Taylor & Francis
Taylor & Francis Group

Journal of Nonparametric Statistics

ISSN: 1048-5252 (Print) 1029-0311 (Online) Journal homepage: https://www.tandfonline.com/loi/gnst20

Incomplete interdirections and lift-interdirections

Šárka Hudecová, Jana Klicnarová & Miroslav Šiman

To cite this article: Šárka Hudecová, Jana Klicnarová & Miroslav Šiman (2019): Incomplete interdirections and lift-interdirections, Journal of Nonparametric Statistics, DOI: 10.1080/10485252.2019.1700255

To link to this article: https://doi.org/10.1080/10485252.2019.1700255

JOURNAL OF NONPARAMETRIC STATISTICS
https://doi.org/10.1080/10485252.2019.1700255

Taylor & Francis
Taylor & Francis Group

Incomplete interdirections and lift-interdirections

Šárka Hudecová[a], Jana Klicnarová[b,c] and Miroslav Šiman[b]

[a]Department of Probability and Statistics, Faculty of Mathematics and Physics, Charles University, Praha 8, Czech Republic; [b]The Czech Academy of Sciences, Institute of Information Theory and Automation, Praha 8, Czech Republic; [c]Faculty of Economics, University of South Bohemia in České Budějovice, České Budějovice, Czech Republic

ABSTRACT

The article presents, discusses, and explores incomplete variants of interdirections, lift-interdirections, and symmetrised lift-interdirections (with a few incomplete designs). Although they are easier to compute in high-dimensional spaces than the originals, they can still replace them in many optimal statistical procedures based on signs and ranks without significantly changing their properties. This is proved theoretically and confirmed empirically in a small simulation study dealing with the canonical examples of multivariate sign and signed-rank one-sample tests applied to high-dimensional data sets.

ARTICLE HISTORY
Received 16 May 2019
Accepted 28 November 2019

KEYWORDS
Interdirection; multivariate rank; multivariate sign; signed-rank test; robustness

1. Introduction

Simple, robust, and powerful univariate sign and rank statistical procedures with weak assumptions have already inspired many multivariate generalisations. Current literature on multivariate rank statistics is largely dominated by spatial ranks and signs (Möttönen and Oja 1995; Oja 2010), followed by component-wise rank and sign vectors (Puri and Sen 1971), Oja multivariate ranks and signs (Oja 1999), the ranks of pseudo-Mahalanobis distances (Hallin and Paindaveine 2002b), and various depth concepts (Zuo and Serfling 2000); see also Chaudhuri and Sengupta (1993) and Chernozhukov, Galichon, Hallin, and Henry (2017). Unfortunately, the signs and ranks based on the purely geometric hyperplane-based concepts of interdirections (Randles 1989) and lift-interdirections (Hettmansperger, Möttönen, and Oja 1999; Oja and Paindaveine 2005) seem to fall out of fashion due to their high computational demands, even though they avoid any estimation of the shape matrix.

This article can be viewed as an attempt to revive these two appealing concepts and to bring them back to the forefront. In particular, it presents incomplete interdirections, incomplete lift-interdirections and incomplete symmetrised lift-interdirections based on incomplete U-statistics. They allow for various random and deterministic designs, they are relatively quick to compute in high-dimensional spaces, and they can replace their

CONTACT Šárka Hudecová hudecova@karlin.mff.cuni.cz Department of Probability and Statistics, Faculty of Mathematics and Physics, Charles University, Sokolovská 83, 186 75 Praha 8, Czech Republic

complete counterparts in many sign and signed-rank test statistics without changing their asymptotic behaviour.

In other words, this article addresses the problem of too many hyperplanes passing through certain number of observations by working only with some of them. Such an approach is hardly surprising, and it was already suggested in the concluding section of Oja and Paindaveine (2005). This article shows that there is a clever way to do the selection, analyses it, and illustrates it with multidimensional data.

The signs and ranks based on incomplete interdirections and incomplete (symmetrised) lift-interdirections might still be used in the large number of statistical tests mentioned in Oja and Paindaveine (2005), including the test of randomness against VARMA dependence, the two-sample and multi-sample tests, the Durbin-Watson test and the test of the order of a VARMA model. In particular, the incomplete interdirections might be used in any sign test where the original interdirections appear; see, e.g. Randles (1989), Randles and Peters (1990), Jan and Randles (1996), Gieser and Randles (1997), Um and Randles (1998), Ghoush and Sengupta (2001), Hallin and Paindaveine (2002a, 2002b, 2004), Taskinen, Oja, and Randles (2005), and Paindaveine (2009). Here, it is proved rigorously only for the one-sample tests of Randles (1989) and Oja and Paindaveine (2005). Note also that all the resulting tests can be expected to be robust with respect to both radial and angular outliers like other hyperplane-based testing procedures (Oja and Paindaveine 2005).

Multivariate rank tests often need some symmetry to work. The sign tests discussed here work for all distributions with elliptical directions (Randles 1989) while the signed-rank tests require elliptical symmetry as those in Oja and Paindaveine (2005); see Serfling (2006) for a review of the most important concepts of multivariate symmetry. These assumptions permit even some distributions that are not unimodal. Such distributions may arise easily in the context of mixtures even in the univariate case; see, e.g. Došlá (2009).

Next Section 2 introduces incomplete interdirections, lift-interdirections, and their symmetric variant. Section 3 then describes their properties and justifies their use in the one-sample tests of Randles (1989) and Oja and Paindaveine (2005), and Section 4 employs them in a brief comparative simulation study using also large sample sizes and dimensions. Section 5 discusses the results and achievements. Final Appendix collects the proofs.

2. Definitions and notation

Let $\mathcal{X}_n$ be a random sample consisting of n p-dimensional observations $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n \in \mathbb{R}^p$, $p \geq 2$. Any ordered k-tuple $\boldsymbol{q}(n,k) = (q_1, \ldots, q_k)$ of distinct integer indices $1 \leq q_1 < q_2 < \cdots < q_k \leq n$ then defines subsample $\mathcal{X}^{\boldsymbol{q}(n,k)} = (\boldsymbol{X}_{q_1}, \ldots, \boldsymbol{X}_{q_k})$. The set of all possible k-tuples $\boldsymbol{q}(n,k)$ will be denoted by $\mathcal{Q}(n,k)$. It has exactly $\binom{n}{k}$ elements.

If $k = p$, then $H^{\boldsymbol{q}(n,p)} \subset \mathbb{R}^p$ is defined as the hyperplane that contains all the observations from $\mathcal{X}^{\boldsymbol{q}(n,p)}$. If $k = p-1$, then $H^{\boldsymbol{q}(n,p-1)} \subset \mathbb{R}^p$ is defined as the hyperplane containing all the observations from $\mathcal{X}^{\boldsymbol{q}(n,p-1)}$ and the origin.

Any hyperplane $H^{\boldsymbol{q}(n,p-1)}$ can be expressed by the equation

$$\det\left(\mathbb{M}^{\boldsymbol{q}(n,p-1)}(\boldsymbol{y})\right) = \boldsymbol{d}^{\boldsymbol{q}(n,p-1)'}\boldsymbol{y} = 0$$

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

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

Similarly, any hyperplane $H^{\boldsymbol{q}(n,p)}$ can be expressed by the equation

$$\det\left(\mathbb{M}^{\boldsymbol{q}(n,p)}(\boldsymbol{y})\right) = d_0^{\boldsymbol{q}(n,p)} + \boldsymbol{d}^{\boldsymbol{q}(n,p)'}\boldsymbol{y} = 0$$

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

$$\mathbb{M}^{\boldsymbol{q}(n,p)}(\boldsymbol{y}) = \begin{pmatrix} 1 & 1 & \cdots & 1 & 1 \\ \boldsymbol{X}_{q_1} & \boldsymbol{X}_{q_2} & \cdots & \boldsymbol{X}_{q_p} & \boldsymbol{y} \end{pmatrix}.$$

The signs

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

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

If $\mathcal{X}_n$ comes from the spherically symmetric distribution centred around the origin, then the angular distance

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

between any two points $\boldsymbol{y}_1, \boldsymbol{y}_2 \in \mathbb{R}^p$ can be measured by means of (affine invariant) interdirections

$$C_{\boldsymbol{y}_1,\boldsymbol{y}_2}(\mathcal{X}_n) = \sum_{\boldsymbol{q}\in\mathcal{Q}(n,p-1)} \left(1 - S^{\boldsymbol{q}}(\boldsymbol{y}_1)S^{\boldsymbol{q}}(\boldsymbol{y}_2)\right)/2;$$

see Randles (1989). (They are defined by means of all hyperplanes passing through the origin and $p-1$ observations, but only those hyperplanes separating $\boldsymbol{y}_1$ from $\boldsymbol{y}_2$ are actually counted.) That is to say that

$$a_{\boldsymbol{y}_1,\boldsymbol{y}_2} = \pi\, C_{\boldsymbol{y}_1,\boldsymbol{y}_2}(\mathcal{X}_n) \Big/ \binom{n}{p-1}$$

is then an affine invariant consistent estimator of $\alpha(\boldsymbol{y}_1, \boldsymbol{y}_2)$. Needless to say that the angles and their estimators play a crucial role in various multivariate sign and signed-rank tests.

Oja and Paindaveine (2005) suggest to measure the empirical distance between two points $\boldsymbol{y}_1$ and $\boldsymbol{y}_2$ by means of (affine invariant) lift-interdirection

$$L_{\boldsymbol{y}_1,\boldsymbol{y}_2}(\mathcal{X}_n) = \sum_{\boldsymbol{q}\in\mathcal{Q}(n,p)} \left(1 - S^{\boldsymbol{q}}(\boldsymbol{y}_1)S^{\boldsymbol{q}}(\boldsymbol{y}_2)\right)/2$$

that counts all hyperplanes passing through p observations and separating $\boldsymbol{y}_1$ from $\boldsymbol{y}_2$. They consider especially the particular case $L_{\boldsymbol{y}}(\mathcal{X}_n) = L_{\boldsymbol{y},-\boldsymbol{y}}(\mathcal{X}_n)$, $\boldsymbol{y} \in \mathbb{R}^p$, and its symmetrised

version

$$\underline{L}_{\boldsymbol{y}}(\mathcal{X}_n) = \sum_{\substack{\boldsymbol{q}\in\mathcal{Q}(n,p)\\ \boldsymbol{s}\in\{-1,1\}^p}}\sum \frac{1-\operatorname{sign}(d_{0\boldsymbol{s}}^{\boldsymbol{q}(n,p)} + \boldsymbol{d}_{\boldsymbol{s}}^{\boldsymbol{q}(n,p)'}\boldsymbol{y})\operatorname{sign}(d_{0\boldsymbol{s}}^{\boldsymbol{q}(n,p)} - \boldsymbol{d}_{\boldsymbol{s}}^{\boldsymbol{q}(n,p)'}\boldsymbol{y})}{2}$$

where $(d_{0\boldsymbol{s}}^{\boldsymbol{q}(n,p)}, \boldsymbol{d}_{\boldsymbol{s}}^{\boldsymbol{q}(n,p)'})'$ is the vector of cofactors of the last column of matrix

$$\mathbb{M}_{\boldsymbol{s}}^{\boldsymbol{q}(n,p)}(\boldsymbol{y}) = \begin{pmatrix} 1 & 1 & \cdots & 1 & 1 \\ s_1\boldsymbol{X}_{q_1} & s_2\boldsymbol{X}_{q_2} & \cdots & s_p\boldsymbol{X}_{q_p} & \boldsymbol{y} \end{pmatrix},$$

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. That is to say that the symmetrised lift-interdirections $\underline{L}_{\boldsymbol{y}}(\mathcal{X}_n)$ are invariant with respect to affine data transformations, permutations of the observations, and reflections of the observations with respect to the origin; see Oja and Paindaveine (2005). The ranks of $\underline{L}_{\boldsymbol{y}}(\mathcal{X}_n)$ can thus replace the ranks of pseudo-Mahalanobis distances in many tests based on them; see ibidem.

The cardinality of $\mathcal{Q}(n,p-1)$ may quickly become impractical for the computation of $a_{\boldsymbol{y}_1,\boldsymbol{y}_2}$ with growing n and p, which is why it seems advantageous to define *incomplete interdirections* as incomplete U-statistics (see, e.g. Lee 1990)

$$\widetilde{C}_{\boldsymbol{y}_1,\boldsymbol{y}_2}(\mathcal{X}_n) = \sum_{\boldsymbol{q}\in\mathcal{Q}_S(n,p-1)} \left(1 - S^{\boldsymbol{q}}(\boldsymbol{y}_1)S^{\boldsymbol{q}}(\boldsymbol{y}_2)\right)/2$$

considering only some hyperplanes determined by some (design) $\mathcal{Q}_S(n,p-1) \subset \mathcal{Q}(n,p-1)$ with $m_{n,p-1}$ elements.

One can also analogously define *incomplete lift-interdirections* $\widetilde{L}_{\boldsymbol{y}_1,\boldsymbol{y}_2}(\mathcal{X}_n)$

$$\widetilde{L}_{\boldsymbol{y}_1,\boldsymbol{y}_2}(\mathcal{X}_n) = \sum_{\boldsymbol{q}\in\mathcal{Q}_S(n,p)} \left(1 - S^{\boldsymbol{q}}(\boldsymbol{y}_1)S^{\boldsymbol{q}}(\boldsymbol{y}_2)\right)/2$$

for some (design) $\mathcal{Q}_S(n,p) \subset \mathcal{Q}(n,p)$ with $m_{n,p}$ elements, and *incomplete symmetrised lift-interdirections*:

$$\underline{\widetilde{L}}_{\boldsymbol{y}}(\mathcal{X}_n) = \sum_{\substack{\boldsymbol{q}\in\mathcal{Q}_S^s(n,p)\\ \boldsymbol{s}\in\{-1,1\}^p}}\sum \left(1-\operatorname{sign}(d_{0\boldsymbol{s}}^{\boldsymbol{q}(n,p)} + \boldsymbol{d}_{\boldsymbol{s}}^{\boldsymbol{q}(n,p)'}\boldsymbol{y})\operatorname{sign}(d_{0\boldsymbol{s}}^{\boldsymbol{q}(n,p)} - \boldsymbol{d}_{\boldsymbol{s}}^{\boldsymbol{q}(n,p)'}\boldsymbol{y})\right)/2$$

for some (design) $\mathcal{Q}_S^s(n,p)$ of elements from $\mathcal{Q}(n,p)$ with $m_{n,p}^s$ elements.

To be more specific, the designs $\mathcal{Q}_S(n,p-1)$, $\mathcal{Q}_S(n,p)$, and $\mathcal{Q}_S^s(n,p)$ are arbitrary collections of predefined deterministic sizes $m_{n,p-1}$, $m_{n,p}$, and $m_{n,p}^s$ containing (not necessarily distinct) elements from $\mathcal{Q}(n,p-1)$, $\mathcal{Q}(n,p)$, and $\mathcal{Q}(n,p)$ in this order. They are called sets or denoted as subsets by slight but useful abuse of terminology. The subscript S indicates the (design) subsets and the index s indicates a relation to the symmetrised lift-interdirections.

These considerations lead to a few meaningful rank concepts in the multivariate space because one can easily write R_i, $\widetilde{R}_i$, and $\underline{\widetilde{R}}_i$ for the ranks of $\boldsymbol{X}_i'\boldsymbol{\Sigma}^{-1}\boldsymbol{X}_i$, $\widetilde{L}_{\boldsymbol{X}_i}(\mathcal{X}_n)$, and $\underline{\widetilde{L}}_{\boldsymbol{X}_i}(\mathcal{X}_n)$ in the corresponding samples of the same quantities for $i = 1,\ldots,n$. Of course, these ranks would be meaningful only for centred observations.

3. Theoretical considerations

Fortunately, both $\widetilde{C}_{\boldsymbol{y}_1,\boldsymbol{y}_2}(\mathcal{X}_n)/m_{n,p-1}$ and $\widetilde{L}_{\boldsymbol{y}_1,\boldsymbol{y}_2}(\mathcal{X}_n)/m_{n,p}$, $\boldsymbol{y}_1 \neq \boldsymbol{y}_2$, are non-degenerate incomplete U-statistics (with kernel $h_{\boldsymbol{y}_1,\boldsymbol{y}_2}$) in the form

$$\widetilde{U}_n^{\boldsymbol{y}_1,\boldsymbol{y}_2} = \frac{1}{m_{n,k}} \sum_{\boldsymbol{q}(n,k)\in\mathcal{Q}_S(n,k)} h_{\boldsymbol{y}_1,\boldsymbol{y}_2}(\boldsymbol{q}(n,k)), \quad m_{n,k} = |\mathcal{Q}_S(n,k)|,$$

(Note also that even $\widetilde{L}_{\boldsymbol{y}_1,\boldsymbol{y}_2}(\mathcal{X}_n)$ may be viewed as an incomplete U-statistic if the design set is really in the form specified above.)

The properties of $\widetilde{U}_n^{\boldsymbol{y}_1,\boldsymbol{y}_2}$ and relations to the corresponding complete U-statistics

$$U_n^{\boldsymbol{y}_1,\boldsymbol{y}_2} = \frac{1}{\binom{n}{k}} \sum_{\boldsymbol{q}(n,k)\in\mathcal{Q}(n,k)} h_{\boldsymbol{y}_1,\boldsymbol{y}_2}(\boldsymbol{q}(n,k))$$

have already been thoroughly investigated in the literature; see Blom (1976) and Janson (1984) for most of the results presented here. All the references below nevertheless point to the book Lee (1990) for simplicity. Note also that both $\widetilde{U}_n^{\boldsymbol{y}_1,\boldsymbol{y}_2}$ and $U_n^{\boldsymbol{y}_1,\boldsymbol{y}_2}$ are unbiased.

Here, the kernel is

$$h_{\boldsymbol{y}_1,\boldsymbol{y}_2}(\boldsymbol{q}(n,k)) = \left(1 - S^{\boldsymbol{q}(n,k)}(\boldsymbol{y}_1)S^{\boldsymbol{q}(n,k)}(\boldsymbol{y}_2)\right)/2$$

and k stands for the kernel order equal to the dimension of $\boldsymbol{q}(n,k)$: $k = p-1$ for $\widetilde{C}_{\boldsymbol{y}_1,\boldsymbol{y}_2}(\mathcal{X}_n)$, and $k = p$ for $\widetilde{L}_{\boldsymbol{y}_1,\boldsymbol{y}_2}(\mathcal{X}_n)$. The kernel, despite the simplified notation, actually depends on k independent and identically distributed random variables from the random sample $\mathcal{X}_n$. Similarly, all the arguments and indices may be omitted when no confusion is possible. In particular, $Q_S(n,k)$ may be shortened to Q_S and $m_{n,k}$ to m_n.

The first question concerns the choice of the design set $\mathcal{Q}_S$ and its influence on the asymptotic relative efficiency (ARE) of $\widetilde{U}_n$ with respect to U_n:

$$ARE(\widetilde{U}_n, U_n) = \lim_{n\to\infty} \frac{\mathrm{var}U_n}{\mathrm{var}\widetilde{U}_n}.$$

The following discussion is limited to the three most important cases when the design elements of $\mathcal{Q}_S$ are set deterministically or chosen randomly with or without replacement. Then the theory of incomplete U-statistics answers the problem satisfactorily (Lee 1990, Section 4.3).

Define

$$\begin{aligned} h_c(\boldsymbol{x}_1,\ldots,\boldsymbol{x}_c) &:= \mathrm{E}h(\boldsymbol{x}_1,\ldots,\boldsymbol{x}_c,\boldsymbol{X}_{c+1},\ldots,\boldsymbol{X}_k), \\ p_c &:= \mathrm{P}(h_c(\boldsymbol{X}_1,\ldots,\boldsymbol{X}_c) = 1), \quad \text{and} \\ \sigma_c^2 &:= \mathrm{var}(h_c(\boldsymbol{X}_1,\ldots,\boldsymbol{X}_c)), \end{aligned}$$

$c = 1,\ldots,k$. Generally, $\sigma_d^2 \geq (d/c)\sigma_c^2$ for any $k \geq d \geq c \geq 1$ (Lee 1990, Theorem 4 on p. 15). In the special context considered here, each h_c follows the Bernoulli distribution $B(1,p_c)$ and, therefore, $\sigma_c^2 = p_c(1-p_c)$, which also depends on $\boldsymbol{y}_1$ and $\boldsymbol{y}_2$. These quantities influence the asymptotic relative efficiencies stated below.

Recall that $\lim_{n\to\infty} nvarU_n = k^2\sigma_1^2$ (Lee 1990, Theorem 3 on p. 12). Furthermore, Theorem 4 on p. 193 there makes it possible to obtain the variance of $\widetilde{U}_n$ for the two special random designs considered. Consequently, if $\mathcal{Q}_S$ is chosen from $\mathcal{Q}$ by random sampling with replacement, then

$$ARE(\widetilde{U}_n, U_n) = \lim_{n\to\infty} \left(\frac{n}{k^2 m_n} \frac{\sigma_k^2}{\sigma_1^2} + 1 - \frac{1}{m_n} \right)^{-1},$$

and if $\mathcal{Q}_S$ is chosen from $\mathcal{Q}$ by random sampling without replacement, then

$$ARE(\widetilde{U}_n, U_n) = \lim_{n\to\infty} \left(\frac{(C_{n,k} - m_n)}{(C_{n,k} - 1)} \frac{n}{k^2 m_n} \frac{\sigma_k^2}{\sigma_1^2} + \frac{C_{n,k}}{C_{n,k} - 1} \left(1 - \frac{1}{m_n} \right) \right)^{-1},$$

where $C_{n,k} = \binom{n}{k}$. The difference between the two sampling schemes is thus minimal if both m_n and n are large but $m_n \ll C_{n,k}$.

Fixed designs are interesting only when they are simple or optimal. In the latter case, they should minimise the variance of $\widetilde{U}_n$ and, therefore, maximise its $ARE(\widetilde{U}_n, U_n)$. In view of symmetry, it seems natural to consider only equireplicate (or, balanced) designs when each index appears in the same number, say r, of the elements of $\mathcal{Q}_S$. Then $mk = nr$.

If the simplest balanced design $\mathcal{Q}_S = \{(1, \ldots, k), (k+1, \ldots, 2k), \ldots, (mk - k + 1, \ldots, n)\}$ is considered for $n = mk$, then

$$ARE(\widetilde{U}_n, U_n) = k\frac{\sigma_1^2}{\sigma_k^2} \leq 1.$$

In general, $\mathcal{Q}_S$ is a minimum variance design, if, for each $\nu = 1, 2, \ldots, k-1$, every ν-subset of $\{1, 2, \ldots, n\}$ is contained in the same number $m\binom{k}{\nu}/\binom{n}{\nu}$ of elements of $\mathcal{Q}_S$ (Lee 1990, Theorem 1 on p. 195). For example, if $\mathcal{Q}_S(n, 2)$ is a balanced design of $\widetilde{U}_n$, then it must be optimal and

$$ARE(\widetilde{U}_n, U_n) = \frac{2r}{2(r-1) + \frac{\sigma_2^2}{\sigma_1^2}} \leq 1$$

increases with r (Lee 1990, p. 196).

Write θ for the common expectation of unbiased $\widetilde{U}_n$ and U_n: $\theta = \mathrm{E}\widetilde{U}_n = \mathrm{E}U_n$.

For the two random designs considered here, the justification for the use of incomplete statistics instead of the complete ones follows from Theorem 1 on p. 200 of Lee (1990) restated here:

Theorem 3.1: *Assume* $\min(n, m_n) \to \infty$, $m_n/n \to \infty$, *and that* $\mathcal{Q}_S$ *results from the random sampling with or without replacement. Then* $\sqrt{n}(\widetilde{U}_n - \theta)$ *and* $\sqrt{n}(U_n - \theta)$ *have the same asymptotic distribution.*

Analogous Theorem 3 on p. 211 of Lee (1990) for fixed balanced designs denotes with f_{cn} the number of pairs $\boldsymbol{q}_1, \boldsymbol{q}_2 \in \mathcal{Q}_S$ that have exactly c elements in common, $c = 0, 1, \ldots, k$:

Theorem 3.2: *If $\mathcal{Q}_S$ is a balanced design*, $\min(n, m_n) \to \infty$, $nf_{cn}/m_n^2 \to 1$ *for* $c = 0$, *and* $nf_{cn}/m_n^2 \to 0$ *for* $c = 1, \ldots, k$, *then* $\sqrt{n}(\widetilde{U}_n - \theta)$ *and* $\sqrt{n}(U_n - \theta)$ *have the same asymptotic distribution.*

Note that the assumption of Theorem 3.1 is satisfied even for $m_n \ll \binom{n}{k}$ such as $m_n = O(n\log(\log(\log(n))))$, irrespective of the fixed kernel order k. This fact seems rather unappreciated in the field of multivariate statistics.

Now assume that $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n$ is a random sample from an elliptical distribution with median vector $\boldsymbol{\theta} \in \mathbb{R}^p$, positive definite scatter matrix $\boldsymbol{\Sigma} \in \mathbb{R}^{p\times p}$, and density proportionate to

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

where function $f : [0, \infty) \to [0, \infty)$ satisfies $\int_0^\infty z^{p-1}f(z)\,\mathrm{d}z < \infty$. Then $\|\boldsymbol{\Sigma}^{-1/2}(\boldsymbol{X}_1 - \boldsymbol{\theta})\|$ has cumulative distribution function F_r with density $f_r(z) \propto z^{p-1}f(z)I[z > 0]$.

Without any loss of generality, consider the one-sample testing problem with the null hypothesis $H_0 : \boldsymbol{\theta} = \boldsymbol{0}$ and alternative $H_1 : \boldsymbol{\theta} \neq \boldsymbol{0}$. Randles (1989) proposed a test for H_0 using interdirections and showed that the test statistic

$$V_B := \frac{p}{n}\sum_{i=1}^{n}\sum_{j=1}^{n}\cos(a_{\boldsymbol{X}_i,\boldsymbol{X}_j})$$

has an asymptotic (distribution-free) χ_p^2 null distribution with p degrees of freedom in the whole class of elliptical distributions.

It appears that the test remains valid even when the interdirections are replaced with incomplete interdirections.

Proposition 3.3: *Assume a random sample $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n$ from an elliptical distribution and consider incomplete interdirections $\widetilde{C}_{\boldsymbol{X}_i,\boldsymbol{X}_j}(\mathcal{X}_n)$, $i, j = 1, \ldots, n$, based on $\mathcal{Q}_S(n, p-1)$ (with cardinality $|\mathcal{Q}_S(n, p-1)| = m_{n,p-1}$) chosen randomly with or without replacement from $\mathcal{Q}(n, p-1)$. If H_0 holds and $m_{n,p-1}/n \to \infty$, then*

$$S_B := \frac{p}{n}\sum_{i=1}^{n}\sum_{j=1}^{n}\cos(\tilde{a}_{\boldsymbol{X}_i,\boldsymbol{X}_j}^{m_{n,p-1}}) \xrightarrow{\mathcal{D}} \chi_p^2 \quad \text{for } n \to \infty \tag{1}$$

where $\tilde{a}_{\boldsymbol{X}_i,\boldsymbol{X}_j}^{m_{n,p-1}} = \pi\widetilde{C}_{\boldsymbol{X}_i,\boldsymbol{X}_j}(\mathcal{X}_n)/m_{n,p-1}$.

Note that Proposition 3.3 holds even for the distributions with elliptical directions (Randles 1989) because of the same arguments contained in that article.

Also the three types of ranks considered here are asymptotically equivalent.

Proposition 3.4: *If $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n$ is a random sample of size n from an elliptical distribution*, $m_{n,p}/n \to \infty$, *and* $m_{n,p}^s/n \to \infty$, *then*

$$\frac{\widetilde{R}_i}{n+1} = \frac{R_i}{n+1} + o_P(1), \quad \text{and} \quad \frac{\underline{\widetilde{R}}_i}{n+1} = \frac{R_i}{n+1} + o_P(1)$$

as $n \to \infty$.

Oja and Paindaveine (2005) proposed a test statistic for H_0 that also has the asymptotic χ^2_p null distribution for all elliptical distributions and that uses both interdirections and ranks of symmetrised lift-interdirections which can still be replaced with their incomplete variants without changing the limiting null distribution under mild assumptions.

Assumption A: Assume a function K that is continuous almost everywhere on (0,1) and satisfies $(1/n)\sum_{i=1}^{n} |K(i/(n+1))|^{2+\delta} \to \int_0^1 |K(u)|^{2+\delta}\,\mathrm{d}u < \infty$ for some $\delta > 0$.

Proposition 3.5: *Assume that a random sample $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n$ comes from an elliptical distribution and that a function K satisfies Assumption A. Consider incomplete interdirections $\widetilde{C}_{\boldsymbol{X}_i,\boldsymbol{X}_j}(\mathcal{X}_n)$ and incomplete symmetrised lift-interdirections $\widetilde{L}_{\boldsymbol{X}_i}(\mathcal{X}_n)$, $i,j = 1,\ldots,n$, that are based, respectively, on the design sets $\mathcal{Q}_S(n,p-1)$, $|\mathcal{Q}_S(n,p-1)| = m_{n,p-1}$, and $\mathcal{Q}^s_S(n,p)$, $|\mathcal{Q}^s_S(n,p)| = m^s_{n,p}$, chosen independently and randomly with or without replacement. If H_0 holds, $m_{n,p-1}/n \to \infty$, and $m^s_{n,p}/n \to \infty$, then*

$$S_A := \frac{p}{n\,\mathrm{E}\,K^2(V)} \sum_{i=1}^{n}\sum_{j=1}^{n} K\left(\frac{\widetilde{R}_i}{n+1}\right) K\left(\frac{\widetilde{R}_j}{n+1}\right) \cos(\tilde{a}^{m_{n,p-1}}_{\boldsymbol{X}_i,\boldsymbol{X}_j}) \xrightarrow{\mathcal{D}} \chi^2_p \tag{2}$$

as $n \to \infty$ where $\tilde{a}^{m_{n,p-1}}_{\boldsymbol{X}_i,\boldsymbol{X}_j} = \pi \widetilde{C}_{\boldsymbol{X}_i,\boldsymbol{X}_j}(\mathcal{X}_n)/m_{n,p-1}$ and V is uniformly distributed on $[0,1]$.

All the propositions stated above are proved in the Appendix.

4. Simulation study

Consider a p-dimensional random sample $\mathcal{X}_n$ of size n, $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n$, that comes from an elliptical distribution with location parameter $\boldsymbol{\theta}$. Use random selection with replacement to independently construct design sets $\mathcal{Q}_S(n,p-1)$ (of size N_H) and $\mathcal{Q}^s_S(n,p)$ (also of size N_H) and compute incomplete interdirections $\widetilde{C}_{\boldsymbol{X}_i,\boldsymbol{X}_j}(\mathcal{X}_n)$ and incomplete symmetrised lift-interdirections $\widetilde{L}_{\boldsymbol{X}_i}(\mathcal{X}_n)$, $i,j = 1,\ldots,n$. Use van der Waerden's scores, i.e., set $K = \sqrt{F^{-1}}$ where F^{-1} stands for the quantile function of the χ^2_p distribution (with p-degrees of freedom).

Then the null hypothesis $H_0 : \boldsymbol{\theta} = \boldsymbol{0}$ can be checked by means of the χ^2_p signed-rank test T_A and sign test T_B that are based on the statistics S_A (2) and S_B (1), respectively. Their asymptotic null distributions are assumed χ^2_p in both cases (for p fixed and $n \to \infty$), which is justified in the previous section and proved in the Appendix.

The tests T_A and T_B are compared with the benchmark Hotelling's T^2 test (as implemented in the ICSNP package (Nordhausen, Sirkia, Oja, and Tyler 2015) for R (R Development Core Team 2008)) in Figures 1 and 2 for a few elliptical distributions and parameters n, p, and N_H in terms of empirical power based on 1000 independent replications. In particular, the simulation study uses multivariate canonical Student t null distributions with $q = 1$, $q = 2.01$, and $q = \infty$ degrees of freedom, $n = 100$ or $n = 1000$ p-dimensional observations with fixed p such as $p = 10$ or $p = 100$ always less than n, and $N_H = 5n$. Every simulated data set was used for evaluating the tests under both H_0 and all ten shift alternatives considered.

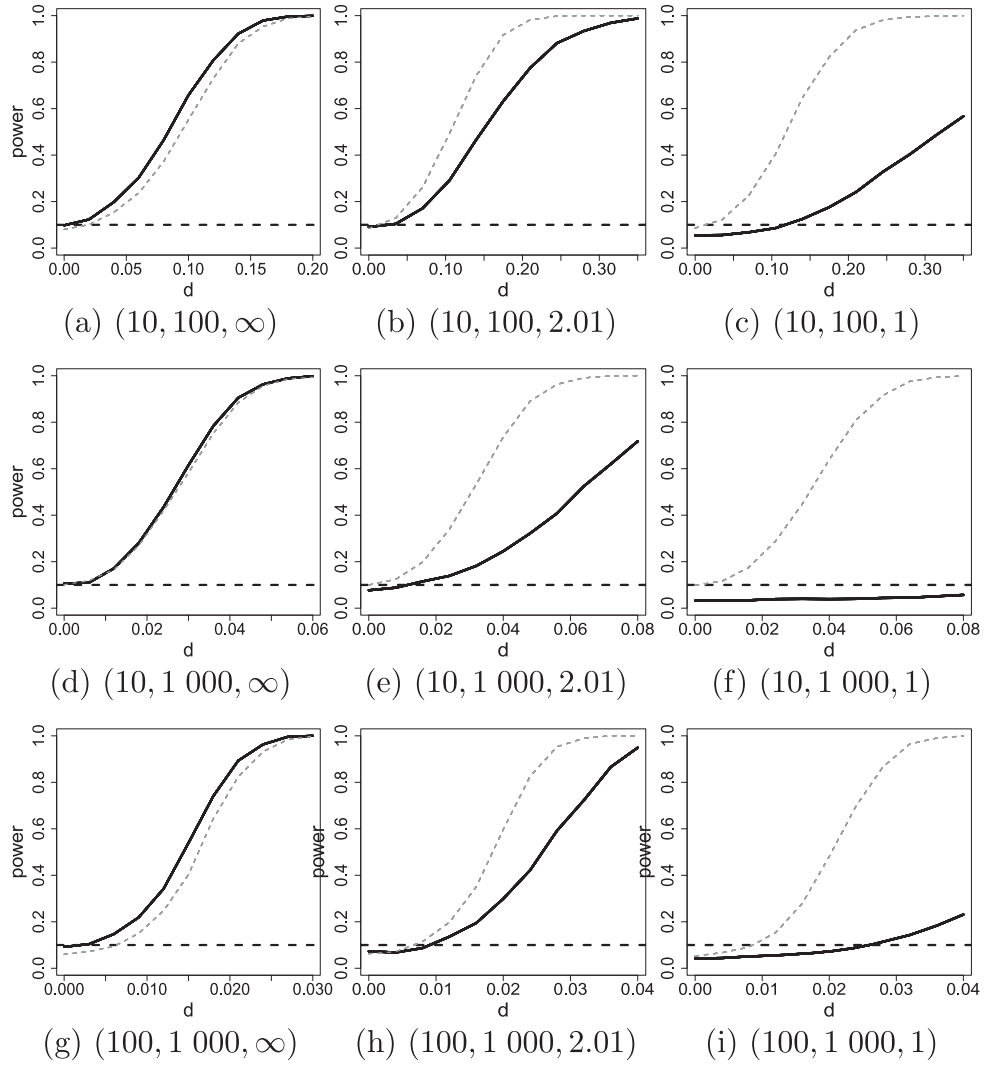

Figure 1. *Power comparison of two one-sample tests at significance level* $\alpha = 0.10$. The plots display empirical powers of the sign test T_B (dotted line) and the benchmark Hotelling's test (thick solid line) with common significance level $\alpha = 0.10$ (horizontal thick dashed line). The sign test based on $N_H = 5n$ hyperplanes was applied to 1000 p-dimensional random samples of size n from the multivariate canonical Student t distribution with q degrees of freedom shifted by $(d, \ldots, d)'$. The parametric combination (p, n, q) characterising each experiment is stated below individual pictures. Each column of subfigures corresponds to one null distribution.

Apparently, even N_H mildly higher than n already leads to tests T_A and T_B with power comparable to the benchmark for normally distributed data and with far better performance for heavy-tailed distributions.

In fact, Figure 3 illustrates the sensitivity of test T_B to the choice of N_H and confirms that even $N_H = 25$ may be satisfactory for common sample sizes ($n = 100$ or $n = 500$) to achieve a desirable test performance, at least in case of not too large p such as $p = 5$ and the multivariate canonical Student t null distribution with only one degree of freedom.

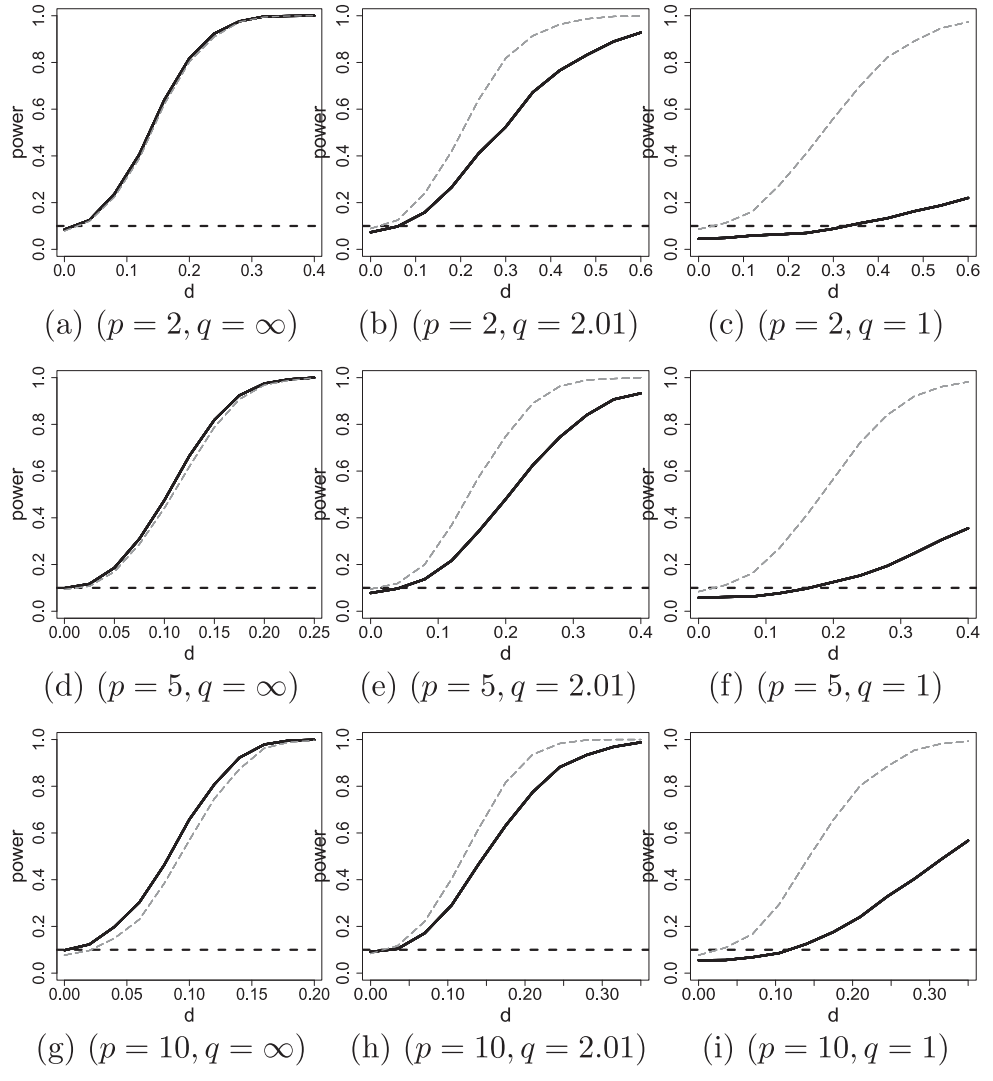

Figure 2. *Power comparison of two one-sample tests at significance level* $\alpha = 0.10$. The plots display empirical powers of the signed-rank test T_A (dashed line) and the benchmark Hotelling's test (thick solid line) with common significance level $\alpha = 0.10$ (horizontal thick dashed line). The signed-rank test based on $2^p N_H$ hyperplanes, $N_H = 5n$, was applied to 1000 p-dimensional random samples of size $n = 100$ from the multivariate canonical Student t distribution with q degrees of freedom shifted by $(d, \ldots, d)'$. The parametric combination (p, q) characterising each experiment is stated below individual pictures. Each column of subfigures corresponds to one null distribution.

The simulation studies do not include the classic tests (based on complete interdirections/symmetrised lift-interdirections), because their computation would be too time consuming or even impossible for most settings considered. Speed comparison is omitted for the same reason. Except for some rather special cases of low n and p, the difference would be not only in the speed of computation but also in its feasibility.

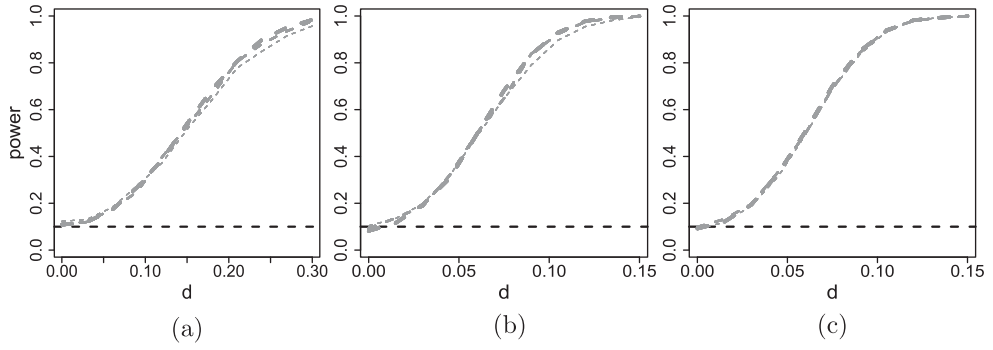

Figure 3. *Test dependence on N_H.* The plots display empirical powers of the sign test T_B (dashed line) with common significance level $\alpha = 0.10$ (horizontal thick dashed line). The test based on N_H^1 (thin), N_H^2 (normal) or N_H^3 (thick) hyperplanes was applied to 1000 five-dimensional ($p = 5$) random samples of size (a) $n = 100$ or (b,c) $n = 500$ from the multivariate canonical Student t distribution with one degree of freedom shifted by $(d, \ldots, d)'$. In particular, $N_H^1 = cn/4$, $N_H^2 = cn/2$ and $N_H^3 = cn$ where $c = 1/5$ in (b) and $c = 1$ in (a) and (c), which leads to the same N_H^i in (a) and (b), $i = 1, 2, 3$. Apparently, there is hardly any visible difference in the test performance if the number of included hyperplanes is not too small.

To sum up the results, the incomplete interdirections and incomplete lift-interdirections make it possible to obtain powerful and computationally feasible tests of high-dimensional data.

5. Discussion

The classic tests based on interdirections or (symmetrised) lift-interdirections are computationally feasible only if the number n or the dimension p of the observations is small. The article uses incomplete U-statistics to define incomplete versions of those hyperplane-based concepts and shows their usefulness for nonparametric statistical inference in spaces with large but fixed dimensions.

In particular, the number of hyperplanes considered by the incomplete interdirections does not grow with the dimension of observations at all while it is equal to $\binom{n}{p-1}$ in case of their complete counterparts. Consequently, even dimensions like $p = 100$ can be handled easily.

As for the incomplete symmetrised lift-interdirections, the number of processed hyperplanes grows with the dimension by factor 2^p, which still may make the computation challenging for $p > 10$ or so. But the improvement achieved by them is still considerable.

The classic tests should be preferable due to their non-stochastic nature given the observations. When they are not feasible to compute, then the incomplete variants provide a way how to proceed further. The results for hyperplanes sampled at random with or without replacement are then virtually the same because the chance of choosing a hyperplane at least twice in the former case is negligible. Therefore, the first option is recommended due to its simplicity.

The brief simulation study confirms that the exact number N_H of hyperplanes considered by the incomplete variants is not too important and that even N_H as low as $N_H = 5n$ can still produce reliable results in common settings with $n \leq 1000$ and $p \leq 100$ as far as

$p \leq n/10$ (ideally $p \leq n/20$). In fact, even $N_H = 25$ was shown satisfactory for small p and not too large n in a very limited comparison. Nevertheless, further simulations are needed to confirm and extend the rules to other contexts and to explore the incomplete variants in small samples.

This article also shows that incomplete U-statistics and simple geometric concepts of multivariate signs and ranks may be very useful for nonparametric inference and play important role in the future.

Disclosure statement

No potential conflict of interest was reported by the authors.

Funding

The research of Miroslav Šiman and Jana Klicnarová was supported by the Czech Science Foundation (Grantová agentura České republiky) project GA17-07384S. The work of Šárka Hudecová was supported by the Czech Science Foundation project GA18-01781Y.

References

Blom, G. (1976), 'Some Properties of Incomplete U-Statistics', *Biometrika*, 63, 573–580.

Chaudhuri, P., and Sengupta, D. (1993), 'Sign Tests in Multidimension: Inference Based on the Geometry of the Data Cloud', *Journal of the American Statistical Association*, 88, 1363–1370.

Chernozhukov, V., Galichon, A., Hallin, M., and Henry, M. (2017), 'Monge-Kantorovich Depth, Quantiles, Ranks, and Signs', *The Annals of Statistics*, 45, 223–256.

Došlá, Š. (2009), 'Conditions for Bimodality and Multimodality of a Mixture of Two Unimodal Densities', *Kybernetika*, 45, 279–292.

Ghoush, S.K., and Sengupta, D. (2001), 'Testing for Proportionality of Multivariate Dispersion Structures Using Interdirections', *Journal of Nonparametric Statistics*, 13, 331–349.

Gieser, P.W., and Randles, R.H. (1997), 'A Nonparametric Test of Independence Between Two Vectors', *Journal of the American Statistical Association*, 92, 561–567.

Hallin, M., and Paindaveine, D. (2002a), 'Optimal Procedures Based on Interdirections and Pseudo-Mahalanobis Ranks for Testing Multivariate Elliptic White Noise Against ARMA Dependence', *Bernoulli*, 8, 787–815.

Hallin, M., and Paindaveine, D. (2002b), 'Optimal Tests for Multivariate Location Based on Interdirections and Pseudo-Mahalanobis Ranks', *The Annals of Statistics*, 30, 1103–1133.

Hallin, M., and Paindaveine, D. (2004), 'Multivariate Signed-Rank Tests in Vector Autoregressive Order Identification', *Statistical Science*, 19, 697–711.

Hettmansperger, T.P., Möttönen, J., and Oja, H. (1999), 'The Geometry of the Affine Invariant Multivariate Sign and Rank Methods', *Journal of Nonparametric Statistics*, 11, 271–285.

Jan, S.L., and Randles, R.H. (1996), 'Interdirection Tests for Simple Repeated-Measures Designs', *Journal of the American Statistical Association*, 91, 1611–1618.

Janson, S. (1984), 'The Asymptotic Distributions of Incomplete U-Statistics', *Zeitschrift für Wahrscheinlichkeitstheorie und Verwandte Gebiete*, 66, 495–505.

Lee, A.J. (1990), *U-Statistics: Theory and Practice*, New York: CRC Press.

Möttönen, J., and Oja, H. (1995), 'Multivariate Spatial Sign and Rank Methods', *Journal of Nonparametric Statistics*, 5, 201–213.

Nordhausen, K., Sirkia, S., Oja, H., and Tyler, D.E. (2015), *ICSNP: Tools for Multivariate Nonparametrics*, R package version 1.1-0, Available at https://CRAN.R-project.org/package = ICSNP.

Oja, H. (1999), 'Affine Invariant Multivariate Sign and Rank Tests and Corresponding Estimates: A Review', *Scandinavian Journal of Statistics*, 26, 319–343.

Oja, H. (2010), *Multivariate Nonparametric Methods With R: An Approach Based on Spatial Signs and Ranks*, New York: Springer. ISBN 9781441904676.

Oja, H., and Paindaveine, D. (2005), 'Optimal Signed-Rank Tests Based on Hyperplanes', *Journal of Statistical Planning and Inference*, 135, 300–323.
Paindaveine, D. (2009), 'On Multivariate Runs Tests for Randomness', *Journal of the American Statistical Association*, 104, 1525–1538.
Puri, M.L., and Sen, P.K. (1971), *Nonparametric Methods in Multivariate Analysis*, New York: Wiley. ISBN 0471702404.
R Development Core Team (2008), *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria. Available at http://www.R-project.org, ISBN 3900051070.
Randles, R.H. (1989), 'A Distribution-Free Multivariate Sign Test Based on Interdirections', *Journal of the American Statistical Association*, 84, 1045–1050.
Randles, R.H., and Peters, D. (1990), 'Multivariate Rank Tests for the Two-Sample Location Problem', *Communications in Statistics – Theory and Methods*, 19, 4225–4238.
Serfling, R.J. (2006), 'Multivariate Symmetry and Asymmetry', in *Encyclopedia of Statistical Sciences* (Vol. 8, 2nd ed.), eds. S. Kotz, N. Balakrishnan, C.B. Read, and B. Vidakovic, pp. 5338–5345. New York: Wiley.
Taskinen, S., Oja, H., and Randles, R.H. (2005), 'Multivariate Nonparametric Tests of Independence', *Journal of the American Statistical Association*, 100, 916–925.
Um, Y., and Randles, R.H. (1998), 'Nonparametric Tests for the Multivariate Multi-Sample Location Problem', *Statistica Sinica*, 8, 801–812.
Zuo, Y., and Serfling, R. (2000), 'General Notions of Statistical Depth Functions', *The Annals of Statistics*, 28, 461–482.

Appendix. Proofs

All the expectations involved in the proofs are taken with respect to the probability distribution of $\boldsymbol{X}$ and with respect to the random designs involved due to the use of incomplete interdirections and symmetrised lift-interdirections. For the sake of simplicity, this is stressed out with a subscript only where confusion would be possible.

Proof of Proposition 3.3: Thanks to the invariance properties of incomplete interdirections, it is sufficient to consider the situation when $\boldsymbol{X}_i = \boldsymbol{U}_i$, $i = 1, \ldots, n$, where $\boldsymbol{U}_1, \ldots, \boldsymbol{U}_n$ are independent variables uniformly distributed on the unit sphere in $\mathbb{R}^p$. The same arguments as in the proof of Theorem A.1. in Randles (1989) imply

$$\mathrm{E}\left[S_B - \frac{p}{n}\sum_{i=1}^{n}\sum_{j=1}^{n}\boldsymbol{U}_i^\top \boldsymbol{U}_j\right]^2 \leq 2\pi^2 p^2 \mathrm{E}\left[\frac{\widetilde{C}_{\boldsymbol{U}_1,\boldsymbol{U}_2}(\mathcal{X}_n)}{m_{n,p-1}} - \frac{1}{\pi}\arccos(\boldsymbol{U}_1^\top \boldsymbol{U}_2)\right]^2.$$

The right unconditional expectation converges to zero for $n \to \infty$ because all the conditional expectations given $\boldsymbol{U}_1$, $\boldsymbol{U}_2$ and $\mathcal{Q}_S(n, p-1)$ do so owing to Theorem 3.1 and Lemma 1 in Hallin and Paindaveine (2002b). Consequently, S_B must have the same asymptotic distribution as $(p/n)\sum_{i=1}^{n}\sum_{j=1}^{n}\boldsymbol{U}_i^\top \boldsymbol{U}_j$, which is known to be χ^2_p. ■

Proof of Proposition 3.4.: Consider only $\widetilde{R}_i$ (as the proof for $\underline{\widetilde{R}}_i$ is analogous), arbitrarily fix $\boldsymbol{x} \in \mathbb{R}^p$, and assume $\boldsymbol{\Sigma} = \mathbf{I}_p$ without any loss of generality. If $\boldsymbol{X}_i = \boldsymbol{x}$, then $\widetilde{R}_i \equiv \widetilde{R}_i(\boldsymbol{x}) = \sum_{j=1}^{n} I[\widetilde{L}_{\boldsymbol{x}}(\mathcal{X}_n) \geq \widetilde{L}_{\boldsymbol{X}_j}(\mathcal{X}_n)]$ and $R_i \equiv R_i(\boldsymbol{x}) = \sum_{j=1}^{n} I[\|\boldsymbol{x}\| \geq \|\boldsymbol{X}_j\|]$. It suffices to show that $n^{-2}\mathrm{E}[\widetilde{R}_i(\boldsymbol{x}) - R_i(\boldsymbol{x})]^2 \to 0$ as $n \to \infty$, which follows, as in the proof of Proposition 2 in Oja and Paindaveine (2005), from

$$\int_{\mathbb{R}^p} g_j(\boldsymbol{x}, \boldsymbol{z}) f(\boldsymbol{z})\, \mathrm{d}\boldsymbol{z} = o(1), \quad j = 1, 2, \tag{A1}$$

where $g_1(\boldsymbol{x}, \boldsymbol{z}) = \mathrm{EI}[\widetilde{L}_{\boldsymbol{x}}(\mathcal{X}_n) \geq \widetilde{L}_{\boldsymbol{z}}(\mathcal{X}_n), \|\boldsymbol{x}\| < \|\boldsymbol{z}\|]$ and $g_2(\boldsymbol{x}, \boldsymbol{z}) = \mathrm{EI}[\widetilde{L}_{\boldsymbol{x}}(\mathcal{X}_n) < \widetilde{L}_{\boldsymbol{z}}(\mathcal{X}_n), \|\boldsymbol{x}\| \geq \|\boldsymbol{z}\|]$. This is shown in the following.

If $\|\boldsymbol{x}\| \geq \|\boldsymbol{z}\|$, then $g_1(\boldsymbol{x},\boldsymbol{z}) = 0$. Furthermore, Theorem 3.1 and Oja and Paindaveine (2005) imply that $m_{n,p}^{-1}\widetilde{L}_{\boldsymbol{x}}(\mathcal{X}_n)$ converges in quadratic mean to the theoretical lift-interdirection $l(\boldsymbol{x}) = \mathrm{E}m_{n,p}^{-1}\widetilde{L}_{\boldsymbol{x}}(\mathcal{X}_n)$, which is a strictly increasing function of $\|\boldsymbol{x}\|$. If $\|\boldsymbol{x}\| < \|\boldsymbol{z}\|$, then the Chebyshev inequality consequently results in

$$\begin{aligned} g_1(\boldsymbol{x},\boldsymbol{z}) &= \mathsf{P}(m_{n,p}^{-1}[\widetilde{L}_{\boldsymbol{x}}(\mathcal{X}_n) - \widetilde{L}_{\boldsymbol{z}}(\mathcal{X}_n)] \geq 0) \\ &\leq \mathsf{P}\left(\left|m_{n,p}^{-1}[\widetilde{L}_{\boldsymbol{x}}(\mathcal{X}_n) - \widetilde{L}_{\boldsymbol{z}}(\mathcal{X}_n)] - (l(\boldsymbol{x}) - l(\boldsymbol{z}))\right| \geq l(\boldsymbol{z}) - l(\boldsymbol{x})\right) \\ &\leq \frac{\mathrm{var}[m_{n,p}^{-1}\widetilde{L}_{\boldsymbol{x}}(\mathcal{X}_n) - m_{n,p}^{-1}\widetilde{L}_{\boldsymbol{z}}(\mathcal{X}_n)]}{(l(\boldsymbol{z}) - l(\boldsymbol{x}))^2} = O(1/n) = o(1) \end{aligned}$$

as $n \to \infty$ for all $\boldsymbol{x},\boldsymbol{z} \in \mathbb{R}^p$, which implies the result because g_1 is bounded. The proof for g_2 is analogous. ■

Proof of Proposition 3.5: The proof mimics that of Proposition 3 in Hallin and Paindaveine (2002b), but some steps require different justification due to the use of incomplete versions of both interdirections and symmetrised lift-interdirections. These arguments are gathered in the concluding Lemma A.1.

Consider distances $d_i = \|\boldsymbol{\Sigma}^{-1/2}\boldsymbol{X}_i\|$ and independent random vectors $\boldsymbol{U}_i = d_i^{-1}\boldsymbol{\Sigma}^{-1/2}\boldsymbol{X}_i$ uniformly distributed on the unit sphere in $\mathbb{R}^p$, $i = 1,\ldots,n$. Assume $\boldsymbol{\Sigma} = \mathbf{I}_p$ without any loss of generality and define:

$$T_1^n = \frac{1}{n}\sum_{i,j=1}^n K\left(\frac{\widetilde{R}_i}{n+1}\right)K\left(\frac{\widetilde{R}_j}{n+1}\right)\left[\cos(\tilde{a}^{m_{n,p-1}}_{\boldsymbol{X}_i,\boldsymbol{X}_j}) - \boldsymbol{U}_i^\top\boldsymbol{U}_j\right],$$

$$T_2^n = \frac{1}{n}\sum_{i,j=1}^n \left[K\left(\frac{\widetilde{R}_i}{n+1}\right)K\left(\frac{\widetilde{R}_j}{n+1}\right) - K(F_r(d_i))K(F_r(d_j))\right]\boldsymbol{U}_i^\top\boldsymbol{U}_j,$$

$$\mathbf{T}_n = \frac{1}{\sqrt{n}}\sum_{i=1}^n K(F_r(d_i))\boldsymbol{U}_i,$$

$$\mathbf{S}_n = \frac{1}{\sqrt{n}}\sum_{i=1}^n K(R_i/(n+1))\boldsymbol{U}_i, \quad \text{and} \quad \widehat{\mathbf{S}}_n = \frac{1}{\sqrt{n}}\sum_{i=1}^n K(\widetilde{R}_i/(n+1))\boldsymbol{U}_i$$

where R_i is the rank of d_i, $i = 1,\ldots n$.

To sum up, the proof approximates S_A with

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

whose limit distribution follows easily from the central limit theorem for $\mathbf{T}_n$. The difference $S_A - S_A^0 \equiv (p/\mathrm{E}[K^2(V)])(T_1^n + T_2^n)$ must be $o_P(1)$ because both $T_1^n \to 0$ and $T_2^n \to 0$ in probability as $n \to \infty$, which is proved below.

As for T_2^n, Hallin and Paindaveine (2002b) proved that $\mathrm{E}\|\mathbf{T}_n - \mathbf{S}_n\|^2 = o(1)$ and $\mathrm{E}[K(R_i/(n+1)) - K(F_r(d_i))]^2 = o(1)$ as $n \to \infty$. Lemma A.1(1) further implies that

$$\mathrm{E}\|\mathbf{S}_n - \widehat{\mathbf{S}}_n\|^2 = \frac{1}{n}\sum_{i=1}^n \mathrm{E}\left[K\left(\frac{R_i}{n+1}\right) - K\left(\frac{\widetilde{R}_i}{n+1}\right)\right]^2 \to 0$$

because the squared differences in the summands are both uniformly integrable thanks to Assumption A and $o(1)$ in probability thanks to Proposition 3.4 and the continuity of K almost everywhere. They are thus $o(1)$ in quadratic mean. Hence, $\mathrm{E}\|\mathbf{T}_n - \widehat{\mathbf{S}}_n\|^2 \to 0$, too. Therefore, $\mathrm{E}\|\mathbf{T}_n\|^2 < \infty$

implies $\mathrm{E}\|\widehat{\mathbf{S}}_n\|^2 < \infty$. The Cauchy-Schwartz inequality then concludes that

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

for $n \to \infty$, which proves that $T_2^n \to 0$ both in L^1 and in probability.

Now turn to T_1^n and define

$$D_i := K\left(\frac{\widetilde{R}_i}{n+1}\right) \quad \text{and} \quad C_{ij} := \cos(\tilde{a}^{m_{n,p-1}}_{\boldsymbol{X}_i,\boldsymbol{X}_j}) - \boldsymbol{U}_i^\top \boldsymbol{U}_j.$$

Then

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

for the particular $\delta > 0$ from Assumption A thanks to $C_{ii} = 0$ almost surely, $i = 1, \ldots, n$, Lemma A.1(2,3) and the Hölder inequality.

The conditional expectation of bounded C_{12} given the design, $\boldsymbol{X}_1$ and $\boldsymbol{X}_2$ always converges to 0 in quadratic mean for the same reasons as in Proposition 3.3. Therefore, $\mathrm{E}|C_{12}|^2 \to 0$ and also $\mathrm{E}|C_{12}|^{2(2+\delta)/\delta} \to 0$ as $n \to \infty$. Furthermore, $\mathrm{E}|D_1 D_2|^{2+\delta} < \infty$ due to Lemma A.1(2). Consequently, $T_1^n \to 0$ in L^2 (and in probability) as $n \to \infty$. ■

Lemma A.1: (1) *If $i \neq j$, then*

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

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

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

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

Proof: Recall that $\widetilde{R}_i$ and $\widetilde{C}_{\boldsymbol{X}_i,\boldsymbol{X}_j}(\mathcal{X}_n)$ respectively depend on independent and equally probable design sets $Q_S^s(n,p)$ and $Q_S(n,p-1)$ that are also independent of $\mathcal{X}$.

(1) Let $i \neq j$. The statement holds when

$$\mathrm{E}K(R_i/(n+1))K(R_j/(n+1))\boldsymbol{U}_i^\top \boldsymbol{U}_j = 0,$$
$$\mathrm{E}K(\widetilde{R}_i/(n+1))K(\widetilde{R}_j/(n+1))\boldsymbol{U}_i^\top \boldsymbol{U}_j = 0, \quad \text{and}$$
$$\mathrm{E}K(\widetilde{R}_i/(n+1))K(R_j/(n+1))\boldsymbol{U}_i^\top \boldsymbol{U}_j = 0.$$

Set $B_i := \mathrm{sgn}(X_{i1})$, $\mathbf{Y}_i := B_i \boldsymbol{X}_i$, 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 $\mathsf{P}(B_i = 1) = \mathsf{P}(B_i = -1) = 1/2, i = 1, \ldots, n$. Recall that $\boldsymbol{X}_i = d_i \boldsymbol{U}_i$ and $\boldsymbol{U}_i = \boldsymbol{X}_i/d_i = \mathbf{Y}_i B_i/d_i$ where $d_i = \|\boldsymbol{X}_i\| = \|\mathbf{Y}_i\|$. Hence, d_i is a function of $\mathbf{Y}_i$ and the ranks R_i's are functions of the sample $\mathcal{Y}$ consisting of $\mathbf{Y}_1, \ldots, \mathbf{Y}_n$. Consequently,

$$\mathrm{E}K(R_i/(n+1))K(R_j/(n+1))\boldsymbol{U}_i^\top \boldsymbol{U}_j$$
$$= \mathrm{E}_{\mathcal{Y}}\mathrm{E}_B[K(R_i/(n+1))K(R_j/(n+1))\boldsymbol{U}_i^\top \boldsymbol{U}_j|\mathbf{Y}_1, \ldots, \mathbf{Y}_n] =$$
$$= \mathrm{E}_{\mathcal{Y}}\frac{\mathbf{Y}_i^\top \mathbf{Y}_j}{d_i d_j}K(R_i/(n+1))K(R_j/(n+1))\mathrm{E}_B[B_i B_j|\mathbf{Y}_1, \ldots, \mathbf{Y}_n] = 0.$$

Any design $\mathcal{Q}_S^s(n,p)$ independent of $\mathcal{X}$ leads to the symmetrised lift-interdirections $\widetilde{\underline{L}}_{\boldsymbol{X}_i}(\mathcal{X}_n)$, $i = 1,\dots,n$, that are the same for any $s_1\boldsymbol{X}_1,\dots,s_n\boldsymbol{X}_n$, $(s_1,\dots,s_n) \in \{-1,1\}^p$. Therefore, they are only functions of $\mathbf{Y}_1,\dots,\mathbf{Y}_n$, and the same thus holds even for their ranks.

Consequently,

$$\begin{aligned}
&\mathrm{E}K(\widetilde{R}_i/(n+1))K(\widetilde{R}_j/(n+1))\boldsymbol{U}_i^\top\boldsymbol{U}_j\\
&\quad= \mathrm{E}_{\mathcal{Y}}\mathrm{E}_{\mathcal{Q}_S^s}\mathrm{E}_B[K(\widetilde{R}_i/(n+1))K(\widetilde{R}_j/(n+1))\boldsymbol{U}_i^\top\boldsymbol{U}_j|\mathbf{Y}_1,\dots,\mathbf{Y}_n,\mathcal{Q}_S^s]\\
&\quad= \mathrm{E}_{\mathcal{Y}}\mathrm{E}_{\mathcal{Q}_S^s}\frac{\mathbf{Y}_i^\top\mathbf{Y}_j}{d_id_j}K(\widetilde{R}_i/(n+1))K(\widetilde{R}_j/(n+1))\mathrm{E}_B\left[B_iB_j|\mathbf{Y}_1,\dots,\mathbf{Y}_n,\mathcal{Q}_S^s\right] = 0.
\end{aligned}$$

The proof of the last equality is analogous.

(2) Obviously,

$$\begin{aligned}
\mathrm{E}D_iD_jC_{ij} = \frac{1}{|\{Q_S^s\}||\{Q_S\}|}\sum_{Q_S^s}\sum_{Q_S}\mathrm{E}_{\mathcal{X}}K\left(\frac{\widetilde{R}_i}{n+1}\right)K\left(\frac{\widetilde{R}_j}{n+1}\right)\\
\times\left(\cos(\tilde{a}_{\boldsymbol{X}_i,\boldsymbol{X}_j}^{m_{n,p-1}}(\mathcal{X})) - \frac{\boldsymbol{X}_i^\top\boldsymbol{X}_j}{d_id_j}\right).
\end{aligned}$$

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 are properly independent and equiprobable. The following expectation is independent of (i,j), $i \neq j$, for the same reason:

$$\begin{aligned}
\mathrm{E}|D_1D_2|^{2+\delta} &= \frac{1}{|\{Q_S^s\}|}\sum_{Q_S^s}\mathrm{E}_{\mathcal{X}}|K(\widetilde{R}_1/(n+1))K(\widetilde{R}_2/(n+1))|^{2+\delta}\\
&= \sum_{l\neq m}|K(l/(n+1)|^{2+\delta}|K(m/(n+1))|^{2+\delta}\underbrace{\sum_{Q_S^s}\frac{1}{|\{Q_S^s\}|}\mathsf{P}(\widetilde{R}_1 = l,\widetilde{R}_2 = m)}_{p(l,m)}.
\end{aligned}$$

As $p(l,m)$ does not depend on l and m, necessarily $p = [n(n-1)]^{-1}$. The finiteness of $\mathrm{E}|D_1D_2|^{2+\delta}$ then follows from the arguments in the proof of Proposition 3 in Hallin and Paindaveine (2002b).

(3) Define $C_{ij}^{\mathcal{Y}} := \cos(\pi\widetilde{C}_{\mathbf{Y}_i,\mathbf{Y}_j}(\mathcal{Y}_n)/m_{n,p-1}) - \mathbf{Y}_i^\top\mathbf{Y}_j/(d_id_j)$, and realise that

$$\cos(\pi\widetilde{C}_{\boldsymbol{X}_i,\boldsymbol{X}_j}(\mathcal{X}_n)/m_{n,p-1}) = B_iB_j\cos(\pi\widetilde{C}_{\mathbf{Y}_i,\mathbf{Y}_j}(\mathcal{Y}_n)/m_{n,p-1}).$$

Then

$$\begin{aligned}
&\mathrm{E}D_iD_jC_{ij}D_kD_lC_{kl}\\
&\quad= \mathrm{E}_{\mathcal{Y}}\mathrm{E}_{\mathcal{Q}_S^s}\mathrm{E}_{\mathcal{Q}_S}C_{ij}^{\mathcal{Y}}C_{kl}^{\mathcal{Y}}\mathrm{E}_B[D_iD_jD_kD_lB_iB_jB_kB_l|\mathbf{Y}_1,\dots,\mathbf{Y}_n,\mathcal{Q}_S^s,\mathcal{Q}_S]\\
&\quad= \mathrm{E}_{\mathcal{Y}}\mathrm{E}_{\mathcal{Q}_S^s}\mathrm{E}_{\mathcal{Q}_S}D_iD_jC_{ij}^{\mathcal{Y}}D_kD_lC_{kl}^{\mathcal{Y}}\mathrm{E}_B[B_iB_jB_kB_l],
\end{aligned}$$

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