Distribution-free Measures of Association Based on Optimal Transport

Chenyu Kong

Kuangyaming Honors School
Nanjing University

February 14, 2025

Overview

1. Outline

- Reproducing kernel Hilbert space
- Oracle Measure
- Empirical Approximation
  - Graphical Functional
  - Optimal Transport
- Conclusion

2. More Comments

- Intuition
- Future Work

Reproducing kernel Hilbert space

The authors mainly use RKHS to embed probability measures into a Hilbert space, in which case the problem is more convenient to deal with:

$$\langle f, m_K(\mu_Y)\rangle_{\mathcal{H}_K} = \int_{\mathcal{Y}} f(y)\, d\mu_Y(y), \quad f \in \mathcal{H}_K$$

in other words,

$$m_K(\mu_Y) = \int K(\cdot, y)\, d\mu_Y(y) = \mathbb{E}_{\mu_Y}[K(\cdot, Y)]$$

for

$$\mu_Y \in \mathcal{M}_K^1(\mathcal{Y}) := \left\{ \nu \in \mathcal{P}(\mathcal{Y}) : \int_{\mathcal{Y}} K(y, y)\, d\nu(y) < \infty \right\}$$

Oracle Measure

The authors use $\eta_K(\mu)$ defined below as the oracle association measure:

$$\eta_K(\mu) := 1 - \frac{\mathbb{E}\left\|K(\cdot, Y') - K\left(\cdot, \widetilde{Y'}\right)\right\|_{\mathcal{H}_K}^2}{\mathbb{E}\left\|K(\cdot, Y_1) - K(\cdot, Y_2)\right\|_{\mathcal{H}_K}^2} = \frac{\mathbb{E}K\left(Y', \widetilde{Y'}\right) - \mathbb{E}K(Y_1, Y_2)}{\mathbb{E}K(Y_1, Y_1) - \mathbb{E}K(Y_1, Y_2)}$$

where Y_1, Y_2 are independent, Y', $\widetilde{Y'}$ are conditionally independent w.r.t. X', and $Y'|X, \widetilde{Y'}|X' \sim \mu_{Y|X'}$. The oracle measure can be rewritten more intuitively as follows:

$$\eta_K(\mu) = \frac{\mathbb{E}_{\mu_X}\left[\mathrm{MMD}_K^2\left(\mu_{Y|X}, \mu_Y\right)\right]}{\mathbb{E}\|K(\cdot, Y) - \mathbb{E}K(\cdot, Y)\|_{\mathcal{H}_K}^2}$$

where the maximum mean discrepancy is the distance of embedded measures in RKHS.

Empirical Approximation

Next it suffices to find a empirical measure to approximate the oracle in probability. We can use

$$\frac{1}{n}\sum_{i=1}^{n} K(Y_i, Y_i) - \frac{1}{n(n-1)}\sum_{i\neq j} K(Y_i, Y_j)$$

to approximate the denominator, while the numerator is trickier to estimate since we don't observe Y' or $\widetilde{Y}'$. However, by the generating process, given X_i, Y_i can be used as the analogue for Y' and some certain Y_j for $\widetilde{Y}'$ when X_j is close to X_i so as not to lose much information contained in X_i.

Empirical Approximation

The authors use graphical functional to generalized this idea. For each X_i, find the closest k X_j s and use the average of corresponding Y_j s as the analogue of $\widetilde{Y'}$. Applying the graphic language gives the empirical measure

$$\widehat{\eta}_n := \frac{\frac{1}{n}\sum_{i=1}^{n} d_i^{-1} \sum_{j:(i,j)\in\mathcal{E}(\mathcal{G}_n)} K(Y_i, Y_j) - \frac{1}{n(n-1)}\sum_{i\neq j} K(Y_i, Y_j)}{\frac{1}{n}\sum_{i=1}^{n} K(Y_i, Y_i) - \frac{1}{n(n-1)}\sum_{i\neq j} K(Y_i, Y_j)}$$

Empirical Approximation

Next the authors molify the empirical measure using optimal transport so as to make the empirical measure possess distribution-free property. To be more precise, the authors transport the initial two probability measures to uniform distribution in $[0,1]^d$ with least information loss. To achieve this, choose $\mathcal{H}_n^d := \left\{h_1^d, \ldots, h_n^d\right\}$ in $[0,1]^d$ s.t.

$$n^{-1} \sum_{i=1}^{n} \delta_{h_i^d} \rightarrow \nu$$

Then rearranging X_i s.t. $(X_1, \cdots, X_n)$ is closest to $(h_1^d, \cdots, h_n^d)$, i.e.

$$\widehat{R}_n\left(X_i\right) = h_{\widehat{\sigma}_n(i)}^d, \quad i = 1, \ldots, n$$

where

$$\widehat{\sigma}_n := \underset{\sigma \in S_n}{\operatorname{argmin}} \sum_{i=1}^{n} \left\| X_i - h_{\sigma(i)}^d \right\|^2$$

Empirical Approximation

At last the authors get a multivariate rank based empirical measure

$$\widehat{\eta}_n^{\mathrm{RANK}} := \frac{n^{-1} \sum_{i} d_i^{-1} \sum_{j:(i,j) \in \mathcal{E}\left(\mathcal{G}_n^{\mathsf{RANK}}\right)} K\left(\widehat{R}_n^Y\left(Y_i\right), \widehat{R}_n^Y\left(Y_j\right)\right) - F_n}{n^{-1} \sum_{i=1}^{n} K\left(\widehat{R}_n^Y\left(Y_i\right), \widehat{R}_n^Y\left(Y_i\right)\right) - F_n}$$

where

$$F_n := (n(n-1))^{-1} \sum_{i \neq j} K\left(\widehat{R}_n^Y\left(Y_i\right), \widehat{R}_n^Y\left(Y_j\right)\right)$$

Conclusion

Finally, the authors show that the multivariate rank based empirical measure *indeed* converges to the oracle measure in probability under certain conditions.

Theorem (Main Theorem)

- $$r(x_1, x_2) := \mathbb{E}\left[K\left(R^Y(Y_1), R^Y(Y_2)\right) \mid R^X(X_1) = x_1, R^X(X_2) = x_2\right]$$

  is β*-Hölder continuous w.r.t.* $x_1, x_2 \in [0,1]^{d_1}$
- $\mathcal{G}_n^{RANK}$ *satisfies*

  $$\limsup_{n\to\infty} \frac{\max_{i=1}^n d_i}{\min_{i=1}^n d_i} < \infty$$

  $$\frac{1}{n \min_{i=1}^n d_i} \sum_{e\in\mathcal{E}\left(\mathcal{G}_n^{\mathrm{RANK}}\right)} |e|^\beta \to 0$$
- $\mathcal{H}_n^{d_1}$ *and* $\mathcal{H}_n^{d_2}$ *converge weakly to* $\mathcal{U}[0,1]^{d_1}$ *and* $\mathcal{U}[0,1]^{d_2}$ *respectively*

(continued on next page)

Conclusion

Theorem (Continued Main Theorem)

Then

$$\widehat{\eta}_n^{\mathit{RANK}} \xrightarrow{\mathbb{P}} 1 - \frac{\mathbb{E}\left\| K\left(\cdot, R^Y\left(Y'\right)\right) - K\left(\cdot, R^Y\left(\widetilde{Y'}\right)\right)\right\|_{\mathcal{H}_K}^2}{\mathbb{E}\left\| K\left(\cdot, R^Y\left(Y_1\right)\right) - K\left(\cdot, R^Y\left(Y_2\right)\right)\right\|_{\mathcal{H}_K}^2} := \eta_K^{\mathrm{RANK}}$$

Conclusion

In addition, the authors show a central limit theorem for this multivariate rank based empirical measure when $\mu = \mu_X \otimes \mu_Y$, i.e. the initial two probability distributions are independent.

Theorem (CLT)

Consider

$$\mathcal{J}_\theta := \left\{ \widetilde{\mathcal{G}} : \limsup_{n\to\infty} \left(\max_{1\le i\le n} \frac{\widetilde{d}_i}{(\log n)^\gamma} + \frac{\max_{i=1}^n \widetilde{d}_i}{\min_{i=1}^n \widetilde{d}_i} \right) \le D, \mathrm{Var}\left(N_n^{\mathrm{RANK}} \right) \ge \epsilon, \forall n \ge D \right\}$$

Then

$$\lim_{n\to\infty} \sup_{\widetilde{\mathcal{G}}\in\mathcal{J}_\theta} \sup_{z\in\mathbb{R}} \left| \mathbb{P}\left(\frac{N_n^{\mathrm{RANK}}}{\widetilde{S}_n} \le z \right) - \Phi(z) \right| = 0$$

Intuition of $\widehat{\eta}_n$

The authors have explained some of intuitions of construction of $\widehat{\eta}_n$ in their previous work [Deb et al., 2020], consider

$$S_n = \sum_{i=1}^{n} d_i^{-1} \sum_{j \in \mathcal{N}_i} \| Y_i - Y_j \|_2$$

When X and Y are independent, $\mathbb{E} S_n \geq \frac{n}{k} \mathbb{E} \| Y_1 - Y_2 \|_2$; if $X = g(Y)$, by Lusin's Theorem, we might assume g is continuous, then

$$\mathbb{E} \left[\sum_{j \in \mathcal{N}_i} d_i^{-1} \| Y_i - Y_j \|_2 \right] = o(1)$$

and $\mathbb{E} S_n = o(n)$.

Intuition of $\widehat{\eta}_n$

From the explanations before it's natural to guess that the term S_n is connected to the relation between distributions of X and Y.

Next the authors introduce RKHS for an extension such that the data samples need not come from Euclidean spaces, only topological spaces.

Moreover, the authors extend NNG to general graph functional and as said in the theorem, as long as the graph satisfies several conditions, $\widehat{\eta}_n$ is a proper empirical statistic for η_K.

Finally, by transferring initial distributions to $\mathcal{U}[0,1]^d$, $\widehat{\eta}_n$ has distribution-free property when $\mu = \mu_X \otimes \mu_Y$.

Future Work

1. Monotonicity of $\widehat{\eta}_n$. It's natural to guess that the relation between X and Y grows monotonically as $\widehat{\eta}_n$ vary from 0 to 1. It is similar with the monotonicity in [Auddy et al., 2024].
2. CLT when $\mu \neq \mu_X \otimes \mu_Y$. Here some debiasing technique will be needed.
3. Local power analysis. Here we consider the independence test problem. [Bhattacharya, 2020] and [Lin and Han, 2021] could be helpful.
4. Other reference distributions. In this paper, the authors use $\mathcal{U}[0,1]^d$ as the reference distribution. Actually, the proof applies with any reference distributions with bounded support. There are also work applying uniform distribution on hypersphere as the reference distribution with center-outward rank, see [Shi et al., 2025]. For distributions with unbounded support, other techniques such as truncation may help.
5. Numerical performance of this statistics combining graph-based technique and multivariate ranks.

References

Auddy, A., Deb, N., and Nandy, S. (2024).
Exact detection thresholds and minimax optimality of Chatterjee's correlation coefficient.
Bernoulli, 30(2):1640 – 1668.

Bhattacharya, B. B. (2020).
Asymptotic distribution and detection thresholds for two-sample tests based on geometric graphs.
The Annals of Statistics, 48(5):2879 – 2903.

Deb, N., Ghosal, P., and Sen, B. (2020).
Measuring association on topological spaces using kernels and geometric graphs.

Lin, Z. and Han, F. (2021).
On boosting the power of chatterjee's rank correlation.
Biometrika.

Shi, H., Drton, M., Hallin, M., and Han, F. (2025).
Distribution-free tests of multivariate independence based on center-outward quadrant, Spearman, Kendall, and van der Waerden statistics.
Bernoulli, 31(1):106 – 129.

The End