Optimal transport in statistics and Pitman efficient multivariate distribution-free testing

Nabarun Deb
Department of Statistics, Columbia University

Kantorovich Initiative Retreat
March 18, 2022

Multivariate distribution-free nonparametric testing

Consider the following nonparametric hypothesis testing problem:

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

- **Data**: $\{X_i\}_{i=1}^m$ iid P on $\mathbb{R}^d$; $\{Y_j\}_{j=1}^n$ iid Q on $\mathbb{R}^d$, $d \geq 1$
- Test if the two-samples came from the same distribution, i.e.,

$$H_0 : P = Q \qquad \text{versus} \qquad H_1 : P \neq Q$$

Multivariate distribution-free nonparametric testing

Consider the following nonparametric hypothesis testing problem:

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

- **Data**: $\{X_i\}_{i=1}^m$ iid P on $\mathbb{R}^d$; $\{Y_j\}_{j=1}^n$ iid Q on $\mathbb{R}^d$, $d \geq 1$
- Test if the two-samples came from the same distribution, i.e.,

$$H_0 : P = Q \qquad \text{versus} \qquad H_1 : P \neq Q$$

- When $d = 1$: Student's t-test (1908), Wilcoxon rank-sum (1947), Cramér-von Mises (1928), Wald and Wolfowitz (1940), Mann and Whitney (1947), Kolmogorov-Smirnov (1939)
- When $d > 1$: Hotelling's T^2-statistic (1931), Weiss (1960), Anderson (1962), Friedman and Raksky (1979), Schilling (1986), Rosenbaum (2005), **Gretton et al. (2012), Székely and Rizzo (2013),** Biswas et al. (2014), Li and Yuan (2019)

Multivariate distribution-free nonparametric testing

Consider the following nonparametric hypothesis testing problem:

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

- **Data**: $\{X_i\}_{i=1}^m$ iid P on $\mathbb{R}^d$; $\{Y_j\}_{j=1}^n$ iid Q on $\mathbb{R}^d$, $d \geq 1$
- Test if the two-samples came from the same distribution, i.e.,

$$H_0 : P = Q \qquad \text{versus} \qquad H_1 : P \neq Q$$

- When $d = 1$: Student's t-test (1908), Wilcoxon rank-sum (1947), Cramér-von Mises (1928), Wald and Wolfowitz (1940), Mann and Whitney (1947), Kolmogorov-Smirnov (1939)
- When $d > 1$: Hotelling's T^2-statistic (1931), Weiss (1960), Anderson (1962), Friedman and Raksky (1979), Schilling (1986), Rosenbaum (2005), **Gretton et al. (2012), Székely and Rizzo (2013),** Biswas et al. (2014), Li and Yuan (2019)

When $d = 1$

- **Two-sample t-test**: Compares $\bar{X}_m$ and $\bar{Y}_n$

When $d = 1$

- **Two-sample t-test**: Compares $\bar{X}_m$ and $\bar{Y}_n$
- Reject if the statistic is larger than the $(1 - \alpha)$-th quantile of t_{m+n-2} (or use a permutation test)
- Approximate (not valid for small sample sizes) level α test, requires additional moment assumptions

When $d = 1$

- **Two-sample t-test**: Compares $\bar{X}_m$ and $\bar{Y}_n$
- Reject if the statistic is larger than the $(1-\alpha)$-th quantile of t_{m+n-2} (or use a permutation test)
- Approximate (not valid for small sample sizes) level α test, requires additional moment assumptions

- **Distribution-free tests:** Null distribution of the test statistic T_n is universal, i.e., $\mathbb{P}(T_n \geq c_{n,\alpha}) = \alpha$ where $c_{n,\alpha}$, the deterministic rejection threshold can be obtained before observing the data

When $d = 1$

- **Two-sample *t*-test**: Compares $\bar{X}_m$ and $\bar{Y}_n$
- Reject if the statistic is larger than the $(1-\alpha)$-th quantile of t_{m+n-2} (or use a permutation test)
- Approximate (not valid for small sample sizes) level α test, requires additional moment assumptions

- **Distribution-free tests:** Null distribution of the test statistic T_n is universal, i.e., $\mathbb{P}(T_n \geq c_{n,\alpha}) = \alpha$ where $c_{n,\alpha}$, the deterministic rejection threshold can be obtained before observing the data
- They are exact tests and valid for all sample sizes

When $d = 1$

- **Two-sample *t*-test**: Compares $\bar{X}_m$ and $\bar{Y}_n$

- Reject if the statistic is larger than the $(1-\alpha)$-th quantile of t_{m+n-2} (or use a permutation test)

- Approximate (not valid for small sample sizes) level α test, requires additional moment assumptions

- **Distribution-free tests:** Null distribution of the test statistic T_n is universal, i.e., $\mathbb{P}(T_n \geq c_{n,\alpha}) = \alpha$ where $c_{n,\alpha}$, the deterministic rejection threshold can be obtained before observing the data

- They are exact tests and valid for all sample sizes

- Based on univariate ranks — advent of classical nonparametrics

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

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

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

- WRS test is distribution-free and exact for all F continuous

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

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

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

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

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

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

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

- WRS test is distribution-free and exact for all F continuous
- WRS test has 0.95 Pitman 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.)

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

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

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

- WRS test is distribution-free and exact for all F continuous
- WRS test has 0.95 Pitman 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.)
- Non-trivial efficiency lower bound of 1 w.r.t. t-test [Chernoff and Savage (1958)] when the following revised statistic is used:

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

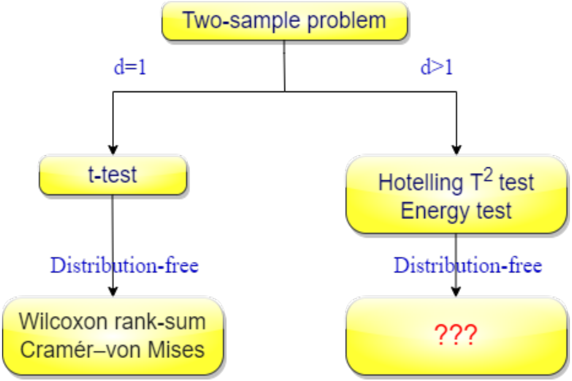
Generalize distribution-freeness, efficiency to multivariate data

Question

Can we construct multivariate nonparametric distribution-free tests?

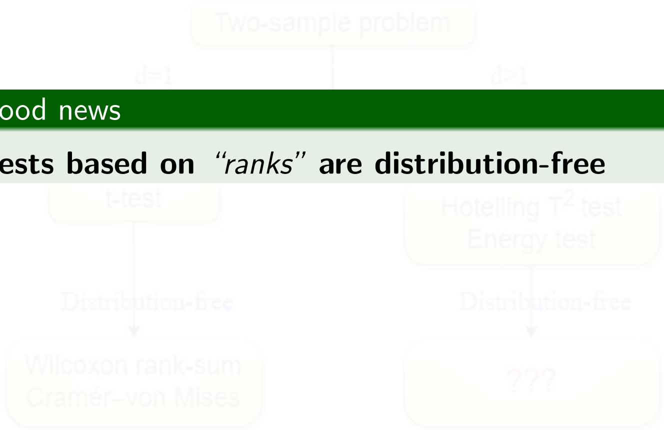
Question

Can we construct multivariate nonparametric distribution-free tests?



Question

Can we construct multivariate nonparametric distribution-free tests?

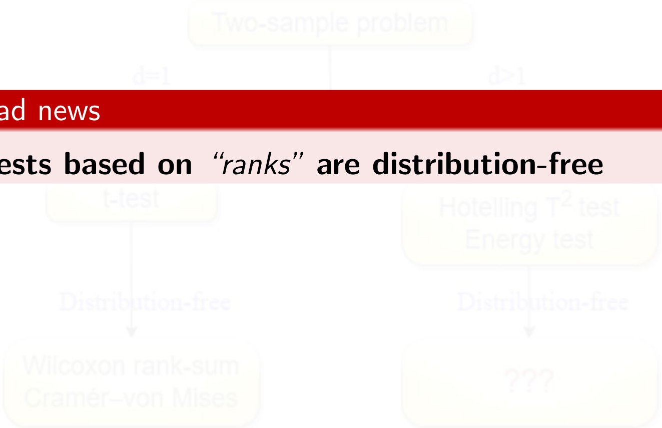


Good news

Tests based on *"ranks"* are distribution-free

Question

Can we construct multivariate nonparametric distribution-free tests?



Bad news

Tests based on *“ranks”* **are distribution-free**

Question

Can we construct multivariate nonparametric distribution-free tests?

Bad news

Tests based on *"ranks"* are distribution-free

How do we define multivariate ranks which lead to distribution-free tests?

What about their statistical efficiency?

Wilcoxon rank-sum
Cramér–von Mises

???

Question

Can we construct multivariate nonparametric distribution-free tests?

Bad news

Tests based on *"ranks"* **are distribution-free**

How do we define multivariate ranks which lead to distribution-free tests?

What about their statistical efficiency?

Optimal transport!

Wilcoxon rank-sum
Cramér–von Mises

???

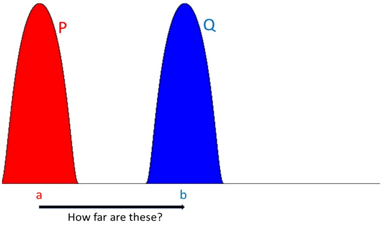
Outline

1. A (very) brief introduction to optimal transport
2. Multivariate ranks using optimal transport
3. Multivariate distribution-free tests using optimal transport
    - Rank Hotelling T^2 test and Pitman efficiency
    - Pitman efficiency, comparison with Hotelling T^2

Outline

1. A (very) brief introduction to optimal transport
2. Multivariate ranks using optimal transport
3. Multivariate distribution-free tests using optimal transport
   - Rank Hotelling T^2 test and Pitman efficiency
   - Pitman efficiency, comparison with Hotelling T^2

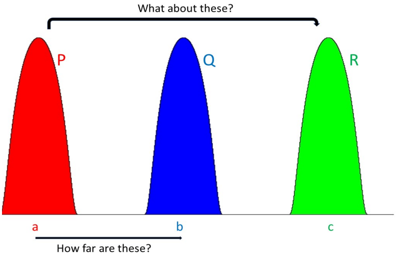
Optimal (measure) transportation



$$KL(P||Q) = \int \log\left(\frac{p}{q}\right) p = \infty$$

$$TV(P, Q) = \frac{1}{2} \int |p - q| = 1$$

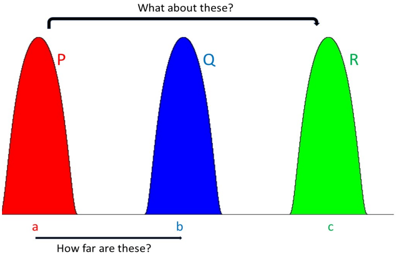
Optimal (measure) transportation



$$KL(P||Q) = \int \log\left(\frac{p}{q}\right) p = \infty = KL(P||R)$$

$$TV(P,Q) = \frac{1}{2}\int |p-q| = 1 = TV(P,R)$$

Optimal (measure) transportation



$$KL(P||Q) = \int \log\left(\frac{p}{q}\right) p = \infty = KL(P||R)$$

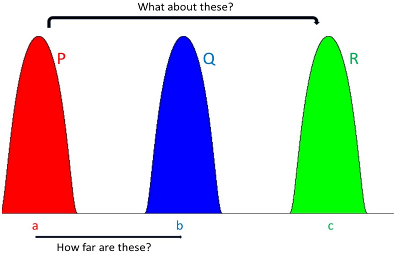
$$TV(P,Q) = \frac{1}{2}\int |p-q| = 1 = TV(P,R)$$

Need a notion of distance that is sensitive to geometry

Monge's approach (1781): Given probability measures P, Q on $\mathbb{R}^d$, find an "optimal" map $T_0 : \mathbb{R}^d \to \mathbb{R}^d$ satisfying

$$\min_{T\#P=Q} \int \|x - T(x)\|^2 \, dP(x), \quad T\#P = Q \Leftrightarrow X \sim P, \; T(X) \sim Q$$

Optimal (measure) transportation



$$KL(P||Q) = \int \log\left(\frac{p}{q}\right) p = \infty = KL(P||R)$$

$$TV(P,Q) = \frac{1}{2}\int |p-q| = 1$$

Need a notion of distance that is sensitive to geometry

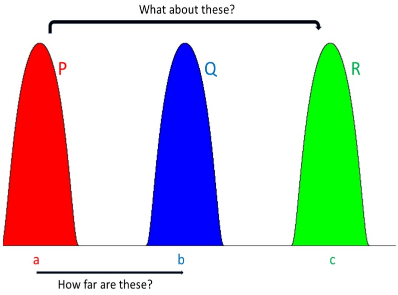
Monge's approach (1781): Given probability measures P, Q on $\mathbb{R}^d$, find an "optimal" map $T_0 : \mathbb{R}^d \to \mathbb{R}^d$ satisfying

$$\mathbf{W}_2^2(P,Q) = \min_{T\#P=Q} \int \|x - T(x)\|^2 \, dP(x), \quad T\#P = Q \Leftrightarrow X \sim P,\ T(X) \sim Q$$

Call optimizer $T_0^{P,Q} \equiv T_0$ (if it exists) — optimal transport (OT) map

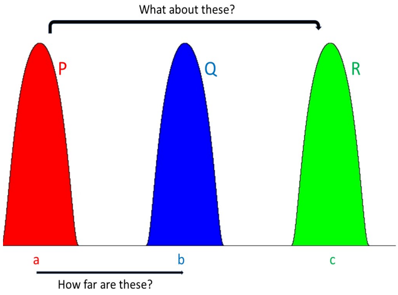
$\mathbf{W}_2^2(P,Q)$ — squared Wasserstein distance

Optimal (Measure) Transportation



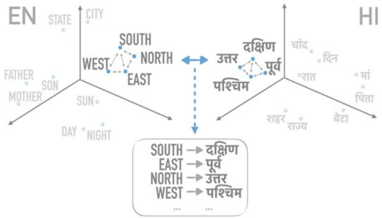
- $\mathbf{W}_2^2(P, Q) = \|b - a\|^2$, $\mathbf{W}_2^2(P, R) = \|c - a\|^2$

Optimal (Measure) Transportation



- $\mathbf{W}_2^2(P, Q) = \|b - a\|^2$, $\mathbf{W}_2^2(P, R) = \|c - a\|^2$
- $T_0^{P,Q}(x) = x + b - a$, $T_0^{P,R}(x) = x + c - a$

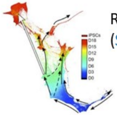
Applications of optimal transport — $X \sim P, \; T(X) \sim Q$



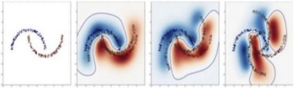
Translation (Mellis and Jaakkola, 2019)

Generative Modelling (Rout et al., 2021)

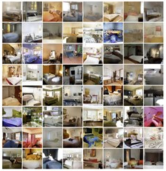
Color transfer (Rabin et al., 2010)



RNA sequencing (Schiebinger et al., 2019)



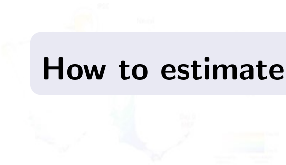
Domain adaptation (Courty et al., 2017)



Wasserstein GAN (Arjovsky et al., 2017)

Translation (Mellis and Jaakkola, 2019)

Generative Modelling (Rout et al, 2021)



(Schiebinger et al, 2019)

Color transfer (Rabin et al, 2010)

How to estimate the optimal transport map?

Domain adaptation (Courty et al, 2017)

Image retrieval (Papadakis, 2015)

Estimation — a plug-in approach

$$T_0 = \operatorname*{arg\,min}_{T\#P=Q} \int \|x - T(x)\|^2 \, dP(x), \quad T\#P = Q \Leftrightarrow X \sim P,\ T(X) \sim Q$$

T_0 is unique (P a.s.) if P, Q are absolutely continuous (McCann (1995))

Estimation — a plug-in approach

$$T_0 = \underset{T\#P=Q}{\arg\min} \int \|x - T(x)\|^2 \, dP(x), \quad T\#P = Q \Leftrightarrow X \sim P, \ T(X) \sim Q$$

T_0 is unique (P a.s.) if P, Q are absolutely continuous (McCann (1995))

Data: $X_1, X_2, \ldots, X_m$ iid P (unknown, absolutely continuous) and $Y_1, \ldots, Y_n$ iid Q (unknown, absolutely continuous)

Estimation — a plug-in approach

$$T_0 = \operatorname*{arg\,min}_{T\#P=Q} \int \|x - T(x)\|^2 \, dP(x), \quad T\#P = Q \Leftrightarrow X \sim P,\ T(X) \sim Q$$

T_0 is unique (P a.s.) if P, Q are absolutely continuous (McCann (1995))

Data: $X_1, X_2, \ldots, X_m$ iid P (unknown, absolutely continuous) and $Y_1, \ldots, Y_n$ iid Q (unknown, absolutely continuous)

Empirical distributions: $P_m := \frac{1}{m} \sum_{i=1}^{m} \delta_{X_i}, \qquad Q_n := \frac{1}{n} \sum_{j=1}^{n} \delta_{Y_j}$

Estimation — a plug-in approach

$$T_0 = \underset{T\#P=Q}{\arg\min} \int \|x - T(x)\|^2 \, dP(x), \quad T\#P = Q \Leftrightarrow X \sim P, \; T(X) \sim Q$$

T_0 is unique (P a.s.) if P, Q are absolutely continuous (McCann (1995))

Data: $X_1, X_2, \ldots, X_m$ iid P (unknown, absolutely continuous) and $Y_1, \ldots, Y_n$ iid Q (unknown, absolutely continuous)

Empirical distributions: $P_m := \frac{1}{m} \sum_{i=1}^m \delta_{X_i}, \qquad Q_n := \frac{1}{n} \sum_{j=1}^n \delta_{Y_j}$

When $m = n$

$$\hat{T} := \underset{T\#P_n=Q_n}{\arg\min} \int \|x - T(x)\|^2 \, dP_n(x)$$

Estimation — a plug-in approach

$$T_0 = \underset{T\#P=Q}{\arg\min} \int \|x - T(x)\|^2 \, dP(x), \quad T\#P = Q \Leftrightarrow X \sim P, \; T(X) \sim Q$$

T_0 is unique (P a.s.) if P, Q are absolutely continuous (McCann (1995))

Data: $X_1, X_2, \ldots, X_m$ iid P (unknown, absolutely continuous) and $Y_1, \ldots, Y_n$ iid Q (unknown, absolutely continuous)

Empirical distributions: $P_m := \frac{1}{m}\sum_{i=1}^m \delta_{X_i}, \qquad Q_n := \frac{1}{n}\sum_{j=1}^n \delta_{Y_j}$

When $m = n$

$$\hat{T} := \underset{T\#P_n=Q_n}{\arg\min} \int \|x - T(x)\|^2 \, dP_n(x) = \underset{T\#P_n=Q_n}{\arg\min} \frac{1}{n}\sum_{i=1}^n \|X_i - T(X_i)\|^2$$

Recall $T\#P_n = Q_n$ means if $X \sim P_n$, then $T(X) \sim Q_n$

Estimation — a plug-in approach

$$T_0 = \underset{T\#P=Q}{\arg\min} \int \|x - T(x)\|^2 \, dP(x), \quad T\#P = Q \Leftrightarrow X \sim P, \; T(X) \sim Q$$

T_0 is unique (P a.s.) if P, Q are absolutely continuous (McCann (1995))

Data: $X_1, X_2, \ldots, X_m$ iid P (unknown, absolutely continuous) and $Y_1, \ldots, Y_n$ iid Q (unknown, absolutely continuous)

Empirical distributions: $P_m := \frac{1}{m}\sum_{i=1}^{m} \delta_{X_i}$, $\qquad Q_n := \frac{1}{n}\sum_{j=1}^{n} \delta_{Y_j}$

When $m = n$

$$\hat{T} := \underset{T\#P_n=Q_n}{\arg\min} \int \|x - T(x)\|^2 \, dP_n(x) = \underset{T\#P_n=Q_n}{\arg\min} \frac{1}{n}\sum_{i=1}^{n} \|X_i - T(X_i)\|^2$$

Recall $T\#P_n = Q_n$ means if $X \sim P_n$, then $T(X) \sim Q_n$

$T\#P_n = Q_n$: $(T(X_1), \ldots, T(X_n))$ is some permutation of $(Y_1, \ldots, Y_n)$

Estimation — a plug-in approach

$$T_0 = \underset{T\#P=Q}{\arg\min} \int \|x - T(x)\|^2 \, dP(x), \quad T\#P = Q \Leftrightarrow X \sim P, \; T(X) \sim Q$$

T_0 is unique (P a.s.) if P, Q are absolutely continuous (McCann (1995))

Data: $X_1, X_2, \ldots, X_m$ iid P (unknown, absolutely continuous) and $Y_1, \ldots, Y_n$ iid Q (unknown, absolutely continuous)

Empirical distributions: $P_m := \frac{1}{m} \sum_{i=1}^{m} \delta_{X_i}, \qquad Q_n := \frac{1}{n} \sum_{j=1}^{n} \delta_{Y_j}$

When $m = n$

$$\hat{T} := \underset{T\#P_n=Q_n}{\arg\min} \int \|x - T(x)\|^2 \, dP_n(x) = \underset{T\#P_n=Q_n}{\arg\min} \frac{1}{n} \sum_{i=1}^{n} \|X_i - T(X_i)\|^2$$

Recall $T\#P_n = Q_n$ means if $X \sim P_n$, then $T(X) \sim Q_n$

Assignment problem (linear program – exact algorithm with complexity $O(n^3)$; parallel computing – Date and Nagi (2016))

Estimation — a plug-in approach (Continued)

What happens when $m < n$?

Can we still define

$$\hat{T} := \underset{T\#P_m = Q_n}{\arg\min} \int \|x - T(x)\|^2 \, dP_m(x)??$$

Estimation — a plug-in approach (Continued)

What happens when $m < n$?

Can we still define

$$\hat{T} := \underset{T\#P_m=Q_n}{\arg\min} \int \|x - T(x)\|^2 \, dP_m(x)??$$

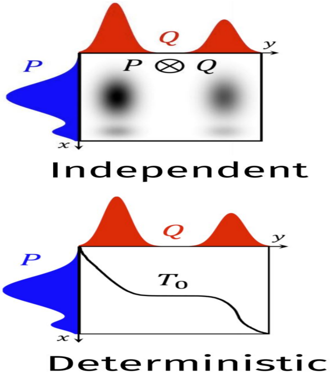
NOT FEASIBLE!

There is no function T such that $T\#P_m = Q_n$

- (Kantorovic relaxation)
  Let $\Pi(P, Q)$ be the set of probability measures (coupling) on $\mathbb{R}^d \times \mathbb{R}^d$, with marginals P, Q.
  Then

  $$W_2^2(P, Q) = \inf_{\gamma \in \Pi(P,Q)} \int \|x - y\|^2 \, d\gamma(x, y)$$

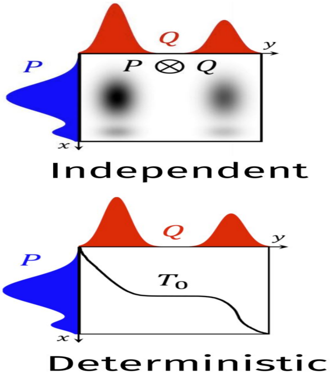
  Examples: $\pi \equiv P \otimes Q$,
  $\pi(x, y) \propto \mathbf{1}(y = T_0(x))$

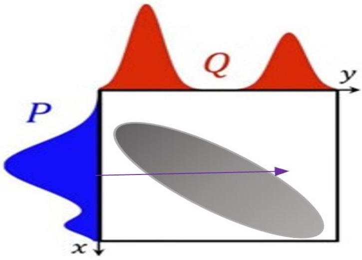


- (Kantorovic relaxation)
  Let $\Pi(P,Q)$ be the set of probability measures (coupling) on $\mathbb{R}^d \times \mathbb{R}^d$, with marginals P, Q.
  Then

  $$W_2^2(P,Q) = \inf_{\gamma \in \Pi(P,Q)} \int \|x-y\|^2 \, d\gamma(x,y)$$

  Examples: $\pi \equiv P \otimes Q$, $\pi(x,y) \propto \mathbf{1}(y = T_0(x))$
- Always has a minimizer which matches T_0 if P is absolutely continuous



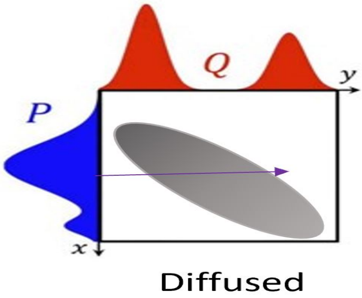


Diffused

- Solve

$$\widehat{\gamma} \in \underset{\gamma \in \Pi(P_m, Q_n)}{\arg\min} \int \|x - y\|^2 \, d\gamma(x, y)$$

via a linear program



- Solve

$$\widehat{\gamma} \in \underset{\gamma \in \Pi(P_m, Q_n)}{\arg\min} \int \|x - y\|^2 \, d\gamma(x, y)$$

 via a linear program

- D., Ghosal, and Sen (NeurIPS, 2021): Define our estimator (barycentric projection) as

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

Both definitions coincide when $m = n$

What is the rate of convergence of $\hat{T}$ to T_0?

What is the rate of convergence of $\hat{T}$ to T_0?

When $m = n$ ...

Empirical OT map:

$$\hat{T} := \underset{T\#P_n = Q_n}{\arg\min} \int \|x - T(x)\|^2 \, dP_n(x)$$

Population OT map:

$$T_0 := \underset{T\#P = Q}{\arg\min} \int \|x - T(x)\|^2 \, dP(x)$$

What is the rate of convergence of $\hat{T}$ to T_0?

When $m = n$ ...

Empirical OT map:

$$\hat{T} := \underset{T\#P_n=Q_n}{\arg\min} \int \|x - T(x)\|^2 \, dP_n(x)$$

Population OT map:

$$T_0 := \underset{T\#P=Q}{\arg\min} \int \|x - T(x)\|^2 \, dP(x)$$

Different parameter spaces

What is the rate of convergence of $\hat{T}$ to T_0?

Rate of convergence (D., Ghosal, and Sen (NeurIPS, 2021))

Assume that T_0 is Lipschitz, and both P and Q are compactly supported (can be relaxed). Then, for $d \geq 4$,

$$\frac{1}{m}\sum_{i=1}^{m} \mathbb{E}\| \hat{T}(X_i) - T_0(X_i)\|^2 \lesssim m^{-\frac{2}{d}} + n^{-\frac{2}{d}}.$$

What is the rate of convergence of $\hat{T}$ to T_0?

Rate of convergence (D., Ghosal, and Sen (NeurIPS, 2021))

Assume that T_0 is Lipschitz, and both P and Q are compactly supported (can be relaxed). Then, for $d \geq 4$,

$$\frac{1}{m}\sum_{i=1}^{m} \mathbb{E}\|\hat{T}(X_i) - T_0(X_i)\|^2 \lesssim m^{-\frac{2}{d}} + n^{-\frac{2}{d}}.$$

- The proof requires convex analysis, chaining and Talagrand's concentration arguments
- Minimax optimal for $d \geq 4$ (Hütter and Rigollet (2019))

What is the rate of convergence of $\hat{T}$ to T_0?

Rate of convergence (D., Ghosal, and Sen (NeurIPS, 2021))

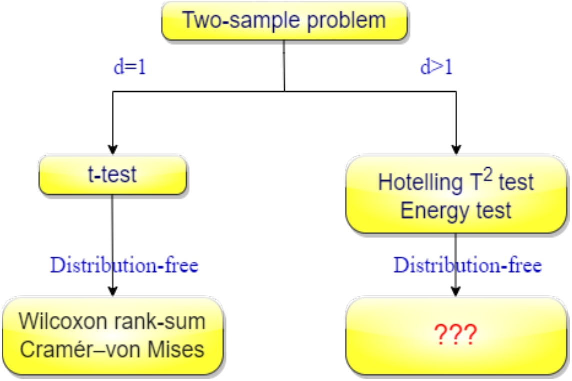
Assume that T_0 is Lipschitz, and both P and Q are compactly supported (can be relaxed). Then, for $d \geq 4$,

$$\frac{1}{m}\sum_{i=1}^{m}\mathbb{E}\|\hat{T}(X_i) - T_0(X_i)\|^2 \lesssim m^{-\frac{2}{d}} + n^{-\frac{2}{d}}.$$

- For $d = 1, 2, 3$, $m = n$, rates are $n^{-4/5}$, $n^{-2/3}$, $n^{-4/7}$ (ongoing work)
- The proof requires convex analysis, chaining and Talagrand's concentration arguments
- Minimax optimal for $d \geq 4$ (Hütter and Rigollet (2019))
- These are the first rates for a practically computable estimator of the OT map T_0 (note $\hat{T}$ requires no tuning) skip

Question

Can we construct multivariate distribution-free tests?



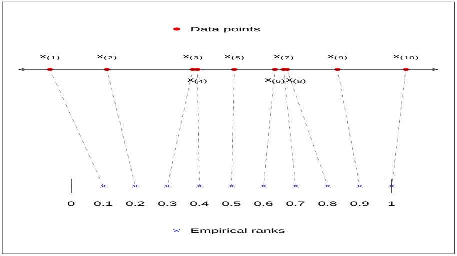
Outline

Ranks: When $d = 1$

- Rank map $\widehat{R}_n$ assigns $\{X_1, X_2, \ldots, X_n\}$ to elements of $\{\frac{1}{n}, \frac{2}{n}, \ldots, \frac{n}{n}\}$
- Define $\nu_n := \frac{1}{n}\sum_{i=1}^{n} \delta_{X_i}$ and $\mu_n := \frac{1}{n}\sum_{j=1}^{n} \delta_{\frac{j}{n}}$

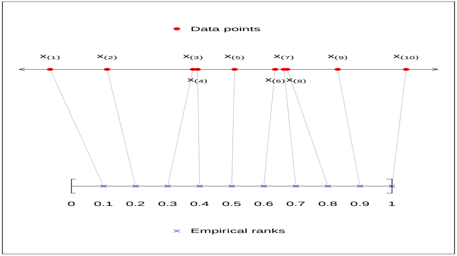
Ranks: When $d = 1$

- Rank map $\widehat{R}_n$ assigns $\{X_1, X_2, \ldots, X_n\}$ to elements of $\{\frac{1}{n}, \frac{2}{n}, \ldots, \frac{n}{n}\}$
- Define $\nu_n := \frac{1}{n}\sum_{i=1}^{n} \delta_{X_i}$ and $\mu_n := \frac{1}{n}\sum_{j=1}^{n} \delta_{\frac{j}{n}}$



Ranks: When $d = 1$

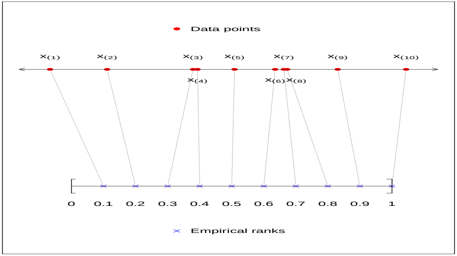
- Rank map $\widehat{R}_n$ assigns $\{X_1, X_2, \ldots, X_n\}$ to elements of $\{\frac{1}{n}, \frac{2}{n}, \ldots, \frac{n}{n}\}$
- Define $\nu_n := \frac{1}{n}\sum_{i=1}^{n} \delta_{X_i}$ and $\mu_n := \frac{1}{n}\sum_{j=1}^{n} \delta_{\frac{j}{n}}$



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

Ranks: When $d = 1$

- Rank map $\widehat{R}_n$ assigns $\{X_1, X_2, \ldots, X_n\}$ to elements of $\{\frac{1}{n}, \frac{2}{n}, \ldots, \frac{n}{n}\}$
- Define $\nu_n := \frac{1}{n}\sum_{i=1}^{n} \delta_{X_i}$ and $\mu_n := \frac{1}{n}\sum_{j=1}^{n} \delta_{\frac{j}{n}}$



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

$\widehat{R}_n$ is the **empirical OT map** from ν_n to μ_n

Multivariate ranks ($d \geq 1$)

- Empirical rank map assigns $\{X_1, \ldots, X_n\} \to \{c_1, \ldots, c_n\} \subset \mathbb{R}^d$ — grid of "reference" points (e.g., a random sample from Unif$[0,1]^d$, $\mathcal{N}(0, I_d)$ distribution, deterministic quasi-Monte Carlo sequences)

Multivariate ranks ($d \geq 1$)

- Empirical rank map assigns $\{X_1, \ldots, X_n\} \to \{c_1, \ldots, c_n\} \subset \mathbb{R}^d$ — grid of "reference" points (e.g., a random sample from Unif$[0,1]^d$, $\mathcal{N}(0, I_d)$ distribution, deterministic quasi-Monte Carlo sequences)

- Sample rank map (Hallin (2017)) is defined as the **empirical OT map**:

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

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

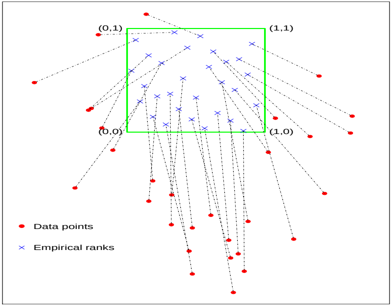
Multivariate ranks ($d \geq 1$)

- Empirical rank map assigns $\{X_1, \ldots, X_n\} \to \{c_1, \ldots, c_n\} \subset \mathbb{R}^d$ — grid of "reference" points (e.g., a random sample from Unif$[0,1]^d$, $\mathcal{N}(0, I_d)$ distribution, deterministic quasi-Monte Carlo sequences)

- Sample rank map (Hallin (2017)) is defined as the **empirical OT map**:

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

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



Multivariate rank function as an OT map

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

Multivariate rank function as an OT map

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

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

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

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

Multivariate rank function as an OT map

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

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

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

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

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

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

Multivariate rank function as an OT map

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

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

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

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

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

- $R(\cdot)$ characterizes distribution: $R_1(x) = R_2(x)\ \forall\, x \in \mathbb{R}^d$ iff $P_1 = P_2$
- $R(\cdot)$ is the gradient of a convex function and smoothly invertible

Multivariate rank function as an OT map

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

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

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

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

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

- $R(\cdot)$ characterizes distribution: $R_1(x) = R_2(x)\ \forall\, x \in \mathbb{R}^d$ iff $P_1 = P_2$
- $R(\cdot)$ is the gradient of a convex function and smoothly invertible
- When $d = 1$, $R(\cdot)$ is the CDF of X, when $\mu =$ Unif$([0,1])$

Properties [D. and Sen (JASA 2020); D., Bhattacharya, and Sen (2021)]

- Distribution-freeness: If ν is absolutely continuous, then

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

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

Properties [D. and Sen (JASA 2020); D., Bhattacharya, and Sen (2021)]

- Distribution-freeness: If ν is absolutely continuous, then

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

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

- Consistency: If $\mu_n := \frac{1}{n}\sum_{j=1}^n \delta_{c_j} \xrightarrow{d} \mu$ (abs. cont.), then

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

where R is the unique OT map from ν to μ.

No moment assumptions needed on the model

Outline

Outline

Testing for equality of two multivariate distributions

- **Data**: $\{X_i\}_{i=1}^m$ iid P on $\mathbb{R}^d$; $\{Y_j\}_{j=1}^n$ iid Q on $\mathbb{R}^d$, $d \geq 1$
- Test if the two samples came from the same distribution, i.e.,

$$\mathrm{H}_0 : P = Q \qquad \text{versus} \qquad \mathrm{H}_1 : P \neq Q$$

Testing for equality of two multivariate distributions

- **Data**: $\{X_i\}_{i=1}^m$ iid P on $\mathbb{R}^d$; $\{Y_j\}_{j=1}^n$ iid Q on $\mathbb{R}^d$, $d \geq 1$
- Test if the two samples came from the same distribution, i.e.,

$$\mathrm{H}_0 : P = Q \qquad \text{versus} \qquad \mathrm{H}_1 : P \neq Q$$

- Let $N = m + n$ and assume $\frac{m}{N} \to \lambda \in (0, 1)$

Testing for equality of two multivariate distributions

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

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

$$\mathrm{H}_0 : P = Q \qquad \text{versus} \qquad \mathrm{H}_1 : P \neq Q$$

- Let $N = m + n$ and assume $\frac{m}{N} \to \lambda \in (0, 1)$

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

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

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

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

Testing for equality of two multivariate distributions

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

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

$$\mathrm{H}_0 : P = Q \qquad \text{versus} \qquad \mathrm{H}_1 : P \neq Q$$

- Let $N = m + n$ and assume $\frac{m}{N} \to \lambda \in (0, 1)$

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

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

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

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

- Consistency: $\mathbb{P}(\mathrm{T}^2_{m,n} > c_\alpha) \to 1$ when $\mathbb{E}[X_1] \neq \mathbb{E}[Y_1]$

Data: $\{X_i\}_{i=1}^m$ iid P (abs. cont.), $\{Y_j\}_{j=1}^n$ iid Q on $\mathbb{R}^d$, $d \geq 1$

Reference dist.: μ on $\mathcal{S} \subset \mathbb{R}^d$ (abs. cont.; e.g., $\mu = \text{Unif}([0,1]^d)$)

Proposed tests [D. & Sen (JASA, 2020); D., Bhattacharya & Sen (2021)]

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

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

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

Data: $\{X_i\}_{i=1}^m$ iid P (abs. cont.), $\{Y_j\}_{j=1}^n$ iid Q on $\mathbb{R}^d$, $d \geq 1$

Reference dist.: μ on $\mathcal{S} \subset \mathbb{R}^d$ (abs. cont.; e.g., $\mu = \text{Unif}([0,1]^d)$)

Proposed tests [D. & Sen (JASA, 2020); D., Bhattacharya & Sen (2021)]

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

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

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

- General principle: Start with a "good" test and replace the X_i's and Y_j's with their pooled multivariate ranks

Data: $\{X_i\}_{i=1}^m$ iid P (abs. cont.), $\{Y_j\}_{j=1}^n$ iid Q on $\mathbb{R}^d$, $d \geq 1$

Reference dist.: μ on $\mathcal{S} \subset \mathbb{R}^d$ (abs. cont.; e.g., $\mu = \text{Unif}([0,1]^d)$)

Proposed tests [D. & Sen (JASA, 2020); D., Bhattacharya & Sen (2021)]

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

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

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

- General principle: Start with a "good" test and replace the X_i's and Y_j's with their pooled multivariate ranks

- This yields the Wilcoxon rank-sum test when applied to the t-test. Therefore $\mathrm{RT}^2_{m,n}$ is equivalent to Wilcoxon rank-sum when $d = 1$

Data: $\{X_i\}_{i=1}^m$ iid P (abs. cont.), $\{Y_j\}_{j=1}^n$ iid Q on $\mathbb{R}^d$, $d \geq 1$

Reference dist.: μ on $\mathcal{S} \subset \mathbb{R}^d$ (abs. cont.; e.g., $\mu = \text{Unif}([0,1]^d)$)

Proposed tests [D. & Sen (JASA, 2020); D., Bhattacharya & Sen (2021)]

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

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

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

- General principle: Start with a "good" test and replace the X_i's and Y_j's with their pooled multivariate ranks

- This yields the Wilcoxon rank-sum test when applied to the t-test. Therefore $\mathrm{RT}^2_{m,n}$ is equivalent to Wilcoxon rank-sum when $d = 1$

Distribution-freeness [D. & Sen (JASA, 2020)]

Under H_0, distributions of $\mathrm{RT}^2_{m,n}$ are free of $P \equiv Q$

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

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

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

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

Power (D., Bhattacharya, and Sen, 2021)

Under location shift alternatives, we have

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

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

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

Power (D., Bhattacharya, and Sen, 2021)

Under location shift alternatives, we have

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

Asymptotic null distribution (D., Bhattacharya, and Sen, 2021)

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

$$\mathrm{RT}^2_{m,n} \xrightarrow{d} \chi^2_d.$$

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

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

Power (D., Bhattacharya, and Sen, 2021)

Under location shift alternatives, we have

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

Asymptotic null distribution (D., Bhattacharya, and Sen, 2021)

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

$$\mathrm{RT}^2_{m,n} \xrightarrow{d} \chi^2_d.$$

Goal

How does rank Hotelling test compare with Hotelling T^2 test?

Outline

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

In principle, $\mathrm{ARE}\,(S_N, T_N)$ can depend on α and β, but in many interesting cases they don't

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

$\mathrm{ARE}\,(\mathrm{RT}^2_{m,n}, \mathrm{T}^2_{m,n})$ can be derived from the distribution of both test statistics under above alternatives

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

$\mathrm{ARE}\,(\mathrm{RT}^2_{m,n}, \mathrm{T}^2_{m,n})$ can be derived from the distribution of both test statistics under above alternatives

Some observations

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

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

$\mathrm{ARE}\,(\mathrm{RT}^2_{m,n}, \mathrm{T}^2_{m,n})$ can be derived from the distribution of both test statistics under above alternatives

Some observations

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

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

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

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

Independent coordinates case

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

Theorem (D., Bhattacharya, and Sen (2021))

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

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

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

Independent coordinates case

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

Theorem (D., Bhattacharya, and Sen (2021))

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

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

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

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

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

Independent coordinates case

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

Theorem (D., Bhattacharya, and Sen (2021))

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

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

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

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

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

Elliptically symmetric distributions

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

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

Elliptically symmetric distributions

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

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

Theorem (D., Bhattacharya, and Sen (2021))

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

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

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

Model for Independent Component Analysis (ICA)

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

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

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

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

$$X = A\, W$$

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

Model for Independent Component Analysis (ICA)

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

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

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

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

$$X = A\,W$$

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

Theorem (D., Bhattacharya, and Sen (2021))

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

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

Recap

- Exact distribution-free, nonparametric test, gives uniformly level α test and high efficiency compared to Hotelling T^2 test
- Provides the first comprehensive extension of classical nonparametric testing to the multivariate setting skip
- Using Gaussian reference distribution ensures $\mathrm{ARE}\,(\mathrm{RT}^2_{m,n}, \mathrm{T}^2_{m,n}) \geq 1$ for many popular subfamilies. Note that the test is agnostic to the underlying subfamily skip

Recap

- Exact distribution-free, nonparametric test, gives uniformly level α test and high efficiency compared to Hotelling T^2 test

- Provides the first comprehensive extension of classical nonparametric testing to the multivariate setting skip

- Using Gaussian reference distribution ensures $\mathrm{ARE}\left(\mathrm{RT}^2_{m,n}, \mathrm{T}^2_{m,n}\right) \geq 1$ for many popular subfamilies. Note that the test is agnostic to the underlying subfamily skip

- Robust against outliers and better finite-sample performance under heavy-tailed distributions skip

Recap

- Exact distribution-free, nonparametric test, gives uniformly level α test and high efficiency compared to Hotelling T^2 test
- Provides the first comprehensive extension of classical nonparametric testing to the multivariate setting skip
- Using Gaussian reference distribution ensures $\mathrm{ARE}\,(\mathrm{RT}^2_{m,n}, \mathrm{T}^2_{m,n}) \geq 1$ for many popular subfamilies. Note that the test is agnostic to the underlying subfamily skip
- Robust against outliers and better finite-sample performance under heavy-tailed distributions skip

Thank you. Questions?

Multivariate two-sample goodness-of-fit test

- Recall general strategy: Start with a "good" test and replace the X_i's and Y_j's with their pooled multivariate ranks

Multivariate two-sample goodness-of-fit test

- Start with a "good" consistent test, say the energy statistic (Székely and Rizzo, 2013) and apply our general strategy

Multivariate two-sample goodness-of-fit test

- Start with a "good" consistent test, say the energy statistic (Székely and Rizzo, 2013) and apply our general strategy
- Suppose $X, X' \stackrel{iid}{\sim} P$, $Y, Y' \stackrel{iid}{\sim} Q$, $K(s,t) := \|s - t\|$, then energy dist. (or kernel MMD (see Gretton et al., 2008)):

$$\mathrm{E}^2(P, Q) := 2\,\mathbb{E}K(X, Y) - \mathbb{E}K(X, X') - \mathbb{E}K(Y, Y') \geq 0$$

Multivariate two-sample goodness-of-fit test

- Start with a "good" consistent test, say the energy statistic (Székely and Rizzo, 2013) and apply our general strategy
- Suppose $X, X' \overset{iid}{\sim} P$, $Y, Y' \overset{iid}{\sim} Q$, $K(s,t) := \|s - t\|$, then energy dist. (or kernel MMD (see Gretton et al., 2008)):

$$\mathrm{E}^2(P, Q) := 2\,\mathbb{E}K(X, Y) - \mathbb{E}K(X, X') - \mathbb{E}K(Y, Y') \geq 0$$

- Characterizes equality of distributions: $E(P, Q) = 0$ iff $P = Q$
- **E-statistic**: $\mathrm{E}^2_{m,n}\left(\{X_i\}_{i=1}^m, \{Y_j\}_{j=1}^n\right) := 2A - B - C$ where

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

Multivariate two-sample goodness-of-fit test

- Start with a "good" consistent test, say the energy statistic (Székely and Rizzo, 2013) and apply our general strategy
- Suppose $X, X' \overset{iid}{\sim} P$, $Y, Y' \overset{iid}{\sim} Q$, $K(s,t) := \|s - t\|$, then energy dist. (or kernel MMD (see Gretton et al., 2008)):

$$\mathrm{E}^2(P, Q) := 2\,\mathbb{E}K(X, Y) - \mathbb{E}K(X, X') - \mathbb{E}K(Y, Y') \geq 0$$

- Characterizes equality of distributions: $E(P, Q) = 0$ iff $P = Q$
- **E-statistic**: $\mathrm{E}^2_{m,n}\left(\{X_i\}_{i=1}^m, \{Y_j\}_{j=1}^n\right) := 2A - B - C$ where

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

- **Energy test**: Reject H_0 if $\mathrm{E}_{m,n}\left(\{X_i\}_{i=1}^m, \{Y_j\}_{j=1}^n\right) > \kappa_\alpha$ (depends on P; we can use permutation test)

Proposed statistic

Rank energy statistic [D. and Sen (JASA, 2020)]

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

$$\hat{\mathbf{R}}_{m,n} : \{X_1, \ldots, X_m, Y_1, \ldots, Y_n\} \to \{c_1, \ldots, c_{m+n}\} \subset [0,1]^d$$

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

Proposed statistic

Rank energy statistic [D. and Sen (JASA, 2020)]

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

$$\hat{\mathbf{R}}_{m,n} : \{X_1, \ldots, X_m, Y_1, \ldots, Y_n\} \to \{c_1, \ldots, c_{m+n}\} \subset [0,1]^d$$

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

Distribution-freeness

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

- Dist. of $\mathrm{RE}_{m,n}$ just depends on c_i's, m, n and d
- **Rank energy test**: Reject H_0 if $\mathrm{RE}_{m,n} > \kappa_\alpha$ (universal threshold, free of $P = Q$)

Proposed statistic

Rank energy statistic [D. and Sen (JASA, 2020)]

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

$$\hat{\mathbf{R}}_{m,n} : \{X_1, \ldots, X_m, Y_1, \ldots, Y_n\} \to \{c_1, \ldots, c_{m+n}\} \subset [0,1]^d$$

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

Distribution-freeness

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

- Dist. of $\mathrm{RE}_{m,n}$ just depends on c_i's, m, n and d
- **Rank energy test**: Reject H_0 if $\mathrm{RE}_{m,n} > \kappa_\alpha$ (universal threshold, free of $P = Q$)

Simplification when $d = 1$

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

Power [D. and Sen (JASA, 2020)]

Under (ii) and $P \neq Q$, if $\frac{m}{m+n} \approx \lambda \in (0,1)$ then,

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

Proposed test has asymptotic power 1, against all fixed alternatives

Power [D. and Sen (JASA, 2020)]

Under (ii) and $P \neq Q$, if $\frac{m}{m+n} \approx \lambda \in (0,1)$ then,

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

Proposed test has asymptotic power 1, against all fixed alternatives

Limiting distribution under H_0 [D. and Sen (JASA, 2020)]

If (i) $P \equiv Q$ is abs. cont., and

(ii) $\frac{1}{N}\sum_{i=1}^{N} \delta_{c_i} \xrightarrow{d} \mu$ a.s. $(N = m + n)$

Then, under H_0, $\exists$ a universal distribution s.t.

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

Pitman efficiency

Some observations

- Efficiency of $\mathrm{RE}_{m,n}$ w.r.t. $\mathrm{E}_{m,n}$ depends on the type I error α and power β, which makes it hard obtain efficiency lower bounds

- Existing tests which are both consistent and distribution-free usually do not have Pitman efficiency

Pitman efficiency

Some observations

- Efficiency of $\mathrm{RE}_{m,n}$ w.r.t. $\mathrm{E}_{m,n}$ depends on the type I error α and power β, which makes it hard obtain efficiency lower bounds
- Existing tests which are both consistent and distribution-free usually do not have Pitman efficiency

- Fix a level parameter $\alpha \in (0, 1)$. Consider:

$$\mathrm{H}_0 : \theta_2 - \theta_1 = 0 \qquad \text{versus} \qquad \mathrm{H}_1 : \theta_2 - \theta_1 = hN^{-1/2}$$

Pitman efficiency

Some observations

- Efficiency of $\mathrm{RE}_{m,n}$ w.r.t. $\mathrm{E}_{m,n}$ depends on the type I error α and power β, which makes it hard obtain efficiency lower bounds

- Existing tests which are both consistent and distribution-free usually do not have Pitman efficiency

- Fix a level parameter $\alpha \in (0, 1)$. Consider:

$$\mathrm{H}_0 : \theta_2 - \theta_1 = 0 \qquad \text{versus} \qquad \mathrm{H}_1 : \theta_2 - \theta_1 = hN^{-1/2}$$

- In Bhattacharya (2019), the author showed that for many asymp. distribution-free tests

$$\mathbb{P}_{\mathrm{H}_1}(T_{m,n} \text{ rejects } \mathrm{H}_0) \longrightarrow \alpha \quad \text{(powerless)}$$

Pitman efficiency

Some observations

- Efficiency of $\mathrm{RE}_{m,n}$ w.r.t. $\mathrm{E}_{m,n}$ depends on the type I error α and power β, which makes it hard obtain efficiency lower bounds
- Existing tests which are both consistent and distribution-free usually do not have Pitman efficiency
- Fix a level parameter $\alpha \in (0, 1)$. Consider:

$$\mathrm{H}_0 : \theta_2 - \theta_1 = 0 \qquad \text{versus} \qquad \mathrm{H}_1 : \theta_2 - \theta_1 = hN^{-1/2}$$

- In Bhattacharya (2019), the author showed that for many asymp. distribution-free tests

$$\mathbb{P}_{\mathrm{H}_1}(T_{m,n} \text{ rejects } \mathrm{H}_0) \longrightarrow \alpha \quad \text{(powerless)}$$

Rank Energy $\mathrm{RE}_{m,n}$ [D., Bhattacharya, and Sen (working paper)]

- $$\lim_{m,n\to\infty} \mathbb{P}_{\mathrm{H}_1}(\mathrm{RE}_{m,n} \text{ rejects } \mathrm{H}_0) > \alpha$$
- Only consistent, exactly dist.-free test that can distinguish H_0 & H_1

Summary

- Introduced optimal transport and obtained first tuning-free, minimax optimal estimator of optimal transport map

Summary

- Introduced optimal transport and obtained first tuning-free, minimax optimal estimator of optimal transport map
- Multivariate distribution-free nonparametric testing procedures with high efficiency, constructed using optimal transport skip

Summary

- Introduced optimal transport and obtained first tuning-free, minimax optimal estimator of optimal transport map
- Multivariate distribution-free nonparametric testing procedures with high efficiency, constructed using optimal transport skip
- Proposed a general framework, other examples include independence testing, testing for symmetry, testing equality of K-distributions ...

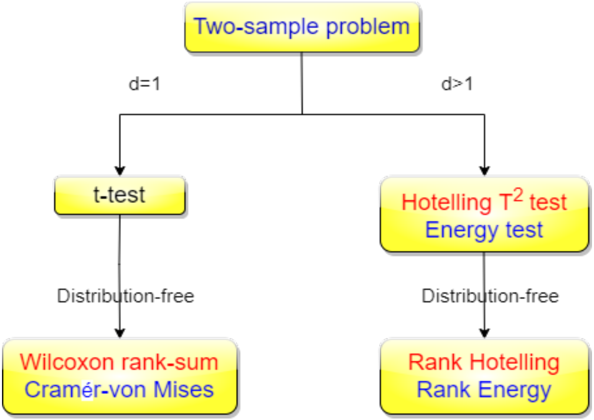
Summary

- Introduced optimal transport and obtained first tuning-free, minimax optimal estimator of optimal transport map
- Multivariate distribution-free nonparametric testing procedures with high efficiency, constructed using optimal transport ⏩ skip
- Proposed a general framework, other examples include independence testing, testing for symmetry, testing equality of K-distributions ...
- Independence testing: In D., Bhattacharya, and Sen (2021) we obtain multivariate distribution-free extensions of Spearman's correlation and kernel tests of dependence; obtain similar results

Summary

- Introduced optimal transport and obtained first tuning-free, minimax optimal estimator of optimal transport map
- Multivariate distribution-free nonparametric testing procedures with high efficiency, constructed using optimal transport ▸▸ skip
- Proposed a general framework, other examples include independence testing, testing for symmetry, testing equality of K-distributions ...
- Independence testing: In D., Bhattacharya, and Sen (2021) we obtain multivariate distribution-free extensions of Spearman's correlation and kernel tests of dependence; obtain similar results
- Tuning-free, robust, computationally feasible procedures, performs particularly well under heavy-tailed data ▸▸ skip

Summary


Two-sample problem
d=1
d>1
t-test
Hotelling T2 test
Energy test
Distribution-free
Distribution-free
Wilcoxon rank-sum
Cramér-von Mises
Rank Hotelling
Rank Energy

References

D. and Sen, (2020). https://arxiv.org/pdf/1909.08733 (JASA, to appear)

D., Ghosal, and Sen (2021). https://arxiv.org/abs/2107.01718 (NeurIPS, 2021)

D., Bhattacharya, and Sen (2021). https://arxiv.org/abs/2104.01986

D., Bhattacharya and Sen (2021+). (working paper)

Nonparametric association/conditional association

- Suppose

$$\begin{pmatrix} X \\ Y \end{pmatrix} \sim \mathcal{N}_2 \left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix} \right)$$

The correlation ρ measures association (linear) between X and Y

Nonparametric association/conditional association

- Suppose

$$\begin{pmatrix} X \\ Y \end{pmatrix} \sim \mathcal{N}_2 \left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix} \right)$$

  The correlation ρ measures association (linear) between X and Y
- $\rho = 0$ iff X and Y are independent
- $\rho = \pm 1$ iff Y is an exact, measurable (linear) function of X

Nonparametric association/conditional association

- Suppose
$$\begin{pmatrix} X \\ Y \end{pmatrix} \sim \mathcal{N}_2\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix}\right)$$
The correlation ρ measures association (linear) between X and Y
- $\rho = 0$ iff X and Y are independent
- $\rho = \pm 1$ iff Y is an exact, measurable (linear) function of X

Can we find a nonparametric measure of association between random elements on topological spaces?

Nonparametric association/conditional association

- Suppose

$$\begin{pmatrix} X \\ Y \end{pmatrix} \sim \mathcal{N}_2 \left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix} \right)$$

  The correlation ρ measures association (linear) between X and Y
- $\rho = 0$ iff X and Y are independent
- $\rho = \pm 1$ iff Y is an exact, measurable (linear) function of X

Can we find a nonparametric measure of association between random elements on topological spaces?

Goal: Given random elements $(X_1, Y_1), \ldots, (X_n, Y_n) \overset{i.i.d.}{\sim} (X, Y)$, define $T(X, Y)$ and T_n (estimator), such that:

Nonparametric association/conditional association

- Suppose
$$\begin{pmatrix} X \\ Y \end{pmatrix} \sim \mathcal{N}_2\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix}\right)$$
The correlation ρ measures association (linear) between X and Y
- $\rho = 0$ iff X and Y are independent
- $\rho = \pm 1$ iff Y is an exact, measurable (linear) function of X

Can we find a nonparametric measure of association between random elements on topological spaces?

Goal: Given random elements $(X_1, Y_1), \ldots, (X_n, Y_n) \overset{i.i.d.}{\sim} (X, Y)$, define $T(X, Y)$ and T_n (estimator), such that:

- $T_n, T(X, Y) \in [0, 1]$
- $T(X, Y) = 0$ iff X and Y are independent

Nonparametric association/conditional association

- Suppose

$$\begin{pmatrix} X \\ Y \end{pmatrix} \sim \mathcal{N}_2\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix}\right)$$

  The correlation ρ measures association (linear) between X and Y

- $\rho = 0$ iff X and Y are independent
- $\rho = \pm 1$ iff Y is an exact, measurable (linear) function of X

Can we find a nonparametric measure of association between random elements on topological spaces?

Goal: Given random elements $(X_1, Y_1), \ldots, (X_n, Y_n) \overset{i.i.d.}{\sim} (X, Y)$, define $T(X, Y)$ and T_n (estimator), such that:

- $T_n, T(X, Y) \in [0, 1]$
- $T(X, Y) = 0$ iff X and Y are independent
- $T(X, Y) = 1$ iff $Y = f(X)$ for (unknown) meas. function $f(\cdot)$

Nonparametric association/conditional association

- Suppose

$$\begin{pmatrix} X \\ Y \end{pmatrix} \sim \mathcal{N}_2 \left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix} \right)$$

  The correlation ρ measures association (linear) between X and Y
- $\rho = 0$ iff X and Y are independent
- $\rho = \pm 1$ iff Y is an exact, measurable (linear) function of X

Can we find a nonparametric measure of association between random elements on topological spaces?

Goal: Given random elements $(X_1, Y_1), \ldots, (X_n, Y_n) \overset{i.i.d.}{\sim} (X, Y)$, define $T(X, Y)$ and T_n (estimator), such that:

- $T_n, T(X, Y) \in [0, 1]$
- $T(X, Y) = 0$ iff X and Y are independent
- $T(X, Y) = 1$ iff $Y = f(X)$ for (unknown) meas. function $f(\cdot)$
- $T_n \xrightarrow{a.s.} T(X, Y)$

Measure of Association

We answer this question in the affirmative by combining ideas from reproducing kernel Hilbert spaces (RKHS) and geometric graphs (e.g., nearest neighbors, minimum spanning trees), to come up with a large class of such measures

- Our measures are completely nonparametric (unlike ρ)
- We can also extend this to measuring conditional association with applications in variable selection, conditional independence testing ...

Measure of Association

We answer this question in the affirmative by combining ideas from reproducing kernel Hilbert spaces (RKHS) and geometric graphs (e.g., nearest neighbors, minimum spanning trees), to come up with a large class of such measures

- Our measures are completely nonparametric (unlike ρ)
- We can also extend this to measuring conditional association with applications in variable selection, conditional independence testing ...

References:

D., Ghosal, and Sen (2020). https://arxiv.org/pdf/2010.01768.pdf (RR at AoS)

Huang, D., and Sen (2021). https://arxiv.org/pdf/2012.14804.pdf (JMLR, to appear)

Auddy, D., and Nandy (2021). https://arxiv.org/pdf/2104.15140.pdf (RR at Bernoulli)

Other works

Applied probability (Ising type models):

D. & Mukherjee (2020). https://arxiv.org/pdf/2005.00710.pdf (AoAP, to appear)

D., Mukherjee, Mukherjee & Yuan (2021). https://arxiv.org/abs/2012.05784

Other works

Applied probability (Ising type models):

D. & Mukherjee (2020). https://arxiv.org/pdf/2005.00710.pdf (AoAP, to appear)

D., Mukherjee, Mukherjee & Yuan (2021). https://arxiv.org/abs/2012.05784

Measuring (conditional) association on topological data:

D., Ghosal & Sen (2020). https://arxiv.org/pdf/2010.01768.pdf (RR at AoS)
Huang, D. & Sen (2021). https://arxiv.org/pdf/2012.14804.pdf (JMLR, to appear)

Auddy, D. & Nandy (2021). https://arxiv.org/pdf/2104.15140.pdf (RR at Bernoulli)

Other works

Applied probability (Ising type models):

D. & Mukherjee (2020). https://arxiv.org/pdf/2005.00710.pdf (AoAP, to appear)

D., Mukherjee, Mukherjee & Yuan (2021). https://arxiv.org/abs/2012.05784

Measuring (conditional) association on topological data:

D., Ghosal & Sen (2020). https://arxiv.org/pdf/2010.01768.pdf (RR at AoS)
Huang, D. & Sen (2021). https://arxiv.org/pdf/2012.14804.pdf (JMLR, to appear)

Auddy, D. & Nandy (2021). https://arxiv.org/pdf/2104.15140.pdf (RR at Bernoulli)

Gaussian mixture models and multiple testing:

D., Saha, Guntuboyina & Sen (2020).

https://www.tandfonline.com/doi/full/10.1080/01621459.2021.1888739

(JASA, published online)

Causal inference:

Ghosh, D., Karmakar & Sen (2021). https://arxiv.org/abs/2111.15524 ⏩ skip

Rate of convergence result

$$T_0 = \operatorname*{arg\,min}_{T\#P=Q} \int \|x - T(x)\|^2 \, dP(x),$$

$$W_2^2(P, Q) = \min_{T\#P=Q} \int \|x - T(x)\|^2 \, dP(x)$$

Data: $X_1, X_2, \ldots, X_n$ iid P and $Y_1, \ldots, Y_n$ iid Q

Rate of convergence result

$$T_0 = \operatorname*{arg\,min}_{T\#P=Q} \int \|x - T(x)\|^2 \, dP(x),$$

$$W_2^2(P, Q) = \min_{T\#P=Q} \int \|x - T(x)\|^2 \, dP(x)$$

Data: $X_1, X_2, \ldots, X_n$ iid P and $Y_1, \ldots, Y_n$ iid Q

Estimator:

$$\hat{T} := \operatorname*{arg\,min}_{T\#P_n=Q_n} \int \|x - T(x)\|^2 \, dP_n(x)$$

Rate of convergence result

$$T_0 = \mathop{\arg\min}_{T\#P=Q} \int \|x - T(x)\|^2 \, dP(x),$$

$$W_2^2(P, Q) = \min_{T\#P=Q} \int \|x - T(x)\|^2 \, dP(x)$$

Data: $X_1, X_2, \ldots, X_n$ iid P and $Y_1, \ldots, Y_n$ iid Q

Estimator:

$$\hat{T} := \mathop{\arg\min}_{T\#P_n=Q_n} \int \|x - T(x)\|^2 \, dP_n(x)$$

Rate of convergence (D., Ghosal, Sen, NeurIPS, 2021)

Assume that T_0 is Lipschitz, and both P and Q are compactly supported (can be relaxed). Then,

$$\frac{1}{n} \sum_{i=1}^{n} \| \hat{T}(X_i) - T_0(X_i)\|^2 \lesssim n^{-\frac{2}{d}}$$

for $d \geq 4$. For $d = 1, 2, 3$, the rates are n^{-1}, $n^{-2/3}$ and $n^{-4/7}$.

Nonparametric regression: Say $Y_i = f_0(X_i) + \epsilon_i$, $i = 1, \ldots, n$, where ϵ_i's are iid $\mathcal{N}(0,1)$, and $f_0 \in \mathcal{F}$

$$\hat{f}_n := \arg\min_{f \in \mathcal{F}} \sum_{i=1}^{n} (Y_i - f(X_i))^2$$

Both $\hat{f}_n$ and f_0 belong to $\mathcal{F}$, which yields the basic inequality:

$$\sum_{i=1}^{n} (Y_i - \hat{f}_n(X_i))^2 \leq \sum_{i=1}^{n} (Y_i - f_0(X_i))^2$$

Nonparametric regression: Say $Y_i = f_0(X_i) + \epsilon_i$, $i = 1, \ldots, n$, where ϵ_i's are iid $\mathcal{N}(0,1)$, and $f_0 \in \mathcal{F}$

$$\hat{f}_n := \arg\min_{f \in \mathcal{F}} \sum_{i=1}^{n} (Y_i - f(X_i))^2$$

Both $\hat{f}_n$ and f_0 belong to $\mathcal{F}$, which yields the basic inequality:

$$\sum_{i=1}^{n} (f_0(X_i) - \hat{f}_n(X_i))^2 \leq 2 \sum_{i=1}^{n} \epsilon_i (\hat{f}_n(X_i) - f_0(X_i))$$

Nonparametric regression: Say $Y_i = f_0(X_i) + \epsilon_i$, $i = 1, \ldots, n$, where ϵ_i's are iid $\mathcal{N}(0, 1)$, and $f_0 \in \mathcal{F}$

$$\hat{f}_n := \arg\min_{f \in \mathcal{F}} \sum_{i=1}^{n} (Y_i - f(X_i))^2$$

Both $\hat{f}_n$ and f_0 belong to $\mathcal{F}$, which yields the basic inequality:

$$\sum_{i=1}^{n} (Y_i - \hat{f}_n(X_i))^2 \leq \sum_{i=1}^{n} (Y_i - f_0(X_i))^2$$

Nonparametric regression: Say $Y_i = f_0(X_i) + \epsilon_i$, $i = 1, \ldots, n$, where ϵ_i's are iid $\mathcal{N}(0,1)$, and $f_0 \in \mathcal{F}$

$$\hat{f}_n := \arg\min_{f \in \mathcal{F}} \sum_{i=1}^{n} (Y_i - f(X_i))^2$$

Both $\hat{f}_n$ and f_0 belong to $\mathcal{F}$, which yields the basic inequality:

$$\sum_{i=1}^{n} (Y_i - \hat{f}_n(X_i))^2 \leq \sum_{i=1}^{n} (Y_i - f_0(X_i))^2$$

OT problem:

$$\hat{T} := \arg\min_{T \# P_n = Q_n} \int \|x - T(x)\|^2 \, dP_n(x)$$

- Constraint set: $\mathcal{T}_n := \{T : T \# P_n = Q_n\}$.
- $\hat{T}_n \in \mathcal{T}_n$ but $T_0 \notin \mathcal{T}_n$

Dual form

Alternatively,

$$W_2^2(P_n, Q_n) = \min_{T\#P_n = Q_n} \int \|x - T(x)\|^2 \, dP_n(x) = \min_{f,g} \int f \, dP_n + \int g \, dQ_n$$

such that $f(x) + g(y) \leq \|x - y\|^2$ for all $x, y \in \mathbb{R}^d$.

Dual form

Alternatively,

$$W_2^2(P_n, Q_n) = \min_{T\#P_n = Q_n} \int \|x - T(x)\|^2 \, dP_n(x) = \min_{f,g} \int f \, dP_n + \int g \, dQ_n$$

such that $f(x) + g(y) \leq \|x - y\|^2$ for all $x, y \in \mathbb{R}^d$.

Note that the constraints are not data driven.

Basic inequality (D., Ghosal and Sen, 2021)

Suppose T_0 is Lipschitz. Write $T_0 = \nabla \varphi_0$ and $\bar{Q}_n := T_0 \# P_n$. Then,

$$\frac{1}{n} \sum_{i=1}^{N} \| \hat{T}_n(X_i) - T_0(X_i)\|^2 \lesssim W_2^2(P_n, Q_n) - W_2^2(P_n, \bar{Q}_n) + \int g \, d(Q_n - \bar{Q}_n)$$

where $g(y) = \varphi_0^*(y) - (1/2)\|y\|^2$, $\varphi_0^*(y) := \sup_{x \in \mathbb{R}^d} (\langle x, y \rangle - \varphi_0(x))$ (Legendre-Fenchel dual of $\varphi_0(\cdot)$)

Proof requires arguments from convex analysis

Dual form

Alternatively,

$$W_2^2(P_n, Q_n) = \max_{f,g} \int f \, dP_n + \int g \, dQ_n$$

such that $f(x) + g(y) \leq \|x - y\|^2$ for all $x, y \in \mathbb{R}^d$.

Note that the constraints are not data driven.

Basic inequality (D., Ghosal and Sen, 2021)

Suppose T_0 is Lipschitz. Write $T_0 = \nabla \varphi_0$ and $\bar{Q}_n := T_0 \# P_n$. Then,

$$\frac{1}{n} \sum_{i=1}^{N} \| \hat{T}_n(X_i) - T_0(X_i)\|^2 \lesssim W_2^2(P_n, Q_n) - W_2^2(P_n, \bar{Q}_n) + \int g \, d(Q_n - \bar{Q}_n)$$

where $g(y) = \varphi_0^*(y) - (1/2)\|y\|^2$, $\varphi_0^*(y) := \sup_{x \in \mathbb{R}^d}(\langle x, y \rangle - \varphi_0(x))$ (Legendre-Fenchel dual of $\varphi_0(\cdot)$)

Proof requires arguments from convex analysis

Using the dual form of $W_2^2(\cdot, \cdot)$, coupled with chaining and Talagrand's concentration inequality proves the rate of convergence result

Thank you. Questions?

Properties

$$T_0 \overset{???}{:=} \underset{T\#P=Q}{\arg\min} \int \|x-T(x)\|^2 \, dP(x), \quad T\#P = Q \Leftrightarrow X \sim P,\ T(X) \sim Q.$$

- Does a solution always exist?
- Is the solution unique?

Properties

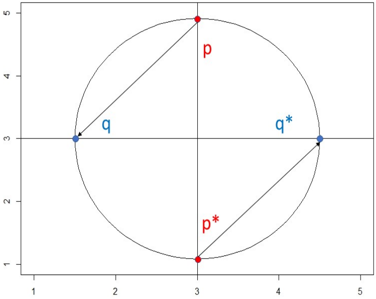
$$T_0 \overset{???}{:=} \underset{T\#P=Q}{\arg\min} \int \|x - T(x)\|^2 \, dP(x), \quad T\#P = Q \Leftrightarrow X \sim P, \ T(X) \sim Q.$$

- Does a solution always exist?

  No! Take $P = \delta_0$ and $Q = \text{Unif}[0,1]$.

- Is the solution unique?

  No! Take $P = 0.5\delta_p + 0.5\delta_{p^*}$ and $Q = 0.5\delta_q + 0.5\delta_{q^*}$.



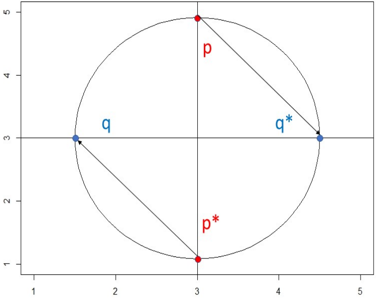
$$T_0 \overset{???}{:=} \underset{T\#P=Q}{\arg\min} \int \|x - T(x)\|^2 \, dP(x), \quad T\#P = Q \Leftrightarrow X \sim P,\ T(X) \sim Q.$$

- Does a solution always exist?

  No! Take $P = \delta_0$ and $Q = \text{Unif}[0,1]$.

- Is the solution unique?

  No! Take $P = 0.5\delta_p + 0.5\delta_{p^*}$ and $Q = 0.5\delta_q + 0.5\delta_{q^*}$.



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:

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

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)

Power plot with varying sample size

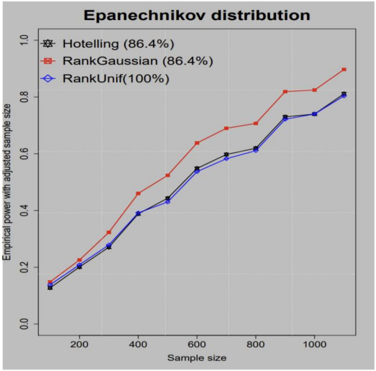


Figure: X_1, Y_1 are i.i.d. Epanechnikov with location parameters 0 and 0.1 respectively. $X_2, X_3 \sim X_1$, $Y_2, Y_3 \sim Y_1$ and $X := (X_1, X_2, X_3)$, $Y := (Y_1, Y_2, Y_3)$. Here

$$\text{eff}(\text{RankUnif}, \text{Hotelling}) = 0.864$$

and $\text{eff}(\text{RankGaussian}, \text{Hotelling}) > 1$ skip

Power plot with varying location parameter

Log-normal location problem (slightly heavy-tailed)

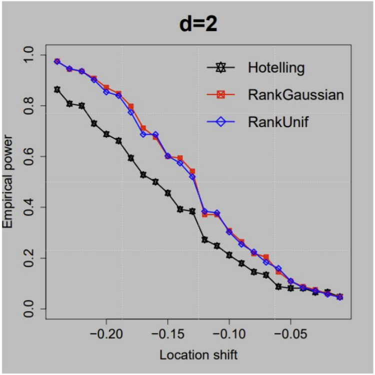


Figure: U_1, U_2 are iid standard normal, and V_1, V_2 are normal with variance 1 and varying mean. Define $X_i := \exp(U_i)$ and $Y_i := \exp(V_i)$. Set $X := (X_1, X_2)$ and $Y := (Y_1, Y_2)$ — sample size $n = 200$ skip

Power plot with varying location parameter

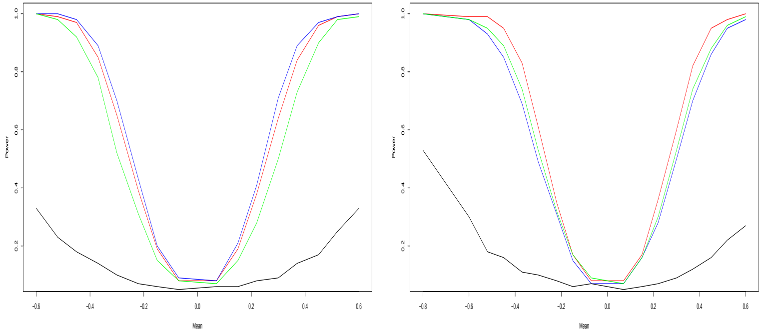


Figure: (Left panel) X_1, Y_1 are i.i.d. normal with mean 0 and μ respectively (and unit variance). $X_2, X_3 \sim X_1$, Y_2, $Y_3 \sim Y_1$ and $X := (X_1, X_2, X_3)$. Similarly define Y. ⏩ skip

(Right panel) $U := (U_1, U_2, U_3)$ and $V := (V_1, V_2, V_3)$ 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).

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 $n = 100, 300, 500, 700, 900,\ d = 8$.

What happens for $d = 1$?

$$T_0 \overset{???}{:=} \underset{T\#P=Q}{\arg\min} \int |x - T(x)|^2 \, dP(x).$$

- Assume $Q = \text{Unif}[0,1]$ and $X \sim P$ with cdf F

What happens for $d = 1$?

$$T_0 \overset{???}{:=} \operatorname*{arg\,min}_{T\#P=Q} \int |x - T(x)|^2 \, dP(x).$$

- Assume $Q = \text{Unif}[0,1]$ and $X \sim P$ with cdf F
- Given $x_1 \leq x_2 \in \mathbb{R}$, note that

$$(x_1 - T_0(x_1))^2 + (x_2 - T_0(x_2))^2 \leq (x_1 - T_0(x_2))^2 + (x_2 - T_0(x_1))^2$$
$$\Leftrightarrow T_0(x_1) \leq T_0(x_2)$$

  Expect $T_0(\cdot)$ to be monotone and $T_0(X) \overset{d}{=} \text{Unif}[0,1]$.

- $T_0(\cdot)$ is the distribution function of X, say $F(\cdot)$

What happens for $d = 1$?

$$\mathbf{F} = \underset{T\#P=Q}{\arg\min} \int |x - T(x)|^2 \, dP(x).$$

- Assume $Q = \text{Unif}[0,1]$ and $X \sim P$ with cdf F
- Given $x_1 \le x_2 \in \mathbb{R}$, note that

$$(x_1 - T_0(x_1))^2 + (x_2 - T_0(x_2))^2 \le (x_1 - T_0(x_2))^2 + (x_2 - T_0(x_1))^2$$
$$\Leftrightarrow T_0(x_1) \le T_0(x_2)$$

  Expect $T_0(\cdot)$ to be monotone and $T_0(X) \stackrel{d}{=} \text{Unif}[0,1]$.

- $T_0(\cdot)$ is the distribution function of X, say $F(\cdot)$
- Note that increasing functions can be viewed as "derivatives" of convex functions.

- Suppose $T_0 \in C^\alpha$ (Hölder or Sobolev class), $\alpha > 1$
- The minimax rate of convergence is

$$n^{-\frac{2\alpha}{2\alpha-2+d}} + n^{-1}$$

- Suppose $T_0 \in C^\alpha$ (Hölder or Sobolev class), $\alpha > 1$
- The minimax rate of convergence is

$$n^{-\frac{2\alpha}{2\alpha-2+d}} + n^{-1}$$

- In ongoing work, we can show that using a kernel density based estimator yields the optimal rate (up to log-factors)

$$\frac{1}{n}\sum_{i=1}^{n} \mathbb{E}\|\hat{T}(X_i) - T_0(X_i)\|^2 \lesssim n^{-\frac{2\alpha}{2\alpha-2+d}} + n^{-1}$$

- In Manole et al. (2021), Hütter and Rigollet (2019), wavelet based estimators have been used to get optimal rates skip

- $T_0(X) \sim Y$, $X \sim \mu$ with density f, $Y \sim \nu$ with density g. Then change of variable formula implies

$$g(T(x))\det(J_{T_0}(x)) = f(x)$$

- Estimating anti-derivative of f, g related to estimating T_0

- $T_0(X) \sim Y$, $X \sim \mu$ with density f, $Y \sim \nu$ with density g. Then change of variable formula implies

$$g(T(x))\det(J_{T_0}(x)) = f(x)$$

- Estimating anti-derivative of f, g related to estimating T_0

- (Caffarelli Regularity, 1992, 1996) — $T_0 \in C^\alpha$ corresponds to $f, g \in C^{\alpha-1}$, $\alpha > 1$

- $T_0(X) \sim Y$, $X \sim \mu$ with density f, $Y \sim \nu$ with density g. Then change of variable formula implies

$$g(T(x))\mathrm{det}(J_{T_0}(x)) = f(x)$$

- Estimating anti-derivative of f, g related to estimating T_0

- (Caffarelli Regularity, 1992, 1996) — $T_0 \in C^\alpha$ corresponds to $f, g \in C^{\alpha-1}$, $\alpha > 1$

- Goal is to estimate the anti-derivative of $C^{\alpha-1}$ functions

- $T_0(X) \sim Y$, $X \sim \mu$ with density f, $Y \sim \nu$ with density g. Then change of variable formula implies

$$g(T(x))\det(J_{T_0}(x)) = f(x)$$

- Estimating anti-derivative of f, g related to estimating T_0

- (Caffarelli Regularity, 1992, 1996) — $T_0 \in C^\alpha$ corresponds to $f, g \in C^{\alpha-1}$, $\alpha > 1$

- Goal is to estimate the anti-derivative of $C^{\alpha-1}$ functions

- (Müller and Gasser, 1979) — optimal minimax lower bounds for estimating k-th derivative of β-smooth functions is

$$n^{-\frac{2(\beta-k)}{2\beta+d}}$$

- $T_0(X) \sim Y$, $X \sim \mu$ with density f, $Y \sim \nu$ with density g. Then change of variable formula implies

$$g(T(x))\det(J_{T_0}(x)) = f(x)$$

- Estimating anti-derivative of f, g related to estimating T_0

- (Caffarelli Regularity, 1992, 1996) — $T_0 \in C^\alpha$ corresponds to $f, g \in C^{\alpha-1}$, $\alpha > 1$

- Goal is to estimate the anti-derivative of $C^{\alpha-1}$ functions

- (Müller and Gasser, 1979) — optimal minimax lower bounds for estimating k-th derivative of β-smooth functions is

$$n^{-\frac{2(\beta-k)}{2\beta+d}}$$

- Use $\beta = \alpha - 1$ and $k = -1$ (anti-derivative), the lower bound reduces to $\frac{2\alpha}{2\alpha-2+d}$ skip

- Choose the reference distribution μ as *spherical uniform*

- Choose the reference distribution μ as *spherical uniform*

- If X is spherically symmetric, then

$$R(X) = \frac{X}{\|X\|} G(\|X\|)$$

where G is the dist. fn. of $\|X\|$

- Choose the reference distribution μ as *spherical uniform*
- If X is spherically symmetric, then

$$R(X) = \frac{X}{\|X\|} G(\|X\|)$$

where G is the dist. fn. of $\|X\|$

- Pooled: $\|X_1\|, \ldots, \|X_m\|, \|Y_1\|, \ldots, \|Y_n\|$. Let $G_{m,n}$ be the empirical cdf of the pooled data

- Choose the reference distribution μ as *spherical uniform*
- If X is spherically symmetric, then

$$R(X) = \frac{X}{\|X\|} G(\|X\|)$$

where G is the dist. fn. of $\|X\|$

- Pooled: $\|X_1\|, \ldots, \|X_m\|, \|Y_1\|, \ldots, \|Y_n\|$. Let $G_{m,n}$ be the empirical cdf of the pooled data
- Modified Rank Hotelling T^2:

$$\mathrm{RT}^2_{m,n} := T^2_{m,n}\left(\frac{X_1}{\|X_1\|} G_{m,n}(\|X_1\|), \ldots, \frac{Y_1}{\|Y_1\|} G_{m,n}(\|Y_1\|), \ldots, \right)$$

- Test is distribution-free — if $Y \stackrel{d}{=} X - \theta$ then detection boundary at $\|\theta\| \sim \sqrt{d/n}$ skip

$d > 1$ — A step in the right direction

Brenier, '91, McCann '95, Polar Factorization Theorem

Assume that P is absolutely continuous on $\mathbb{R}^d$, Then there exists a unique (P a.s.) $T_0 : \mathbb{R}^d \to \mathbb{R}^d$ such that $T_0(\cdot)$ is the gradient of a convex function and

$$T_0 \# P = Q.$$

If both P and Q have finite second moments, then $T_0(\cdot)$ solves

$$\min_{T\#P=Q} \int \|x - T(x)\|^2 \, dP(x).$$

- Existence of $T_0(\cdot)$ does not require any moment assumptions
- Uniqueness: $T_0 \# P = Q$ and $T_0 \# R = Q$ will imply $P = R$

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 $U \sim \text{Uniform}([0,1])$

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

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

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

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

- 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

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

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

Multivariate rank functions as transport maps

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

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

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

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

Multivariate rank functions as transport maps

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

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

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

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

- **Data**: $X_1, \ldots, X_n$ iid ν (abs. cont.) on $\mathbb{R}^d$
- $\{c_1, \ldots, c_n\} \subset \mathbb{R}^d$ — grid of "reference" points

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

$$\hat{\mathbf{R}}_{\mathbf{n}} = \underset{\sigma \in S_n}{\arg\min} \frac{1}{n} \sum_{i=1}^{n} \|X_i - c_{\sigma(i)}\|^2 \equiv \underset{\mathbf{T}}{\arg\min} \frac{1}{n} \sum_{i=1}^{n} \|X_i - \mathbf{T}(X_i)\|^2$$

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

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

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

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

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

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

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

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

- Even when $\mathbb{E}\|X\|^2 = +\infty$, 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

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

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

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

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

- Even when $\mathbb{E}\|X\|^2 = +\infty$, 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

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

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

Regularity: L_2-convergence [D. and Sen, JASA 2020]

$X_1, \ldots, X_n$ iid ν (abs. cont.). If $\frac{1}{n}\sum_{i=1}^{n} \delta_{c_i} \xrightarrow{w} \text{Unif}([0,1]^d)$, then

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

Result gives the required regularity of the empirical multivariate rank map

Population version

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

Rank energy distance [D. and Sen, JASA 2020]

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

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

where $\mathbf{Z} \sim \lambda P + (1-\lambda) Q$.

Population version

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

Rank energy distance [D. and Sen, JASA 2020]

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

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

  where $\mathbf{Z} \sim \lambda P + (1-\lambda) Q$.

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

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

Population version

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

Rank energy distance [D. and Sen, JASA 2020]

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

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

  where $\mathbf{Z} \sim \lambda P + (1-\lambda) Q$.

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

- $\mathrm{RE}_\lambda = 0$ iff $P = Q$ provided P, Q 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, ...

Population version

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

Rank energy distance [D. and Sen, JASA 2020]

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

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

  where $\mathbf{Z} \sim \lambda P + (1-\lambda) Q$.

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

- $\mathrm{RE}_\lambda = 0$ iff $P = Q$ provided P, Q 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 $X_1, \ldots, X_n \sim \mathrm{P}_{\theta_1}$ and $Y_1, \ldots, Y_m \sim \mathrm{P}_{\theta_2}$, with $m/(m+n) = \lambda \in (0,1)$. We want to test:

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

Pitman efficiency

- Consider $X_1, \ldots, X_n \sim \mathrm{P}_{\theta_1}$ and $Y_1, \ldots, Y_m \sim \mathrm{P}_{\theta_2}$, with $m/(m+n) = \lambda \in (0,1)$. We want to test:

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

- Fix α (size) and $\gamma > \alpha$ (power).
- Two test functions $T_{m,n}$ and $S_{m,n}$.

Pitman efficiency

- Consider $X_1, \dots, X_n \sim \mathrm{P}_{\theta_1}$ and $Y_1, \dots, Y_m \sim \mathrm{P}_{\theta_2}$, with $m/(m+n) = \lambda \in (0,1)$. We want to test:

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

- 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:

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

Pitman efficiency

- Consider $X_1, \ldots, X_n \sim \mathrm{P}_{\theta_1}$ and $Y_1, \ldots, Y_m \sim \mathrm{P}_{\theta_2}$, with $m/(m+n) = \lambda \in (0,1)$. We want to test:

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

- 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:

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

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

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