Computational Statistics and Data Analysis 172 (2022) 107480

Contents lists available at ScienceDirect

Computational Statistics and Data Analysis

www.elsevier.com/locate/csda

Multivariate ranks based on randomized lift-interdirections

Šárka Hudecová [a,*], Miroslav Šiman [b]

[a] Department of Probability and Statistics, Faculty of Mathematics and Physics, Charles University. Sokolovská 83, 186 75 Praha 8, Czech Republic

[b] The Czech Academy of Sciences, Institute of Information Theory and Automation. Pod Vodárenskou věží 4, 182 00 Praha 8, Czech Republic

ARTICLE INFO

Article history:
Received 2 March 2021
Received in revised form 18 March 2022
Accepted 19 March 2022
Available online 24 March 2022

Keywords:
Multivariate rank
Lift-interdirection
Interdirection
Rank test
One-sample test
Robustness

ABSTRACT

Every multivariate sign and rank test needs a workable concept of ranks for multivariate data. Unfortunately, multidimensional spaces lack natural ordering and, consequently, there are no universally accepted ways how to rank vector observations. Existing proposals usable beyond small dimensions are very few in number, and each of them has its own advantages and drawbacks. Therefore, new multivariate ranks based on randomized lift-interdirections are presented, discussed and investigated. These naturally robust and invariant hyperplane-based ranks can be computed quickly and easily even in relatively high-dimensional spaces, and they can be used for nonparametric statistical inference in some existing optimal statistical procedures without altering their asymptotic behavior under null hypotheses or changing their performance under local alternatives. This is not only proved theoretically in case of the canonical sign and rank one-sample test for elliptically distributed observations, but also illustrated empirically in a small simulation study.

1. Introduction

The future of data analysis seems high-dimensional and largely nonparametric. This is why there already exist many multivariate analogues to univariate ranks and signs such as spatial ranks and signs (Möttönen and Oja, 1995; Oja, 2005), Oja ranks and signs (Oja, 1999), component-wise ranks and signs (Puri and Sen, 1971), the ranks and signs based on the measure transportation approach (Chernozhukov et al., 2017; Hallin et al., 2021), the ranks of pseudo-Mahalanobis distances (Hallin and Paindaveine, 2002), numerous variants of data depth (Zuo and Serfling, 2000), and other interesting concepts (Chaudhuri and Sengupta, 1993).

The hyperplane-based ranks (Hettmansperger et al., 1999; Oja and Paindaveine, 2005) and signs (Randles, 1989) appear especially appealing due to their simplicity, clear geometric interpretation, full affine invariance, weak moment assumptions, complete avoidance of any shape matrix estimators and robustness to both radial and angular outliers (Oja and Paindaveine, 2005). Unfortunately, these original concepts rely on *ordinary* interdirections (leading to signs) and *ordinary* symmetrized lift-interdirections (leading to ranks) that are too computationally demanding to be useful in most real-life applications. Hudecová et al. (2020) have recently addressed the issue by defining *incomplete* interdirections and *incomplete* symmetrized lift-interdirections to be used instead. Although the resulting signs are already satisfactory in terms of computational speed, the resulting ranks are still somewhat slow to compute because they still require full symmetrization of the observations,

* Corresponding author.
E-mail address: hudecova@karlin.mff.cuni.cz (Š. Hudecová).

https://doi.org/10.1016/j.csda.2022.107480
0167-9473/

which leads to their computational complexity growing exponentially with data dimension. Consequently, this article further improves the incomplete symmetrized lift-interdirections by introducing *randomized* lift-interdirections (i.e., incomplete interdirections with random symmetrization) that are finally quick to compute in high-dimensional spaces, too. Both the randomized lift-interdirections and incomplete interdirections may thus give rise to optimal sign and rank statistical procedures suitable even for high-dimensional data whose dimension is sufficiently smaller than the number of observations. Although it is here rigorously proved and empirically demonstrated only for the canonical one-sample signed-rank test, it is likely to hold for the other tests mentioned in Oja and Paindaveine (2005) as well.

It should be pointed out that the hyperplane-based ranks are known to be meaningful only for elliptically distributed data. Consequently, all the tests using hyperplane-based ranks of any kind heavily rely on the assumption of elliptical symmetry, which may be viewed as too restrictive. But on the other hand, the assumption is often natural and makes it possible to bypass the curse of dimensionality.

Next Section 2 introduces necessary notation, terminology and definitions, Section 3 investigates the ranks based on randomized lift-interdirections, their use in the canonical one-sample test of Oja and Paindaveine (2005), and various theoretical properties of the resulting new test. Section 4 illustrates the new one-sample test with an application to real data and with a small representative simulation study involving both small and large sample sizes and dimensions. It also shows that the ranks of randomized interdirections are really quite fast to compute. The last Section 5 collects concluding comments. The technical proofs can be found in Appendix A.

2. Definitions and notation

Consider n independent and identically distributed p-dimensional stochastic vectors $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n \in \mathbb{R}^p$, $p \geq 2$, forming a random sample $\mathcal{X}_n$ from a continuous distribution. Then the data points are in general position with probability one. In other words, no hyperplane contains more than p observations almost surely. Any k-tuple $\boldsymbol{q}(n,k) = (q_1, \ldots, q_k)$ of distinct integer indices $1 \leq q_1, \ldots, q_k \leq n$ can be associated with the subsample $\mathcal{X}^{\boldsymbol{q}(n,k)} = (\boldsymbol{X}_{q_1}, \ldots, \boldsymbol{X}_{q_k})$. The set of all $\binom{n}{k}$ possible k-tuples $\boldsymbol{q}(n,k)$ will be denoted by $\mathcal{Q}(n,k)$.

Any subsample $\mathcal{X}^{\boldsymbol{q}(n,p)}$ almost surely defines the hyperplane $H^{\boldsymbol{q}(n,p)} \subset \mathbb{R}^p$ containing its observations:

$$\det\big(\mathbb{M}^{\boldsymbol{q}(n,p)}(\boldsymbol{x})\big) = d_0^{\boldsymbol{q}(n,p)} + \boldsymbol{d}^{\boldsymbol{q}(n,p)\prime}\boldsymbol{x} = 0$$

where $\boldsymbol{d}^{\boldsymbol{q}(n,p)} = (d_1^{\boldsymbol{q}(n,p)}, \ldots, d_p^{\boldsymbol{q}(n,p)})'$ and $d_j^{\boldsymbol{q}(n,p)}$, $j = 0, \ldots, p$, is the cofactor of the $(j+1)$th element in the last column of matrix

$$\mathbb{M}^{\boldsymbol{q}(n,p)}(\boldsymbol{x}) = ((1, \boldsymbol{X}'_{q_1})', (1, \boldsymbol{X}'_{q_2})', \ldots, (1, \boldsymbol{X}'_{q_p})', (1, \boldsymbol{x}')').$$

Similarly, any subsample $\mathcal{X}^{\boldsymbol{q}(n,p-1)}$ almost surely defines the hyperplane $H^{\boldsymbol{q}(n,p-1)} \subset \mathbb{R}^p$ containing all of its observations and the origin:

$$\det\big(\mathbb{M}^{\boldsymbol{q}(n,p-1)}(\boldsymbol{x})\big) = \boldsymbol{d}^{\boldsymbol{q}(n,p-1)\prime}\boldsymbol{x} = 0$$

where $\boldsymbol{d}^{\boldsymbol{q}(n,p-1)} = (d_1^{\boldsymbol{q}(n,p-1)}, \ldots, d_p^{\boldsymbol{q}(n,p-1)})'$ and $d_j^{\boldsymbol{q}(n,p-1)}$, $j = 1, \ldots, p$, is the cofactor of the jth element in the last column of matrix

$$\mathbb{M}^{\boldsymbol{q}(n,p-1)}(\boldsymbol{x}) = (\boldsymbol{X}_{q_1}, \boldsymbol{X}_{q_2}, \ldots, \boldsymbol{X}_{q_{p-1}}, \boldsymbol{x}).$$

Obviously, the signs

$$S^{\boldsymbol{q}(n,p-1)}(\boldsymbol{x}) = \operatorname{sign}(\boldsymbol{d}^{\boldsymbol{q}(n,p-1)\prime}\boldsymbol{x}) \text{ and}$$
$$S^{\boldsymbol{q}(n,p)}(\boldsymbol{x}) = \operatorname{sign}(d_0^{\boldsymbol{q}(n,p)} + \boldsymbol{d}^{\boldsymbol{q}(n,p)\prime}\boldsymbol{x})$$

indicate the position of $\boldsymbol{x} \in \mathbb{R}^p$ with respect to $H^{\boldsymbol{q}(n,p-1)}$ and $H^{\boldsymbol{q}(n,p)}$, respectively.

Any couple of points $\boldsymbol{y}_1, \boldsymbol{y}_2 \in \mathbb{R}^p$ has two fundamental affine-invariant hyperplane-based characteristics, namely interdirection

$$C_{\boldsymbol{y}_1,\boldsymbol{y}_2} = C_{\boldsymbol{y}_1,\boldsymbol{y}_2}(\mathcal{X}_n) = \mathrm{I}(\boldsymbol{y}_1 \neq \boldsymbol{y}_2) \cdot \sum_{\boldsymbol{q} \in \mathcal{Q}_C} \frac{1 - S^{\boldsymbol{q}}(\boldsymbol{y}_1) S^{\boldsymbol{q}}(\boldsymbol{y}_2)}{2},$$

defined for $\mathcal{Q}_C \subset \mathcal{Q}(n, p-1)$, and lift-interdirection

$$L_{\boldsymbol{y}_1,\boldsymbol{y}_2} = L_{\boldsymbol{y}_1,\boldsymbol{y}_2}(\mathcal{X}_n) = \sum_{\boldsymbol{q} \in \mathcal{Q}_L} \frac{1 - S^{\boldsymbol{q}}(\boldsymbol{y}_1) S^{\boldsymbol{q}}(\boldsymbol{y}_2)}{2},$$

defined for $\mathcal{Q}_L \subset \mathcal{Q}(n, p)$. Any point $\boldsymbol{y} \in \mathbb{R}^p$ (with its reflection $-\boldsymbol{y}$) also gives rise to symmetrized lift-interdirection

$$\underline{L}_{\boldsymbol{y}} = \underline{L}_{\boldsymbol{y}}(\mathcal{X}_n) = \sum_{\substack{\boldsymbol{q}\in\mathcal{Q}_L \\ \boldsymbol{s}\in\mathcal{S}(\boldsymbol{q})}} \sum \frac{1 - S_{\boldsymbol{s}}^{\boldsymbol{q}(n,p)}(\boldsymbol{y}) S_{\boldsymbol{s}}^{\boldsymbol{q}(n,p)}(-\boldsymbol{y})}{2}$$

where $S_{\boldsymbol{s}}^{\boldsymbol{q}(n,p)}(\boldsymbol{x}) = \operatorname{sign}(d_{0\boldsymbol{s}}^{\boldsymbol{q}(n,p)} + \boldsymbol{d}_{\boldsymbol{s}}^{\boldsymbol{q}(n,p)'}\boldsymbol{x})$, $(d_{0\boldsymbol{s}}^{\boldsymbol{q}(n,p)}, \boldsymbol{d}_{\boldsymbol{s}}^{\boldsymbol{q}(n,p)'})'$ is nothing but the vector of cofactors of the last column of matrix

$$\mathbb{M}_{\boldsymbol{s}}^{\boldsymbol{q}(n,p)}(\boldsymbol{x}) = ((1, s_1\boldsymbol{X}'_{q_1})', (1, s_2\boldsymbol{X}'_{q_2})', \ldots, (1, s_p\boldsymbol{X}'_{q_p})', (1, \boldsymbol{x}')'),$$

$\mathcal{S}(\boldsymbol{q}) \subset \{-1,1\}^p$ for any $\boldsymbol{q} \in \mathcal{Q}_L$, and $\{-1,1\}^p$ is the set of all 2^p p-dimensional vectors $\boldsymbol{s} = (s_1, \ldots, s_p)'$ with individual coordinates equal to either 1 or -1. The sets $\mathcal{Q}_C$ and $\mathcal{Q}_L$ will be called design sets hereinafter.

The ordinary concepts of Randles (1989) and Oja and Paindaveine (2005) correspond to $\mathcal{Q}_C = \mathcal{Q}(n, p-1)$, $\mathcal{Q}_L = \mathcal{Q}(n,p)$, and $\mathcal{S}(\boldsymbol{q}) = \{-1,1\}^p$, $\boldsymbol{q} \in \mathcal{Q}_L$. Observe that ordinary $C_{\boldsymbol{y}_1,\boldsymbol{y}_2}(\mathcal{X}_n)$ (resp. $L_{\boldsymbol{y}_1,\boldsymbol{y}_2}(\mathcal{X}_n)$) counts all the hyperplanes separating $\boldsymbol{y}_1$ from $\boldsymbol{y}_2$ that pass through the origin and $p-1$ observations (resp. through p observations). Ordinary $\underline{L}_{\boldsymbol{y}}(\mathcal{X}_n)$ is then only a symmetrized version of $L_{\boldsymbol{y},-\boldsymbol{y}}(\mathcal{X}_n)$ with desirable invariance with respect to the reflections of the observations around the origin.

The sets $\mathcal{Q}(n,p-1)$ and $\mathcal{Q}(n,p)$ generate too many hyperplanes and effectively prohibit the computation of the ordinary characteristics in multidimensional spaces. Nevertheless, one can consider only some hyperplanes and still obtain meaningful results. This possibility leads to incomplete modifications with possibly random $\mathcal{Q}_C \subset \mathcal{Q}(n,p-1)$ and $\mathcal{Q}_L \subset \mathcal{Q}(n,p)$, but still with $\mathcal{S}(\boldsymbol{q}) = \{-1,1\}^p$, $\boldsymbol{q} \in \mathcal{Q}_L$. They have been analyzed by means of the theory of incomplete U-statistics in Hudecová et al. (2020). Unfortunately, the use of all sign vectors in the symmetrization still makes the computation of incomplete $\underline{L}_{\boldsymbol{y}}(\mathcal{X}_n)$ too demanding for large p because the number of all sign vectors grows with p exponentially. The incomplete variants and corresponding quantities such as ranks and angular estimators may be denoted with $\widetilde{\ }$.

Finally, the randomized lift-interdirections presented in this article correspond to $\mathcal{Q}_L \subset \mathcal{Q}(n,p)$ and $\mathcal{S}(\boldsymbol{q}) = \{\boldsymbol{S}_{\boldsymbol{q}}\}$, $\boldsymbol{S}_{\boldsymbol{q}} \in \{-1,1\}^p$, where $\boldsymbol{S}_{\boldsymbol{q}}$ is chosen independently of $\boldsymbol{q}$, $\mathcal{X}_n$, $\mathcal{Q}_C$ and $\mathcal{Q}_L$ at random from the uniform distribution on $\{-1,1\}^p$. Compared to the case of incomplete symmetrized lift-interdirections, the size of $\mathcal{S}(\boldsymbol{q})$ is here reduced from 2^p to 1 for each $\boldsymbol{q} \in \mathcal{Q}_L$, which may obviously speed the computation of $\underline{L}_{\boldsymbol{y}}$ considerably. However, the resulting $\underline{L}_{\boldsymbol{y}}$ is not an incomplete U-statistic any more. The randomized lift-interdirections and their ranks may be distinguished with $\widehat{\ }$ in the following text.

The concept of interdirections and symmetrized lift-interdirections is very useful for statistical inference whenever

$$a_{\boldsymbol{y}_1,\boldsymbol{y}_2} := \pi C_{\boldsymbol{y}_1,\boldsymbol{y}_2}(\mathcal{X}_n)/|\mathcal{Q}_C|$$

is a consistent estimator of the angle $\alpha(\boldsymbol{y}_1, \boldsymbol{y}_2)$ between $\boldsymbol{y}_1$ and $\boldsymbol{y}_2$,

$$\alpha(\boldsymbol{y}_1, \boldsymbol{y}_2) := \arccos\left(\frac{\boldsymbol{y}_1'\boldsymbol{y}_2}{\|\boldsymbol{y}_1\|\|\boldsymbol{y}_2\|}\right),$$

and when the normalized ranks $R_i/(n+1)$ and $\underline{R}_i/(n+1)$ of $L_{\boldsymbol{X}_i,-\boldsymbol{X}_i}(\mathcal{X}_n)$ and $\underline{L}_{\boldsymbol{X}_i}(\mathcal{X}_n)$ in the corresponding samples of the same quantities closely approximate the normalized ranks $R_i^M/(n+1)$ of (pseudo-)Mahalanobis distances of $\boldsymbol{X}_i$'s from the center. All that appears to happen for elliptical distributions, see Proposition 1 below.

Recall that the density of any elliptical distribution $\mathcal{E}_{\boldsymbol{\theta},\boldsymbol{\Sigma},f}$ with median vector $\boldsymbol{\theta} \in \mathbb{R}^p$ and positive definite scatter matrix $\boldsymbol{\Sigma} \in \mathbb{R}^{p\times p}$ is proportional to

$$f\left(\sqrt{(\boldsymbol{x}-\boldsymbol{\theta})'\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\theta})}\right), \quad \boldsymbol{x} \in \mathbb{R}^p, \tag{1}$$

for some function $f : [0,\infty) \to [0,\infty)$ satisfying $\int_0^\infty z^{p-1} f(z)dz < \infty$. If a random vector $\boldsymbol{X}$ follows such distribution, then $r(\boldsymbol{X}) := \|\boldsymbol{\Sigma}^{-1/2}(\boldsymbol{X}-\boldsymbol{\theta})\|$ has cumulative distribution function $\tilde{F}_f$ with corresponding density $\tilde{f}_f(z)$ proportional to $z^{p-1}f(z)I[z>0]$, and R_i^M's are the ranks of the $r(\boldsymbol{X}_i)$'s.

The following assumption collects some requirements on the (design) sets used for defining and computing the different variants of interdirections and (symmetrized) lift-interdirections.

Assumption A.

1. *Either $\mathcal{Q}_C = \mathcal{Q}(n,p-1)$ or $\mathcal{Q}_C$ has a fixed deterministic size $|\mathcal{Q}_C|$, its elements are sampled from $\mathcal{Q}(n,p-1)$ randomly with or without replacement and $|\mathcal{Q}_C|/n \to \infty$ for $n \to \infty$.*
2. *Either $\mathcal{Q}_L = \mathcal{Q}(n,p)$ or $\mathcal{Q}_L$ has a fixed deterministic size $|\mathcal{Q}_L|$, its elements are sampled from $\mathcal{Q}(n,p)$ randomly with or without replacement and $|\mathcal{Q}_L|/n \to \infty$ for $n \to \infty$.*
3. *For any $\boldsymbol{q}$, $\mathcal{S}(\boldsymbol{q}) = \{-1,1\}^p$.*
4. *For any $\boldsymbol{q}$, $\mathcal{S}(\boldsymbol{q}) = \{\boldsymbol{S}_{\boldsymbol{q}}\}$ where $\boldsymbol{S}_{\boldsymbol{q}}$ is a vector sampled independently and randomly from the uniform distribution on $\{-1,1\}^p$.*

Assumption B. *(Score) function $K : (0, 1) \to \mathbb{R}$ is continuous and satisfies*

$$\frac{1}{n}\sum_{i=1}^{n} |K(i/(n+1))|^{2+\delta} \to \int_0^1 |K(u)|^{2+\delta} du < \infty$$

for some $\delta > 0$.

3. Theory

Next Proposition 1 summarizes some relevant and useful results established in Randles (1989), Oja and Paindaveine (2005), and Hudecová et al. (2020).

Proposition 1. *Let $\mathcal{X}_n$ be a p-dimensional random sample $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n$ of size n from the elliptical distribution $\mathcal{E}_{\boldsymbol{\theta},\boldsymbol{\Sigma},f}$, and assume that $\mathcal{X}_n$, $\mathcal{Q}_C$ and $\mathcal{Q}_L$ are independent.*

1. *If $\mathcal{Q}_C$ meets Assumption A1 and $\boldsymbol{\theta} = \boldsymbol{0}$, then $a_{\boldsymbol{y}_1,\boldsymbol{y}_2} = \alpha(\boldsymbol{y}_1, \boldsymbol{y}_2) + o_P(1)$ as $n \to \infty$.*
2. *If $\mathcal{Q}_L$ meets Assumption A2, then $\frac{R_i}{n+1} = \frac{R_i^M}{n+1} + o_P(1)$ as $n \to \infty$, $i = 1, \ldots, n$.*
3. *If $\mathcal{Q}_L$ and $\mathcal{S}(\boldsymbol{q})$ meet Assumptions A2 and A3, then $\frac{R_i}{n+1} = \frac{R_i^M}{n+1} + o_P(1)$ as $n \to \infty$, $i = 1, \ldots, n$.*
4. *Let Assumptions A1–A3 be satisfied for the sets, and consider a function K satisfying Assumption B. Define*

$$S = \frac{p}{n\,\mathrm{E}\,K^2(V)} \sum_{i,j=1}^{n} K\left(\frac{R_i}{n+1}\right) K\left(\frac{R_j}{n+1}\right) \cos(a_{\boldsymbol{X}_i,\boldsymbol{X}_j})$$

where V is uniformly distributed on $[0, 1]$ and either only the ordinary or only the incomplete variants of interdirections and symmetrized lift-interdirections are used. Then $S \xrightarrow{\mathcal{D}} \chi^2_p$ under $H_0 : \boldsymbol{\theta} = \boldsymbol{0}$ for $n \to \infty$.

This article extends the last two statements of Proposition 1 to the randomized lift-interdirections and their ranks. In particular, the new results state that claims 3. and 4. of Proposition 1 hold even for the ranks computed from randomized lift-interdirections, i.e., with Assumption A4 instead of Assumption A3.

Proposition 2. *Let $\mathcal{X}_n$ be a p-dimensional random sample $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n$ of size n from the elliptical distribution $\mathcal{E}_{\boldsymbol{\theta},\boldsymbol{\Sigma},f}$.*

1. *If Assumptions A2 and A4 hold with $\mathcal{X}_n$ and $\mathcal{Q}_L$ independent, then*

$$\frac{\widehat{R}_i}{n+1} = \frac{R_i^M}{n+1} + o_P(1) \tag{2}$$

as $n \to \infty$, $i = 1, \ldots, n$.

2. *Let Assumptions A1, A2, and A4 hold with $\mathcal{X}_n$, $\mathcal{Q}_C$ and $\mathcal{Q}_L$ independent, and consider a function K satisfying Assumption B. Define*

$$S = \frac{p}{n\,\mathrm{E}\,K^2(V)} \sum_{i,j=1}^{n} K\left(\frac{\widehat{R}_i}{n+1}\right) K\left(\frac{\widehat{R}_j}{n+1}\right) \cos(a_{\boldsymbol{X}_i,\boldsymbol{X}_j}) \tag{3}$$

where V is uniformly distributed on $[0, 1]$ and the ranks are computed from the randomized lift-interdirections. Then $S \xrightarrow{\mathcal{D}} \chi^2_p$ under $H_0 : \boldsymbol{\theta} = \boldsymbol{0}$ for $n \to \infty$.

The technical proof of Proposition 2 can be found in Appendix A.

Note that the choice $K = 1$ leads to the one-sample multivariate sign statistic of Randles (1989) that is equivalent to the traditional sign-test statistic for $p = 1$. The van der Waerden type of the test statistic results from $K = \sqrt{G_p^{-1}}$ where G_p^{-1} stands for the quantile function of the χ^2_p distribution.

Now denote as $\mathcal{H}(\boldsymbol{\theta}, \boldsymbol{\Sigma}, f)$ the hypothesis that $\mathcal{X}_n$ is a random sample from elliptical distribution $\mathcal{E}_{\boldsymbol{\theta},\boldsymbol{\Sigma},f}$. According to Proposition 2, the test statistic S of (3) is asymptotically distribution-free under the null hypothesis

$$H_0 : \boldsymbol{\theta} = \boldsymbol{0}, \text{ more precisely } H_0 = \cup_{\boldsymbol{\Sigma}} \cup_f \mathcal{H}(\boldsymbol{0}, \boldsymbol{\Sigma}, f). \tag{4}$$

(The unions will always be taken over the largest sets compatible with assumptions.) The null hypothesis is rejected if S exceeds $\chi^2_{p,1-\alpha}$, which is the $(1-\alpha)$th quantile of the χ^2_p distribution. Obviously, the hypothesis assuming general $\boldsymbol{\theta}_0$ could be tested as H_0 on the shifted sample $\boldsymbol{X}_1 - \boldsymbol{\theta}_0, \ldots, \boldsymbol{X}_n - \boldsymbol{\theta}_0$.

Next Proposition 3 clarifies the behavior of test statistic S of (3), computed from the ranks of randomized lift-interdirections, under local alternatives and its asymptotic relative efficiency (ARE) with respect to the classical Hotelling test. All the assertions easily follow from the asymptotic representation of S established here in the proof of Proposition 2 and from Propositions 3 to 5 of Hallin and Paindaveine (2002).

Assumption C. *The elliptical distribution $\mathcal{E}_{\boldsymbol{\theta},\boldsymbol{\Sigma},f}$, with associated functions f, $\tilde{F}_f$ and $\tilde{f}_f$, has absolutely continuous f with derivative f' existing almost everywhere and satisfying $\int_0^\infty [f'(z)]^2[f(z)]^{-1}z^{p-1}dz < \infty$. Furthermore, $\varphi_f := -f'/f$.*

Proposition 3. *Let $\mathcal{X}_n$ be a p-dimensional random sample $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n$ of size n from the elliptical distribution $\mathcal{E}_{\boldsymbol{\theta},\boldsymbol{\Sigma},f}$ satisfying Assumption C, and let Assumptions A1, A2 and A4 hold with independent $\mathcal{X}_n$, $\mathcal{Q}_C$ and $\mathcal{Q}_L$.*

1. *Consider function K satisfying Assumption B and the sequence of local alternatives $\mathcal{H}^{(n)}(n^{-1/2}\boldsymbol{\tau}, \boldsymbol{\Sigma}, f)$ for some $\boldsymbol{\tau} \in \mathbb{R}^p$. Then $S \xrightarrow{\mathcal{D}} \chi^2_p(\lambda)$ as $n \to \infty$, where $\chi^2_p(\lambda)$ is the noncentral χ^2_p distribution with noncentrality parameter λ,*

$$\lambda = \frac{1}{p\,\mathrm{E}\,K^2(V)}\boldsymbol{\tau}^\top\boldsymbol{\Sigma}^{-1}\boldsymbol{\tau}c_f^2 \tag{5}$$

for V uniformly distributed on $[0,1]$ and

$$c_f = \mathrm{E}\,K(V)\varphi_f(\tilde{F}_f^{-1}(V)) = \int_0^1 K(z)\varphi_f(\tilde{F}_f^{-1}(z))dz. \tag{6}$$

2. *If Assumption B holds for $K := \varphi_g \circ \tilde{F}_g^{-1}$ for some g corresponding to the elliptical distribution $\mathcal{E}_{\boldsymbol{\theta},\boldsymbol{\Sigma},g}$ satisfying Assumption C, then the sequence of tests rejecting H_0 of (4) for $S > \chi^2_{p,1-\alpha}$ is locally asymptotically maximin at asymptotic level α against alternatives of the form $\cup_{\boldsymbol{\theta}\neq\boldsymbol{0}} \cup_{\boldsymbol{\Sigma}} \mathcal{H}(\boldsymbol{\theta}, \boldsymbol{\Sigma}, g)$.*
3. *If K satisfies Assumption B and $\int_0^\infty z^{p+1}f(z)dz < \infty$ for a given f, then the asymptotic relative efficiency (ARE) of the sequence of tests rejecting H_0 of (4) for $S > \chi^2_{p,1-\alpha}$ with respect to the Hotelling test is*

$$\mathrm{ARE}_{p,K,f} = \frac{d_f c_f^2}{p^2\,\mathrm{E}\,K^2(V)}$$

where c_f is defined in (6) and $d_f = \mathrm{E}[\tilde{F}_f^{-1}(V)]^2$ for V uniformly distributed on $[0,1]$.

Note that $c_f < \infty$ and $d_f < \infty$ thanks to the assumptions. And see Lehman and Romano (2005) for the information on maximin tests and asymptotic relative efficiency.

The Cauchy-Swartz inequality implies that the noncentrality parameter λ in (5) is maximal for $K = \varphi_f \circ \tilde{F}_f^{-1}$. If the van der Waerden scores $K = \sqrt{G_p^{-1}}$ are considered in Claim 3. of Proposition 3, then always $\mathrm{ARE}_{p,K,f} \geq 1$ and the equality holds only for the multivariate normal distribution, all that according to Proposition 6 of Hallin and Paindaveine (2002). In other words, the test based on S is then uniformly no worse than the Hotelling one in terms of power against local alternatives.

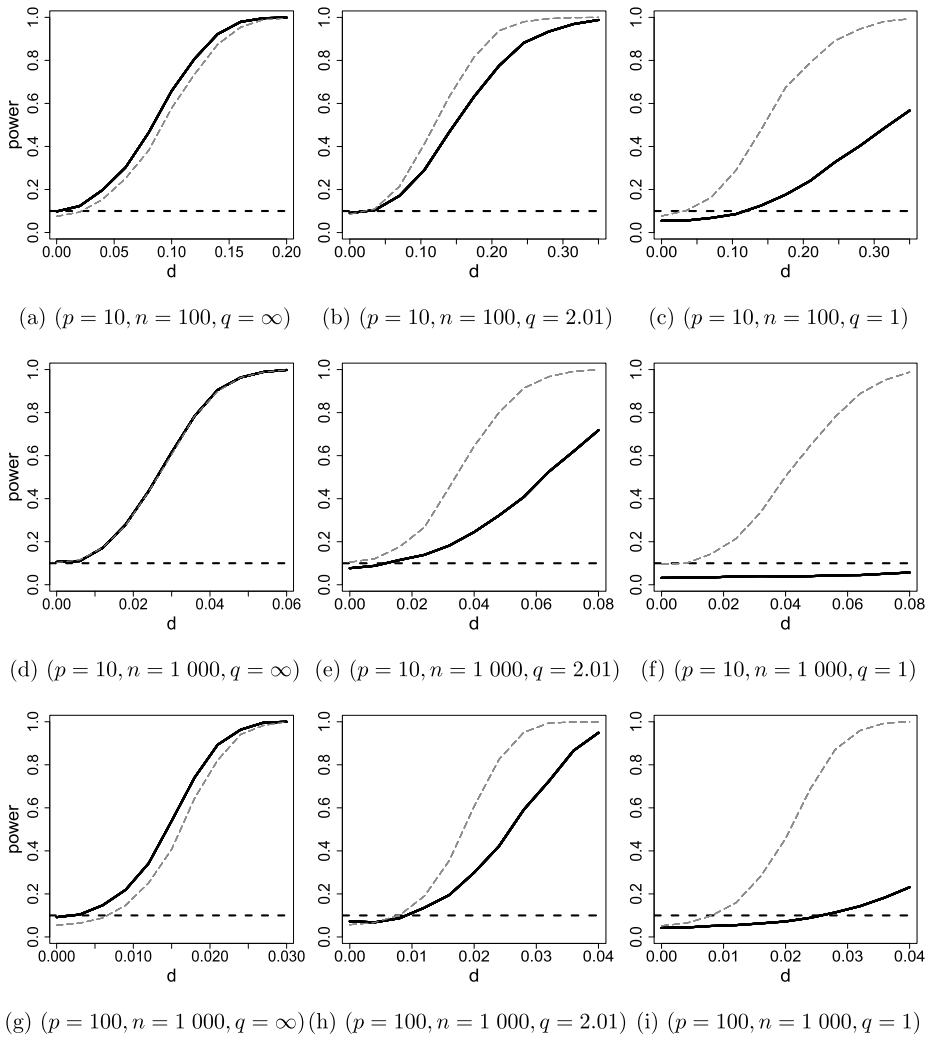
4. Simulation study

The simulation example considers p-dimensional random samples $\mathcal{X}_n$ from (elliptical) canonical t_q distributions with various degrees of freedom q and medians $\boldsymbol{\theta}$. It tests the null hypothesis $H_0 : \boldsymbol{\theta} = \boldsymbol{0}$ against $H_1 : \boldsymbol{\theta} \neq \boldsymbol{0}$ by means of the statistic S of (3) and its asymptotic χ^2_p null distribution (for fixed p and $n \to \infty$). The test, say T_S, uses random selection with replacement to independently choose sets $\mathcal{Q}_C$ (of size $|\mathcal{Q}_C| = 5n$), $\mathcal{Q}_L$ (also of size $|\mathcal{Q}_L| = 5n$) and $\mathcal{S}(\boldsymbol{q}) = \{\boldsymbol{S}_{\boldsymbol{q}}\}$, and it employs van der Waerden's scores.

Fig. 1 here is similar to Figures 1 and 2 of Hudecová et al. (2020) and provides a power comparison between the one-sample test T_S and the benchmark Hotelling T^2 test (as implemented in the *ICSNP* package (Nordhausen et al., 2018) for *R* (R Core Team, 2019)) for a few parameters p, n and q. The empirical power was always computed from $N = 1\,000$ independent replications. Every simulated data set was used for evaluating the tests under both H_0 and all ten shift alternatives considered.

The new test clearly outperforms the benchmark for heavy-tailed distributions and remains competitive even for normally distributed data where it may sometimes lose the comparison due to lower empirical size.

Fig. 2 somewhat analogously applies the same two tests T^2 and T_S to small samples of size $n = 25$ coming from the multivariate normal distribution in $\mathbb{R}^p$, $p = 3, 4, 5$. Then the superiority of T^2 is slightly more evident but, even for such small n, there is no visible difference in the performance of T_S between using the ranks based on randomized or incomplete symmetrized lift-interdirections and between $|\mathcal{Q}_L| = 10n$ and $|\mathcal{Q}_L| = 1000n$. The signs are always estimated by means of incomplete interdirections with $|\mathcal{Q}_C| = 100n$.

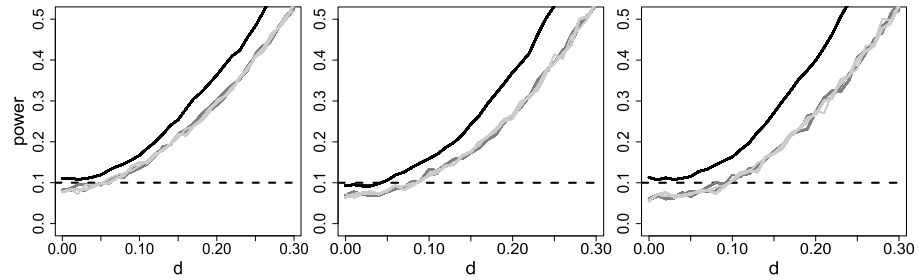


(a) $(p = 10, n = 100, q = \infty)$ (b) $(p = 10, n = 100, q = 2.01)$ (c) $(p = 10, n = 100, q = 1)$

(d) $(p = 10, n = 1\,000, q = \infty)$ (e) $(p = 10, n = 1\,000, q = 2.01)$ (f) $(p = 10, n = 1\,000, q = 1)$

(g) $(p = 100, n = 1\,000, q = \infty)$ (h) $(p = 100, n = 1\,000, q = 2.01)$ (i) $(p = 100, n = 1\,000, q = 1)$

Fig. 1. Power comparison of two one-sample tests at significance level $\alpha = 0.10$. The plots show empirical powers of the benchmark Hotelling test (thick solid black line) and of the signed-rank test T_S (thin gray dashed line) with van der Waerden's scores. The tests were applied to 1 000 p-dimensional random samples of size n from the multivariate canonical Student t distribution with q degrees of freedom shifted by $(d, \dots, d)' \in \mathbb{R}^p$. The parametric combination (p, n, q) characterizing each experiment is stated below individual pictures.



(a) $p = 3$ (b) $p = 4$ (c) $p = 5$

Fig. 2. Small-sample comparison of one-sample tests at significance level $\alpha = 0.10$. The plots compare the benchmark Hotelling test (thick solid black line) and the signed-rank test T_S with the van der Waerden scores in terms of size and power by applying them to 1 000 independent p-dimensional random samples of size $n = 25$ from the multivariate standard normal distribution shifted by $(d, \dots, d)' \in \mathbb{R}^p$. The test T_S uses incomplete interdirections with $|\mathcal{Q}_C| = 100n$ for computing the signs and incomplete symmetrized lift-interdirections (thin) or randomized lift-interdirections (thick) with $|\mathcal{Q}_L| = 10n$ (dark gray) or $|\mathcal{Q}_L| = 1000n$ (light gray) for computing the ranks. All the four power curves for different variants of T_S virtually coincide.

Table 1
The table lists average times (in seconds) needed for ranking n observations (uniformly distributed in $[0,1]^p$) by means of randomized lift-interdirections for various data dimensions p and counts n. The design size $|\mathcal{Q}_L|$ was set to $n/10$ or n.

Average times (in seconds)

n	$p=2$	$p=4$	$p=8$	$p=16$	$p=32$	$p=64$	$p=128$	$p=256$
				$\lvert\mathcal{Q}_L\rvert = 0.1n$				
200	0.005	0.005	0.005	0.005	0.006	0.009	0.025	–
400	0.010	0.010	0.010	0.010	0.012	0.019	0.051	0.276
800	0.021	0.022	0.022	0.024	0.028	0.043	0.110	0.568
1600	0.052	0.055	0.058	0.065	0.072	0.102	0.247	1.205
3200	0.158	0.181	0.231	0.235	0.259	0.330	0.639	2.613
				$\lvert\mathcal{Q}_L\rvert = n$				
200	0.039	0.039	0.041	0.044	0.053	0.081	0.243	–
400	0.085	0.086	0.090	0.097	0.113	0.177	0.509	2.684
800	0.218	0.224	0.220	0.245	0.270	0.403	1.128	5.414
1600	0.642	0.653	0.671	0.703	0.797	1.099	2.437	11.415
3200	1.851	1.888	1.948	2.000	2.232	2.995	5.999	25.368

Table 2
The table lists average times (in seconds) needed for ranking n observations (uniformly distributed in $[0,1]^p$) by means of incomplete symmetrized lift-interdirections for various data dimensions p and counts n. The design size $|\mathcal{Q}_L|$ was set to $n/10$ or n.

Average times (in seconds)

n	$p=2$	$p=4$	$p=8$	$p=2$	$p=4$	$p=8$
		$\lvert\mathcal{Q}_L\rvert = 0.1n$			$\lvert\mathcal{Q}_L\rvert = n$	
25	0.002	0.006	0.088	0.018	0.067	1.116
50	0.004	0.014	0.226	0.034	0.136	2.232
100	0.007	0.028	0.465	0.070	0.278	4.590
200	0.015	0.058	1.046	0.147	0.604	9.726
400	0.033	0.132	2.249	0.359	1.352	22.226
800	0.080	0.386	5.476	0.892	3.340	–

The ranks based on randomized lift-interdirections are interesting not only because of their theoretical properties and the powerful tests they may lead to, but also because of their computational speed, illustrated in Table 1. The table reports average (elapsed) times (based on 1 000 replications) needed for the ranking of uniformly distributed data. Even a naive implementation in R 3.5.3 on a standard notebook (64bit Win 10 with RAM 8 GB and CPU Intel Core i5-8300H 2.3 GHz) leads to small computational times of a few seconds even for $n > 1\,000$ observations, dimensions $p > 100$, and design sizes $|\mathcal{Q}_L| = 0.1n$ or n. As expected, the times show no exponential increase in p and (almost) linear increase in $|\mathcal{Q}_L|$, which can be used for predicting the computational times for other settings.

On the other hand, analogous computational times for the ranks based on incomplete symmetrized lift-interdirections, obtained in the same way and presented in Table 2, are (almost) exponential in p, though still (roughly) linear in $|\mathcal{Q}_L|$. It is because the incomplete symmetrized lift-interdirections use 2^p-times more hyperplanes than the randomized lift-interdirections with the same design size and because the evaluation of hyperplanes for not too small or small-dimensional data sets usually consumes most of the total computational time even in the computation of randomized lift-interdirections.

Consequently, the incomplete symmetrized lift-interdirections could not be computed in a reasonable time at all for many of the combinations (n, p) considered in Table 1. The same holds even more so for the ordinary symmetrized lift-interdirections that typically use even far more hyperplanes than the incomplete ones.

As a real data example, consider the changes in three pulmonary function characteristics (FVC - forced vital capacity, FEV_3 - forced expiratory volume, CC - closing capacity) of $n = 12$ workers after 6 hours of exposure to cotton dust, as recorded in the *pulmonary* data set contained in the *R* package *ICSNP* (Nordhausen et al., 2018).

The data can be expected to follow an elliptical distribution and result in the p-value $p = 0.05123$ of the benchmark Hotelling's test. As the focus is on ranks, let only the ordinary interdirections be used for estimating the signs. Then the van der Waerden version of the one-sample test of no change leads to p-value $p = 0.0388$ with the ranks based on symmetrized lift-interdirections. If the ranks of incomplete symmetrized lift-interdirections ($\widetilde{\underline{R}}_i$'s) or randomized lift-interdirections ($\widehat{\underline{R}}_i$'s) are used instead, $i = 1, \ldots, 12$, then the resulting p-value becomes slightly stochastic because of the random choices involved in ranking the data. Table 3 reports the means and standard deviations of such p-values based on 100 independent random selections of $|\mathcal{Q}_L|$ hyperplanes (and possibly also sign vectors) involved in the computation.

The results make it clear that both the random hyperplanes and random signs introduce some uncertainty to the statistical inference, but its amount does not affect the mean p-value too much, diminishes with growing $|\mathcal{Q}_L|$ and quickly becomes reasonably small. For example, the p-value would be less than 0.05 with probability higher than some 0.85 already for $|\mathcal{Q}_L| = 5n = 60$.

In real-life applications, one would usually compute only one such stochastic p-value and consider data sets large enough to make the asymptotic test reliable.

Table 3
The table lists the means and standard deviations of empirical p-values obtained for 100 independent choices of $|\mathcal{Q}_L|$ hyperplanes (and possibly also sign vectors) from the van der Waerden one-sample test of three-dimensional pulmonary data of size $n = 12$ using the ordinary interdirections and two types of ranks: $\widetilde{R}_i$'s (based on the incomplete symmetrized lift-interdirections) and $\widehat{R}_i$'s (based on the randomized lift-interdirections), $i = 1, \ldots, n$.

The one-sample test of pulmonary data

$\lvert\mathcal{Q}_L\rvert/n$	p-value characteristics			
	mean		standard deviation	
	$\widetilde{R}_i$	$\widehat{R}_i$	$\widetilde{R}_i$	$\widehat{R}_i$
2.5	0.0414	0.0431	0.00655	0.00994
5	0.0395	0.0415	0.00514	0.00756
10	0.0386	0.0403	0.00392	0.00624
20	0.0374	0.0389	0.00284	0.00478
40	0.0372	0.0379	0.00183	0.00287
80	0.0367	0.0369	0.00186	0.00203
160	0.0370	0.0368	0.00186	0.00189

5. Concluding remarks

This article shows that the (highly robust) ranks of randomized lift-interdirections may lead to powerful and computationally feasible sign and rank tests for high-dimensional data and that they may replace other ranks in the canonical multivariate one-sample sign and rank test statistic without changing its limiting distribution and asymptotic behavior under local alternatives. Such rank replacement is likely to be possible even in other sign and rank tests.

The computational benefit achieved in comparison with the use of ordinary or incomplete symmetrized lift-interdirections and their ranks is substantial and indisputable.

Acknowledgements

The research of Šárka Hudecová was supported by the Czech Science Foundation project GA22-01639K. The research of Miroslav Šiman was supported by the Czech Science Foundation project GA21-05325S.

Appendix A. Proofs and auxiliary results

Proof of Proposition 2. As a rule, the expectations will be taken over everything stochastic, including the random design sets and sign vectors used by incomplete interdirections or randomized lift-interdirections. It is highlighted with a subscript to prevent confusion only when it is deemed important for understanding the proof. Everything is greatly simplified by the assumption that both the design sets, the sign vectors and the observations are mutually independent.

The proof of claim (1) derives from the proof of Proposition 2 in Oja and Paindaveine (2005) for ordinary lift-interdirections and closely mimics the proof of Proposition 3.4 of Hudecová et al. (2020) for the incomplete variants. The main difference is that randomized lift-interdirections $\widehat{L}_{\boldsymbol{x}}(\mathcal{X}_n)$ are not U-statistics any more and that the random character of the sign vector must also be taken into consideration. It is stated here in full length for the sake of clarity and completeness.

Fix $\boldsymbol{x} \in \mathbb{R}^p$ arbitrarily. It is sufficient to consider only spherical distribution $\mathcal{E}_{\mathbf{0},\mathrm{I},f}$ without any loss of generality thanks to the affine invariance of $\widehat{L}_{\boldsymbol{x}}(\mathcal{X}_n)$.

Each design set $\mathcal{Q}$ (either $\mathcal{Q}_L$ or $\mathcal{Q}_C$) of fixed size is equiprobable in the set $\{\mathcal{Q}\}$ of all such sets of the same size. This must be considered in the expectations below.

Like in Proposition 2 in Oja and Paindaveine (2005), for fixed design size $m = |\mathcal{Q}_L|$,

$$\mathrm{E}\left(\frac{1}{|\{\mathcal{Q}_L\}|}\widehat{L}_{\boldsymbol{x}}\right) = \frac{1}{2^{pm}|\{\mathcal{Q}_L\}|} \sum_{\substack{\mathcal{Q}_L=\{\boldsymbol{q}_k,k=1,\ldots,m\}\in\{\mathcal{Q}_L\},\\ \boldsymbol{s}(\boldsymbol{q}_k)\in\{-1,1\}^p,\ k=1,\ldots,m}} \mathrm{E}(\widehat{L}_{\boldsymbol{x}}|\boldsymbol{s}(\boldsymbol{q}_1),\ldots\boldsymbol{s}(\boldsymbol{q}_m),\mathcal{Q}_L) = l(\boldsymbol{x})$$

where $l(\boldsymbol{x})$, monotonically increasing in $\|\boldsymbol{x}\|$, is the theoretical lift-interdirection of $\boldsymbol{x}$ from Proposition 1 of Oja and Paindaveine (2005). Similarly,

$$\mathrm{var}\Big(\frac{1}{|\mathcal{Q}_L|}\widehat{L}_{\boldsymbol{x}}\Big) = O\,(n^{-1}) \tag{A.1}$$

because of the relation of $\widehat{L}_{\boldsymbol{x}}$ to $\widetilde{L}_{\boldsymbol{x},-\boldsymbol{x}}/|\mathcal{Q}_L|$ or $\widetilde{L}_{\boldsymbol{x},-\boldsymbol{x}}/|\mathcal{Q}_L|$, which are nothing but incomplete U-statistics with bounded kernel. Consequently, $\widehat{L}_{\boldsymbol{x}}/|\mathcal{Q}_L| \xrightarrow{L_2} l(\boldsymbol{x})$.

If $\boldsymbol{X}_i = \boldsymbol{x}$, then $\widehat{\underline{R}}_i = \sum_{j=1}^n \mathrm{I}[\widehat{\underline{L}}_{\boldsymbol{x}}(\mathcal{X}_n) \geq \widehat{\underline{L}}_{\boldsymbol{X}_j}(\mathcal{X}_n)]$, $R_i^M = \sum_{j=1}^n \mathrm{I}[\|\boldsymbol{x}\| \geq \|\boldsymbol{X}_j\|]$, and

$$\widehat{\underline{R}}_i - R_i^M = \sum_{j=1}^n \left(\mathrm{I}_{[\widehat{\underline{L}}_{\boldsymbol{x}} \geq \widehat{\underline{L}}_{\boldsymbol{X}_j}]} - \mathrm{I}_{[\|\boldsymbol{x}\| \geq \boldsymbol{X}_j\|]}\right) = \sum_{j=1}^n \left(\mathrm{I}_{[\widehat{\underline{L}}_{\boldsymbol{x}} \geq \widehat{\underline{L}}_{\boldsymbol{X}_j}]}\, \mathrm{I}_{[\|\boldsymbol{x}\| < \|\boldsymbol{X}_j\|]} - \mathrm{I}_{[\widehat{\underline{L}}_{\boldsymbol{x}} < \widehat{\underline{L}}_{\boldsymbol{X}_j}]}\, \mathrm{I}_{[\|\boldsymbol{x}\| \geq \|\boldsymbol{X}_j\|]}\right).$$

Consequently, $\mathrm{E}\left[(n+1)^{-2}\left(\widehat{\underline{R}}_i - R_i\right)^2\right]$ is asymptotically negligible for $n \to \infty$ if all the terms

$$\mathrm{E}\left[\mathrm{I}_{[\widehat{\underline{L}}_{\boldsymbol{x}} \geq \widehat{\underline{L}}_{\boldsymbol{X}_j}]}\, \mathrm{I}_{[\|\boldsymbol{x}\| < \|\boldsymbol{X}_j\|]}\, \mathrm{I}_{[\widehat{\underline{L}}_{\boldsymbol{x}} \geq \widehat{\underline{L}}_{\boldsymbol{X}_k}]}\, \mathrm{I}_{[\|\boldsymbol{x}\| < \|\boldsymbol{X}_k\|]}\right],$$
$$\mathrm{E}\left[\mathrm{I}_{[\widehat{\underline{L}}_{\boldsymbol{x}} \geq \widehat{\underline{L}}_{\boldsymbol{X}_j}]}\, \mathrm{I}_{[\|\boldsymbol{x}\| < \|\boldsymbol{X}_j\|]}\, \mathrm{I}_{[\widehat{\underline{L}}_{\boldsymbol{x}} < \widehat{\underline{L}}_{\boldsymbol{X}_k}]}\, \mathrm{I}_{[\|\boldsymbol{x}\| \geq \|\boldsymbol{X}_k\|]}\right], \text{ and}$$
$$\mathrm{E}\left[\mathrm{I}_{[\widehat{\underline{L}}_{\boldsymbol{x}} < \widehat{\underline{L}}_{\boldsymbol{X}_j}]}\, \mathrm{I}_{[\|\boldsymbol{x}\| \geq \|\boldsymbol{X}_j\|]}\, \mathrm{I}_{[\widehat{\underline{L}}_{\boldsymbol{x}} < \widehat{\underline{L}}_{\boldsymbol{X}_k}]}\, \mathrm{I}_{[\|\boldsymbol{x}\| \geq \|\boldsymbol{X}_k\|]}\right],$$

$j, k = 1, \ldots, n$, are $o(1)$, which is in turn implied by the Cauchy-Schwartz inequality if both $\mathrm{E}\, g_1(\boldsymbol{X}_j)$ and $\mathrm{E}\, g_2(\boldsymbol{X}_j)$ are $o(1)$ where

$$g_1(\boldsymbol{X}_j) = \mathrm{I}_{[\widehat{\underline{L}}_{\boldsymbol{x}} \geq \widehat{\underline{L}}_{\boldsymbol{X}_j}]}\, \mathrm{I}_{[\|\boldsymbol{x}\| < \|\boldsymbol{X}_j\|]} \text{ and } g_2(\boldsymbol{X}_j) = \mathrm{I}_{[\widehat{\underline{L}}_{\boldsymbol{x}} < \widehat{\underline{L}}_{\boldsymbol{X}_j}]}\, \mathrm{I}_{[\|\boldsymbol{x}\| \geq \|\boldsymbol{X}_j\|]},$$

$j = 1, \ldots, n$. This will be implied by the dominated convergence theorem (using the boundedness of both g_1 and g_2) from the asymptotic negligibility of conditional expectations $\mathrm{E}\left(g_1(\boldsymbol{X}_j) | \boldsymbol{X}_j = \boldsymbol{z}\right)$ and $\mathrm{E}\left(g_2(\boldsymbol{X}_j) | \boldsymbol{X}_j = \boldsymbol{z}\right)$.

Obviously, $g_1(\boldsymbol{z})$ may be non-zero only for $\|\boldsymbol{z}\| > \|\boldsymbol{x}\|$ when necessarily $l(\boldsymbol{z}) \geq l(\boldsymbol{x})$ due to the monotonicity of l, and the Chebyshev inequality implies

$$\begin{aligned}
\mathrm{E}\left(g_1(\boldsymbol{X}_j) | \boldsymbol{X}_j = \boldsymbol{z}\right) &= \mathrm{P}\left(|\mathcal{Q}_L|^{-1}(\widehat{\underline{L}}_{\boldsymbol{x}} - \widehat{\underline{L}}_{\boldsymbol{z}}) \geq 0\right) \\
&\leq \mathrm{P}\left(\left||\mathcal{Q}_L|^{-1}(\widehat{\underline{L}}_{\boldsymbol{x}} - \widehat{\underline{L}}_{\boldsymbol{z}}) - (l(\boldsymbol{x}) - l(\boldsymbol{z}))\right| \geq l(\boldsymbol{z}) - l(\boldsymbol{x})\right) \\
&\leq \frac{\mathrm{var}\left(|\mathcal{Q}_L|^{-1}(\widehat{\underline{L}}_{\boldsymbol{x}} - \widehat{\underline{L}}_{\boldsymbol{z}})\right)}{(l(\boldsymbol{z}) - l(\boldsymbol{x}))^2} = O(1/n) = o(1)
\end{aligned}$$

because $\mathrm{var}\left(|\mathcal{Q}_L|^{-1}\widehat{\underline{L}}_{\boldsymbol{x}}\right)$ is $O(1/n)$ according to (A.1) and the same analogously holds even for $\mathrm{var}\left(|\mathcal{Q}_L|^{-1}\widehat{\underline{L}}_{\boldsymbol{z}}\right)$.

As for $g_2(\boldsymbol{z})$, it is non-zero only for $\|\boldsymbol{x}\| \geq \|\boldsymbol{z}\|$ when necessarily $l(\boldsymbol{x}) \geq l(\boldsymbol{z})$. However, $\mathrm{P}(\boldsymbol{X}_j = \boldsymbol{x} = \boldsymbol{X}_i) = 0$ and one can thus proceed as for $g_1(\boldsymbol{z})$ almost surely.

Now it is clear that (2) holds conditionally on $\boldsymbol{X}_i = \boldsymbol{x}$ (and on the design). The desired unconditional statement

$$\frac{\widehat{\underline{R}}_i}{n+1} = \frac{R_i^M}{n+1} + o_P(1) \text{ as } n \to \infty$$

follows from integrating the conditional expectation

$$\mathrm{E}\left[\left(\frac{\widehat{\underline{R}}_i}{n+1} - \frac{R_i^M}{n+1}\right)^2 \mid \boldsymbol{X}_{\boldsymbol{i}} = \boldsymbol{x}, \ldots\right]$$

over the stochastic conditions thanks to Lebesgue's dominated convergence theorem.

The proof of claim 2 derives from the proof of Lemma 3 in Hallin and Paindaveine (2002) and closely mimics the proof of Proposition 3.5 of Hudecová et al. (2020) for incomplete interdirections and lift-interdirections. The complex proof is stated here in full detail for the sake of clarity and completeness. One only has to justify some steps differently due to the stochastic nature of the sign vector appearing in the definition of randomized interdirections.

Assume again elliptical distribution with $\boldsymbol{\theta} = \boldsymbol{0}$ and $\boldsymbol{\Sigma} = \mathrm{I}_p$ without any loss of generality. Define distances $d_i = \|\boldsymbol{X}_i\|$ with ranks R_i^M and consider independent random vectors $\boldsymbol{U}_i = d_i^{-1}\boldsymbol{X}_i$ uniformly distributed on the unit sphere in $\mathbb{R}^p$, $i = 1, \ldots, n$. Basically, the proof approximates S of (3) with

$$S^0 = \frac{p}{n\,\mathrm{E}\,K^2(V)} \sum_{i,j=1}^n K(\tilde{F}_f(d_i))K(\tilde{F}_f(d_j))\boldsymbol{U}_i^\top \boldsymbol{U}_j$$

whose limit distribution follows easily from the central limit theorem for

$$\mathbf{T}_n = \frac{1}{\sqrt{n}} \sum_{i=1}^n K(\tilde{F}_f(d_i))\boldsymbol{U}_i.$$

The difference can be written as $S - S^0 = p\{\mathrm{E}[K^2(V)]\}^{-1}(T_1^n + T_2^n)$ where

$$T_1^n = \frac{1}{n}\sum_{i,j=1}^{n} D_i D_j G_{ij} \text{ and}$$

$$T_2^n = \frac{1}{n}\sum_{i,j=1}^{n} \left[D_i D_j - K(\tilde{F}_f(d_i))K(\tilde{F}_f(d_j))\right] \boldsymbol{U}_i^\top \boldsymbol{U}_j$$

for $D_i := K\left(\underline{R}_i/(n+1)\right)$ and $G_{ij} := \cos(a_{\boldsymbol{X}_i,\boldsymbol{X}_j}) - \boldsymbol{U}_i^\top \boldsymbol{U}_j$, $i, j = 1, \ldots, n$.

Evidently, $T_2^n = \widehat{\mathbf{S}}_n^\top \widehat{\mathbf{S}}_n - \mathbf{T}_n^\top \mathbf{T}_n = (\widehat{\mathbf{S}}_n + \mathbf{T}_n)^\top (\widehat{\mathbf{S}}_n - \mathbf{T}_n)$ for

$$\widehat{\mathbf{S}}_n = \frac{1}{\sqrt{n}}\sum_{i=1}^{n} D_i \boldsymbol{U}_i \text{ and } \mathbf{T}_n = \frac{1}{\sqrt{n}}\sum_{i=1}^{n} K(\tilde{F}_f(d_i))\boldsymbol{U}_i.$$

It turns out that both $\mathbf{T}_n$ and $\widehat{\mathbf{S}}_n$ may be approximated with a simpler statistic

$$\mathbf{S}_n = \frac{1}{\sqrt{n}}\sum_{i=1}^{n} K(R_i^M/(n+1))\boldsymbol{U}_i,$$

which will be exploited advantageously. The goal is to show both $T_1^n = o_P(1)$ and $T_2^n = o_P(1)$ for $n \to \infty$, which then implies $S - S^0 = o_P(1)$.

Some partial results can easily be adopted from the aforementioned proofs because $\mathbf{T}_n$ and $\mathbf{S}_n$ are the same as in Hallin and Paindaveine (2002), and because G_{ij} is the same as in Hudecová et al. (2020). In particular, the proof in Hallin and Paindaveine (2002) establishes $\mathrm{E}\,\|\mathbf{T}_n - \mathbf{S}_n\|^2 = o(1)$ and $\mathrm{E}[K(R_i^M/(n+1)) - K(\tilde{F}_f(d_i))]^2 = o(1)$ as $n \to \infty$, $i = 1, \ldots, n$, and the proof in Hudecová et al. (2020) provides $\mathrm{E}\,|G_{ij}|^{2(2+\delta)/\delta} \to 0$ for $n \to \infty$ for any $\delta > 0$ and $i \neq j$. Further technicalities necessary for the proof are collected in subsequent Lemma 1.

The assumption on K implies, for $n \to \infty$, $\mathrm{E}\,\|\mathbf{T}_n\|^2 \to \mathrm{E}\,K^2(V) < \infty$ as well as

$$\mathrm{E}\,\|\mathbf{S}_n - \widehat{\mathbf{S}}_n\|^2 = \frac{1}{n}\sum_{i=1}^{n} \mathrm{E}\left[K\left(R_i^M/(n+1)\right) - D_i\right]^2 \to 0 \tag{A.2}$$

where the first equality follows from Lemma 1(1). The squared differences in the summands are then both uniformly integrable and $o(1)$ in probability (thanks to (2) and K continuous almost everywhere), therefore $o(1)$ in quadratic mean.

Consequently, $\mathrm{E}\,\|\mathbf{T}_n - \widehat{\mathbf{S}}_n\|^2 \le \mathrm{E}\,\|\mathbf{T}_n - \mathbf{S}_n\|^2 + \mathrm{E}\,\|\mathbf{S}_n - \widehat{\mathbf{S}}_n\|^2 \to 0$ for $n \to \infty$, and $\mathrm{E}\,\|\widehat{\mathbf{S}}_n + \mathbf{T}_n\|^2$ is both bounded and finite. As a result,

$$\mathrm{E}\,|T_2^n| \le \sqrt{\mathrm{E}\,\|\widehat{\mathbf{S}}_n + \mathbf{T}_n\|^2}\sqrt{\mathrm{E}\,\|\widehat{\mathbf{S}}_n - \mathbf{T}_n\|^2} \to 0$$

for $n \to \infty$ owing to the Cauchy-Schwartz inequality. Therefore, $T_2^n \to 0$ with $n \to \infty$ both in L^1 and in probability.

As for T_1^n,

$$\begin{aligned}\mathrm{E}\,\|T_1^n\|^2 &= \frac{1}{n^2}\mathrm{E}\left(\sum_{i,j=1}^{n} D_i D_j G_{ij}\right)^2 = \frac{2}{n^2}\sum_{i\neq j}\mathrm{E}[D_i D_j G_{ij}]^2 = \frac{2(n-1)}{n}\mathrm{E}[D_1 D_2 G_{12}]^2 \\ &\le \frac{2(n-1)}{n}(\mathrm{E}\,|D_1 D_2|^{2+\delta})^{2/(2+\delta)}(\mathrm{E}\,|G_{12}|^{2(2+\delta)/\delta})^{\delta/(2+\delta)}\end{aligned}$$

for the particular $\delta > 0$ from the assumption. It is mainly because of Lemma 1(2), Lemma 1(3), $G_{ii} = 0$, $i = 1, \ldots, n$, and the Hölder inequality. As $\mathrm{E}\,|D_1 D_2|^{2+\delta} < \infty$ due to Lemma 1(2), $T_1^n \to 0$ in L^2 (and in probability) for $n \to \infty$. □

Lemma 1.

(1) If $i \neq j$, $i, j = 1, \ldots, n$, then

$$\mathrm{E}\left[K\left(R_i^M/(n+1)\right) - D_i\right]\left[K\left(R_j^M/(n+1)\right) - D_j\right]\boldsymbol{U}_i^\top \boldsymbol{U}_j = 0.$$

(2) $\mathrm{E}\,|D_1 D_2|^{2+\delta} < \infty$ and $\mathrm{E}\,D_i D_j G_{ij} = \mathrm{E}\,D_1 D_2 G_{12}$ for $i \neq j$.

(3) If $(i, j) \neq (k, l)$ and $(i, j) \neq (l, k)$, then

$$\mathrm{E}\,D_i D_j G_{ij} D_k D_l G_{kl} = 0.$$

Proof. Set $B_i := \operatorname{sgn}(X_{i1})$, $\mathbf{Y}_i := B_i \boldsymbol{X}_i$, and note that B_i are (for spherically distributed $\boldsymbol{X}_i = (X_{i1}, \ldots, X_{ip})'$) independent of $\mathbf{Y}_i$, mutually independent and identically distributed with $\mathrm{P}(B_i = 1) = \mathrm{P}(B_i = -1) = 1/2$, $i = 1, \ldots, n$. Then $\boldsymbol{X}_i = d_i \boldsymbol{U}_i$ and $\boldsymbol{U}_i = \boldsymbol{X}_i/d_i = \mathbf{Y}_i B_i/d_i$ where $d_i = \|\boldsymbol{X}_i\| = \|\mathbf{Y}_i\|$. Consequently, both d_i's and their ranks R_i^M's are functions of the sample $\mathcal{Y}_n$ consisting of $\mathbf{Y}_1, \ldots, \mathbf{Y}_n$. Analogously, $\mathcal{B}_n$ will denote the sample of $B_1, \ldots, B_n$.
(1) Let $i \neq j$. The statement holds when

$$\begin{aligned}
&\mathrm{E}\, K(R_i^M/(n+1))K(R_j^M/(n+1))\boldsymbol{U}_i^\top \boldsymbol{U}_j = 0,\\
&\mathrm{E}\, K(\underline{\widehat{R}}_i/(n+1))K(\underline{\widehat{R}}_j/(n+1))\boldsymbol{U}_i^\top \boldsymbol{U}_j = 0, \text{ and}\\
&\mathrm{E}\, K(\underline{\widehat{R}}_i/(n+1))K(R_j^M/(n+1))\boldsymbol{U}_i^\top \boldsymbol{U}_j = 0.
\end{aligned}$$

The first equality has been dealt with in Hallin and Paindaveine (2002). It follows from

$$\begin{aligned}
&\mathrm{E}K(R_i^M/(n+1))K(R_j^M/(n+1))\boldsymbol{U}_i^\top \boldsymbol{U}_j\\
&= \mathrm{E}_{\mathcal{Y}_n}\, \mathrm{E}_{B_n}[K(R_i^M/(n+1))K(R_j^M/(n+1))\boldsymbol{U}_i^\top \boldsymbol{U}_j | \mathcal{Y}_n, \mathcal{B}_n]\\
&= \mathrm{E}_{\mathcal{Y}_n} \frac{\mathbf{Y}_i^\top \mathbf{Y}_j}{d_i d_j} K(R_i^M/(n+1))K(R_j^M/(n+1))\, \mathrm{E}_{B_n}[B_i B_j | \mathcal{Y}_n, \mathcal{B}_n] = 0.
\end{aligned}$$

Furthermore, for $|\mathcal{Q}_L| = m$,

$$\begin{aligned}
&\mathrm{E}_{\mathcal{Q}_L=\{\boldsymbol{q}_1,\ldots\boldsymbol{q}_m\},\boldsymbol{s}(\boldsymbol{q}_1),\ldots,\boldsymbol{s}(\boldsymbol{q}_m)}[K(\underline{\widehat{R}}_i/(n+1))K(\underline{\widehat{R}}_j/(n+1))|\mathcal{Y}_n, \mathcal{B}_n]\\
&\qquad = \frac{1}{2^{pm}|\{\mathcal{Q}_L\}|} \sum_{\substack{\mathcal{Q}_L=\{\boldsymbol{q}_k, k=1,\ldots,m\}\in\{\mathcal{Q}_L\},\\ \boldsymbol{s}(\boldsymbol{q}_k)\in\{-1,1\}^p,\ k=1,\ldots,m}} K(\underline{\widehat{R}}_i/(n+1))K(\underline{\widehat{R}}_j/(n+1)) \qquad \text{(A.3)}
\end{aligned}$$

is the same for any $u_1\boldsymbol{X}_1, \ldots, u_n\boldsymbol{X}_n$, $(u_1, \ldots, u_n) \in \{-1, 1\}^n$. Therefore, it is only a function of $\mathcal{Y}_n$, say $\rho(\mathcal{Y}_n)$, and

$$\mathrm{E}\, K(\underline{\widehat{R}}_i/(n+1))K(\underline{\widehat{R}}_j/(n+1))\boldsymbol{U}_i^\top \boldsymbol{U}_j = \mathrm{E}_{\mathcal{Y}_n} \frac{\mathbf{Y}_i^\top \mathbf{Y}_j}{d_i d_j} \rho(\mathcal{Y}_n)\, \mathrm{E}_{\mathcal{B}_n}[B_i B_j | \mathcal{Y}_n, \mathcal{B}_n] = 0.$$

The third equality could be proved analogously.
(2) Obviously, for $|\mathcal{Q}_L| = m$,

$$\begin{aligned}
\mathrm{E}\, D_i D_j G_{ij} = \frac{1}{2^{pm}|\{\mathcal{Q}_C\}||\{\mathcal{Q}_L\}|} &\sum_{\mathcal{Q}_C\in\{\mathcal{Q}_C\}} \sum_{\substack{\mathcal{Q}_L=\{\boldsymbol{q}_k, k=1,\ldots,m\}\in\{\mathcal{Q}_L\},\\ \boldsymbol{s}(\boldsymbol{q}_k)\in\{-1,1\}^p,\ k=1,\ldots,m}} \mathrm{E}_{\mathcal{X}_n} K\left(\frac{\widehat{R}_i}{n+1}\right) K\left(\frac{\widehat{R}_j}{n+1}\right)\\
&\times \left(\cos(a_{\boldsymbol{X}_i,\boldsymbol{X}_j}) - \frac{\boldsymbol{X}_i^\top \boldsymbol{X}_j}{d_i d_j}\right).
\end{aligned}$$

The expectation is the same for all pairs (i, j) of distinct indices because the $\boldsymbol{X}_i$'s are i.i.d. and the design sets and sign vectors are properly independent and equiprobable. The following expectation is independent of (i, j), $i \neq j$, for the same reason:

$$\begin{aligned}
\mathrm{E}|D_i D_j|^{2+\delta} &= \frac{1}{2^{pm}|\{\mathcal{Q}_L\}|} \sum_{\substack{\mathcal{Q}_L=\{\boldsymbol{q}_k, k=1,\ldots,m\}\in\{\mathcal{Q}_L\},\\ \boldsymbol{s}(\boldsymbol{q}_k)\in\{-1,1\}^p,\ k=1,\ldots,m}} \mathrm{E}_{\mathcal{X}_n} \left| K\left(\frac{\widehat{R}_i}{n+1}\right) K\left(\frac{\widehat{R}_j}{n+1}\right)\right|^{2+\delta}\\
&= \sum_{l_1\neq l_2} |K(l_1/(n+1))|^{2+\delta} |K(l_2/(n+1))|^{2+\delta} \underbrace{\frac{1}{2^{pm}|\{\mathcal{Q}_L\}|} \sum_{\substack{\mathcal{Q}=\{\boldsymbol{q}_k,\ k=1,\ldots,m\}\in\{\mathcal{Q}_L\},\\ \boldsymbol{s}(\boldsymbol{q}_k)\in\{-1,1\}^p, k=1,\ldots,m}} \mathrm{P}(\underline{\widehat{R}}_1 = l_1, \underline{\widehat{R}}_2 = l_2)}_{p(l_1,l_2)}.
\end{aligned}$$

As $p(l_1, l_2)$ does not depend on l_1 and l_2, necessarily $p(l_1, l_2) = [n(n-1)]^{-1}$. The finiteness of $\mathrm{E}\,|D_1 D_2|^{2+\delta}$ then follows from the arguments based on the Riemann sums of (Hallin and Paindaveine, 2002, Proof of Proposition 3).
(3) Realize that $G_{ij}^{\mathcal{Y}_n} := \cos(\pi \widetilde{C}_{\mathbf{Y}_i,\mathbf{Y}_j}(\mathcal{Y}_n)/|\mathcal{Q}_C|) - \mathbf{Y}_i^\top \mathbf{Y}_j/(d_i d_j)$ and thus $G_{ij}^{\mathcal{Y}_n} = B_i B_j G_{ij}^{\mathcal{X}_n}$. Analogously as in (A.3), it can be shown that $\mathrm{E}_{\mathcal{Q}_C}\, \mathrm{E}_{\mathcal{Q}_L=\{\boldsymbol{q}_1,\ldots\boldsymbol{q}_m\},\boldsymbol{s}(\boldsymbol{q}_1),\ldots,\boldsymbol{s}(\boldsymbol{q}_m)}[D_i D_j G_{ij}^{\mathcal{Y}_n} D_k D_l G_{kl}^{\mathcal{Y}_n} | \mathcal{Y}_n, \mathcal{B}_n]$ is only a function of $\mathcal{Y}_n$, say $\rho(\mathcal{Y}_n)$, and

$$\mathrm{E}\, D_i D_j G_{ij} D_k D_l G_{kl} = \mathrm{E}_{\mathcal{Y}_n}\, \rho(\mathcal{Y}_n)\, \mathrm{E}_{\mathcal{B}_n}[B_i B_j B_k B_l | \mathcal{Y}_n, \mathcal{B}_n],$$

which is indeed zero if (i, j) is different from both (k, l) and (l, k). □

References

Chaudhuri, P., Sengupta, D., 1993. Sign tests in multidimension: inference based on the geometry of the data cloud. J. Am. Stat. Assoc. 88, 1363–1370.
Chernozhukov, V., Galichon, A., Hallin, M., Henry, M., 2017. Monge-Kantorovich depth, quantiles, ranks, and signs. Ann. Stat. 45, 223–256.
Hallin, M., Del Barrio, E., Cuesta-Albertos, J., Matrán, C., 2021. Distribution and quantile functions, ranks and signs in dimension d: a measure transportation approach. Ann. Stat. 49, 1139–1165.
Hallin, M., Paindaveine, D., 2002. Optimal tests for multivariate location based on interdirections and pseudo-Mahalanobis ranks. Ann. Stat. 30, 1103–1133.
Hettmansperger, T., Möttönen, J., Oja, H., 1999. The geometry of the affine invariant multivariate sign and rank methods. J. Nonparametr. Stat. 11, 271–285.
Hudecová, Š., Klicnarová, J., Šiman, M., 2020. Incomplete interdirections and lift-interdirections. J. Nonparametr. Stat. 32, 93–108.
Lehman, E., Romano, J., 2005. Testing Statistical Hypotheses, 3rd edition. Springer, New York.
Möttönen, J., Oja, H., 1995. Multivariate spatial sign and rank methods. J. Nonparametr. Stat. 5, 201–213.
Nordhausen, K., Sirkia, S., Oja, H., Tyler, D.E., 2018. ICSNP: Tools for Multivariate Nonparametrics, R package version 1.1-1.
Oja, H., 1999. Affine invariant multivariate sign and rank tests and corresponding estimates: a review. Scand. J. Stat. 26, 319–343.
Oja, H., 2005. Multivariate Nonparametric Methods with R: An Approach Based on Spatial Signs and Ranks. Springer, New York.
Oja, H., Paindaveine, D., 2005. Optimal signed-rank tests based on hyperplanes. J. Stat. Plan. Inference 135, 300–323.
Puri, M., Sen, P., 1971. Nonparametric Methods in Multivariate Analysis. Wiley.
R Core Team, 2019. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria.
Randles, R., 1989. A distribution-free multivariate sign test based on interdirections. J. Am. Stat. Assoc. 84, 1045–1050.
Zuo, Y., Serfling, R., 2000. General notions of statistical depth functions. Ann. Stat. 28, 461–482.