

London School of Economics and Political Science

Lecture Notes for Statistics Seminar

2021 FALL

(Scribed and Edited by YANG XUZH1)

LASTEST UPDATED: October 24, 2021, 10:23

Contents

1	Preliminary	1
1.1	Statistical Depth	1
1.2	Optimal Transport	4
1.2.1	Background and Formulations	4
1.2.2	Kantorovich Duality	5
2	Monge-Kantorovich Depth, Quantiles, Ranks and Signs (Chernozhukov et al. 2017)	9
2.1	Motivation	9
2.2	Idea	9
2.3	Definition of MK Depth, Quantiles, Ranks and Signs	10
2.3.1	Step 1: Find an optimal transport	10
2.3.2	Step 2: Define MK vector quantiles and vector ranks	12
2.3.3	Step 3: Define MK depth, quantiles, ranks and signs	12
2.4	Empirical Depth, Ranks and Quantiles	13
2.4.1	Empirical vector quantiles and ranks	13
3	Optimal Transport and Its Application on Statistical Methods (Deb et al. 2021)	15
3.1	Motivation	15
3.2	Contributions	15
3.3	Definition of Multivariate Rank Based on OT.	15
3.3.1	Population rank and quantile	15
3.3.2	Empirical rank and quantile	15
3.4	Pointwise Convergence	16
3.5	Multivariate Rank-Based Nonparametric Testing	16
3.5.1	Rank distance covariance (mutual independence testing)	16
	Bibliography	18

Lecture 1


Preliminary

1.1 Statistical Depth

The concept of statistical depth was introduced in order to overcome the lack of a canonical ordering in \mathbb{R}^d for $d > 1$, hence the absence of the related notions of quantile and distribution functions, ranks and signs. The earliest and most popular depth concept is *Tukey's halfspace depth* (see [Tukey \[1975\]](#)).

Definition 1.1.1

Let $X = \{x_1, x_2, \dots, x_n\}$ be a finite set of data points in \mathbb{R}^d and let x be an arbitrary point, not necessary in X . The *depth* of x relative to X is defined as the smallest number of points of X lying in any closed halfspace determined by a line through x (see Figure 1.1 from [Miller et al. \[2003\]](#)).

 **Remark 1.** In detail, the halfspace depth $D_p^{Tukey}(x)$ can be written as

$$D_p^{Tukey}(x) \triangleq \min_{\phi \in S^{d-1}} \# \left\{ i : (x_i - x)^T \phi \geq 0 \right\},$$

where S^{d-1} denotes the the unit sphere in \mathbb{R}^d .

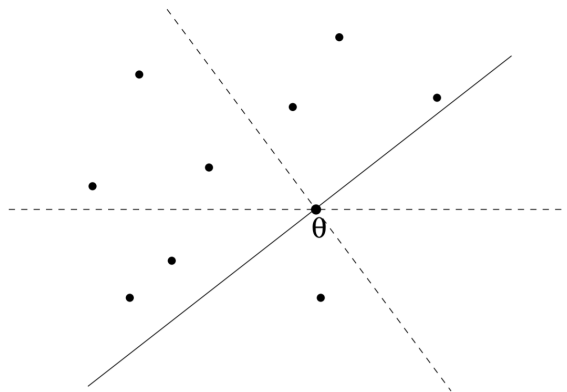


Figure 1.1: The point x (which is not a data point) as depth 1.

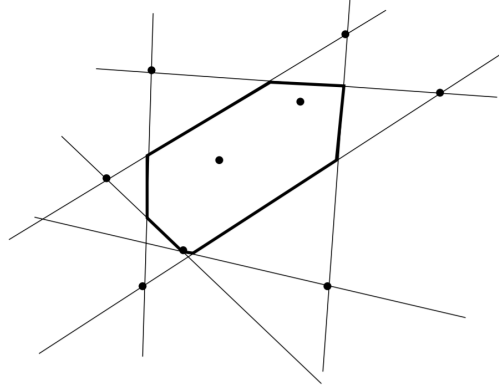



Figure 1.2: The boundary of depth contour 2 is drawn in bold.

Remark 2. It can be seen that the more x is centrally located, the higher its depth. And the depth can be at most $\lfloor n/2 \rfloor$ when X is symmetric about x , e.g., in 1-d case, the median of X has the largest value of depth. Hence, the depth ordering qualifies as a *center-outward ordering* of points in \mathbb{R}^d relative to the center given by the set of the deepest points, $\arg \sup_{x \in \mathbb{R}^d} D_P^{\text{Tukey}}(x)$. 

Then Tukey considered the use of *contour* of depth for indicating the shape of two-dimensional datasets.

Definition 1.1.2

For a fixed positive integer k , the *k -th depth region* $\mathbb{C}_P(k)$ is the set of all $x \in \mathbb{R}^d$ with $D_P^{\text{Tukey}}(x) \geq k$ (see Figure 1.2 from Miller et al. [2003]).


Remark 3. By construction, the depth region are nested:

$$\forall (d, d') \in \mathbb{R}_+^2, \quad d' \geq d \Rightarrow \mathbb{C}_P(d') \subset \mathbb{C}_P(d).$$



Remark 4. In fact, $\mathbb{C}_P(k)$ is the intersection of all the d -dimensional halfspaces containing $n - k + 1$ points of dataset X . In detail,

$$\begin{aligned} \mathbb{C}_P(k) &= \left\{ x \in \mathbb{R}^d : D_P^{\text{Tukey}}(x) \geq k \right\} \\ &= \left\{ x \in \mathbb{R}^d : \min_{\phi \in \mathcal{S}^{d-1}} \# \left\{ i : (x_i - x)^T \phi \geq 0 \right\} \geq k \right\} \\ &= \bigcap_{\phi \in \mathcal{S}^{d-1}} \left\{ x \in \mathbb{R}^d : \# \left\{ i : (x_i - x)^T \phi \geq 0 \right\} \geq k \right\}. \end{aligned} \quad (1.1.1)$$

Thus, $\mathbb{C}_P(k)$ is convex. This implies that halfspace depth contours *cannot* pick non-convex features in the geometry of the underlying distribution, as illustrated in Figure 1.3. This feature is shared by most existing depth concepts and might be considered undesirable for distributions with non-convex support. 

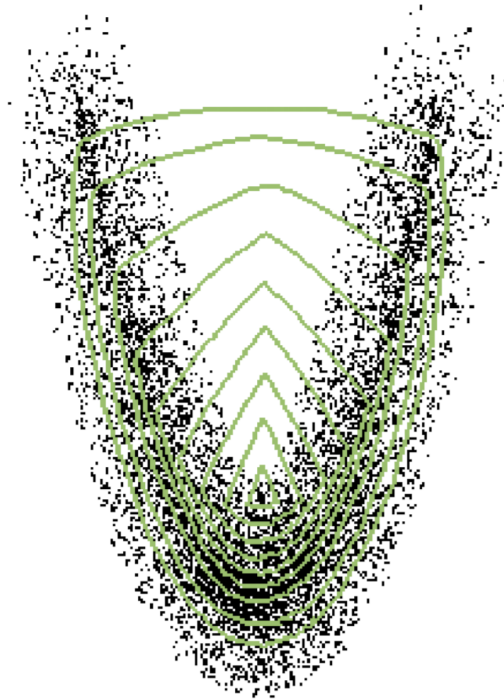




Figure 1.3: Tukey halfspace depth contours for a banana-shaped distribution.

Definition 1.1.3

For $\tau \in [0, 1]$, *the depth region with probability content at least τ* is

$$\mathbb{K}_P(\tau) \triangleq \mathbb{C}_P(d(\tau)), \quad d(\tau) \triangleq \sup\{d \in \mathbb{R} : P(\mathbb{C}(d)) \geq \tau\};$$

the corresponding contour region is the boundary $\mathcal{K}_P(\tau) \triangleq \partial\mathbb{K}_P(\tau)$

 **Remark 5.** The $d(\tau)$ can be taken as the minimal depth d such that the probability content of $\mathbb{C}_P(d)$ is larger than or equal to τ . Thus for any $x_0 \in \mathbb{K}_P(\tau)$, x_0 should lie in the cone. 

Four axioms are proposed by [Liu \[1990\]](#) and [Zuo and Serfling \[2000\]](#) to unify the diverse depth functions $D_P(x)$:

- (Affine invariance) $D_{P_{Ax+b}}(Ax + b) = D_{P_X}(x)$, for any $x \in \mathbb{R}^d$, any non-singular $d \times d$ matrix A , and any $b \in \mathbb{R}^d$.
- (Maximality at the center) If x_0 is a center of symmetry for P , it is deepest.
- (Linear monotonicity relative to the deepest point) If x_0 is deepest, then $D_P(x) \leq D_P((1 - \alpha)x_0 + \alpha x)$ for all $\alpha \in [0, 1]$
- (Vanishing at infinity) $\lim_{\|x\| \rightarrow \infty} D_P(x) = 0$.

It can be shown that halfspace depth relative to any distribution with non-vanishing density on \mathbb{R}^d satisfies the four axioms above.

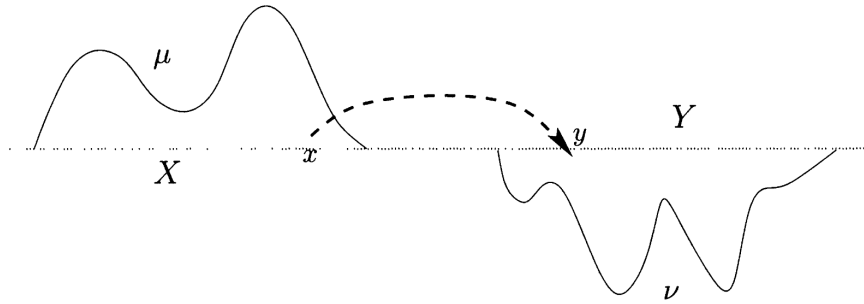


Figure 1.4: The Monge's mass transportation problem.

1.2 Optimal Transport

1.2.1 Background and Formulations

Case 1: Monge Problem (Monge formulation)

Assuming that we are giving a pile of sand, and a hole that we have to completely fill up with the sand. We shall model the pile and the hole with measure space (X, \mathcal{U}, μ) and (Y, \mathcal{Y}, ν) respectively (Figure 1.4). Moving the sand around needs some effort, which is modelled by a measurable cost function $c(x, y)$ defined on $X \times Y$.

? Problem 1.2.1

We aim to find a measurable **map** $T : X \rightarrow Y$ to minimize

$$I(T) = \int_X c(x, T(x)) d\mu(x),$$

under the constrain: $\mu(T^{-1}(A)) = \nu(A)$, for any $A \in \mathcal{Y}$, we shall write as $T\#\mu = \nu$ and say that ν is the **push-forward** of μ by T .

Case 2: Mines and Factories (Kantorovich formulation)

Suppose that we have a collection of m mines mining iron ore and a collection of n factories. We still seek for the most economical way to transport the iron ore to the factories. However, because T is a map and the number of mines is not coincide with the number of factories, Monge's formulation is not feasible for this situation. Hence, instead of seeking for a map, we seek for a **transportation plan**.

We can use a probability measure $\pi(x, y)$ on the product space $X \times Y$ to model the transport plan. Thus, the problem becomes


? Problem 1.2.2

Minimize

$$I(\pi) = \int_{X \times Y} c(x, y) d\pi(x, y),$$

where $\pi \in \Pi(\mu, \nu) = \{\pi \in P(X \times Y) : \mu \text{ and } \nu \text{ are marginal of } \pi\}$.

Remark 6. The set $\Pi(\mu, \nu)$ is not empty, since the tensor product of μ and ν lies in $\Pi(\mu, \nu)$ (this seems to be the most stupid transport plan: every unit sand of X , regardless of its location, is dis-

tributed over all the factories). In this way, we do not exclude the probability that some mass located at a point x may be split into several parts. 

1.2.2 Kantorovich Duality

Theorem 1.2.1

Let X and Y be polish spaces, let $\pi \in P(X \times Y)$, $\mu \in P(X)$ and $\nu \in P(Y)$, and let $c : X \times Y \rightarrow \mathbb{R}_+ \cup \{+\infty\}$ be a lower semi-continous cost function. Whenever $(\phi, \psi) \in L^1(\mu) \times L^1(\nu)$, define



$$J(\phi, \psi) = \int_X \phi d\mu + \int_Y \psi d\nu,$$

and

$$\Phi_c = \left\{ (\phi, \psi) \in L^1(\mu) \times L^1(\nu) : \phi(x) + \psi(y) \leq c(x, y), \text{ a.s. } - \mu, \text{ a.s. } - \nu \right\}.$$

Then

$$\inf_{\Pi(\mu, \nu)} I[\pi] = \sup_{\Phi_c} J(\phi, \psi) \quad (1.2.1)$$

 **Remark 7.** Here is an intuitive interpretation of Theorem 1.2.1 from Villani [2021]. Suppose for instance that you are both a mathematician and an industrialist, and want to transfer a huge amount of coal from your mines to your factories. You can hire trucks to do this transportation problem, but you have to pay them $c(x, y)$ for each ton of coal which is transported from place x to place y . Both the amount of coal which you can extract from each mine, and the amount which each factory should receive, are fixed. As you are trying to solve the associated Monge-Kantorovich problem in order to minimize the price you have to pay, another mathematician comes to you and tells you "My friend, let me handle this for you: I will ship all your coal with my own trucks and you won't have to worry about what goes where. I will just set a price $\phi(x)$ for loading one ton of coal at place x , and a price $\psi(y)$ for unloading it at destination y . I will set the prices in such a way that your financial interest will be to let me handle all your transportation! Indeed, you can check very easily that for any x and y , the sum $\phi(x) + \psi(y)$ will always be less than the cost $c(x, y)$. 

Proof. We can rewrite the left hand side of 1.2.1 as

$$\inf_{\pi \in \Pi} I(\pi) = \inf_{\pi} \left[I(\pi) + \begin{cases} 0 & \text{if } \pi \in \Pi \\ +\infty & \text{else} \end{cases} \right].$$

Note that

$$\begin{cases} 0 & \text{if } \pi \in \Pi \\ +\infty & \text{else} \end{cases} = \sup_{(\phi, \psi) \in C_b(X) \times C_b(Y)} \left\{ \int \phi(x) d\mu(x) + \int \psi d\nu(y) - \int [\phi(x) + \psi(y)] d\pi(x, y) \right\}.$$

Thus take **minimax principle** as granted, we have

$$\inf_{\pi \in \Pi} I(\pi) = \inf_{\pi} \sup_{(\phi, \psi) \in C_b(X) \times C_b(Y)} \left[I(\pi) + \int \phi(x) d\mu(x) + \int \psi d\nu(y) - \int [\phi(x) + \psi(y)] d\pi(x, y) \right]$$

$$\begin{aligned}
&= \sup_{(\phi, \psi) \in C_b(X) \times C_b(Y)} \inf_{\pi} \left[\int \phi(x) d\mu(x) + \int \psi(y) d\nu(y) - \int [\phi(x) + \psi(y) - c(x, y)] d\pi(x, y) \right] \\
&= \sup_{(\phi, \psi) \in C_b(X) \times C_b(Y)} \left\{ \int \phi(x) d\mu(x) + \int \psi(y) d\nu(y) - \sup_{\pi} \int [\phi(x) + \psi(y) - c(x, y)] d\pi(x, y) \right\} \\
&= \sup_{(\phi, \psi) \in C_b(X) \times C_b(Y)} \left\{ J(x, y) - \begin{cases} 0 & \text{if } (\phi, \psi) \in \Phi_c \\ +\infty & \text{else} \end{cases} \right\} \\
&= \sup_{(\phi, \psi) \in \Phi_c} J(x, y).
\end{aligned}$$

✎

Another way to motivate Kantorovich's duality is by analogy with the finite dimensional case (see [Evans \[1997\]](#)). Firstly, we introduce finite dimensional linear programming by using minimax principle again.

✎ Example 1. Finite dimensional linear programming

We may prove the following linear programming duality by using minimax principle:

$$\sup_{Ax \leq b} c \cdot x = \inf_{y \geq 0, A^\top y = c} b \cdot y, \quad (1.2.2)$$

where c and b are constant vectors in \mathbb{R}^n .

$$\text{LHS} = \sup_x \left[c \cdot x + \begin{cases} 0 & \text{if } Ax \leq b \\ -\infty & \text{else} \end{cases} \right], \quad (1.2.3)$$

where we can rewrite function appears inside the bracket as

$$\begin{cases} 0 & \text{if } Ax \leq b \\ -\infty & \text{else} \end{cases} = \inf_{y \geq 0} [y \cdot (b - Ax)].$$

Thus, we have

$$\begin{aligned}
(1.2.3) &= \sup_x \left\{ c \cdot x + \inf_{y \geq 0} [y \cdot (b - Ax)] \right\} \\
&= \sup_x \inf_{y \geq 0} \{ c \cdot x + [y \cdot (b - Ax)] \} \\
&= \inf_{y \geq 0} \sup_x \{ y \cdot b + c \cdot x - y \cdot (Ax) \} \\
&= \inf_{y \geq 0} \left\{ y \cdot b + \sup_x [(c^\top - y^\top A)x] \right\} \\
&= \inf_{y \geq 0} \left[y \cdot b + \begin{cases} 0 & \text{if } A^\top y = c \\ +\infty & \text{else} \end{cases} \right] \\
&= \inf_{y \geq 0, A^\top y = c} b \cdot y.
\end{aligned}$$

The finite dimensional version of Problem 1.2.2 can be formulated as follow. Assume $X = \{x_i\}_{i=1}^n$ and $Y = \{y_j\}_{j=1}^m$ are two discrete sample spaces with measure μ and ν , where $\mu(x_i) = \mu_i$ and $\nu(y_j) = \nu_j$

are positive. And the product space $(X \otimes Y, \mathcal{X} \otimes \mathcal{Y}, \mu \otimes \nu)$, where we define $\mu \otimes \nu(x_i, y_j) = \pi_{ij}$. Thus, the finite dimensional version of Problem 1.2.2 can be write as

$$\min_{\pi_{ij}} \sum_{i=1}^n \sum_{j=1}^m c_{ij} \pi_{ij}, \quad (1.2.4)$$

under the constraint

$$\sum_{j=1}^m \pi_{ij} = \mu_i, \quad \sum_{i=1}^n \pi_{ij} = \nu_j,$$

where c_{ij} are postive constants. In fact, this formula can be rewritten as (1.2.2), by taking

$$\begin{aligned} b &= (c_{11}, c_{12}, \dots, c_{nm})^\top, \\ y &= (\pi_{11}, \pi_{12}, \dots, \pi_{nm})^\top, \\ c &= (\mu_1, \mu_2, \dots, \mu_n, \nu_1, \dots, \nu_m)^\top, \end{aligned}$$

and

$$A = \begin{bmatrix} \mathbb{1}_m & 0 & 0 & 0 & \dots & 0 & \mathbf{e}_1^{(n)} & \mathbf{e}_2^{(n)} & \dots & \mathbf{e}_m^{(n)} \\ 0 & \mathbb{1}_m & 0 & 0 & \dots & 0 & \mathbf{e}_1^{(n)} & \mathbf{e}_2^{(n)} & \dots & \mathbf{e}_m^{(n)} \\ 0 & 0 & \mathbb{1}_m & 0 & \dots & 0 & \mathbf{e}_1^{(n)} & \mathbf{e}_2^{(n)} & \dots & \mathbf{e}_m^{(n)} \\ & & & & \dots & & & & & \\ 0 & 0 & 0 & 0 & \dots & \mathbb{1}_m & \mathbf{e}_1^{(n)} & \mathbf{e}_2^{(n)} & \dots & \mathbf{e}_m^{(n)} \end{bmatrix}.$$

Thus, take $x = (u_1, \dots, u_n, v_1, \dots, v_m)^\top$ and according to LHS of (1.2.2), the dual problem of (1.2.4) is

$$\sup_{u, v} \left[\sum_{i=1}^n \mu_i u_i + \sum_{j=1}^m \nu_j v_j \right]$$

under the constraint

$$u_i + v_j \leq c_{ij}, \text{ for all } i, j,$$

which is discrete version of kantorovich duality.

Lecture 2

Monge-Kantorovich Depth, Quantiles, Ranks and Signs

(Chernozhukov et al. 2017)

2.1 Motivation

“An important feature of halfspace depth is the convexity of its contours, which thus satisfy the star-convexity requirement embodied in the linear monotonicity axiom. That feature is shared by most existing depth concepts and might be considered undesirable for distributions with non-convex supports or level contours, and multi-modal ones”

2.2 Idea

Case 1: Univariate Distribution Family \mathcal{P}^1 .

The first case is that of the family \mathcal{P}^1 of all univariate distributions with non-vanishing Lebesgue densities. Here, the definition of quantile and distribution function, ranks and signs are the "classical" univariate ones.



Case 2: Elliptical Distribution Family \mathcal{P}_{ell}^d over \mathbb{R}^d .

Let $X \sim P_{\mu, \Sigma, g} \in \mathcal{P}_{ell}^d$, where μ and Σ are location parameter and scatter parameter respectively, and g is radial density (with radial distribution G). Then the following are equivalent:

$$X \sim P_{\mu, \Sigma, g} \iff Y \triangleq \Sigma^{-1/2}(X - \mu) \sim P_{0, I, g} \iff R_P(Y) \triangleq \frac{Y}{\|Y\|} G(\|Y\|) \sim U_d,$$

where U_d denotes the spherical uniform distribution (RP: $U_d = r\phi, r \perp \phi$).

- Note that the spherical contours with $P_{\mu, I, g}$ -probability contents τ coincide with the halfspace depth contours.
- Thus we can naturally define a τ -quantile contour as the contour of the depth region with probability content at least τ (contour of equal depth of τ).
- $R_P(Y)$, $R_P(Y)/\|R_P(Y)\|$ and $\|R_P(Y)\|$ play the roles of *vector ranks*, *signs* and *ranks*, respectively.
- The inverse map $u \longrightarrow Q_P(u)$ of the vector rank map $y \longrightarrow R_P(y)$ is called *vector quantile map*

 **Remark 8.** Take a closer look at the case, we can observe that the τ -quantile contour (contour of the depth region with probability content τ) is $Q_P(\mathcal{S}(\tau))$, while the depth region with probability content τ is $Q_P(\mathbb{S}(\tau))$. 

Case 3: General Distribution P over \mathbb{R}^d .

From the insights we gained through the traditional cases, for general distribution P over \mathbb{R}^d , we also want to define a map Q_P that transform the spherical uniform distribution U_d into the target distribution. The following theorem due to [Brenier \[1991\]](#) and [McCann \[1995\]](#) guarantees the existence of such map.



2.3 Definition of MK Depth, Quantiles, Ranks and Signs

2.3.1 Step 1: Find an optimal transport

Theorem 2.3.1

Let P and F be two distributions on \mathbb{R}^d .

- (1) If F is absolutely continuous with respect to the Lebesgue measure on \mathbb{R}^d , the following holds: for any convex set $\mathcal{U} \subset \mathbb{R}^d$ containing the support of F , there exists a convex function $\psi : \mathcal{U} \rightarrow \mathbb{R} \cup \{+\infty\}$ such that $\nabla\psi \# F = P$. The gradient $\nabla\psi$ of that function exists and unique, F -a.e.
- (2) If, in addition, P is absolutely continuous on \mathbb{R}^d the following holds: for any convex set $\mathcal{Y} \subset \mathbb{R}^d$ containing the support of P , there exists a convex function $\psi^* : \mathcal{Y} \rightarrow \mathbb{R} \cup \{+\infty\}$ such that $\nabla\psi^* \# P = F$. The gradient $\nabla\psi^*$ of ψ^* exists and unique, and $\nabla\psi^* = \nabla\psi^{-1}$, P -a.e.

 **Remark 9.** Note that although ψ and ψ^* may not be unique, the map $\nabla\psi$ and $\nabla\psi^*$ is unique a.e. 

We will see that if P and F have **finite second moments**, then $\nabla\psi$ is an **optimal transport plan** from F to P under *quadratic cost*. That is,

$$\nabla\psi = \arg \inf_{Q \# F = P} \int (u - Q(u))^2 dF(y) \quad (2.3.1)$$

The following proposition will help us to establish a dual problem of (2.3.1).

Proposition 2.3.1

If μ and ν be two probability measures on \mathbb{R}^d with **finite second moments**, then (1.2.1) with quadratic cost becomes

$$\sup_{\Pi(\mu, \nu)} \int (x \cdot y) d\pi(x, y) = \inf_{\Phi} J(\phi, \psi), \quad (2.3.2)$$

where $\Phi = \{(\phi, \psi) \in L^1(d\mu) \times L^1(d\nu) : x \cdot y \leq \phi(x) + \psi(y) \text{ a.e.}\}$.

Proof. From the definition of Φ_C , we have for any $(\phi, \psi) \in \Phi_C$,

$$\phi(x) + \psi(y) \leq \frac{|x - y|^2}{2},$$

this can be rewritten as

$$x \cdot y \leq \left[\frac{|x|^2}{2} - \phi(x) \right] + \left[\frac{|y|^2}{2} - \psi(y) \right].$$


If we define

$$\tilde{\phi}(x) = \frac{|x|^2}{2} - \phi(x), \quad \tilde{\psi}(y) = \frac{|y|^2}{2} - \psi(y),$$

and forget about the \sim symbol, then we can see

$$\inf_{\Pi(\mu, \nu)} I[\pi] = M_2 - \sup_{\Pi(\mu, \nu)} \int (x \cdot y) d\pi(x, y) \quad (2.3.3)$$

$$\sup_{\Phi_c} J(\phi, \psi) = M_2 - \inf_{\tilde{\Phi}} J(\phi, \psi), \quad (2.3.4)$$


where $M_2 \triangleq \int_{\mathbb{R}^n} \frac{|x|^2}{2} d\mu(x) + \int_{\mathbb{R}^n} \frac{|y|^2}{2} d\nu(y) < +\infty$. Thus from (2.3.3) and (2.3.4) we obtained (2.3.2). 

Definition 2.3.1

For any function $\psi : \mathcal{U} \rightarrow \mathbb{R} \cup \{+\infty\}$, the **conjugate** $\psi^* : \mathcal{Y} \rightarrow \mathbb{R} \cup \{+\infty\}$ of ψ is defined as

$$\psi^*(y) \triangleq \sup_{u \in \mathcal{U}} [u \cdot y - \psi(u)], \quad y \in \mathcal{Y}. \quad (2.3.5)$$

Call (ψ, ψ^*) **conjugate pair of potentials** over $(\mathcal{U}, \mathcal{Y})$.

 **Remark 10.** The transformation in (2.3.5) is called **Legendre Transformation**. From the definition we can see that any $(\psi, \psi^*) \in \tilde{\Phi}$. It can be proved that if ψ is differential and strict convex, we have

$$\nabla \psi^* = (\nabla \psi)^{-1}, \quad (2.3.6)$$

the proof can be found at the Wikipedia page of Legendre Transformation. 

The following theorem states that the infimum of J on $\tilde{\Phi}$ is unchanged if one restricts J to the very small subset of $\tilde{\Phi}$ which is made of all the conjugate pairs.

Theorem 2.3.2

Let μ, ν be two "proper" probability measures on \mathbb{R}^n , with finite second moments.


- (i) Then there exists a pair (ψ, ψ^*) of lower semi-continuous proper conjugate convex function on \mathbb{R}^n such that

$$\inf_{\tilde{\Phi}} J = J(\psi, \psi^*); \quad (2.3.7)$$

- (ii) $\nabla\psi$ and $\nabla\psi^*$ are the unique solutions of the Monge problem for transport μ to ν and transport ν to μ with a quadratic cost function, respectively;
- (iii) according to (2.3.6), for $d\mu$ -almost all x and $d\nu$ -almost all y ,

$$\nabla\psi^* \circ \nabla\psi(x) = x$$

$$\nabla\psi \circ \nabla\psi^*(y) = y$$

 **Remark 11.** From this theorem, we know

- (i) where do the $\nabla\psi$ and $\nabla\psi^*$ in Theorem 2.3.1 come from and how to calculate them;
- (ii) if P and F in Theorem 2.3.1 have finite second moment condition, then $\nabla\psi$ and $\nabla\psi^*$ are optimal transport under quadratic cost function.
- (iii) why the gradient $\nabla\psi^*$ is the inverse of $\nabla\psi$.



2.3.2 Step 2: Define MK vector quantiles and vector ranks

To define a multivariate rank on the interest distribution P based on the reference distribution F , we only need to find the unique transport R_P from P to F , and the corresponding inverse transport Q_P will naturally lead to the definition of quantile function.

Definition 2.3.2

Denote by $\nabla\psi$ the F -almost surely unique gradient of a convex function ψ in (1) of Theorem 2.3.1, and let ψ^* be the conjugate of ψ . Then we define *vector quantile* Q_P and *vector ranks* R_P as

$$Q_P(u) \in \arg \sup_{y \in \mathcal{Y}} [y^\top u - \psi^*(y)], \quad u \in \mathcal{U}$$

$$R_P(y) \in \arg \sup_{u \in \mathcal{U}} [y^\top u - \psi(u)], \quad y \in \mathbb{R}^d$$

 **Remark 12.** By the envelope theorem and Rademacher's theorem (see Villani [2021]), we have

$$Q_P = \nabla\psi \text{ a.e. on } \mathcal{U}, \quad R_P = \nabla\psi^* \text{ a.e. on } \mathcal{Y},$$

and the under some regularity conditions the "a.e." can become "for all".



2.3.3 Step 3: Define MK depth, quantiles, rankds and signs

Let's take $F = U_d$ and P be an arbitrary distribution to deliver the multivariate notion of quantiles and ranks through R_P and Q_P .

Definition 2.3.3

- (1) The MK rank of $y \in \mathbb{R}^d$ is $\|R_P(y)\|$ and the MK sign is $R_P(y)/\|R_P(y)\|$.
- (2) The MK τ -quantile contour is the set $Q_P(\mathcal{S}(\tau))$, and the MK depth region with probability

content τ is $Q_P(\mathbb{S}(\tau))$.

(3) The MK depth of $y \in \mathbb{R}^d$ with respect to P is the depth of $R_P(y)$ under D_P^{Tukey} :

$$D_P^{\text{MK}}(y) := D_{U_d}^{\text{Tukey}}(R_P(y))$$

2.4 Empirical Depth, Ranks and Quantiles

Now, we are ready to discuss the empirical version of MK quantiles, ranks and depth based on estimator \hat{P}_n for P and \hat{F}_n for F .

2.4.1 Empirical vector quantiles and ranks

Definition 2.4.1

The empirical vector quantile \hat{Q}_n and vector rank \hat{R}_n are any pair of functions satisfying, for each $u \in \mathcal{U}$ and $y \in \mathcal{Y}$

$$\begin{aligned}\hat{Q}_n &\in \arg \sup_{y \in \mathcal{Y}} [y^\top u - \hat{\psi}_n^*(y)], \\ \hat{R}_n &\in \arg \sup_{u \in \mathcal{U}} [y^\top u - \hat{\psi}_n(u)],\end{aligned}$$



where $\hat{\psi}_n^*(y)$ and $\hat{\psi}_n(u)$ is the empirical counterpart of (2.3.7).

Now, we introduce a Glivenko-Cantelli Theorem (*uniform convergence*) for empirical MK ranks and quantiles.

Theorem 2.4.1

Suppose that the sets \mathcal{U} and \mathcal{Y} are two compact subsets of \mathbb{R}^d , and the probability measure P and F are absolutely continuous with respect to Lebesgue measure, with $\text{spt}(P) \subset \mathcal{Y}$ and $\text{spt}(F) \subset \mathcal{U}$. Then, as $n \rightarrow \infty$, for any closed set $K \subset \text{int}(\text{spt}(P))$ and any closed set $K' \subset \text{int}(\text{spt}(F))$,

$$\begin{aligned}\sup_{u \in K} \|\hat{Q}_n(y) - Q_P(u)\| &\xrightarrow{P} 0, \\ \sup_{y \in K'} \|\hat{R}_n(y) - R_P(y)\| &\xrightarrow{P} 0.\end{aligned}$$

 **Remark 13.** Although Theorem 2.4.1 gives us a Glivenko-Cantelli consistency, the compactly supported distributions assumption is a relatively strong assumption, which makes it not exact distribution free. 

Lecture 3

Optimal Transport and Its Application on Statistical Methods

(Deb et al. 2021)

3.1 Motivation

Based on Chernozhukov et al. [2017]

3.2 Contributions

The main contributions in this work are listed as followed:

- (I) Construct a novel definition of multivariate rank and quantile from the perspective of assignment problem, based on optimal transport technique, and lead to exact distribution-free, computing feasible testing procedures
- (II) A holistic idea is developed to construct multivariate rank-based distribution-free test.

3.3 Definition of Multivariate Rank Based on OT.

The pioneering work of Chernozhukov et al. [2017] has developed novel definitions of multivariate rank and quantile based on the mathematical solution for *Kantorovich's Problem*. However, from the statistical perspective, we usually only care about the measure transportation between the empirical distribution and the empirical rank distribution, which will lead to a Monge's Problem. Thus the measure transportation can be thought as an assignment problem instead.

3.3.1 Population rank and quantile

3.3.2 Empirical rank and quantile

In standard statistical applications, the population rank map is not available to the practitioner. Thus, given iid random samples $\mathcal{D}_n^X := \{\mathbf{X}_1, \dots, \mathbf{X}_n\}$, and let $\mathcal{H}_n^d := \{\mathbf{h}_1^d, \dots, \mathbf{h}_n^d\}$ denote the fixed set of sample *multivariate rank vectors*. In practice, we may take \mathcal{H}_n^d to be the d -dimensional **Halton sequence** of size n . Let μ_n^X and ν_n denote the empirical distribution of \mathcal{D}_n^X and \mathcal{H}_n^d , respectively.


Definition 3.3.1

We define the empirical rank function $\hat{\mathbf{R}}_n : \mathcal{D}_n^{\mathbf{X}} \rightarrow \mathcal{H}_n^d$ as the optimal transport map which pushes $\mu_n^{\mathbf{X}}$ to ν_n , that is,

$$\hat{\mathbf{R}}_n = \arg \inf_F \int ||\mathbf{X} - F(\mathbf{X})||^2 d\mu_n^{\mathbf{X}}, \quad \text{subject to } F\# \mu_n^{\mathbf{X}} = \nu_n \quad (3.3.1)$$

Remark 14. Note that by F pushes μ to ν , we mean that $F(\mathbf{X}) \sim \nu$, where $\mathbf{X} \sim \mu$. Thus (3.3.1) can be rewritten as

$$\hat{\mathbf{c}}_n = \arg \inf_{\sigma \in \mathcal{S}_n} \frac{1}{n} \sum_{i=1}^n ||\mathbf{X}_i - \mathbf{h}_{\sigma(i)}^d||^2. \quad (3.3.2)$$

The optimization problem (3.3.2) allows us to view it as an **assignment problem** for which algorithms with worst case complexity $\mathcal{O}(n^3)$. 

3.4 Pointwise Convergence

Based on (3.3.2), observe that the sample rank map $\hat{\mathbf{R}}_n$ satisfies

$$\hat{\mathbf{R}}_n(\mathbf{X}_i) = \mathbf{h}_{\hat{\sigma}_n(i)}^d, \quad \text{for } i = 1, \dots, n,$$

and the following theorem shows that the sample rank map converges to its population counterpart under minimal assumptions.

Theorem 3.4.1

Assume $\mathbf{X}_1, \dots, \mathbf{X}_n \stackrel{iid}{\sim} \mu \in \mathcal{P}_{ac}^d$, and $\nu_n \xrightarrow{w} \mathcal{U}^d$, then

$$\frac{1}{n} \sum_{i=1}^n ||\hat{\mathbf{R}}_n(\mathbf{X}_i) - \mathbf{R}_n(\mathbf{X}_i)||_2^2 \xrightarrow{a.s.} 0$$

Remark 15. 

3.5 Multivariate Rank-Based Nonparametric Testing

3.5.1 Rank distance covariance (mutual independence testing)

Part I: Measure

Suppose that Z_1 and Z_2 are real-valued absolutely continuous random variables with distribution function $G_1(\cdot)$ and $G_2(\cdot)$. It can be prove that $Z_1 \perp\!\!\!\perp Z_2$ iff. $G_1(Z_1) \perp\!\!\!\perp G_2(Z_2)$. Thus, it is natural to establish the following measure of dependence:

$$\begin{aligned} \mathcal{R}_w := & \iint | \mathbb{E} \exp (i G_{t,s}(\mathbf{Z})) \\ & - \mathbb{E} \exp (i t G_1\left(Z_1\right)) \mathbb{E} \exp (i s G_2\left(Z_2\right)) |^2 w(t, s) d t d s, \end{aligned} \quad (3.5.1)$$

where $\mathbf{Z} = (Z_1, Z_2)$, $G_{ts} = tG_1(Z_1) + sG_2(Z_2)$ and w is a weight function such that $\mathcal{R}_w < +\infty$.

? Problem 3.5.1


Can we extend \mathcal{R}_w beyond $d = 1$? How to choose the weight function w ?

For the first problem, we only need to replace G_1 and G_2 with the population multivariate ranks. For the second one, we will borrow the weight function used in the *distance covariance*.

Q Definition 3.5.1



Suppose that $\mathbf{Z}_1 \sim \mu_1 \in \mathcal{P}_{ac}(\mathbb{R}^{d_1})$ and $\mathbf{Z}_2 \sim \mu_2 \in \mathcal{P}_{ac}(\mathbb{R}^{d_2})$. Let $\mathbf{R}_1(\cdots)$ and \mathbf{R}_2 denote the corresponding population rank maps. We define the rank distance covariance (RdCov^2) as

$$\begin{aligned} \text{RdCov}^2(\mathbf{Z}_1, \mathbf{Z}_2) &:= \int_{\mathbb{R}^{d_1+d_2}} \\ &\times \frac{|\mathbb{E} \exp(i\mathbf{R}_{\mathbf{t},\mathbf{s}}(\mathbf{Z})) - \mathbb{E} \exp(it^\top \mathbf{R}_1(\mathbf{Z}_1)) \mathbb{E} \exp(is^\top \mathbf{R}_2(\mathbf{Z}_2))|^2}{c(d_1) c(d_2) \|\mathbf{t}\|^{1+d_1} \|\mathbf{s}\|^{1+d_2}} \\ &\times d\mathbf{t} d\mathbf{s}. \end{aligned} \quad (3.5.2)$$

 **Remark 16.** Suppose that $(\mathbf{Z}_1^1, \mathbf{Z}_2^1), (\mathbf{Z}_1^2, \mathbf{Z}_2^2), (\mathbf{Z}_1^3, \mathbf{Z}_2^3)$ are independent observations having the same distribution as $(\mathbf{Z}_1^1, \mathbf{Z}_2^1)$. Then, we have alternative definition

$$\begin{aligned} &\text{RdCov}^2(\mathbf{Z}_1, \mathbf{Z}_2) \\ &= \mathbb{E} \left[\left\| \mathbf{R}_1(\mathbf{Z}_1^1) - \mathbf{R}_1(\mathbf{Z}_1^2) \right\| \left\| \mathbf{R}_2(\mathbf{Z}_2^1) - \mathbf{R}_2(\mathbf{Z}_2^2) \right\| \right] \\ &+ \mathbb{E} \left[\left\| \mathbf{R}_1(\mathbf{Z}_1^1) - \mathbf{R}_1(\mathbf{Z}_1^3) \right\| \right] \mathbb{E} \left[\left\| \mathbf{R}_2(\mathbf{Z}_2^1) - \mathbf{R}_2(\mathbf{Z}_2^3) \right\| \right] \\ &- 2\mathbb{E} \left[\left\| \mathbf{R}_1(\mathbf{Z}_1^1) - \mathbf{R}_1(\mathbf{Z}_1^2) \right\| \left\| \mathbf{R}_2(\mathbf{Z}_2^1) - \mathbf{R}_2(\mathbf{Z}_2^3) \right\| \right] \end{aligned} \quad (3.5.3)$$



 **Remark 17.** Unlike the distance covariance, (3.5.3) does not require any moment assumptions on \mathbf{Z}_1 and \mathbf{Z}_2 . 

Part II: Test procedure

Suppose that $(\mathbf{X}_1, \mathbf{Y}_1), \dots, (\mathbf{X}_n, \mathbf{Y}_n)$ are iid observations from some distribution $\mu \in \mathcal{P}(\mathbb{R}^{d_1+d_2})$ with marginals μ_X and μ_Y . We are interested in testing the hypothesis: $H_0 : \mu = \mu_X \otimes \mu_Y$. The author proposed a procedure that is exactly distribution-free and guarantees consistency against all fixed alternatives based on the rank distance covariance.

We consider the empirical version of (3.5.3) as

$$\text{RdCov}_n^2 := S_1 + S_2 - 2S_3,$$

where

$$S_1 := \frac{1}{n^2} \sum_{k,l=1}^n \left\| \hat{R}_n^X(\mathbf{X}_k) - \hat{R}_n^X(\mathbf{X}_l) \right\| \left\| \hat{R}_n^Y(\mathbf{Y}_k) - \hat{R}_n^Y(\mathbf{Y}_l) \right\|$$

$$S_2 := \left(\frac{1}{n^2} \sum_{k,l=1}^n \left\| \hat{R}_n^{\mathbf{X}}(\mathbf{X}_k) - \hat{R}_n^{\mathbf{X}}(\mathbf{X}_l) \right\| \right) \times \left(\frac{1}{n^2} \sum_{k,l=1}^n \left\| \hat{R}_n^{\mathbf{Y}}(\mathbf{Y}_k) - \hat{R}_n^{\mathbf{Y}}(\mathbf{Y}_l) \right\| \right)$$

$$S_3 := \frac{1}{n^3} \sum_{k,l,m=1}^n \left\| \hat{R}_n^{\mathbf{X}}(\mathbf{X}_k) - \hat{R}_n^{\mathbf{X}}(\mathbf{X}_l) \right\| \left\| \hat{R}_n^{\mathbf{Y}}(\mathbf{Y}_k) - \hat{R}_n^{\mathbf{Y}}(\mathbf{Y}_m) \right\|$$

? Problem 3.5.2

- (a) What is the limiting distribution of the test statistic?
- (b) Is the test consistent against all fixed alternatives, as the sample size grows?

🎓 Theorem 3.5.1

We assume

- (1) $\mu_{\mathbf{X}} \in \mathcal{P}_{ac}(\mathbb{R}^{d_1})$ and $\mu_{\mathbf{Y}} \in \mathcal{P}_{ac}(\mathbb{R}^{d_2})$
- (2) The empirical distribution of $\mathcal{H}_n^{d_1}$ and $\mathcal{H}_n^{d_2}$ converge weakly to \mathcal{U}^{d_1} and \mathcal{U}^{d_2} , respectively.

Then under the null hypothesis, there exists universal nonnegative constants (η_1, η, \dots) such that

$$n\text{RdCov}_n^2 \xrightarrow{w} \sum_{j=1}^{\infty} \eta_j Z_j^2, \quad \text{as } n \rightarrow +\infty,$$

where Z_1, Z_2, \dots are iid standard Gaussian random variables.

🎓 Theorem 3.5.2

Under the same assumptions in Theorem 3.5.1, $\text{RdCov}_n^2 \xrightarrow{a.s.} \text{RdCov}^2(\mathbf{X}, \mathbf{Y})$, where $(\mathbf{X}, \mathbf{Y}) \sim \mu$. Moreover, $\mathbb{P}\left(n\text{RdCov}_n^2 \geq c_n\right) \rightarrow 1$, as $n \rightarrow +\infty$, provided $\mu \neq \mu_{\mathbf{X}} \otimes \mu_{\mathbf{Y}}$.

Bibliography

- Yann Brenier. Polar factorization and monotone rearrangement of vector-valued functions. *Communications on pure and applied mathematics*, 44(4):375–417, 1991.
- Victor Chernozhukov, Alfred Galichon, Marc Hallin, and Marc Henry. Monge–kantorovich depth, quantiles, ranks and signs. *The Annals of Statistics*, 45(1):223–256, 2017.
- Lawrence C Evans. Partial differential equations and monge-kantorovich mass transfer. *Current developments in mathematics*, 1997(1):65–126, 1997.
- Regina Y Liu. On a notion of data depth based on random simplices. *The Annals of Statistics*, pages 405–414, 1990.
- Robert J McCann. Existence and uniqueness of monotone measure-preserving maps. *Duke Mathematical Journal*, 80(2):309–323, 1995.
- Kim Miller, Suneeta Ramaswami, Peter Rousseeuw, J Antoni Sellares, Diane Souvaine, Ileana Streinu, and Anja Struyf. Efficient computation of location depth contours by methods of computational geometry. *Statistics and Computing*, 13(2):153–162, 2003.
- John W Tukey. Mathematics and the picturing of data. In *Proceedings of the International Congress of Mathematicians, Vancouver, 1975*, volume 2, pages 523–531, 1975.
- Cédric Villani. *Topics in optimal transportation*, volume 58. American Mathematical Soc., 2021.
- Yijun Zuo and Robert Serfling. General notions of statistical depth function. *Annals of statistics*, pages 461–482, 2000.