Multivariate rank-based distribution-free nonparametric testing using optimal transportation

Nabarun Deb
Department of Statistics, Columbia University
Berkeley-Columbia Meeting in Engineering and Statistics, Feb 28, 2020.

Joint work with Dr. Bodhisattva Sen, ongoing work with Dr. Bhaswar Bhattacharya

Multivariate distribution-free nonparametric testing

Consider the following nonparametric hypothesis testing problem:

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.,

$$H_0 : P_1 = P_2 \qquad \text{versus} \qquad H_1 : P_1 \neq P_2.$$

- When $d = 1$: **Smirnov (1939)**, Smirnov (1939), Wald and Wolfowitz (1940), Mann and Whitney (1947), Wilcoxon (1947).

- When $d > 1$: Weiss (1960), Anderson (1962), Schilling (1986), Rosenbaum (2005), Gretton et al. (2012), Székely and Rizzo (2013), Biswas et al. (2014), Chen and Friedman (2017), Li and Yuan (2019).

- Exact distribution-freeness: A statistic is said to be exactly distribution-free if its null distribution is universal (free of the underlying data generation mechanism).

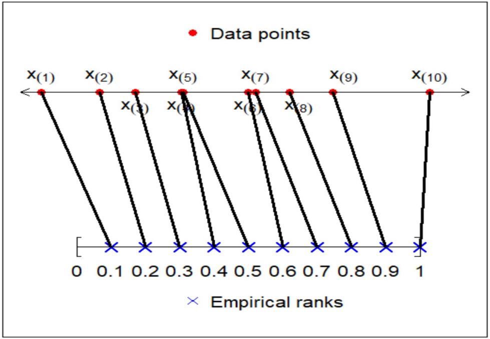
- The tests should be consistent under minimal assumptions and also be computationally feasible.

- We can also handle testing for mutual independence and testing for multivariate symmetry.

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 $\{\frac{1}{n}, \frac{2}{n}, \ldots, \frac{n}{n}\}$.

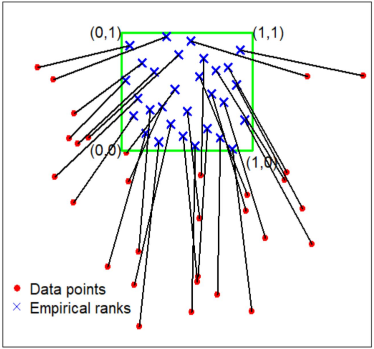


$$\hat{\sigma} := \underset{\sigma=(\sigma(1),\ldots,\sigma(n)) \in S_n}{\arg\min} \sum_{i=1}^{n} \left| X_i - \frac{\sigma(i)}{n} \right|^2.$$

where S_n is the set of all permutations of $\{1, 2, \ldots, n\}$.

Multivariate ranks ($d \geq 1$)

- **Data**: $\mathbf{X}_1, \ldots, \mathbf{X}_n$ iid on $\mathbb{R}^d$ (abs. cont. distribution)
  Empirical rank map assigns $\{\mathbf{X}_1, \ldots, \mathbf{X}_n\} \to \{\mathbf{c}_1, \ldots, \mathbf{c}_n\} \subset [0,1]^d$
  — sequence of "uniform-like" points (quasi-Monte Carlo sequence)



$$\hat{\sigma} := \operatorname*{arg\,min}_{\sigma=(\sigma(1),\ldots,\sigma(n)) \in S_n} \sum_{i=1}^{n} \|\mathbf{X}_i - \mathbf{c}_{\sigma(i)}\|^2$$

- Assignment problem (can be reduced to a linear program; can be exactly solved using $O(n^3)$ Hungarian algorithm; some approximations in **Agarwal and Sharathkumar (2014)**).

- **Data**: $\mathbf{X}_1, \ldots, \mathbf{X}_n$ i.i.d. on $\mathbb{R}^d$ (abs. cont. distribution)
- $\{\mathbf{c}_1, \ldots, \mathbf{c}_n\} \subset [0,1]^d$ — sequence of "uniform-like" points
- $$\hat{\sigma} := \underset{\sigma=(\sigma(1),\ldots,\sigma(n)) \in S_n}{\arg\min} \sum_{i=1}^n \|\mathbf{X}_i - \mathbf{c}_{\sigma(i)}\|^2$$
- **Sample rank map**: $\hat{\mathbf{R}}_n : \{\mathbf{X}_1, \ldots, \mathbf{X}_n\} \to \{\mathbf{c}_1, \ldots, \mathbf{c}_n\}$ where

$$\hat{\mathbf{R}}_n(\mathbf{X}_i) = \mathbf{c}_{\hat{\sigma}(i)}, \qquad i = 1, \ldots, n$$

Distribution-free property (Similar result in Hallin 2017)

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

Multivariate two-sample goodness-of-fit test

Testing for equality of 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 $\mathbb{R}^d$, **both absolutely continuous**, $d \geq 1$
- Test if the two-samples come from the same distribution, i.e.,

$$H_0 : P_1 = P_2 \qquad \text{versus} \qquad H_1 : P_1 \neq P_2$$

- Start with a "good" test, say the energy statistic (Székely and Rizzo, 2013).
- Suppose $\mathbf{X}, \mathbf{X}' \overset{iid}{\sim} P_1$, $\mathbf{Y}, \mathbf{Y}' \overset{iid}{\sim} P_2$ and set $h(\mathbf{s}, \mathbf{t}) := \|\mathbf{s} - \mathbf{t}\|$, then energy distance between P_1 and P_2:

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

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

- **E-statistic**: $E^2_{m,n}\left(\{\mathbf{X}_i\}_{i=1}^m, \{\mathbf{Y}_j\}_{j=1}^n\right) := 2A - B - C$ 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**: Reject H_0 if $E_{m,n}\left(\{\mathbf{X}_i\}_{i=1}^m, \{\mathbf{Y}_j\}_{j=1}^n\right) > \kappa_\alpha$

- $E^2_{m,n} \xrightarrow{a.s.} E^2$ under appropriate moment assumptions.

- Critical value κ_α depends on $P_1 = P_2$!

Proposed statistic

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

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 $\mathrm{RE}_{m,n} > \kappa_\alpha$ (universal threshold, free of $P_1 = P_2$).
- The only other computationally feasible distribution-free test in this context was proposed in Rosenbaum (2005). Another distribution-free test from Biswas et al. (2014) is NP-hard.

Simplification for $d = 1$

$\mathrm{RE}^2_{m,n}$ is exactly equivalent (constant multiple of) to the two-sample Cramér-von Mises statistic.

Limiting distribution under H_0

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

Then, under H_0, $\exists$ a universal distribution $\mathbb{D}_d$ s.t.

$$\frac{mn}{m+n}\mathrm{RE}^2_{m,n} \xrightarrow{d} \sum_{j=1}^{\infty} \lambda_j Z_j^2 \qquad \text{as} \quad \min\{m,n\} \to \infty \quad \text{where } \lambda_j \geq 0.$$

Power

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

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

In fact, $\mathrm{RE}_{m,n} \xrightarrow{a.s.} 0$ a.s. iff $P_1 = P_2$.

Proposed test has asymptotic power 1, against all fixed alternatives

Pitman asymptotics

- Consider $\mathbf{X}_1, \ldots, \mathbf{X}_m \sim \mathrm{P}_{\theta_1}$ and $\mathbf{Y}_1, \ldots, \mathbf{Y}_n \sim \mathrm{P}_{\theta_2}$, with $m/(m+n) = \lambda \in (0,1)$. We want to test:

$$\mathrm{H}_0 : \theta_2 - \theta_1 = 0 \qquad \text{versus} \qquad \mathrm{H}_1 : \theta_2 - \theta_1 = \mathbf{h}(m+n)^{-1/2}.$$

- Fix a level parameter α and assume $m/(m+n) = \lambda \in (0,1)$.
- Given a statistic $T_{m,n}$, one is interested in showing that:

$$\mathbb{P}_{\mathrm{H}_1}(T_{m,n} \text{ rejects } \mathrm{H}_0) \longrightarrow \alpha + g(\mathbf{h})$$

where $g(\mathbf{h}) > 0$ if $\mathbf{h} \neq 0$.

Crossmatch test (Rosenbaum 2005)

Pitman asymptotics for crossmatch test (Rosenbaum 2005)

Consider the testing set-up from before (with additional regularity assumptions). Then, for any **h**, we have:

$$\lim_{m,n\to\infty} \mathbb{P}_{\mathrm{H}_1}(T_{m,n} \text{ rejects } \mathrm{H}_0) = \alpha.$$

- Therefore, crossmatch test does not distinguish between the null and the alternative at the contiguous scale.
- The same phenomena happens for many other graph-based asymptotically distribution-free tests, see Bhattacharya 2019, Theorem 3.1

Rank energy test

Efficiency for rank energy test

Consider the testing set-up from before (with additional regularity assumptions). Then, for any **h**, we have:

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

where $\tilde{Z}_j^2$ has a non-central chi-squared distribution with non-centrality parameter depending on **h**. In particular,

$$\lim_{m,n\to\infty} \mathbb{P}_{\mathrm{H}_1}(T_{m,n} \text{ rejects } \mathrm{H}_0) > \alpha.$$

- Therefore, rank energy test **does** distinguish between the null and the alternative at the contiguous scale.
- In particular, the Pitman efficiency of rank energy test with respect to the crossmatch test is therefore infinite.

Asymptotic stabilization

	(100)	(300)	(500)	(700)	(900)
0.05	0.39	0.40	0.39	0.40	0.40
0.1	0.36	0.36	0.36	0.36	0.36

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

	(100)	(300)	(500)	(700)	(900)
0.05	1.37	1.38	1.38	1.38	1.38
0.1	1.34	1.35	1.35	1.35	1.35

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

Summary

- Multivariate distribution-free nonparametric testing procedures

- Based on multivariate ranks defined using optimal transportation (see Chernozhukhov et al. (2017), Hallin (2019).

- Proposed a general framework, other examples may include testing for symmetry, testing the equality of K-distributions, independence testing ...

- 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

- The corresponding paper — https://arxiv.org/pdf/1909.08733.pdf.


The End

Power plot with varying location parameter

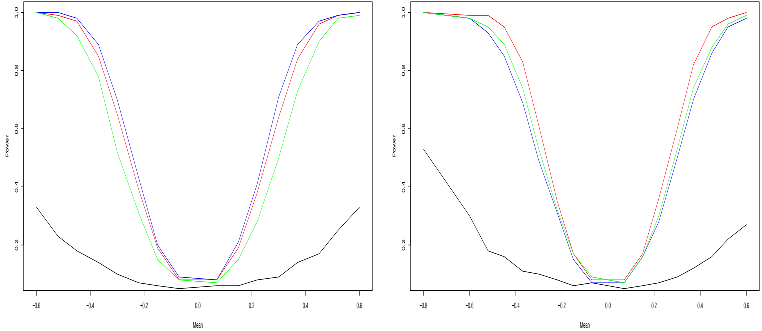


Figure: (Left panel) X_1, Y_1 are i.i.d. normal with mean 0 and μ respectively (and unit variance). $X_2, X_3 \sim X_1$, $Y_2, Y_3 \sim Y_1$ and $\mathbf{X} := (X_1, X_2, X_3)$. Similarly define $\mathbf{Y}$.
(Right panel) $\mathbf{U} := (U_1, U_2, U_3)$ and $\mathbf{V} := (V_1, V_2, V_3)$ where $U_i = \exp(X_i)$, $V_i = \exp(Y_i)$ 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

	(RB)	(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 RB - Rosenbaum's crossmatch test (Rosenbaum, 2005), HHG - Heller, Heller and Gorfine (Heller et al., 2013), En - energy statistic (Székely and Rizzo, 2013).

Rank functions as transport maps: When $d = 1$

- $X \sim F$ on $\mathbb{R}$, F abs. cont. c.d.f.
- **Rank**: The rank of $x \in \mathbb{R}$ is $F(x)$ (aka the c.d.f. at x)
- **Property**: $F(X) \sim \text{Uniform}([0,1])$
- Thus, F transports the distribution of X to $U \sim \text{Uniform}([0,1])$
- In fact, if $\mathbb{E}[X^2] < \infty$, c.d.f. F is the optimal transport map as

$$F = \underset{T:T(X)\overset{d}{=}U}{\arg\min}\ \mathbb{E}|X - T(X)|^2$$

- Sample rank map (aka empirical c.d.f.) is also a transport map:

$$\hat{R}_n := \underset{\sigma \in S_n}{\arg\min} \frac{1}{n} \sum_{i=1}^{n} \left| X_i - \frac{\sigma(i)}{n} \right|^2 = \underset{T}{\arg\min} \frac{1}{n} \sum_{i=1}^{n} |X_i - T(X_i)|^2$$

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

Multivariate rank functions as transport maps

- $\mathbf{X} \sim \nu$; ν is a probability measure in $\mathbb{R}^d$ (abs. cont.)
- $\mathbf{U} \sim$ Uniform$([0,1]^d)$

- **Goal**: Find the "optimal" transport map $\mathbf{T}$ s.t. $\mathbf{T}(\mathbf{X}) \stackrel{d}{=} \mathbf{U}$

- If $\mathbb{E}\|\mathbf{X}\|^2 < \infty$, the population rank function $\mathbf{R}(\cdot)$ is the transport map s.t.

$$\mathbf{R} := \operatorname*{arg\,min}_{\mathbf{T}:\mathbf{T}(\mathbf{X})\stackrel{d}{=}\mathbf{U},\, \mathbf{X}\sim\nu} \mathbb{E}\|\mathbf{X} - \mathbf{T}(\mathbf{X})\|^2$$

- **Data**: $\mathbf{X}_1, \ldots, \mathbf{X}_n$ iid ν (abs. cont.) on $\mathbb{R}^d$
- $\{\mathbf{c}_1, \ldots, \mathbf{c}_n\} \subset [0,1]^d$ — sequence of "uniform-like" points

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

$$\hat{\mathbf{R}}_{\mathbf{n}} = \operatorname*{arg\,min}_{\sigma\in S_n} \frac{1}{n}\sum_{i=1}^{n} \|\mathbf{X}_i - \mathbf{c}_{\sigma(i)}\|^2 \equiv \operatorname*{arg\,min}_{\mathbf{T}} \frac{1}{n}\sum_{i=1}^{n} \|\mathbf{X}_i - \mathbf{T}(\mathbf{X}_i)\|^2$$

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

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

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

- Even when $\mathbb{E}\|\mathbf{X}\|^2 = +\infty$, population rank function $\mathbf{R}(\cdot)$ can also be defined More details

- Sample multivariate rank map $\hat{\mathbf{R}}_{\mathbf{n}}(\cdot)$ is defined as

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

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

Regularity: L_2-convergence [Deb and S. (2019)]

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

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

Result gives the required regularity of the empirical multivariate rank map

Population version

Assume $m/(m+n) = \lambda \in (0,1)$.

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

- Joint rank map: The "pooled" population rank map:

$$\mathrm{R}_\lambda : \mathrm{R}_\lambda(\mathbf{Z}) \sim \mathrm{Uniform}([0,1]^d)$$

  where $\mathbf{Z} \sim \lambda P_1 + (1-\lambda) P_2$.

- **Rank energy**: $\mathrm{RE}^2_\lambda(P_1, P_2) := E^2(R_\lambda(\mathbf{X}), R_\lambda(\mathbf{Y}))$.

- $\mathrm{RE}_\lambda = 0$ iff $P_1 = P_2$ provided P_1, P_2 are absolutely continuous.

- Our general principle could have been used with any other procedure for testing equality of distributions, e.g., the MMD statistic [Gretton et al. (2008)] which uses ideas from RKHS, ...

- For $d = 1$, we prove that $\mathrm{RE}^2_{m,n}$ and RE^2_λ are exactly equivalent to the sample and population two-sample Cramér-von Mises statistic.

Pitman efficiency

- Consider $\mathbf{X}_1, \ldots, \mathbf{X}_n \sim \mathrm{P}_{\theta_1}$ and $\mathbf{Y}_1, \ldots, \mathbf{Y}_m \sim \mathrm{P}_{\theta_2}$, with $m/(m+n) = \lambda \in (0, 1)$. We want to test:

$$\mathrm{H}_0 : \theta_2 - \theta_1 = 0 \qquad \text{versus} \qquad \mathrm{H}_1 : \theta_2 - \theta_1 = \mathbf{h}(m+n)^{-1/2}.$$

- Fix α (size) and $\gamma > \alpha$ (power).
- Two test functions $T_{m,n}$ and $S_{m,n}$.
- $K(T_{m,n})$ denotes minimum number of samples such that:

$$\mathbb{E}_{\mathrm{H}_0}(T_{m,n}) \le \alpha \qquad \text{and} \qquad \mathbb{E}_{\mathrm{H}_1}(T_{m,n}) \ge \gamma.$$

- The Pitman efficiency of $S_{m,n}$ with respect to $T_{m,n}$ is given by

$$\lim_{m+n \to \infty} \frac{K(T_{m,n})}{K(S_{m,n})}.$$