Multivariate Distribution-free Nonparametric Testing using Optimal Transport

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

Statistical Machine Learning Bootcamp
Columbia University
New York

16 January, 2020

[1]Supported by NSF grants DMS-1712822

Nonparametric testing

Consider the following two nonparametric hypothesis testing problems

Testing for equality of distributions (two-sample goodness-of-fit (GoF))

- **Data**: $\{\mathbf{X}_i\}_{i=1}^m$ iid P_1 on $\mathbb{R}^d$; $\{\mathbf{Y}_j\}_{j=1}^n$ 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 $d = 1$: Smirnov (1939), Wald and Wolfowitz (1940), Wilcoxon (1945), Mann and Whitney (1947), Anderson (1962), ...
- When $d > 1$: 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), ...

Nonparametric testing

Testing for mutual independence

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

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

- 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), ...

Two-sample goodness-of-fit testing: when $d = 1$

- **Data**: $\{X_i\}_{i=1}^m$ iid $N(\mu_1, \sigma^2)$; $\{Y_j\}_{j=1}^n$ iid $N(\mu_2, \sigma^2)$
- **Unknowns**: $\mu_1, \mu_2 \in \mathbb{R}$, $\sigma^2 > 0$
- **Test**: $\mathrm{H}_0 : \mu_2 = \mu_1$ versus $\mathrm{H}_1 : \mu_2 > \mu_1$

Two-sample t-test for normal populations

- Test statistic based on, for $\bar{X}_m = \frac{1}{m}\sum_{i=1}^m X_i$ and $\bar{Y}_n = \frac{1}{n}\sum_{j=1}^n Y_j$,

$$\bar{Y}_n - \bar{X}_m \sim N\left(0, \sigma^2\left(\frac{1}{n} + \frac{1}{m}\right)\right) \qquad \text{(under } \mathrm{H}_0\text{)}$$

- **Test statistic**: $T_{m,n} := \frac{\sqrt{m+n-2}\,(\bar{Y}_n - \bar{X}_m)}{\sqrt{S_X^2 + S_Y^2}\sqrt{\frac{1}{m}+\frac{1}{n}}} \sim t_{m+n-2}$ (under H_0)

  where $S_X^2 := \sum_{i=1}^m (X_i - \bar{X}_m)^2$, and $S_Y^2 := \sum_{j=1}^n (Y_j - \bar{Y}_n)^2$

- **Reject** H_0 if $T_{m,n} > q_{1-\alpha}^{(m+n-2)}$, $\alpha \in (0,1)$

- $q_{1-\alpha}^{(m+n-2)} = (1-\alpha)$-th quantile of t-distribution with $m+n-2$ d.f.

- **Data**: $\{X_i\}_{i=1}^m$ iid c.d.f. $F(\cdot)$; $\{Y_j\}_{j=1}^n$ iid c.d.f. $G(\cdot) \equiv F(\cdot - \theta)$
- **Unknowns**: $F(\cdot)$ (unknown continuous c.d.f.) and θ
- **Test**: $\mathrm{H}_0 : \theta = 0$ versus $\mathrm{H}_1 : \theta > 0$

Two sample t-test

- **Test statistic**: $T_{m,n} := \frac{\sqrt{m+n-2}\,(\bar{Y}_n - \bar{X}_m)}{\sqrt{S_X^2 + S_Y^2}\sqrt{\frac{1}{m}+\frac{1}{n}}} \overset{\text{approx}}{\sim} t_{m+n-2}$ (under H_0)

  where $S_X^2 := \sum_{i=1}^m (X_i - \bar{X}_m)^2$, and $S_Y^2 := \sum_{j=1}^n (Y_j - \bar{Y}_n)^2$

- Reject H_0 if $T_{m,n} > q_{1-\alpha}^{(m+n-2)}$, $\alpha \in (0,1)$

- $q_{1-\alpha}^{(m+n-2)} = (1-\alpha)$-th quantile of t-distribution with $m+n-2$ d.f.

- Approximate level $1-\alpha$ test, i.e., when $F \equiv G$,

$$\mathbb{P}(T_{m,n} > q_{1-\alpha}^{(m+n-2)}) \approx \alpha$$

Question: Can we find a robust, distribution-free test in this case?

- **Data**: $\{X_i\}_{i=1}^m$ iid c.d.f. $F(\cdot)$; $\{Y_j\}_{j=1}^n$ iid c.d.f. $G(\cdot) \equiv F(\cdot - \theta)$
- **Unknowns**: $F(\cdot)$ (unknown continuous c.d.f.) and θ
- **Test**: $\mathrm{H}_0 : \theta = 0$ versus $\mathrm{H}_1 : \theta > 0$

Wilcoxon rank-sum test (1945) [a.k.a. Mann-Whitney test (1947)]

- Rank the pooled sample: $Z_1 < Z_2 < \ldots < Z_N$; here $N = m + n$
- Pooled-rank of X_i is $R_i \in \{\frac{1}{N}, \frac{2}{N}, \ldots, \frac{N}{N}\}$; pooled-rank of Y_j is Q_j
- **Reject** H_0 if $\sum_{j=1}^n Q_j > \kappa_\alpha^{(m,n)}$
- Equivalent to the 2-sample t-test based on the pooled ranks
- $\sum_{j=1}^n Q_j$ is distribution-free (under H_0), i.e., no dependence on F
- $(R_1, \ldots, R_m, Q_1, \ldots, Q_n)$ is distributed uniformly over the $N!$ permutations of $\{\frac{1}{N}, \frac{2}{N}, \ldots, \frac{N}{N}\}$
- Leads to universal critical value $\kappa_\alpha^{(m,n)}$ (only dependent on α, m, n)

Comparison of Wilcoxon rank-sum (WRS) test with 2-sample t-test

- WRS test is distribution-free and exact for all F continuous
- WRS test has 0.95 efficiency w.r.t. t-test when F 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.)

Classical nonparametrics

- The paper Wilcoxon (1945) led to the development of this field
- Procedures valid under weak set of assumptions
- Yet have high efficiency compared to classical methods; robust to outliers & contamination
- Lead to distribution-free methods

What are generalizations of this approach for multivariate data?

Multivariate distribution-free nonparametric testing

- A general framework for multivariate distribution-free nonparametric testing based on ranks [Deb and S. (2019), Shi et al. (2019), Ghosal and S. (2019), Beirlant et al. (2019), Hallin et al. (2019), ...]

- Multivariate ranks obtained using the theory of optimal transport [Hallin (2017), del Barrio et al. (2018), Chernozhukov et al. (2017), Easton and McCulloch (1990), ...]

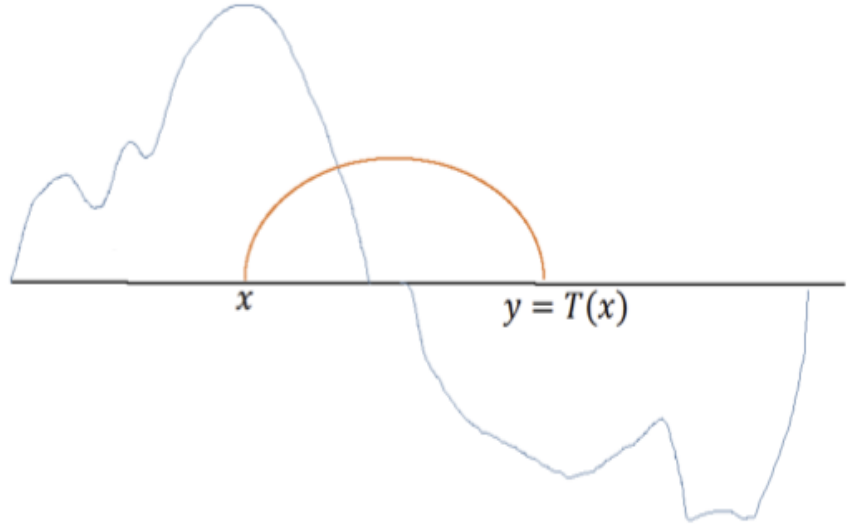
Why ranks?

- In one-dimension, ranks lead to distribution-free tests

- **Examples**: Wilcoxon rank-sum test [Wilcoxon (1945)], Spearman's rank correlation [Spearman (1904)], two-sample Kolmogorov-Smirnov test [Smirnoff (1933)], two-sample Cramér-von Mises statistic [Anderson (1962)], Wilcoxon signed-rank test [Wilcoxon (1945)], Wald-Wolfowitz runs test [Wald and Wolfowitz (1940)], Mann-Whitney rank-sum test [Mann and Whitney (1947)], Kruskal-Wallis test [Kruskal (1952)], Hoeffding's D-test [Hoeffding (1948)], etc. ...

1. Distribution-free Nonparametric Testing using Optimal Transport
   - Nonparametric Testing: Introduction
   - Optimal Transport: Monge's Problem
2. Multivariate Two-sample Goodness-of-fit Testing
3. Testing Independence of Two Random Vectors

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

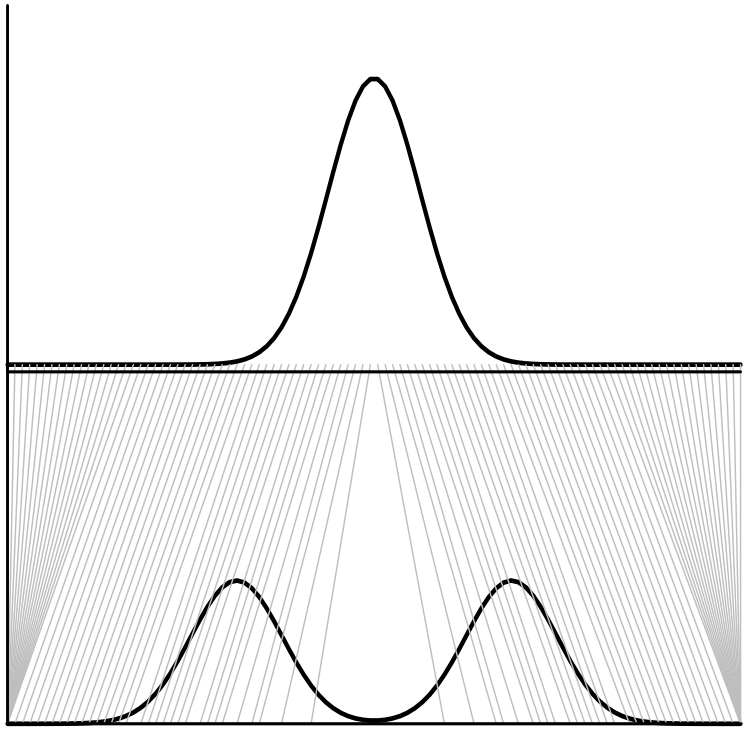


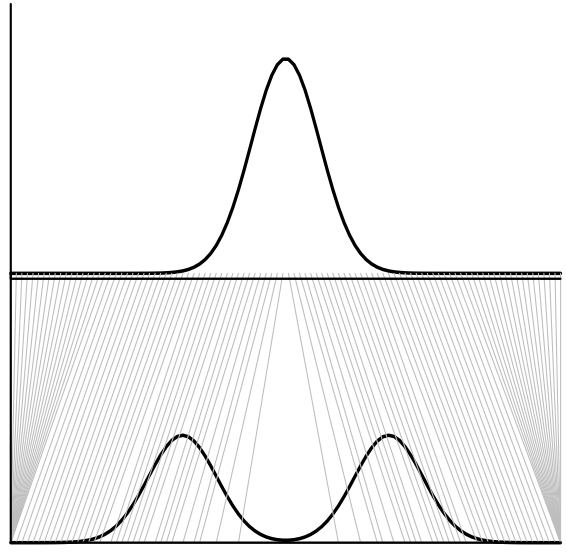
Goal: $\inf_{T:T(X)\sim\mu} \mathbb{E}_\nu[c(X, T(X))] \qquad X \sim \nu$

- ν (on $\mathcal{X}$) and μ (on $\mathcal{Y}$) probability measures, $\int_{\mathcal{X}} d\nu(x) = \int_{\mathcal{Y}} d\mu(y) = 1$
- $c(x, y) \geq 0$: cost of transporting x to y (e.g., $c(x, y) = \|x - y\|^2$)
- $T(X) \sim \mu$ where $X \sim \nu$; $\quad T$ transports ν to μ

One-dimensional optimal transport

- Suppose $\mathcal{X}, \mathcal{Y} \subset \mathbb{R}$; μ, ν abs. cont.; F_μ and F_ν c.d.f.'s
- **Goals**: (i) Transport ν to μ; i.e., find T s.t. if $X \sim \nu$ then $T(X) \sim \mu$
- (ii) T minimizes cost $\mathbb{E}_\nu[(X - T(X))^2]$; assume $c(x, y) = (x - y)^2$





- The minimizing T must satisfy (Why?)

$$(x_0 - T(x_0))^2 + (x_1 - T(x_1))^2 \leq (x_0 - T(x_1))^2 + (x_1 - T(x_0))^2$$

- This means that if $x_1 > x_0$ then $T(x_1) \geq T(x_0)$
- So T must be a monotone nondecreasing function

- As $T(\cdot)$ is nondecreasing, for $x \in \mathbb{R}$,

$$\mathbb{P}\Big(X \in (-\infty, x]\Big) = \mathbb{P}\Big(T(X) \in (-\infty, T(x)]\Big) \quad \Rightarrow \; F_\nu(x) = F_\mu(T(x))$$

- Thus, $T = F_\mu^{-1} \circ F_\nu$ (and this map T is unique)

Optimal transportation when $d = 1$

- $\mathcal{X}, \mathcal{Y} \subset \mathbb{R}$; μ, ν abs. cont.; F_μ and F_ν c.d.f.'s
- **Goals**: (i) Transport ν to μ; i.e., find T s.t. if $X \sim \nu$ then $T(X) \sim \mu$
  (ii) T minimizes cost $\mathbb{E}_\nu[(X - T(X))^2]$
- **Solution**: $T = F_\mu^{-1} \circ F_\nu$ (and this map T is unique)

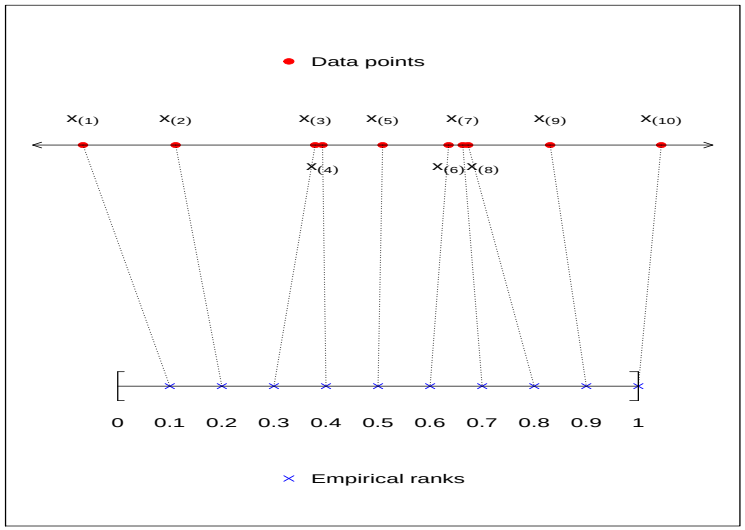
Rank function as a transport map: when $d = 1$

- $X \sim \nu$ (abs. cont.) on $\mathbb{R}$, $F \equiv F_\nu$ c.d.f. of ν
- **Rank**: The 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
- In fact, if $\mathbb{E}[X^2] < \infty$, c.d.f. F is the optimal transport map as

$$F = \underset{T : T(X) \sim \mu}{\arg\min}\ \mathbb{E}_\nu[(X - T(X))^2]$$

Sample ranks: When $d = 1$

- **Data**: $X_1, \ldots, X_n$ iid on $\mathbb{R}$ (having a cont. distribution)
- Rank map assigns $\{X_1, X_2, \ldots, X_n\}$ to elements of $\{\frac{1}{n}, \frac{2}{n}, \ldots, \frac{n}{n}\}$



Sample rank map is also a transport map:

$$\hat{R}_n := \arg\min_T \frac{1}{n}\sum_{i=1}^n |X_i - T(X_i)|^2 = \arg\max_T \frac{1}{n}\sum_{i=1}^n X_{(i)} \cdot T(X_{(i)})$$

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

Multivariate rank functions as transport maps

- $\mathbf{X} \sim \nu$; ν is a probability measure in $\mathbb{R}^d$ (abs. cont.)
- Find "optimal" transport map $\mathbf{T}$ s.t. $\mathbf{T}(\mathbf{X}) \stackrel{d}{=} \mathbf{U} \sim \text{Unif}([0,1]^d) \equiv \mu$

Population rank function

If $\mathbb{E}\|\mathbf{X}\|^2 < \infty$, rank function $\mathbf{R} : \mathbb{R}^d \to [0,1]^d$ is the transport map s.t.

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

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

- $\mathbf{R}(\cdot)$ characterizes distribution: $\mathbf{R}_1(\mathbf{x}) = \mathbf{R}_2(\mathbf{x})\ \forall\, \mathbf{x} \in \mathbb{R}^d$ iff $P_1 = P_2$
- $\mathbf{R}(\cdot)$ is invertible, i.e., there exists $\mathbf{Q}(\cdot)$ such that

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

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

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

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

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

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

Suppose $\mathbf{X} \sim \nu$ abs. cont. There is an ν-a.e. unique measurable mapping $\mathbf{R} : \mathbb{R}^d \to [0,1]^d$, transporting $\mathbf{X}$ to $\mathbf{U}$ (i.e., $\mathbf{R}(\mathbf{X}) \overset{d}{=} \mathbf{U}$), of the form

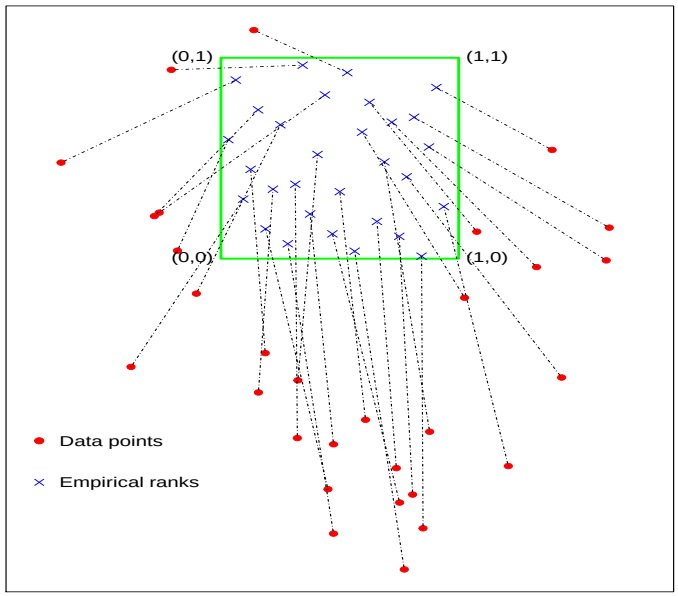
$$\mathbf{R}(\mathbf{x}) = \nabla\varphi(\mathbf{x}), \qquad \text{for } \nu\text{-a.e. } \mathbf{x}, \tag{2}$$

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

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

- **Data**: $\mathbf{X}_1, \ldots, \mathbf{X}_n$ iid ν on $\mathbb{R}^d$ (abs. cont. distribution)

- Empirical rank map $\hat{\mathbf{R}}_n : \{\mathbf{X}_1, \ldots, \mathbf{X}_n\} \to \{\mathbf{c}_1, \ldots, \mathbf{c}_n\} \subset [0,1]^d$ — sequence of "uniform-like" points (quasi-Monte Carlo sequence)



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

$$\hat{\mathbf{R}}_n = \arg\min_{\mathbf{T}} \frac{1}{n} \sum_{i=1}^{n} \|\mathbf{X}_i - \mathbf{T}(\mathbf{X}_i)\|^2$$

where $\mathbf{T}$ transports $\nu_n := \frac{1}{n}\sum_{i=1}^n \delta_{\mathbf{X}_i}$ to $\mu_n := \frac{1}{n}\sum_{i=1}^n \delta_{\mathbf{c}_i}$

- Assignment problem (can be reduced to a linear program — $O(n^3)$)

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

Suppose that $\mathbf{X}_1, \ldots, \mathbf{X}_n$ iid on $\mathbb{R}^d$ with abs. cont. distribution. Then,

$$(\hat{\mathbf{R}}_n(\mathbf{X}_1), \ldots, \hat{\mathbf{R}}_n(\mathbf{X}_n))$$

is uniformly distributed over the $n!$ permutations of $\{\mathbf{c}_1, \ldots, \mathbf{c}_n\}$.

This is the first step to obtaining distribution-free tests

Regularity: a.s.-convergence [Deb and S. (2019)]

$\mathbf{X}_1, \ldots, \mathbf{X}_n$ iid ν (abs. cont.). If $\frac{1}{n}\sum_{i=1}^n \delta_{\mathbf{c}_i} \xrightarrow{d} \text{Unif}([0,1]^d)$, then

$$\frac{1}{n}\sum_{i=1}^n \|\hat{\mathbf{R}}_n(\mathbf{X}_i) - \mathbf{R}(\mathbf{X}_i)\| \xrightarrow{a.s.} 0 \qquad \text{as } n \to \infty.$$

Result gives the required regularity to the empirical multivariate rank map

Open research question: What is the rate of convergence of $\hat{\mathbf{R}}_n$ to $\mathbf{R}$? [Hütter and Rigollet (2019)]

Multivariate two-sample goodness-of-fit test

Testing for equality of two multivariate distributions

- **Data**: $\{\mathbf{X}_i\}_{i=1}^m$ iid P_1 on $\mathbb{R}^d$; $\{\mathbf{Y}_j\}_{j=1}^n$ iid P_2 on $\mathbb{R}^d$, $d \geq 1$
- Test if the two samples come from the same distribution, i.e.,

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

- Start with a "good" test, say the energy statistic [Székely (2003), Székely and Rizzo (2013)]; can also use any kernel test (MMD) [Gretton et al. (2012), Sejdinovic et al. (2013)]
- Suppose $\mathbf{X}, \mathbf{X}' \overset{iid}{\sim} P_1$, $\mathbf{Y}, \mathbf{Y}' \overset{iid}{\sim} P_2$ and set $h(\mathbf{s}, \mathbf{t}) := \|\mathbf{s} - \mathbf{t}\|$
- The energy distance between P_1 and P_2:

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

- Characterizes equality of distributions: $\mathrm{E}(P_1, P_2) = 0$ iff $P_1 = P_2$

- The energy distance between P_1 and P_2:

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

- **E-statistic**: $\mathrm{E}^2_{m,n}\left(\{\mathbf{X}_i\}_{i=1}^m, \{\mathbf{Y}_j\}_{j=1}^n\right) := 2A - B - C$ where

$$A = \frac{1}{mn}\sum_{i,j=1}^{m,n} h(\mathbf{X}_i, \mathbf{Y}_j), \quad B = \frac{1}{m^2}\sum_{i,j=1}^{m} h(\mathbf{X}_i, \mathbf{X}_j), \quad C = \frac{1}{n^2}\sum_{i,j=1}^{n} h(\mathbf{Y}_i, \mathbf{Y}_j)$$

Energy test [Székely (2003)]

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

- **Test**: Reject H_0 if $\mathrm{E}_{m,n}\left(\{\mathbf{X}_i\}_{i=1}^m, \{\mathbf{Y}_j\}_{j=1}^n\right) > c_\alpha$

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

Rank energy test

Rank energy statistic [Deb and S. (2019)]

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

$$\hat{\mathbf{R}}_{m,n} : \{\mathbf{X}_1, \ldots, \mathbf{X}_m, \mathbf{Y}_1, \ldots, \mathbf{Y}_n\} \to \{\mathbf{c}_1, \ldots, \mathbf{c}_{m+n}\} \subset [0,1]^d$$

- **Rank energy**: $\mathrm{RE}^2_{m,n} := \mathrm{E}^2_{m,n}\left(\{\hat{\mathbf{R}}_{m,n}(\mathbf{X}_i)\}_{i=1}^m, \{\hat{\mathbf{R}}_{m,n}(\mathbf{Y}_j)\}_{j=1}^n\right)$

Distribution-freeness

Under H_0, distribution of $\mathrm{RE}_{m,n}$ is free of $P_1 \equiv P_2$, if P_1 is abs. cont.

- Dist. of $\mathrm{RE}_{m,n}$ just depends on $\mathbf{c}_i$'s, m, n and d
- **Rank energy test**: Reject H_0 if

$$\mathrm{RE}_{m,n} > \kappa_\alpha^{(m,n)}$$

where $\kappa_\alpha^{(m,n)}$ is a universal threshold (free of $P_1 \equiv P_2$)

Limiting distribution under $\mathrm{H}_0 : P_1 = P_2$

If (i) $P_1 \equiv P_2$ is abs. cont., and
(ii) $\frac{1}{n}\sum_{i=1}^{n} \delta_{\mathbf{c}_i} \xrightarrow{d} \text{Uniform}([0,1]^d)$,

then, under H_0, for some universal $\{\lambda_j \geq 0 : j \geq 1\}$,

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

where $\{Z_j\}_{j\geq 1}$ are iid $N(0,1)$.

The choice of the $\mathbf{c}_i$'s have no effect for large m, n

Power

Under (ii) and $P_1 \neq P_2$, if $m/(m+n) \to \lambda \in (0,1)$ then,

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

Proposed test has asymptotic power 1, against all fixed alternatives

When $d = 1$

When $d = 1$, $\mathrm{RE}_{m,n}$ is equivalent to two-sample Cramér-von Mises statistic [Anderson (1962)]:

$$\frac{1}{2}\mathrm{RE}^2_{m,n} = \int \left\{\mathbb{F}^X_m(t) - \mathbb{F}^Y_n(t)\right\}^2 d\mathbb{F}_{m+n}(t)$$

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

- Our general principle could have been used with any other procedure for testing equality of distributions, e.g., the MMD statistic [Gretton et al. (2012)] which uses ideas from RKHS, ...

- For example, take "any" kernel $K(\cdot, \cdot)$ in

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

  and all the results hold almost verbatim

Power plot with varying location parameter

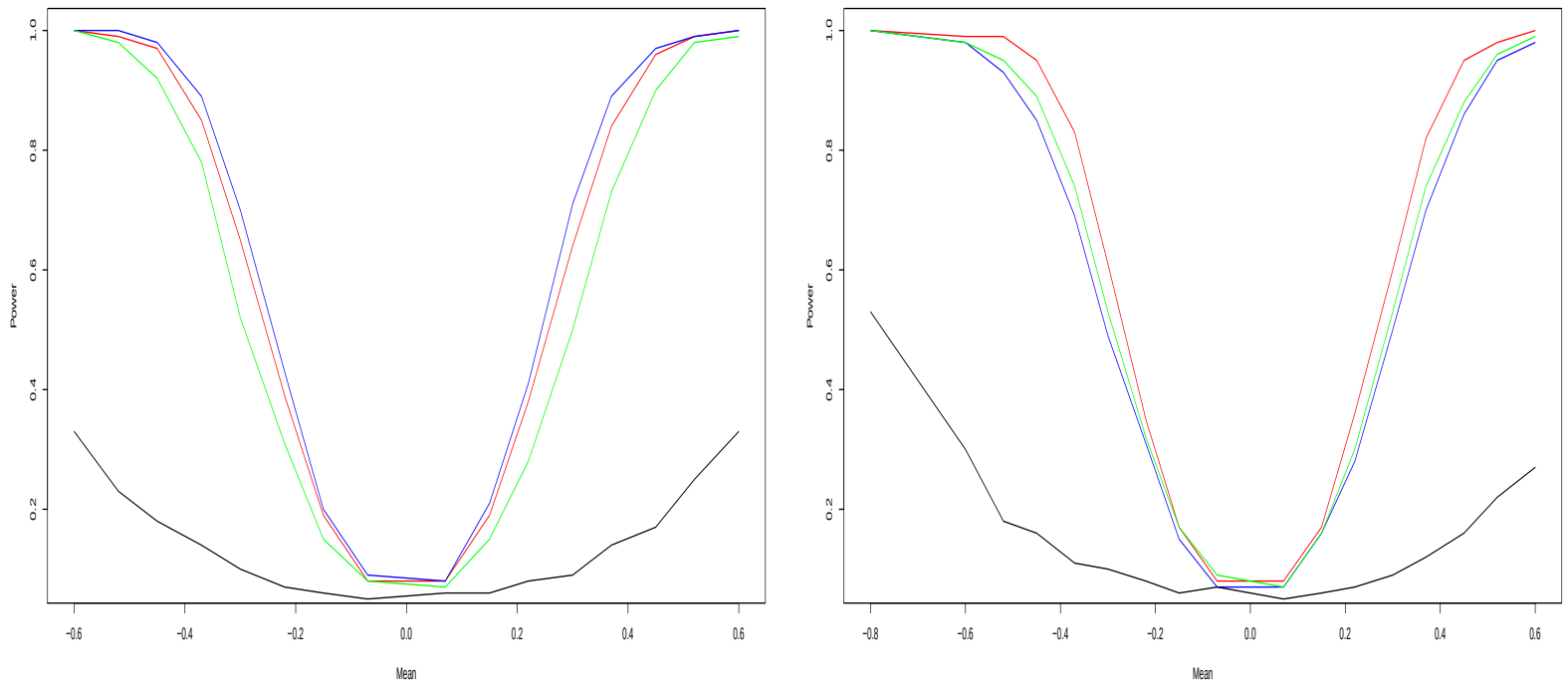


Figure: (Left panel) X_1, Y_1 are i.i.d. normal with mean 0 and μ respectively (and unit variance). $X_2, X_3 \sim X_1$, $Y_2, Y_3 \sim Y_1$ and $\mathbf{X} := (X_1, X_2, X_3)$. Similarly define $\mathbf{Y}$.
(Right panel) $\mathbf{U} := (U_1, U_2, U_3)$ and $\mathbf{V} := (V_1, V_2, V_3)$ where $U_i = \exp(X_i)$, $V_i = \exp(Y_i)$ and $X_1, X_2, X_3, Y_1, Y_2, Y_3$ has the same distribution as above.
Red - Rank energy, **Black - Crossmatch**, **Blue - Energy**, **Green - 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.

Asymptotic stabilization of critical values

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

Testing for mutual independence

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

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

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

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

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

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

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

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

$$-2\,\mathbb{E}\big[h(\mathbf{X}, \mathbf{X}')h(\mathbf{Y}, \mathbf{Y}'')\big] \geq 0$$

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

$$S_1 = \frac{1}{n^2}\sum_{i,j=1}^{n} h(\mathbf{X}_i, \mathbf{X}_j)h(\mathbf{Y}_i, \mathbf{Y}_j), \qquad S_3 = \frac{1}{n^3}\sum_{i,j,k=1}^{n} h(\mathbf{X}_i, \mathbf{X}_j)h(\mathbf{Y}_i, \mathbf{Y}_k),$$

$$S_2 = \Big(\frac{1}{n^2}\sum_{i,j=1}^{n} h(\mathbf{X}_i, \mathbf{X}_j)\Big)\Big(\frac{1}{n^2}\sum_{i,j=1}^{n} h(\mathbf{Y}_i, \mathbf{Y}_j)\Big)$$

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

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

$$\mathrm{dCov}_n(\{(\mathbf{X}_i, \mathbf{Y}_i)\}_{i=1}^n) > c_\alpha$$

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

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

$$\mathrm{dCov}_n(\{(\mathbf{X}_i, \mathbf{Y}_i)\}_{i=1}^n) > c_\alpha$$

- Critical value c_α depends on n, P_X, P_Y!

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

- Sample rank of $\mathbf{X}_i$: $\hat{\mathbf{R}}_n^X : \{\mathbf{X}_1, \ldots, \mathbf{X}_n\} \to \{\mathbf{c}_1^{(1)}, \ldots, \mathbf{c}_n^{(1)}\} \subset [0,1]^{d_1}$
- Sample rank of $\mathbf{Y}_i$: $\hat{\mathbf{R}}_n^Y : \{\mathbf{Y}_1, \ldots, \mathbf{Y}_n\} \to \{\mathbf{c}_1^{(2)}, \ldots, \mathbf{c}_n^{(2)}\} \subset [0,1]^{d_2}$
- **Rank distance cov.**: $\mathrm{RdCov}_n = \mathrm{dCov}_n\left(\left\{(\hat{\mathbf{R}}_n^X(\mathbf{X}_i), \hat{\mathbf{R}}_n^Y(\mathbf{Y}_i))\right\}_{i=1}^n\right)$

Distribution-freeness

$\mathbf{X}$ and $\mathbf{Y}$ 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 $\mathbf{c}_i^{(k)}$'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

Suppose: (i) $\mathbf{X}$ and $\mathbf{Y}$ are abs. cont., and

(ii) $\frac{1}{n}\sum_{i=1}^n \delta_{\mathbf{c}_i^{(k)}} \xrightarrow{d}$ 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 $\mathbf{c}_i^{(k)}$'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 $\mathbf{c}_i^{(k)}$'s have no effect for large n

Power

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

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

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)):

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

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. (2008)] which uses ideas from RKHS, ...

Summary

- Multivariate distribution-free nonparametric testing procedures
- Based on multivariate ranks defined using optimal transport
- Developed a general framework; other examples may include testing for multivariate symmetry, testing the equality of K-distributions, independence testing of K-vectors...
- The proposed tests are: (i) distribution-free and have good efficiency in general, (ii) are more powerful for distributions with heavy tails, and (iii) are robust to outliers & contamination
- Deb and S. (2019). `https://arxiv.org/pdf/1909.08733.pdf`
- Asymptotic efficiency? Other applications?

Thank you very much!

Questions?