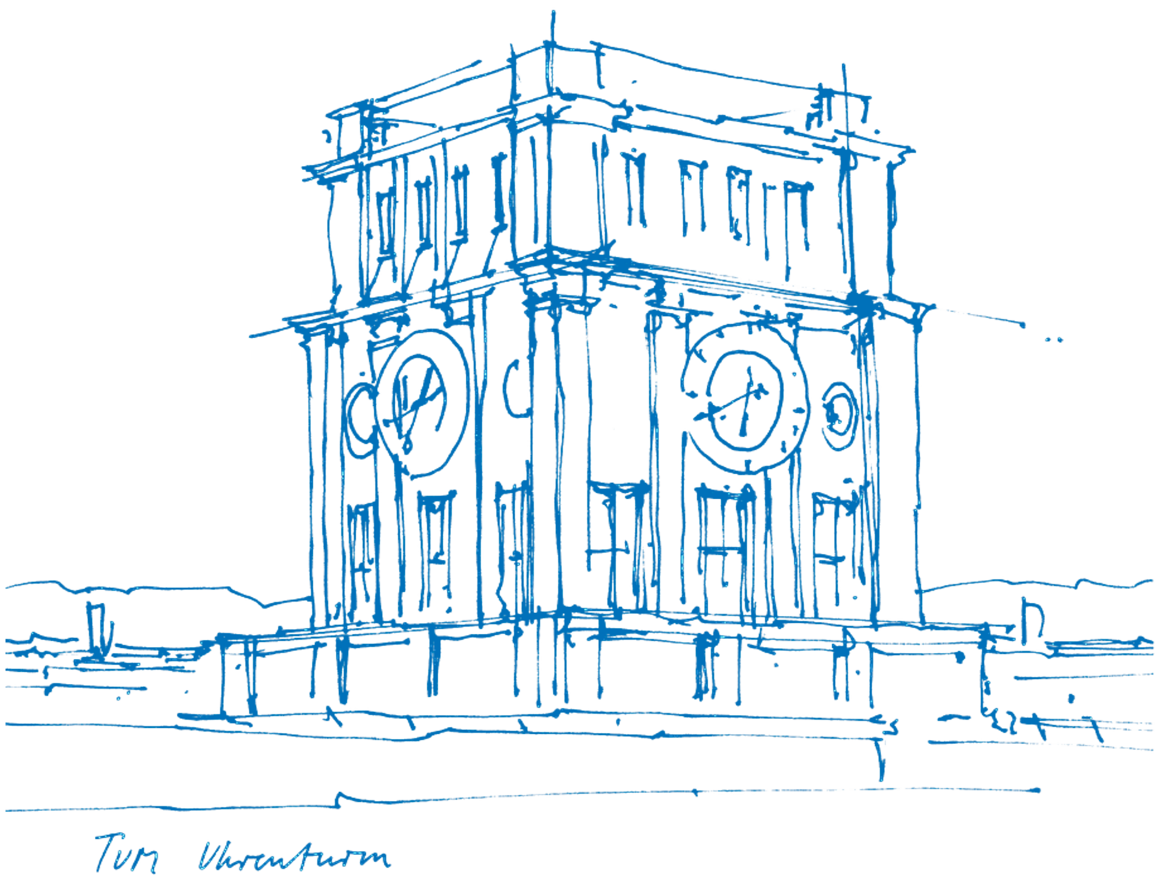
On universally consistent and fully distribution-free rank tests of vector independence

Hongjian Shi

Technical University of Munich

ETH-UCPH-TUM Workshop

12 October 2022



Collaborators

Marc Hallin

Mathias Drton

Fang Han

Problem

We consider testing independence of two random vectors with absolutely continuous distributions based on finite observations.

$$\boldsymbol{X} = \underbrace{(X_1, X_2, \ldots, X_p)^\top}_{p \text{ covariates}} \qquad \boldsymbol{Y} = \underbrace{(Y_1, Y_2, \ldots, Y_q)^\top}_{q \text{ covariates}}$$

$$H_0 : \boldsymbol{X} \text{ is independent of } \boldsymbol{Y}$$

Data:

$$(\boldsymbol{X}_1, \boldsymbol{Y}_1)$$
$$(\boldsymbol{X}_2, \boldsymbol{Y}_2)$$
$$\vdots$$
$$(\boldsymbol{X}_n, \boldsymbol{Y}_n)$$

n independent copies of $(\boldsymbol{X}, \boldsymbol{Y})$

Paradigm

Criteria:

- The test should be distribution-free, and directly implementable without the need of permutation.
- The test should be consistent in a certain sense.

a long-standing problem

- The test should be optimal under certain standard.

rate-optimality

- ~~The dimension p should be allowed to be much larger than the sample size n.~~

future work

Bivariate case

- **Data:** $\{(X_i \in \mathbb{R}, Y_i \in \mathbb{R}), i \in [n]\}$ i.i.d.
- **Aim:** testing if "$H_0 : X \perp\!\!\!\perp Y$" is true.

The test should be distribution-free

The ranks of $X_1, \ldots, X_n$ are uniformly distributed on the set of all permutations of $[n]$.

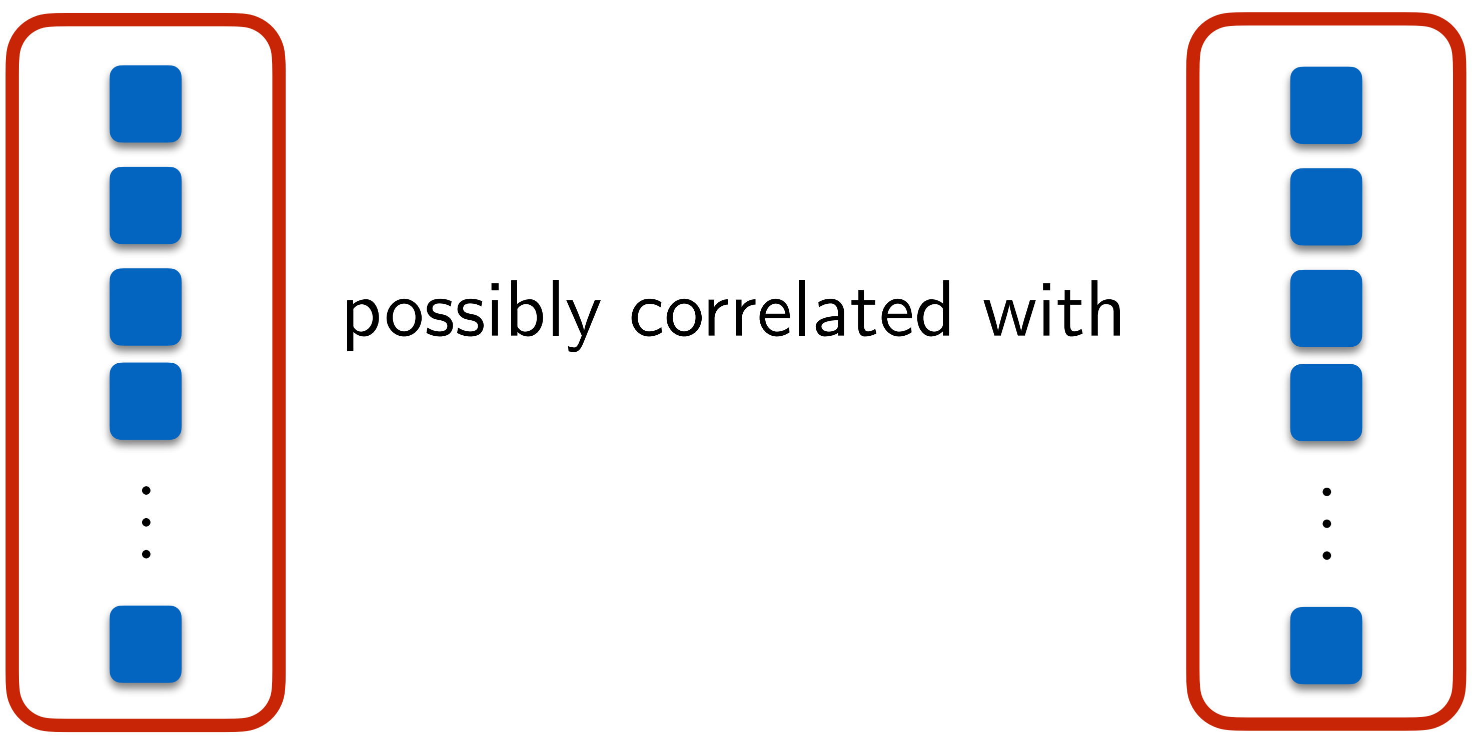
Under H_0, the marginal ranks of $\{X_i, i \in [n]\}$ and $\{Y_i, i \in [n]\}$ are independent.

For any test statistic based on ranks, its null distribution is both determined and independent of $P_{X,Y}$, i.e., the test is distribution-free.

(Marginal) rank tests?

- **Data:** $\{(\boldsymbol{X}_i \in \mathbb{R}^p, \boldsymbol{Y}_i \in \mathbb{R}^q), i \in [n]\}$ i.i.d.
- **Aim:** testing if "$H_0 : \boldsymbol{X} \perp\!\!\!\perp \boldsymbol{Y}$" is true.

Tests built on marginal ranks are no longer distribution-free:



ranks of $\{X_{1,j}, \ldots, X_{n,j}\}$ ranks of $\{X_{1,k}, \ldots, X_{n,k}\}$

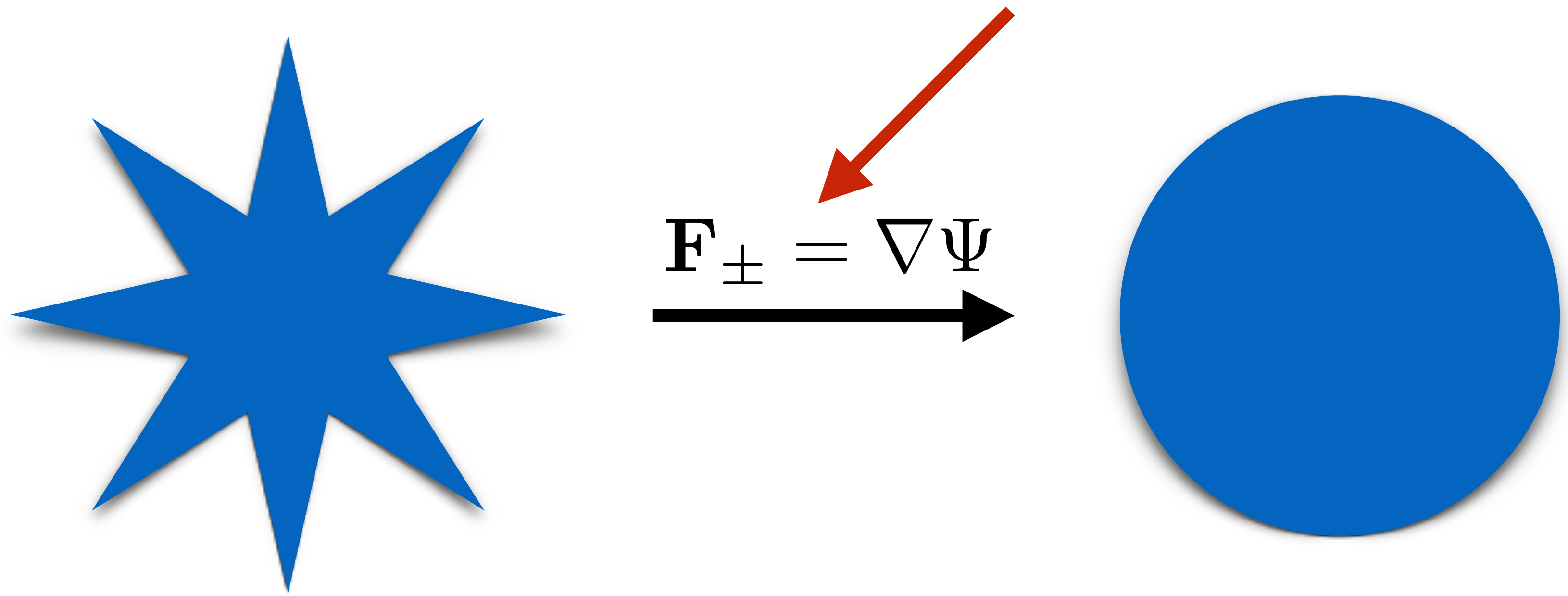
Outline

- **Center-outward multivariate rank**
- **The proposed test**
- **Discussion**

Solution

- **Center-outward CDF in general dimension**

Center-outward population distribution function



P_d

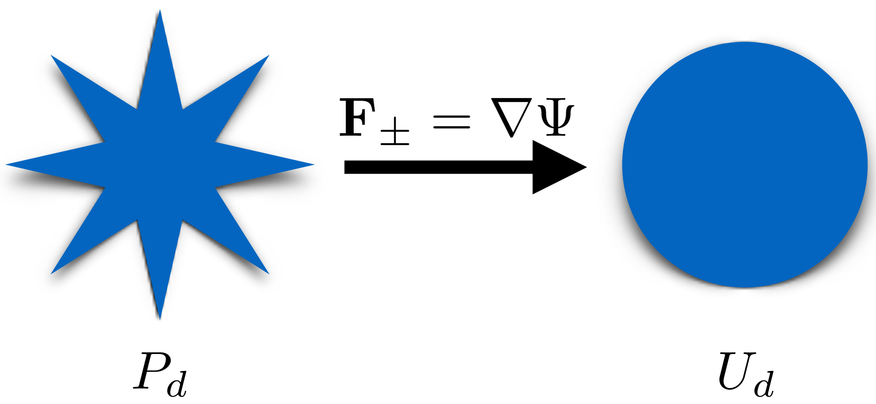
a general absolutely continuous probability measure in dimenion *d*

U_d

spherical uniform measure over *d*-dimensional unit ball

Solution

- **Center-outward CDF in general dimension**



$$\inf_{T} \int_{\mathbb{R}^d} \left\| T(\boldsymbol{x}) - \boldsymbol{x} \right\|_2^2 dP_d$$

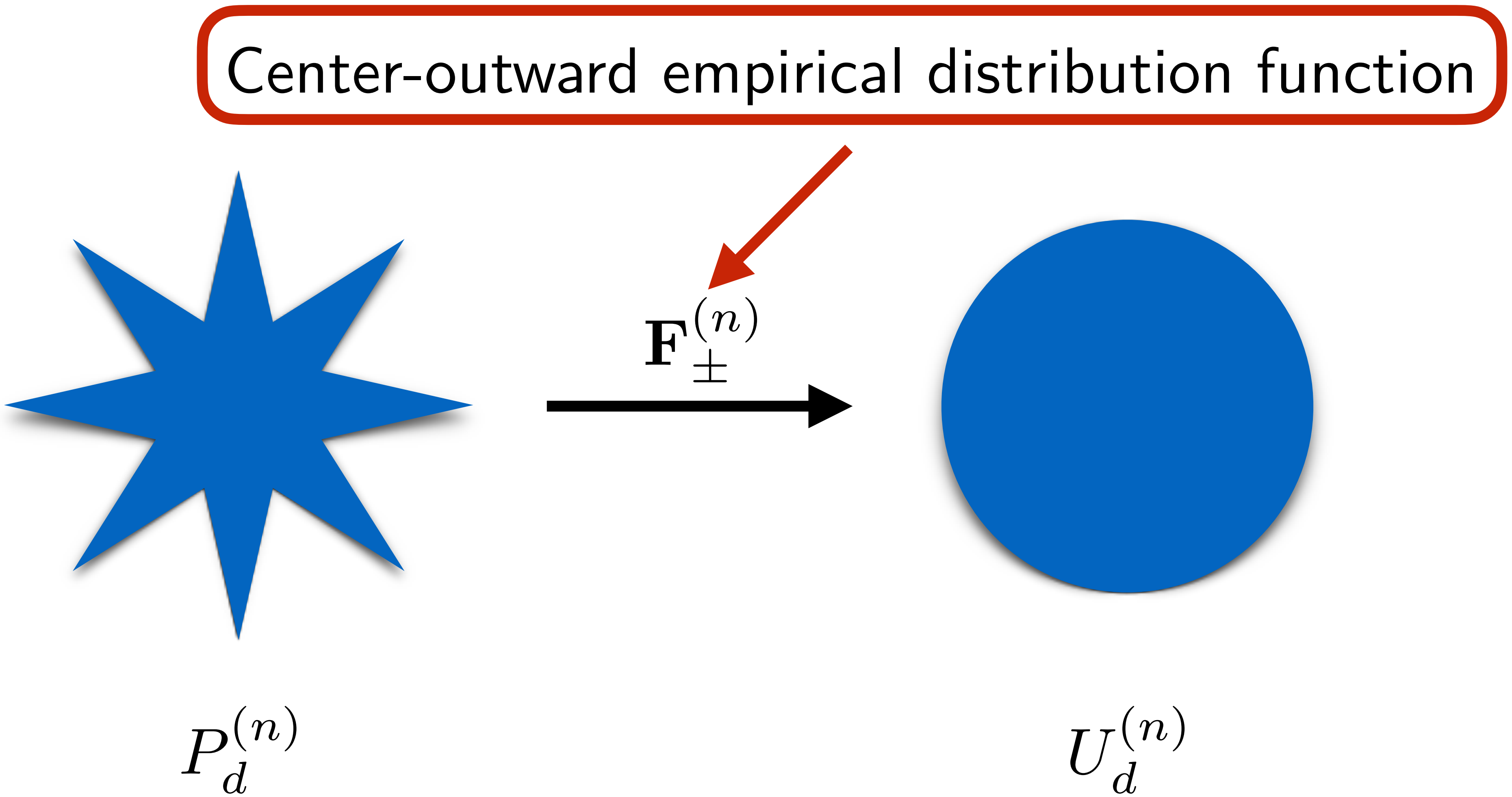
subject to $T_\sharp P_d = U_d$

optimal transport problem

- Existence and uniqueness: Main Theorem in McCann (1995)

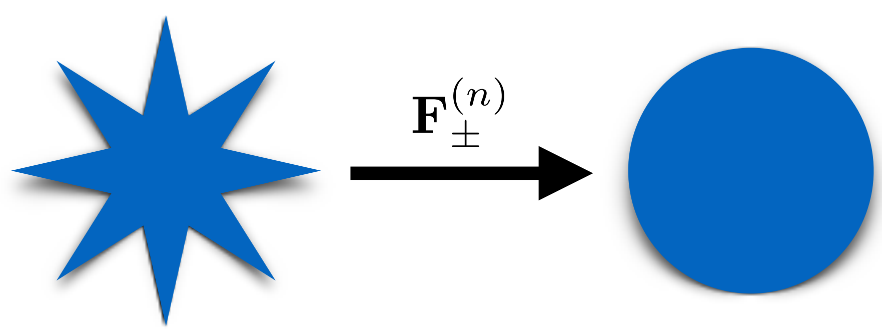
Solution

- **Center-outward Empirical CDF in general dimension**



Solution

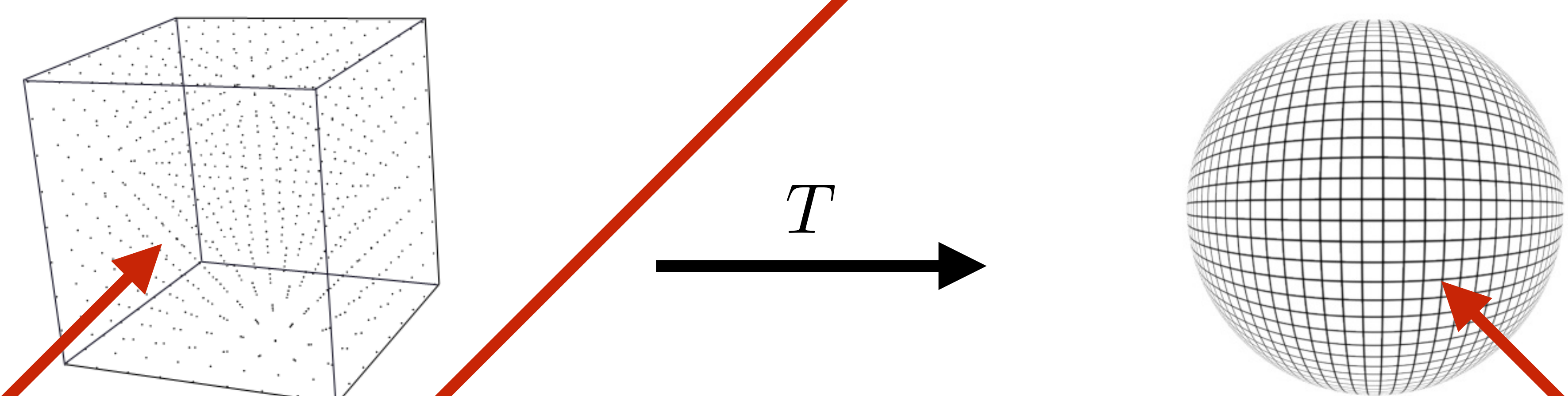
- **Center-outward Empirical CDF in general dimension**



$P_d^{(n)} \xrightarrow{\mathbf{F}_\pm^{(n)}} U_d^{(n)}$

$$\mathbf{F}_\pm^{(n)} := \operatorname*{argmin}_{T \in \mathcal{T}} \sum_{i=1}^{n} \left\| \boldsymbol{X}_i - T(\boldsymbol{X}_i) \right\|_2^2$$

assignment problem



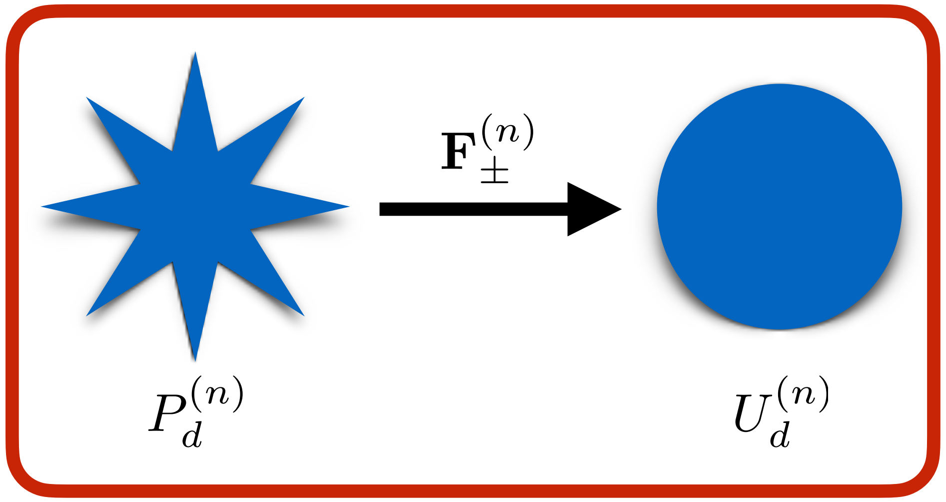
$\{\boldsymbol{X}_i\}_{i=1}^n \subseteq \mathbb{R}^d \xrightarrow{T} \mathbb{S}_d^{(n)}$

the collection of all bijective mappings between $\{\boldsymbol{X}_i\}_{i=1}^n$ and $\{\boldsymbol{u}_i\}_{i=1}^n$

consisting of n points $\{\boldsymbol{u}_i\}_{i=1}^n \subseteq \mathbb{R}^d$ that approximate the spherical uniform measure over the unit ball

Solution

- **Center-outward Empirical CDF in general dimension**



$$\mathbf{F}_{\pm}^{(n)} := \underset{T \in \mathcal{T}}{\operatorname{argmin}} \sum_{i=1}^{n} \left\| \boldsymbol{X}_i - T(\boldsymbol{X}_i) \right\|_2^2$$

assignment problem

Center-outward multi-rank is strongly consistent and distribution-free:

Hallin et al. (2021): Let $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n$ be i.i.d. with absolutely continous distribution P_d. Then

$$\left\| \mathbf{F}_{\pm}^{(n)}(\boldsymbol{X}_i) - \mathbf{F}_{\pm}(\boldsymbol{X}_i) \right\|_2 \xrightarrow{\text{a.s.}} 0 \quad \text{as } n \to \infty,$$

and $(\mathbf{F}_{\pm}^{(n)}(\boldsymbol{X}_1), \ldots, \mathbf{F}_{\pm}^{(n)}(\boldsymbol{X}_n))$ is uniformly distributed over all permutations of $\mathbb{S}_d^{(n)}$.

Outline

- **Center-outward multivariate rank**
- **The proposed test**
- **Discussion**

Distance covariance

- **SRB's insight:**

  The squared distance covariance between two random vectors $\boldsymbol{X} \in \mathbb{R}^p$ and $\boldsymbol{Y} \in \mathbb{R}^q$ with finite first moments, introduced by Székely, Rizzo and Bakirov (2007), is defined as

$$\mathrm{dCov}^2(\boldsymbol{X}, \boldsymbol{Y}) := \frac{1}{c_p c_q} \int_{\mathbb{R}^{p+q}} \frac{\|\phi_{\boldsymbol{X},\boldsymbol{Y}}(\boldsymbol{t}, \boldsymbol{s}) - \phi_{\boldsymbol{X}}(\boldsymbol{t})\phi_{\boldsymbol{Y}}(\boldsymbol{s})\|_2^2}{\|\boldsymbol{t}\|_2^{1+p}\|\boldsymbol{s}\|_2^{1+q}} \mathrm{d}\boldsymbol{t}\mathrm{d}\boldsymbol{s},$$

  where $\phi_{\boldsymbol{X}}$, $\phi_{\boldsymbol{Y}}$ and $\phi_{\boldsymbol{X},\boldsymbol{Y}}$ are the individual and joint characteristic functions of X and Y respectively.

  The sample squared distance covariance is defined as

$$\mathrm{dCov}_n^2\Big((\boldsymbol{X}_i)_{i=1}^n, (\boldsymbol{Y}_i)_{i=1}^n\Big) := \binom{n}{4}^{-1}$$

$$\sum_{i_1 \neq \cdots \neq i_4} \frac{1}{4 \cdot 4!} g(\boldsymbol{X}_{i_1}, \boldsymbol{X}_{i_2}, \boldsymbol{X}_{i_3}, \boldsymbol{X}_{i_4}) g(\boldsymbol{Y}_{i_1}, \boldsymbol{Y}_{i_2}, \boldsymbol{Y}_{i_3}, \boldsymbol{Y}_{i_4}),$$

  where $g(\boldsymbol{z}_1, \boldsymbol{z}_2, \boldsymbol{z}_3, \boldsymbol{z}_4)$

$$:= \|\boldsymbol{z}_1 - \boldsymbol{z}_2\|_2 + \|\boldsymbol{z}_3 - \boldsymbol{z}_4\|_2 - \|\boldsymbol{z}_1 - \boldsymbol{z}_3\|_2 - \|\boldsymbol{z}_2 - \boldsymbol{z}_4\|_2.$$

Test

- **Data:** $\{(\boldsymbol{X}_i \in \mathbb{R}^p, \boldsymbol{Y}_i \in \mathbb{R}^q), i \in [n]\}$ i.i.d. distributed with absolutely continuous probability measures $P_{\boldsymbol{X}}, P_{\boldsymbol{Y}}$.
- **Aim:** testing if "$H_0 : \boldsymbol{X} \perp\!\!\!\perp \boldsymbol{Y}$" is true.

Proposal:

- Calculate center-outward ranks $\mathbf{F}^{(n)}_{\boldsymbol{X},\pm}(\boldsymbol{X}_1), \dots, \mathbf{F}^{(n)}_{\boldsymbol{X},\pm}(\boldsymbol{X}_n)$ and $\mathbf{F}^{(n)}_{\boldsymbol{Y},\pm}(\boldsymbol{Y}_1), \dots, \mathbf{F}^{(n)}_{\boldsymbol{Y},\pm}(\boldsymbol{Y}_n)$;
- Combine center-outward ranks with distance covariance, obtaining the test statistic

$$\widehat{M}_n := n \cdot \mathrm{dCov}^2_n\big((\mathbf{F}^{(n)}_{\boldsymbol{X},\pm}(\boldsymbol{X}_i))_{i=1}^n, (\mathbf{F}^{(n)}_{\boldsymbol{Y},\pm}(\boldsymbol{Y}_i))_{i=1}^n\big);$$

- Reject H_0 if $\widehat{M}_n$ is large enough.

Theory

- **Multivariate Hájek asymptotic representation**
  - Consider the "population" center-outward signed-ranks $\mathbf{F}_{\boldsymbol{X},\pm}(\boldsymbol{X}_i) \sim U_p$ and $\mathbf{F}_{\boldsymbol{Y},\pm}(\boldsymbol{Y}_i) \sim U_q$;
  - Lead to the "oracle" test statistic:

$$\widetilde{M}_n := n \cdot \mathrm{dCov}_n^2((\mathbf{F}_{\boldsymbol{X},\pm}(\boldsymbol{X}_i))_{i=1}^n, (\mathbf{F}_{\boldsymbol{Y},\pm}(\boldsymbol{Y}_i))_{i=1}^n);$$

  - Standard exercise (e.g. Jakobsen (2017, Theorem 5.10)) gives, under H_0,

$$\widetilde{M}_n \stackrel{\mathrm{d}}{\to} \sum_{k=1}^{\infty} \lambda_k(\xi_k^2 - 1),$$

where $\lambda_1, \lambda_2, \ldots$ are positive constants depending only on p, q.

Theory

- **Multivariate Hájek asymptotic representation**

Main Theorem (SHDH 2022). Let $(\boldsymbol{X}_1, \boldsymbol{Y}_1), \ldots, (\boldsymbol{X}_n, \boldsymbol{Y}_n)$ be independent copies of $(\boldsymbol{X}, \boldsymbol{Y})$ with absolutely continuous probability measures $P_{\boldsymbol{X}}, P_{\boldsymbol{Y}}$ and $\boldsymbol{X}$ and $\boldsymbol{Y}$ are independent. Then it holds that

$$\widehat{M}_n - \widetilde{M}_n = o_{\mathrm{P}}(1),$$

where

$$\widehat{M}_n := n \cdot \mathrm{dCov}_n^2((\mathbf{F}_{\boldsymbol{X},\pm}^{(n)}(\boldsymbol{X}_i))_{i=1}^n, (\mathbf{F}_{\boldsymbol{Y},\pm}^{(n)}(\boldsymbol{Y}_i))_{i=1}^n)$$
$$\widetilde{M}_n := n \cdot \mathrm{dCov}_n^2((\mathbf{F}_{\boldsymbol{X},\pm}(\boldsymbol{X}_i))_{i=1}^n, (\mathbf{F}_{\boldsymbol{Y},\pm}(\boldsymbol{Y}_i))_{i=1}^n).$$

As an immediate corollary,

$$\widehat{M}_n \xrightarrow{\mathsf{d}} \sum_{k=1}^{\infty} \lambda_k(\xi_k^2 - 1).$$

Theory

- **Consistency**

The test based on $\widehat{M}_n$ takes the form

$$\mathsf{T}_\alpha := \mathbb{1}\big(\widehat{M}_n > q_{1-\alpha}\big),$$

where $q_{1-\alpha}$ is the $(1-\alpha)$-quantile of $\sum_{k=1}^{\infty} \lambda_k(\xi_k^2 - 1)$.

Proposition (Uniform validity and consistency). The test T_α is uniformly valid in the sense that

$$\lim_{n\to\infty} \sup_{\mathrm{P}\in\mathcal{P}_p^{\mathrm{ac}}\otimes\mathcal{P}_q^{\mathrm{ac}}} \mathrm{P}(\mathsf{T}_\alpha = 1) = \alpha.$$

Moreover, for fixed alternative such that $\boldsymbol{X} \in \mathcal{P}_p^{\mathrm{ac}}$ and $\boldsymbol{Y} \in \mathcal{P}_q^{\mathrm{ac}}$ are dependent, it holds that

$$\lim_{n\to\infty} \mathrm{P}(\mathsf{T}_\alpha = 1) = 1.$$

Theory

- **Local power**

$$\begin{pmatrix} \boldsymbol{X} \\ \boldsymbol{Y} \end{pmatrix} := \begin{pmatrix} \mathbf{I}_p & \delta\,\mathbf{M} \\ \delta\,\mathbf{M}' & \mathbf{I}_q \end{pmatrix} \begin{pmatrix} \boldsymbol{X}^* \\ \boldsymbol{Y}^* \end{pmatrix} = \mathbf{A}_\delta \begin{pmatrix} \boldsymbol{X}^* \\ \boldsymbol{Y}^* \end{pmatrix}$$

Sequence of local alternatives:

$$H_{1,n}(\delta_0) : \delta = \delta_n, \qquad \text{where } \delta_n := n^{-1/2}\delta_0$$

Proposition (SHDH 2022). If some regularity assumption holds, we have that for any number $\beta > 0$ satisfying $\alpha + \beta < 1$ there exists a constant $c_\beta > 0$ only depending on β such that as long as $|\delta_0| < c_\beta$

$$\inf_{\overline{\mathsf{T}}_\alpha \in \mathcal{T}_\alpha} \mathrm{P}\{\overline{\mathsf{T}}_\alpha = 0 \mid H_{1,n}(\delta_0)\} \geq 1 - \alpha - \beta$$

for all sufficiently large n. Here the infimum is taken over all size-α tests.

Theory

- **Local power**

$$\begin{pmatrix} \boldsymbol{X} \\ \boldsymbol{Y} \end{pmatrix} := \begin{pmatrix} \mathbf{I}_p & \delta\,\mathbf{M} \\ \delta\,\mathbf{M}' & \mathbf{I}_q \end{pmatrix} \begin{pmatrix} \boldsymbol{X}^* \\ \boldsymbol{Y}^* \end{pmatrix} = \mathbf{A}_\delta \begin{pmatrix} \boldsymbol{X}^* \\ \boldsymbol{Y}^* \end{pmatrix}$$

Sequence of local alternatives: This is the detection boundary!

$$H_{1,n}(\delta_0) : \delta = \delta_n, \quad \text{where } \delta_n := n^{-1/2}\delta_0$$

Theorem (SHDH 2022). If some regularity assumption holds, then for any number $\beta > 0$, there exists some sufficiently large constant $C_\beta > 0$ only depending on β such that, as long as $|\delta_0| > C_\beta$,

$$\lim_{n\to\infty} \mathrm{P}\{\mathsf{T}_\alpha = 1 \mid H_{1,n}(\delta_0)\} \geq 1 - \beta.$$

multivariate Hájek asymptotic representation and Le Cam's third lemma!

Paradigm

Goals to reach:

- The test should be distribution-free, and directly implementable without the need of permutation.

Yes, by distribution-freeness of multivariate ranks and Hájek asymptotic representation!

- The test should be consistent in a certain sense.

Yes, by consistency of distance covariance and P-a.s. invertibility of center-outward distribution function!

- The test should be optimal under certain standard.

Yes, by multivariate Hájek asymptotic representation and Le Cam's third lemma!

Outline

- **Center-outward multivariate rank**
- **The proposed test (cont.)**
- **Discussion**

Test

- **Data:** $\{(\boldsymbol{X}_i \in \mathbb{R}^p, \boldsymbol{Y}_i \in \mathbb{R}^q), i \in [n]\}$ i.i.d.
- **Aim:** testing if "$H_0 : \boldsymbol{X} \perp\!\!\!\perp \boldsymbol{Y}$" is true.

Proposal:

- Calculate center-outward ranks $\mathbf{F}^{(n)}_{\boldsymbol{X},\pm}(\boldsymbol{X}_1), \ldots, \mathbf{F}^{(n)}_{\boldsymbol{X},\pm}(\boldsymbol{X}_n)$ and $\mathbf{F}^{(n)}_{\boldsymbol{Y},\pm}(\boldsymbol{Y}_1), \ldots, \mathbf{F}^{(n)}_{\boldsymbol{Y},\pm}(\boldsymbol{Y}_n)$;
- Compute $\widehat{\mathbf{W}}^{(n)}_J := \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \mathbf{J}_{\boldsymbol{X}}\Big(\mathbf{F}^{(n)}_{\boldsymbol{X};\pm}(\boldsymbol{X}_i)\Big)\mathbf{J}_{\boldsymbol{Y}}\Big(\mathbf{F}^{(n)}_{\boldsymbol{Y};\pm}(\boldsymbol{Y}_i)\Big)$, where $\mathbf{J}(\mathbf{u}) := J(\|\mathbf{u}\|)\frac{\mathbf{u}}{\|\mathbf{u}\|}\mathbf{1}_{[\|\mathbf{u}\|\neq 0]}$;

  e.g. $J(u) = 1$ (sign);
  $J(u) = u$ (Wilcoxon);
  $J(u) = \big(F^{-1}_{\chi^2_d}(u)\big)^{1/2}$ (vdW).
- Reject H_0 if $\|\widehat{\mathbf{W}}^{(n)}_J\|^2_{\mathrm{F}}$ is large enough.

Pitman efficiency

- **Pitman efficiency**

$$\begin{pmatrix} \boldsymbol{X} \\ \boldsymbol{Y} \end{pmatrix} := \begin{pmatrix} \mathbf{I}_p & \delta\,\mathbf{M} \\ \delta\,\mathbf{M}' & \mathbf{I}_q \end{pmatrix} \begin{pmatrix} \boldsymbol{X}^* \\ \boldsymbol{Y}^* \end{pmatrix} = \mathbf{A}_\delta \begin{pmatrix} \boldsymbol{X}^* \\ \boldsymbol{Y}^* \end{pmatrix}$$

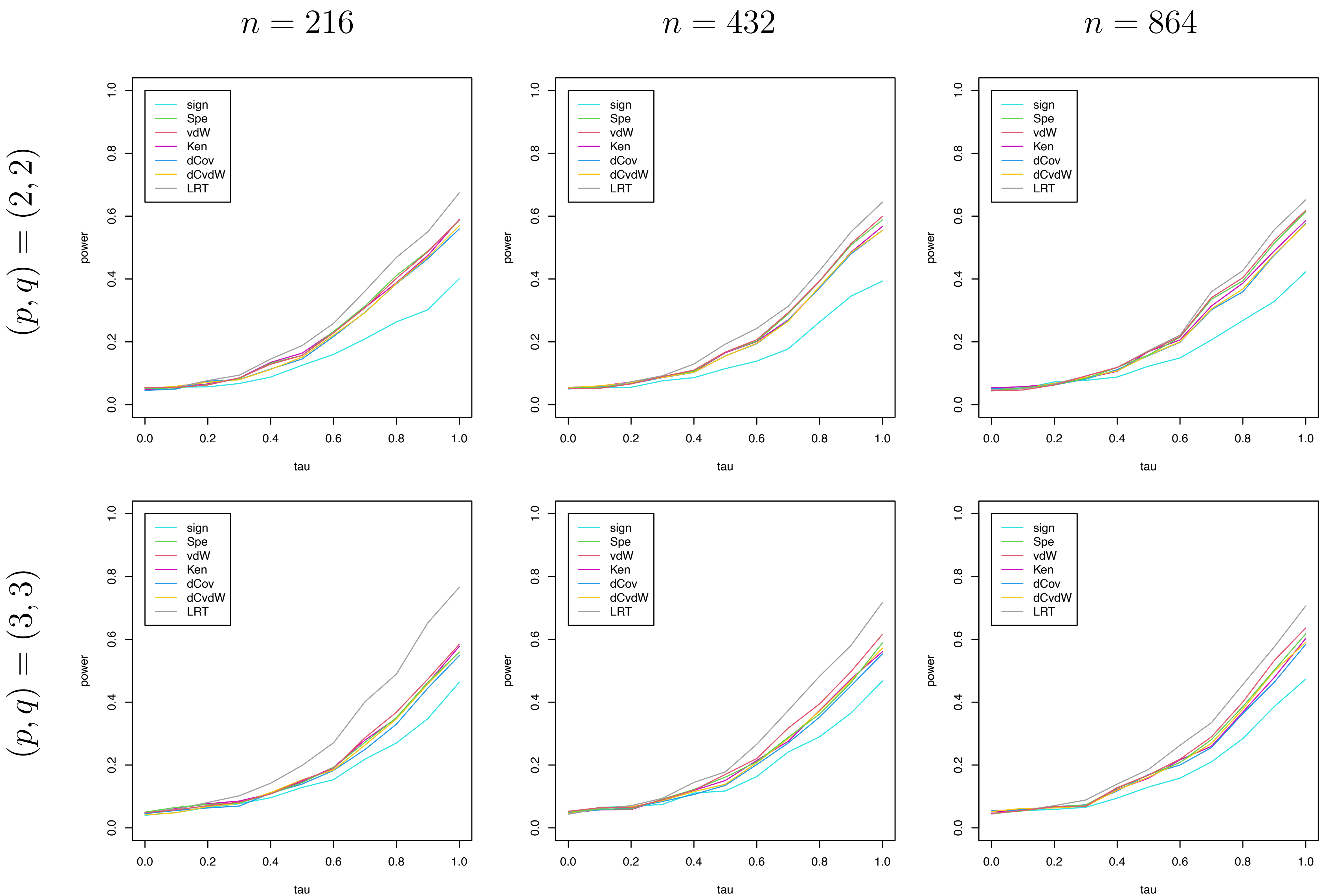
Sequence of local alternatives:

$$H_{1,n}(\delta_0) : \delta = \delta_n, \qquad \text{where } \delta_n := n^{-1/2}\delta_0$$

Theorem (SDHH 2021). If $\boldsymbol{X}^*$ and $\boldsymbol{Y}^*$ are elliptically symmetric distributions satisfying some regularity assumption. Then, the Pitman asymptotic relative efficiency (ARE) of the center-outward test based on the van der Waerden score functions $\left(F^{-1}_{\chi^2_d}(\cdot)\right)^{1/2}$ with respect to Wilks' test is larger than or equal to 1.

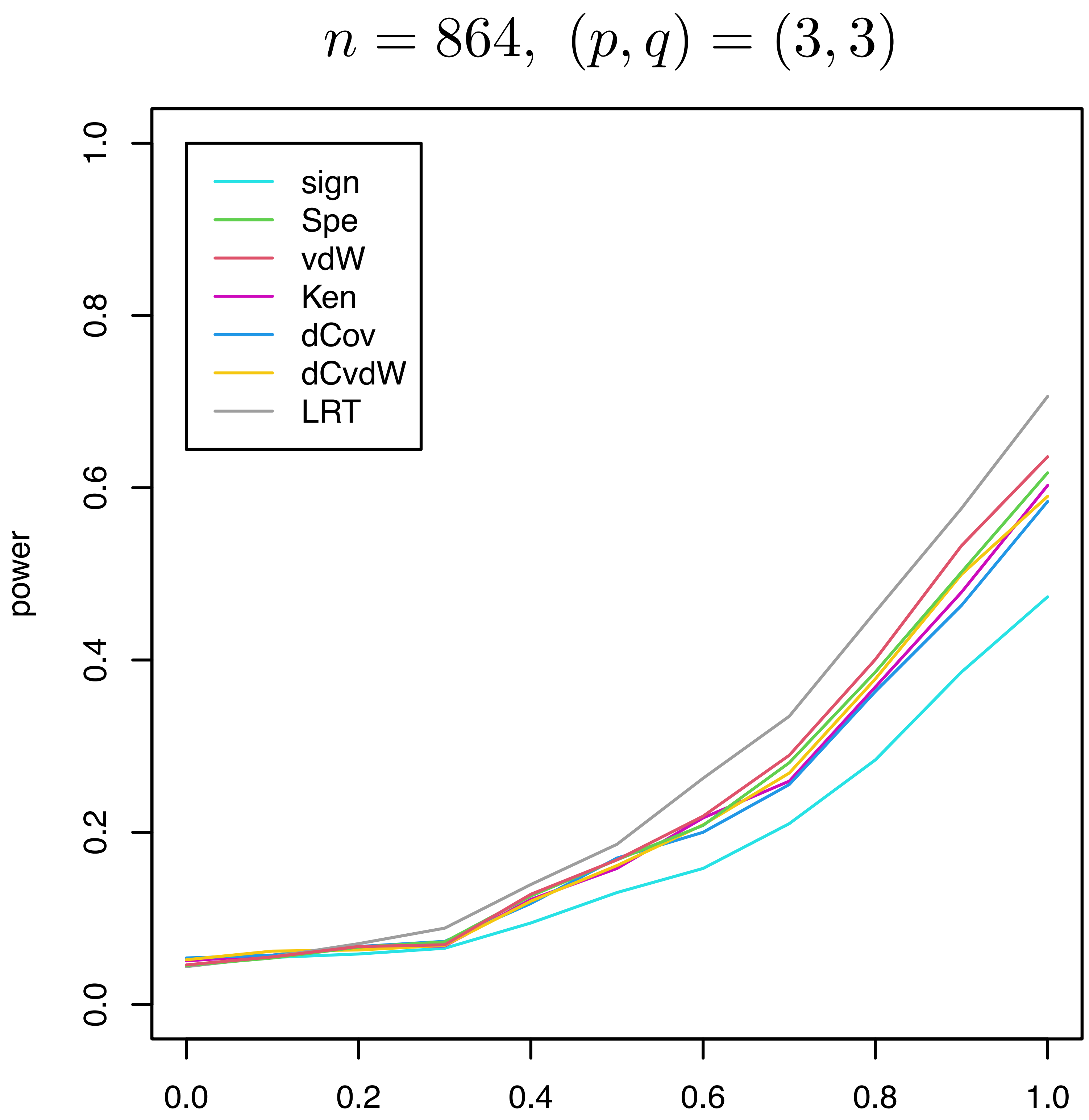
Simulation

- **Gaussian**



Simulation

- **Gaussian**



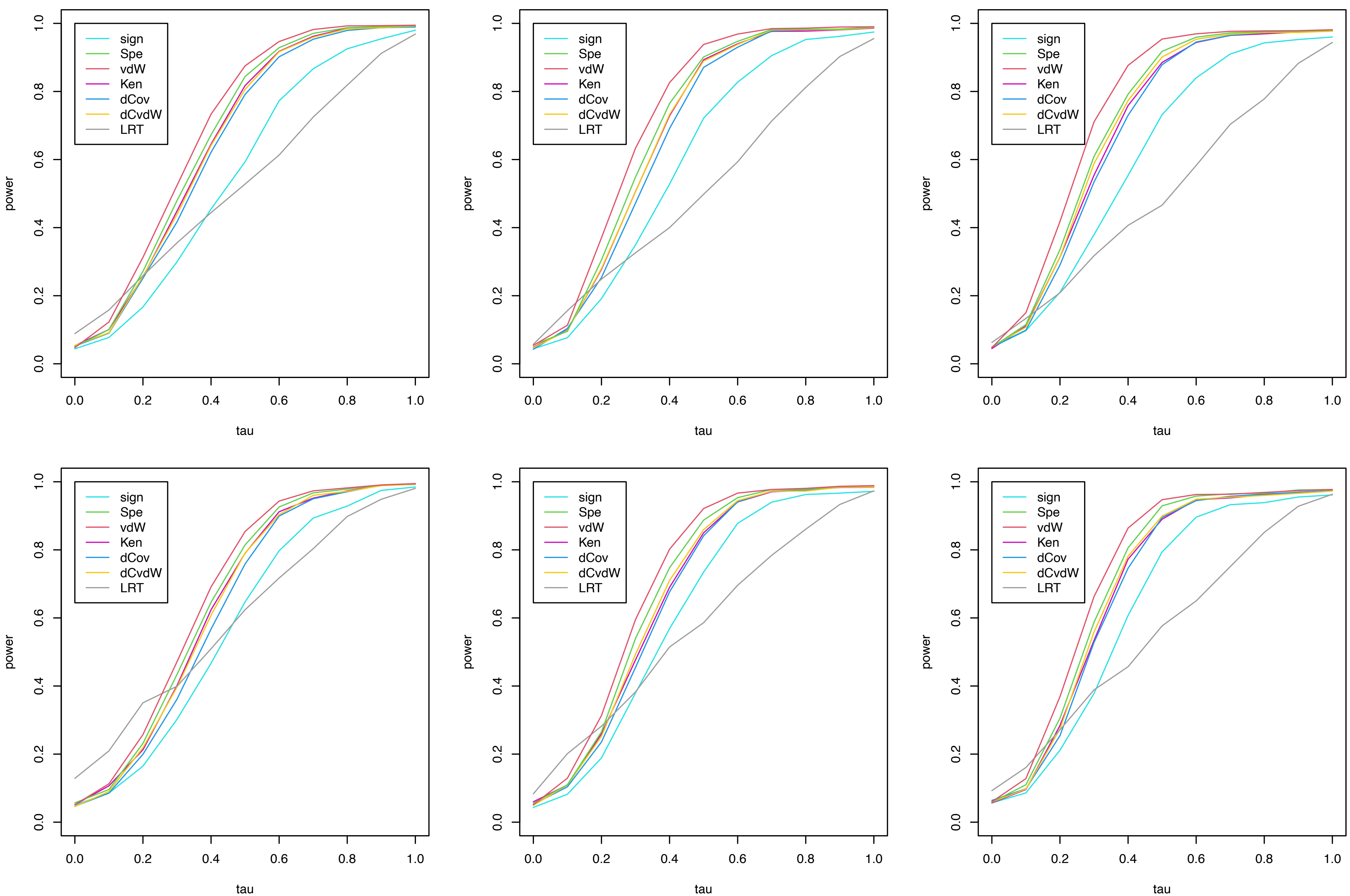
Simulation

- **Cauchy**

$n = 216$ $n = 432$ $n = 864$

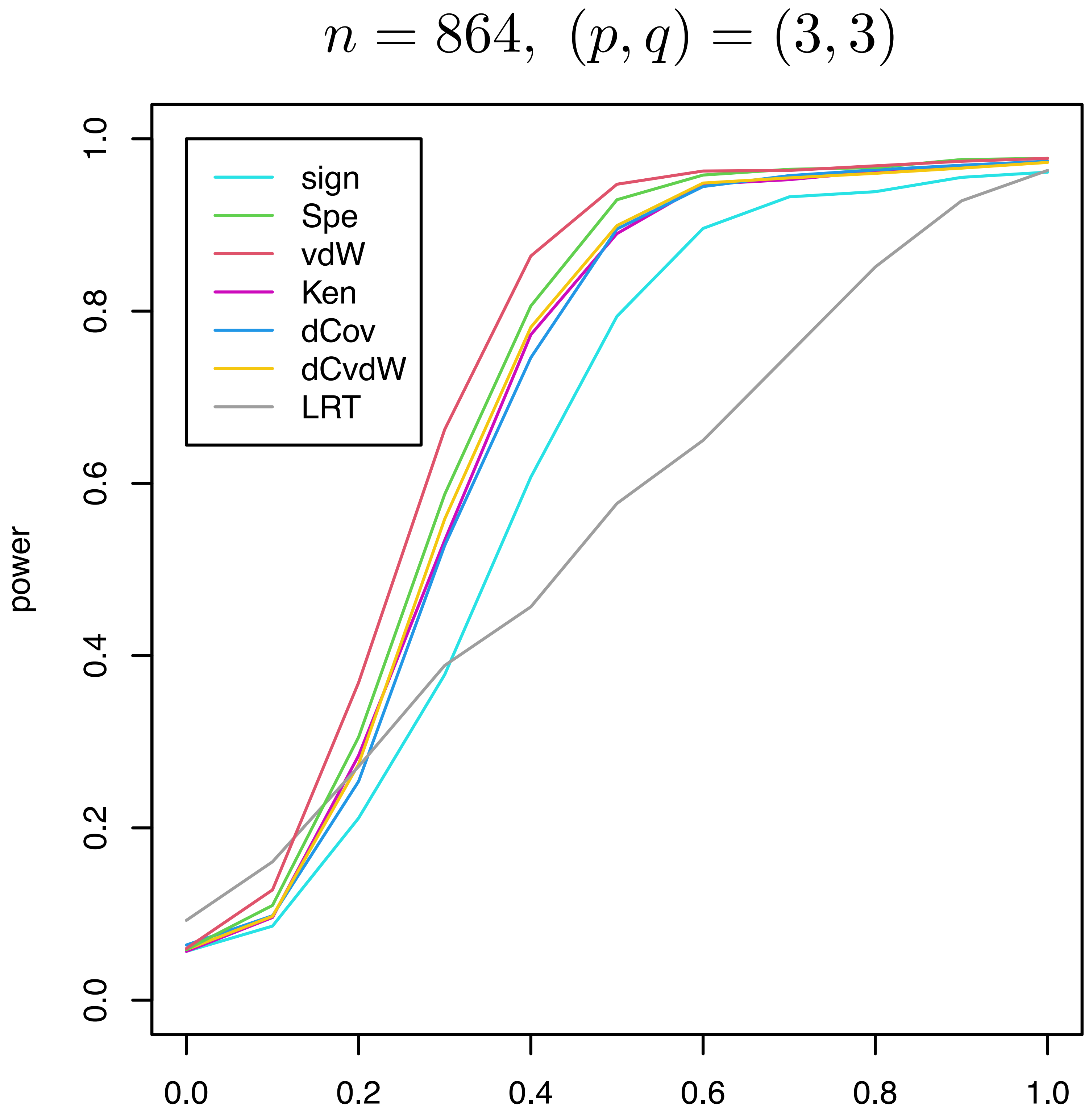
$(p, q) = (2, 2)$

$(p, q) = (3, 3)$



Simulation

- **Cauchy**



Outline

- **Center-outward multivariate rank**
- **The proposed test**
- **Discussion**

Discussion

- Computational complexity of optimal transport-based multivariate ranks.
  General dimension: $\tilde{O}(n^{5/2})$ complexity,
  and $\tilde{O}(n^{3/2})$ complexity if using fast approximation;
  if dimension is 2: $\tilde{O}(n^{3/2+\delta})$ complexity,
  and $\tilde{O}(n^{5/4})$ if using fast approximation.

- Future works: High-dimensional, Conditional independence testing...

- See our papers SDH 2022 (JASA, **117**:395–410), SHDH 2022 (AoS, **50**:1933–1959), and SDHH 2021 (arXiv:2111.15567v1) for more results.

Papers

- Shi, H., Drton, M., and Han, F. (2022). Distribution-free consistent independence tests via center-outward ranks and signs. J. Amer. Statist. Assoc. 117(537):395–410.

- Shi, H., Hallin, M., Drton, M., and Han, F. (2022). On universally consistent and fully distribution-free rank tests of vector independence. Ann. Statist. 50(4):1933–1959.

- Shi, H., Drton, M., Hallin, M., and Han, F. (2021). Center-outward sign-and rank-based quadrant, Spearman, and Kendall tests for multivariate independence. Available at arXiv:2111.15567v1.

Thanks!