test; requires additional moment assumptions; not robust to outliers

Question: Can we find a distribution-free test which is also efficient, and robust to outliers and contamination?

Answer: Wilcoxon rank-sum test [Wilcoxon (1945)]

- Distribution-free: Null distribution is universal — does not depend on the underlying distribution of the data
- Exact test valid for all sample sizes; robust to outliers
- Based on univariate ranks — advent of classical nonparametrics

Comparison of Wilcoxon rank-sum (WRS) test with two-sample t-test

Pool $(X_1, \ldots, X_m, Y_1, \ldots, Y_n)$: (scaled) ranks $\widehat{R}_{m,n}(X_i)$'s and $\widehat{R}_{m,n}(Y_j)$'s

$$\frac{1}{n}\sum_{j=1}^{n}\widehat{R}_{m,n}(Y_j) - \frac{1}{m}\sum_{i=1}^{m}\widehat{R}_{m,n}(X_i)$$

- WRS test is distribution-free and exact for all $P_1 = P_2$ continuous,

Comparison of Wilcoxon rank-sum (WRS) test with two-sample t-test

Pool $(X_1, \ldots, X_m, Y_1, \ldots, Y_n)$: (scaled) ranks $\widehat{R}_{m,n}(X_i)$'s and $\widehat{R}_{m,n}(Y_j)$'s

$$\frac{1}{n}\sum_{j=1}^{n}\widehat{R}_{m,n}(Y_j) - \frac{1}{m}\sum_{i=1}^{m}\widehat{R}_{m,n}(X_i)$$

- WRS test is distribution-free and exact for all $P_1 = P_2$ continuous, as under H_0, $\left(\widehat{R}_{m,n}(X_1), \ldots, \widehat{R}_{m,n}(X_m), \widehat{R}_{m,n}(Y_1), \ldots, \widehat{R}_{m,n}(Y_n)\right)$ is distributed uniformly over the $(m+n)!$ permutations of $\left\{\frac{1}{m+n}, \frac{2}{m+n}, \ldots, 1\right\}$

Comparison of Wilcoxon rank-sum (WRS) test with two-sample t-test

Pool $(X_1, \ldots, X_m, Y_1, \ldots, Y_n)$: (scaled) ranks $\widehat{R}_{m,n}(X_i)$'s and $\widehat{R}_{m,n}(Y_j)$'s

$$\frac{1}{n}\sum_{j=1}^{n} \widehat{R}_{m,n}(Y_j) - \frac{1}{m}\sum_{i=1}^{m} \widehat{R}_{m,n}(X_i)$$

- WRS test is distribution-free and exact for all $P_1 = P_2$ continuous, as under H_0, $\left(\widehat{R}_{m,n}(X_1), \ldots, \widehat{R}_{m,n}(X_m), \widehat{R}_{m,n}(Y_1), \ldots, \widehat{R}_{m,n}(Y_n)\right)$ is distributed uniformly over the $(m+n)!$ permutations of $\left\{\frac{1}{m+n}, \frac{2}{m+n}, \ldots, 1\right\}$

- WRS test has 0.95 Pitman efficiency w.r.t t-test when P_1 is Gaussian

Comparison of Wilcoxon rank-sum (WRS) test with two-sample t-test

Pool $(X_1, \ldots, X_m, Y_1, \ldots, Y_n)$: (scaled) ranks $\widehat{R}_{m,n}(X_i)$'s and $\widehat{R}_{m,n}(Y_j)$'s

$$\frac{1}{n}\sum_{j=1}^{n} \widehat{R}_{m,n}(Y_j) - \frac{1}{m}\sum_{i=1}^{m} \widehat{R}_{m,n}(X_i)$$

- WRS test is distribution-free and exact for all $P_1 = P_2$ continuous, as under H_0, $\left(\widehat{R}_{m,n}(X_1), \ldots, \widehat{R}_{m,n}(X_m), \widehat{R}_{m,n}(Y_1), \ldots, \widehat{R}_{m,n}(Y_n)\right)$ is distributed uniformly over the $(m+n)!$ permutations of $\left\{\frac{1}{m+n}, \frac{2}{m+n}, \ldots, 1\right\}$
- WRS test has 0.95 Pitman efficiency w.r.t t-test when P_1 is Gaussian
- Non-trivial efficiency lower bound of 0.864 w.r.t t-test [Hodges and Lehmann (1956)]; efficiency can be $+\infty$ (for heavy-tailed dist.)

Comparison of Wilcoxon rank-sum (WRS) test with two-sample t-test

Pool $(X_1, \ldots, X_m, Y_1, \ldots, Y_n)$: (scaled) ranks $\widehat{R}_{m,n}(X_i)$'s and $\widehat{R}_{m,n}(Y_j)$'s

$$\frac{1}{n}\sum_{j=1}^{n} \widehat{R}_{m,n}(Y_j) - \frac{1}{m}\sum_{i=1}^{m} \widehat{R}_{m,n}(X_i)$$

- WRS test is distribution-free and exact for all $P_1 = P_2$ continuous, as under H_0, $\left(\widehat{R}_{m,n}(X_1), \ldots, \widehat{R}_{m,n}(X_m), \widehat{R}_{m,n}(Y_1), \ldots, \widehat{R}_{m,n}(Y_n)\right)$ is distributed uniformly over the $(m+n)!$ permutations of $\left\{\frac{1}{m+n}, \frac{2}{m+n}, \ldots, 1\right\}$

- WRS test has 0.95 Pitman efficiency w.r.t t-test when P_1 is Gaussian

- Non-trivial efficiency lower bound of 0.864 w.r.t t-test [Hodges and Lehmann (1956)]; efficiency can be $+\infty$ (for heavy-tailed dist.)

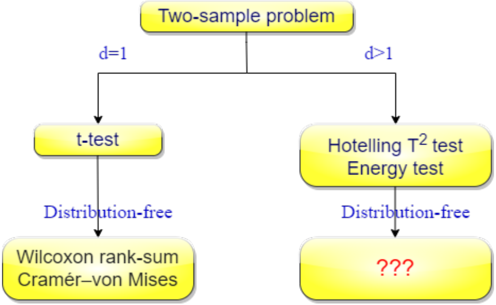
- Non-trivial efficiency lower bound of 1 w.r.t t-test [Chernoff and Savage (1958)] when the following revised statistic is used:

$$\frac{1}{n}\sum_{j=1}^{n} \Phi^{-1}(\widehat{R}_{m,n}(Y_j)) - \frac{1}{m}\sum_{i=1}^{m} \Phi^{-1}(\widehat{R}_{m,n}(X_i))$$

Generalize all these properties to multivariate data

Question: Can we construct multivariate robust distribution-free tests?

Question: Can we construct multivariate robust distribution-free tests?



- When $d = 1$ tests based on "ranks" are distribution-free
- How do we define multivariate ranks that lead to distribution-free tests?
- What about their statistical efficiency?

Optimal transport!

Outline

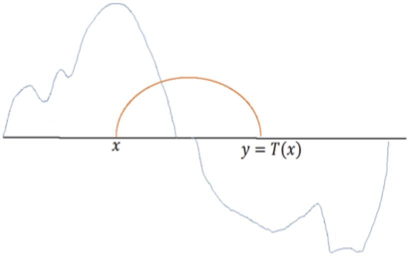
1. Optimal Transport: Monge's Problem
   - Introduction
   - Multivariate Ranks via Optimal Transport
2. Multivariate Two-sample Goodness-of-fit Testing
   - Hotelling T^2 and Kernel MMD
   - Distribution-free Testing
   - Lower bounds on Asymptotic (Pitman) Relative Efficiency
3. Testing for Independence Between Two Random Vectors
   - Distance Covariance
   - Distribution-free Testing

Outline

1. Optimal Transport: Monge's Problem
   - Introduction
   - Multivariate Ranks via Optimal Transport
2. Multivariate Two-sample Goodness-of-fit Testing
   - Hotelling T^2 and Kernel MMD
   - Distribution-free Testing
   - Lower bounds on Asymptotic (Pitman) Relative Efficiency
3. Testing for Independence Between Two Random Vectors
   - Distance Covariance
   - Distribution-free Testing

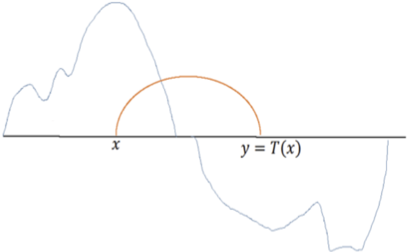
Optimal transport: Monge's problem

Gaspard Monge (1781): What is the cheapest way to transport a pile of sand to cover a sinkhole?



Optimal transport: Monge's problem

Gaspard Monge (1781): What is the cheapest way to transport a pile of sand to cover a sinkhole?



Goal: $$\inf_{T:T(X)\sim\mu} \mathbb{E}_\nu[c(X, T(X))] \qquad X \sim \nu$$

- ν ("data" dist.) and μ ("reference" dist.)
- $c(x, y) \geq 0$: cost of transporting x to y (e.g., $c(x, y) = \|x - y\|^2$)
- T transports ν to μ: $T\#\nu = \mu$ (i.e., $T(X) \sim \mu$ where $X \sim \nu$)

Rank function as the optimal transport (OT) map: when $d = 1$

- $X \sim \nu$ (continuous dist.) on $\mathbb{R}$, $F \equiv F_\nu$ c.d.f. of ν
- **Rank**: The population rank of $x \in \mathbb{R}$ is $F(x)$ (a.k.a. the c.d.f. at x)
- **Property**: $F(X) \sim \text{Uniform}([0, 1]) \equiv \mu$; i.e., F transports ν to μ

Rank function as the optimal transport (OT) map: when $d = 1$

- $X \sim \nu$ (continuous dist.) on $\mathbb{R}$, $F \equiv F_\nu$ c.d.f. of ν
- **Rank**: The population rank of $x \in \mathbb{R}$ is $F(x)$ (a.k.a. the c.d.f. at x)
- **Property**: $F(X) \sim \text{Uniform}([0,1]) \equiv \mu$; i.e., F transports ν to μ
- If $\mathbb{E}_\nu[X^2] < \infty$, the c.d.f. F is the optimal transport (OT) map as

$$F = \operatorname*{arg\,min}_{T : T\#\nu = \mu} \mathbb{E}_\nu[(X - T(X))^2]$$

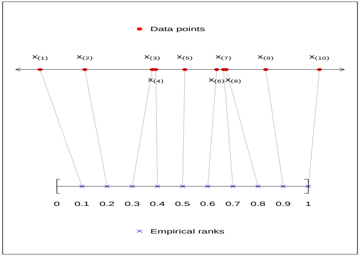
where

$$c(x, y) = (x - y)^2$$

Sample rank map: when $d = 1$

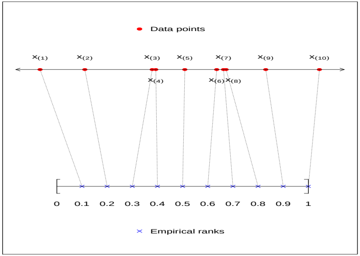
- **Data**: $X_1, \ldots, X_n$ iid ν (cont. distribution) on $\mathbb{R}$

- **Sample rank map**: $\hat{R}_n : \{X_1, X_2, \ldots, X_n\} \longrightarrow \{\frac{1}{n}, \frac{2}{n}, \ldots, \frac{n}{n}\}$



Sample rank map: when $d = 1$

- **Data**: $X_1, \ldots, X_n$ iid ν (cont. distribution) on $\mathbb{R}$
- **Sample rank map**: $\hat{R}_n : \{X_1, X_2, \ldots, X_n\} \longrightarrow \{\frac{1}{n}, \frac{2}{n}, \ldots, \frac{n}{n}\}$



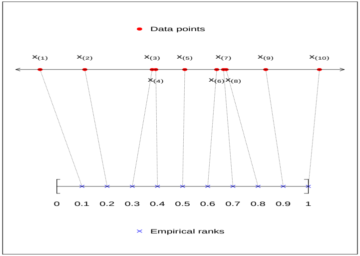
Sample rank map $\hat{R}_n$ is the OT map that transports

$$\nu_n := \frac{1}{n}\sum_{i=1}^{n} \delta_{X_i} \qquad \text{to} \qquad \mu_n := \frac{1}{n}\sum_{j=1}^{n} \delta_{\frac{j}{n}},$$

i.e., $$\hat{R}_n := \underset{T:T\#\nu_n=\mu_n}{\arg\min} \ \frac{1}{n}\sum_{i=1}^{n} |X_i - T(X_i)|^2$$

Sample rank map: when $d = 1$

- **Data**: $X_1, \ldots, X_n$ iid ν (cont. distribution) on $\mathbb{R}$
- **Sample rank map**: $\hat{R}_n : \{X_1, X_2, \ldots, X_n\} \longrightarrow \{\frac{1}{n}, \frac{2}{n}, \ldots, \frac{n}{n}\}$



Sample rank map $\hat{R}_n$ is the OT map that transports

$$\nu_n := \frac{1}{n}\sum_{i=1}^{n} \delta_{X_i} \qquad \text{to} \qquad \mu_n := \frac{1}{n}\sum_{j=1}^{n} \delta_{\frac{j}{n}},$$

i.e., $$\hat{R}_n := \operatorname*{arg\,min}_{T:T\#\nu_n=\mu_n} \frac{1}{n}\sum_{i=1}^{n} |X_i - T(X_i)|^2 = \operatorname*{arg\,max}_{T:T\#\nu_n=\mu_n} \frac{1}{n}\sum_{i=1}^{n} X_{(i)} \cdot T(X_{(i)})$$

- $X \sim \nu$; ν is a probability measure in $\mathbb{R}^d$ (abs. cont.)
- **Reference dist.**: $U \sim \mu$ on $\mathcal{S} \subset \mathbb{R}^d$ ($\mu = \text{Unif}([0,1]^d)$, $N(0, I_d)$)
- Find OT map T s.t. $T(X) \stackrel{d}{=} U \sim \mu$ (μ abs. cont.)

- $X \sim \nu$; ν is a probability measure in $\mathbb{R}^d$ (abs. cont.)
- **Reference dist.**: $U \sim \mu$ on $\mathcal{S} \subset \mathbb{R}^d$ ($\mu = \text{Unif}([0,1]^d)$, $N(0, I_d)$)
- Find OT map T s.t. $T(X) \overset{d}{=} U \sim \mu$ (μ abs. cont.)

Population rank function (a.k.a OT map) [Chernozhukov et al. (2017)]

If $\mathbb{E}_\nu \|X\|^2 < \infty$, rank function $R : \mathbb{R}^d \to \mathcal{S}$ is the transport map s.t.

$$R := \underset{T: T\#\nu=\mu}{\arg\min}\ \mathbb{E}_\nu \|X - T(X)\|^2$$

- $X \sim \nu$; ν is a probability measure in $\mathbb{R}^d$ (abs. cont.)
- **Reference dist.**: $U \sim \mu$ on $\mathcal{S} \subset \mathbb{R}^d$ ($\mu =$ Unif$([0,1]^d)$, $N(0, I_d)$)
- Find OT map T s.t. $T(X) \stackrel{d}{=} U \sim \mu$ (μ abs. cont.)

Population rank function (a.k.a OT map) [Chernozhukov et al. (2017)]

If $\mathbb{E}_\nu \|X\|^2 < \infty$, rank function $R : \mathbb{R}^d \to \mathcal{S}$ is the transport map s.t.

$$R := \underset{T : T \# \nu = \mu}{\arg\min}\ \mathbb{E}_\nu \|X - T(X)\|^2$$

Properties of population rank function [Brenier (1991), McCann (1995)]

- $R(\cdot)$ characterizes distribution: $R_1(x) = R_2(x)\ \forall\ x \in \mathbb{R}^d$ iff $P_1 = P_2$

- $X \sim \nu$; ν is a probability measure in $\mathbb{R}^d$ (abs. cont.)
- **Reference dist.**: $U \sim \mu$ on $\mathcal{S} \subset \mathbb{R}^d$ ($\mu =$ Unif$([0,1]^d)$, $N(0, I_d)$)
- Find OT map T s.t. $T(X) \stackrel{d}{=} U \sim \mu$ (μ abs. cont.)

Population rank function (a.k.a OT map) [Chernozhukov et al. (2017)]

If $\mathbb{E}_\nu \|X\|^2 < \infty$, rank function $R : \mathbb{R}^d \to \mathcal{S}$ is the transport map s.t.

$$R := \underset{T : T \# \nu = \mu}{\arg\min}\ \mathbb{E}_\nu \|X - T(X)\|^2$$

Properties of population rank function [Brenier (1991), McCann (1995)]

- $R(\cdot)$ characterizes distribution: $R_1(x) = R_2(x)\ \forall\ x \in \mathbb{R}^d$ iff $P_1 = P_2$
- $R(\cdot)$ is invertible, i.e., there exists unique $Q(\cdot)$ s.t.

$$R \circ Q(u) = u \ \ (\mu\text{-a.e.}) \qquad \text{and} \qquad Q \circ R(x) = x \ \ (\nu\text{-a.e.})$$

- $X \sim \nu$; ν is a probability measure in $\mathbb{R}^d$ (abs. cont.)
- **Reference dist.**: $U \sim \mu$ on $\mathcal{S} \subset \mathbb{R}^d$ ($\mu = \text{Unif}([0,1]^d)$, $N(0, I_d)$)
- Find OT map T s.t. $T(X) \overset{d}{=} U \sim \mu$ (μ abs. cont.)

Population rank function (a.k.a OT map) [Chernozhukov et al. (2017)]

If $\mathbb{E}_\nu \|X\|^2 < \infty$, rank function $R : \mathbb{R}^d \to \mathcal{S}$ is the transport map s.t.

$$R := \underset{T: T\#\nu=\mu}{\arg\min}\; \mathbb{E}_\nu \|X - T(X)\|^2$$

Properties of population rank function [Brenier (1991), McCann (1995)]

- $R(\cdot)$ characterizes distribution: $R_1(x) = R_2(x)\ \forall\, x \in \mathbb{R}^d$ iff $P_1 = P_2$
- $R(\cdot)$ is invertible, i.e., there exists unique $Q(\cdot)$ s.t.

$$R \circ Q(u) = u \ \ (\mu\text{-a.e.}) \qquad \text{and} \qquad Q \circ R(x) = x \ \ (\nu\text{-a.e.})$$

- Both $R(\cdot)$ and $Q(\cdot)$ and gradients of convex functions

- If $\mathbb{E}_\nu \|X\|^2 < \infty$, the population rank function $R(\cdot)$ is defined as

$$R := \operatorname*{arg\,min}_{T : T\#\nu = \mu} \mathbb{E}_\nu \|X - T(X)\|^2 \tag{1}$$

- Even when $\mathbb{E}_\nu \|X\|^2 = +\infty$, $R(\cdot)$ can still be defined

- If $\mathbb{E}_\nu \|X\|^2 < \infty$, the population rank function $R(\cdot)$ is defined as

$$R := \underset{T: T\#\nu=\mu}{\arg\min}\ \mathbb{E}_\nu \|X - T(X)\|^2 \tag{1}$$

- Even when $\mathbb{E}_\nu \|X\|^2 = +\infty$, $R(\cdot)$ can still be defined

Characterization of the population rank function [McCann (1995)]

Suppose $X \sim \nu$ abs. cont. on $\mathbb{R}^d$. Then $\exists$ ν-a.e. unique meas. mapping $R : \mathbb{R}^d \to \mathcal{S}$, transporting ν to μ (i.e., $R\#\nu = \mu$), of the form

$$R(x) = \nabla\varphi(x), \qquad \text{for } \nu\text{-a.e. } x, \tag{2}$$

where $\varphi : \mathbb{R}^d \to \mathbb{R} \cup \{+\infty\}$ is a convex function (cf. when $d = 1$).

- If $\mathbb{E}_\nu\|X\|^2 < \infty$, the population rank function $R(\cdot)$ is defined as

$$R := \underset{T:T\#\nu=\mu}{\arg\min}\ \mathbb{E}_\nu\|X - T(X)\|^2 \tag{1}$$

- Even when $\mathbb{E}_\nu\|X\|^2 = +\infty$, $R(\cdot)$ can still be defined

Characterization of the population rank function [McCann (1995)]

Suppose $X \sim \nu$ abs. cont. on $\mathbb{R}^d$. Then $\exists$ ν-a.e. unique meas. mapping $R : \mathbb{R}^d \to \mathcal{S}$, transporting ν to μ (i.e., $R\#\nu = \mu$), of the form

$$R(x) = \nabla\varphi(x), \qquad \text{for } \nu\text{-a.e. } x, \tag{2}$$

where $\varphi : \mathbb{R}^d \to \mathbb{R} \cup \{+\infty\}$ is a convex function (cf. when $d = 1$).

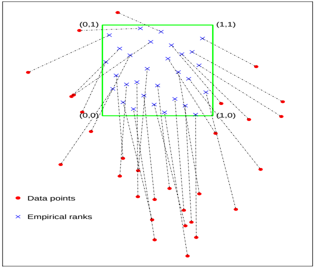
Moreover, when $\mathbb{E}_\nu\|X\|^2 < \infty$, $R(\cdot)$ as defined in (2) also satisfies (1).

- **Data**: $X_1, \ldots, X_n$ iid ν on $\mathbb{R}^d$ (abs. cont.); $\mu \sim$ Unif($[0,1]^d$)

- Empirical rank map $\hat{R}_n$: $\{X_1, \ldots, X_n\} \to \{c_1, \ldots, c_n\} \subset [0,1]^d$ — sequence of "uniform-like" points (or quasi-Monte Carlo sequence)

- **Data**: $X_1, \ldots, X_n$ iid ν on $\mathbb{R}^d$ (abs. cont.); $\mu \sim \mathrm{Unif}([0,1]^d)$

- Empirical rank map $\hat{R}_n$: $\{X_1, \ldots, X_n\} \to \{c_1, \ldots, c_n\} \subset [0,1]^d$ — sequence of "uniform-like" points (or quasi-Monte Carlo sequence)



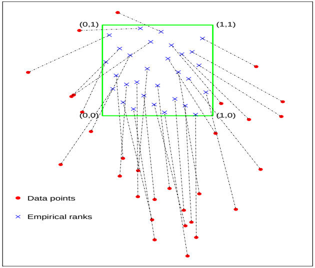
- Sample multivariate rank map is defined as the OT map s.t.

$$\hat{R}_n := \underset{T : T\#\nu_n = \mu_n}{\arg\min} \frac{1}{n} \sum_{i=1}^{n} \|X_i - T(X_i)\|^2$$

where T transports $\nu_n := \frac{1}{n}\sum_{i=1}^{n} \delta_{X_i}$ to $\mu_n := \frac{1}{n}\sum_{j=1}^{n} \delta_{c_j}$

- **Data**: $X_1, \ldots, X_n$ iid ν on $\mathbb{R}^d$ (abs. cont.); $\mu \sim \text{Unif}([0,1]^d)$

- Empirical rank map $\hat{R}_n$: $\{X_1, \ldots, X_n\} \to \{c_1, \ldots, c_n\} \subset [0,1]^d$ — sequence of "uniform-like" points (or quasi-Monte Carlo sequence)

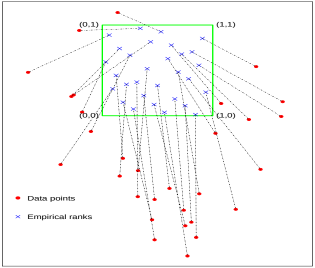


- Sample multivariate rank map is defined as the OT map s.t.

$$\hat{R}_n := \underset{T:T\#\nu_n=\mu_n}{\arg\min} \frac{1}{n}\sum_{i=1}^{n} \|X_i - T(X_i)\|^2$$

where T transports $\nu_n := \frac{1}{n}\sum_{i=1}^{n}\delta_{X_i}$ to $\mu_n := \frac{1}{n}\sum_{j=1}^{n}\delta_{c_j}$

Computation: Assignment problem



$$\hat{R}_n := \arg\min_{T:T\#\nu_n=\mu_n} \frac{1}{n}\sum_{i=1}^{n} \|X_i - T(X_i)\|^2$$

- Assignment problem (can be reduced to a linear program; the Hungarian algorithm has worst case time complexity $O(n^3)$)
- Various near linear time approximation algorithms exist for this problem — Drake & Hougardya (2005), Agarwal & Varadarajan (2004), Sharathkumar & Agarwal (2012), Agarwal et al. (2022)

Distribution-free property [Hallin (2017), Deb and S. (2019)]

Suppose that $X_1, \ldots, X_n$ iid on $\mathbb{R}^d$ with abs. cont. distribution. Then,

$$(\hat{R}_n(X_1), \ldots, \hat{R}_n(X_n))$$

is uniformly distributed over the $n!$ permutations of $\{c_1, \ldots, c_n\}$.

Distribution-free property [Hallin (2017), Deb and S. (2019)]

Suppose that $X_1, \ldots, X_n$ iid on $\mathbb{R}^d$ with abs. cont. distribution. Then,

$$(\hat{R}_n(X_1), \ldots, \hat{R}_n(X_n))$$

is uniformly distributed over the $n!$ permutations of $\{c_1, \ldots, c_n\}$.

The first step to obtaining distribution-free tests [Hallin et al. (2021)]

Distribution-free property [Hallin (2017), Deb and S. (2019)]

Suppose that $X_1, \ldots, X_n$ iid on $\mathbb{R}^d$ with abs. cont. distribution. Then,

$$(\hat{R}_n(X_1), \ldots, \hat{R}_n(X_n))$$

is uniformly distributed over the $n!$ permutations of $\{c_1, \ldots, c_n\}$.

The first step to obtaining distribution-free tests [Hallin et al. (2021)]

Consistency [Deb and S. (2019), Deb, Bhattacharya and S. (2021)]

$X_1, \ldots, X_n$ iid ν (abs. cont.). If $\mu_n := \frac{1}{n}\sum_{j=1}^n \delta_{c_j} \xrightarrow{d} \mu$ (abs. cont.), then

$$\frac{1}{n}\sum_{i=1}^{n} \|\hat{R}_n(X_i) - R(X_i)\|^2 \xrightarrow{p} 0 \qquad \text{as} \quad n \to \infty.$$

Regularity to the empirical multivariate rank/OT map

Question: What is the rate of convergence of $\hat{R}_n$?

Assume $\int \|x\|^2 \, d\nu(x) < \infty$, $\int \|y\|^2 \, d\mu(y) < \infty$; $\quad R\#\nu = \mu$, $\hat{R}_n\#\nu_n = \mu_n$

Rate of convergence [Deb, Ghosal and S. (2021)]

Proof of this result

Suppose the population rank map $R(\cdot)$ is Lipschitz. Then, under appropriate conditions on μ_n,

$$\mathbb{E}\left[\frac{1}{n}\sum_{i=1}^{n}\|\hat{R}_n(X_i) - R(X_i)\|^2\right] \lesssim \begin{cases} n^{-1/2} & d = 2,3, \\ n^{-1/2}\log n & d = 4, \\ n^{-2/d} & d > 4. \end{cases}$$

Question: What is the rate of convergence of $\hat{R}_n$?

Assume $\int \|x\|^2 \, d\nu(x) < \infty$, $\int \|y\|^2 \, d\mu(y) < \infty$; $\quad R\#\nu = \mu$, $\hat{R}_n \# \nu_n = \mu_n$

Rate of convergence [Deb, Ghosal and S. (2021)] Proof of this result

Suppose the population rank map $R(\cdot)$ is Lipschitz. Then, under appropriate conditions on μ_n,

$$\mathbb{E}\left[\frac{1}{n}\sum_{i=1}^{n}\|\hat{R}_n(X_i) - R(X_i)\|^2\right] \lesssim \begin{cases} n^{-1/2} & d = 2, 3, \\ n^{-1/2}\log n & d = 4, \\ n^{-2/d} & d > 4. \end{cases}$$

This is the optimal rate for $d \geq 4$ [Hütter & Rigollet (2019)]

Question: What is the rate of convergence of $\hat{R}_n$?

Assume $\int \|x\|^2 \, d\nu(x) < \infty$, $\int \|y\|^2 \, d\mu(y) < \infty$; $\quad R\#\nu = \mu$, $\hat{R}_n \# \nu_n = \mu_n$

Rate of convergence [Deb, Ghosal and S. (2021)] Proof of this result

Suppose the population rank map $R(\cdot)$ is Lipschitz. Then, under appropriate conditions on μ_n,

$$\mathbb{E}\left[\frac{1}{n}\sum_{i=1}^{n}\|\hat{R}_n(X_i) - R(X_i)\|^2\right] \lesssim \begin{cases} n^{-1/2} & d = 2, 3, \\ n^{-1/2}\log n & d = 4, \\ n^{-2/d} & d > 4. \end{cases}$$

This is the optimal rate for $d \geq 4$ [Hütter & Rigollet (2019)]

Estimation of the OT map R ($R\#\nu = \mu$) Barycentric Projection

- When $\{X_i\}_{i=1}^n$ and $\{c_j\}_{j=1}^m$ may have unequal sample sizes, R can be estimated using the barycentric projection $\tilde{R}$ (of the optimal coupling in the 2-Wasserstein distance between $\{X_i\}$ and $\{c_j\}$)
- Under additional smoothness assumptions, $\tilde{R}$ can have faster rates (by smoothing ν_n and μ_n)

Outline

Multivariate two-sample goodness-of-fit test

Testing for equality of two multivariate distributions

- **Data**: $\{X_i\}_{i=1}^m$ iid P_1 on $\mathbb{R}^d$; $\{Y_j\}_{j=1}^n$ iid P_2 on $\mathbb{R}^d$, $d \geq 1$

- Test if the two samples came from the same distribution, i.e.,

$$\mathrm{H}_0 : P_1 = P_2 \qquad \text{versus} \qquad \mathrm{H}_1 : P_1 \neq P_2$$

Multivariate two-sample goodness-of-fit test

Testing for equality of two multivariate distributions

- **Data**: $\{X_i\}_{i=1}^m$ iid P_1 on $\mathbb{R}^d$; $\{Y_j\}_{j=1}^n$ iid P_2 on $\mathbb{R}^d$, $d \geq 1$

- Test if the two samples came from the same distribution, i.e.,

$$\mathrm{H}_0 : P_1 = P_2 \qquad \text{versus} \qquad \mathrm{H}_1 : P_1 \neq P_2$$

- **Hotelling T^2 statistic** [Hotelling (1931)]: The multivariate analogue of Student's t-statistic, given by

$$\mathrm{T}^2_{m,n} := \frac{mn}{m+n} \left(\bar{X} - \bar{Y}\right)^\top S_{m,n}^{-1} \left(\bar{X} - \bar{Y}\right);$$

  where $S_{m,n}$ is pooled covariance matrix

- Reject H_0 iff $\mathrm{T}^2_{m,n} > c_\alpha$ [asymp. cut-off c_α: $(1-\alpha)$ quantile of χ^2_d]

Kernel two-sample test [Gretton et al. (2012)]

- The maximum mean discrepancy (MMD) btw. P_1 and P_2:

  $$\mathrm{MMD}^2(P_1, P_2) := \mathbb{E}[K(X, X')] + \mathbb{E}[K(Y, Y')] - 2\,\mathbb{E}[K(X, Y)] \geq 0,$$

  $K : \mathbb{R}^d \times \mathbb{R}^d$ is a kernel function[2]; $\quad X, X' \overset{iid}{\sim} P_1; \;\; Y, Y' \overset{iid}{\sim} P_2$

- $\mathrm{MMD}^2(P_1, P_2) = 0 \quad$ iff $\quad P_1 = P_2$ (if K is characteristic)

---

[2] Gaussian kernel: $K(x, y) = \exp(-\|x - y\|^2)$; Distance kernel: $K(x, y) = \frac{1}{2}\{\|x\| + \|y\| - \|x - y\|\}$

Kernel two-sample test [Gretton et al. (2012)]

- The maximum mean discrepancy (MMD) btw. P_1 and P_2:

  $$\mathrm{MMD}^2(P_1, P_2) := \mathbb{E}[K(X, X')] + \mathbb{E}[K(Y, Y')] - 2\,\mathbb{E}[K(X, Y)] \geq 0,$$

  $K : \mathbb{R}^d \times \mathbb{R}^d$ is a kernel function[2]; $X, X' \overset{iid}{\sim} P_1$; $Y, Y' \overset{iid}{\sim} P_2$

- $\mathrm{MMD}^2(P_1, P_2) = 0$ iff $P_1 = P_2$ (if K is characteristic)

- **Estimator**: $\mathrm{MMD}^2_{m,n}\left(\{X_i\}_{i=1}^m, \{Y_j\}_{j=1}^n\right) := A + B - 2C$ where

$$A := \frac{1}{m^2}\sum_{i,j=1}^{m} K(X_i, X_j), \quad B := \frac{1}{n^2}\sum_{i,j=1}^{n} K(Y_i, Y_j), \quad C := \frac{1}{mn}\sum_{i,j=1}^{m,n} K(X_i, Y_j)$$

[2] Gaussian kernel: $K(x, y) = \exp(-\|x - y\|^2)$; Distance kernel: $K(x, y) = \frac{1}{2}\{\|x\| + \|y\| - \|x - y\|\}$

Kernel two-sample test [Gretton et al. (2012)]

- The maximum mean discrepancy (MMD) btw. P_1 and P_2:

$$\mathrm{MMD}^2(P_1, P_2) := \mathbb{E}[K(X, X')] + \mathbb{E}[K(Y, Y')] - 2\,\mathbb{E}[K(X, Y)] \geq 0,$$

$K : \mathbb{R}^d \times \mathbb{R}^d$ is a kernel function[2]; $\quad X, X' \overset{iid}{\sim} P_1; \quad Y, Y' \overset{iid}{\sim} P_2$

- $\mathrm{MMD}^2(P_1, P_2) = 0 \quad$ iff $\quad P_1 = P_2$ (if K is characteristic)

- **Estimator**: $\mathrm{MMD}^2_{m,n}\left(\{X_i\}_{i=1}^m, \{Y_j\}_{j=1}^n\right) := A + B - 2C$ where

$$A := \frac{1}{m^2} \sum_{i,j=1}^{m} K(X_i, X_j), \quad B := \frac{1}{n^2} \sum_{i,j=1}^{n} K(Y_i, Y_j), \quad C := \frac{1}{mn} \sum_{i,j=1}^{m,n} K(X_i, Y_j)$$

- Reject $\mathrm{H}_0 : P_1 = P_2$ iff $\qquad \mathrm{MMD}^2_{m,n} > \kappa_\alpha$

- Critical value κ_α depends on $P_1 = P_2$! (but can be by-passed by using a permutation test)

[2] Gaussian kernel: $K(x, y) = \exp(-\|x - y\|^2)$; $\qquad$ Distance kernel: $K(x, y) = \frac{1}{2}\{\|x\| + \|y\| - \|x - y\|\}$

Data: $\{X_i\}_{i=1}^m$ iid P_1 (abs. cont.), $\{Y_j\}_{j=1}^n$ iid P_2 on $\mathbb{R}^d$, $d \geq 1$
Reference dist.: μ on $\mathcal{S} \subset \mathbb{R}^d$ (abs. cont.; e.g., $\mu = \mathrm{Unif}([0,1]^d)$)

Proposed tests [Deb and S. (2019), Deb, Bhattacharya and S. (2021)]

- Joint rank map: The sample ranks of the pooled observations:

$$\hat{R}_{m,n} : \{X_1, \ldots, X_m, Y_1, \ldots, Y_n\} \to \{c_1, \ldots, c_{m+n}\} \subset \mathcal{S}$$

- **Rank Hotelling**: $\mathrm{RT}^2_{m,n} := \mathrm{T}^2_{m,n}\left(\{\hat{R}_{m,n}(X_i)\}, \{\hat{R}_{m,n}(Y_j)\}\right)$
- **Rank MMD**: $\mathrm{RMMD}^2_{m,n} := \mathrm{MMD}^2_{m,n}\left(\{\hat{R}_{m,n}(X_i)\}, \{\hat{R}_{m,n}(Y_j)\}\right)$

Data: $\{X_i\}_{i=1}^m$ iid P_1 (abs. cont.), $\{Y_j\}_{j=1}^n$ iid P_2 on $\mathbb{R}^d$, $d \geq 1$
Reference dist.: μ on $\mathcal{S} \subset \mathbb{R}^d$ (abs. cont.; e.g., $\mu = \text{Unif}([0,1]^d)$)

Proposed tests [Deb and S. (2019), Deb, Bhattacharya and S. (2021)]

- Joint rank map: The sample ranks of the pooled observations:

$$\hat{R}_{m,n} : \{X_1, \ldots, X_m, Y_1, \ldots, Y_n\} \to \{c_1, \ldots, c_{m+n}\} \subset \mathcal{S}$$

- **Rank Hotelling**: $\mathrm{RT}^2_{m,n} := \mathrm{T}^2_{m,n}\left(\{\hat{R}_{m,n}(X_i)\}, \{\hat{R}_{m,n}(Y_j)\}\right)$
- **Rank MMD**: $\mathrm{RMMD}^2_{m,n} := \mathrm{MMD}^2_{m,n}\left(\{\hat{R}_{m,n}(X_i)\}, \{\hat{R}_{m,n}(Y_j)\}\right)$

- In general, our principle is to start with a "good" test and replace the X_i's and Y_j's with their pooled multivariate ranks

- This yields the Wilcoxon rank-sum test when applied to the t-test

Distribution-freeness [Deb and S. (2019)]

Under H_0, distributions of $\mathrm{RT}^2_{m,n}, \mathrm{RMMD}^2_{m,n}$ are free of $P_1 \equiv P_2$

Rank Hotelling test [Deb, Bhattacharya, and S. (2021)]

Rank Hotelling test: $\phi_{m,n} \equiv \mathbf{1}\{\mathrm{RT}^2_{m,n} > \kappa_\alpha^{(m,n)}\}$ — distribution-free

$\kappa_\alpha^{(m,n)}$ depends on c_j's, m, n and d

Rank Hotelling test [Deb, Bhattacharya, and S. (2021)]

Rank Hotelling test: $\phi_{m,n} \equiv \mathbf{1}\{\mathrm{RT}^2_{m,n} > \kappa_\alpha^{(m,n)}\}$ — distribution-free

$\kappa_\alpha^{(m,n)}$ depends on c_j's, m, n and d

Asymptotic null distribution (Deb, Bhattacharya, and S., 2021)

Under H_0, if $\mu_n := \frac{1}{n}\sum_{j=1}^{n} \delta_{c_j} \xrightarrow{d} \mu$, then,

$$\mathrm{RT}^2_{m,n} \xrightarrow{d} \chi^2_d \qquad \text{as} \quad \min\{m,n\} \to \infty.$$

The choice of the c_j's have no effect for large m, n

Rank Hotelling test [Deb, Bhattacharya, and S. (2021)]

Rank Hotelling test: $\phi_{m,n} \equiv \mathbf{1}\{\mathrm{RT}^2_{m,n} > \kappa_\alpha^{(m,n)}\}$ — distribution-free

$\kappa_\alpha^{(m,n)}$ depends on c_j's, m, n and d

Asymptotic null distribution (Deb, Bhattacharya, and S., 2021)

Under H_0, if $\mu_n := \frac{1}{n}\sum_{j=1}^{n} \delta_{c_j} \xrightarrow{d} \mu$, then,

$$\mathrm{RT}^2_{m,n} \xrightarrow{d} \chi^2_d \qquad \text{as} \quad \min\{m, n\} \to \infty.$$

The choice of the c_j's have no effect for large m, n

Power (Deb, Bhattacharya, and S., 2021)

Under location shift alternatives ($P_1 \neq P_2$), if (i) $\mu_n \xrightarrow{d} \mu$, and (ii) $\frac{m}{m+n} \to \lambda \in (0,1)$, then,

$$\lim_{m,n\to\infty} \mathbb{E}_{\mathrm{H}_1}[\phi_{m,n}] = 1.$$

Rank Hotelling test [Deb, Bhattacharya, and S. (2021)]

Rank Hotelling test: $\phi_{m,n} \equiv \mathbf{1}\{\mathrm{RT}^2_{m,n} > \kappa_\alpha^{(m,n)}\}$ — distribution-free

$\kappa_\alpha^{(m,n)}$ depends on c_j's, m, n and d

Asymptotic null distribution (Deb, Bhattacharya, and S., 2021)

Under H_0, if $\mu_n := \frac{1}{n}\sum_{j=1}^n \delta_{c_j} \xrightarrow{d} \mu$, then,

$$\mathrm{RT}^2_{m,n} \xrightarrow{d} \chi^2_d \qquad \text{as} \quad \min\{m, n\} \to \infty.$$

The choice of the c_j's have no effect for large m, n

Power (Deb, Bhattacharya, and S., 2021)

Under location shift alternatives ($P_1 \neq P_2$), if (i) $\mu_n \xrightarrow{d} \mu$, and (ii) $\frac{m}{m+n} \to \lambda \in (0, 1)$, then,

$$\lim_{m,n\to\infty} \mathbb{E}_{\mathrm{H}_1}[\phi_{m,n}] = 1.$$

Question: How does rank Hotelling $\mathrm{RT}^2_{m,n}$ compare with Hotelling $\mathrm{T}^2_{m,n}$?

Rank MMD test [Deb and S. (2019)]

- **Rank MMD**: $\mathrm{RMMD}^2_{m,n} = \mathrm{MMD}^2_{m,n}\left(\{\hat{R}_{m,n}(X_i)\}, \{\hat{R}_{m,n}(Y_j)\}\right)$
- **Rank MMD test**: Reject H_0 iff $\mathrm{RMMD}^2_{m,n} > \kappa_\alpha^{(m,n)}$;
  $\kappa_\alpha^{(m,n)}$ is a universal threshold (free of $P_1 \equiv P_2$)
- Dist. of $\mathrm{RMMD}^2_{m,n}$ (under H_0) just depends on c_j's, m, n and d

Rank MMD test [Deb and S. (2019)]

- **Rank MMD**: $\mathrm{RMMD}^2_{m,n} = \mathrm{MMD}^2_{m,n}\left(\{\hat{R}_{m,n}(X_i)\}, \{\hat{R}_{m,n}(Y_j)\}\right)$
- **Rank MMD test**: Reject H_0 iff $\mathrm{RMMD}^2_{m,n} > \kappa_\alpha^{(m,n)}$;
  $\kappa_\alpha^{(m,n)}$ is a universal threshold (free of $P_1 \equiv P_2$)
- Dist. of $\mathrm{RMMD}^2_{m,n}$ (under H_0) just depends on c_j's, m, n and d

Limiting distribution under $\mathrm{H}_0 : P_1 = P_2$ [Deb and S. (2019)]

If (i) $P_1 \equiv P_2$ is abs. cont., and (ii) $\mu_n := \frac{1}{n}\sum_{j=1}^n \delta_{c_j} \xrightarrow{d} \mu$,

then, under H_0, for universal $\{\lambda_j \geq 0 : j \geq 1\}$ and $\{Z_j\}_{j\geq 1}$ iid $N(0,1)$,

$$\frac{mn}{m+n}\mathrm{RMMD}^2_{m,n} \xrightarrow{d} \sum_{j=1}^{\infty} \lambda_j Z_j^2 \qquad \text{as } \min\{m,n\} \to \infty.$$

The choice of the c_j's has no effect for large m, n

Asymptotic stabilization of critical values

Critical values: $\kappa_\alpha^{(m,n)}$

	$n = 100$	300	500	700	900
$\alpha = 0.05$	0.39	0.40	0.39	0.40	0.40
$\alpha = 0.10$	0.36	0.36	0.36	0.36	0.36

Table: Thresholds for $\alpha = 0.05$, 0.1 & $m = n = 100, 300, 500, 700, 900$, $d = 2$.

	$n = 100$	300	500	700	900
$\alpha = 0.05$	1.37	1.38	1.38	1.38	1.38
$\alpha = 0.10$	1.34	1.35	1.35	1.35	1.35

Table: Thresholds for $\alpha = 0.05$, 0.1 & $m = n = 100, 300, 500, 700, 900$, $d = 8$.

Connection to the two-sample Cramér-von Mises statistic when $d = 1$

When $d = 1$, $\mathrm{RMMD}^2_{m,n}$ is equivalent to two-sample Cramér-von Mises statistic [Anderson (1962)] when distance kernel[a] is used [Székely (2003)]:

$$\mathrm{RMMD}^2_{m,n} = 2 \int \left\{ \mathbb{F}^X_m(t) - \mathbb{F}^Y_n(t) \right\}^2 d\mathbb{F}_{m+n}(t)$$

where $\mathbb{F}^X_n$, $\mathbb{F}^Y_n$, $\mathbb{F}_{m+n}$ are empirical cdf's of the X's, Y's, and pooled sample.

Connection to the two-sample Cramér-von Mises statistic when $d = 1$

When $d = 1$, $\mathrm{RMMD}^2_{m,n}$ is equivalent to two-sample Cramér-von Mises statistic [Anderson (1962)] when distance kernel[a] is used [Székely (2003)]:

$$\mathrm{RMMD}^2_{m,n} = 2 \int \left\{ \mathbb{F}^X_m(t) - \mathbb{F}^Y_n(t) \right\}^2 d\mathbb{F}_{m+n}(t)$$

where $\mathbb{F}^X_n$, $\mathbb{F}^Y_n$, $\mathbb{F}_{m+n}$ are empirical cdf's of the X's, Y's, and pooled sample.

[a] $K(x, y) = 2^{-1}(|x| + |y| - |x - y|)$

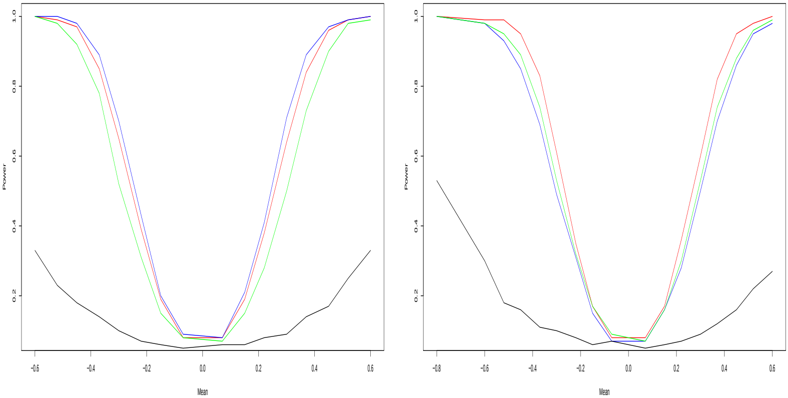
Power [Deb and S. (2019)]

Under $P_1 \neq P_2$, if (i) $\mu_n \xrightarrow{d} \mu$, and (ii) $\frac{m}{m+n} \to \lambda \in (0, 1)$, then,

$$\mathbb{P}\left(\mathrm{RMMD}_{m,n} > \kappa_\alpha^{(m,n)}\right) \to 1 \qquad \text{as } m, n \to \infty.$$

Proposed test has asymptotic power 1, against all fixed alternatives

Question: Can we quantify the power of these OT-based tests?



Left panel: $\begin{pmatrix} X_1 \\ X_2 \\ X_3 \end{pmatrix} \sim N_3(0, I_3)$; $\begin{pmatrix} Y_1 \\ Y_2 \\ Y_3 \end{pmatrix} \sim N_3(\mu 1_3, I_3)$ as $\mu \in \mathbb{R}$ varies

Right panel: $U = (U_1, U_2, U_3)$, $V = (V_1, V_2, V_3)$, $U_i = e^{X_i}$, $V_i = e^{Y_i}$

Performance of 4 tests: **Energy**, **Rank energy**, **Crossmatch**, **HHG**

More simulations

	(C)	(HHG)	(EN)	(REN)
V1	0.13	0.15	0.13	0.34
V2	0.34	0.94	0.94	0.89
V3	0.41	0.34	0.34	0.46
V4	0.34	0.31	0.33	0.32
V5	0.73	0.70	0.56	0.93
V6	0.90	0.88	0.82	0.99
V7	0.13	0.51	0.65	0.63
V8	0.11	0.39	0.35	0.43
V9	0.06	1.00	0.97	1.00
V10	0.28	0.99	1.00	0.59

Table: Proportion of times the null hypothesis was rejected across 10 settings. Here $n = 200$, $d = 3$. Here (C) – Rosenbaum's crossmatch test [Rosenbaum (2005)], (HHG) – Heller, Heller and Gorfine [Heller et al. (2013)], (EN) – energy statistic [Székely and Rizzo (2013)], (REN) – rank energy test.

1. Optimal Transport: Monge's Problem
   - Introduction
   - Multivariate Ranks via Optimal Transport
2. Multivariate Two-sample Goodness-of-fit Testing
   - Hotelling T^2 and Kernel MMD
   - Distribution-free Testing
   - Lower bounds on Asymptotic (Pitman) Relative Efficiency
3. Testing for Independence Between Two Random Vectors
   - Distance Covariance
   - Distribution-free Testing

- **Question**: How to compare two consistent tests S_N and T_N?
- **Asymptotic relative (Pitman) efficiency** (ARE) [Pitman (1948), Serfling (1980), Lehmann & Romano (2005), van der Vaart (1998)]

- **Question**: How to compare two consistent tests S_N and T_N?
- **Asymptotic relative (Pitman) efficiency** (ARE) [Pitman (1948), Serfling (1980), Lehmann & Romano (2005), van der Vaart (1998)]

- $X_1, \ldots, X_m \overset{iid}{\sim} \mathrm{P}_{\theta_1}$ & $Y_1, \ldots, Y_n \overset{iid}{\sim} \mathrm{P}_{\theta_2}$; $\quad N = m + n$; $\frac{m}{N} \approx \lambda \in (0, 1)$
- $\{P_\theta\}_{\theta \in \Theta \subset \mathbb{R}^p}$: "smooth" (satisfies DQM) parametric family

- **Question**: How to compare two consistent tests S_N and T_N?
- **Asymptotic relative (Pitman) efficiency** (ARE) [Pitman (1948), Serfling (1980), Lehmann & Romano (2005), van der Vaart (1998)]

- $X_1, \ldots, X_m \overset{iid}{\sim} \mathrm{P}_{\theta_1}$ & $Y_1, \ldots, Y_n \overset{iid}{\sim} \mathrm{P}_{\theta_2}$; $N = m + n$; $\frac{m}{N} \approx \lambda \in (0,1)$
- $\{P_\theta\}_{\theta \in \Theta \subset \mathbb{R}^p}$: "smooth" (satisfies DQM) parametric family

- **Test** $\mathrm{H}_0 : \theta_2 = \theta_1$ vs. $\mathrm{H}_1 : \theta_2 = \theta_1 + \Delta$; $\Delta \to 0$

- **Question**: How to compare two consistent tests S_N and T_N?
- **Asymptotic relative (Pitman) efficiency** (ARE) [Pitman (1948), Serfling (1980), Lehmann & Romano (2005), van der Vaart (1998)]

- $X_1, \ldots, X_m \overset{iid}{\sim} \mathrm{P}_{\theta_1}$ & $Y_1, \ldots, Y_n \overset{iid}{\sim} \mathrm{P}_{\theta_2}$; $\quad N = m + n$; $\frac{m}{N} \approx \lambda \in (0, 1)$
- $\{P_\theta\}_{\theta \in \Theta \subset \mathbb{R}^p}$: "smooth" (satisfies DQM) parametric family

- **Test** $\mathrm{H}_0 : \theta_2 = \theta_1 \quad$ vs. $\quad \mathrm{H}_1 : \theta_2 = \theta_1 + \Delta; \quad \Delta \to 0$
- Fix $\alpha \in (0, 1)$ (level) and $\beta \in (\alpha, 1)$ (power)

- **Question**: How to compare two consistent tests S_N and T_N?
- **Asymptotic relative (Pitman) efficiency** (ARE) [Pitman (1948), Serfling (1980), Lehmann & Romano (2005), van der Vaart (1998)]

- $X_1, \ldots, X_m \overset{iid}{\sim} \mathrm{P}_{\theta_1}$ & $Y_1, \ldots, Y_n \overset{iid}{\sim} \mathrm{P}_{\theta_2}$; $\quad N = m + n$; $\frac{m}{N} \approx \lambda \in (0,1)$
- $\{P_\theta\}_{\theta \in \Theta \subset \mathbb{R}^p}$: "smooth" (satisfies DQM) parametric family

- **Test** $\mathrm{H}_0 : \theta_2 = \theta_1 \quad$ vs. $\quad \mathrm{H}_1 : \theta_2 = \theta_1 + \Delta; \quad \Delta \to 0$
- Fix $\alpha \in (0,1)$ (level) and $\beta \in (\alpha, 1)$ (power)

- Let $N_\Delta(T_\cdot) \equiv N_\Delta$ denote the minimum number of samples s.t.:

$$\mathbb{E}_{\mathrm{H}_0}[T_{N_\Delta}] = \alpha \qquad \text{and} \qquad \mathbb{E}_{\mathrm{H}_1}[T_{N_\Delta}] \geq \beta$$

- **Question**: How to compare two consistent tests S_N and T_N?
- **Asymptotic relative (Pitman) efficiency** (ARE) [Pitman (1948), Serfling (1980), Lehmann & Romano (2005), van der Vaart (1998)]

- $X_1, \ldots, X_m \overset{iid}{\sim} \mathrm{P}_{\theta_1}$ & $Y_1, \ldots, Y_n \overset{iid}{\sim} \mathrm{P}_{\theta_2}$; $\quad N = m + n$; $\frac{m}{N} \approx \lambda \in (0, 1)$
- $\{P_\theta\}_{\theta \in \Theta \subset \mathbb{R}^p}$: "smooth" (satisfies DQM) parametric family

- **Test** $\mathrm{H}_0 : \theta_2 = \theta_1 \quad$ vs. $\quad \mathrm{H}_1 : \theta_2 = \theta_1 + \Delta; \quad \Delta \to 0$
- Fix $\alpha \in (0, 1)$ (level) and $\beta \in (\alpha, 1)$ (power)
- Let $N_\Delta(T_.) \equiv N_\Delta$ denote the minimum number of samples s.t.:

$$\mathbb{E}_{\mathrm{H}_0}[T_{N_\Delta}] = \alpha \qquad \text{and} \qquad \mathbb{E}_{\mathrm{H}_1}[T_{N_\Delta}] \geq \beta$$

- The asymptotic (Pitman) efficiency of S_N w.r.t. T_N is given by

$$\mathrm{ARE}(S_N, T_N) := \lim_{\Delta \to 0} \frac{N_\Delta(T_.)}{N_\Delta(S_.)}$$

- **Question**: How to compare two consistent tests S_N and T_N?
- **Asymptotic relative (Pitman) efficiency** (ARE) [Pitman (1948), Serfling (1980), Lehmann & Romano (2005), van der Vaart (1998)]

- $X_1, \ldots, X_m \overset{iid}{\sim} \mathrm{P}_{\theta_1}$ & $Y_1, \ldots, Y_n \overset{iid}{\sim} \mathrm{P}_{\theta_2}$; $N = m + n$; $\frac{m}{N} \approx \lambda \in (0,1)$
- $\{P_\theta\}_{\theta \in \Theta \subset \mathbb{R}^p}$: "smooth" (satisfies DQM) parametric family

- **Test** $\mathrm{H}_0 : \theta_2 = \theta_1$ vs. $\mathrm{H}_1 : \theta_2 = \theta_1 + \Delta$; $\Delta \to 0$
- Fix $\alpha \in (0,1)$ (level) and $\beta \in (\alpha, 1)$ (power)
- Let $N_\Delta(T_.) \equiv N_\Delta$ denote the minimum number of samples s.t.:

$$\mathbb{E}_{\mathrm{H}_0}[T_{N_\Delta}] = \alpha \qquad \text{and} \qquad \mathbb{E}_{\mathrm{H}_1}[T_{N_\Delta}] \geq \beta$$

- The asymptotic (Pitman) efficiency of S_N w.r.t. T_N is given by

$$\mathrm{ARE}\,(S_N, T_N) := \lim_{\Delta \to 0} \frac{N_\Delta(T_.)}{N_\Delta(S_.)}$$

$\mathrm{ARE}\,(S_N, T_N)$ can depend on α and β, but in some cases it doesn't!

Hotelling T^2: $\mathrm{T}^2_{m,n}(\{X_i\},\{Y_j\}) = \frac{mn}{m+n}\left(\bar{X}-\bar{Y}\right)^\top S^{-1}_{m,n}\left(\bar{X}-\bar{Y}\right)$

Rank Hotelling: $\mathrm{RT}^2_{m,n} = \mathrm{T}^2_{m,n}\left(\{\hat{R}_{m,n}(X_i)\},\{\hat{R}_{m,n}(Y_j)\}\right)$

- $X_1,\ldots,X_m \overset{iid}{\sim} \mathrm{P}_{\theta_1}$ & $Y_1,\ldots,Y_n \overset{iid}{\sim} \mathrm{P}_{\theta_2}$; $\quad N = m+n$
- $\{P_\theta\}_{\theta\in\Theta\subset\mathbb{R}^p}$: "smooth" (satisfies DQM) parametric family
- Consider $\mathrm{H}_0 : \theta_2 = \theta_1$ vs. $\mathrm{H}_1 : \theta_2 = \theta_1 + hN^{-1/2}$; $\quad h \neq 0 \in \mathbb{R}^p$

$\mathrm{ARE}\left(\mathrm{RT}^2_{m,n}, \mathrm{T}^2_{m,n}\right)$ can be derived under the above alternatives

Hotelling T^2: $\mathrm{T}^2_{m,n}(\{X_i\},\{Y_j\}) = \frac{mn}{m+n}\left(\bar{X}-\bar{Y}\right)^\top S^{-1}_{m,n}\left(\bar{X}-\bar{Y}\right)$

Rank Hotelling: $\mathrm{RT}^2_{m,n} = \mathrm{T}^2_{m,n}\left(\{\hat{R}_{m,n}(X_i)\},\{\hat{R}_{m,n}(Y_j)\}\right)$

- $X_1,\ldots,X_m \overset{iid}{\sim} \mathrm{P}_{\theta_1}$ & $Y_1,\ldots,Y_n \overset{iid}{\sim} \mathrm{P}_{\theta_2}$; $N = m+n$
- $\{P_\theta\}_{\theta\in\Theta\subset\mathbb{R}^p}$: "smooth" (satisfies DQM) parametric family
- Consider $\mathrm{H}_0 : \theta_2 = \theta_1$ vs. $\mathrm{H}_1 : \theta_2 = \theta_1 + hN^{-1/2}$; $h \neq 0 \in \mathbb{R}^p$

$\mathrm{ARE}\left(\mathrm{RT}^2_{m,n}, \mathrm{T}^2_{m,n}\right)$ can be derived under the above alternatives

Some observations

- Expression of $\mathrm{ARE}\left(\mathrm{RT}^2_{m,n}, \mathrm{T}^2_{m,n}\right)$ does not depend on α and β
- Asymp. dist. of $\mathrm{RT}^2_{m,n}$ can depend on choice of μ (reference dist.)

Hotelling T^2: $\mathrm{T}^2_{m,n}(\{X_i\},\{Y_j\}) = \frac{mn}{m+n}\left(\bar{X}-\bar{Y}\right)^\top S^{-1}_{m,n}\left(\bar{X}-\bar{Y}\right)$

Rank Hotelling: $\mathrm{RT}^2_{m,n} = \mathrm{T}^2_{m,n}\left(\{\hat{R}_{m,n}(X_i)\},\{\hat{R}_{m,n}(Y_j)\}\right)$

- $X_1,\dots,X_m \overset{iid}{\sim} \mathrm{P}_{\theta_1}$ & $Y_1,\dots,Y_n \overset{iid}{\sim} \mathrm{P}_{\theta_2}$; $N = m+n$
- $\{P_\theta\}_{\theta\in\Theta\subset\mathbb{R}^p}$: "smooth" (satisfies DQM) parametric family
- Consider $\mathrm{H}_0 : \theta_2 = \theta_1$ vs. $\mathrm{H}_1 : \theta_2 = \theta_1 + hN^{-1/2}$; $h \neq 0 \in \mathbb{R}^p$

$\mathrm{ARE}\left(\mathrm{RT}^2_{m,n}, \mathrm{T}^2_{m,n}\right)$ can be derived under the above alternatives

Some observations

- Expression of $\mathrm{ARE}\left(\mathrm{RT}^2_{m,n}, \mathrm{T}^2_{m,n}\right)$ does not depend on α and β
- Asymp. dist. of $\mathrm{RT}^2_{m,n}$ can depend on choice of μ (reference dist.)

Can we lower bound ARE for sub-classes of multivariate dists., i.e.,

$$\min_{\mathcal{F}} \mathrm{ARE}\left(\mathrm{RT}^2_{m,n}, \mathrm{T}^2_{m,n}\right) = ??$$

$X_1, \ldots, X_m \overset{iid}{\sim} \mathrm{P}_{\theta_1}$ & $Y_1, \ldots, Y_n \overset{iid}{\sim} \mathrm{P}_{\theta_2}$; $\quad N = m + n$

Independent coordinates case

$\mathcal{F}_{\mathrm{ind}} = \{P_\theta\}_{\theta \in \Theta}$ has density $p_\theta(z_1, \ldots, z_d) = \prod_{i=1}^d f_i(z_i - \theta_i)$, $\theta \in \mathbb{R}^d$

Theorem [Deb, Bhattacharya, and S. (2021)]

Suppose $\frac{m}{N} \to \lambda \in (0, 1)$. If $\mu_N := \frac{1}{N} \sum_{j=1}^N \delta_{c_j} \xrightarrow{d} \mathrm{Unif}([0,1]^d) \equiv \mu$, then

$$\min_{\mathcal{F}_{\mathrm{ind}}} \mathrm{ARE}\left(\mathrm{RT}^2_{m,n}, \mathrm{T}^2_{m,n}\right) = 0.864.$$

$X_1, \ldots, X_m \overset{iid}{\sim} \mathrm{P}_{\theta_1}$ & $Y_1, \ldots, Y_n \overset{iid}{\sim} \mathrm{P}_{\theta_2}$; $\quad N = m + n$

Independent coordinates case

$\mathcal{F}_{\mathrm{ind}} = \{P_\theta\}_{\theta \in \Theta}$ has density $p_\theta(z_1, \ldots, z_d) = \prod_{i=1}^d f_i(z_i - \theta_i)$, $\theta \in \mathbb{R}^d$

Theorem [Deb, Bhattacharya, and S. (2021)]

Suppose $\frac{m}{N} \to \lambda \in (0,1)$. If $\mu_N := \frac{1}{N} \sum_{j=1}^N \delta_{c_j} \xrightarrow{d} \mathrm{Unif}([0,1]^d) \equiv \mu$, then

$$\min_{\mathcal{F}_{\mathrm{ind}}} \mathrm{ARE}\left(\mathrm{RT}^2_{m,n}, \mathrm{T}^2_{m,n}\right) = 0.864.$$

If $\mu_N \xrightarrow{d} N(0, I_d) \equiv \mu$, then

$$\min_{\mathcal{F}_{\mathrm{ind}}} \mathrm{ARE}\left(\mathrm{RT}^2_{m,n}, \mathrm{T}^2_{m,n}\right) = 1.$$

$X_1, \ldots, X_m \overset{iid}{\sim} \mathrm{P}_{\theta_1}$ & $Y_1, \ldots, Y_n \overset{iid}{\sim} \mathrm{P}_{\theta_2}$; $\quad N = m + n$

Independent coordinates case

$\mathcal{F}_{\text{ind}} = \{P_\theta\}_{\theta \in \Theta}$ has density $p_\theta(z_1, \ldots, z_d) = \prod_{i=1}^d f_i(z_i - \theta_i)$, $\theta \in \mathbb{R}^d$

Theorem [Deb, Bhattacharya, and S. (2021)]

Suppose $\frac{m}{N} \to \lambda \in (0,1)$. If $\mu_N := \frac{1}{N} \sum_{j=1}^N \delta_{c_j} \xrightarrow{d} \mathrm{Unif}([0,1]^d) \equiv \mu$, then

$$\min_{\mathcal{F}_{\text{ind}}} \mathrm{ARE}\left(\mathrm{RT}^2_{m,n}, \mathrm{T}^2_{m,n}\right) = 0.864.$$

If $\mu_N \xrightarrow{d} N(0, I_d) \equiv \mu$, then

$$\min_{\mathcal{F}_{\text{ind}}} \mathrm{ARE}\left(\mathrm{RT}^2_{m,n}, \mathrm{T}^2_{m,n}\right) = 1.$$

- Generalizes Hodges & Lehmann (1956), Chernoff & Savage (1958)
- ARE can be arbitrarily large (can tend to $+\infty$) for heavy tailed dists.

Elliptically symmetric distributions

$\mathcal{F}_{\mathrm{ell}} = \{P_\theta\}_{\theta \in \Theta}$ is class of elliptically symmetric distributions on $\mathbb{R}^d$, i.e.,

$$p_\theta(x) \propto (\det(\Sigma))^{-\frac{1}{2}} \underline{f}\left((x-\theta)^\top \Sigma^{-1} (x-\theta)\right), \quad \text{for all } x \in \mathbb{R}^d$$

Elliptically symmetric distributions

$\mathcal{F}_{\mathrm{ell}} = \{P_\theta\}_{\theta \in \Theta}$ is class of elliptically symmetric distributions on $\mathbb{R}^d$, i.e.,

$$p_\theta(x) \propto (\det(\Sigma))^{-\frac{1}{2}} \underline{f}\left((x-\theta)^\top \Sigma^{-1}(x-\theta)\right), \quad \text{for all } x \in \mathbb{R}^d$$

Theorem [Deb, Bhattacharya, and S. (2021)]

Suppose: (i) $\mu_N \xrightarrow{d} N(0, I_d) \equiv \mu$, (ii) $\frac{m}{N} \to \lambda \in (0,1)$. Then,

$$\min_{\mathcal{F}_{\mathrm{ell}}} \mathrm{ARE}\left(\mathrm{RT}^2_{m,n}, \mathrm{T}^2_{m,n}\right) = 1.$$

- This generalizes the famous result of Chernoff and Savage (1958)

Model for Independent Component Analysis (ICA)

$\mathcal{F}_{\mathrm{ICA}} = \{f_1(\cdot - \theta) : f_1 \in \mathcal{F}\}_{\theta \in \mathbb{R}^d}$ where $f_1 \in \mathcal{F}$ has the form

$$f_1(x_1, \ldots, x_d) = \prod_{i=1}^{d} \tilde{f}_i \left(\sum_{j=1}^{d} a_{ji} x_j \right)$$

where $\tilde{f}_1, \tilde{f}_2, \ldots, \tilde{f}_d$ are univariate densities, and $A = (a_{ij})_{d \times d}$ is an orthogonal matrix (unknown)

Thus, f_1 is the density of $X_{d \times 1}$ where

$$X = A\,W$$

with $W_{d \times 1}$ having independent components.

Model for Independent Component Analysis (ICA)

$\mathcal{F}_{\mathrm{ICA}} = \{f_1(\cdot - \theta) : f_1 \in \mathcal{F}\}_{\theta \in \mathbb{R}^d}$ where $f_1 \in \mathcal{F}$ has the form

$$f_1(x_1, \ldots, x_d) = \prod_{i=1}^{d} \tilde{f}_i \left(\sum_{j=1}^{d} a_{ji} x_j \right)$$

where $\tilde{f}_1, \tilde{f}_2, \ldots, \tilde{f}_d$ are univariate densities, and $A = (a_{ij})_{d \times d}$ is an orthogonal matrix (unknown)

Thus, f_1 is the density of $X_{d \times 1}$ where

$$X = A\,W$$

with $W_{d \times 1}$ having independent components.

Theorem [Deb, Bhattacharya, and S. (2021)]

Suppose: (i) $\mu_N \xrightarrow{d} N(0, I_d) \equiv \mu$, (ii) $\frac{m}{N} \to \lambda \in (0,1)$. Then,

$$\min_{\mathcal{F}_{\mathrm{ICA}}} \mathrm{ARE}\left(\mathrm{RT}^2_{m,n}, \mathrm{T}^2_{m,n}\right) = 1.$$

Asymptotic efficiency of the Rank MMD test

Rank MMD: $\mathrm{RMMD}^2_{m,n} = \mathrm{MMD}^2_{m,n}\left(\{\hat{R}_{m,n}(X_i)\}, \{\hat{R}_{m,n}(Y_j)\}\right)$

Test: $\mathrm{H}_0 : \theta_2 = \theta_1$ vs. $\mathrm{H}_1 : \theta_2 = \theta_1 + hN^{-1/2}; h \neq 0 \in \mathbb{R}^p$

Theorem [Deb, Bhattacharya and S. (2021+)]

Under $\mathrm{H}_1 : \theta_2 = \theta_1 + hN^{-1/2}$,

$$\frac{mn}{N}\,\mathrm{RMMD}^2_{m,n} \xrightarrow{d} \sum_{j=1}^{\infty} \lambda_j \tilde{Z}_j^2$$

where $\tilde{Z}_j^2$ has non-central chi-squared distribution (depending on h).

Asymptotic efficiency of the Rank MMD test

Rank MMD: $\mathrm{RMMD}^2_{m,n} = \mathrm{MMD}^2_{m,n}\left(\{\hat{R}_{m,n}(X_i)\}, \{\hat{R}_{m,n}(Y_j)\}\right)$

Test: $\mathrm{H}_0 : \theta_2 = \theta_1$ vs. $\mathrm{H}_1 : \theta_2 = \theta_1 + hN^{-1/2}; h \neq 0 \in \mathbb{R}^p$

Theorem [Deb, Bhattacharya and S. (2021+)]

Under $\mathrm{H}_1 : \theta_2 = \theta_1 + hN^{-1/2}$,

$$\frac{mn}{N}\,\mathrm{RMMD}^2_{m,n} \xrightarrow{d} \sum_{j=1}^{\infty} \lambda_j \tilde{Z}_j^2$$

where $\tilde{Z}_j^2$ has non-central chi-squared distribution (depending on h).

- Let T_N denote the level α test based on the $\mathrm{RMMD}^2_{m,n}$
- Then, $\mathbb{E}_{\mathrm{H}_0}[T_N] = \alpha$ and $\lim_{\|h\|\to\infty} \lim_{N\to\infty} \mathbb{E}_{\mathrm{H}_1}[T_N] = 1$
- Rank MMD test has non-trivial power at the contiguous $N^{-1/2}$-scale
- Rank MMD has non-zero ARE compared to kernel MMD

Other (asymptotically) distribution-free GoF tests

- Crossmatch test of Rosenbaum (2005) is a distribution-free, consistent, and computationally feasible GoF test

- The crossmatch test S_N does not distinguish between the null and the alternative at the contiguous $N^{-1/2}$-scale, i.e., for any h:

$$\mathbb{E}_{\mathrm{H}_0}[S_N] = \alpha \qquad \text{and} \qquad \mathbb{E}_{\mathrm{H}_1}[S_N] \longrightarrow \alpha$$

- Pitman efficiency of rank MMD w.r.t. crossmatch is $+\infty$

Other (asymptotically) distribution-free GoF tests

- Crossmatch test of Rosenbaum (2005) is a distribution-free, consistent, and computationally feasible GoF test

- The crossmatch test S_N does not distinguish between the null and the alternative at the contiguous $N^{-1/2}$-scale, i.e., for any h:

$$\mathbb{E}_{\mathrm{H}_0}[S_N] = \alpha \qquad \text{and} \qquad \mathbb{E}_{\mathrm{H}_1}[S_N] \longrightarrow \alpha$$

- Pitman efficiency of rank MMD w.r.t. crossmatch is $+\infty$

- Many other graph-based[a] (asymptotically distribution-free) tests are also asymptotically powerless at $N^{-1/2}$-scale [Bhattacharya (2019)]

- The data depth-based (asymptotically distribution-free) tests have power at $N^{-1/2}$-scale, but computationally infeasible as d increases

[a] including Friedman & Rafsky (1979)'s MST based test; Schilling (1988) and Henze (1988) used k-nearest neighbor (k-NN) graph

Outline

Testing for mutual independence

- $(X, Y) \sim P$ on $\mathbb{R}^{d_1} \times \mathbb{R}^{d_2}$; $d_1, d_2 \geq 1$
- **Data**: n iid observations $\{(X_i, Y_i)\}_{i=1}^n$ from P
- Test if X is independent of Y, i.e.,

$$\mathrm{H}_0 : X \perp\!\!\!\perp Y \qquad \text{versus} \qquad \mathrm{H}_1 : X \not\!\perp\!\!\!\perp Y$$

Testing for mutual independence

- $(X, Y) \sim P$ on $\mathbb{R}^{d_1} \times \mathbb{R}^{d_2}$; $d_1, d_2 \geq 1$
- **Data**: n iid observations $\{(X_i, Y_i)\}_{i=1}^n$ from P
- Test if X is independent of Y, i.e.,

$$\mathrm{H}_0 : X \perp\!\!\!\perp Y \qquad \text{versus} \qquad \mathrm{H}_1 : X \not\perp\!\!\!\perp Y$$

- When $d_1 = d_2 = 1$: Pearson (1904), Spearman (1904), Kendall (1938), Hoeffding (1948), Blomqvist (1950), Blum et al. (1961), Rosenblatt (1975), Feuerverger (1993), ...
- When $d_1 > 1$ or $d_2 > 1$: Friedman and Rafsky (1979), Székely et al. (2007), Gretton et al. (2008), Oja (2010), Heller et al. (2013), Biswas et al. (2016), Berrett and Samworth (2019), ...

Can also test for K-vector/sample analogues of these problems

Testing for mutual independence

- $(X, Y) \sim P$ on $\mathbb{R}^{d_1} \times \mathbb{R}^{d_2}$, $\quad X \sim P_X$, $\quad Y \sim P_Y$, $\quad d_1, d_2 \geq 1$
- **Data**: $\{(X_i, Y_i) : 1 \leq i \leq n\}$ iid P
- **Test**: $\quad \mathrm{H}_0 : X \perp\!\!\!\perp Y \quad$ vs. $\quad \mathrm{H}_1 : X \not\perp\!\!\!\perp Y$

Testing for mutual independence

- $(X, Y) \sim P$ on $\mathbb{R}^{d_1} \times \mathbb{R}^{d_2}$, $\quad X \sim P_X$, $\quad Y \sim P_Y$, $\quad d_1, d_2 \geq 1$
- **Data**: $\{(X_i, Y_i) : 1 \leq i \leq n\}$ iid P
- **Test**: $\quad \mathrm{H}_0 : X \perp\!\!\!\perp Y \quad$ vs. $\quad \mathrm{H}_1 : X \not\!\perp\!\!\!\perp Y$

Distance Covariance [Szekely et al. (2007, 2009), Feuerverger (1993)]

- Let $(X, Y), (X', Y'), (X'', Y'') \overset{iid}{\sim} P$ (with finite mean), and set

$$h(s, t) := \|s - t\|$$

- **Distance covariance**: $\mathrm{dCov}(X, Y)$ is defined as

$$\mathrm{dCov}(X, Y) := \mathbb{E}\big[h(X, X')h(Y, Y')\big] + \mathbb{E}\big[h(X, X')\big]\mathbb{E}\big[h(Y, Y')\big] - 2\,\mathbb{E}\big[h(X, X')h(Y, Y'')\big] \geq 0$$

Testing for mutual independence

- $(X, Y) \sim P$ on $\mathbb{R}^{d_1} \times \mathbb{R}^{d_2}$, $\quad X \sim P_X$, $\quad Y \sim P_Y$, $\quad d_1, d_2 \geq 1$

- **Data**: $\{(X_i, Y_i) : 1 \leq i \leq n\}$ iid P

- **Test**: $\quad \mathrm{H}_0 : X \perp\!\!\!\perp Y \quad$ vs. $\quad \mathrm{H}_1 : X \not\perp\!\!\!\perp Y$

Distance Covariance [Szekely et al. (2007, 2009), Feuerverger (1993)]

- Let $(X, Y), (X', Y'), (X'', Y'') \stackrel{iid}{\sim} P$ (with finite mean), and set

$$h(s, t) := \|s - t\|$$

- **Distance covariance**: $\mathrm{dCov}(X, Y)$ is defined as

$$\begin{aligned}\mathrm{dCov}(X, Y) := \mathbb{E}\big[h(X, X')h(Y, Y')\big] \; + \; \mathbb{E}\big[h(X, X')\big]\mathbb{E}\big[h(Y, Y')\big] \\ - 2\,\mathbb{E}\big[h(X, X')h(Y, Y'')\big] \geq 0\end{aligned}$$

- Characterizes independence: $\mathrm{dCov}(X, Y) = 0 \quad$ iff $\quad X \perp\!\!\!\perp Y$

- $\mathrm{dCov}(X, Y) := \mathbb{E}\big[h(X, X')h(Y, Y')\big] \;+\; \mathbb{E}\big[h(X, X')\big]\mathbb{E}\big[h(Y, Y')\big]$
$$- 2\,\mathbb{E}\big[h(X, X')h(Y, Y'')\big] \geq 0$$
- **Sample distance covariance**: $\mathrm{dCov}_n = S_1 + S_2 - 2S_3$ where

$$S_1 = \frac{1}{n^2}\sum_{i,j=1}^{n} h(X_i, X_j)h(Y_i, Y_j), \qquad S_3 = \frac{1}{n^3}\sum_{i,j,k=1}^{n} h(X_i, X_j)h(Y_i, Y_k),$$

$$S_2 = \Big(\frac{1}{n^2}\sum_{i,j=1}^{n} h(X_i, X_j)\Big)\Big(\frac{1}{n^2}\sum_{i,j=1}^{n} h(Y_i, Y_j)\Big)$$

- $\mathrm{dCov}(X, Y) := \mathbb{E}\big[h(X, X')h(Y, Y')\big] + \mathbb{E}\big[h(X, X')\big]\mathbb{E}\big[h(Y, Y')\big]$
$$- 2\,\mathbb{E}\big[h(X, X')h(Y, Y'')\big] \geq 0$$

- **Sample distance covariance**: $\mathrm{dCov}_n = S_1 + S_2 - 2S_3$ where

$$S_1 = \frac{1}{n^2}\sum_{i,j=1}^{n} h(X_i, X_j)h(Y_i, Y_j), \qquad S_3 = \frac{1}{n^3}\sum_{i,j,k=1}^{n} h(X_i, X_j)h(Y_i, Y_k),$$

$$S_2 = \Big(\frac{1}{n^2}\sum_{i,j=1}^{n} h(X_i, X_j)\Big)\Big(\frac{1}{n^2}\sum_{i,j=1}^{n} h(Y_i, Y_j)\Big)$$

- **Test**: $\mathrm{H}_0 : X \perp\!\!\!\perp Y$ vs. $\mathrm{H}_1 : X \not\!\perp\!\!\!\perp Y$

- **Distance covariance test**: Reject H_0 if

$$\mathrm{dCov}_n(\{(X_i, Y_i)\}_{i=1}^{n}) > c_\alpha$$

- Critical value c_α depends on n, P_X, P_Y! (can use permutation test)

1. Optimal Transport: Monge's Problem
   - Introduction
   - Multivariate Ranks via Optimal Transport
2. Multivariate Two-sample Goodness-of-fit Testing
   - Hotelling T^2 and Kernel MMD
   - Distribution-free Testing
   - Lower bounds on Asymptotic (Pitman) Relative Efficiency
3. Testing for Independence Between Two Random Vectors
   - Distance Covariance
   - Distribution-free Testing

- **Test**: $\mathrm{H}_0 : X \perp\!\!\!\perp Y$ vs. $\mathrm{H}_1 : X \not\perp\!\!\!\perp Y$
- **Distance covariance test**: Reject H_0 if

$$\mathrm{dCov}_n(\{(X_i, Y_i)\}_{i=1}^n) > c_\alpha$$

- Critical value c_α depends on n, P_X, P_Y! (can use permutation test)

- **Test**: $\mathrm{H}_0 : X \perp\!\!\!\perp Y$ vs. $\mathrm{H}_1 : X \not\!\perp\!\!\!\perp Y$
- **Distance covariance test**: Reject H_0 if

$$\mathrm{dCov}_n(\{(X_i, Y_i)\}_{i=1}^n) > c_\alpha$$

- Critical value c_α depends on n, P_X, P_Y! (can use permutation test)
- Take $\mu_1 = \text{Uniform}([0,1]^{d_1})$ and $\mu_2 = \text{Uniform}([0,1]^{d_2})$

Rank distance covariance [Deb and S. (2019)]

- Sample rank of X_i: $\hat{R}_n^X : \{X_1, \ldots, X_n\} \to \{c_1^{(1)}, \ldots, c_n^{(1)}\} \subset [0,1]^{d_1}$
- Sample rank of Y_i: $\hat{R}_n^Y : \{Y_1, \ldots, Y_n\} \to \{c_1^{(2)}, \ldots, c_n^{(2)}\} \subset [0,1]^{d_2}$

- **Test**: $\mathrm{H}_0 : X \perp\!\!\!\perp Y$ vs. $\mathrm{H}_1 : X \not\!\perp\!\!\!\perp Y$
- **Distance covariance test**: Reject H_0 if

$$\mathrm{dCov}_n(\{(X_i, Y_i)\}_{i=1}^n) > c_\alpha$$

- Critical value c_α depends on n, P_X, P_Y! (can use permutation test)
- Take $\mu_1 = \text{Uniform}([0,1]^{d_1})$ and $\mu_2 = \text{Uniform}([0,1]^{d_2})$

Rank distance covariance [Deb and S. (2019)]

- Sample rank of X_i: $\hat{R}_n^X : \{X_1, \ldots, X_n\} \to \{c_1^{(1)}, \ldots, c_n^{(1)}\} \subset [0,1]^{d_1}$
- Sample rank of Y_i: $\hat{R}_n^Y : \{Y_1, \ldots, Y_n\} \to \{c_1^{(2)}, \ldots, c_n^{(2)}\} \subset [0,1]^{d_2}$
- **Rank distance cov.**: $\mathrm{RdCov}_n = \mathrm{dCov}_n\left(\left\{(\hat{R}_n^X(X_i), \hat{R}_n^Y(Y_i))\right\}_{i=1}^n\right)$

Distribution-freeness

X and Y abs. cont. Under H_0, the dist. of RdCov_n is free of P_X and P_Y.

- Under H_0, distribution of RdCov_n just depends on $c_j^{(k)}$'s, n, d_1, d_2
- **Rank distance covariance test**: Reject H_0 if $\quad \mathrm{RdCov}_n > \kappa_\alpha^{(n)}$

- Under H_0, distribution of RdCov_n just depends on $c_j^{(k)}$'s, n, d_1, d_2

- **Rank distance covariance test**: Reject H_0 if $\mathrm{RdCov}_n > \kappa_\alpha^{(n)}$

Limiting distribution under H_0 [Deb and S. (2019)]

Suppose: (i) X and Y are abs. cont., and

(ii) $\frac{1}{n}\sum_{j=1}^{n}\delta_{c_j^{(k)}} \xrightarrow{d}$ Uniform$([0,1]^{d_k})$, for $k = 1, 2$.

Then, under H_0, $\exists$ universal distribution $\mathbb{L}_{d_1,d_2}$ (not depending on $c_j^{(k)}$'s) s.t.

$$n \cdot \mathrm{Rdcov}_n \xrightarrow{d} \mathbb{L}_{d_1,d_2} \qquad \text{as } n \to \infty.$$

The choice of the $c_j^{(k)}$'s have no effect for large n

- Under H_0, distribution of RdCov_n just depends on $c_j^{(k)}$'s, n, d_1, d_2
- **Rank distance covariance test**: Reject H_0 if $\mathrm{RdCov}_n > \kappa_\alpha^{(n)}$

Limiting distribution under H_0 [Deb and S. (2019)]

Suppose: (i) X and Y are abs. cont., and

(ii) $\frac{1}{n}\sum_{j=1}^{n} \delta_{c_j^{(k)}} \xrightarrow{d}$ Uniform$([0,1]^{d_k})$, for $k = 1, 2$.

Then, under H_0, $\exists$ universal distribution $\mathbb{L}_{d_1,d_2}$ (not depending on $c_j^{(k)}$'s) s.t.

$$n \cdot \mathrm{Rdcov}_n \xrightarrow{d} \mathbb{L}_{d_1,d_2} \qquad \text{as } n \to \infty.$$

The choice of the $c_j^{(k)}$'s have no effect for large n

Power

Suppose $X \not\perp\!\!\!\perp Y$, and (i) & (ii) hold. Then,

$$\mathbb{P}\left(\mathrm{RdCov}_n > \kappa_\alpha^{(n)}\right) \to 1 \qquad \text{as } n \to \infty.$$

Proposed test has asymptotic power 1, against all fixed alternatives

When $d_1 = d_2 = 1$

When $d_1 = d_2 = 1$, RdCov_n has close connections to Hoeffding's D-statistic [Hoeffding (1948)] (see Blum et al. (1961)):

$$\frac{1}{4}\mathrm{RdCov}_n = \int \left\{\mathbb{F}_n(x,y) - \mathbb{F}_n^X(x)\mathbb{F}_n^Y(y)\right\}^2 d\mathbb{F}_n^X(x)\, d\mathbb{F}_n^Y(y)$$

where $\mathbb{F}_n$, $\mathbb{F}_n^X$, and $\mathbb{F}_n^Y$ are the empirical c.d.f.'s of (X, Y), X and Y.

When $d_1 = d_2 = 1$

When $d_1 = d_2 = 1$, RdCov_n has close connections to Hoeffding's D-statistic [Hoeffding (1948)] (see Blum et al. (1961)):

$$\frac{1}{4}\mathrm{RdCov}_n = \int \left\{\mathbb{F}_n(x,y) - \mathbb{F}_n^X(x)\mathbb{F}_n^Y(y)\right\}^2 d\mathbb{F}_n^X(x)\, d\mathbb{F}_n^Y(y)$$

where $\mathbb{F}_n$, $\mathbb{F}_n^X$, and $\mathbb{F}_n^Y$ are the empirical c.d.f.'s of (X, Y), X and Y.

- Our general principle could have been used with any other procedure for mutual independence testing, e.g., the HSIC statistic [Gretton et al. (2005)] which uses ideas from RKHS, ...

When $d_1 = d_2 = 1$

When $d_1 = d_2 = 1$, RdCov_n has close connections to Hoeffding's D-statistic [Hoeffding (1948)] (see Blum et al. (1961)):

$$\frac{1}{4}\mathrm{RdCov}_n = \int \left\{\mathbb{F}_n(x,y) - \mathbb{F}_n^X(x)\mathbb{F}_n^Y(y)\right\}^2 d\mathbb{F}_n^X(x)\, d\mathbb{F}_n^Y(y)$$

where $\mathbb{F}_n$, $\mathbb{F}_n^X$, and $\mathbb{F}_n^Y$ are the empirical c.d.f.'s of (X, Y), X and Y.

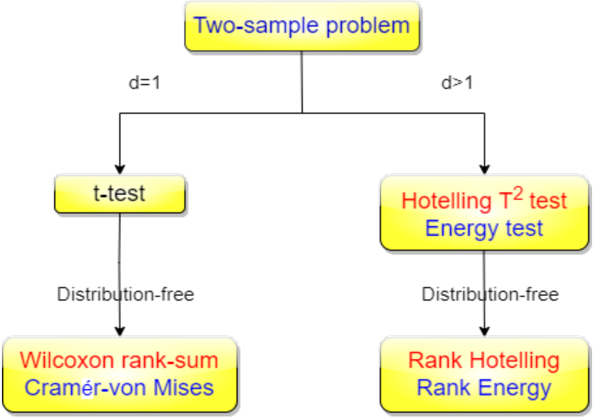
- Our general principle could have been used with any other procedure for mutual independence testing, e.g., the HSIC statistic [Gretton et al. (2005)] which uses ideas from RKHS, ...
- The other computationally feasible distribution-free test in the context was proposed in Heller et al. (2012); however they do not guarantee consistency against all fixed alternatives

Summary

- Multivariate distribution-free testing procedures
- Based on multivariate ranks defined via optimal transport

Summary

- Multivariate distribution-free testing procedures
- Based on multivariate ranks defined via optimal transport
- Proposed a general framework, other examples may include testing for symmetry, testing the equality of K-distributions, independence testing of K-vectors, ...
- The proposed tests are: (i) distribution-free and have good efficiency, (ii) computationally feasible, (iii) more powerful for distributions with heavy tails, and (iv) robust to outliers & contamination



Ghosal and S. (2019). https://arxiv.org/abs/1905.05340 (AoS, to appear)
Deb and S. (2019). https://arxiv.org/pdf/1909.08733 (JASA, to appear)
Deb, Ghosal and S. (2021). https://arxiv.org/pdf/2107.01718. NeurIPS
Deb, Bhattacharya and S. (2021). https://arxiv.org/abs/2104.01986
Deb, Bhattacharya and S. (2021+). (working paper)

Thank you very much!

Questions?

- $\nu_n := \frac{1}{n}\sum_{i=1}^{n} \delta_{X_i}, \quad \mu_n := \frac{1}{n}\sum_{j=1}^{n} \delta_{c_j}$

- **OT maps**: $R\#\nu = \mu, \qquad \hat{R}_n\#\nu_n = \mu_n$

- $\nu_n := \frac{1}{n}\sum_{i=1}^n \delta_{X_i}, \quad \mu_n := \frac{1}{n}\sum_{j=1}^n \delta_{c_j}$
- **OT maps**: $R\#\nu = \mu, \qquad \hat{R}_n\#\nu_n = \mu_n$
- Suppose $R = \nabla\varphi$, where $\varphi : \mathbb{R}^d \to \mathbb{R} \cup \{+\infty\}$ is convex
- **Legendre-Fenchel dual of** φ: $\varphi^*(y) := \sup_{x\in\mathbb{R}^d}[x^\top y - \varphi(x)]$
- **Fact 1**: R is $\frac{1}{\lambda}$-Lipschitz iff φ^* is λ-strongly convex
- φ^* is λ-strongly convex if, for all $x, y \in \mathrm{Dom}(\varphi^*)$,
$$\varphi^*(y) \geq \varphi^*(x) + \nabla\varphi^*(x)^\top(y-x) + \frac{\lambda}{2}\|y-x\|^2$$

- $\nu_n := \frac{1}{n}\sum_{i=1}^n \delta_{X_i}, \quad \mu_n := \frac{1}{n}\sum_{j=1}^n \delta_{c_j}$
- **OT maps**: $R\#\nu = \mu, \qquad \hat{R}_n\#\nu_n = \mu_n$
- Suppose $R = \nabla\varphi$, where $\varphi : \mathbb{R}^d \to \mathbb{R} \cup \{+\infty\}$ is convex
- **Legendre-Fenchel dual of** φ: $\varphi^*(y) := \sup_{x\in\mathbb{R}^d}[x^\top y - \varphi(x)]$
- **Fact 1**: R is $\frac{1}{\lambda}$-Lipschitz iff φ^* is λ-strongly convex
- φ^* is λ-strongly convex if, for all $x, y \in \mathrm{Dom}(\varphi^*)$,
$$\varphi^*(y) \geq \varphi^*(x) + \nabla\varphi^*(x)^\top(y-x) + \frac{\lambda}{2}\|y-x\|^2.$$
- **Fact 2**: $\nabla\varphi^*(R(x)) = x$ a.e.
- The 2-Wasserstein distance (squared) between ν and μ is defined as:
$$W_2^2(\nu,\mu) := \min_{\pi\in\Pi(\nu,\mu)} \int \|x-y\|^2\, d\pi(x,y),$$
where $\Pi(\nu,\mu) := \{$distributions on $\mathbb{R}^d \times \mathbb{R}^d$ with marginals ν & $\mu\}$.

Estimation of OT map [Deb, Ghosal and S. (2021)] Rate of convergence

If the population rank map $R(\cdot)$ is $\frac{1}{\lambda}$-Lipschitz, then

$$\lambda \int \|\hat{R}_n(x) - R(x)\|^2 \, d\nu_n(x) \leq W_2^2(\nu_n, \tilde{\mu}_n) - W_2^2(\nu_n, \mu_n) + 2 \int g \, d(\mu_n - \tilde{\mu}_n)$$

where $\tilde{\mu}_n := \frac{1}{n} \sum_{i=1}^n \delta_{R(X_i)}$ and $g(y) := \varphi^*(y) - \frac{1}{2}\|y\|^2$.

Estimation of OT map [Deb, Ghosal and S. (2021)]

If the population rank map $R(\cdot)$ is $\frac{1}{\lambda}$-Lipschitz, then

$$\lambda \int \|\hat{R}_n(x) - R(x)\|^2 \, d\nu_n(x) \leq W_2^2(\nu_n, \tilde{\mu}_n) - W_2^2(\nu_n, \mu_n) + 2 \int g \, d(\mu_n - \tilde{\mu}_n)$$

where $\tilde{\mu}_n := \frac{1}{n} \sum_{i=1}^n \delta_{R(X_i)}$ and $g(y) := \varphi^*(y) - \frac{1}{2}\|y\|^2$.

- Then, recalling $\nu_n := \frac{1}{n} \sum_{i=1}^n \delta_{X_i}$ and $\mu_n := \frac{1}{n} \sum_{j=1}^n \delta_{c_j}$,

$$\begin{aligned}
D_1 &:= \int \varphi^* d\mu_n - \int \varphi^* d\tilde{\mu}_n \\
&= \int [\varphi^*(\hat{R}_n(x)) - \varphi^*(R(x))] d\nu_n(x) \qquad (\text{as } \hat{R}_n \# \nu_n = \mu_n) \\
&\overset{(a)}{\geq} \int \left\{ \nabla \varphi^*(R(x))^\top (\hat{R}_n(x) - R(x)) + \frac{\lambda}{2} \|\hat{R}_n(x) - R(x)\|^2 \right\} d\nu_n(x) \\
&\overset{(b)}{=} \underbrace{\int x^\top (\hat{R}_n(x) - R(x)) d\nu_n(x)}_{D_2} + \frac{\lambda}{2} \int \|\hat{R}_n(x) - R(x)\|^2 d\nu_n(x)
\end{aligned}$$

- **Fact 3**: $2D_2 = W_2^2(\nu_n, \tilde{\mu}_n) - W_2^2(\nu_n, \mu_n) + \int \|y\|^2 \, d(\mu_n - \tilde{\mu}_n)(y)$

- Then 2-Wasserstein (squared) distance between ν and μ is:

$$W_2^2(\nu, \mu) := \min_{\pi \in \Pi(\nu,\mu)} \int \|x - y\|^2 \, d\pi(x, y), \tag{3}$$

where $\Pi(\nu, \mu) := \{$distributions on $\mathbb{R}^d \times \mathbb{R}^d$ with marginals ν & $\mu\}$.

- Then 2-Wasserstein (squared) distance between ν and μ is:

$$W_2^2(\nu,\mu) := \min_{\pi \in \Pi(\nu,\mu)} \int \|x-y\|^2 \, d\pi(x,y), \tag{3}$$

where $\Pi(\nu,\mu) := \{\text{distributions on } \mathbb{R}^d \times \mathbb{R}^d \text{ with marginals } \nu \text{ \& } \mu\}$.

- Let γ be a minimizer of (3). The barycentric projection of γ is

$$T(x) := \frac{\int_y y \, d\gamma(x,y)}{\int_y d\gamma(x,y)} = \mathbb{E}_\gamma[Y|X=x].$$

Thus, $T(x)$ is the conditional mean of Y given $X = x$ under γ.

- Then 2-Wasserstein (squared) distance between ν and μ is:

$$W_2^2(\nu, \mu) := \min_{\pi \in \Pi(\nu,\mu)} \int \|x - y\|^2 \, d\pi(x, y), \tag{3}$$

where $\Pi(\nu, \mu) := \{\text{distributions on } \mathbb{R}^d \times \mathbb{R}^d \text{ with marginals } \nu \ \& \ \mu\}$.

- Let γ be a minimizer of (3). The barycentric projection of γ is

$$T(x) := \frac{\int_y y \, d\gamma(x, y)}{\int_y d\gamma(x, y)} = \mathbb{E}_\gamma[Y|X = x].$$

Thus, $T(x)$ is the conditional mean of Y given $X = x$ under γ.

- When $\exists$ an OT map R such that $R\#\nu = \mu$, then $R = T$

Estimation of T using Barycentric projection

- Let $\nu_n := \frac{1}{n}\sum_{i=1}^n \delta_{X_i}$ and $\mu_m := \frac{1}{m}\sum_{j=1}^m \delta_{c_j}$
- Let $\tilde{\gamma} := \underset{\pi \in \Pi(\nu_n, \mu_m)}{\arg\min} \int \|x - y\|^2 \, d\pi(x, y)$ — optimal coupling
- Define $\tilde{R}$ as the barycentric projection of $\tilde{\gamma}$