Distribution-free tests of multivariate independence based on center-outward quadrant, Spearman, Kendall, and van der Waerden statistics

Chenyu Kong

Kuangyaming Honors School
Nanjing University

February 17, 2025

Overview

Center-outward Distributions

The authors conduct multivariate independence test by introducing analogues of classical statistics such as Spearman, Kendall leveraging optimal transport, which has attracted increasing attention recently. The core idea is similar to the one in [Deb et al., 2024]. However, in this article, the authors transport the distribution to uniform distribution on unit ball in $\mathbb{R}^d$ instead of $\mathcal{U}[0,1]^d$. Though the unit ball possesses smooth boundary geometrically while the unit cube has vertices, it's an interested question which one as the reference distribution is better or there is another reference distribution outperforms both.

Center-outward Distribution

The authors use a main theorem in [McCann, 1995], which is fundamental in optimal transport:

Theorem (McCann's Theorem)

Let μ, ν be Borel probability measures and suppose μ vanishes on Borel subsets with Hausdorff dimension $d - 1$. Then there exists a convex function $\psi : \mathbb{R}^d \to \mathbb{R}$ with gradient $\nabla\psi$ pushing μ forward to ν and $\nabla\psi$ is μ-unique. Here $\nabla\psi$ pushing μ forward to ν means

$$\int_{-\infty}^{x} d\mu = \int_{-\infty}^{\nabla\psi(x)} d\nu$$

Center-outward Distribution

Applying McCann's Theorem, for any $P \in \mathcal{P}^d$ Lebesgue absolutely continuous distributions on $\mathbb{R}^d$, there exists a P-unique $\nabla\phi$ s.t. $\nabla\phi(Z) \sim U_d$ if and only if $Z \sim P$, where U_d is the spherical uniform distribution. Denote this P-unique $\nabla\phi$ by $F_\pm$ called center-outward distribution, and it suffices to find an empirical analogue.

Given n sample data $Z^{(n)} = \left(Z_1^{(n)}, \cdots, Z_n^{(n)} \right)$ with each component being a sample point in $\mathbb{R}^d$. Choose a grid $\mathfrak{G}_n$ in the unit ball in $\mathbb{R}^d$ such that the uniform discrete distribution on $\mathfrak{G}_n$ converges weakly to U_d, and let

$$F_\pm^{(n)} \in \operatorname*{argmin}_{T:\left(Z_1^{(n)}, \cdots, Z_n^{(n)} \right) \to \mathfrak{G}_n} \sum_{i=1}^{n} \left\| T\left(Z_i^{(n)} \right) - Z_i^{(n)} \right\|^2$$

Grid Selection

The authors provide a way to choose $\mathfrak{G}_n$ since U_d is the product of a uniform over the distances to the origin and a uniform over the unit sphere. Let $n = n_R n_S + n_0$ with $n_0 < \min(n_R, n_S)$ and $\min(n_R, n_S) \to \infty$. Here n_0 denotes the number of origins in $\mathfrak{G}_n$, and n_S denotes the number of regular points $\{s_i^{n_S} : i = 1, \cdots, n_S\}$ on sphere S^{d-1}. Therefore

$$\mathfrak{G}_n = \left\{ \left(\frac{r}{n_R + 1} \right) s_s^{n_S} : r = 1, \cdots, n_R; s = 1, \cdots, n_S \right\} \bigcup \{0_i : i = 1, \cdots, n_0\}$$

Moreover, given that U_d is a symmetric distribution, we can choose $\mathfrak{G}_n$ such that $\mathbf{u} \in \mathfrak{G}_n$ implies $-\mathbf{u} \in \mathfrak{G}_n$, i.e. $\mathfrak{G}_n$ is a symmetric grid.

Center-outward rank and sign

Denote the center-outward rank of $Z_i^{(n)}$ by

$$R_{i;\pm}^{(n)} = (n_R + 1) \left\| \mathbf{F}_\pm^{(n)} \left(\mathbf{Z}_i^{(n)} \right) \right\|$$

Rescaled center-outward rank of $Z_i^{(n)}$:

$$\widetilde{R}_{i;\pm}^{(n)} := \left\| \mathbf{F}_\pm^{(n)} \left(\mathbf{Z}_i^{(n)} \right) \right\|, \quad i = 1, \dots, n$$

Center-outward sign of $Z_i^{(n)}$:

$$\mathbf{S}_{i;\pm}^{(n)} := \frac{\mathbf{F}_\pm^{(n)} \left(\mathbf{Z}_i^{(n)} \right)}{\left\| \mathbf{F}_\pm^{(n)} \left(\mathbf{Z}_i^{(n)} \right) \right\|} \quad \mathbf{F}_\pm^{(n)} \left(\mathbf{Z}_i^{(n)} \right) \neq \mathbf{0}$$

Center-outward rank and sign

If $\mathbf{Z}^{(n)}$ consists of i.i.d. samples from P, then $\left(\mathbf{F}_{\pm}^{(n)}\left(\mathbf{Z}_1^{(n)}\right), \ldots, \mathbf{F}_{\pm}^{(n)}\left(\mathbf{Z}_n^{(n)}\right)\right)$ is uniformly distributed over the $n!$ permutations of $\mathfrak{G}_n$, and pointwise convergence holds:

$$\left\|\mathbf{F}_{\pm}^{(n)}\left(\mathbf{Z}_i^{(n)}\right) - \mathbf{F}_{\pm}\left(\mathbf{Z}_i^{(n)}\right)\right\| \to 0 \quad a.s.$$

Moreover, if P satisfies further conditions, it can be showed that the convergence holds uniformly:

$$\max_{1 \leq i \leq n} \left\|\mathbf{F}_{\pm}^{(n)}\left(\mathbf{Z}_i^{(n)}\right) - \mathbf{F}_{\pm}\left(\mathbf{Z}_i^{(n)}\right)\right\| \longrightarrow 0 \quad a.s.$$

Independence Test

In terms of independence test, we consider the hypothesis test problem:

$$H_0 : X_1 \text{ and } X_2 \text{ are independent} \qquad H_1 : X_1 \text{ and } X_2 \text{ are not independent}$$

As a result, given the i.i.d. samples $(\mathbf{X}'_{11}, \mathbf{X}'_{21})', (\mathbf{X}'_{12}, \mathbf{X}'_{22})', \ldots, (\mathbf{X}'_{1n}, \mathbf{X}'_{2n})'$ from $\mathbb{R}^{d_1+d_2}$, the authors introduce matrices for independence test computing from center-outward ranks and signs, i.e. the center-outward version of sign, Spearman, Kendall statistics.

Statistics

Let $\text{sign}[\mathbf{M}]$ stand for the matrix collecting signs of each entry of $\mathbf{M}$, and let $J_k : [0,1) \to \mathbb{R}$ denote any continuous and square-integrable score function, introduce

$$\underset{\sim}{\mathbf{W}}_{\text{sign}}^{(n)} := \frac{1}{n} \sum_{i=1}^{n} \mathbf{S}_{1i;\pm}^{(n)} \mathbf{S}_{2i;\pm}^{(n)\prime}$$

$$\underset{\sim}{\mathbf{W}}_{\text{Spearman}}^{(n)} := \frac{1}{n} \sum_{i=1}^{n} \widetilde{R}_{1i;\pm}^{(n)} \widetilde{R}_{2i;\pm}^{(n)} \mathbf{S}_{1i;\pm}^{(n)} \mathbf{S}_{2i;\pm}^{(n)\prime}$$

$$\underset{\sim}{\mathbf{W}}_{\text{Kendall}}^{(n)} := \binom{n}{2}^{-1} \sum_{i<i'} \text{sign} \left[\left(\widetilde{R}_{1i;\pm}^{(n)} \mathbf{S}_{1i;\pm}^{(n)} - \widetilde{R}_{1i';\pm}^{(n)} \mathbf{S}_{1i';\pm}^{(n)} \right) \left(\widetilde{R}_{2i;\pm}^{(n)} \mathbf{S}_{2i;\pm}^{(n)} - \widetilde{R}_{2i';\pm}^{(n)} \mathbf{S}_{2i';\pm}^{(n)} \right)' \right]$$

$$\underset{\sim}{\mathbf{W}}_{J}^{(n)} := \frac{1}{n} \sum_{i=1}^{n} J_1 \left(\widetilde{R}_{1i;\pm}^{(n)} \right) J_2 \left(\widetilde{R}_{2i;\pm}^{(n)} \right) \mathbf{S}_{1i;\pm}^{(n)} \mathbf{S}_{2i;\pm}^{(n)\prime}$$

Statistics

Then the authors point out the asymptotic representations of these matrices and show a version of central limit theorem for them. The asymptotic representations are:

$$\mathbf{W}_{\mathsf{sign}}^{(n)} := \frac{1}{n} \sum_{i=1}^{n} \mathbf{S}_{1i;\pm} \mathbf{S}_{2i;\pm}'$$

$$\mathbf{W}_{\mathsf{Spearman}}^{(n)} := \frac{1}{n} \sum_{i=1}^{n} \mathbf{F}_{1;\pm} (\mathbf{X}_{1i}) \mathbf{F}_{2;\pm}' (\mathbf{X}_{2i})$$

$$\mathbf{W}_{\mathsf{Kendall}}^{(n)} := \binom{n}{2}^{-1} \sum_{i<i'} \mathrm{sign} \left[(\mathbf{F}_{1;\pm} (\mathbf{X}_{1i}) - \mathbf{F}_{1;\pm} (\mathbf{X}_{1i'})) (\mathbf{F}_{2;\pm} (\mathbf{X}_{2i}) - \mathbf{F}_{2;\pm} (\mathbf{X}_{2i'}))' \right]$$

$$\mathbf{W}_{J}^{(n)} := \frac{1}{n} \sum_{i=1}^{n} J_1 \left(\|\mathbf{F}_{1;\pm} (\mathbf{X}_{1i})\| \right) J_2 \left(\|\mathbf{F}_{2;\pm} (\mathbf{X}_{2i})\| \right) \mathbf{S}_{1i;\pm} \mathbf{S}_{2i;\pm}'$$

Statistics

After introducing these matrices, the test statistics are

$$T_{\sim \text{sign}}^{(n)} := nd_1 d_2 \left\| \mathbf{W}_{\sim \text{sign}}^{(n)} \right\|_{\text{F}}^2$$

$$T_{\sim \text{Spearman}}^{(n)} := 9nd_1 d_2 \left\| \mathbf{W}_{\sim \text{Spearman}}^{(n)} \right\|_{\text{F}}^2$$

$$T_{\sim \text{Kendall}}^{(n)} := \frac{9n}{4} \left\| \mathbf{W}_{\sim \text{Kendall}}^{(n)} \right\|_{\text{F}}^2$$

and

$$T_{\sim J}^{(n)} := \frac{nd_1 d_2}{\sigma_{J_1}^2 \sigma_{J_2}^2} \left\| \mathbf{W}_{\sim J}^{(n)} \right\|_{\text{F}}^2$$

Statistics

If taking $J_k^{\mathrm{vdW}}(u) := \left(F_{\chi^2_{d_k}}^{-1}(u) \right)^{1/2}$, we get the van der Waerden test statistic.

These tests are strictly distribution-free, and exact critical values can be computed or simulated as well. They are the extensions of traditional univariate tests.

Generalized Konijn Families

Let $\mathbf{X}^* = (\mathbf{X}_1^{*\prime}, \mathbf{X}_2^{*\prime})'$ with independent component from $\mathbb{R}^{d_1}$ and $\mathbb{R}^{d_2}$ respectively. Consider

$$\mathbf{X} = \begin{pmatrix} \mathbf{X}_1 \\ \mathbf{X}_2 \end{pmatrix} := \mathbf{M}_\delta \begin{pmatrix} \mathbf{X}_1^* \\ \mathbf{X}_2^* \end{pmatrix} := \begin{pmatrix} (1-\delta)\mathbf{I}_{d_1} & \delta\mathbf{M}_1 \\ \delta\mathbf{M}_2 & (1-\delta)\mathbf{I}_{d_2} \end{pmatrix} \begin{pmatrix} \mathbf{X}_1^* \\ \mathbf{X}_2^* \end{pmatrix}$$

The distribution $\mathrm{P}_\delta^{\mathbf{X}}(\mathrm{P}_1, \mathrm{P}_2; \mathbf{M}_1, \mathbf{M}_2)$ of $\mathbf{X}$ belongs to

$$\mathcal{P}_{\mathrm{P}_1, \mathrm{P}_2; \mathbf{M}_1, \mathbf{M}_2}^{\mathbf{X}} := \left\{ \mathrm{P}_\delta^{\mathbf{X}}(\mathrm{P}_1, \mathrm{P}_2; \mathbf{M}_1, \mathbf{M}_2) \mid \delta \in \mathbb{R} \right\}$$

called generalized Konijn family.

Introduce the score $\boldsymbol{\varphi} := (\boldsymbol{\varphi}_1', \boldsymbol{\varphi}_2')' := (\varphi_{1;1}, \ldots, \varphi_{1;d_1}, \varphi_{2;1}, \ldots, \varphi_{2;d_2})'$ with

$$\varphi_{k;\ell} := -2 \frac{D_\ell \left[(f_k)^{1/2} \right]}{(f_k)^{1/2}}$$

Generalized Konijn Families

Let $P_\delta^{(n)}$ denote the distribution of $\mathbf{X}^{(n)} := \left(\mathbf{X}_1^{(n)}, \ldots, \mathbf{X}_n^{(n)}\right)$ with each component being i.i.d. sampled from $\mathbf{X} = (\mathbf{X}_1', \mathbf{X}_2')' \sim P_\delta^{\mathbf{X}} \left(P_1, P_2; \mathbf{M}_1, \mathbf{M}_2\right) \in \mathcal{P}_{P_1, P_2; \mathbf{M}_1, \mathbf{M}_2}^{\mathbf{X}}$. Then $\left\{ P_\delta^{(n)} \mid \delta \in \mathbb{R} \right\}$ is LAN(local asymptotic normal) at $\delta = 0$. In other words, when $\delta = 0$,

$$\Lambda^{(n)} \left(\mathbf{X}^{(n)}\right) := \log \frac{dP_{n^{-1/2}\tau}^{(n)} \left(\mathbf{X}^{(n)}\right)}{dP_0^{(n)}} = \tau \Delta^{(n)} \left(\mathbf{X}^{(n)}\right) - \frac{1}{2}\tau^2\gamma^2 + o_P(1)$$

and

$$\begin{aligned}
\Delta^{(n)} \left(\mathbf{X}^{(n)}\right) := & n^{-1/2} \sum_{i=1}^n \left[\mathbf{X}_{1i}^{(n)\prime} \mathbf{M}_2' \boldsymbol{\varphi}_2 \left(\mathbf{X}_{2i}^{(n)}\right) + \mathbf{X}_{2i}^{(n)\prime} \mathbf{M}_1' \boldsymbol{\varphi}_1 \left(\mathbf{X}_{1i}^{(n)}\right) \right. \\
& \left. - \left(\mathbf{X}_{1i}^{(n)\prime} \boldsymbol{\varphi}_1 \left(\mathbf{X}_{1i}^{(n)}\right) - d_1\right) - \left(\mathbf{X}_{2i}^{(n)\prime} \boldsymbol{\varphi}_2 \left(\mathbf{X}_{2i}^{(n)}\right) - d_2\right) \right]
\end{aligned}$$

are asymptotic normal with mean 0 and variance γ^2.

Generalized Konijn Families

In particular, if

$$f_k \left(\mathbf{x}_k \right) \propto \left(\det \left(\mathbf{\Sigma}_k \right) \right)^{-1/2} \phi_k \left(\sqrt{\mathbf{x}_k' \mathbf{\Sigma}_k^{-1} \mathbf{x}_k} \right), \quad k = 1, 2$$

denote the distribution by $\mathrm{P}_\delta^{\mathrm{ell}} \left(\phi_1, \phi_2, \mathbf{\Sigma}_1, \mathbf{\Sigma}_2; \mathbf{M}_1, \mathbf{M}_2 \right)$. Moreover, we call $\mathrm{P}_\delta^{\mathrm{ell}} \left(\phi_1, \phi_2, \mathbf{I}_{d_1}, \mathbf{I}_{d_2}; \mathbf{M}, \mathbf{M}' \right)$ elliptical Konijn alternatives, and corresponding $\mathcal{P}_{\phi_1, \phi_2, \mathbf{I}_{d_1}, \mathbf{I}_{d_2}; \mathbf{M}, \mathbf{M}'}^{\mathrm{ell}}$ elliptical Konijn families.

Limiting Distributions

In this part, the authors establish the Chernoff–Savage and Hodges–Lehmann results for the AREs(asymptotic relative efficiency), relative to Wilks' test, of their center-outward van der Waerden and Spearman tests. and Spearman tests, respectively.

Theorem (Limiting Distributions)

If $\mathbf{X} \sim \mathrm{P}^{\mathbf{X}}_{n^{-1/2}\tau}\left(\mathrm{P}_1, \mathrm{P}_2; \mathbf{M}_1, \mathbf{M}_2\right)$,

1. The limiting distribution of $\underset{\sim}{T}_J^{(n)}$ is noncentral chi-square with $d_1 d_2$ degrees of freedom and noncentrality parameter

$$C_J(\tau) \left\| \mathrm{E}_{H_0}\left[\mathbf{J}_1\left(\mathbf{F}_{1;\pm}\left(\mathbf{X}_1\right)\right) \mathbf{R} \mathbf{J}_2\left(\mathbf{F}_{2;\pm}\left(\mathbf{X}_2\right)\right)' \right] \right\|^2_{\mathrm{F}}$$

2. The limiting distribution of $\underset{\sim\mathrm{Kendall}}{T^{(n)}}(n)$ is noncentral chi-square with $d_1 d_2$ degrees of freedom and noncentrality parameter

$$9\tau^2 \left\| \mathrm{E}_{H_0}\left[\mathbf{F}^{\square}_{1;\pm}\left(\mathbf{X}_1\right) \mathbf{R} \mathbf{F}^{\square}_{2;\pm}\left(\mathbf{X}_2\right)' \right] \right\|^2_{\mathrm{F}}$$

Limiting Distributions

In the limiting distribution theorem above,
$C_J(\tau) := \frac{\tau^2 d_1 d_2}{\sigma_{J_1}^2 \sigma_{J_2}^2}, \mathbf{R} := \mathbf{X}_1' \mathbf{M}_2' \boldsymbol{\varphi}_2 (\mathbf{X}_2) + \mathbf{X}_2' \mathbf{M}_1' \boldsymbol{\varphi}_1 (\mathbf{X}_1), \mathbf{J}_k(\mathbf{u}) := J_k(\|\mathbf{u}\|) \frac{\mathbf{u}}{\|\mathbf{u}\|} \mathbf{1}_{[\|\mathbf{u}\| \neq 0]}$, and

$$\left(\mathbf{F}_{k;\pm}^{\square} (\mathbf{X}_k)\right)_\ell := 2 F_{k\ell} \left(\left(\mathbf{F}_{k;\pm} (\mathbf{X}_k)\right)_\ell\right) - 1, \quad k = 1, 2, \quad \ell = 1, \dots, d_k$$

where F_{kl} is the marginal cumulative distribution function of $\left(\mathbf{F}_{k;\pm} (\mathbf{X}_k)\right)_\ell$.

Wilk's Test

Moreover, recall that Wilk's test statistic takes the form

$$T_{\text{Wilks}}^{(n)} := n \log V^{(n)}$$

with

$$V^{(n)} := \frac{\det\left(\mathbf{S}_1^{(n)}\right) \det\left(\mathbf{S}_2^{(n)}\right)}{\det\left(\mathbf{S}^{(n)}\right)}$$

where $\mathbf{S}_k^{(n)}$ is the sample covariance matrix for $\mathbf{X}_{k1}, \ldots, \mathbf{X}_{kn}$.

Under alternatives of the form $\mathbf{X} \sim \mathrm{P}_{n^{-1/2}\tau}^{\text{ell}}(\phi_1, \phi_2, \boldsymbol{\Sigma}_1, \boldsymbol{\Sigma}_2; \mathbf{M}_1, \mathbf{M}_2)$, the limiting distribution of Wilk's statistics is noncentral chi-square with $d_1 d_2$ degrees of freedom and noncentrality parameter

$$\tau^2 \left\| \Sigma_1^{1/2} \mathbf{M}_2' \boldsymbol{\Sigma}_2^{-1/2} + \boldsymbol{\Sigma}_1^{-1/2} \mathbf{M}_1 \boldsymbol{\Sigma}_2^{1/2} \right\|_{\text{F}}^2$$

Pitman Efficiency

Theorem (Pitman ARE)

The Pitman ARE in $\mathcal{P}^{\mathrm{ell}}_{\phi_1,\phi_2,\Sigma_1,\Sigma_2;\mathbf{M}_1,\mathbf{M}_2}$ of center-outward test $\psi_J^{(n)}$ relative to Wilk's test $\psi_{\mathrm{Wilks}}^{(n)}$ is

$$\mathrm{ARE}\left(\psi_J^{(n)}, \psi_{\textit{Wilks}}^{(n)}\right) = \frac{\left\| D_1 C_2 \mathbf{\Sigma}_1^{1/2} \mathbf{M}_2' \mathbf{\Sigma}_2^{-1/2} + D_2 C_1 \mathbf{\Sigma}_1^{-1/2} \mathbf{M}_1 \mathbf{\Sigma}_2^{1/2} \right\|_{\mathrm{F}}^2}{d_1 d_2 \sigma_{J_1}^2 \sigma_{J_2}^2 \left\| \mathbf{\Sigma}_1^{1/2} \mathbf{M}_2' \mathbf{\Sigma}_2^{-1/2} + \mathbf{\Sigma}_1^{-1/2} \mathbf{M}_1 \mathbf{\Sigma}_2^{1/2} \right\|_{\mathrm{F}}^2}$$

where

$$C_k = C_k\left(J_k, \phi_k\right) := \mathrm{E}\left[J_k(U) \rho_k\left(\widetilde{F}_k^{-1}(U) \right) \right] \qquad D_k = D_k\left(J_k, \phi_k\right) := \mathrm{E}\left[J_k(U) \widetilde{F}_k^{-1}(U) \right]$$

with $\rho_k := \frac{-\phi_k'}{\phi_k}$, $\widetilde{F}_k$ being the cumulative distribution function of $\left\| \mathbf{\Sigma}_k^{-1/2} \mathbf{X}_k \right\|$ and U being a random variable uniformly distributed in $(0, 1)$.

Pitman Efficiency

In particular, under the family $\mathcal{P}^{\text{ell}}_{\phi_1, \phi_2, \mathbf{I}_{d_1}, \mathbf{I}_{d_2}; \mathbf{M}, \mathbf{M}'}$, we have

1. $\mathrm{ARE}\left(\psi^{(n)}_{\text{vdW}}, \psi^{(n)}_{\text{Wilks}}\right) \geq 1$ with equality under P_1 and P_2 Gaussian.

2. $\mathrm{ARE}\left(\psi^{(n)}_{\text{Spearman}}, \psi^{(n)}_{\text{Wilks}}\right) \geq \Omega\left(d_1, d_2\right) \geq 9/16$, where Wilcoxon score function $J^w_k(u) = u$, and

$$\Omega\left(d_1, d_2\right) := 9\left(2c^2_{d_1} + d_1 - 1\right)^2 \left(2c^2_{d_2} + d_2 - 1\right)^2 / 1024 d_1 d_2 c^2_{d_1} c^2_{d_2}$$

with

$$c_d := \inf\left\{x > 0 \mid \left(\sqrt{x} B_{\sqrt{2d-1}/2}(x)\right)' = 0\right\},$$

$$B_a(x) := \sum_{m=0}^{\infty} \frac{(-1)^m}{m!\Gamma(m+a+1)}\left(\frac{x}{2}\right)^{2m+a}$$

Comments

This article mainly conducts tests of independence using spherical uniform distribution as the reference distribution and construct extended statistics based on center-outward ranks and signs. In the local asymptotic power analysis part, the authors mainly focus on the ARE of their statistics against classical Wilk's test by calculating the limit distributions and noncentrality parameter. Introducing the generalized Konijn families is actually introducing local alternatives.

References

Deb, N., Ghosal, P., and Sen, B. (2024).
Distribution-free measures of association based on optimal transport.

McCann, R. J. (1995).
Existence and uniqueness of monotone measure-preserving maps.
Duke Mathematical Journal, 80(2):309 – 323.

The End