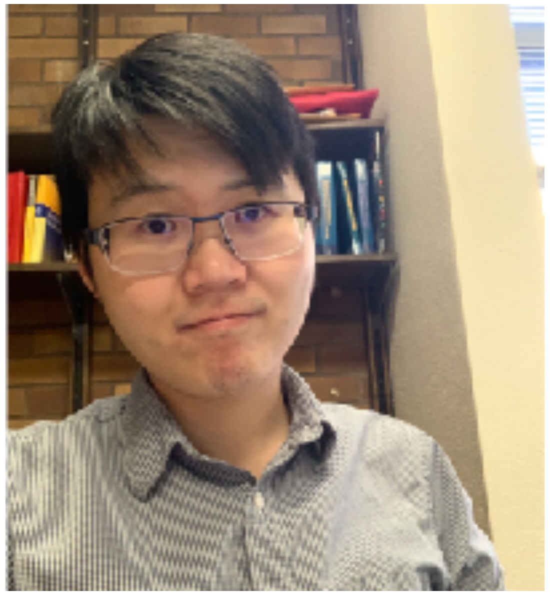
Marginal and multivariate ranks, optimal transport theory, and Le Cam

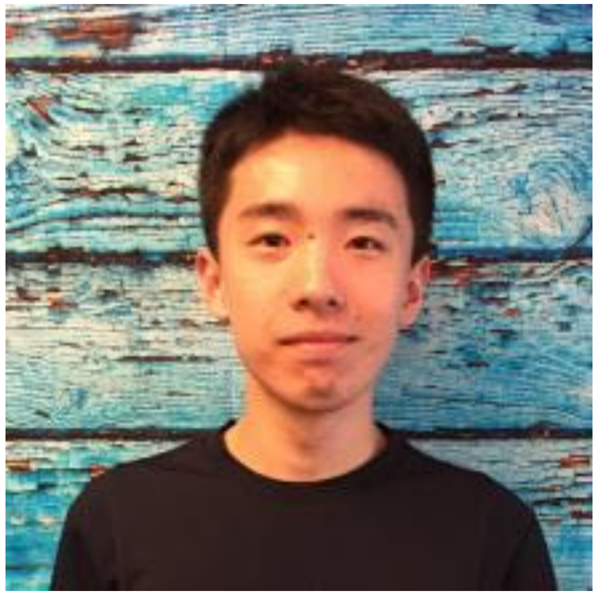
Fang Han

UNIVERSITY of WASHINGTON

Mathias Drton



Fang Han



Hongjian Shi

Two problems

- Problem 1: marginal rank
- Problem 2: multivariate rank

Problem 1

We consider **testing mutual independence** of **many covariates** based on limited information.

$$\boldsymbol{X} = \underbrace{(X_1, X_2, \ldots, X_p)^\top}_{p \text{ covariates}}$$

$$H_0 : X_1, \ldots, X_p \text{ are mutually independent}$$

Data:

$$\boxed{\begin{array}{l} \boldsymbol{X}_1 = (X_{1,1}, X_{1,2}, \ldots, X_{1,p})^\top \\ \boldsymbol{X}_2 = (X_{2,1}, X_{2,2}, \ldots, X_{2,p})^\top \\ \quad \vdots \qquad\quad \vdots \qquad \vdots \qquad\qquad \vdots \\ \boldsymbol{X}_n = (X_{n,1}, X_{n,2}, \ldots, X_{n,p})^\top \end{array}}$$

n independent copies of $\boldsymbol{X}$

Paradigm

Goals to reach:

- the dimension p should be allowed to be much larger than the sample size n;
- the test should be distribution-free, hence directly implementable without the need of permutation;
- the test should be consistent in a certain sense;
- the test should be optimal under certain standard.

Outline

- Bivariate case
- High dimensional case
- Discussion

Bivariate case

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

- the test should be distribution-free, hence directly implementable without the need of permutation;

Tests built on ranks are hence desirable:

The ranks of $X_1, \ldots, X_n$ are uniformly distributed on $\mathcal{S}_n$, the set of all permutations of $[n]$

$\Downarrow$

Under H_0, the marginal ranks of $\{X_i, i \in [n]\}$ and $\{Y_i, i \in [n]\}$ are independent with margins uniformly distributed on $\mathcal{S}_n$.

$\Downarrow$

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

Bivariate case

- the test should be consistent in a certain sense;

with unlimited information, the test shall be able to reject null iff it is not true

Some *Prob101* facts:

- Zero correlation does NOT mean independence;

~~Pearson covariance/correlation~~

- Zero Kendall/Spearman rank correlation does NOT mean independence;

~~Kendall's tau/Spearman's rho~~

Solution

- **Hoeffding's insight: measure of dependence**

(JOC/EFR April 2019)

$$D(X,Y) := \int (\boxed{F} - \boxed{F_1 F_2})^2 dF$$

bivariate CDF of (X,Y)

marginals CDFs of X and Y

Hoeffding (1948): $D(X,Y) \geq 0$. If in addition F is absolutely continuous, then $D(X,Y) = 0$ iff $F = F_1 F_2$.

Solution

- **Hoeffding's insight: estimation**

(JOC/EFR April 2019)

$$\int (F - F_1F_2)^2 dF$$
$$= \int F^2 dF - 2\int FF_1F_2 dF + \int F_1^2F_2^2 dF$$
$$= \mathbb{E}\mathbb{1}(X_1 \le X_3, Y_1 \le Y_3)\mathbb{1}(X_2 \le X_3, Y_2 \le Y_3)$$
$$-$$
$$2\mathbb{E}\mathbb{1}(X_1 \le X_4, Y_1 \le Y_4)\mathbb{1}(X_2 \le X_4)\mathbb{1}(Y_3 \le Y_4)$$
$$+$$
$$\mathbb{E}\mathbb{1}(X_1 \le X_5)\mathbb{1}(X_2 \le X_5)\mathbb{1}(Y_3 \le Y_5)\mathbb{1}(Y_4 \le Y_5)$$

$$\widehat{D}_n = \binom{n}{5}^{-1} \sum_{1 \le i_1 < i_2 < \cdots < i_5 \le n} h_D\left\{\binom{X_{i_1}}{Y_{i_1}}, \ldots, \binom{X_{i_5}}{Y_{i_5}}\right\}$$

a U-statistics

Test based on Hoeffding's *D*

Theorem (Hoeffding, 1948). Under H_0, we have

$$\binom{5}{2}^{-1}(n-1)\widehat{D}_n \xrightarrow{\mathsf{d}} \frac{3}{\pi^4}\sum_{i,j=1}^{\infty}\frac{1}{i^2j^2}(\xi_{ij}^2-1),$$

where $\{\xi_{ij}, i,j=1,2,\ldots\}$ are i.i.d. standard Gaussian random variables.

- The above is a standard non-central limit theorem for *degenerate* U-statistics; the convergence is a mixture of chi-squares instead of simple Gaussian distribution.
- A one-sided and directly implementable test of H_0 can be immediately obtained from the above theorem.

$D(X,Y) \geq 0$

Outline

- Bivariate case
- High dimensional case
- Discussion

High dimensional case

- **Data:** $\{(\boldsymbol{X}_i = (X_{i,1}, \ldots, X_{i,p})^\top \in \mathbb{R}^p, i \in [n]\}$ i.i.d. distributed with continuous marginal CDFs.

- **Aim:** testing if the following null hypothesis is true:

$$H_0 : X_1, \ldots, X_p \text{ are mutually independent.}$$

Proposal:

- Look at each pair of covariates in the data: $\{(X_{i,j}, X_{i,k})^\top, j < k \in [p], i \in [n]\}$;

- Calculate the Hoeffding's D statistic $\widehat{D}_{jk}$ of each pair;

- The proposed test statistic then is $\max\limits_{j<k} \widehat{D}_{jk}$.

Theory

What is the null distribution of $\max_{j<k} \widehat{D}_{jk}$ as possibly $p \gg n$?

Theorem (Hoeffding, 1948). Under H_0, we have

$$\binom{5}{2}^{-1}(n-1)\widehat{D}_n \xrightarrow{\mathsf{d}} \frac{3}{\pi^4}\sum_{i,j=1}^{\infty}\frac{1}{i^2 j^2}(\xi_{ij}^2 - 1),$$

where $\{\xi_{ij}, i,j = 1,2,\ldots\}$ are i.i.d. standard Gaussian random variables.

Proposition (DHS, 2019). Let $Y_1, \ldots, Y_d$ be $d = p(p-1)/2$ independent copies of $Y \overset{d}{=} \sum_{v=1}^{\infty}\lambda_v(\xi_v^2 - 1)$ with $\lambda_v \geq 0$ and $\Lambda := \sum_v \lambda_v < \infty$. Then, as $p \to \infty$,

$$\max_{j\in[d]}\frac{Y_j}{\lambda_1} - 4\log p - (\mu_1 - 2)\log\log p + \frac{\Lambda}{\lambda_1} \xrightarrow{\mathsf{d}} G.$$

Here G follows a Gumbel distribution with distribution function

$$\exp\Big\{-\frac{2^{\mu_1/2-2}\kappa}{\Gamma(\mu_1/2)}\exp\Big(-\frac{y}{2}\Big)\Big\},$$

where μ_1 is the multiplicity of the largest eigenvalue λ_1 in the sequence $\{\lambda_1, \lambda_2, \ldots\}$, $\kappa := \prod_{v=\mu_1+1}^{\infty}(1-\lambda_v/\lambda_1)^{-1/2}$, and $\Gamma(z) := \int_0^{\infty} x^{z-1}e^{-x}dx$ is the gamma function.

Theory

The major technical obstacle:

how fast is each $\widehat{D}_{jk}$ weakly converging to the limit?

2.1.3 Cramér's Moderate Deviation Theorem

The Berry–Esseen inequality gives a bound on the absolute error in approximating the distribution of W_n by the standard normal distribution. The usefulness of the bound may be limited when $\Phi(x)$ is close to 0 or 1. Cramér's theory of moderate deviations provides the relative errors. Petrov (1975, pp. 219–228) gives a comprehensive treatment of the theory and introduces the *Cramér series*, which is a power series whose coefficients can be expressed in terms of the cumulants of the underlying distribution and which is used in part (a) of the following theorem.

Theorem 2.13.

(a) Let $X_1, X_2, \ldots$ be i.i.d. random variables with $E(X_1) = 0$ and $Ee^{t_0|X_1|} < \infty$ for some $t_0 > 0$. Then for $x \ge 0$ and $x = o(n^{1/2})$,

$$\frac{P(W_n \ge x)}{1 - \Phi(x)} = \exp\left\{x^2 \lambda\left(\frac{x}{\sqrt{n}}\right)\right\}\left(1 + O\left(\frac{1+x}{\sqrt{n}}\right)\right), \tag{2.8}$$

where $\lambda(t)$ is the Cramér series.

(b) If $Ee^{t_0\sqrt{|X_1|}} < \infty$ for some $t_0 > 0$, then

$$\frac{P(W_n > x)}{1 - \Phi(x)} \to 1 \quad \text{as } n \to \infty \text{ uniformly in } x \in \left[0, o(n^{1/6})\right). \tag{2.9}$$

(c) The converse of (b) is also true; that is, if (2.9) holds, then $Ee^{t_0\sqrt{|X_1|}} < \infty$ for some $t_0 > 0$.

In parts (a) and (b) of Theorem 2.13, $P(W_n \ge x)/(1 - \Phi(x))$ can clearly be replaced by $P(W_n \le -x)/\Phi(-x)$. Moreover, similar results are also available for standardized sums S_n/B_n of independent but not necessarily identically distributed random variables with bounded moment generating functions in some neighborhood of the origin; see Petrov (1975). In Chap. 7, we establish Cramér-type moderate deviation results for *self-normalized* (rather than standardized) sums of independent random variables under much weaker conditions.

We need a CMDT for degenerate U-statistics!

However, it is long-standing open…

Theory

The major technical obstacle:

how fast is each $\widehat{D}_{jk}$ weakly converging to the limit?

Theorem (DHS, 2019). **CMDT for degenerate U-statistics:** We have, for any sequence of positive scalars $e_n \to 0$,

$$\lim_{n\to\infty} \sup_{x_n \in [-\Lambda, e_n n^{\frac{1}{8}-\epsilon}]} \left| \frac{\Pr\left\{ \binom{5}{2}^{-1}(n-1)\widehat{D}_{jk} > x_n \right\}}{\Pr\left\{ \frac{3}{\pi^4}\sum_{i,j=1}^{\infty} \frac{1}{i^2 j^2}(\xi_{ij}^2 - 1) > x_n \right\}} - 1 \right| = 0,$$

where $\{\xi_{ij}, i, j = 1, 2, \ldots\}$ are i.i.d. standard Gaussian and ϵ is an arbitrarily small universal constant.

Theory

Theorem (DHS, 2019). **CMDT for degenerate U-statistics:** We have, for any sequence of positive scalars $e_n \to 0$,

$$\lim_{n\to\infty} \sup_{x_n \in [-\Lambda, e_n n^{\frac{1}{8}-\epsilon}]} \left| \frac{\Pr\left\{ \binom{5}{2}^{-1} (n-1) \widehat{D}_{jk} > x_n \right\}}{\Pr\left\{ \frac{3}{\pi^4} \sum_{i,j=1}^{\infty} \frac{1}{i^2 j^2} (\xi_{ij}^2 - 1) > x_n \right\}} - 1 \right| = 0,$$

where $\{\xi_{ij}, i, j = 1, 2, \ldots\}$ are i.i.d. standard Gaussian and ϵ is an arbitrarily small universal constant.

$\Downarrow$

Corollary (DHS, 2019). We have, as $p, n \to \infty$ and $\log p = o(n^{1/8-\epsilon})$,

$$\lim_{n,p\to\infty} \Pr\left\{ \frac{\pi^4 (n-1)}{30} \max_{j<k} \widehat{D}_{jk} - 4\log p + \log\log p + \frac{\pi^4}{36} > Q_{D,\alpha} \right\} = \alpha$$

with $Q_{D,\alpha} := \log\{\kappa_D^2/(8\pi)\} - 2\log\log(1-\alpha)^{-1}$ and $\kappa_D := \left\{ 2 \prod_{n=2}^{\infty} \frac{\pi/n}{\sin(\pi/n)} \right\}^{1/2} \approx 2.466.$

Paradigm

Goals to reach:

- the dimension p should be allowed to be much larger than the sample size n;

$$\boxed{\log p = o(n^{1/8-\epsilon})}$$

- the test should be distribution-free, hence directly implementable without the need of permutation;

$$\boxed{\mathsf{T}_{D,\alpha} := \mathbb{1}\left\{\frac{\pi^4(n-1)}{30}\max_{j<k}\widehat{D}_{jk} - 4\log p + \log\log p + \frac{\pi^4}{36} > Q_{D,\alpha}\right\}}$$

- the test should be consistent in a certain sense;

permit consistent assessment of pairwise independence

- the test should be optimal under certain standard.

Paradigm

Goals to reach:

- the test should be optimal under certain standard.

$$\mathcal{V}(C) := \left\{ \mathbf{M} \in \mathbb{R}^{p\times p} \colon \mathbf{M} \succeq 0, \mathrm{diag}(\mathbf{M}) = \mathbf{I}_p, \mathbf{M} = \mathbf{M}^\top, \max_{j\neq k} |M_{jk}| \geq C\sqrt{\frac{\log p}{n}} \right\}$$

sparse alternative class

The above is the critical power range.

Theorem (Theorem 5, HCL (2017)). There exists a universal constant $c_0 > 0$ such that for any number $\beta > 0$ satisfying $\alpha + \beta < 1$, in any asymptotic regime with $p \to \infty$ as $n \to \infty$ but $\log p/n = o(1)$, it holds for all sufficiently large n and p that

$$\inf_{\overline{T}_\alpha \in \mathcal{T}_\alpha} \sup_{\mathbf{\Sigma} \in \mathcal{V}(c_0)} \Pr_{\mathbf{\Sigma}}(\overline{T}_\alpha = 0) \geq 1 - \alpha - \beta.$$

Here the infimum is taken over all size-α tests, and the supremum is taken over all centered Gaussian distributions with (Pearson) covariance matrix $\mathbf{\Sigma}$.

Paradigm

Goals to reach:

- the test should be optimal under certain standard.

$$\mathcal{V}(C) := \left\{ \mathbf{M} \in \mathbb{R}^{p\times p} \colon \mathbf{M} \succeq 0, \operatorname{diag}(\mathbf{M}) = \mathbf{I}_p, \mathbf{M} = \mathbf{M}^\top, \max_{j\neq k} |M_{jk}| \geq C\sqrt{\frac{\log p}{n}} \right\}$$

sparse alternative class

The proposed test is right on the boundary (up to constant)!

Theorem (DHS, 2019). For a sufficiently large universal constant $C_0 > 0$, we have, as long as $n, p \to \infty$,

$$\inf_{\boldsymbol{\Sigma} \in \mathcal{V}(C_0)} \Pr_{\boldsymbol{\Sigma}}(T_{D,\alpha} = 1) = 1 - o(1)$$

where the infimum is over centered Gaussian distributions with (Pearson) covariance matrix $\boldsymbol{\Sigma}$.

Extensions

- **Hoeffding's measure of dependence**

U-statistic of order 5

$$D(X,Y) := \int (F - F_1F_2)^2 dF$$

Hoeffding (1948): $D(X,Y) \geq 0$. If in addition F is absolutely continuous, then $D(X,Y) = 0$ iff $F = F_1F_2$.

- **Blum-Kiefer-Rosenblatt's modification**

$$\boxed{R(X,Y) := \int (F - F_1F_2)^2 d(F_1 \otimes F_2)}$$

BKR (1961): without any distributional assumption, we have $R(X,Y) \geq 0$ and $R(X,Y) = 0$ iff $F = F_1F_2$.

$$\boxed{\widehat{R}_n = \binom{n}{6}^{-1} \sum_{1 \leq i_1 < i_2 < \cdots < i_6 \leq n} h_R\left\{\binom{X_{i_1}}{Y_{i_1}}, \ldots, \binom{X_{i_6}}{Y_{i_6}}\right\}}$$

U-statistic of order 6

Extensions

- **Bergsma and Dassios's proposal**

$$\widehat{\tau}_n^* = \binom{n}{4}^{-1} \sum_{1 \le i_1 < i_2 < \cdots < i_4 \le n} h_{\tau^*}\left\{ \binom{X_{i_1}}{Y_{i_1}}, \ldots, \binom{X_{i_4}}{Y_{i_4}} \right\}$$

$$\begin{aligned} h_{\tau^*}(\boldsymbol{z}_1, \ldots, \boldsymbol{z}_4) := \frac{1}{16} \sum_{(i_1,\ldots,i_4) \in \mathcal{P}_4} & \Big\{ \mathbb{1}(z_{i_1,1}, z_{i_3,1} < z_{i_2,1}, z_{i_4,1}) + \mathbb{1}(z_{i_2,1}, z_{i_4,1} < z_{i_1,1}, z_{i_3,1}) \\ & - \mathbb{1}(z_{i_1,1}, z_{i_4,1} < z_{i_2,1}, z_{i_3,1}) - \mathbb{1}(z_{i_2,1}, z_{i_3,1} < z_{i_1,1}, z_{i_4,1}) \Big\} \\ & \Big\{ \mathbb{1}(z_{i_1,2}, z_{i_3,2} < z_{i_2,2}, z_{i_4,2}) + \mathbb{1}(z_{i_2,2}, z_{i_4,2} < z_{i_1,2}, z_{i_3,2}) \\ & - \mathbb{1}(z_{i_1,2}, z_{i_4,2} < z_{i_2,2}, z_{i_3,2}) - \mathbb{1}(z_{i_2,2}, z_{i_3,2} < z_{i_1,2}, z_{i_4,2}) \Big\} \end{aligned}$$

Bergsma and Dassios (2014): $\mathbb{E}h_\tau^* \ge 0$, and if (X, Y) are absolutely continuous or discrete or a mixture of both, then $\mathbb{E}h_\tau^* = 0$ iff X, Y are independent.

a mysterious kernel…

What we knew

- **Hoeffding, BKR, and Bergsma-Dassios are really alike each other…**

The Statistician (2003)
52, *Part* 1, *pp.* 41–57

On the conventional wisdom regarding two consistent tests of bivariate independence

Govind S. Mudholkar
University of Rochester, USA

and Gregory E. Wilding
State University of New York, Buffalo, USA

[Received October 2001. Final revision July 2002]

Summary. Hoeffding's test of bivariate independence and its asymptotic equivalent due to Blum, Kiefer and Rosenblatt are well known to be consistent against all dependence alternatives. However, the two tests, which are often treated as interchangeable, are rarely used in data analysis mainly because their finite sample null distributions are unavailable, and little is known about their operating characteristics. In this paper the conventional wisdom regarding the equivalence of these tests and their distributions is examined by first tabulating their null distributions for sample sizes $n = 5, 6, \ldots, 25, 30, \ldots, 50, 60, \ldots, 100$, and then studying their power functions empirically. The power functions are compared with those of the commonly used methods based on the product moment correlation, the rank correlation and Kendall's τ, for bivariate normal and log-normal populations, as well as a variety of dependence models such as the well-known copulas due to Morgenstern, Gumbel, Plackett, Marshall and Olkin, Raftery, Clayton and Frank. It is seen that the Blum, Kiefer and Rosenblatt test is generally preferable in terms of power against positive dependence alternatives and that the conventional wisdom deserves a revision.

Keywords: Blum, Kiefer and Rosenblatt test; Copulas; Hoeffding test; Kendall's τ; Product moment correlation; Rank correlation

Electronic Journal of Statistics
Vol. 10 (2016) 2287–2311
ISSN: 1935-7524
DOI: 10.1214/16-EJS1165

Large-sample theory for the Bergsma-Dassios sign covariance

Preetam Nandy[*,†]

Seminar for Statistics, ETH Zürich, Switzerland
e-mail: nandy@stat.math.ethz.ch

Luca Weihs[*] **and Mathias Drton**

Department of Statistics, University of Washington, Seattle, WA, USA
e-mail: lucaw@uw.edu; md5@uw.edu

Abstract: The Bergsma-Dassios sign covariance is a recently proposed extension of Kendall's tau. In contrast to tau or also Spearman's rho, the new sign covariance τ^* vanishes if and only if the two considered random variables are independent. Specifically, this result has been shown for continuous as well as discrete variables. We develop large-sample distribution theory for the empirical version of τ^*. In particular, we use theory for degenerate U-statistics to derive asymptotic null distributions under independence and demonstrate in simulations that the limiting distributions give useful approximations.

MSC 2010 subject classifications: Primary 62G10, 62G20; secondary 62G30.
Keywords and phrases: Test of independence, asymptotics, U-statistics, nonparametric correlation, degeneracy, Hoeffding's D.

Received February 2016.

- **As a matter of fact, they are asymptotically equivalent up to scaling!**

What we now know

- **There exists an identity between Hoeffding, BKR, and Bergsma-Dassios!**

$$\boxed{3\widehat{D}_n + 2\widehat{R}_n = 5\widehat{\tau}_n^*}$$

as long as there is no tie

(哲学系若手研究者育成プロジェクト)

ON MEASURES OF ASSOCIATION AND A RELATED PROBLEM

TAKEMI YANAGIMOTO

(Received Feb. 17, 1969; revised Aug. 20, 1969)

Two measures of association $M_1(F)$ and $M_2(F)$ are discussed, which are defined by the expectations of certain rank statistics, T_1 and T_2, respectively. W. Hoeffding [1] has introduced the measure $M_1(F)$ and some of its properties. $M_i(F)$, $i=1, 2$, have desirable properties as the measures of association, for example, $M_i(F)=0$, iff $F(x, y)$ is independent, and $M_i(\Phi\rho)$ is a monotone increasing function of $|\rho|$, when $\Phi\rho$ is the d.f. of two-dimensional normal distribution with correlation coefficient ρ. In Section 2 precise properties are obtained under mild conditions. In Section 3, using these measures, we give a complete result on a relation between equiprobable rankings and independence, which is an improvement of a result by Hoeffding [1].

Outline

- Bivariate case
- High dimensional case
- Discussion

Discussion

- The first optimal consistent test of independence in high dimensions.
- A Cramer moderate deviation theorem for general degenerate U-statistics has also been derived.
- Room still left for improvement.
- Open doors to new problems?
- We note that computing all three is super fast; O{n Pol(log n)} complexity.

High dimensional independence testing with maxima of rank correlations

Mathias Drton, Fang Han, Hongjian Shi

(Submitted on 14 Dec 2018)

Testing mutual independence for high dimensional observations is a fundamental statistical challenge. Popular tests based on linear and simple rank correlations are known to be incapable of detecting non-linear, non-monotone relationships, calling for methods that can account for such dependences. To address this challenge, we propose a family of tests that are constructed using maxima of pairwise rank correlations that permit consistent assessment of pairwise independence. Built upon a newly developed Cramér-type moderate deviation theorem for degenerate U-statistics, our results cover a variety of rank correlations including Hoeffding's D, Blum-Kiefer-Rosenblatt's R, and Bergsma-Dassios-Yanagimoto's τ^*. The proposed tests are distribution-free, implementable without the need for permutation, and are shown to be rate-optimal against sparse alternatives under the Gaussian copula model. As a by-product of the study, we reveal an identity between the aforementioned three rank correlation statistics, and hence make a step towards proving a conjecture of Bergsma and Dassios.

Comments: 43 pages
Subjects: Statistics Theory (math.ST)
Cite as: arXiv:1812.06189 [math.ST]
(or arXiv:1812.06189v1 [math.ST] for this version)

Two problems

- **Problem 1: marginal rank**
- **Problem 2: multivariate rank**

Problem 2

We consider **testing independence** of **two random vectors of fixed probability measure** based on limited information.

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

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

Data:

$$\boxed{\begin{array}{c} (\boldsymbol{X}_1, \boldsymbol{Y}_1) \\ (\boldsymbol{X}_2, \boldsymbol{Y}_2) \\ \vdots \\ (\boldsymbol{X}_n, \boldsymbol{Y}_n) \end{array}}$$

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

Paradigm

Goals to reach:

- the test should be distribution-free, hence directly implementable without the need of permutation;
- the test should be consistent in a certain sense;
- the test should be optimal under certain standard.

a long-standing open problem!

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

impossible in certain sense

Outline

- Multivariate rank
- The proposed test
- Discussion

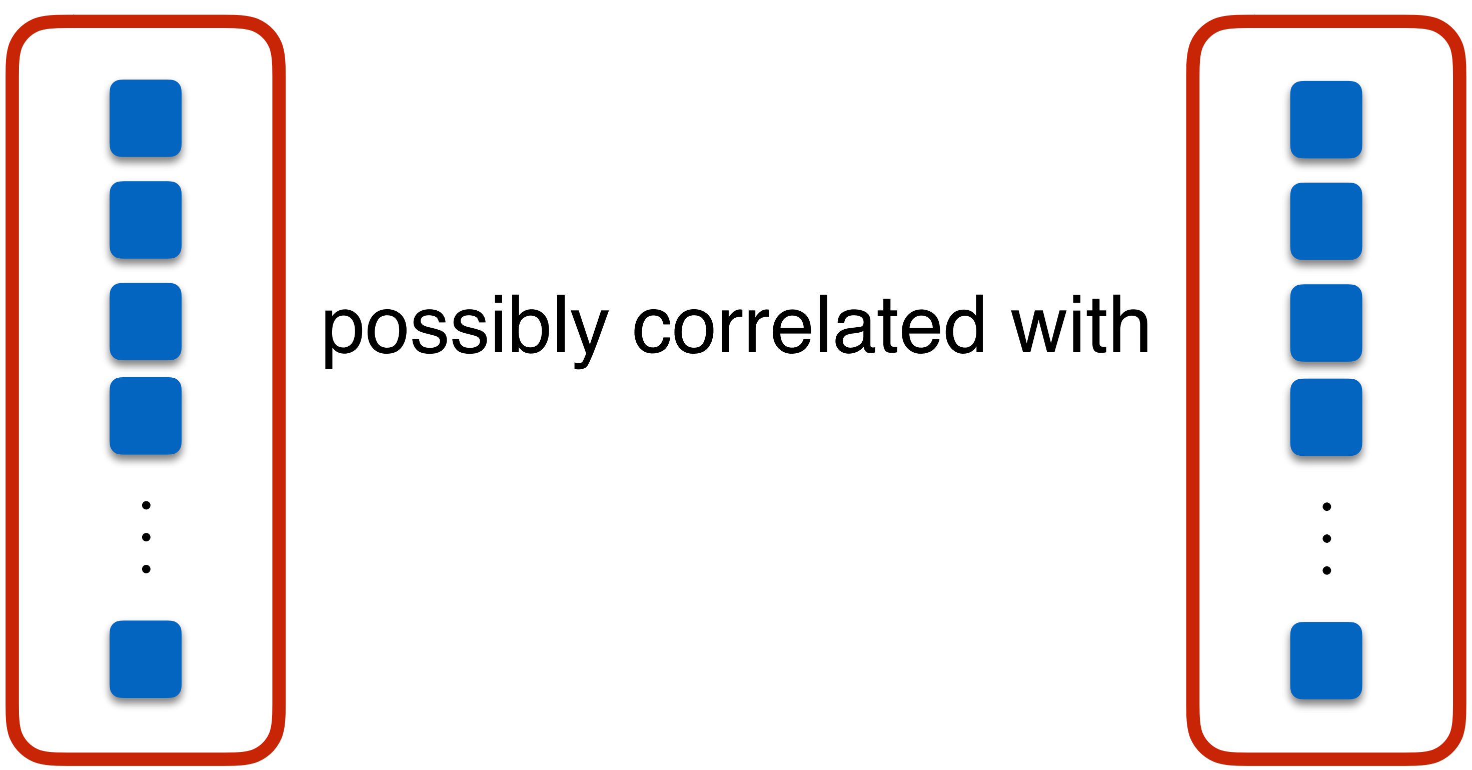
Hallin's multivariate rank

- **Data:** $\{(\boldsymbol{X}_i \in \mathbb{R}^p, \boldsymbol{Y}_i \in \mathbb{R}^q), i \in [n]\}$ i.i.d. distributed with fixed nonvanishing probability measures for $\boldsymbol{X}, \boldsymbol{Y}$.

- **Aim:** testing if "$H_0 : \boldsymbol{X} \perp\!\!\!\perp \boldsymbol{Y}$" is true.

- the test should be distribution-free, hence directly implementable without the need of permutation;

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



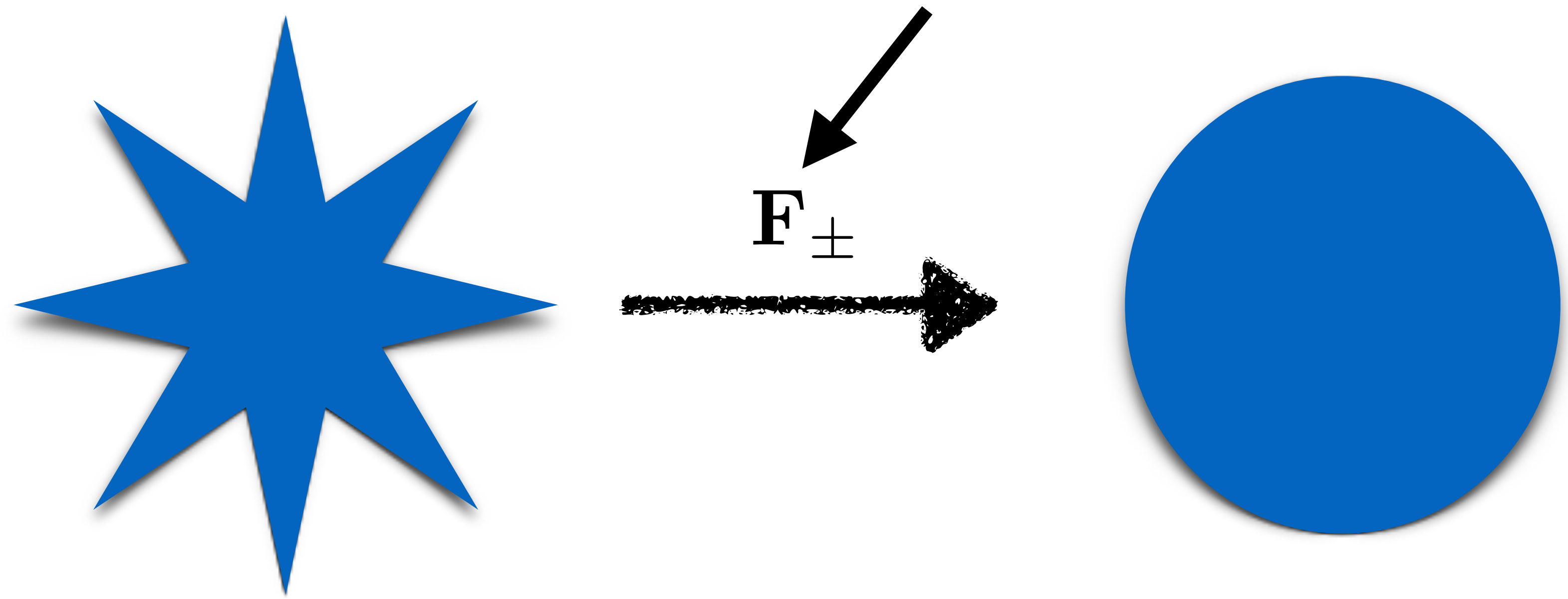
ranks of $\{X_{1,j}, \dots, X_{n,j}\}$ ranks of $\{X_{1,k}, \dots, X_{n,k}\}$

Solution

- **Insight: redefining rank in general dimension**

(Marc Hallin;
ULB MATHEMATICAL
STATISTICS GROUP)

"Hallin's" population distribution function



$\mathbf{F}_{\pm}$

P_d

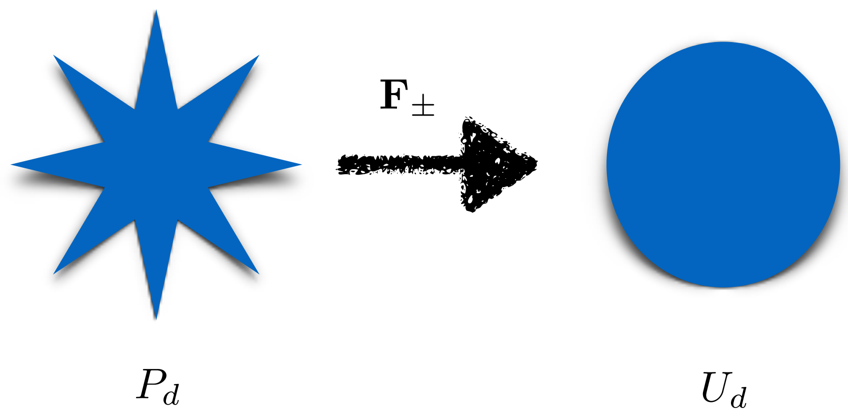
a general measure satisfying the non-vanishing property

U_d

uniform distribution over *d*-dimensional unit-ball

Solution

- **Insight: redefining rank 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, uniqueness, and bijection: Main Theorem in McCann (1995, p. 310)

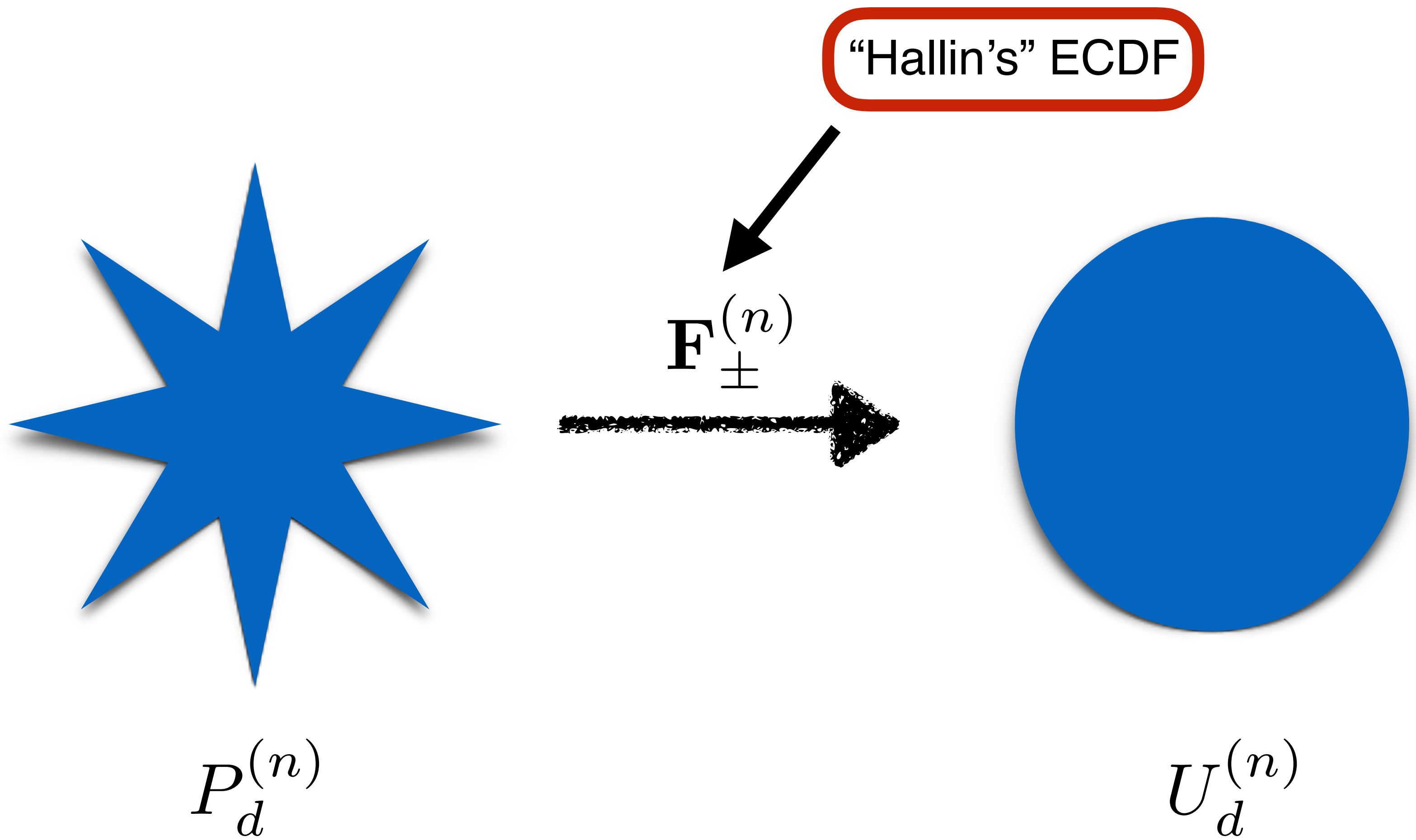
- Homeomorphisms: Theorem 1.1 in Figalli (2018, p. 2):

"The set $\mathbf{F}_{\pm}^{-1}(\mathbf{0})$ is compact and has Lebesgue measure zero; the restrictions of $\mathbf{F}_{\pm}(\cdot)$ and $\mathbf{F}_{\pm}^{-1}(\cdot)$ to $\mathbb{R}^d \backslash \mathbf{F}_{\pm}^{-1}(\mathbf{0})$ and $\mathbb{S}_d \backslash \{\mathbf{0}\}$ are homeomorphisms between $\mathbb{R}^d \backslash \mathbf{F}_{\pm}^{-1}(\mathbf{0})$ and $\mathbb{S}_d \backslash \{\mathbf{0}\}$".

Solution

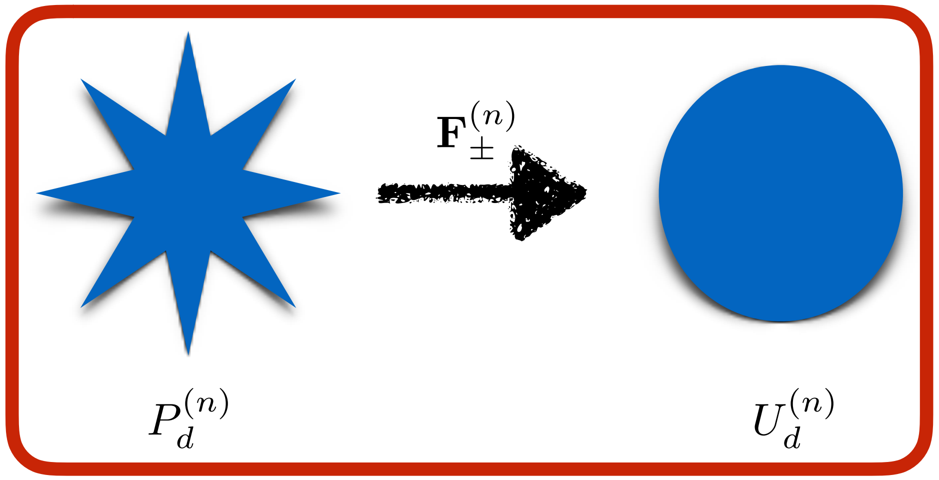
- **Insight: how to estimate**

(Marc Hallin;
ULB MATHEMATICAL
STATISTICS GROUP)

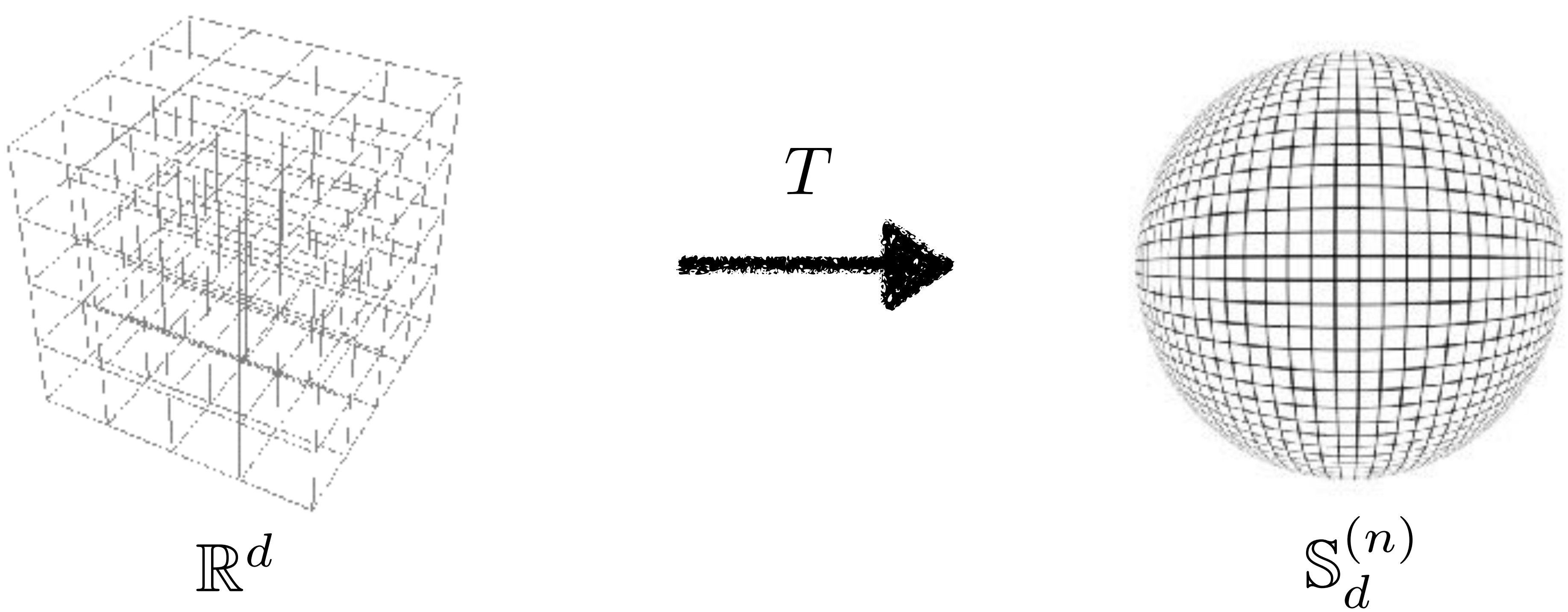


Solution

- **Insight: redefining rank in general dimension**



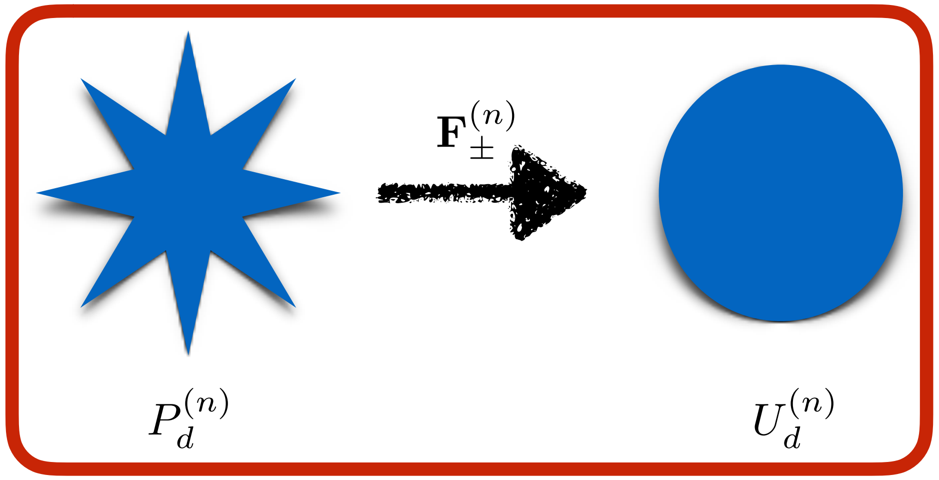
$$\mathbf{F}_{\pm}^{(n)} := \underset{T \in \mathcal{T}}{\operatorname{argmin}} \sum_{i=1}^{n} \left\| \boldsymbol{X}_i - T(\boldsymbol{X}_i) \right\|_2^2$$



consisting of *n* points that approximate the unit-ball

Solution

- **Insight: redefining rank in general dimension**



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

"Hallin's" rank is distribution-free:

Hallin 2017 (Proposition 6.1): Let $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n$ be i.i.d. with nonvanishing distribution P_d. Then $(\mathbf{F}_{\pm}^{(n)}(\boldsymbol{X}_1), \ldots, \mathbf{F}_{\pm}^{(n)}(\boldsymbol{X}_n))$ is uniformly distributed over all permutations of $\mathbb{S}_d^{(n)}$.

And we also have a G-C theorem to quantify convergence.

Outline

- Hallin's multivariate rank
- The proposed test
- Discussion

Generalized BKR Test

- **Data:** $\{(\boldsymbol{X}_i \in \mathbb{R}^p, \boldsymbol{Y}_i \in \mathbb{R}^q), i \in [n]\}$ i.i.d. distributed with fixed nonvanishing probability measures for $\boldsymbol{X}, \boldsymbol{Y}$.

- **Aim:** testing if "$H_0 : \boldsymbol{X} \perp\!\!\!\perp \boldsymbol{Y}$" is true.

Proposal:

- Calculate Hallin's ranks $\mathbf{F}^{(n)}_{\boldsymbol{X},\pm}(\boldsymbol{X}_1), \ldots, \mathbf{F}^{(n)}_{\boldsymbol{X},\pm}(\boldsymbol{X}_n)$ and $\mathbf{F}^{(n)}_{\boldsymbol{Y},\pm}(\boldsymbol{Y}_1), \ldots, \mathbf{F}^{(n)}_{\boldsymbol{Y},\pm}(\boldsymbol{Y}_n)$;

- Combine Hallin's 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.

converge to BKR in one-dimension

Theory

What is the null distribution of $\widehat{M}_n$?

- Consider the "population" Hallin's 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:

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

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

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

  where $\lambda_1, \lambda_2, \ldots$ are nonnegative constants only depend on p, q.

Theory

Main Theorem (SDH 2019+). Let $(\boldsymbol{X}_1, \boldsymbol{Y}_1), \ldots, (\boldsymbol{X}_n, \boldsymbol{Y}_n)$ be independent copies of $(\boldsymbol{X}, \boldsymbol{Y})$ with fixed nonvanishing probability measures P_X, P_Y and $\boldsymbol{X}$ and $\boldsymbol{Y}$ are independent. Then it holds that

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

Hallin 2017 (Proposition 6.1): Let $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n$ be i.i.d. with nonvanishing distribution P_d. Then $(\mathbf{F}_{\pm}^{(n)}(\boldsymbol{X}_1), \ldots, \mathbf{F}_{\pm}^{(n)}(\boldsymbol{X}_n))$ is uniformly distributed over all permutations of $\mathbb{S}_d^{(n)}$.

$\widehat{M}_n := n \cdot \mathrm{dCov}_n^2((\mathbf{F}_{\boldsymbol{X},\pm}^{(n)}(\boldsymbol{X}_i))_{i=1}^n, (\mathbf{F}_{\boldsymbol{Y},\pm}^{(n)}(\boldsymbol{Y}_i))_{i=1}^n)$ is relying on the product space of two uniform permutation measures.

The Main Theorem is hence intrinsically a combinatorial non-central limit theorem.

Combinatorial CLT

A COMBINATORIAL CENTRAL LIMIT THEOREM[1]

By Wassily Hoeffding

Institute of Statistics, University of North Carolina

1. Summary. Let $(Y_{n1}, \cdots, Y_{nn})$ be a random vector which takes on the $n!$ permutations of $(1, \cdots, n)$ with equal probabilities. Let $c_n(i, j), i, j = 1, \cdots, n$, be n^2 real numbers. Sufficient conditions for the asymptotic normality of

$$S_n = \sum_{i=1}^{n} c_n(i, Y_{ni})$$

are given (Theorem 3). For the special case $c_n(i, j) = a_n(i)b_n(j)$ a stronger version of a theorem of Wald, Wolfowitz and Noether is obtained (Theorem 4). A condition of Noether is simplified (Theorem 1).

2. Introduction and statement of results. An example of what is here meant by a combinatorial central limit theorem is a solution of the following problem. For every positive integer n there are given $2n$ real numbers $a_n(i), b_n(i), i = 1, \cdots, n$. It is assumed that the $a_n(i)$ are not all equal and the $b_n(i)$ are not all equal. Let $(Y_{n1}, \cdots, Y_{nn})$ be a random vector which takes on the $n!$ permutations of $(1, \cdots, n)$ with equal probabilities $1/n!$. Under what conditions is

$$S_n = \sum_{i=1}^{n} a_n(i)b_n(Y_{ni}) \tag{1}$$

asymptotically normally distributed as $n \to \infty$?

Throughout this paper a random variable S_n will be called asymptotically normal or asymptotically normally distributed if

$$\lim_{n\to\infty} \Pr\{S_n - ES_n \le x\sqrt{\operatorname{var} S_n}\} = \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{x} e^{-\frac{1}{2}y^2}\,dy, \qquad -\infty < x < \infty,$$

where ES_n and $\operatorname{var} S_n$ are the mean and the variance of S_n.

In the particular case $a_n(i) = b_n(i) = i$ the asymptotic normality of S_n was proved by Hotelling and Pabst [2]. The first general result is due to Wald and Wolfowitz [6], who showed that S_n is asymptotically normal if, as $n \to \infty$,

$$\frac{\frac{1}{n}\sum_{i=1}^{n}(a_n(i) - \bar{a}_n)^r}{\left[\frac{1}{n}\sum_{i=1}^{n}(a_n(i) - \bar{a}_n)^2\right]^{r/2}} = O(1), \qquad r = 3, 4, \cdots, \tag{2}$$

and

$$\frac{\frac{1}{n}\sum_{i=1}^{n}(b_n(i) - \bar{b}_n)^r}{\left[\frac{1}{n}\sum_{i=1}^{n}(b_n(i) - \bar{b}_n)^2\right]^{r/2}} = O(1), \qquad r = 3, 4, \cdots, \tag{3}$$

[1] Work done under the sponsorship of the Office of Naval Research.

The Annals of Statistics
1997, Vol. 25, No. 5, 2210–2227

ERROR BOUND IN A CENTRAL LIMIT THEOREM OF DOUBLE-INDEXED PERMUTATION STATISTICS

By Lincheng Zhao[1], Zhidong Bai[2], Chern-Ching Chao[2] and Wen-Qi Liang[2]

University of Science and Technology of China, National Sun Yat-Sen University, Academia Sinica and Academia Sinica

An error bound in the normal approximation to the distribution of the double-indexed permutation statistics is derived. The derivation is based on Stein's method and on an extension of a combinatorial method of Bolthausen. The result can be applied to obtain the convergence rate of order $n^{-1/2}$ for some rank-related statistics, such as Kendall's tau, Spearman's rho and the Mann–Whitney–Wilcoxon statistic. Its applications to graph-related nonparametric statistics of multivariate observations are also mentioned.

1. Introduction. Let $\xi(i, j, k, l)$, $i, j, k, l \in N = \{1, \ldots, n\}$, be real numbers depending on n. We are interested in the double-indexed permutation statistics (DIPS) of the general form $\sum_{i,j}\xi(i, j, \pi(i), \pi(j))$, where π is uniformly distributed on the set $\mathscr{P}_n$ of all permutations of N. The DIPS of the restricted form $\sum_{i,j} a_{ij} b_{\pi(i)\pi(j)}$ was first investigated by Daniels (1944) in the study of a generalized correlation coefficient with Kendall's tau and Spearman's rho being special cases. Daniels gave a set of sufficient conditions for their asymptotic normality as $n \to \infty$. Further investigations along this direction have been done by Bloemena (1964), Jogdeo (1968), Abe (1969), Shapiro and Hubert (1979), Barbour and Eagleson (1986) and Pham, Möcks and Sroka (1989). In these contexts, the so-called scores a_{ij} and b_{ij} are either symmetric ($a_{ij} = a_{ji}, b_{ij} = b_{ji}$) or skew-symmetric ($a_{ij} = -a_{ji}, b_{ij} = -b_{ji}$). The uses of DIPS have diversely been suggested by Friedman and Rafsky (1979, 1983) and Schilling (1986) in multivariate nonparametric tests, by Hubert and Schultz (1976) in clustering studies, by Mantel and Valand (1970) in biometry, and by Cliff and Ord (1981) in geography.

The purpose of this paper is to derive a bound for the error in the normal approximation to the distribution of the DIPS of the general form

Received October 1995; revised December 1996.

[1] Research partially supported by National Natural Science Foundation of China, Ph.D. Program Foundation of National Education Committee of China and Special Foundation of Academia Sinica.

[2] Research partially supported by National Science Council of the Republic of China.

AMS 1991 subject classifications. Primary 60F05, 62E20; secondary 62H20.

Key words and phrases. Asymptotic normality, correlation coefficient, graph theory, multivariate association, permutation statistics, Stein's method.

Combinatorial non-CLT

$$\widehat{D}^{(n)} = \binom{n}{2}^{-1} \sum_{1 \le j_1 < j_2 \le n} g_1^{(n)}\left(\boldsymbol{z}_{1;j_1}^{(n)}, \boldsymbol{z}_{1;j_2}^{(n)}\right) g_2^{(n)}\left(\boldsymbol{z}_{2;\pi_{j_1}}^{(n)}, \boldsymbol{z}_{2;\pi_{j_2}}^{(n)}\right)$$

Theorem 4.1. *Assume that for each $i = 1, 2$, the functions $g_i^{(n)}$, $n \in \mathbb{Z}_+$, and g_i satisfy the following conditions:*

(i) each $g_i^{(n)}$ is symmetric, i.e., $g_i^{(n)}(\boldsymbol{z}, \boldsymbol{z}') = g_i^{(n)}(\boldsymbol{z}', \boldsymbol{z})$ for all $\boldsymbol{z}, \boldsymbol{z}' \in \Omega_i$;

(ii) the family $g_i^{(n)}$, $n \in \mathbb{Z}_+$, is equicontinuous;

(iii) each $g_i^{(n)}$ is non-negative definite, i.e.,

$$\sum_{j_1, j_2 = 1}^{\ell} c_{j_1} c_{j_2} g_i^{(n)}(\boldsymbol{z}_{j_1}, \boldsymbol{z}_{j_2}) \ge 0$$

for all $c_1, \ldots, c_\ell \in \mathbb{R}$, $\boldsymbol{z}_1, \ldots, \boldsymbol{z}_\ell \in \Omega_i$, $\ell \in \mathbb{Z}_+$;

(iv) each $g_i^{(n)}$ has $\mathbb{E}(g_i^{(n)}(\boldsymbol{z}, \boldsymbol{Z}_i^{(n)})) = 0$;

(v) each $g_i^{(n)}$ has $\mathbb{E}(g_i^{(n)}(\boldsymbol{Z}_i^{(n)}, \boldsymbol{Z}_i^{(n)\prime})^2) \in (0, +\infty)$;

(vi) as $n \to \infty$, the functions $g_i^{(n)}$ converge uniformly on Ω_i to g_i, with $\mathbb{E}(g_i(\boldsymbol{Z}_i, \boldsymbol{Z}_i')^2) \in (0, +\infty)$.

It then holds that

$$n\widehat{D}^{(n)} \xrightarrow{\mathrm{d}} \sum_{k_1, k_2 = 1}^{\infty} \lambda_{1,k_1} \lambda_{2,k_2} (\xi_{k_1,k_2}^2 - 1)$$

as $n \to \infty$, where ξ_{k_1,k_2}, $k_1, k_2 \in \mathbb{Z}_+$, are i.i.d. standard Gaussian, and the $\lambda_{i,k} \ge 0$, $k \in \mathbb{Z}_+$, are eigenvalues of the Hilbert-Schmidt integral operator given by g_i. So, for each i the $\lambda_{i,k}$ solve the integral equations

$$\mathbb{E}(g_i(\boldsymbol{z}_i, \boldsymbol{Z}_i) e_{i,k}(\boldsymbol{Z}_i)) = \lambda_{i,k} e_{i,k}(\boldsymbol{z}_i)$$

for a system of orthonormal eigenfunctions $e_{i,k}$.

Optimality

(wikipedia page of Lucien Le Cam)

(youtube clip of Otomar Hájek)

6.6 Theorem. *Let P_n and Q_n be sequences of probability measures on measurable spaces $(\Omega_n, \mathcal{A}_n)$, and let $X_n : \Omega_n \mapsto \mathbb{R}^k$ be a sequence of random vectors. Suppose that $Q_n \lhd P_n$ and*

$$\left(X_n, \frac{dQ_n}{dP_n}\right) \overset{P_n}{\rightsquigarrow} (X, V).$$

Then $L(B) = \mathrm{E} 1_B(X) V$ defines a probability measure, and $X_n \overset{Q_n}{\rightsquigarrow} L$.

- from "Asymptotic Statistics" by van der Vaart

Optimality

(wikipedia page of Lucien Le Cam)

(youtube clip of Otomar Hájek)

6.7 *Example (Le Cam's third lemma).* The name *Le Cam's third lemma* is often reserved for the following result. If

$$\left(X_n, \log \frac{dQ_n}{dP_n}\right) \overset{P_n}{\rightsquigarrow} N_{k+1}\left(\begin{pmatrix} \mu \\ -\frac{1}{2}\sigma^2 \end{pmatrix}, \begin{pmatrix} \Sigma & \tau \\ \tau^T & \sigma^2 \end{pmatrix}\right),$$

then

$$X_n \overset{Q_n}{\rightsquigarrow} N_k(\mu + \tau, \Sigma).$$

- from "Asymptotic Statistics" by van der Vaart

$$\widehat{M}_n \xrightarrow{\text{d}} \sum_{k=1}^{\infty} \lambda_k(\xi_k^2 - 1)$$ non-normal limit!

Optimality

6.6 Theorem. *Let P_n and Q_n be sequences of probability measures on measurable spaces $(\Omega_n, \mathcal{A}_n)$, and let $X_n : \Omega_n \mapsto \mathbb{R}^k$ be a sequence of random vectors. Suppose that $Q_n \triangleleft P_n$ and*

$$\left(X_n, \frac{dQ_n}{dP_n}\right) \overset{P_n}{\rightsquigarrow} (X, V).$$

Then $L(B) = \mathrm{E}1_B(X)\, V$ defines a probability measure, and $X_n \overset{Q_n}{\rightsquigarrow} L$.

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

$$P_n := N_{p+q}\left(\begin{bmatrix}0\\0\end{bmatrix}, \begin{bmatrix}I_p & 0\\ 0 & I_q\end{bmatrix}\right)$$

null hypothesis

$$Q_n := N_{p+q}\left(\begin{bmatrix}0\\0\end{bmatrix}, \begin{bmatrix}I_p & C/\sqrt{n}\\ C/\sqrt{n} & I_q\end{bmatrix}\right)$$

local alternative hypothesis

Optimality

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

$$\boxed{P_n := N_{p+q}\left(\begin{bmatrix}0\\0\end{bmatrix}, \begin{bmatrix}I_p & 0\\ 0 & I_q\end{bmatrix}\right)}$$

$$\boxed{Q_n := N_{p+q}\left(\begin{bmatrix}0\\0\end{bmatrix}, \begin{bmatrix}I_p & C/\sqrt{n}\\ C/\sqrt{n} & I_q\end{bmatrix}\right)}$$

Theorem. (1) As C is small enough, it is impossible to differentiate Q_n from P_n.
(2) Considering the threshold $t > 0$ such that

$$P_n(\widehat{M}_n > t) \to \alpha,$$

then for any $\beta > \alpha$, there exists a $C > 0$ only depending on β such that

$$Q_n(\widehat{M}_n > t) > \beta$$

for all n large enough.

Paradigm

Goals to reach:

- the test should be distribution-free, hence directly implementable without the need of permutation;

  Yes, by Hallin's result and our combinatorial non-center limit theorem!

- the test should be consistent in a certain sense;

  Yes, by consistency of dCov, G-C Theorem, and homeomorphisms of Hallin's DF!

- the test should be optimal under certain standard.

  Yes, by a use of Le Cam (and Hajak)'s contiguity lemma!

Discussion

- The first distribution-free consistent test of independence in general dimension.
- A combinatorial non-central limit theorem is derived.
- The ranks can be computed fast.
  General dimension:
  O(n^{5/2} Poly(log n)) complexity;
  if dimension is 2:
  O(n^{3/2+delta} Poly(log n)) complexity

Distribution-free consistent independence tests via Hallin's multivariate rank

Hongjian Shi, Mathias Drton, Fang Han

(Submitted on 22 Sep 2019 (v1), last revised 3 Oct 2019 (this version, v2))

This paper investigates the problem of testing independence of two random vectors of general dimensions. For this, we give for the first time a distribution-free consistent test. Our approach combines distance covariance with a new multivariate rank statistic recently introduced by Hallin (2017). In technical terms, the proposed test is consistent and distribution-free in the family of multivariate distributions with nonvanishing (Lebesgue) probability densities. Exploiting the (degenerate) U-statistic structure of the distance covariance and the combinatorial nature of Hallin's rank, we are able to derive the limiting null distribution of our test statistic. The resulting asymptotic approximation is accurate already for moderate sample sizes and makes the test implementable without requiring permutation. The limiting distribution is derived via a more general result that gives a new type of combinatorial non-central limit theorem for double- and multiple-indexed permutation statistics.

Comments: More references were added and differences between ours and an independent work by Deb and Sen were explained
Subjects: Statistics Theory (math.ST)
Cite as: arXiv:1909.10024 [math.ST]
(or arXiv:1909.10024v2 [math.ST] for this version)

Thanks!