Multivariate Rank-based Distribution-free Nonparametric Testing using Optimal Transport

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

Oberwolfach Workshop: "Statistics meets Machine Learning" (26 January – 1 February 2020)
Germany

Joint work with Nabarun Deb (Columbia U)

30 January, 2020

Preprint available at `https://arxiv.org/abs/1909.08733`

[1]Supported by NSF grant DMS-1712822

Outline

Outline

Multivariate 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.,

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

Multivariate 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), ...

Multivariate nonparametric testing

Testing for mutual independence

- $(\mathbf{X}, \mathbf{Y}) \sim P$ on $\mathbb{R}^{d_1} \times \mathbb{R}^{d_2}$; $d_1, d_2 \geq 1$
- **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.,

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

Multivariate nonparametric testing

Testing for mutual independence

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

We can also handle testing for K-vector/sample analogues of these problems and can also test for multivariate symmetry

Multivariate distribution-free nonparametric testing

- Two-sample GoF testing: $\mathrm{H}_0 : P_1 = P_2$ vs. $\mathrm{H}_1 : P_1 \neq P_2$
- Testing for independence: $\mathrm{H}_0 : \mathbf{X} \perp\!\!\!\perp \mathbf{Y}$ vs. $\mathrm{H}_1 : \mathbf{X} \not\perp\!\!\!\perp \mathbf{Y}$

Multivariate distribution-free nonparametric testing

- Two-sample GoF testing: $\mathrm{H}_0 : P_1 = P_2$ vs. $\mathrm{H}_1 : P_1 \neq P_2$
- Testing for independence: $\mathrm{H}_0 : \mathbf{X} \perp\!\!\!\perp \mathbf{Y}$ vs. $\mathrm{H}_1 : \mathbf{X} \not\!\perp\!\!\!\perp \mathbf{Y}$

Our contributions: summary

- Develop exactly distribution-free multivariate tests (i.e., null distributions of the test statistics are free of the underlying (unknown) data generating distributions, for all sample sizes)

Multivariate distribution-free nonparametric testing

- Two-sample GoF testing: $\mathrm{H}_0 : P_1 = P_2$ vs. $\mathrm{H}_1 : P_1 \neq P_2$
- Testing for independence: $\mathrm{H}_0 : \mathbf{X} \perp\!\!\!\perp \mathbf{Y}$ vs. $\mathrm{H}_1 : \mathbf{X} \not\perp\!\!\!\perp \mathbf{Y}$

Our contributions: summary

- Develop exactly distribution-free multivariate tests (i.e., null distributions of the test statistics are free of the underlying (unknown) data generating distributions, for all sample sizes)
- Consistent against all fixed alternatives (i.e., power of the test converges to 1 as sample size increases)

Multivariate distribution-free nonparametric testing

- Two-sample GoF testing: $\mathrm{H}_0 : P_1 = P_2$ vs. $\mathrm{H}_1 : P_1 \neq P_2$
- Testing for independence: $\mathrm{H}_0 : \mathbf{X} \perp\!\!\!\perp \mathbf{Y}$ vs. $\mathrm{H}_1 : \mathbf{X} \not\perp\!\!\!\perp \mathbf{Y}$

Our contributions: summary

- Develop exactly distribution-free multivariate tests (i.e., null distributions of the test statistics are free of the underlying (unknown) data generating distributions, for all sample sizes)
- Consistent against all fixed alternatives (i.e., power of the test converges to 1 as sample size increases)
- Computationally feasible ($O(n^3)$ algorithm)

Multivariate distribution-free nonparametric testing

- Two-sample GoF testing: $\mathrm{H}_0 : P_1 = P_2$ vs. $\mathrm{H}_1 : P_1 \neq P_2$
- Testing for independence: $\mathrm{H}_0 : \mathbf{X} \perp\!\!\!\perp \mathbf{Y}$ vs. $\mathrm{H}_1 : \mathbf{X} \not\perp\!\!\!\perp \mathbf{Y}$

Our contributions: summary

- Develop exactly distribution-free multivariate tests (i.e., null distributions of the test statistics are free of the underlying (unknown) data generating distributions, for all sample sizes)
- Consistent against all fixed alternatives (i.e., power of the test converges to 1 as sample size increases)
- Computationally feasible ($O(n^3)$ algorithm)

Most existing tests do not satisfy the above three desirable properties

Multivariate distribution-free nonparametric testing

- A general framework for multivariate distribution-free nonparametric testing based on ranks
- Multivariate ranks obtained using the theory of optimal transport [Hallin (2017), Chernozhukov et al. (2017), del Barrio et al. (2018), Ghosal and S. (2019), Deb and S. (2019), ...]

Multivariate distribution-free nonparametric testing

- A general framework for multivariate distribution-free nonparametric testing based on ranks

- Multivariate ranks obtained using the theory of optimal transport [Hallin (2017), Chernozhukov et al. (2017), del Barrio et al. (2018), Ghosal and S. (2019), Deb and S. (2019), ...]

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)], Wald-Wolfowitz runs test [Wald and Wolfowitz (1940)], Hoeffding's D-test [Hoeffding (1948)], etc. ...

Multivariate distribution-free nonparametric testing

- A general framework for multivariate distribution-free nonparametric testing based on ranks

- Multivariate ranks obtained using the theory of optimal transport [Hallin (2017), Chernozhukov et al. (2017), del Barrio et al. (2018), Ghosal and S. (2019), Deb and S. (2019), ...]

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)], Wald-Wolfowitz runs test [Wald and Wolfowitz (1940)], Hoeffding's D-test [Hoeffding (1948)], etc. ...

- In general, rank-based tests are: (i) distribution-free and have good efficiency, (ii) are more powerful for distributions with heavy tails, and (iii) are robust to outliers & contamination

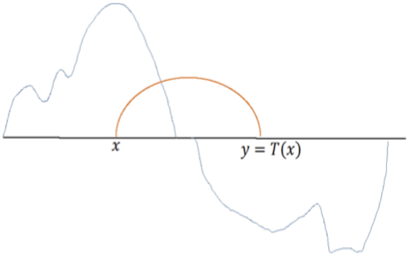
1. Multivariate Rank-based Distribution-free Nonparametric Testing
   - Nonparametric Testing: Introduction
   - Optimal Transport: Monge's Problem

2. Multivariate Two-sample Goodness-of-fit Testing
   - Distribution-free Testing
   - Asymptotic (Pitman) Efficiency

3. Testing for Independence Between Two Random Vectors
   - Distribution-free Testing

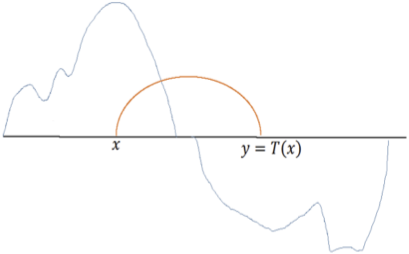
Optimal Transport: Monge's problem

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 (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 $X \sim \nu$; T transports ν to μ

Rank function as the optimal 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

Rank function as the optimal 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 $\mathbb{E}_\nu[X^2] < \infty$) the c.d.f. F is the optimal transport map as

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

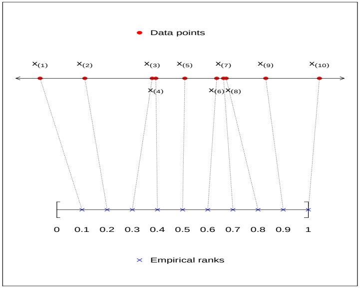
where

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

Sample rank: when $d = 1$

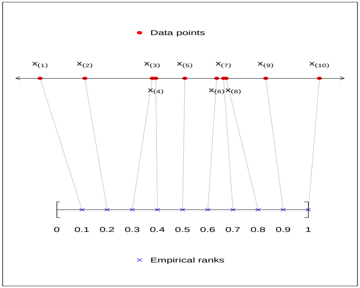
- **Data**: $X_1, \ldots, X_n$ iid ν (cont. distribution) on $\mathbb{R}$

- **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: when $d = 1$

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



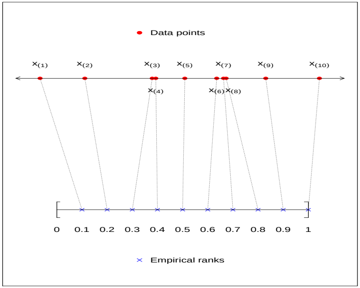
Sample rank map $\hat{R}_n$ is also a transport map that transports

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

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

Sample rank: 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 also a transport map that transports

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

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

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. $\mathbf{T}(\mathbf{X}) \stackrel{d}{=} \mathbf{U} \sim \text{Unif}([0,1]^d) \equiv \mu$

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

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

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 $\mathbb{E}_\nu \|\mathbf{X}\|^2 < \infty$, 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 $P_1 = P_2$

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 $\mathbb{E}_\nu \|\mathbf{X}\|^2 < \infty$, 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 unique $\mathbf{Q}(\cdot)$ s.t.

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

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 $\mathbb{E}_\nu \|\mathbf{X}\|^2 < \infty$, 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 unique $\mathbf{Q}(\cdot)$ s.t.

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

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

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

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

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

Suppose $\mathbf{X} \sim \nu$ abs. cont. on $\mathbb{R}^d$. Then $\exists$ ν-a.e. unique meas. mapping $\mathbf{R} : \mathbb{R}^d \to [0,1]^d$, transporting $\mathbf{X}$ to $\mathbf{U}$ (i.e., $\mathbf{R}(\mathbf{X}) \stackrel{d}{=} \mathbf{U}$), 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 $d = 1$).

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

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

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

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

Suppose $\mathbf{X} \sim \nu$ abs. cont. on $\mathbb{R}^d$. Then $\exists$ ν-a.e. unique meas. mapping $\mathbf{R} : \mathbb{R}^d \to [0,1]^d$, transporting $\mathbf{X}$ to $\mathbf{U}$ (i.e., $\mathbf{R}(\mathbf{X}) \stackrel{d}{=} \mathbf{U}$), of the form

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

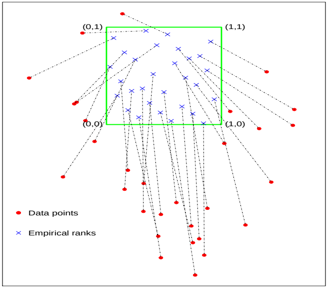
where $\varphi : \mathbb{R}^d \to \mathbb{R} \cup \{+\infty\}$ is a convex function (cf. when $d = 1$).

Moreover, when $\mathbb{E}_\nu\|\mathbf{X}\|^2 < \infty$, $\mathbf{R}(\cdot)$ 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)

- **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)



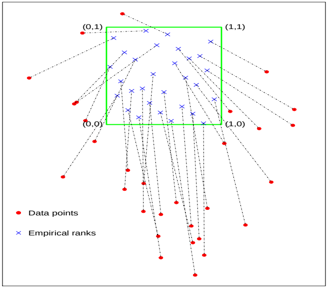
- Sample multivariate rank map is defined as the tranport map s.t.

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

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

- **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 $\{\mathbf{c}_1, \ldots, \mathbf{c}_n\}$.

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

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

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 $\frac{1}{n}\sum_{i=1}^n \delta_{\mathbf{c}_i} \xrightarrow{d} \mathrm{Unif}([0,1]^d)$, 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)]

Outline

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.,

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

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:

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

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

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

- 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)

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

Our proposal: 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

Our proposal: 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 $\mathrm{RE}_{m,n} > \kappa_\alpha^{(m,n)}$;
  $\kappa_\alpha^{(m,n)}$ is a universal threshold (free of $P_1 \equiv P_2$)

Our proposal: 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 $\mathrm{RE}_{m,n} > \kappa_\alpha^{(m,n)}$;
  $\kappa_\alpha^{(m,n)}$ is a universal threshold (free of $P_1 \equiv P_2$)

- The only other computationally feasible distribution-free test in this context was proposed by Rosenbaum (2005)

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

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

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

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

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

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 $\{\lambda_j \geq 0 : j \geq 1\}$,

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

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

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 $\{\lambda_j \geq 0 : j \geq 1\}$,

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

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 $\frac{m}{m+n} \to \lambda \in (0,1)$, 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 (under minimal assumptions)

Rank energy distance: Population version

- Assume $\frac{m}{m+n} \to \lambda \in (0,1)$
- $\mathbf{X} \sim P_1$ and $\mathbf{Y} \sim P_2$ (on $\mathbb{R}^d$); $\mathbf{Z} \sim \lambda P_1 + (1-\lambda) P_2$

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

- "Pooled" population rank map R_λ s.t. $\mathrm{R}_\lambda(\mathbf{Z}) \sim \mathrm{Uniform}([0,1]^d)$

Rank energy distance: Population version

- Assume $\frac{m}{m+n} \to \lambda \in (0,1)$
- $\mathbf{X} \sim P_1$ and $\mathbf{Y} \sim P_2$ (on $\mathbb{R}^d$); $\mathbf{Z} \sim \lambda P_1 + (1-\lambda) P_2$

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

- "Pooled" population rank map R_λ s.t. $\mathrm{R}_\lambda(\mathbf{Z}) \sim \mathrm{Uniform}([0,1]^d)$
- **Rank energy distance**: $\mathrm{RE}_\lambda^2(P_1, P_2) := \mathrm{E}^2(R_\lambda(\mathbf{X}), R_\lambda(\mathbf{Y}))$
- **Result**: $\mathrm{RE}_\lambda = 0$ iff $P_1 = P_2$ provided P_1, P_2 are abs. cont.

Rank energy distance: Population version

- Assume $\frac{m}{m+n} \to \lambda \in (0,1)$
- $\mathbf{X} \sim P_1$ and $\mathbf{Y} \sim P_2$ (on $\mathbb{R}^d$); $\mathbf{Z} \sim \lambda P_1 + (1-\lambda)P_2$

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

- "Pooled" population rank map R_λ s.t. $\mathrm{R}_\lambda(\mathbf{Z}) \sim \mathrm{Uniform}([0,1]^d)$
- **Rank energy distance**: $\mathrm{RE}^2_\lambda(P_1, P_2) := \mathrm{E}^2(R_\lambda(\mathbf{X}), R_\lambda(\mathbf{Y}))$
- **Result**: $\mathrm{RE}_\lambda = 0$ iff $P_1 = P_2$ provided P_1, P_2 are abs. cont.

Almost sure convergence

If $\frac{1}{n}\sum_{i=1}^n \delta_{\mathbf{c}_i} \xrightarrow{d} \mathrm{Uniform}([0,1]^d)$, then

$$\mathrm{RE}^2_{m,n} \xrightarrow{a.s.} \mathrm{RE}^2_\lambda(P_1, P_2).$$

When $d = 1$

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

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

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

When $d = 1$

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

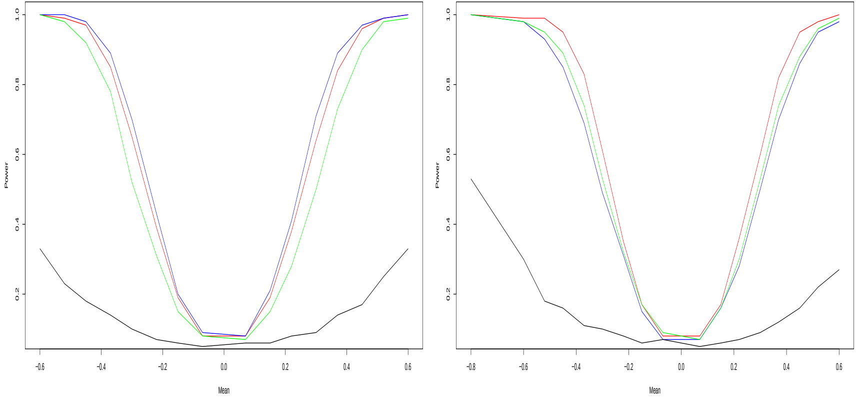
$$\frac{1}{2}\mathrm{RE}^2_{m,n} = \int \left\{\mathbb{F}^X_m(t) - \mathbb{F}^Y_n(t)\right\}^2 d\mathbb{F}_{m+n}(t)$$

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



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

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

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.

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

Asymptotic (Pitman) efficiency

$\mathbf{X}_1, \ldots, \mathbf{X}_m \overset{iid}{\sim} \mathrm{P}_{\theta_1}$ & $\mathbf{Y}_1, \ldots, \mathbf{Y}_n \overset{iid}{\sim} \mathrm{P}_{\theta_2}$; $N = m + n$; $m/N = \lambda \in (0,1)$

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

Asymptotic (Pitman) efficiency

$\mathbf{X}_1, \ldots, \mathbf{X}_m \overset{iid}{\sim} \mathrm{P}_{\theta_1}$ & $\mathbf{Y}_1, \ldots, \mathbf{Y}_n \overset{iid}{\sim} \mathrm{P}_{\theta_2}$; $N = m + n$; $m/N = \lambda \in (0, 1)$

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

Pitman efficiency

- Fix α (size) and $\gamma > \alpha$ (power); two test functions — T_N and S_N

Asymptotic (Pitman) efficiency

$\mathbf{X}_1, \ldots, \mathbf{X}_m \overset{iid}{\sim} \mathrm{P}_{\theta_1}$ & $\mathbf{Y}_1, \ldots, \mathbf{Y}_n \overset{iid}{\sim} \mathrm{P}_{\theta_2}$; $N = m + n$; $m/N = \lambda \in (0,1)$

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

Pitman efficiency

- Fix α (size) and $\gamma > \alpha$ (power); two test functions — T_N and S_N
- $K(T_N)$ denotes minimum number of samples such that:

$$\mathbb{E}_{\mathrm{H}_0}[T_N] \leq \alpha \qquad \text{and} \qquad \mathbb{E}_{\mathrm{H}_1}[T_N] \geq \gamma$$

- The Pitman efficiency of S_N w.r.t. to T_N is given by

$$\lim_{N \to \infty} \frac{K(T_N)}{K(S_N)}$$

Asymptotic (Pitman) efficiency

$\mathbf{X}_1, \ldots, \mathbf{X}_m \overset{iid}{\sim} \mathrm{P}_{\theta_1}$ & $\mathbf{Y}_1, \ldots, \mathbf{Y}_n \overset{iid}{\sim} \mathrm{P}_{\theta_2}$; $\quad N = m + n$; $\; m/N = \lambda \in (0,1)$

Test: $\mathrm{H}_0 : \theta_2 = \theta_1$ versus $\mathrm{H}_1 : \theta_2 = \theta_1 + \mathbf{h}N^{-1/2}$; $\quad \mathbf{h} \neq 0 \in \mathbb{R}^p$

Pitman efficiency

- Fix α (size) and $\gamma > \alpha$ (power); two test functions — T_N and S_N
- $K(T_N)$ denotes minimum number of samples such that:

$$\mathbb{E}_{\mathrm{H}_0}[T_N] \leq \alpha \qquad \text{and} \qquad \mathbb{E}_{\mathrm{H}_1}[T_N] \geq \gamma$$

- The Pitman efficiency of S_N w.r.t. to T_N is given by

$$\lim_{N \to \infty} \frac{K(T_N)}{K(S_N)}$$

In general, a test has non-trivial Pitman efficiency if it has non-trivial asymptotic power for testing against the above local alternatives

Asymptotic efficiency for rank energy test

Want to test: $\mathrm{H}_0 : \theta_2 = \theta_1$ versus $\mathrm{H}_1 : \theta_2 = \theta_1 + \mathbf{h}N^{-1/2}$; $\mathbf{h} \neq 0 \in \mathbb{R}^p$

Theorem [Deb, Bhattacharya and S. (2020+)]

Assume regularity conditions; e.g., $\{P_\theta\}$ satisfies DQM. Then, under $\mathrm{H}_1 : \theta_2 = \theta_1 + \mathbf{h}N^{-1/2}$,

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

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

Asymptotic efficiency for rank energy test

Want to test: $\mathrm{H}_0 : \theta_2 = \theta_1$ versus $\mathrm{H}_1 : \theta_2 = \theta_1 + \mathbf{h}N^{-1/2}$; $\mathbf{h} \neq 0 \in \mathbb{R}^p$

Theorem [Deb, Bhattacharya and S. (2020+)]

Assume regularity conditions; e.g., $\{P_\theta\}$ satisfies DQM. Then, under $\mathrm{H}_1 : \theta_2 = \theta_1 + \mathbf{h}N^{-1/2}$,

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

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

- Let T_N denote the test based on the rank energy statistic $\mathrm{RE}^2_{m,n}$
- Then, $\lim_{N\to\infty} \mathbb{E}_{\mathrm{H}_0}[T_N] = \alpha$ and $\lim_{\|\mathbf{h}\|\to\infty} \lim_{N\to\infty} \mathbb{E}_{\mathrm{H}_1}[T_N] = 1$
- Therefore, rank energy test does distinguish between the null and the alternative (has non-trivial power) at the contiguous scale

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 energy test 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 $\mathbf{h}$:

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

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

What about other asymptotically distribution-free tests?

- Many other graph-based[a] (asymptotically distribution-free) tests are also asymptotically powerless [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

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

Testing for mutual independence

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

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

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

Testing for mutual independence

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

- **Data**: $\{(\mathbf{X}_i, \mathbf{Y}_i) : 1 \leq i \leq n\}$ iid P

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

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

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

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

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

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

- **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! (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! (can use permutation test)

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 $\mathbf{Y}_i$: $\hat{\mathbf{R}}_n^Y : \{\mathbf{Y}_1, \ldots, \mathbf{Y}_n\} \to \{\mathbf{c}_1^{(2)}, \ldots, \mathbf{c}_n^{(2)}\} \subset [0,1]^{d_2}$

- **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)

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 $\mathrm{RdCov}_n > \kappa_\alpha^{(n)}$

- 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 [Deb and S. (2019)]

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

(ii) $\frac{1}{n}\sum_{i=1}^n \delta_{\mathbf{c}_i^{(k)}} \xrightarrow{d} \text{Uniform}([0,1]^{d_k})$, 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

- 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 [Deb and S. (2019)]

Suppose: (i) **X** and **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,

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

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)]:

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

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)]:

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

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)]:

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

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 nonparametric testing procedures
- Based on multivariate ranks defined using optimal transport

Summary

- Multivariate distribution-free nonparametric testing procedures
- Based on multivariate ranks defined using optimal transport
- Proposed a general framework, other examples may include testing for symmetry, testing the equality of K-distributions, independence testing of K-vectors, ...
- Tuning-free, computationally feasible procedures

Summary

- Multivariate distribution-free nonparametric testing procedures
- Based on multivariate ranks defined using optimal transport
- Proposed a general framework, other examples may include testing for symmetry, testing the equality of K-distributions, independence testing of K-vectors, ...
- Tuning-free, computationally feasible procedures
- 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

Thank you very much!

Questions?