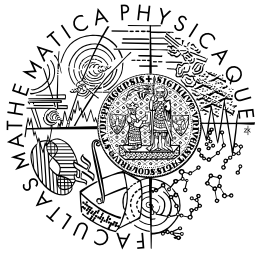

FACULTY
OF MATHEMATICS
AND PHYSICS
Charles University

MASTER THESIS

Bc. Veronika Roubínová

Center-outward ranks and signs and their application in statistical tests

Department of Probability and Mathematical Statistics

Supervisor of the master thesis: RNDr. Šárka Hudecová, Ph.D.

Study programme: Mathematics

Study branch: Probability, Mathematical Statistics and Econometrics

Prague 2023

I declare that I carried out this master thesis independently, and only with the cited sources, literature and other professional sources. It has not been used to obtain another or the same degree.

I understand that my work relates to the rights and obligations under the Act No. 121/2000 Sb., the Copyright Act, as amended, in particular the fact that the Charles University has the right to conclude a license agreement on the use of this work as a school work pursuant to Section 60 subsection 1 of the Copyright Act.

In ..... date ....

Author's signature

I would like to thank my supervisor RNDr. Šárka Hudecová, Ph.D. for the provided guidance and all the helpful remarks. On top of that, I would like to thank my parents and my partner for their endless support throughout my studies. Last but not least, I would like to thank Katka for the beneficial language proofreading.

Title: Center-outward ranks and signs and their application in statistical tests

Author: Bc. Veronika Roubínová

Department: Department of Probability and Mathematical Statistics

Supervisor: RNDr. Šárka Hudecová, Ph.D., Department of Probability and Mathematical Statistics

Abstract: This thesis describes the theory of multivariate rank tests based on center-outward ranks and signs. The definition of the center-outward ranks and signs is based on the measure transportation problem and depends highly on the chosen underlying grid. Several ways to generate such grids are suggested. Center-outward ranks and signs are then used to construct various test statistics for one-sample testing of location. The main contribution of the work is the introduction of new variants of the one-sample test of location. The proposed test statistics are based on randomized signs and added zero with the usage of the permutation tests for obtaining p-values. The tests are constructed under the assumption of both central or angular symmetry of the underlying distribution. In the end, a simulation study is performed to illustrate the performance of the proposed tests under different settings for several alternatives.

Keywords: center-outward distribution function, rank tests, multivariate ranks, multivariate signs, one-sample test of location

Contents

Introduction

Rank-based tests are well known and are used mostly for their convenient properties which include distribution-freeness. The concept of ranks and signs is connected with order statistics and therefore with the ordering of the real line. It is natural to try and generalize these ideas into their multidimensional versions. Unfortunately, the ordering of points on a real line cannot be easily transferred to multidimensional spaces. Therefore, the mentioned concepts are hard to define in $\mathbb{R}^d, d \geq 2$.

There are already many proposed ways how to introduce multivariate versions of ranks and signs and therefore also quantile functions. One possible way is to use ranks and signs componentwise, but this idea has several problems, including the fact that the ranks are not asymptotically distribution-free. Another way is to use depth-based ranks. These concepts have been studied by many statisticians, for more information see for example Liu & Singh (1993) or Li & Liu (2004).

Another approach built upon measure transportation has been recently proposed by Chernozhukov et al. (2017). Based on these concepts, the center-outward ranks and signs were defined. The center-outward ranks and signs have properties like distribution-freeness and essential maximal ancillarity, making them a useful tool in statistical testing, see Hallin et al. (2021) or Hallin, Liu & Verdebout (2022).

This thesis provides the definition of the center-outward ranks and signs based on the existing theory. Subsequently, our aim was to propose several test statistics for one-sample testing of location. The introduced semiparametric tests are useful in situations where normality or even central symmetry of underlying distribution cannot be assumed. Most of the presented concepts are supplemented with practical examples and illustrating figures. The thesis provides recommendations of specific tests useful in multiple various settings and compares the newly proposed one-sample tests with the already well-known Hotelling's test of location. The main contribution of this work is a proposal of one-sample tests of location under central and angular symmetry and their comparison with known tests in the simulation study.

In the first chapter, the basic properties of the rank-based tests are stated. The second chapter provides a definition of the center-outward ranks and signs, their main properties, and the construction of grids for the empirical center-outward distribution function. Then, test statistics based on the previously described theory are introduced in the third chapter and their asymptotic behavior is studied. Both, the two-sample and one-sample tests of location are discussed together with the proposal of a new test statistic based on adding zero for the one-sample problem. The one-sample tests are constructed under the assumption of central and angular symmetry. The last chapter contains a simulation study that tries to compare some of the proposed test statistics under different settings and using different test alternatives.

1. Preliminaries

In this section, we will introduce the basics of ranks and rank-based test statistics to be able to define more complex concepts of center-outward ranks and signs for multivariate data. These concepts generalize the ideas from the univariate case and using some additional theory we construct statistical instruments such as ranks, signs, and quantile functions. First, we need to introduce a notation for the basic tools we use in the following chapters. Most of the elementary concepts in the first two chapters are based on the work Hallin et al. (2021). Therefore, we use the same notation.

Let us use μ_d as a denote for the Lebesgue measure over $\mathbb{R}^d$ with its Borel σ-field $\mathcal{B}_d$. Let $\mathcal{P}_d$ denote the family of Lebesgue absolutely continuous distributions over $(\mathbb{R}_d, \mathcal{B}_d)$ and $\mathcal{F}_d$ denote the corresponding family of densities. We will consider $\mathcal{B}_d^n$ as the n-fold product $\mathcal{B}_d \times \cdots \times \mathcal{B}_d$, and let $P^{(n)}$ or $P_f^{(n)}$ be the distribution of an i.i.d. n-tuple with marginals $P = P_f \in \mathcal{P}_d$ and $\mathcal{P}_d^{(n)}$ be the corresponding collection $\{P_f^{(n)}, f \in \mathcal{F}_d\}$. Using $\mathcal{P}_d^{(n)}$-a.s. means $P^{(n)}$-a.s. for all $P \in \mathcal{P}_d^{(n)}$. We denote the support of P and its interior by $\overline{\text{spt}}(P)$ and $\text{spt}(P)$, respectively. Some of the most used terms throughout the thesis are the unit sphere and the open and closed unit ball in $\mathbb{R}^d$. Therefore, we present their definitions.

Definition 1.1. *The $(d-1)$-dimensional* ***unit sphere*** *is defined by*

$$\mathcal{S}_{d-1} = \{x \in \mathbb{R}^d \mid \|x\| = 1\}.$$

In the same way, we define the d-dimensional ***open and closed unit ball*** *by*

$$\mathbb{S}_d = \{x \in \mathbb{R}^d \mid \|x\| < 1\},$$

$$\overline{\mathbb{S}}_d = \{x \in \mathbb{R}^d \mid \|x\| \leq 1\},$$

respectively.

Our main aim is to construct semiparametric tests for multivariate distributions. One of the main advantages of these tests is their distribution-freeness. We would like to transfer this property from the one-dimensional case to the multidimensional one. In the next parts, at first, we revisit the theory behind ranks and signs and the test statistics based on them. Then, we present concepts of center-outward ranks. Finally, the aim of this thesis is to construct several testing tools based on the previous theory and illustrate the according behavior in a simulation study.

1.1 Ranks, distribution and quantile function in one-dimensional space

In dimension one, concepts of distribution, quantile functions, and ranks are well-known. We will go through them only briefly before getting to their multivariate extensions.

Let us consider F to be the distribution function of a random variable Z with distribution $P \in \mathcal{P}_1$ and let us suppose we have a sample $Z_1^{(n)}, \ldots, Z_n^{(n)}$ from the distribution $P \in \mathcal{P}_1$ which has, with probability one, n distinct values. Then we can define the following

- **the ranks** of $Z_1^{(n)}, \ldots, Z_n^{(n)}$ denoted by $\boldsymbol{R}^{(n)} \coloneqq (R_1^{(n)}, \ldots, R_n^{(n)})^\top$, where

$$R_i^{(n)} = \sum_{j=1}^{n} \mathbb{1}[Z_j^{(n)} \leq Z_i^{(n)}],$$

- **the order statistic** is denoted by $\boldsymbol{Z}_{(\cdot)}^{(n)} \coloneqq \left(Z_{(1)}^{(n)}, \ldots, Z_{(n)}^{(n)}\right)^\top$, where

$$Z^{(n)}_{\left(R_i^{(n)}\right)} = Z_i^{(n)}, i = 1, \ldots, n,$$

- **the empirical distribution function** at $Z_i^{(n)}$ is given by

$$F^{(n)}(Z_i^{(n)}) = \frac{R_i^{(n)}}{n+1} = \frac{1}{n+1} \sum_{j=1}^{n} \mathbb{1}[Z_j^{(n)} \leq Z_i^{(n)}].$$

The value $n+1$ in the denominator of the empirical distribution function is used so that $F^{(n)}(Z_i^{(n)})$ takes values in the open interval (0,1) instead of the closed one $[0, 1]$. The mapping $Z_i^{(n)} \mapsto R_i^{(n)}/(n+1)$ is non-decreasing with the domain given by the sample and the range in a regular grid

$$\left\{\frac{1}{n+1}, \ldots, \frac{n}{n+1}\right\}.$$

The empirical distribution function $F^{(n)}$ as a function over $\mathbb{R}$ is then given by

$$F^{(n)}(t) = \frac{1}{n+1} \sum_{j=1}^{n} \mathbb{1}[Z_j^{(n)} \leq t], \quad t \in \mathbb{R}. \tag{1.1}$$

Glivenko-Cantelli's result connects empirical distribution functions to their population versions. Here, we present its well-known formulation, and we try to provide similar results for the center-outward version in the following chapters.

Theorem 1.1 (Cantelli-Glivenko)**.** *Let P be a probability measure on Borel sets of the real line $\mathbb{R}$ with the distribution function F. Let $X_1, \ldots, X_n$ be i.i.d. random variables with the distribution function F. Let F_n be the empirical distribution function computed as $F_n(x) = \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}[X_i \leq x]$. Consequently, it almost surely holds that*

$$\sup_{x \in \mathbb{R}} |F_n(x) - F(x)| \to 0 \text{ as } n \to \infty.$$

Proof. The formulation of the Cantelli-Glivenko theorem and its proof can be found in Dudley (2014) Theorem 1.3.

□

From the Cantelli-Glivenko theorem, we can conclude

$$\max_{1\leq i\leq n}\left|F_n(Z_i^{(n)}) - F(Z_i^{(n)})\right| \to 0 \text{ a.s. as } n\to\infty.$$

1.2 Rank-based tests and their main properties

In this part, we present rank-based tests and their properties. The following description of the problem is based on Section 1.1 in Hallin et al. (2021).

Let us consider a semiparametric model of some real-valued observation

$$\boldsymbol{X} = (X_1,\ldots,X_n)^\top, n\in\mathbb{N}$$

with distribution $\mathsf{P}^{(n)}_{\boldsymbol{\theta},f}$, which depends on finite-dimensional parameter $\boldsymbol{\theta}\in\boldsymbol{\Theta}$, and on density $f\in\mathcal{F}_1$ (where $\mathcal{F}_1$ is the family of Lebesque densities over $\mathbb{R}$) of $Z_i(\boldsymbol{\theta})$ being some unobserved univariate noise. We write $\boldsymbol{X}\sim\mathsf{P}^{(n)}_{\boldsymbol{\theta},f}$ if and only if $Z_1(\boldsymbol{\theta}),\ldots,Z_n(\boldsymbol{\theta})$ are independent and identically distributed with density f. We call $\boldsymbol{Z^{(n)}}(\boldsymbol{\theta}) \coloneqq (Z_1(\boldsymbol{\theta}),\ldots,Z_n(\boldsymbol{\theta}))$ $\boldsymbol{\theta}$-residuals. In the described model, the problem of testing the null hypothesis $H_0^{(n)}:\boldsymbol{\theta}=\boldsymbol{\theta}_0$ can be transformed to testing whether $Z_1(\boldsymbol{\theta}_0),\ldots,Z_n(\boldsymbol{\theta}_0)$ are independent and identically distributed with unspecified density $f\in\mathcal{F}_1$. For that, rank tests based on $\boldsymbol{R}^{(n)}(\boldsymbol{\theta}_0) \coloneqq (R_1^{(n)}(\boldsymbol{\theta}_0),\ldots,R_n^{(n)}(\boldsymbol{\theta}_0))^\top$ ranks of $Z_1(\boldsymbol{\theta}_0),\ldots,Z_n(\boldsymbol{\theta}_0)$ are available. These tests are *distribution-free* because, under the null hypothesis, the distribution of the ranks is uniform over $n!$ permutations of the set $\{1,\ldots,n\}$, regardless of the density $f\in\mathcal{F}_1$. This property is one of the main advantages of univariate rank-based tests.

We illustrate the above presented concepts in a univariate case.

Consider $X_1,\ldots,X_n$ independent random variables such that X_i has a density $f(x-c_i\theta)$, where $c_i, i=1,\ldots,n$ are known constants and θ is an unknown parameter. If θ_0 is the true value of the parameter, then random variables

$$Z_i(\theta_0) = X_i - c_i\theta_0$$

are independent and identically distributed with density f. The density f is unspecified, and the model is, therefore, semiparametric.

Without loss of generality, we assume $\theta_0=0$. We want to test the null hypothesis

$$H_0:\theta=0 \text{ versus } H_1:\theta>0.$$

Let $R_i^{(n)}$ denote ranks of $Z_i \coloneqq Z_i(0)$. Then based on Hájek et al. (1999) or Anděl (2007), a suitable test statistic takes form

$$S=\sum_{i=1}^n c_i\, a(R_i^{(n)}), \tag{1.2}$$

where $a(i) = J\left(\frac{i}{n+1}\right), i = 1, \ldots, n$ and $J : (0,1) \to \mathbb{R}$ is a function called *score function*, and numbers $c_i, i = 1, \ldots, n$ are referred to as *regression coefficients*. The statistic S is called *simple linear rank statistic*.

Same as in Section 2.2.4 in Hájek et al. (1999), let us denote function $\varphi(u, f)$ of density f corresponding to distribution function F by

$$\varphi(u, f) = -\frac{f'(F^{-1}(u))}{f(F^{-1}(u))} \quad , 0 < u < 1.$$

Next, we define scores in the same way as Hájek et al. (1999) in Section 3.4.3.

Definition 1.2. *Let us consider $U_{(1)}^{(n)}, \ldots, U_{(n)}^{(n)}$ an ordered sample from the uniform distribution on $[0, 1]$. Then the numbers*

$$a_n(i, f) = \mathsf{E}\, \varphi(U_{(i)}^{(n)}, f), \quad 1 \leq i \leq n,$$

*are called **scores**, corresponding to the density f.*

If $X_{(1)}^{(n)}, \ldots, X_{(n)}^{(n)}$ is an ordered sample from the distribution with density f, we can rewrite $a_n(i, f)$ as

$$a_n(i, f) = \mathsf{E}\left\{-\frac{f'(X_{(i)}^{(n)})}{f(X_{(i)}^{(n)})}\right\},$$

and the test statistic (1.2) with J chosen as φ takes form

$$S = \sum_{i=1}^{n} c_i\, a_n(R_i^{(n)}, f). \tag{1.3}$$

It is usually difficult to compute $a_n(i, f)$. Therefore, we can replace the scores with the approximation

$$a_n^*(i, f) = \varphi(\mathsf{E}\, U_{(i)}^{(n)}, f) = \varphi\left(\frac{i}{n+1}, f\right).$$

Consequently, we get a test statistic

$$S^* = \sum_{i=1}^{n} c_i a_n^*(R_i^{(n)}(\theta), f).$$

It is possible to show (Theorem a and Lemma a in Section V.1.6 in Hájek & Šidák (1967)) that S^* and S are asymptotically equivalent.

From Theorem 1 in Section 3.4.6 in Hájek et al. (1999), we have the following theorem.

Theorem 1.2. *If f is absolutely continuous and*

$$\int_{-\infty}^{\infty} |f'(x)|\, \mathrm{d}x < \infty$$

holds, then the test with the critical region

$$\sum_{i=1}^{n} c_i a_n(R_i^{(n)}, f) \geq k$$

is the locally most powerful rank test at the given level of significance for

$$H_0 : \theta = 0 \text{ } \textit{versus} \text{ } H_1 : \theta > 0.$$

The level of significance mentioned in the previous theorem is

$$\alpha = \mathrm{P}\left(\sum_{i=1}^{n} c_i a_n(R_i^{(n)}, f) \geq k\right),$$

where probability is computed under the null hypothesis H_0.

We provide the following example to illustrate the above-presented construction of θ-residuals and the test statistic. It is based on Section 11.2.1. in Anděl (2007).

Example 1.1 (Two-sample test of location). Let us consider a random sample $X_1, \ldots, X_{n_1}$ from a continuous distribution with distribution function F_1 and density f_1. Also, consider another independent random sample $Y_1, \ldots, Y_{n_2}$ from a continuous distribution with distribution function F_2 and density f_2. We want to test whether the two densities coincide. We assume that under the alternative, there is a difference in location, i.e., that $f_1(x) = f(x - \theta)$ and $f_2(x) = f(x)$ for a particular density f. The null hypothesis and the alternative are

$$H_0 : \theta = 0 \text{ versus } H_1 : \theta > 0.$$

This is a special case of the previous problem with the following regression coefficients

$$c_i = \begin{cases} 1, & \text{if } i = 1, \ldots, n_1, \\ 0, & \text{if } i = n_1 + 1, \ldots, n_1 + n_2. \end{cases}$$

Let us consider a sample $Z_1(\theta), \ldots, Z_n(\theta), n = n_1 + n_2$, which under the null hypothesis H_0 can be computed as follows:

$$Z_i \coloneqq Z_i(0) = \begin{cases} X_i, & \text{if } i = 1, \ldots, n_1, \\ Y_{i-n_1}, & \text{if } i = n_1 + 1, \ldots, n. \end{cases} \tag{1.4}$$

The test statistic is of the form

$$S = \sum_{i=1}^{n} c_i \, a_n(R_i^{(n)}, f) = \sum_{i=1}^{n_1} a_n(R_i^{(n)}, f),$$

where $R_i^{(n)}$ are ranks of $Z_i, i = 1, \ldots, n$, i.e., ranks computed for the joint sample.

For the special case of a logistic distribution, we have

$$f(x) = \frac{e^{-x}}{(1+e^{-x})^2}$$

and

$$F(x) = \frac{1}{1+e^{-x}}.$$

Then

$$f'(x) = \frac{-e^{-x}(1+e^{-x}) - e^{-x}2(1+e^{-x})(-e^{-x})}{(1+e^{-x})^4} = \\ = \frac{-e^{-x} + 2e^{-2x}}{(1+e^{-x})^3} = -\frac{e^{-x}(1-e^{-x})}{(1+e^{-x})^3}$$

and

$$F^{-1}(u) = \ln \frac{u}{1-u}.$$

From that we can conclude $\varphi(u, f) = 2u - 1$ so

$$\mathsf{E}\ \varphi(U_i^{(n)}, f) = \frac{2i}{n+1} - 1.$$

The corresponding test statistic is

$$S = \sum_{i=1}^{n_1} \left(\frac{2R_i^{(n)}}{n+1} - 1 \right) = \frac{2}{n+1} \sum_{i=1}^{n_1} R_i^{(n)} - n_1.$$

This test statistic is a linear transformation of the two-sample Wilcoxon test statistic

$$W = \sum_{i=1}^{n_1} R_i^{(n)},$$

so the tests based on these two statistics are equivalent, meaning the null hypothesis is rejected by one test if and only if it is rejected by the other. Therefore, based on Theorem 1.2, we can conclude that the two-sample Wilcoxon test is the locally most powerful rank test in the case of the logistic distribution.

In the same way, for standard normal distribution, we get

$$f(x) = \frac{1}{\sqrt{2\pi}} e^{-x^2/2}, \quad F(x) = \Phi(x),$$

and

$$f'(x) = -xf(x), \quad \varphi(u, f) = \Phi^{-1}(u).$$

We use the approximated scores

$$a_N^*(i, f) = \Phi^{-1}\left(\frac{i}{N+1} \right).$$

The corresponding test statistic is

$$S = \sum_{i=1}^{n_1} \Phi^{-1}\left(\frac{R_i^{(n)}}{n+1}\right).$$

Therefore, the van der Waerden test is the asymptotically locally most powerful rank test in the case of normal distribution. We talk about the "asymptotically" locally most powerful rank test because of the usage of the approximated scores.

The asymptotic normality of S from (1.3) is given by Theorem a in Section V.1.5 in Hájek & Šidák (1967).

Theorem 1.3. *Consider scores $a_n(i, f)$ from Definition 1.2 associated with an integrable function $\varphi(u, f)$. We denote*

$$\overline{\varphi} = \int_0^1 \varphi(u, f)\mathrm{d}u$$

and we assume

$$\int_0^1 (\varphi(u, f) - \overline{\varphi})^2 \mathrm{d}u > 0.$$

Under the null hypothesis H_0 and the following condition for regression coefficients

$$\frac{\sum_{i=1}^n (c_i - \overline{c})^2}{\max_{1\le i\le n}(c_i - \overline{c})^2} \to \infty \quad \text{as } n \to \infty,$$

where $\overline{c} \coloneqq n^{-1}\sum_{i=1}^n c_i$, the test statistic

$$S = \sum_{i=1}^n c_i\, a_n(R_i^{(n)}, f)$$

is asymptotically normal $\mathcal{N}(\mu_c, \sigma_c^2)$ with mean

$$\mu_c = \overline{c}\sum_{i=1}^n a_n(i, f)$$

and variance

$$\sigma_c^2 = \left(\sum_{i=1}^n (c_i - \overline{c})^2\right) \int_0^1 (\varphi(u, f) - \overline{\varphi})^2 \,\mathrm{d}u.$$

The statement that the test statistic S is asymptotically normal $\mathcal{N}(\mu_c, \sigma_c^2)$ means that

$$\frac{S - \mu_c}{\sigma_c} \xrightarrow{d} \mathcal{N}(0, 1), \text{ as } n \to \infty.$$

Example 1.2 (Two-sample test). Let us suppose two independent random samples $X_1, \ldots, X_{n_1}$ and $Y_1, \ldots, Y_{n_2}$ with a shift in location θ which is the same as in Example 1.1. We want to test the null hypothesis

$$H_0 : \theta = 0 \text{ versus } H_1 : \theta > 0.$$

The test statistic could then be

$$S = \sum_{i=1}^{n_1} \left(\frac{2R_i^{(n)}}{n+1} - 1 \right) = \frac{2}{n+1} \sum_{i=1}^{n_1} R_i^{(n)} - n_1,$$

where $R_i^{(n)}$ is the rank of X_i in the joint random sample $X_1, \ldots, X_{n_1}, Y_1, \ldots, Y_{n_2}$. For the test statistic S, it holds due to Theorem 1.3

$$\frac{S - \mu_c}{\sigma_c} \xrightarrow{d} \mathcal{N}(0,1).$$

We compute μ_c in the same way as in Theorem 1.3

$$\mu_c = \overline{c} \sum_{i=1}^{n} a_n(i,f) = \frac{n_1}{n} \left(\sum_{i=1}^{n} \frac{2i}{n+1} - n \right) = \frac{n_1}{n} \frac{2}{n+1} \frac{n(n+1)}{2} - n_1 = 0.$$

It holds

$$\sum_{i=1}^{n} (c_i - \overline{c})^2 = \sum_{i=1}^{n_1} \left(1 - \frac{n_1}{n}\right)^2 + \sum_{i=n_1+1}^{n} \left(-\frac{n_1}{n}\right)^2 = n_1 \left(\frac{n_2}{n}\right)^2 + n_2 \left(\frac{n_1}{n}\right)^2 = \frac{n_1 n_2}{n},$$

$$\overline{\varphi} = \int_0^1 \varphi(u,f) \mathrm{d}u = \int_0^1 2u - 1 \mathrm{d}u = 0,$$

and it also holds

$$\int_0^1 (\varphi(u,f) - \overline{\varphi})^2 \mathrm{d}u = \int_0^1 (\varphi(u,f))^2 \mathrm{d}u = \int_0^1 (2u-1)^2 \mathrm{d}u = \int_0^1 4u^2 - 4u + 1 \mathrm{d}u =$$
$$= 4\frac{1}{3} - 4\frac{1}{2} + 1 = \frac{1}{3}.$$

It can be easily derived that the asymptotic variance under the null hypothesis is

$$\sigma_c^2 = \frac{n_1 n_2}{3n}.$$

From this and the previous theorem, we get

$$\frac{\frac{2}{n+1} \sum_{i=1}^{n_1} R_i^{(n)} - n_1}{\sqrt{\frac{n_1 n_2}{3n}}} \xrightarrow{d} \mathcal{N}(0,1),$$

i.e.,

$$\frac{W - \frac{n_1(n+1)}{2}}{\sqrt{\frac{n_1 n_2 (n+1)^2}{12n}}} \xrightarrow{d} \mathcal{N}(0,1),$$

where W is the Wilcoxon test statistic $W = \sum_{i=1}^{n_1} R_i^{(n)}$.

On the other hand, from Theorem c in Hájek et al. (1999), we also obtain

$$\mathsf{E}\, S = \overline{a} \sum_{i=1}^{n} c_i, \quad \mathrm{var} S = \sigma_a^2 \sum_{i=1}^{n} (c_i - \overline{c})^2,$$

where

$$\overline{a} = \frac{1}{n} \sum_{i=1}^{n} a(i) = \frac{1}{n} \left(\sum_{i=1}^{n} \frac{2i}{n+1} - n \right) = \frac{1}{n} \frac{2n(n+1)}{2(n+1)} - 1 = 0,$$

$$\overline{c} = \frac{1}{n}\sum_{i=1}^{n} c_i = \frac{n_1}{n}.$$

Therefore, $\mathsf{E}\, S = 0$ equals μ_c from the previous computations. Moreover,

$$\begin{aligned}
\sigma_a^2 &= \frac{1}{n-1}\sum_{i=1}^{n}(a(i)-\overline{a})^2 = \frac{1}{n-1}\sum_{i=1}^{n}(a(i))^2 = \frac{1}{n-1}\sum_{i=1}^{n}\left(\frac{2i}{n+1}-1\right)^2 = \\
&= \frac{1}{n-1}\left(\sum_{i=1}^{n}\frac{4i^2}{(n+1)^2} - \sum_{i=1}^{n}\frac{4i}{n+1} + n\right) \\
&= \frac{1}{n-1}\left(\frac{4n(n+1)(2n+1)}{6(n+1)^2} - \frac{4n(n+1)}{2(n+1)} + n\right) = \\
&= \frac{1}{n-1}\left(\frac{2n(2n+1)}{3(n+1)} - n\right) = \frac{1}{n-1}\left(\frac{2n(2n+1)-3n(n+1)}{3(n+1)}\right) = \\
&= \frac{1}{n-1}\frac{n^2-n}{3(n+1)} = \frac{n}{3(n+1)}.
\end{aligned}$$

From that can be concluded

$$\operatorname{var} S = \frac{n}{3(n+1)}\frac{n_1 n_2}{n} = \frac{n_1 n_2}{3(n+1)},$$

which behaves asymptotically, the same way as σ_c^2. Then from $\mathsf{E}\, S$ we get

$$0 = \frac{2}{n+1}\,\mathsf{E}\, W - n_1,$$

so $\mathsf{E}\, W = \frac{n_1(n+1)}{2}$, same as in the previous part, and from $\operatorname{var} S$ we get

$$\frac{n_1 n_2}{3(n+1)} = \frac{4}{(n+1)^2}\operatorname{var} W,$$

so $\operatorname{var} W = \frac{n_1 n_2 (n+1)}{12}$ which is asymptotically equivalent to $\frac{n_1 n_2 (n+1)^2}{12n}$ from the previous computations.

2. Center-outward ranks

In this chapter, we first outline the idea of center-outward ranks and signs for the univariate case, and then we move to the d-dimensional generalization. Most of the concepts and notation are taken from the articles Hallin et al. (2021) and Hallin (2022).

2.1 Center-outward ranks in one dimension

The ordering of the real line cannot be expected to hold in multidimensional spaces. Therefore, we will present new concepts based on center-outward orientation. We will assume that the distribution function F of a random variable Z is strictly increasing. Accordingly, the corresponding density f is greater than 0, the quantile function is well-defined, and the median and other quantiles are unique.

Definition 2.1. *For a random variable Z with the distribution $P \in \mathcal{P}_1$, we define the **center-outward distribution function** as $\boldsymbol{F}_{\pm} \coloneqq 2F - 1$, where F is the distribution function corresponding to the distribution P.*

From the definition, it is obvious that $\boldsymbol{F}_{\pm}$ is just a linear transformation of the distribution function F. Then $\|\boldsymbol{F}_{\pm}(z)\| = |2F(z) - 1|$ and we define

$$\boldsymbol{S}_{\pm}(z) \coloneqq \mathbb{I}\left[\boldsymbol{F}_{\pm}(z) \neq 0\right] \boldsymbol{F}_{\pm}(z)/\|\boldsymbol{F}_{\pm}(z)\|.$$

Clearly, $\boldsymbol{S}_{\pm}$ is the sign of the deviation from the median or, said otherwise, a point on the unit sphere $\mathcal{S}_0 = \{-1, 1\}$ because for median M, we have

$$\boldsymbol{F}_{\pm}(M) = 2F(M) - 1 = 0.$$

Definition 2.2. *The inverse of $\boldsymbol{F}_{\pm}$ is called the **center-outward quantile function**. We denote it by $\boldsymbol{Q}_{\pm}$. The sets*

$$\{\boldsymbol{Q}_{\pm}(u) \,|\, |u| = p\}$$

and the intervals

$$\{\boldsymbol{Q}_{\pm}(u) \,|\, |u| \leq p\}$$

*are called **quantile contours** and **quantile regions** respectively, with the quantile level*
$0 \leq p < 1$.

The quantile regions are closed, connected, and nested. Next, we will define the empirical version of the center-outward distribution function. For that, let us have a sample $Z_1^{(n)}, \ldots, Z_n^{(n)}$ which has, with probability one, n distinct values. Therefore, let us, without loss of generality, suppose that the values are distinct. There are $\lfloor n/2 \rfloor$ values on the right side of the sample median which for n even is taken as an average of the two middle values. We order them and assign them ranks $R_{\pm,i}^{(n)}$ with the values $1, \ldots, \lfloor n/2 \rfloor$ and the signs $\boldsymbol{S}_{\pm,i}^{(n)} = 1$. In the same way,

we give to $\lfloor n/2 \rfloor$ values on the left side of the sample median the ranks $R^{(n)}_{\pm,i}$ with the values $\lfloor n/2 \rfloor, \ldots, 1$ and the signs $\boldsymbol{S}^{(n)}_{\pm,i} = -1$.

Definition 2.3. *We call $R^{(n)}_{\pm,i}$ and $\boldsymbol{S}^{(n)}_{\pm,i}$ from the previous paragraph the* ***center-outward ranks*** *and the* ***center-outward signs****. We define the* ***empirical center-outward distribution function*** *as*

$$\boldsymbol{F}^{(n)}_{\pm}(Z^{(n)}_i) \coloneqq \boldsymbol{S}^{(n)}_{\pm,i} \frac{R^{(n)}_{\pm,i}}{\lfloor n/2 \rfloor + 1} = \begin{cases} 2F^{(n)}(Z^{(n)}_i) - 1 & n \text{ odd} \\ \dfrac{n+1}{n+2}\left(2F^{(n)}(Z^{(n)}_i) - 1\right) + \dfrac{1}{n+2} & n \text{ even,} \end{cases}$$

where $F^{(n)}$ is the empirical distribution function from (1.1).

The empirical center-outward distribution function takes values on a regular grid which is the intersection of two unit vectors $\pm\mathbf{1}$ and $\lfloor n/2 \rfloor$ circles, with the center at the origin and radii

$$\frac{1}{\lfloor n/2 \rfloor + 1}, \ldots, \frac{\lfloor n/2 \rfloor}{\lfloor n/2 \rfloor + 1}.$$

For n odd, the grid also contains the origin.

Example 2.1. Let us consider samples $Z^{(n)}_1, \ldots, Z^{(n)}_n$ for $n = 7$ and $n = 8$. Then, the empirical center-outward distribution function $\boldsymbol{F}^{(n)}_{\pm}(Z^{(n)}_i), i = 1, \ldots, n$ takes values from the grid in Figure 2.1.

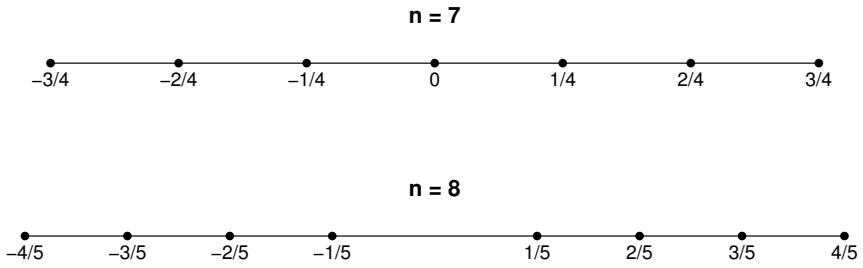

Figure 2.1: A regular grid with values of the empirical center-outward distribution function for $n = 7$ and $n = 8$ in the dimension $d = 1$.

Under the previous assumptions for n even, signs are uniformly distributed over the unit sphere $\mathcal{S}_0$ (for n odd we have the additional sign 0 for the sample median) and independent of the ranks $R^{(n)}_{\pm}$. Ranks are uniformly distributed over the integers $(0, \ldots, \lfloor n/2 \rfloor)$ or $(1, \ldots, \lfloor n/2 \rfloor)$, according to n being odd or even.

Glivenko-Cantelli's result also holds for the empirical center-outward distribution function $\boldsymbol{F}^{(n)}_{\pm}$ yielding

$$\max_{1 \le i \le n} \left\| \boldsymbol{F}^{(n)}_{\pm}(Z^{(n)}_i) - \boldsymbol{F}_{\pm}(Z^{(n)}_i) \right\| \to 0 \text{ a.s. as } n \to \infty.$$

2.2 Center-outward ranks in multi-dimensional space

In this section, we present the definition of the center-outward ranks in multi-dimensional space, here $\mathbb{R}^d$, $d > 1$. For that, we use the concept of measure transportation which is associated with the so-called *Monge-Kantorovich* problem. How should one best move given piles of sand to fill up given holes of the same total volume? That is a very practical question, with which it all started.

2.2.1 Measure transportation

To describe the problem in a formal way, let P_1 and P_2 denote two probability measures over $(\mathbb{R}^d; \mathcal{B}^d)$. Consider a Borel-measurable function $L : \mathbb{R}^{2d} \to [0, 1]$ such that for $x_1, x_2 \in \mathbb{R}^d$ the value $L(x_1, x_2)$ represents the cost of transporting x_1 to x_2. We need to find a measurable mapping $T_{P_1,P_2} : \mathbb{R}^d \to \mathbb{R}^d$ which solves the minimization problem

$$\inf_T \int_{\mathbb{R}^d} L(x, T(x)) \mathrm{d}P_1 \quad \text{subject to} \quad T\#P_1 = P_2, \tag{2.1}$$

where $T\#P_1$ is called a *push forward* of P_1 by T. Explicitly, for any Borel set A, $T\#P_1(A) \coloneqq P_1(T^{-1}(A))$. Equivalently, for $X \sim P_1$, it holds $T(X) \sim P_2$, see Villani (2003). The mapping T_{P_1,P_2} for which the infimum of the problem (2.1) is obtained is called the*optimal transport*.

The main result connected with the presented problematics is *McCann's theorem*, see McCann (1995). It implies that for any given absolutely continuous distributions P_1 and P_2 over $\mathbb{R}^d$, there exists a convex function $\psi : \mathbb{R}^d \to \mathbb{R}$ with almost everywhere gradient $\nabla\psi$ pushing P_1 forward to P_2. The function ψ may not be unique, but $\nabla\psi$ is P_1-a.s. uniquely determined. Under the existence of finite moments of order two, moreover, $\nabla\psi$ is an L^2-optimal (in the Monge-Kantorovich sense) transport pushing P_1 forward to P_2. By L^2-optimal we mean an optimal transport for cost function L given by $L(x_1, x_2) = \|x_1 - x_2\|_2^2$.

2.2.2 Center-outward ranks and signs

In this section, we finally get to the definition of the center-outward distribution function, quantile function, ranks, and signs. All of the following concepts and definitions in this section are taken from Section 2 in Hallin et al. (2021). Let us assume that P has a non-vanishing density over $\mathbb{R}^d$, i.e., density $f \coloneqq \mathrm{d}P/\mathrm{d}\mu_d$ where μ_d is the Lebesgue measure over $\mathbb{R}^d$, such that for all $D > 0$ there exist constants $\lambda^-_{D,P}, \lambda^+_{D,P}$ satisfying

$$0 < \lambda^-_{D,P} \leq f(\boldsymbol{z}) \leq \lambda^+_{D,P} < \infty$$

for all $\boldsymbol{z}$ with the norm $\|\boldsymbol{z}\| \leq D$. We will denote that by $P \in \mathcal{P}^+_d$.

By the uniform distribution over the unit ball we mean the distribution U_d obtained by the product of the uniform measure over the unit sphere and the uniform distribution over the unit interval [0, 1]. Equivalently, X has a uniform

distribution, if and only if it can be written as

$$X \stackrel{d}{=} R\,U,$$

where R is uniformly distributed over [0,1], U is uniformly distributed over the unit sphere, and they are independent. Then we can define the center-outward distribution function as follows.

Definition 2.4. *We define* ***the center-outward distribution function*** $\boldsymbol{F}_{\pm}$ *as the unique gradient of a convex function* ψ*, where the gradient* $\nabla\psi$ *is pushing* P *forward to the uniform distribution over the unit ball* $\mathbb{S}_d$*.*

To define the center-outward quantile function, we use the result given by Figalli (2018) which implies that the center-outward distribution function $\boldsymbol{F}_{\pm}$ is a homeomorphism from the punctured unit ball $\mathbb{S}_d \setminus \{0\}$ onto $\mathbb{R}^d \setminus \boldsymbol{F}_{\pm}^{-1}(\{0\})$. Therefore, $\boldsymbol{F}_{\pm}$ has a well-defined homeomorphic inverse over these domains. We denote it by $\boldsymbol{Q}_{\pm}$ and extend it to the whole unit ball by defining $\boldsymbol{Q}_{\pm}(0) = \boldsymbol{F}_{\pm}^{-1}(\{0\})$.

Definition 2.5. *The homeomorphic inverse* $\boldsymbol{Q}_{\pm}$ *of the center-outward distribution function* $\boldsymbol{F}_{\pm}$ *extended by letting* $\boldsymbol{Q}_{\pm}(0) = \boldsymbol{F}_{\pm}^{-1}(\{0\})$ *is called* ***the center-outward quantile function*** *and* $\boldsymbol{Q}_{\pm}(0)$ *is called* ***the center-outward median****. For* $q \in (0,1)$*, we define* ***the center-outward quantile region*** *as*

$$\mathbb{C}(q) \coloneqq \boldsymbol{Q}_{\pm}(q\,\overline{\mathbb{S}}_d) = \{\boldsymbol{z} \mid \|\boldsymbol{F}_{\pm}(\boldsymbol{z})\| \leq q\}.$$

We denote by

$$\mathcal{C}(q) \coloneqq \boldsymbol{Q}_{\pm}(q\,\mathcal{S}_{d-1}) = \{\boldsymbol{z} \mid \|\boldsymbol{F}_{\pm}(\boldsymbol{z})\| = q\}$$

the center-outward quantile contour *of a given order* q*.*

From the definition of the center-outward distribution function, we move to its empirical version. For that, let $\boldsymbol{Z}^{(n)} \coloneqq (\boldsymbol{Z}_1^{(n)}, \ldots, \boldsymbol{Z}_n^{(n)})$ denote an n-tuple of random vectors. These might be observations or residuals associated with a parameter $\boldsymbol{\theta}$ of interest, see Section 1.2. Let us suppose that $\boldsymbol{Z}_i^{(n)}$'s are i.i.d. with density $f \in \mathcal{F}_d$, distribution P, and the center-outward distribution function $\boldsymbol{F}_{\pm}$.

The extension of the definition of the empirical distribution function $\boldsymbol{F}_{\pm}^{(n)}$ of $\boldsymbol{F}_{\pm}$ from univariate concepts into its multivariate version is connected with a transformation of the data into a grid. Assuming $d \geq 2$, we factorize n into

$$n = n_R n_S + n_0, \quad n_R, n_S, n_0 \in \mathbb{N}, \quad 0 \leq n_0 < \min(n_R, n_S), \tag{2.2}$$

where $n_R \to \infty$ and $n_S \to \infty$ as $n \to \infty$ (implying $n_0/n \to 0$). Using this factorization, we create a "regular" grid. The grid is formed by $n_R n_S$ points in the unit ball $\mathbb{S}_{d-1}$ arising as the intersection between an n_S-tuple $(\boldsymbol{u}_1, \ldots, \boldsymbol{u}_{n_S})$ of unit vectors of $\mathcal{S}_d$ and n_R hyperspheres with the center at $\mathbf{0}$ and radii

$$j/(n_R + 1), j = 1, \ldots, n_R,$$

together with n_0 copies of the origin. The n_S-tuple should be as uniform as possible. For dimension 2, the uniformity can be obtained by dividing the unit

circle into n_S arcs of equal length $2\pi/n_S$. In higher dimensions, this problem becomes more complicated. In the simulation study in Chapter 4, we investigate also other types of grids in different forms. More information about some possible constructions of grids is provided in Section 2.5.

For the resulting grid, we form a discrete distribution with probability masses $1/n$ at each of the $n_R n_S$ non-zero points and a probability mass n_0/n at the origin. This distribution converges weakly to the uniform U_d over the ball $\mathbb{S}_d$.

Definition 2.6. *We define* ***the empirical distribution function*** $\boldsymbol{F}_{\pm}^{(n)}$ *as a mapping* $\boldsymbol{F}_{\pm}^{(n)} : (\boldsymbol{Z}_1^{(n)}, \ldots, \boldsymbol{Z}_n^{(n)}) \mapsto (\boldsymbol{F}_{\pm}^{(n)}(\boldsymbol{Z}_1^{(n)}), \ldots, \boldsymbol{F}_{\pm}^{(n)}(\boldsymbol{Z}_n^{(n)}))$ *so that*

$$\sum_{i=1}^{n} \|\boldsymbol{Z}_i^{(n)} - \boldsymbol{F}_{\pm}^{(n)}(\boldsymbol{Z}_i^{(n)})\|^2 = \min_{\pi \in \Pi} \sum_{i=1}^{n} \|\boldsymbol{Z}_{\pi(i)}^{(n)} - \boldsymbol{F}_{\pm}^{(n)}(\boldsymbol{Z}_i^{(n)})\|^2,$$

where Π *is a set of all* $n!$ *possible permutations of the set* $\{1, \ldots, n\}$ *and*

$$\{\boldsymbol{F}_{\pm}^{(n)}(\boldsymbol{Z}_i^{(n)}), i = 1, \ldots, n\}$$

is a set of the n *points from the grid.*

Along with the definition of the empirical distribution function, we also introduce terms connected with it.

Definition 2.7. *We define* ***the center-outward ranks*** *by*

$$R_{\pm,i}^{(n)} \coloneqq (n_R + 1)\|\boldsymbol{F}_{\pm}^{(n)}(\boldsymbol{Z}_i^{(n)})\|.$$

The empirical center-outward quantile contours and regions *are given by*

$$\mathcal{C}_{\pm,\boldsymbol{Z}^{(n)}}^{(n)}\left(\frac{j}{n_R+1}\right) \coloneqq \{\boldsymbol{Z}_i^{(n)} | R_{\pm,i}^{(n)} = j\} \text{ and } \mathbb{C}_{\pm,\boldsymbol{Z}^{(n)}}^{(n)}\left(\frac{j}{n_R+1}\right) \coloneqq \{\boldsymbol{Z}_i^{(n)} | R_{\pm,i}^{(n)} \leq j\}.$$

Last, we define by

$$\boldsymbol{S}_{\pm,i}^{(n)} \coloneqq \boldsymbol{F}_{\pm}^{(n)}(\boldsymbol{Z}_i^{(n)}) \mathbb{I}\left[\boldsymbol{F}_{\pm}^{(n)}(\boldsymbol{Z}_i^{(n)}) \neq 0\right] / \|\boldsymbol{F}_{\pm}^{(n)}(\boldsymbol{Z}_i^{(n)})\|$$

the center-outward signs*.*

The center-outward sign is a d-dimensional vector which can be interpreted also as a direction.

Example 2.2. Let us suppose we have a sample $\boldsymbol{Z}^{(n)} \coloneqq (\boldsymbol{Z}_1^{(n)}, \ldots, \boldsymbol{Z}_n^{(n)}), n = 24$. We factorize n into $n = n_R n_S + n_0$, where $n_R = 4, n_S = 6$, and $n_0 = 0$. Then the described grid might look like Figure 2.2.

The highlighted point of the grid in Figure 2.2 corresponds to the rank $R_{\pm}^{(n)} = 4$ and the sign

$$\boldsymbol{S}_{\pm}^{(n)} = \begin{pmatrix} \cos(2\pi/3) \\ \sin(2\pi/3) \end{pmatrix}.$$

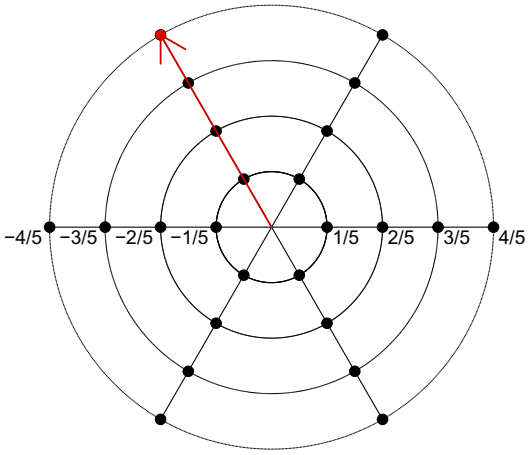

Figure 2.2: A regular grid with values of the empirical center-outward distribution function in dimension 2 for $n_R = 4$ and $n_S = 6$.

The next concept to present is an order statistic. In this case, it is not clear how to order the n-tuple of $\boldsymbol{Z}_i^{(n)}$ values. We fix the notation as follows. Let the order statistic $\boldsymbol{Z}_{(\cdot)}^{(n)}$ be given by the following expression

$$\boldsymbol{Z}_{(\cdot)}^{(n)} = (\boldsymbol{Z}_{(1)}^{(n)}, \ldots, \boldsymbol{Z}_{(n)}^{(n)}),$$

where the first component of $\boldsymbol{Z}_{(i)}^{(n)}$ is the ith order statistic of the n-tuple of the first components of values $\boldsymbol{Z}_{(i)}^{(n)}$.

2.3 Properties of the center-outward ranks

In this section, we present the main properties of the center-outward ranks. The propositions and also their proofs can be found in Section 2 in Hallin et al. (2021). We also present additional examples to illustrate the stated properties.

The following proposition summarizes the properties of the center-outward distribution function and the corresponding quantile function. It is taken from Propositions 2.1 and 2.2 in Hallin et al. (2021).

Proposition 2.1. *For a random variable $\boldsymbol{Z}$ with the distribution $P \in \mathcal{P}_d$, denote by $\boldsymbol{F}_\pm$ the center-outward distribution function of P and by $\boldsymbol{Q}_\pm$ the corresponding center-outward quantile function. Then*

1. *$\boldsymbol{F}_\pm$ takes values in $\overline{\mathcal{S}}_{d-1}$ and $\boldsymbol{F}_\pm \# P = U_d$. $\boldsymbol{F}_\pm$, thus, is a probability-integral transformation,*

2. *$\|\boldsymbol{F}_\pm(\boldsymbol{Z})\|$ is uniform over $[0,1]$, $\boldsymbol{S}(\boldsymbol{Z}) := \boldsymbol{F}_\pm(\boldsymbol{Z})/\|\boldsymbol{F}_\pm(\boldsymbol{Z})\|$ is uniform over $\mathcal{S}_{d-1}$, and they are mutually independent,*

3. *$\boldsymbol{F}_\pm$ entirely characterizes P,*

4. *for $d = 1$, $\boldsymbol{F}_\pm$ coincides with $2F-1$ (F the traditional distribution function),*

5. *$\boldsymbol{Q}_\pm$ pushes U_d forward to P, hence entirely characterizes P,*

6. *the center-outward quantile region $\mathbb{C}(q), 0 < q < 1$, has P-probability content q,*

7. *$\boldsymbol{Q}_\pm(u)$ coincides for $d = 1$ with $\inf\{x|F(x) \geq (1+u) = 2)\}, u \in (-1,1)$, and $\mathbb{C}(q), q \in (0,1)$ with*

$$[\inf\{x|F(x) \geq (1-q)/2\}, \inf\{x|F(x) \geq (1+q)/2\}] \bigcap \overline{\text{spt}}(\text{P}),$$

F meaning the traditional distribution function.

Properties 1 and 2 follow immediately from the construction of $\boldsymbol{F}_\pm$.

Also, the equivalent of the Glivenko-Cantelli theorem holds for the center-outward distribution function. The proposition can be found in Hallin et al. (2021) (Proposition 2.4) and the proof for an even more general version which extends this proposition under sup form to cyclically monotone interpolations of $\boldsymbol{F}_\pm^{(n)}$ is given there (Proposition 3.3).

Proposition 2.2. *Let $\boldsymbol{Z}_1^{(n)}, \ldots, \boldsymbol{Z}_n^{(n)}$ be i.i.d. with distribution $P \in \mathcal{P}_d^+$. Then,*

$$\max_{1 \leq i \leq n} \left\| \boldsymbol{F}_\pm^{(n)}(\boldsymbol{Z}_i^{(n)}) - \boldsymbol{F}_\pm(\boldsymbol{Z}_i^{(n)}) \right\| \to 0 \quad \textit{a.s. as } n \to \infty.$$

Other, no less important, properties are distribution-freeness and maximal ancillarity. These are the properties that make rank-based tests useful. Their validity in one-dimensional spaces has already been described in Chapter 1. Here we present their extension to multidimensional spaces for ranks and order statistics based on the center-outward distribution function, see Proposition 2.5 in Hallin et al. (2021).

Proposition 2.3. *Let $\boldsymbol{Z}_1^{(n)}, \ldots, \boldsymbol{Z}_n^{(n)}$ be i.i.d. with distribution $P \in \mathcal{P}_d$, center-outward distribution function $\boldsymbol{F}_\pm$, order statistic $\boldsymbol{Z}_{(\cdot)}^{(n)}$, and empirical center-outward distribution function $\boldsymbol{F}_\pm^{(n)}$. Then,*

1. *$\boldsymbol{Z}_{(\cdot)}^{(n)}$ is sufficient and complete, hence minimal sufficient, for $\mathcal{P}_d^{(n)}$,*

2. *$\boldsymbol{F}_\pm^{(n)} = (\boldsymbol{F}_\pm^{(n)}(\boldsymbol{Z}_1^{(n)}), \ldots, \boldsymbol{F}_\pm^{(n)}(\boldsymbol{Z}_n^{(n)}))$ is uniformly distributed over the $n!/n_0!$ permutations with repetitions (the origin counted as n_0 indistinguishable points) of the grid described in Section 2.2,*

3. *for $n_0 = 0$, the vectors of center-outward ranks $(R_{\pm,1}^{(n)}, \ldots, R_{\pm,n}^{(n)})$ and signs $(\boldsymbol{S}_{\pm,1}^{(n)}, \ldots, \boldsymbol{S}_{\pm,n}^{(n)})$ are mutually independent. For $n_0 > 0$, the same independence holds for the $(n_R n_S)$-tuple of ranks and signs associated with the (random) set $\{i_1, \ldots, i_{n_R n_S}\}$, such that $\boldsymbol{F}_\pm^{(n)}(\boldsymbol{Z}_{i_j}^{(n)})$,*

4. *for all $P \in \mathcal{P}_d$, $\boldsymbol{Z}_{(\cdot)}^{(n)}$ and $\boldsymbol{F}_\pm^{(n)}(\boldsymbol{Z}_{i_j}^{(n)})$ are mutually P-independent.*

Another property of the center-outward ranks and signs is their invariance with respect to the change of location. Their properties under orthogonal transfor-

mation are summarized in the following proposition. Let $\boldsymbol{F}_{\pm}^{\boldsymbol{Z}}$ denote the center-outward distribution function of a random variable $\boldsymbol{Z}$. The proposition and its proof can be found in Hallin, Hlubinka & Hudecová (2022) (Proposition 2.2, proof in Appendix A.1).

Proposition 2.4. *Let $\boldsymbol{\mu} \in \mathbb{R}^d$ and let $\boldsymbol{O}$ be a $d \times d$ orthogonal matrix. Then,*

1. $\boldsymbol{F}_{\pm}^{\boldsymbol{\mu}+\boldsymbol{O}\boldsymbol{Z}}(\boldsymbol{\mu}+\boldsymbol{O}\boldsymbol{z}) = \boldsymbol{O}\boldsymbol{F}_{\pm}^{\boldsymbol{Z}}(\boldsymbol{z}), \boldsymbol{z} \in \mathbb{R}^d$,

2. *denoting by $\boldsymbol{F}_{\pm}^{\boldsymbol{\mu}+\boldsymbol{O}\boldsymbol{Z},(n)}$ the empirical distribution function of the sample $\boldsymbol{\mu}+\boldsymbol{O}\boldsymbol{Z}_1, \ldots, \boldsymbol{\mu}+\boldsymbol{O}\boldsymbol{Z}_n$, associated with the grid $\boldsymbol{O}\mathcal{G}_n$, analogously by $\boldsymbol{F}_{\pm}^{\boldsymbol{Z},(n)}$ the empirical distribution function of the sample $\boldsymbol{Z}_1, \ldots, \boldsymbol{Z}_n$, associated with the grid $\mathcal{G}_n$, then it holds*

$$\boldsymbol{F}_{\pm}^{\boldsymbol{\mu}+\boldsymbol{O}\boldsymbol{Z},(n)}(\boldsymbol{\mu}+\boldsymbol{O}\boldsymbol{Z}_i) = \boldsymbol{O}\boldsymbol{F}_{\pm}^{\boldsymbol{Z},(n)}(\boldsymbol{Z}_i), \quad i = 1, \ldots, n,$$

3. *the center-outward ranks $R_{\pm,i}^{(n)}$ and the angles between $\boldsymbol{S}_{\pm,i}^{(n)}$ and $\boldsymbol{S}_{\pm,j}^{(n)}, i, j = 1, \ldots, n$ computed from the sample $\boldsymbol{Z}_1, \ldots, \boldsymbol{Z}_n$ and the grid $\mathcal{G}_n$ are the same as those computed from the sample $\boldsymbol{\mu}+\boldsymbol{O}\boldsymbol{Z}_1, \ldots, \boldsymbol{\mu}+\boldsymbol{O}\boldsymbol{Z}_n$ and the grid $\boldsymbol{O}\mathcal{G}_n$.*

Example 2.3. Let us assume a random sample $\boldsymbol{Z}_1, \ldots, \boldsymbol{Z}_n, n = 100$ from a 2-dimensional normal distribution $\mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, where $\boldsymbol{\mu} = (0,0)^\top$ and

$$\boldsymbol{\Sigma} = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}.$$

We factorize n into $n = n_R n_S + n_0$, where $n_R = 10, n_S = 10$ and $n_0 = 0$, and again create a regular grid $\mathcal{G}_n$ with the values of the empirical center-outward distribution function, see Figure 2.3. In Figure 2.3, we randomly chose one observation and highlighted it in the plot corresponding to the sample and in the grid $\mathcal{G}_n$ of the empirical center-outward distribution function $\boldsymbol{F}_{\pm}$.

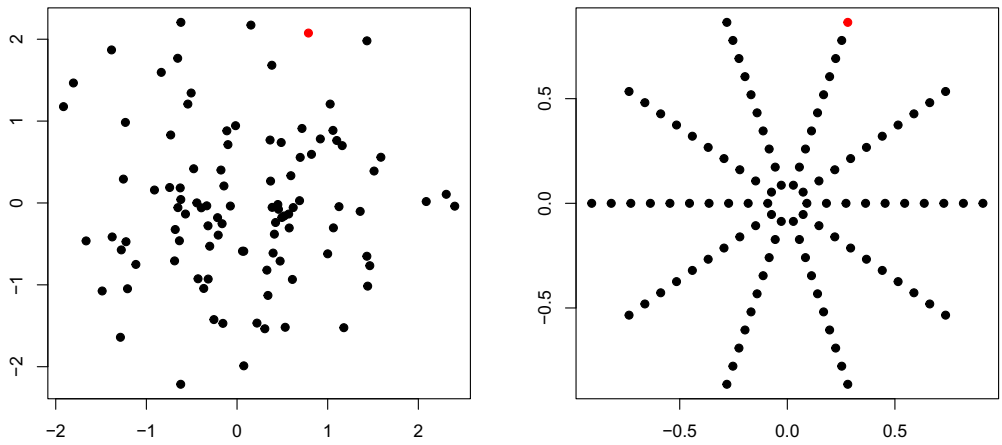

Figure 2.3: A random sample from a 2-dimensional normal distribution and the corresponding regular grid with the values of the empirical center-outward distribution function.

The highlighted point of the grid corresponds to the rank $R_{\pm}^{(n)} = 10$ and the sign

$$\boldsymbol{S}_{\pm}^{(n)} = \begin{pmatrix} \cos(2\pi/5) \\ \sin(2\pi/5) \end{pmatrix}.$$

We want to demonstrate Proposition 2.4. Therefore, we choose $\boldsymbol{\mu} = (3,3)^\top$ and an orthogonal matrix corresponding to the rotation by an angle $\theta = 6\pi/5$

$$\boldsymbol{O} = \begin{pmatrix} \cos(6\pi/5) & \sin(6\pi/5) \\ \sin(6\pi/5) & \cos(6\pi/5) \end{pmatrix}.$$

We take the sample $\boldsymbol{Z}_1, \ldots, \boldsymbol{Z}_n, n = 100$ and transform it to

$$\boldsymbol{X}_i = \boldsymbol{\mu} + \boldsymbol{O}\boldsymbol{Z}_i, i = 1, \ldots, 100.$$

For the new random sample $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n$, we use the same grid $\mathcal{G}_n$ with the values of the corresponding empirical center-outward distribution function from the untransformed sample. The results with the same highlighted point as in the previous part are plotted in Figure 2.4. The previous random sample is plotted in gray, with the highlighted point in red and the new one in black and blue color.

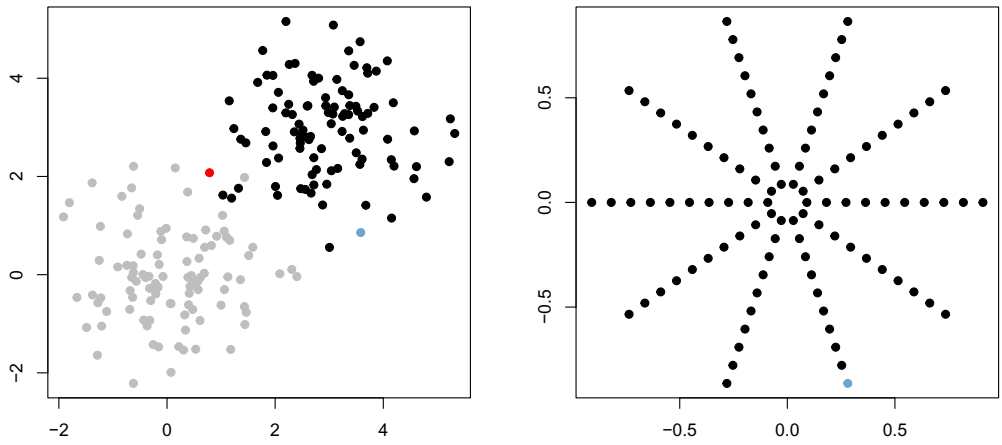

Figure 2.4: Transformed random sample (shifted and rotated 2-dimensional distribution) and the corresponding regular grid with the values of the empirical center-outward distribution function.

Subsequently, denoting by $\boldsymbol{F}_{\pm}^{\boldsymbol{\mu}+\boldsymbol{O}\boldsymbol{Z},(n)}$ the empirical distribution function of the sample $\boldsymbol{\mu}+\boldsymbol{O}\boldsymbol{Z}_1, \ldots, \boldsymbol{\mu}+\boldsymbol{O}\boldsymbol{Z}_n$ and analogously by $\boldsymbol{F}_{\pm}^{\boldsymbol{Z},(n)}$ the empirical distribution function of the sample $\boldsymbol{Z}_1, \ldots, \boldsymbol{Z}_n$, then for the highlighted point it holds

$$\boldsymbol{F}_{\pm}^{\boldsymbol{\mu}+\boldsymbol{O}\boldsymbol{Z},(n)}(\boldsymbol{\mu}+\boldsymbol{O}\boldsymbol{Z}_i) = \boldsymbol{O}\boldsymbol{F}_{\pm}^{\boldsymbol{Z},(n)}(\boldsymbol{Z}_i), \quad i = 1, \ldots, n.$$

In other words, the rank $R_{\pm}^{(n)}$ of the highlighted point in the transformed sample is the same as in the original one, and the sign $\boldsymbol{S}_{\pm}^{(n)}$ of the highlighted point is the same as the previous one rotated by the angle $\theta = 6\pi/5$. It is clear that

the angles between the empirical center-outward distribution function of the two observations are the same as before the transformation. That is consistent with Property 3 in Theorem 2.4.

2.4 Elliptical distributions

In this section, we present the resulting center-outward distribution function for elliptical distributions. First, let us define this type of distribution. Most of the following definitions are taken from Fang et al. (1990).

Definition 2.8. *A d-dimensional random vector $\boldsymbol{Z}$ is said to have a **spherical distribution** if for every $\Gamma \sim \mathcal{O}(n)$, it holds*

$$\boldsymbol{Z} \stackrel{d}{=} \Gamma \boldsymbol{Z},$$

where the sign $\stackrel{d}{=}$ means equality of the distributions, and $\mathcal{O}(n)$ denotes the set of $n \times n$ orthogonal matrices, i.e., the set of real square matrices Γ whose columns and rows are orthonormal vectors meaning

$$\Gamma\Gamma^\top = \Gamma^\top\Gamma = I_n,$$

where I_n is an $n \times n$ identity matrix.

From the corollary of Theorem 2.2 and Theorem 2.3 in Fang et al. (1990), it follows that a random vector $\boldsymbol{Z}$ has a spherical distribution if and only if

$$\boldsymbol{Z} \stackrel{d}{=} R\,U, \tag{2.3}$$

where $R \stackrel{d}{=} \|\boldsymbol{Z}\|, R$ is from a continuous distribution with a distribution function F, with the corresponding density f, and $U \stackrel{d}{=} \boldsymbol{Z}/\|\boldsymbol{Z}\|, U$ has a uniform distribution over the unit sphere $\mathcal{S}_{d-1}$. Moreover, R and U are independent. The distribution function F of $\|\boldsymbol{Z}\|$, and density f, are called *the radial distribution and radial density.*

It holds that $\boldsymbol{Z}$ has a spherical distribution if and only if

$$\boldsymbol{F}_{sp}(\boldsymbol{Z}) \coloneqq F(\|\boldsymbol{Z}\|)\boldsymbol{Z}/\|\boldsymbol{Z}\| \sim U_d.$$

It follows from the representation (2.3) and from the fact that the distribution U_d is obtained by the product of the uniform over the unit interval [0, 1] and the uniform measure over the unit sphere.

Let us suppose a spherical distribution with non-vanishing radial density. The mapping $\boldsymbol{Z} \mapsto \boldsymbol{F}_{sp}(\boldsymbol{Z})$ is such that it pushes the distribution P of the random vector $\boldsymbol{Z}$ forward to the uniform distribution over the unit ball $\mathbb{S}_d$. Because of the uniqueness of $\boldsymbol{F}_{\pm}$, the mapping $\boldsymbol{Z} \mapsto \boldsymbol{F}_{sp}(\boldsymbol{Z})$ coincides with the center-outward distribution function corresponding to $\boldsymbol{Z}$.

Definition 2.9. *Let $\boldsymbol{X}$ be a d-dimensional random vector. We say that $\boldsymbol{X}$ has an **elliptical distribution** $P_{\boldsymbol{\mu},\boldsymbol{\Sigma},f}$ with location $\boldsymbol{\mu} \in \mathbb{R}^d$, positive definite symmetric $d \times d$ scatter matrix $\boldsymbol{\Sigma}$, and radial density f if and only if $\boldsymbol{Z} \coloneqq \boldsymbol{\Sigma}^{-1/2}(\boldsymbol{X} - \boldsymbol{\mu})$ has a spherical distribution with radial density f.*

Analogously as for spherical distribution, we have

$$\boldsymbol{F}_{ell}(\boldsymbol{X}) \coloneqq F(\|\boldsymbol{\Sigma}^{-1/2}(\boldsymbol{X}-\boldsymbol{\mu})\|)(\boldsymbol{\Sigma}^{-1/2}(\boldsymbol{X}-\boldsymbol{\mu}))/\|\boldsymbol{\Sigma}^{-1/2}(\boldsymbol{X}-\boldsymbol{\mu})\| \sim U_d.$$

Same as for spherical distributions, the mapping $\boldsymbol{X} \mapsto \boldsymbol{F}_{ell}(\boldsymbol{X})$ pushes the distribution P of the random vector $\boldsymbol{X}$ forward to U_d. Due to the uniqueness of $\boldsymbol{F}_{\pm}$, $\boldsymbol{F}_{ell}(\boldsymbol{X})$ coincides with the center-outward distribution function corresponding to $\boldsymbol{X}$.

Let $\boldsymbol{X}_1^{(n)}, \ldots, \boldsymbol{X}_n^{(n)}$ be a sample from an elliptical distribution. Consider $\widehat{\boldsymbol{\mu}}^{(n)}$ and $\widehat{\boldsymbol{\Sigma}}^{(n)}$ consistent estimators of $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$. The empirical version of $\boldsymbol{F}_{\boldsymbol{ell}}$ is based on Mahalanobis ranks and signs.

Definition 2.10. *The ranks $R_i^{(n)}$ of the residual moduli*

$$\|\boldsymbol{Z}_i^{(n)}\| \coloneqq \|(\widehat{\boldsymbol{\Sigma}}^{(n)})^{-1/2}(\boldsymbol{X}_i - \widehat{\boldsymbol{\mu}}^{(n)})\|$$

*are called **Mahalanobis ranks**. In the same way, we call the unit vectors*

$$\boldsymbol{S}_i^{(n)} \coloneqq \boldsymbol{Z}_i^{(n)}/\|\boldsymbol{Z}_i^{(n)}\|$$

***Mahalanobis signs**.*

With these definitions, the empirical version of $\boldsymbol{F}_{\boldsymbol{ell}}$ is the following

$$\boldsymbol{F}_{\boldsymbol{ell}}^{(n)}(\boldsymbol{X}_i^{(n)}) \coloneqq (R_i^{(n)}/(n+1))\boldsymbol{S}_i^{(n)}.$$

The consistency in Glivenko-Cantelli sense can be obtained for $\boldsymbol{F}_{\boldsymbol{ell}}^{(n)}$, see Proposition C.1 in Hallin et al. (2021).

Proposition 2.5. *Let $\boldsymbol{X}_i^{(n)}, i = 1, \ldots, n$ be i.i.d. with an elliptical distribution $P_{\boldsymbol{\mu},\boldsymbol{\Sigma},f}$ and assume that $\widehat{\boldsymbol{\mu}}^{(n)}$ and $\widehat{\boldsymbol{\Sigma}}^{(n)}$ are strongly consistent estimators of $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$, respectively. Then, $\boldsymbol{F}_{ell}$ and $\boldsymbol{F}_{\pm}$ coincide and*

$$\max_{1\le i\le n} \left\|\boldsymbol{F}_{ell}^{(n)}(\boldsymbol{X}_i^{(n)}) - \boldsymbol{F}_{\pm}^{(n)}(\boldsymbol{X}_i^{(n)})\right\|, \text{ hence also } \max_{1\le i\le n} \left\|\boldsymbol{F}_{ell}^{(n)}(\boldsymbol{X}_i^{(n)}) - \boldsymbol{F}_{\pm}(\boldsymbol{X}_i^{(n)})\right\|$$

tend to zero a.s. as $n \to \infty$ where $\boldsymbol{F}_{\pm}$ denotes the center-outward distribution function of $P_{\boldsymbol{\mu},\boldsymbol{\Sigma},f}$.

2.5 Construction of grids

The values of ranks, signs, and the empirical center-outward distribution function $\boldsymbol{F}_{\pm}^{(n)}$ completely depend on the choice of the underlying grid $\mathcal{G}_n$. The grids are associated with the factorization

$$n = n_R n_S + n_0, \quad n_R, n_S, n_0 \in \mathbb{N}, \quad 0 \le n_0 < \min(n_R, n_S).$$

We assume a sequence of grids $\{\mathcal{G}_n\}$ such that $n_R \to \infty$ and $n_S \to \infty$ as $n \to \infty$ (implying $n_0/n \to 0$). As mentioned in Section 2.2.2, we want the grids $\{\mathcal{G}_n\}$ to be as "regular" as possible. We also want that for $n_R \to \infty$ and $n_S \to \infty$, the uniform discrete distribution over the grid $\mathcal{G}_n$ converges weakly to the uniform distribution over $\mathbb{S}_d$.

2.5.1 Grids in $\mathbb{R}^2$

In $\mathbb{R}^2$, the points on the grid can be written using polar coordinates in the form

$$g_{i,j} = \begin{pmatrix} r_i \cos(\varphi_j) \\ r_i \sin(\varphi_j) \end{pmatrix}, \quad i = 1, \ldots, n_R, j = 1, \ldots, n_S.$$

Denoting

$$\boldsymbol{s}_j = \begin{pmatrix} \cos(\varphi_j) \\ \sin(\varphi_j) \end{pmatrix}, \quad j = 1, \ldots, n_S,$$

the directions, i.e., unit vectors in $\mathbb{R}^2$, we get

$$g_{i,j} = r_i \boldsymbol{s}_j,$$

where r_i corresponds to the norm of the vector.

One possible way of solving the problem is to construct grids in the same way as in Section 2.2.2. This means dividing the unit circle into n_S arcs of equal length $2\pi/n_S$ and taking the intersections between an n_S-tuple $(\boldsymbol{u}_1, \ldots, \boldsymbol{u}_{n_S})$

$$\boldsymbol{u}_j = \begin{pmatrix} \cos(\varphi_j) \\ \sin(\varphi_j) \end{pmatrix}, j = 1, \ldots, n_S,$$

where $\varphi_j = (2\pi j)/n_S, j = 1, \ldots, n_S$ of created unit vectors of $\mathcal{S}_d$ and n_R hyperspheres with the center at $\mathbf{0}$ and radii $r_j = j/(n_R + 1), j = 1, \ldots, n_R$ along with n_0 copies of the origin. For an example, see Figure 2.2.

Another possible approach is to create a random grid, i.e., to take for example n_S-tuple $(\boldsymbol{u}_1, \ldots, \boldsymbol{u}_{n_S})$, same as before, but for each of the unit vectors to take n_R random variables with uniform distribution over $[0, 1]$ and to create n_R points as intersections of the unit vector and hyperspheres with radii given by the random variables along with n_0 copies of the origin. An example of such grid for $n_R = n_S = 10, n_0 = 1$ is in the left top subplot of Figure 2.5.

The previous method is just one possible way to construct a random grid with the required properties. Let us assume we take n_S random variables $\varphi_1, \ldots, \varphi_{n_S}$ with uniform distribution over $[0, 2\pi]$ and create the unit vectors as

$$(\cos \varphi_i, \sin \varphi_i)^\top, i = 1, \ldots, n_S.$$

Then we can again construct the grid by taking intersections of these unit vectors and n_R hyperspheres with the center at $\mathbf{0}$ and radii $j/(n_R+1), j = 1, \ldots, n_R$ along with n_0 copies of the origin, see the right top subplot of Figure 2.5.

We can also combine the two approaches using randomness and take n_S random variables $\varphi_1, \ldots, \varphi_{n_S}$ with uniform distribution over $[0, 2\pi]$ to create the unit vectors as $(\cos \varphi_i, \sin \varphi_i)^\top, i = 1, \ldots, n_S$ and, moreover, for each of the unit vectors we can take n_R random variables with uniform distribution over $[0, 1]$ and create n_R points as intersections of the unit vector and hyperspheres with radii given by the random variables. This, along with n_0 copies of the origin, creates the grid, see the left bottom subplot of Figure 2.5.

Another possible combination of the two mentioned random approaches is to generate the points of the grid as

$$g_i = r_i s_i, \quad i = 1, \ldots, n,$$

where r_i are independent identically distributed random variables from the uniform distribution over $[0, 1]$ and s_i are independent identically distributed random variables from the uniform distribution over unit sphere $\mathcal{S}_1$ which can be generated in the same way as mentioned in the beginning of the subsection, using angles φ_i. This approach does not need factorization, see the right bottom subplot of Figure 2.5.

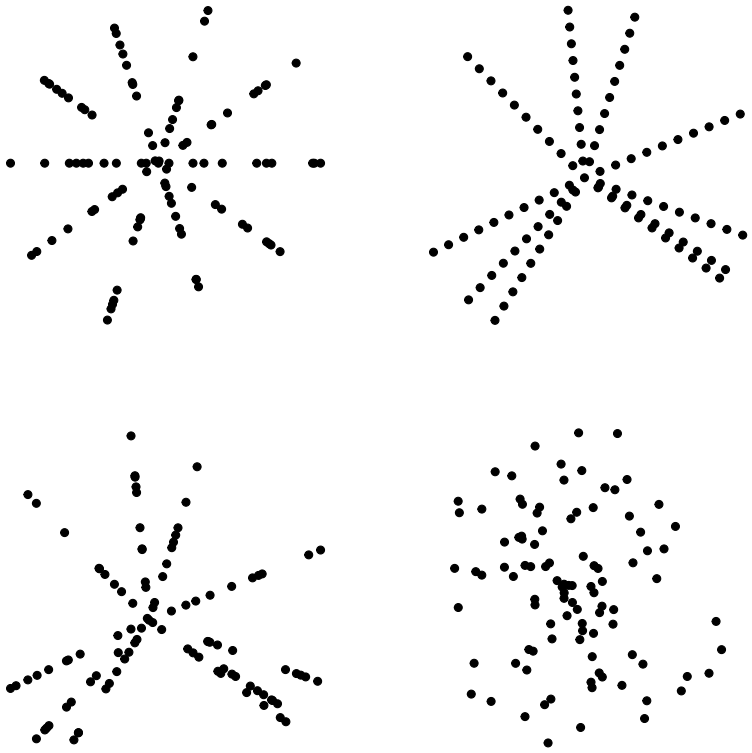

Figure 2.5: Some possible choices of grids for empirical center-outward distribution function $\boldsymbol{F}_{\pm}^{(n)}$, for $n_R = n_S = 10, n_0 = 1$. In the left top subplot, there are random ranks, and in the right top, there are random signs. In the left bottom subplot, there is a combination of both with fixed signs for n_R points, and in the right bottom, there is a combination of both with ranks and signs generated independently.

2.5.2 Grids in spaces with dimension $d > 2$

In the higher dimension, the problem of choosing the grid for the empirical distribution function becomes quite complex. If we have n_S unit vectors created by points on a unit sphere $\mathcal{S}_{d-1}$ in $\mathbb{R}^d$, it is possible to create the grid from the intersections of these unit vectors with n_R hyperspheres with the center at $\mathbf{0}$ and radii $j/(n_R + 1), j = 1, \ldots, n_R$ along with n_0 copies of the origin or to take n_R random variables with uniform distribution over $[0, 1]$ and create n_R points as

intersections of the unit vector and hyperspheres with radii given by the random variables. Therefore, the grid points can be written as

$$g_{i,j} = r_i s_j, i = 1, \ldots, n_R, j = 1, \ldots, n_S,$$

where $r_i, i = 1, \ldots, n_R$ are ranks and $s_j, j = 1, \ldots, n_S$ are directions. r_i, s_j, or both might be random or deterministic (i.e., like the regular grid shown in Figure 2.2).

Again, the choices of the n_S directions should be as "regular" as possible. One way to choose the directions is to generate the unit vector directly from the uniform distribution on $\mathcal{S}_{d-1}$. That can be achieved, for example, by generating a random sample $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n$ from a multivariate normal distribution and then taking $\boldsymbol{X}_1/\|\boldsymbol{X}_1\|, \ldots, \boldsymbol{X}_n/\|\boldsymbol{X}_n\|$, which follows from the properties of spherical distribution presented in Section 2.4. The resulting random sample is uniformly distributed over the unit sphere $\mathcal{S}_{d-1}$. Another possible way is to choose points in the unit cube $\mathbb{C}_{d-1}$ in $\mathbb{R}^{d-1}$ as uniformly as possible and then transform these points onto the unit sphere $\mathcal{S}_{d-1}$ in $\mathbb{R}^d$ to form unit vectors. For that, we can use the method described in Section 1.5.3 in Fang & Wang (1994) and spherical coordinates, which are defined by

$$\begin{aligned} x_j &= \prod_{i=1}^{j-1} S_i C_j, \quad j = 1, \ldots, s-1, \\ x_s &= \prod_{i=1}^{s-1} S_i, \end{aligned}$$

where $S_i = \sin(\pi\varphi_i), C_i = \cos(\pi\varphi_i), i = 1, \ldots, s-2$ and $S_{s-1} = \sin(2\pi\varphi_{s-1})$, $C_{s-1} = \cos(2\pi\varphi_{s-1})$.

This is a transformation of the unit cube in $\mathbb{R}^{d-1}$ onto a unit sphere $\mathcal{S}_{d-1}$ in $\mathbb{R}^d$. In order to get $\boldsymbol{x}$ with uniform distribution on $\mathcal{S}_{d-1}$, the random variables $\varphi_1, \ldots, \varphi_{s-1}$ should be independent, and the density of φ_i is

$$p_i(\varphi) = \frac{\pi}{B\left(\frac{1}{2}, \frac{s-i}{2}\right)} (\sin(\pi\varphi))^{s-i-1}. \tag{2.4}$$

The proof can be found in Appendix B.2 in Fang & Wang (1994). The corresponding cumulative distribution function is

$$F_i(\varphi) = \int_0^{\varphi} p_i(t) \mathrm{d}t. \tag{2.5}$$

To illustrate the previous method, we present the following example inspired by Section 1.5.3 in Fang & Wang (1994).

Example 2.4 (Constructing grid in $\mathbb{R}^3$). In this case, s from the previous theory equals 3. Therefore, we need two random variables φ_1, φ_2. Their density is given

by (2.4) and can be computed as

$$p_1(\varphi) = \frac{\pi}{B\left(\frac{1}{2}, 1\right)}(\sin(\pi\varphi))^1 = \frac{\pi}{2}\sin(\pi\varphi),$$
$$p_2(\varphi) = \frac{\pi}{B\left(\frac{1}{2}, \frac{1}{2}\right)}(\sin(\pi\varphi))^0 = \frac{\pi}{\pi} = 1,$$

and the cumulative distribution functions are given by (2.5)

$$F_1(\varphi) = 1/2(1 - \cos(\pi\varphi)),$$
$$F_2(\varphi) = \varphi.$$

Let us suppose we have a set of points in the unit cube $\mathbb{C}_2$ denoted by $\{(c_{k1}, c_{k2}), k = 1, \ldots, n\}$ and we want to transform them onto the unit sphere $\mathcal{S}_2$ in $\mathbb{R}^3$. For $\{(c_{k1}, c_{k2}), k = 1, \ldots, n\}$ uniformly scattered in the unit cube in $\mathbb{R}^2$, we get, using the inverse of the cumulative distribution function, the points uniformly scattered on the unit sphere $\mathcal{S}_2$. From the inverse of the cumulative distribution functions, we compute

$$\tilde{\varphi}_{k1} = F_1^{-1}(c_{k1}) = \frac{1}{\pi}\arccos(1 - 2c_{k1}),$$
$$\tilde{\varphi}_{k2} = F_2^{-1}(c_{k2}) = c_{k2}.$$

From that and the trigonometric identity, we have

$$\cos(\pi\tilde{\varphi}_{k1}) = 1 - 2c_{k1},$$
$$\sin(\pi\tilde{\varphi}_{k1}) = \sqrt{1 - \cos^2(\pi\tilde{\varphi}_1)} = \sqrt{1 - (1 - 4c_{k1} + 4c_{k1}^2)} = 2\sqrt{c_{k1}(1 - c_{k1})},$$
$$\cos(2\pi\tilde{\varphi}_{k2}) = \cos(2\pi c_{k2}),$$
$$\sin(2\pi\tilde{\varphi}_{k2}) = \sin(2\pi c_{k2}).$$

The points on the unit sphere $\mathcal{S}_2$ are given by

$$x_{k1} = \cos(\pi\tilde{\varphi}_{k1}) = 1 - 2c_{k1},$$
$$x_{k2} = \sin(\pi\tilde{\varphi}_{k1})\cos(2\pi\tilde{\varphi}_{k2}) = 2\sqrt{c_{k1}(1 - c_{k1})}\cos(2\pi c_{k2}),$$
$$x_{k3} = \sin(\pi\tilde{\varphi}_{k1})\sin(2\pi\tilde{\varphi}_{k2}) = 2\sqrt{c_{k1}(1 - c_{k1})}\sin(2\pi c_{k2}).$$

This transformation gives us a way how to map points from the unit cube $\mathbb{C}_{d-1}$ onto the unit sphere $\mathcal{S}_{d-1}$ in $\mathbb{R}^d$ while keeping them uniformly scattered, for more details and formalities see Theorem 1.6 in Fang & Wang (1994).

This method turns the problem of finding a regular grid on the unit sphere into a problem of finding the uniformly scattered points in the unit cube. The procedure described in the previous example can be generalized into higher dimensions. In the following simulation study in Chapter 4, we generate data also in the dimension 4, but the details of the computation of the transformation are similar to the previous example and are omitted here. The low-discrepancy sequences,

such as the Halton sequence, see Section 1.3.3 in Fang & Wang (1994), might be useful to generate points in the unit cube $\mathbb{C}_{d-1}$.

In the following Figure 2.6, we present the Halton sequence generated in the unit cube in $\mathbb{R}^2$ and the corresponding unit sphere in $\mathbb{R}^3$ after the transformation described in this section.

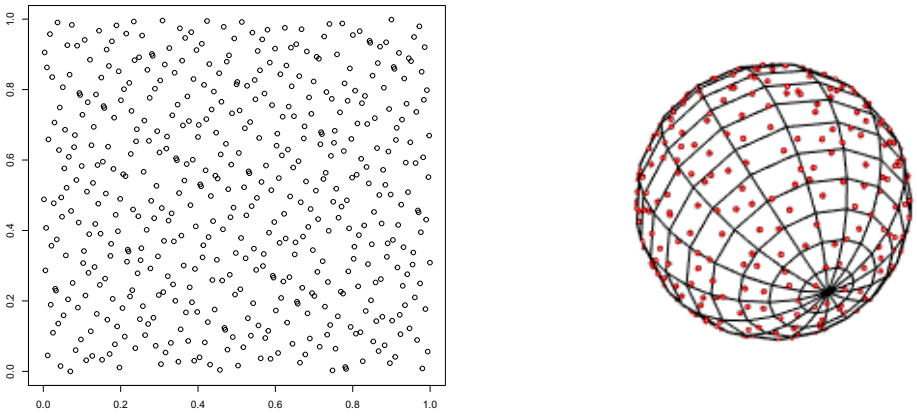

Figure 2.6: Halton sequence transformed from the unit cube in $\mathbb{R}^2$ onto the unit sphere in $\mathbb{R}^3$.

3. Test statistics based on center-outward ranks

In this chapter, we present test statistics based on the center-outward ranks. The main idea comes from the theory of rank statistics in Hájek et al. (1999), and some of the following concepts are taken from Hallin, Hlubinka & Hudecová (2022).

3.1 Multivariate simple rank statistic

In this section, we provide some concepts for testing the difference in location between two samples. The construction of the tests will be based on center-outward ranks and signs.

3.1.1 Assumptions and the definition of the center-outward rank statistic

We generalize the ideas presented in Section 1.2 and Example 1.1 by taking a vector score function $\boldsymbol{J} : \mathbb{S}_d \to \mathbb{R}^d$ and real numbers $c_1^{(n)}, \ldots, c_n^{(n)}, n \in \mathbb{N}$ as regression constants. In order to establish the asymptotic normality of the later presented test statistic, we make the following assumptions, same as in Section 3.1 in Hallin, Hlubinka & Hudecová (2022).

Assumption 3.1. *(i)* $\boldsymbol{J} : \mathbb{S}_d \to \mathbb{R}^d$ *is continuous over* $\mathbb{S}_d$*,*

(ii) for any sequence $\mathfrak{s}^{(n)} = \{s_1^{(n)}, \ldots, s_n^{(n)}\}$ *of* n*-tuples in* $\mathbb{S}_d$ *such that the uniform discrete distribution over* $\mathfrak{s}^{(n)}$ *converges weakly to* U_d *as* $n \to \infty$*, it holds*

$$\lim_{n\to\infty} n^{-1}\mathrm{tr}\sum_{i=1}^{n} \boldsymbol{J}(s_i^{(n)})\boldsymbol{J}'(s_i^{(n)}) = \mathrm{tr}\int_{\mathbb{S}_d} \boldsymbol{J}(\boldsymbol{u})\boldsymbol{J}'(\boldsymbol{u})\mathrm{d}U_d,$$

where $\int_{\mathbb{S}_d} \boldsymbol{J}(\boldsymbol{u})\boldsymbol{J}'(\boldsymbol{u})\mathrm{d}U_d < \infty$ *has full rank.*

We also need to make an assumption about regression constants.

Assumption 3.2. *The* $c_i^{(n)}, i = 1, \ldots, n$ *are not all equal (for given* n*) and satisfy*

$$\frac{\sum_{i=1}^{n}(c_i^{(n)} - \overline{c}^{(n)})^2}{\max_{1\le i\le n}(c_i^{(n)} - \overline{c}^{(n)})^2} \to \infty \quad \text{as } n \to \infty,$$

where $\overline{c}^{(n)} := n^{-1}\sum_{i=1}^{n} c_i^{(n)}$*.*

Let us denote

$$\boldsymbol{T}_a^{(n)} := \left(\sum_{i=1}^{n}(c_i^{(n)} - \overline{c}^{(n)})^2\right)^{-1/2} \sum_{i=1}^{n}(c_i^{(n)} - \overline{c}^{(n)})\boldsymbol{J}(\boldsymbol{F}_{\pm}^{(n)}(\boldsymbol{Z}_i^{(n)})),$$

and

$$\boldsymbol{T}^{(n)} \coloneqq \left(\sum_{i=1}^{n}(c_i^{(n)} - \overline{c}^{(n)})^2\right)^{-1/2} \sum_{i=1}^{n}(c_i^{(n)} - \overline{c}^{(n)})\boldsymbol{J}(\boldsymbol{F}_{\pm}(\boldsymbol{Z}_i^{(n)})).$$

It is typically impossible to compute $\boldsymbol{T}^{(n)}$ in practice because $\boldsymbol{F}_{\pm}$ is unknown. We will use $\boldsymbol{T}^{(n)}$ to derive the asymptotic distribution of $\boldsymbol{T}_a^{(n)}$ which will be used for testing.

We call $\boldsymbol{T}_a^{(n)}$ an *approximate-score linear rank statistic*. It is constructed in the same way as a simple linear rank statistic (1.2) in Section 1.2. Moreover, it is already normalized in some way, see the connection to Theorem 1.3.

We can also assume $\boldsymbol{J}(\boldsymbol{u}), \boldsymbol{u} \in \mathbb{S}_d$, to be of a special form

$$\boldsymbol{J}(\boldsymbol{u}) \coloneqq J(\|\boldsymbol{u}\|)\frac{\boldsymbol{u}}{\|\boldsymbol{u}\|}\mathbb{I}\left[\|\boldsymbol{u}\| \neq 0\right], \quad \boldsymbol{u} \in \mathbb{S}_d, \tag{3.1}$$

where $J : [0,1) \to \mathbb{R}$ is a univariate score function. Then Assumption 3.1 reduces to

Assumption 3.3. *(i) J is continuous,*

(ii) it holds

$$0 < \lim_{n\to\infty} n^{-1}\sum_{r=1}^{n} J^2(r/(n+1)) = \int_0^1 J^2(u)\mathrm{d}u < \infty.$$

The test statistics $\boldsymbol{T}_a^{(n)}$ and $\boldsymbol{T}^{(n)}$ can also be rewritten for score function $\boldsymbol{J}$ of form (3.1). We get

$$\boldsymbol{T}_a^{(n)} \coloneqq \left(\sum_{i=1}^{n}(c_i^{(n)} - \overline{c}^{(n)})^2\right)^{-1/2} \sum_{i=1}^{n}(c_i^{(n)} - \overline{c}^{(n)})J\left(\frac{R_{\pm,i}^{(n)}}{n_R+1}\right)\boldsymbol{S}_{\pm,i}^{(n)} \tag{3.2}$$

and

$$\boldsymbol{T}^{(n)} \coloneqq \left(\sum_{i=1}^{n}(c_i^{(n)} - \overline{c}^{(n)})^2\right)^{-1/2} \sum_{i=1}^{n}(c_i^{(n)} - \overline{c}^{(n)})J\left(\|\boldsymbol{F}_{\pm}(\boldsymbol{Z}_i^{(n)})\|\right)\frac{\boldsymbol{F}_{\pm}(\boldsymbol{Z}_i^{(n)})}{\|\boldsymbol{F}_{\pm}(\boldsymbol{Z}_i^{(n)})\|}. \tag{3.3}$$

3.1.2 Asymptotic normality

The asymptotic normality of the univariate rank-based test statistic was stated in Section 1.2 (Theorem 1.3). In this section, we provide its multivariate version for the test statistics presented above.

Let us consider a random sample $\boldsymbol{Z}_1^{(n)}, \ldots, \boldsymbol{Z}_n^{(n)}, n \in \mathbb{N}$ of d-dimensional random vectors with distribution P from a family of distributions with non-vanishing densities $\mathcal{P}_d^+$. Let $\boldsymbol{F}_{\pm}$ be the corresponding center-outward distribution function and $\boldsymbol{F}_{\pm}^{(n)}$ its empirical version with a range on a grid $\mathcal{G}_n$.

In the following, we will consider a sequence of grids $\{\mathcal{G}_n\}$ such that $n_R \to \infty$ and $n_S \to \infty$ and a uniform distribution on the grid $\mathcal{G}_n$ converges to U_d as $n_R \to \infty$

and $n_S \to \infty$. Then, we have a sequence of empirical center-outward distribution functions $\boldsymbol{F}_{\pm}^{(n)}$ with corresponding grids $\mathcal{G}_n$ satisfying the previous condition. By $n \to \infty$ in the following Theorems 3.1, 3.2, and 3.3 we mean the convergence of $n_R \to \infty$, $n_S \to \infty$, and corresponding grids $\mathcal{G}_n$ and empirical center-outward distribution functions $\boldsymbol{F}_{\pm}^{(n)}$ with factorization $n = n_R n_S + n_0$.

Due to the following theorem the two presented test statistics $\boldsymbol{T}_a^{(n)}$ and $\boldsymbol{T}^{(n)}$ are asymptotically equivalent.

Theorem 3.1. *Let Assumptions 3.1 and 3.2 hold. Then for $\boldsymbol{Z}_1^{(n)}, \ldots, \boldsymbol{Z}_n^{(n)}$ i.i.d. with distribution $P \in \mathcal{P}_d^+$, it holds that $\boldsymbol{T}_a^{(n)} - \boldsymbol{T}^{(n)}$ converges to zero in quadratic mean (hence also in probability) as $n \to \infty$.*

The proof can be found in Appendix A.2 in Hallin, Hlubinka & Hudecová (2022). Along with asymptotic equivalence, we present a theorem about the asymptotic normality of the mentioned test statistics.

Theorem 3.2. *Let Assumptions 3.1 and 3.2 hold. Then for $\boldsymbol{Z}_1^{(n)}, \ldots, \boldsymbol{Z}_n^{(n)}$ i.i.d. with distribution $P \in \mathcal{P}_d$, the test statistics $\boldsymbol{T}_a^{(n)}$ and $\boldsymbol{T}^{(n)}$ are asymptotically normal as $n \to \infty$, with mean $\mathbf{0}$ and covariance matrix*

$$\int_{\mathbb{S}_d} \boldsymbol{J}(\boldsymbol{u})\boldsymbol{J}'(\boldsymbol{u})\mathrm{d}U_d.$$

The main idea of the proof is to take the statistic $\boldsymbol{T}^{(n)}$ (3.3) and use the central limit theorem to prove its asymptotic normality. Then it is sufficient to use the asymptotic equivalence from Theorem 3.1. The full proof can be found in Appendix A.3 in Hallin, Hlubinka & Hudecová (2022), and we will prove it for a simpler form of the score function $\boldsymbol{J}(\boldsymbol{u})$ presented in (3.1).

Theorem 3.3. *Let Assumptions 3.3 and 3.2 hold. Then for $\boldsymbol{Z}_1^{(n)}, \ldots, \boldsymbol{Z}_n^{(n)}$ i.i.d. with distribution $P \in \mathcal{P}_d$, the reduced test statistics $\boldsymbol{T}_a^{(n)}$ and $\boldsymbol{T}^{(n)}$ from equations* (3.2) *and* (3.3) *are asymptotically normal as $n \to \infty$, with mean $\mathbf{0}$ and covariance matrix*

$$\frac{\int_0^1 J^2(u)\mathrm{d}u}{d}\boldsymbol{I}_d,$$

where $\boldsymbol{I}_d$ is $d \times d$ unit matrix.

Proof. We will prove the asymptotic normality of $\boldsymbol{T}^{(n)}$ and the asymptotic normality of $\boldsymbol{T}_a^{(n)}$ then follows from Theorem 3.1. We take $\boldsymbol{F}_{\pm}(\boldsymbol{Z}_1^{(n)})$ which is a random variable such that

$$\boldsymbol{W} \coloneqq \boldsymbol{F}_{\pm}(\boldsymbol{Z}_1^{(n)}) = \|\boldsymbol{F}_{\pm}(\boldsymbol{Z}_1^{(n)})\| \frac{\boldsymbol{F}_{\pm}(\boldsymbol{Z}_1^{(n)})}{\|\boldsymbol{F}_{\pm}(\boldsymbol{Z}_1^{(n)})\|} \stackrel{d}{=} R\boldsymbol{U},$$

where R is uniform over $[0, 1]$, $\boldsymbol{U}$ is uniform over $\mathcal{S}_{d-1}$, and they are mutually independent.

$\boldsymbol{T}^{(n)}$ is a sum of independent random variables and it holds that $\mathsf{E}\,\boldsymbol{U} = \mathbf{0}$ (see Fang et al. (1990), Theorem 2.7). Therefore,

$$\begin{aligned}\mathsf{E}\,\boldsymbol{T}^{(n)} &= \left(\sum_{i=1}^{n}(c_i^{(n)} - \overline{c}^{(n)})^2\right)^{-1/2}\sum_{i=1}^{n}(c_i^{(n)} - \overline{c}^{(n)})\,\mathsf{E}\,[J\,(R)\,\boldsymbol{U}] = \\ &= \left(\sum_{i=1}^{n}(c_i^{(n)} - \overline{c}^{(n)})^2\right)^{-1/2}\sum_{i=1}^{n}(c_i^{(n)} - \overline{c}^{(n)})\,\mathsf{E}\,J\,(R)\,\mathsf{E}\,\boldsymbol{U} = \mathbf{0}.\end{aligned}$$

For variance, it holds from the independence and the fact that $\boldsymbol{T}^{(n)}$ is a sum of i.i.d. random variables that

$$\begin{aligned}\mathrm{var}\,(\boldsymbol{T}^{(n)}) &= \left(\sum_{i=1}^{n}(c_i^{(n)} - \overline{c}^{(n)})^2\right)^{-1}\sum_{i=1}^{n}(c_i^{(n)} - \overline{c}^{(n)})^2\mathrm{var}[J\,(R)\,\boldsymbol{U}] = \\ &= \mathrm{var}\,[J(R)\,\boldsymbol{U}] = \mathsf{E}\,[(J(R)\boldsymbol{U})(J(R)\boldsymbol{U})^\top] - \mathsf{E}\,[J(R)\boldsymbol{U}]\,\mathsf{E}\,[J(R)\boldsymbol{U}]^\top = \\ &= \mathsf{E}\,J^2(R)\,\mathsf{E}\,\boldsymbol{U}\boldsymbol{U}^\top - (\mathsf{E}\,J(R))^2\,\mathsf{E}\,\boldsymbol{U}\,\mathsf{E}\,\boldsymbol{U}^\top = \\ &= \mathsf{E}\,J^2(R)\,\mathsf{E}\,\boldsymbol{U}\boldsymbol{U}^\top - \mathsf{E}\,J^2(R)\,\mathsf{E}\,\boldsymbol{U}\,\mathsf{E}\,\boldsymbol{U}^\top \\ &\quad + \mathsf{E}\,J^2(R)\,\mathsf{E}\,\boldsymbol{U}\,\mathsf{E}\,\boldsymbol{U}^\top - (\mathsf{E}\,J(R))^2\,\mathsf{E}\,\boldsymbol{U}\,\mathsf{E}\,\boldsymbol{U}^\top = \\ &= \mathsf{E}\,J^2(R)\mathrm{var}\,\boldsymbol{U} + \mathrm{var}\,J(R)\,\mathsf{E}\,\boldsymbol{U}\,\mathsf{E}\,\boldsymbol{U}^\top.\end{aligned}$$

From Theorem 2.7 in Fang et al. (1990), we have $\mathsf{E}\,\boldsymbol{U} = \mathbf{0}$ and $\mathrm{var}\,\boldsymbol{U} = \frac{1}{d}\boldsymbol{I}_d$. Therefore,

$$\mathrm{var}\,(\boldsymbol{T}^{(n)}) = \frac{\int_0^1 J^2(u)\mathrm{d}u}{d}\boldsymbol{I}_d.$$

Altogether, we obtain $\boldsymbol{T}^{(n)}$ as a sum of independent variables for which the Feller-Lindeberg condition is satisfied due to Assumption 3.2. Therefore, from the central limit theorem, we get

$$\boldsymbol{T}^{(n)} \xrightarrow{d} \mathcal{N}_d\left(\mathbf{0}, \frac{\int_0^1 J^2(u)\mathrm{d}u}{d}\boldsymbol{I}_d\right), \quad \text{as } n \to \infty,$$

in a way described in the formulation of Theorem 3.3. □

We get back to the situation of the two-sample test of location from Example 1.1 and generalize the idea into a multivariate case based on the theory presented above.

3.2 Two-sample test of location

Let us consider a random sample $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_{n_1}$ from a d-dimensional distribution with a continuous distribution function F_1 and a density f_1. Also, consider another independent random sample $\boldsymbol{Y}_1, \ldots, \boldsymbol{Y}_{n_2}$ from a d-dimensional distribution with a continuous distribution function F_2 and a density f_2. We want to test whether the two densities coincide. We assume that under the alternative, there is a difference in location, i.e., that $f_1(\boldsymbol{x}) = f(\boldsymbol{x} - \boldsymbol{\theta})$ and $f_2(\boldsymbol{x}) = f(\boldsymbol{x}), \boldsymbol{x} \in \mathbb{R}^d$ for a density f. The null hypothesis and the alternative are

$$H_0 : \boldsymbol{\theta} = \mathbf{0} \text{ versus } H_1 : \boldsymbol{\theta} \neq \mathbf{0}.$$

The joint density of the sample $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_{n_1}, \boldsymbol{Y}_1, \ldots, \boldsymbol{Y}_{n_2}$ is

$$q_\theta(\mathbb{z}) = \prod_{i=1}^{n_1+n_2} f(\boldsymbol{z}_i - c_i\boldsymbol{\theta}),$$

where $\mathbb{z} = (\boldsymbol{z_1}, \ldots, \boldsymbol{z_{n_1+n_2}})^\top = (\boldsymbol{x_1}, \ldots, \boldsymbol{x_{n_1}}, \boldsymbol{y_1}, \ldots, \boldsymbol{y_{n_2}})^\top$, and

$$c_i = \begin{cases} 1, & \text{if } i = 1, \ldots, n_1, \\ 0, & \text{if } i = n_1 + 1, \ldots, n_1 + n_2. \end{cases}$$

We can define a sample $\boldsymbol{Z}_1, \ldots, \boldsymbol{Z}_n$ computed as a multivariate version of (1.4) under the null hypothesis H_0:

$$\boldsymbol{Z}_i \coloneqq \boldsymbol{Z}_i(\boldsymbol{0}) = \begin{cases} \boldsymbol{X}_i, & \text{if } i = 1, \ldots, n_1, \\ \boldsymbol{Y}_{i-n_1}, & \text{if } i = n_1 + 1, \ldots, n_1 + n_2. \end{cases} \tag{3.4}$$

In this case, the test statistic $\boldsymbol{T}_a^{(n)}$ based on (3.2) is of form

$$\begin{aligned}
\boldsymbol{T}_a^{(n)} &\coloneqq \left(\sum_{i=1}^{n}(c_i^{(n)} - \overline{c}^{(n)})^2\right)^{-1/2} \sum_{i=1}^{n}(c_i^{(n)} - \overline{c}^{(n)}) J\left(\frac{R_{\pm,i}^{(n)}}{n_R+1}\right) \boldsymbol{S}_{\pm,i}^{(n)} = \\
&= \left(\sum_{i=1}^{n_1}\left(1 - \frac{n_1}{n}\right)^2 + \sum_{i=n_1+1}^{n}\left(-\frac{n_1}{n}\right)^2\right)^{-1/2} \sum_{i=1}^{n}(c_i^{(n)} - \overline{c}^{(n)}) J\left(\frac{R_{\pm,i}^{(n)}}{n_R+1}\right) \boldsymbol{S}_{\pm,i}^{(n)} = \\
&= \left(n_1\left(\frac{n_2}{n}\right)^2 + n_2\left(\frac{n_1}{n}\right)^2\right)^{-1/2} \sum_{i=1}^{n}(c_i^{(n)} - \overline{c}^{(n)}) J\left(\frac{R_{\pm,i}^{(n)}}{n_R+1}\right) \boldsymbol{S}_{\pm,i}^{(n)} = \\
&= \left(\frac{n_1 n_2}{n}\right)^{-1/2} \left(\sum_{i=1}^{n_1} J\left(\frac{R_{\pm,i}^{(n)}}{n_R+1}\right) \boldsymbol{S}_{\pm,i}^{(n)} - \frac{n_1}{n}\sum_{i=1}^{n} J\left(\frac{R_{\pm,i}^{(n)}}{n_R+1}\right) \boldsymbol{S}_{\pm,i}^{(n)}\right),
\end{aligned}$$

where n_R is from factorization (2.2) and $R_{\pm,i}^{(n)}$ and $\boldsymbol{S}_{\pm,i}^{(n)}, i = 1, \ldots, n$ are ranks and signs computed out of the sample $\boldsymbol{Z}_1, \ldots, \boldsymbol{Z}_n$. If J is an identity and the grid $\mathcal{G}_n = \{g_i, i = 1, \ldots, n\}$ is chosen in a way that

$$\sum_{i=1}^{n} g_i = \sum_{i=1}^{n} \frac{R_{\pm,i}^{(n)}}{n_R+1} \boldsymbol{S}_{\pm,i}^{(n)} = \boldsymbol{0},$$

then the test statistic $\boldsymbol{T}_a^{(n)}$ is of form

$$\boldsymbol{T}_a^{(n)} = \left(\frac{n_1 n_2}{n}\right)^{-1/2} \sum_{i=1}^{n_1} \frac{R_{\pm,i}^{(n)}}{n_R+1} \boldsymbol{S}_{\pm,i}^{(n)}.$$

It is a multivariate analogy of Wilcoxon test statistic, i.e., the sum of ranks corresponding to the first sample just multiplied by the sign and partly standardized.

3.2.1 Asymptotic behavior of the test statistic

To test the null hypothesis $H_0 : \boldsymbol{\theta} = \mathbf{0}$ versus $H_1 : \boldsymbol{\theta} \neq \mathbf{0}$, we can use Theorem 3.3 which states

$$\boldsymbol{T}_a^{(n)} \xrightarrow{d} \mathcal{N}_d \left(\mathbf{0}, \frac{\int_0^1 J^2(u)\mathrm{d}u}{d} \boldsymbol{I}_d \right), \quad \text{as } n \to \infty.$$

Therefore, it holds

$$\left(\frac{d}{\int_0^1 J^2(u)\mathrm{d}u} \right)^{1/2} \boldsymbol{T}_a^{(n)} \xrightarrow{d} \mathcal{N}_d \left(\mathbf{0}, \boldsymbol{I}_d \right), \quad \text{as } n \to \infty.$$

We construct the following test statistic as a quadratic form

$$\begin{aligned} Q^{(n)} &:= \left\| \left(\frac{d}{\int_0^1 J^2(u)\mathrm{d}u} \right)^{1/2} \boldsymbol{T}_a^{(n)} \right\|^2 = \\ &= \left(\frac{nd}{n_1 n_2 \int_0^1 J^2(u)\mathrm{d}u} \right) \left\| \sum_{i=1}^{n_1} J\left(\frac{R_{\pm,i}^{(n)}}{n_R + 1} \right) \boldsymbol{S}_{\pm,i}^{(n)} - \frac{n_1}{n} \sum_{i=1}^{n} J\left(\frac{R_{\pm,i}^{(n)}}{n_R + 1} \right) \boldsymbol{S}_{\pm,i}^{(n)} \right\|^2 . \end{aligned} \tag{3.5}$$

From the previous theory, the test statistic $Q^{(n)}$ under H_0 have asymptotically χ_d^2 distribution. Therefore, we reject the null hypothesis at the asymptotic level of significance α as long as $Q^{(n)}$ is greater than $(1-\alpha)-$quantile of χ_d^2 distribution.

To illustrate the theory, we provide an example of behavior under the null hypothesis and the alternative in the case of two Gaussian samples.

Example 3.1. At first, we simulate the behavior under the null hypothesis. Consider two random samples $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_{n_1}$ and $\boldsymbol{Y}_1, \ldots, \boldsymbol{Y}_{n_2}$, where $n_1 = n_2 = 50$ and both are generated from 2-dimensional normal distribution $\mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, where $\boldsymbol{\mu} = (0,0)^\top$ and

$$\boldsymbol{\Sigma} = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}.$$

Let $\boldsymbol{Z}_1, \ldots, \boldsymbol{Z}_n, n = 100$ be a sample, which can be computed under the null hypothesis as follows

$$\boldsymbol{Z}_i = \begin{cases} \boldsymbol{X}_i, & \text{if } i = 1, \ldots, n_1, \\ \boldsymbol{Y}_{i-n_1}, & \text{if } i = n_1 + 1, \ldots, n_1 + n_2. \end{cases}$$

We factorize n into $n = n_R n_S + n_0$, where $n_R = 10, n_S = 10$ and $n_0 = 0$, and create a regular grid $\mathcal{G}_n$ with values of the empirical center-outward distribution function for a random sample $\mathcal{Z}$ given by $\boldsymbol{Z}_1, \ldots, \boldsymbol{Z}_n$. We distinguish the two samples by color ($\boldsymbol{X}_1, \ldots, \boldsymbol{X}_{n_1}$ in red and $\boldsymbol{Y}_1, \ldots, \boldsymbol{Y}_{n_2}$ in blue) and plot them, see the left subplot of Figure 3.1.

We take the multivariate analogy of the Wilcoxon test statistic (so J is the identity) and the grid is chosen in a way that

$$\sum_{i=1}^{n} \frac{R_{\pm,i}^{(n)}}{n_R+1} \boldsymbol{S}_{\pm,i}^{(n)} = \mathbf{0}.$$

The test statistic $\boldsymbol{T}_a^{(n)}$ is of form

$$\boldsymbol{T}_a^{(n)} = \left(\frac{n_1 n_2}{n}\right)^{-1/2} \sum_{i=1}^{n_1} \frac{R_{\pm,i}^{(n)}}{n_R+1} \boldsymbol{S}_{\pm,i}^{(n)}.$$

We compute the test statistic $\boldsymbol{T}_a^{(n)}$ and plot the resulting vector into the grid $\mathcal{G}_n$ with values of the empirical center-outward distribution function, see the right subplot of Figure 3.1. The two samples are again distinguished by color. It can be seen that $\boldsymbol{T}_a^{(n)}$ is a sum of vectors given by red points and multiplied by $\left(\frac{n_1 n_2}{n}\right)^{-1/2}$.

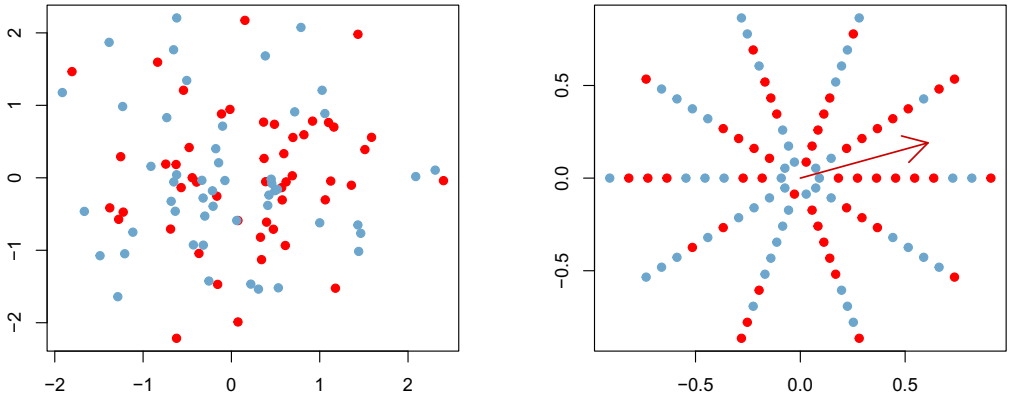

Figure 3.1: Two Gaussian random samples with $\boldsymbol{\mu} = (0,0)^\top$ and $\boldsymbol{\Sigma} = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}$, distinguished by color on the left. The corresponding regular grid with values of the empirical center-outward distribution function is plotted with additional vector statistic $\boldsymbol{T}_a^{(n)}$ in the right subplot.

The value of $\boldsymbol{T}_a^{(n)}$ is $(0.609, 0.190)^\top$ and we compute the test statistic followingly

$$Q^{(n)} = \left(\frac{nd}{n_1 n_2 \int_0^1 J^2(u)\mathrm{d}u}\right) \left\| \sum_{i=1}^{n_1} \left(\frac{R_{\pm,i}^{(n)}}{n_R+1}\right) \boldsymbol{S}_{\pm,i}^{(n)} \right\|^2.$$

The value of $Q^{(n)}$ is 2.444 and the corresponding asymptotic p-value is

$$p = 1 - G(q) \doteq 0.295,$$

where G is the distribution function of χ_2^2 distribution and q is the value of the test statistic $Q^{(n)}$. Therefore, we do not reject the null hypothesis of the same location.

Next, we generate data under the alternative. We have two random samples $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_{n_1}$ and $\boldsymbol{Y}_1, \ldots, \boldsymbol{Y}_{n_2}$, where $n_1 = n_2 = 50$, and both are generated from 2-dimensional normal distribution. The first sample has mean $\boldsymbol{\mu}_1 = (0, 0)^\top$ and the second one $\boldsymbol{\mu}_2 = (1, 1)^\top$. Both samples are generated from a normal distribution with the following covariance matrix

$$\boldsymbol{\Sigma} = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}.$$

We create a sample $\mathcal{Z}$ and grid $\mathcal{G}_n$ in the same way as above. The statistics $\boldsymbol{T}_a^{(n)}$ and $Q^{(n)}$ are computed analogously to the previous case. The value of $\boldsymbol{T}_a^{(n)}$ is $(-1.459, -1.697)^\top$, $Q^{(n)} = 30.042$, and the corresponding p-value is less than 0.001. Therefore, we reject the null hypothesis of the same location. See Figure 3.2 for the plot of both samples, with the grid $\mathcal{G}_n$ with samples distinguished by color and the vector statistic $\boldsymbol{T}_a^{(n)}$ added as the red arrow.

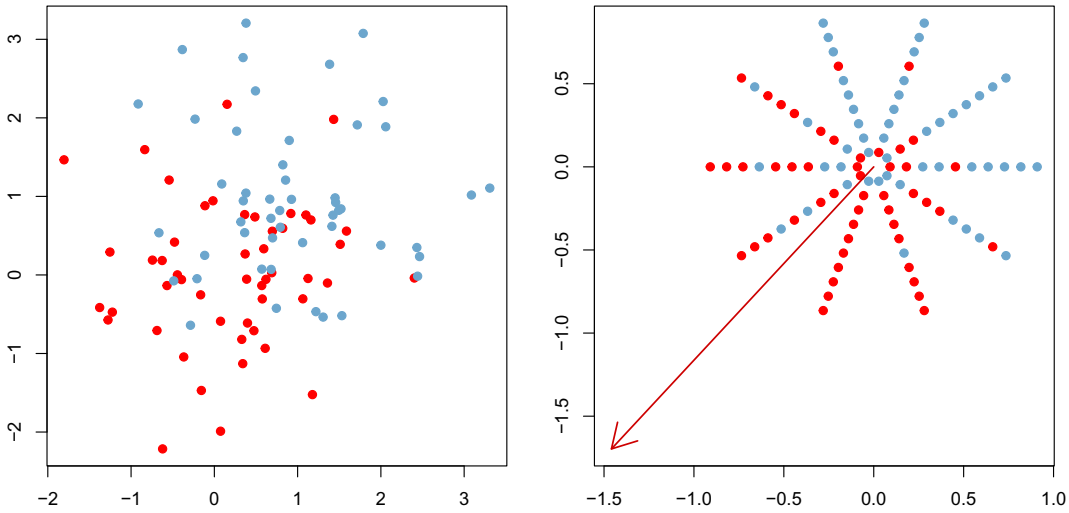

Figure 3.2: Two Gaussian random samples with $\boldsymbol{\mu}_1 = (0, 0)^\top$, $\boldsymbol{\mu}_2 = (1, 1)^\top$, and $\boldsymbol{\Sigma} = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}$, distinguished by color on the left. The corresponding regular grid with values of the empirical center-outward distribution function is plotted with the additional vector statistic $\boldsymbol{T}_a^{(n)}$ in the right subplot.

Due to Figure 3.2, the observations from the first sample, i.e., the red points are accumulated on the left side of the grid $\mathcal{G}_n$ resolving into a vector statistic $\boldsymbol{T}_a^{(n)}$ with large norm.

3.2.2 Permutation test

A permutation test might be useful when n is small and we cannot rely on the asymptotic behavior. It is another point of view on testing the null hypothesis, for basic theory see Davison & Hinkley (1997).

Let us assume a situation from Section 3.2 where we have a sample $\mathcal{Z}$ given by $\boldsymbol{Z}_1, \ldots, \boldsymbol{Z}_n$, computed under the null hypothesis H_0 as (3.4). The idea is to take the permutation π of set $\{1, \ldots, n\}$ and to create a sample $\mathcal{Z}^*$ from $\boldsymbol{Z}_1, \ldots, \boldsymbol{Z}_n$

permuted according to the permutation π. The first n_1 random vectors form the first sample and the other n_2 form the second one to be compared. For sample $\mathcal{Z}^*$, we compute test statistic Q^* analogously to (3.5).

Generally, we reject the null hypothesis H_0 for values of the test statistic $Q^{(n)}$ greater than $(1-\alpha)-$quantile of χ^2_d distribution. Under the null hypothesis H_0, the distribution of Q^* is the same as the distribution of $Q^{(n)}$, and the p-value can be computed as a sample quantile. Under the alternative by the change of the allocation into the groups, we get from the situation in Figure 3.2 to the situation in Figure 3.1. Therefore, the test statistic $Q^{(n)}$ should be greater than Q^* and the test should reflect it in the p-value. The p-value of the corresponding test can be computed by

$$p = \frac{\text{number of permutations such that } Q^* \geq Q^{(n)}}{\text{number of all possible permutations}},$$

where $Q^{(n)}$ is the test statistic computed for the two-sample problem.

With n increasing, the number of permutations increases so fast that it might be impossible to compute them all. Therefore, we approximate p-value by taking B randomly selected permutations. We get samples $\mathcal{Z}^*_1, \ldots, \mathcal{Z}^*_B$ and corresponding test statistics $Q^*_1, \ldots, Q^*_B$ and compute

$$p = \frac{1 + \sum_{b=1}^{B} \mathbb{I}[Q^*_b \geq Q^{(n)}]}{1 + B}.$$

For the permutation test in the two-sample testing of the same location, it is enough to transport the data into the chosen grid just once. The permutations of allocations of the points into the two groups can be done separately. It is enough to recolor the given points on the grid with the empirical distribution function due to the new allocation and compute the test statistic based on the two new groups.

Example 3.2. Let us get back to the situation in Example 3.1. For the same situation and the same data, we also perform a permutation test. We have two random samples $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_{n_1}$ and $\boldsymbol{Y}_1, \ldots, \boldsymbol{Y}_{n_2}$, where $n_1 = n_2 = 50$, and both are generated from 2-dimensional normal distribution.

First, consider the behavior under the null hypothesis. The test statistic $Q^{(n)}$ from Example 3.1 is 2.444 and the corresponding asymptotic p-value is 0.295. There are 100! different permutations of observations. Out of all of them, we choose $B = 999$ permutations and use the previously described method to perform the permutation test. The corresponding p-value for our chosen permutations is 0.284.

We perform the same procedure for $B = 999$ and for the data from Example 3.1 simulated under the alternative. For this scenario, the test statistic $Q^{(n)}$ from Example 3.1 is 30.042 and the corresponding asymptotic p-value is less than 0.001. From the permutation test, we obtain the p-value equal to 0.

3.3 One-sample test of location under central symmetry

This section provides several concepts for testing the location with test statistics based on the center-outward ranks and signs. The idea of location is relatively broad and not specified. We will first deal with location as a center of symmetry.

At first, we define the central symmetry the same as in Serfling (2006).

Definition 3.1. *A d-dimensional random vector* $\boldsymbol{X}$ *is centrally symmetric about* $\boldsymbol{\theta} \in \mathbb{R}^d$ *if*

$$\boldsymbol{X} - \boldsymbol{\theta} \stackrel{d}{=} \boldsymbol{\theta} - \boldsymbol{X}.$$

Consider a random sample $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n$ satisfying Definition 3.1. The parameter $\boldsymbol{\theta}$ is the center of symmetry and we want to test the null hypothesis $H_0 : \boldsymbol{\theta} = \boldsymbol{\theta}_0$. Let us assume without loss of generality $\boldsymbol{\theta}_0 = \mathbf{0}$. Then we test

$$H_0 : \boldsymbol{\theta} = \mathbf{0} \text{ versus } H_1 : \boldsymbol{\theta} \neq \mathbf{0}. \tag{3.6}$$

The following subsection shows one of the ways how to test the null hypothesis (3.6).

3.3.1 Test with randomized signs

Under the null hypothesis (3.6), it holds $\boldsymbol{X} \stackrel{d}{=} -\boldsymbol{X}$. Let us consider a random variable Y independent of $\boldsymbol{X}$ with uniform distribution on $\{-1, 1\}$. It holds that the random variables $\boldsymbol{X}$ and $\boldsymbol{X}^* \coloneqq Y\boldsymbol{X}$ have the same distribution.

Consider a random sample $Y_1, \ldots, Y_n$ independent of the random sample $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n$. Then the distribution of the random sample $\mathcal{X}_n = (\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n)$ and random sample $\mathcal{X}_n^* = (\boldsymbol{X}_1^*, \ldots, \boldsymbol{X}_n^*)$, where $\boldsymbol{X}_i^* \coloneqq Y_i\boldsymbol{X}_i, i = 1, \ldots, n$, coincide under the null hypothesis. Let us denote $\mathcal{I}^+ = \{i \mid Y_i = 1\}$, $n_+ = |\mathcal{I}^+|$ and

$$\mathcal{X}^+ = \{\boldsymbol{X}_i^* \mid i \in \mathcal{I}^+\}, \quad \mathcal{X}^- = \{\boldsymbol{X}_i^* \mid i \notin \mathcal{I}^+\}.$$

It holds that the samples $\mathcal{X}^+$ and $\mathcal{X}^-$ are independent and have the same distribution. This takes us back to the situation of the two-sample test shown in Sections 3.1.1 and 3.1.2. Therefore, we can use the test statistic in the same way as in Section 3.2:

$$\boldsymbol{T}^{(n)} = \left(\frac{n_+(n-n_+)}{n}\right)^{-1/2} \left(\sum_{i\in\mathcal{I}^+} J\left(\frac{R_{\pm,i}^{(n)}}{n_R+1}\right)\boldsymbol{S}_{\pm,i}^{(n)} - \frac{n_+}{n}\sum_{i=1}^{n} J\left(\frac{R_{\pm,i}^{(n)}}{n_R+1}\right)\boldsymbol{S}_{\pm,i}^{(n)}\right), \tag{3.7}$$

where $R_{\pm,i}^{(n)}, \boldsymbol{S}_{\pm,i}^{(n)}$ are ranks and signs of $\boldsymbol{X}_i^*, i = 1, \ldots, n$. It follows from Theorem 3.3 that under Assumptions 3.3 and 3.2, the test statistic $\boldsymbol{T}^{(n)}$ is asymptotically normal as $n \to \infty$ (in such way that n_R and n_S from factorization $n = n_R n_S + n_0$ satisfy $n_R \to \infty$ and $n_S \to \infty$), with mean $\mathbf{0}$ and covariance matrix

$$\frac{\int_0^1 J^2(u)\mathrm{d}u}{d}\boldsymbol{I}_d,$$

where $\boldsymbol{I}_d$ is $d \times d$ unit matrix. Therefore, we reject the null hypothesis $H_0 : \boldsymbol{\theta} = \mathbf{0}$ if

$$Q^{(n)} = \left(\frac{nd}{n_+(n-n_+) \int_0^1 J^2(u)\mathrm{d}u} \right) \left\| \sum_{i \in \mathcal{I}^+} J\left(\frac{R_{\pm,i}^{(n)}}{n_R+1} \right) \boldsymbol{S}_{\pm,i}^{(n)} - \frac{n_+}{n} \sum_{i=1}^{n} J\left(\frac{R_{\pm,i}^{(n)}}{n_R+1} \right) \boldsymbol{S}_{\pm,i}^{(n)} \right\|^2$$

is greater than $(1-\alpha)$-quantile of χ^2_d distribution. n_+ and $n - n_+$ must converge to infinity with the same speed of convergence to satisfy Assumption 3.2. In this case, n_+ and $n - n_+$ are random variables with binomial distribution with mean $n/2$. Therefore, the Assumption 3.2 holds with probability 1.

For a given random sample $\mathcal{X}_n$, the result of the statistical test is random. For another realization of the sample $Y_1, \ldots, Y_n$, we would get another p-value. It might be possible to combine these p-values from more replications with different $Y_1, \ldots, Y_n$, but this problem is beyond the scope of this work.

We provide the following example to illustrate the concept and the behavior under the null hypothesis and the alternative.

Example 3.3 (One-sample test of location under central symmetry). First, let us explore the behavior under the null hypothesis. Consider a sample $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n$, $n = 100$ from a 2-dimensional normal distribution $\mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, where $\boldsymbol{\mu} = (0,0)^\top$ and

$$\boldsymbol{\Sigma} = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}.$$

We generate a random sample $Y_1, \ldots, Y_n, n = 100$, from the uniform distribution on $\{-1, 1\}$ and assign these signs to observations $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n$. The allocation of the signs to the observations is displayed on the left in Figure 3.3. The observations with assigned negative signs are colored blue and the observations with assigned positive signs are colored red. Then, we create the random sample $\mathcal{X}_n^*$ by

$$\boldsymbol{X}_i^* \coloneqq Y_i \, \boldsymbol{X}_i, i = 1, \ldots, n,$$

where the blue points from the left subplot of Figure 3.3 are reflected through the origin.

We take the blue and red points as two separate samples and perform a two-sample test of location. In the situation in Figure 3.3, the null hypothesis is not violated so the test should not reject it.

Next, we simulate the behavior under the alternative. We consider a sample $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n, n = 100$ from a 2-dimensional normal distribution $\mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, where $\boldsymbol{\mu} = (1,1)^\top$ and

$$\boldsymbol{\Sigma} = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}.$$

We again test the null hypothesis $H_0 : \boldsymbol{\theta} = (0,0)^\top$ and we proceed in a similar way as in the previous part. The resulting allocation of signs to the observations and the random sample $\mathcal{X}_n$ colored according to the allocation of signs are displayed in Figure 3.4.

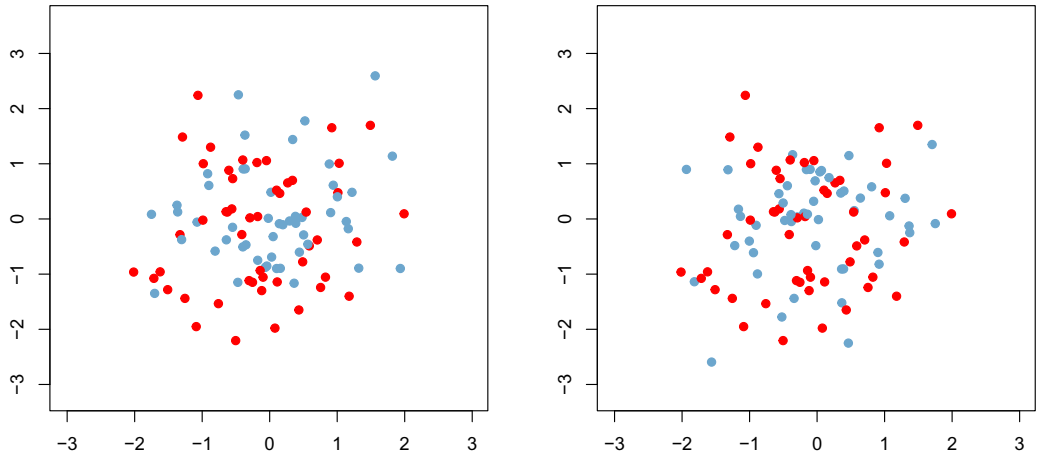

Figure 3.3: The behavior under the null hypothesis. A random sample $\mathcal{X}_n$ colored according to the allocation of signs in the left subplot and a random sample $\mathcal{X}_n^*$ with blue points reflected through the origin on the right.

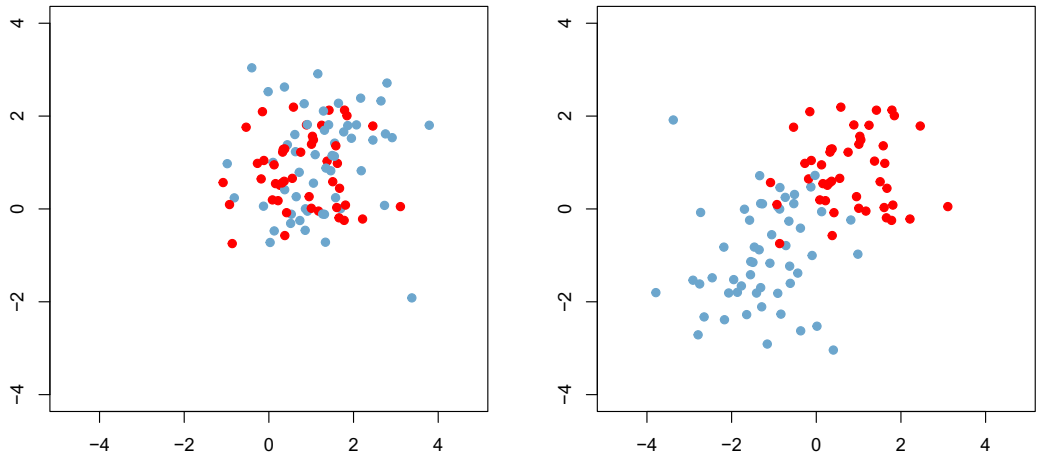

Figure 3.4: The behavior under the alternative hypothesis. A random sample $\mathcal{X}_n$ colored according to the allocation of signs in the left subplot and a random sample $\mathcal{X}_n^*$ with blue points reflected through the origin in the right one.

In this case, it is clear that after the reflection of the blue points through the origin, we get two almost separate samples, see Figure 3.4. Thus, the two-sample test of location should reject the null hypothesis.

3.3.2 Test based on added $\boldsymbol{\theta_0}$

In this subsection, we provide another way to test the location of a random sample. Let us again consider the null hypothesis (3.6).

Let us suppose we have a centrally symmetric random sample $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n$ and, moreover, we add an origin as another observation $\boldsymbol{X}_{n+1} = \mathbf{0}$. We factorize

$$n + 1 = n_R n_S + 1$$

to get a grid $\mathcal{G}_n$ with exactly one copy of the origin. We compute the empirical center-outward distribution function $\boldsymbol{F}_{\pm}^{(n)}$ and the test statistic would be

$$T_0^{(n)} \coloneqq \|\boldsymbol{F}_{\pm}^{(n)}(\mathbf{0})\|.$$

Under the null hypothesis, we assume that $\boldsymbol{F}_{\pm}^{(n)}(\mathbf{0})$ would be in the origin or at least in its close proximity. Therefore, the value of $T_0^{(n)}$ should be small. On the contrary, we expect the norm to be large under the alternative. Therefore, we should reject the null hypothesis because of the large test statistic values. To illustrate intuition, we provide the following example.

Example 3.4. Let us again consider the behavior under the null hypothesis. Suppose we have a sample $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n, n = 100$ from a 2-dimensional normal distribution $\mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, where $\boldsymbol{\mu} = (0, 0)^\top$ and

$$\boldsymbol{\Sigma} = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}.$$

We add an origin as another observation $\boldsymbol{X}_{n+1} = \mathbf{0}$ and factorize $n+1 = n_R n_S + 1$, where $n_R = n_S = 10$. The random sample with added zero and the regular grid with values of the empirical center-outward distribution function $\boldsymbol{F}_{\pm}^{(n)}$ computed from the sample with added zero is plotted in Figure 3.5. The origin is highlighted in red color.

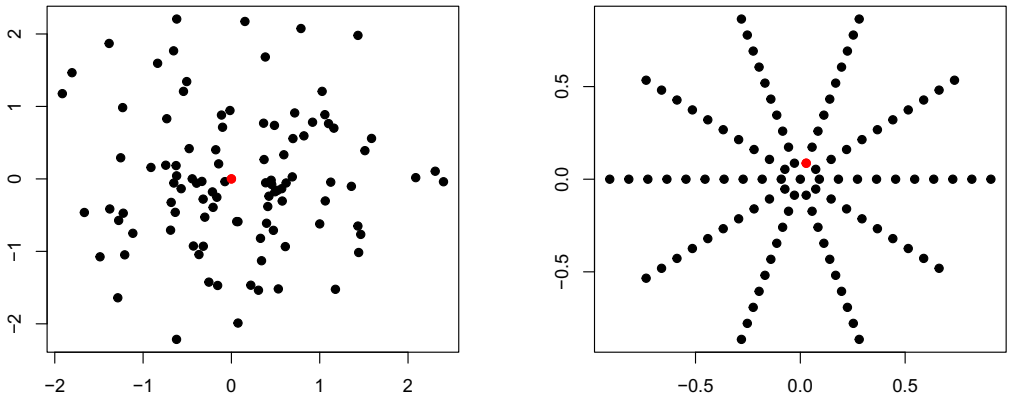

Figure 3.5: The behavior under the null hypothesis. A random sample $\mathcal{X}_n$ with added zero and a regular grid with values of $\boldsymbol{F}_{\pm}^{(n)}$. The origin is highlighted in red color.

In the same way, we can simulate the behavior under the alternative. Let us consider a sample $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n, n = 100$ from a 2-dimensional normal distribution $\mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, where $\boldsymbol{\mu} = (1, 1)^\top$ and

$$\boldsymbol{\Sigma} = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}.$$

We performed the same procedure as under the null hypothesis and plot the random sample with added zero and the regular grid with values of the empirical center-outward distribution function $\boldsymbol{F}_{\pm}^{(n)}$ in Figure 3.6. The origin is again highlighted in red color.

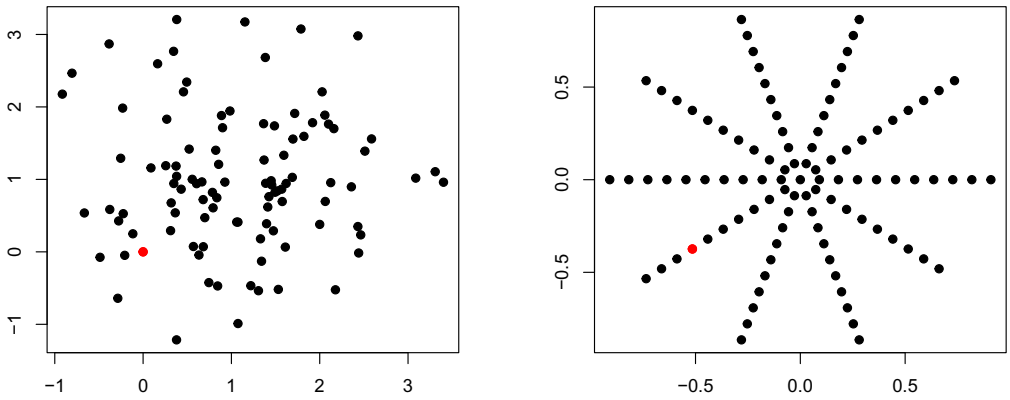

Figure 3.6: The behavior under the alternative hypothesis. A random sample $\mathcal{X}_n$ with added zero and a regular grid with values of $\boldsymbol{F}_{\pm}^{(n)}$. The origin is highlighted in red color.

Everything is consistent with our previous intuition about the test statistic

$$T_0^{(n)} \coloneqq \|\boldsymbol{F}_{\pm}^{(n)}(\mathbf{0})\|.$$

One way to test the null hypothesis H_0 (3.6) is to derive the exact or at least asymptotic distribution of $T_0^{(n)}$, which could be quite difficult. Thus, we use another approach using a permutation test similarly as in Section 3.2.2.

Same as in Section 3.3.1, it holds $\boldsymbol{X} \stackrel{d}{=} -\boldsymbol{X}$ under the null hypothesis. Consider a random sample $Y_1, \ldots, Y_n$ with uniform distribution on $\{-1, 1\}$ independent of a random sample $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n$. Then the distribution of a random sample $\mathcal{X}_n = (\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n)$ and a random sample $\mathcal{X}_n^* = (\boldsymbol{X}_1^*, \ldots, \boldsymbol{X}_n^*)$, where $\boldsymbol{X}_i^* \coloneqq Y_i \boldsymbol{X}_i, i = 1, \ldots, n$, coincide under the null hypothesis. We add an origin as another observation $\boldsymbol{X}_{n+1}^* = \mathbf{0}$ to the random sample $\mathcal{X}_n^*$ and compute

$$T_0^* \coloneqq \|\boldsymbol{F}_{\pm}^{*(n)}(\mathbf{0})\|.$$

Under the null hypothesis H_0, the $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n$ are equally likely to be positive and negative. There are 2^n possible combinations of observations with the given signs. We again approximate the p-value by taking B randomly selected assignments of signs to observations. We get samples $\mathcal{Z}_1^*, \ldots, \mathcal{Z}_B^*$ and the corresponding test statistics $T_{0,1}^*, \ldots, T_{0,B}^*$ and compute

$$p = \frac{1 + \sum_{b=1}^{B} \mathbb{I}[T_{0,b}^* \geq T_0^{(n)}]}{1 + B}.$$

In this case, we need to transform the data into the chosen grid for each permutation separately. Therefore, the computation of the p-value might be computationally expensive.

The result of this approach is dependent on the choice of the grid $\mathcal{G}_n$. For the regular grid we have chosen so far, each rank is common for n_S points. Therefore,

the test statistic $T_0^{(n)}$ is discrete with $n_R + 1$ possible values. Because of that, we might not be able to construct a test on the exact chosen level of significance. Consequently, we suggest using a slightly non-regular grid. We use the grid with random ranks presented in Section 2.5.1.

Example 3.5. Let us suppose the same situation as in the previous Example 3.4 and the behavior under the null hypothesis. In this case, we use a random grid with random ranks described in Section 2.5.1 and plotted in the left top subfigure of Figure 2.5. We add an origin as another observation $\boldsymbol{X}_{n+1} = \mathbf{0}$ and factorize $n + 1 = n_R n_S + 1$, where $n_R = n_S = 10$. We compute the test statistic

$$T_0^{(n)} \coloneqq \|\boldsymbol{F}_{\pm}^{(n)}(\mathbf{0})\|.$$

In this case, $T_0^{(n)} = 0.013$. The random sample with added zero and the non-regular grid with values of the empirical center-outward distribution function $\boldsymbol{F}_{\pm}^{(n)}$ computed from the sample with added zero is plotted in Figure 3.7. The origin is highlighted in red color.

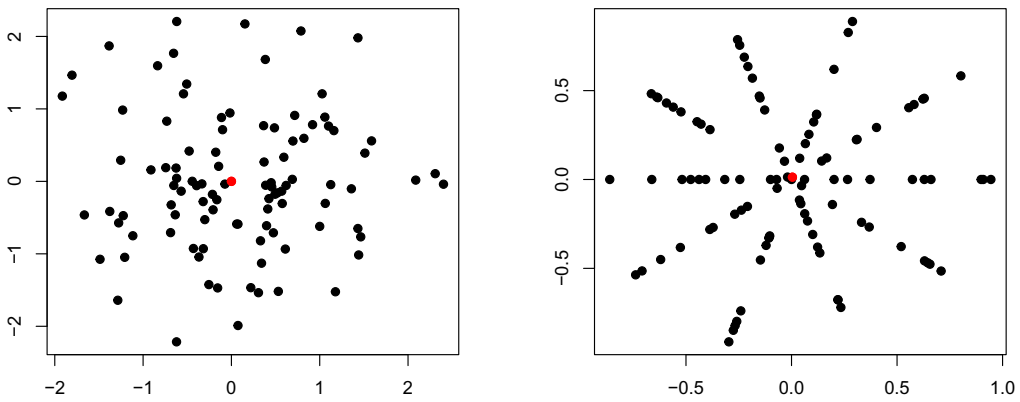

Figure 3.7: The behavior under the null hypothesis. A random sample $\mathcal{X}_n$ with added zero and a non-regular grid with values of $\boldsymbol{F}_{\pm}^{(n)}$. The origin is highlighted in red color.

We perform the permutation test with $B = 999$ based on the theory described above. The random grid was fixed, i.e., the same for all permutations, and we get a p-value equal to 0.741. It is consistent with the fact that we generated the data under the null hypothesis. We also provide the histogram of the values of the test statistic $T_{0,b}^*$ from the permutation test with the value of $T_0^{(n)}$ highlighted in red.

We can also generate the data under the alternative, i.e., a sample $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n$, $n = 100$ from a 2-dimensional normal distribution $\mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, where $\boldsymbol{\mu} = (1, 1)^\top$ and

$$\boldsymbol{\Sigma} = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}.$$

We again add an origin as another observation $\boldsymbol{X}_{n+1} = \mathbf{0}$ and factorize

$$n + 1 = n_R n_S + 1, \text{ where } n_R = n_S = 10.$$

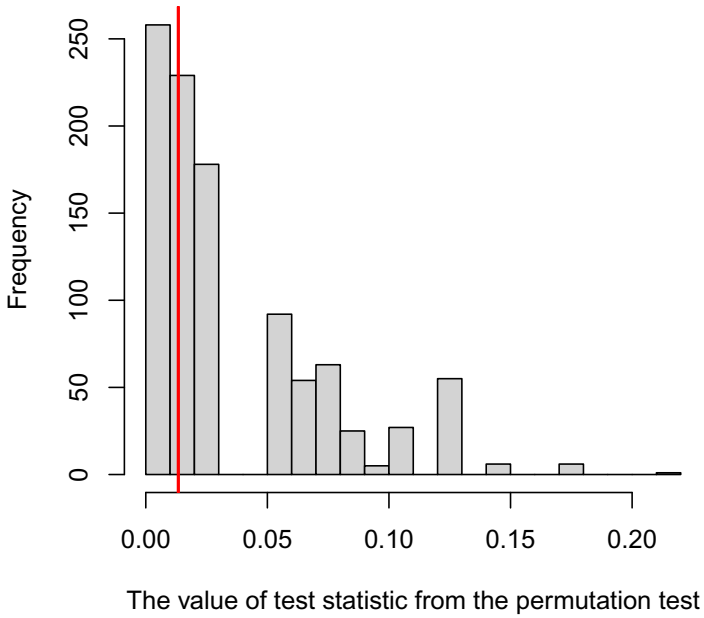

Figure 3.8: The histogram of the values of the test statistic $T^*_{0,b}$ from the permutation test, with the value of $T_0^{(n)}$ highlighted in red.

The random sample with added zero and the regular grid with values of the empirical center-outward distribution function $\boldsymbol{F}_{\pm}^{(n)}$ computed from the sample with added zero is plotted in Figure 3.9. The origin is highlighted in red color.

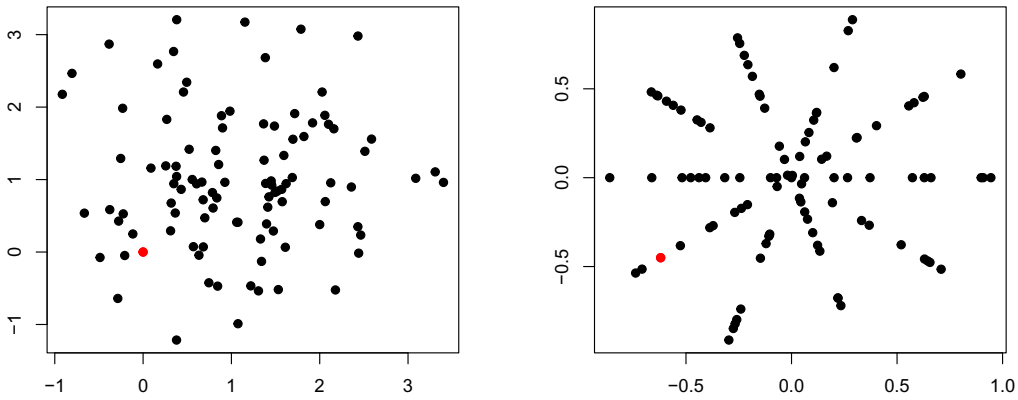

Figure 3.9: The behavior under the alternative. A random sample $\mathcal{X}_n$ with added zero and a non-regular grid with values of $\boldsymbol{F}_{\pm}^{(n)}$. The origin is highlighted in red color.

After performing the permutation test with $B = 999$, we get p-value equal to 0. The random grid was again fixed, i.e., the same for all permutations. The results are consistent with the fact that we generated the data under the alternative.

3.4 One-sample test of the location under angular symmetry

The previous Section 3.3 presents tests based on the assumption of central symmetry. This assumption is not fulfilled for a number of distributions. A slightly more general concept is angular symmetry, which will be covered here. We define the angular symmetry based on Serfling (2006).

Definition 3.2. *A d-dimensional random vector $\boldsymbol{X}$ has a distribution angularly symmetric about $\boldsymbol{\theta} \in \mathbb{R}^d$ if*

$$\frac{\boldsymbol{X}-\boldsymbol{\theta}}{\|\boldsymbol{X}-\boldsymbol{\theta}\|} \stackrel{d}{=} \frac{\boldsymbol{\theta}-\boldsymbol{X}}{\|\boldsymbol{X}-\boldsymbol{\theta}\|}.$$

For $\boldsymbol{X}$ angularly symmetric due to Definition 3.2, it holds that $\dfrac{\boldsymbol{X}-\boldsymbol{\theta}}{\|\boldsymbol{X}-\boldsymbol{\theta}\|}$ has a centrally symmetric distribution.

Consider a random sample $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n$ satisfying Definition 3.2. In this situation, we can again test the null hypothesis (3.6). We denote

$$\boldsymbol{W} = \frac{\boldsymbol{X}}{\|\boldsymbol{X}\|}$$

for a random vector $\boldsymbol{X}$. Under the null hypothesis, it holds $\boldsymbol{W} \stackrel{d}{=} -\boldsymbol{W}$, which takes us back to the situation of the test with randomized signs for one-sample test of location in Section 3.3.1.

The only problem is that the distribution of $\boldsymbol{W}$ is not absolutely continuous over $\mathbb{R}^2$. Therefore, we cannot use the theory described in previous sections. One way how to solve this problem is to work with the data on the unit sphere directly, for details see Hallin, Liu & Verdebout (2022). Another way to work with such data and also with the theory described in the previous section is to denote

$$\widetilde{\boldsymbol{X}} = R\frac{\boldsymbol{X}}{\|\boldsymbol{X}\|},$$

where R is a random variable with continuous distribution over $[0, \infty)$ which is independent of $\dfrac{\boldsymbol{X}}{\|\boldsymbol{X}\|}$. Then we use the following lemma to derive that under the null hypothesis $\widetilde{\boldsymbol{X}} \stackrel{d}{=} -\widetilde{\boldsymbol{X}}$.

Lemma 3.4. *Let us consider a random vector $\boldsymbol{X}$ defined on $\mathbb{R}^d$ satisfying angular symmetry about $\mathbf{0}$ from Definition 3.2 and suppose we have a continuous random variable R defined on $\mathbb{R}$ which is independent of $\dfrac{\boldsymbol{X}}{\|\boldsymbol{X}\|}$, then it holds $\widetilde{\boldsymbol{X}} \stackrel{d}{=} -\widetilde{\boldsymbol{X}}$.*

Proof. We have

$$\frac{\boldsymbol{X}}{\|\boldsymbol{X}\|} \stackrel{d}{=} -\frac{\boldsymbol{X}}{\|\boldsymbol{X}\|},$$

therefore, the equality of the cumulative distribution functions holds

$$F_{\frac{\boldsymbol{X}}{\|\boldsymbol{X}\|}}(\boldsymbol{x}) = F_{-\frac{\boldsymbol{X}}{\|\boldsymbol{X}\|}}(\boldsymbol{x}), \boldsymbol{x} \in \mathbb{R}^d.$$

From this we can conclude $F_R(y)F_{\frac{\boldsymbol{X}}{\|\boldsymbol{X}\|}}(\boldsymbol{x}) = F_R(y)F_{-\frac{\boldsymbol{X}}{\|\boldsymbol{X}\|}}(\boldsymbol{x}), y \in \mathbb{R}, \boldsymbol{x} \in \mathbb{R}^d$. Because R and $\dfrac{\boldsymbol{X}}{\|\boldsymbol{X}\|}$ are independent, the product of the cumulative distribution functions is the cumulative distribution of the product of the random variable and vector. We get $F_{R\frac{\boldsymbol{X}}{\|\boldsymbol{X}\|}}(\boldsymbol{x}) = F_{-R\frac{\boldsymbol{X}}{\|\boldsymbol{X}\|}}(\boldsymbol{x}), \boldsymbol{x} \in \mathbb{R} \times \mathbb{R}^d$. From the equality of the distribution function, we get

$$R\frac{\boldsymbol{X}}{\|\boldsymbol{X}\|} \stackrel{d}{=} -R\frac{\boldsymbol{X}}{\|\boldsymbol{X}\|},$$

which is the wanted result. □

Due to Lemma 3.4, under the null hypothesis it holds $\widetilde{\boldsymbol{X}} \stackrel{d}{=} -\widetilde{\boldsymbol{X}}$ and $\widetilde{\boldsymbol{X}}$ is absolutely continuous.

Therefore, we consider a random sample $Y_1, \ldots, Y_n$ with the uniform distribution on $\{-1, 1\}$ and independent of the random sample $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n$ and a random sample $R_1, \ldots, R_n$ with the uniform distribution over [0,1] which is independent of the random vectors $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n$ and also independent of $Y_1, \ldots, Y_n$. We denote

$$\widetilde{\boldsymbol{X}}_i = R_i\frac{\boldsymbol{X}_i}{\|\boldsymbol{X}_i\|}.$$

Subsequently, the distribution of a random sample $\mathcal{X}_n = (\widetilde{\boldsymbol{X}}_1, \ldots, \widetilde{\boldsymbol{X}}_n)$ and a random sample $\mathcal{X}_n^* = (\widetilde{\boldsymbol{X}}_1^*, \ldots, \widetilde{\boldsymbol{X}}_n^*)$, where $\widetilde{\boldsymbol{X}}_i^* \coloneqq Y_i\widetilde{\boldsymbol{X}}_i, i = 1, \ldots, n$, coincide under the null hypothesis. Let us denote $\mathcal{I}^+ = \{i \mid Y_i = 1\}$, $n_+ = |\mathcal{I}^+|$, and

$$\mathcal{X}^+ = \{\widetilde{\boldsymbol{X}}_i^* \mid i \in \mathcal{I}^+\}, \quad \mathcal{X}^- = \{\widetilde{\boldsymbol{X}}_i^* \mid i \notin \mathcal{I}^+\}.$$

It holds that the samples $\mathcal{X}^+$ and $\mathcal{X}^-$ are independent and have the same distribution under the null hypothesis. Therefore, we can again use the test statistic

$$Q^{(n)} = \left(\frac{nd}{n_+(n-n_+)\int_0^1 J^2(u)\mathrm{d}u}\right)\left\|\sum_{i\in\mathcal{I}^+} J\left(\frac{R_{\pm,i}^{(n)}}{n_R+1}\right)\boldsymbol{S}_{\pm,i}^{(n)} - \frac{n_+}{n}\sum_{i=1}^{n} J\left(\frac{R_{\pm,i}^{(n)}}{n_R+1}\right)\boldsymbol{S}_{\pm,i}^{(n)}\right\|^2,$$

where $R_{\pm,i}^{(n)}, \boldsymbol{S}_{\pm,i}^{(n)}$ are ranks and signs of $\widetilde{\boldsymbol{X}}_i^*, i = 1, \ldots, n$. We reject the null hypothesis $H_0 : \boldsymbol{\theta} = \boldsymbol{0}$ if $Q^{(n)}$ is greater than $(1-\alpha)$-quantile of χ_d^2 distribution.

Remark. We presented in detail the one-sample test of location with randomized signs, but it is also possible to use the one-sample test based on added zero derived under the central symmetry, see Section 3.3.

The result of this statistical test is again random. There are two sources of randomness here, the samples $Y_1, \ldots, Y_n$ and $R_1, \ldots, R_n$.

Same as before, we provide the following example to illustrate the concept and the behavior under the null hypothesis and the alternative.

Example 3.6 (One-sample test of location under angular symmetry). At first, let us explore the behavior under the null hypothesis. Consider a sample $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n$, $n = 100$ from an angularly symmetric distribution. In our case, we created an angularly symmetric distribution by taking a sample from 2-dimensional normal distribution $\mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, where $\boldsymbol{\mu} = (0, 0)^\top$ and

$$\boldsymbol{\Sigma} = \begin{pmatrix} 1 & 3 \times 0.9 \\ 3 \times 0.9 & 9 \end{pmatrix},$$

i.e., with correlation 0.9. Then we multiplied all observations with the first element greater than 0 by a random variable from a uniform distribution over [0,3] and the rest by a random variable from a uniform distribution over [0,1]. Each observation is multiplied by another new realization of the random variables from uniform distributions over [0,3], resp. [0,1]. This distribution is not centrally symmetric, which can be seen in the left subfigure of Figure 3.10. The distributions of $\boldsymbol{X}$(black color) and $-\boldsymbol{X}$(red color) are clearly different. On the other hand, after normalizing the sample, the resulting unit vectors are centrally symmetric. See the right subfigure of Figure 3.10, again with the original normalized sample in black and the normalized sample reflected through the origin colored in red.

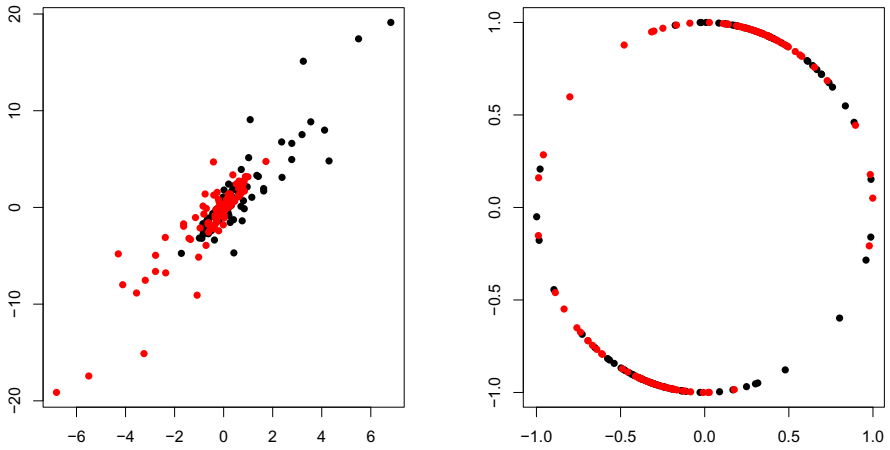

Figure 3.10: The angularly symmetric distribution described in Example 3.6. In the left subplot, there is the sample in black, and its version reflected through the origin in red. In the right subplot, the same method is done for normalized observations of the sample.

So, we have a sample $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n, n = 100$ from an angularly symmetric distribution described above. Moreover, we generate a random sample $Y_1, \ldots, Y_n, n = 100$ from a uniform distribution on $\{-1, 1\}$ and assign these signs to observation $\widetilde{\boldsymbol{X}}_1, \ldots, \widetilde{\boldsymbol{X}}_n$ computed by

$$\widetilde{\boldsymbol{X}}_i = R_i \frac{\boldsymbol{X}_i}{\|\boldsymbol{X}_i\|},$$

where $R_1, \ldots, R_n$ are independent, R_i has a uniform distribution over [0,1], and is independent of $\boldsymbol{X}_i/\|\boldsymbol{X}_i\|$. The allocation of the signs to $\widetilde{\boldsymbol{X}}_i$ is displayed in the left subplot of Figure 3.11. The observations with assigned negative signs are in blue, and the observations with assigned positive signs are in red. Then, we

create the random sample $\mathcal{X}_n^*$ by

$$\widetilde{\boldsymbol{X}}_i^* \coloneqq Y_i \widetilde{\boldsymbol{X}}_i, i = 1, \ldots, n,$$

where the blue points from the left subplot of Figure 3.11 are reflected through the origin. The result is shown in the middle subplot of Figure 3.11.

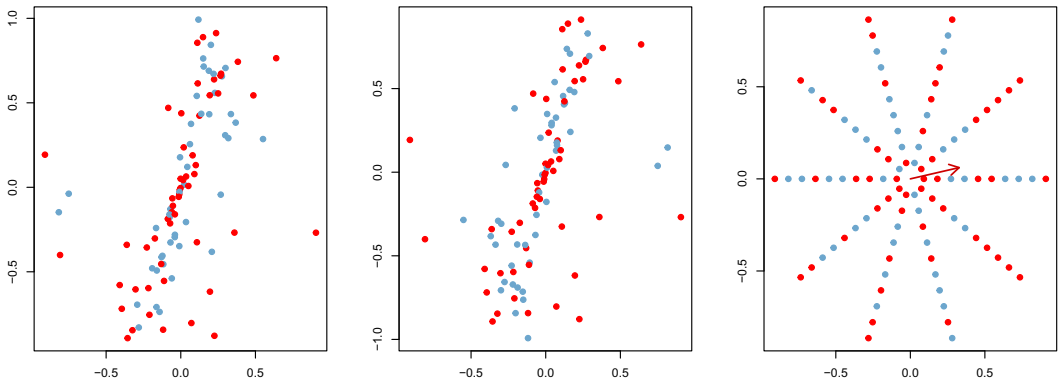

Figure 3.11: The behavior under the null hypothesis. The normalized random sample $\widetilde{\mathcal{X}}_n$ colored according to the allocation of signs in the left subplot. The random sample $\widetilde{\mathcal{X}}_n^*$ with blue points reflected through the origin in the middle. The corresponding regular grid with values of the empirical center-outward distribution function is plotted with additional vector statistic $\boldsymbol{T}^{(n)}$ from (3.7) in the right subplot.

We take the blue and red points as two separate samples and perform a two-sample test of location. In the situation in Figure 3.11, the null hypothesis should not be violated, which corresponds to the fact that the plotted vector $\boldsymbol{T}^{(n)}$ computed as in (3.7) has small norm and $Q^{(n)} = 0.665$ which is smaller than $(1-\alpha)$-quantile of χ_2^2 distribution.

Next, we simulate the behavior under the alternative. We consider a sample $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n, n = 100$, again from an angularly symmetric distribution. We created the sample $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n$ by taking a sample from 2-dimensional normal distribution $\mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, where $\boldsymbol{\mu} = (1,1)^\top$ and

$$\boldsymbol{\Sigma} = \begin{pmatrix} 1 & 3 \times 0.9 \\ 3 \times 0.9 & 9 \end{pmatrix},$$

i.e., with correlation 0.9. Then, we multiplied it by a random variable from a uniform distribution over [0,3], respectively over [0,1], depending on the sign of its first element, same as in the sample created under the null hypothesis. We again test the null hypothesis $H_0 : \boldsymbol{\theta} = (0,0)^\top$ and we proceed in a similar way as in the previous part. The resulting allocation of signs to the observations and the random sample $\mathcal{X}_n$ colored according to the allocation of signs are displayed in Figure 3.12.

In this case, it is clear that after the reflection of the blue points through the origin, we get two almost separate samples, see Figure 3.12. The vector statistic

$\boldsymbol{T}^{(n)}$ is equal to $(3.087, 0.3643)^\top$ and the test statistic is 57.986, which is greater than $(1-\alpha)$-quantile of χ^2_d distribution. Therefore, we reject the null hypothesis (3.6).

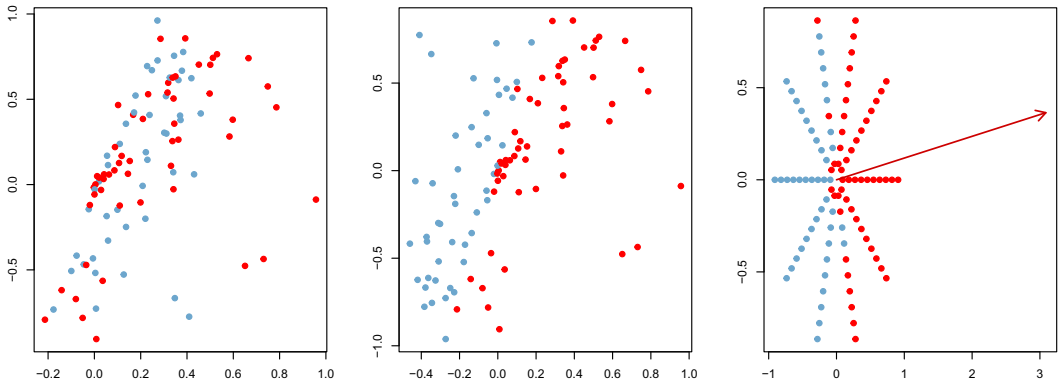

Figure 3.12: The behavior under the alternative. The random sample $\widetilde{\mathcal{X}}_n$ is colored according to the allocation of signs in the left subplot. The random sample $\widetilde{\mathcal{X}}_n^*$ with blue points reflected through the origin is in the middle. The corresponding regular grid with the values of the empirical center-outward distribution function is plotted with the additional vector statistic $\boldsymbol{T}^{(n)}$ from (3.7) in the right subplot.

4. Simulation

In the previous chapters, we presented the theory behind the tests based on the center-outward ranks and signs. In this part, we try to illustrate the performance of mentioned tests with respect to different conditions. The following study was conducted in R Core Team (2022) and will be divided into several sections considering partial tasks. To compute the center-outward distribution function, we use the function `solve_LSAP` from package `clue`, see Hornik (2023) and Hornik (2005). For generating the Halton sequences, the function `ghalton` from library `qrng` was used, see Hofert & Lemieux (2020).

4.1 Factorization and performance of different grids

In this section, we present the results of the simulation study concerning the power of the one-sample test of location with randomized signs with the identity scores function $J(u) = u$, see Section 3.3.1 while using different grids for the empirical center-outward distribution function.

We generate the data under the null hypothesis from a centered multivariate normal distribution with the correlation between all marginals equal to 0.7 and variances of marginals equal to 1. Then, we also generate data from the same distribution under several alternatives by taking a shift in the form of a d-dimensional vector with the first element shifted by δ, i.e., $s = (\delta, 0, \ldots, 0)^\top$ for

$$\delta \in \{0.05, 0.1, 0.15, 0.2, 0.25, 0.3, 0.35, 0.4\}.$$

The following grids in the following dimensions were considered (for more information on the grids, see Section 2.5):

1. For $d = 2$:
    - **regular grid** – created as an intersection of unit vectors creating arcs of equal length $2\pi/n_S$ and spheres with radii $r_j = j/(n_R + 1)$, $j = 1, \ldots, n_R$,
    - **grid with random ranks** – created as an intersection of unit vectors creating arcs of equal length $2\pi/n_S$ and spheres with radii r_j, $j = 1, \ldots, n_R$, where r_j are realizations of n_R random variables with a uniform distribution over $[0, 1]$,
    - **grid with random signs** – created as an intersection of randomly chosen unit vectors and spheres with radii $r_j = j/(n_R + 1), j = 1, \ldots, n_R$,
    - **grid with random ranks and signs** – created as an intersection of randomly chosen unit vectors and spheres with radii $r_j, j = 1, \ldots, n_R$, where r_j are realizations of n_R random variables with a uniform distribution over $[0, 1]$,

- **grid with random ranks and signs separately** – created as $g_i = r_i s_i$, $i = 1, \ldots, n$, where r_i are independent identically distributed random variables from the uniform distribution over [0, 1] and s_i are independent identically distributed random variables from the uniform distribution over unit sphere $\mathcal{S}_1$.

2. For $d = 4$:

   - **grid with uniform signs and regular ranks** – created as an intersection of the unit vectors generated from the uniform distribution on $\mathcal{S}_{d-1}$ and hyperspheres with radii $r_j = j/(n_R + 1), j = 1, \ldots, n_R$,
   - **grid with uniform signs and random ranks** – created as an intersection of the unit vectors generated from the uniform distribution on $\mathcal{S}_{d-1}$ and hyperspheres with radii $r_j, j = 1, \ldots, n_R$, where r_j are realizations of n_R random variables with a uniform distribution over $[0, 1]$,
   - **grid with Halton signs and regular ranks** – created as an intersection of the unit vectors generated by a transformation of the Halton sequence in $\mathbb{R}^3$ onto a hypersphere in $\mathbb{R}^4$ and hyperspheres with radii $r_j = j/(n_R + 1), j = 1, \ldots, n_R$,
   - **grid with Halton signs and random ranks** – created as an intersection of the unit vectors generated by a transformation of the Halton sequence in $\mathbb{R}^3$ onto a hypersphere in $\mathbb{R}^4$ and hyperspheres with radii $r_j, j = 1, \ldots, n_R$, where r_j are realizations of n_R random variables with uniform distribution over $[0, 1]$.
   - **grid with Halton signs and random ranks separately** – created as $g_i = r_i s_i, i = 1, \ldots, n$, where r_i are independent identically distributed random variables from the uniform distribution over $[0, 1]$ and s_i are the unit vectors generated by transforming the Halton sequence in $\mathbb{R}^3$ onto a hypersphere in $\mathbb{R}^4$.

The power of the chosen test was computed from 500 replications and for the following sample sizes and factorizations:

1. For $\boldsymbol{n = 100}$:
   - $n_R = n_S = 10$,
   - $n_R = 20, n_S = 5$,
   - $n_R = 5, n_S = 20$.
2. For $\boldsymbol{n = 400}$:
   - $n_R = n_S = 20$,
   - $n_R = 40, n_S = 10$,
   - $n_R = 10, n_S = 40$.

The results are summarized in Figures 4.1, 4.2, 4.3, 4.4. In Figures 4.1 and 4.3, there are greater differences between the grids compared to Figures 4.2 and 4.4, where $n = 400$. For $d = 2$, we recommend using the grid with random ranks and signs separately. It performs well for $n = 100$, and for $n = 400$ the differences between grids completely disappear. For $d = 4$, we recommend using the grid with random ranks and signs separately or the grid with Halton signs and random ranks separately because of good performance for both $n = 100$ and $n = 400$. For $d = 4$, we observe the differences between the grids even for $n = 400$. The performance of the given test for different factorizations of the grid seems to be quite similar both for $d = 2$ and $d = 4$. For the next parts of the simulation, we chose the square root factorization, i.e., for $n = 100$ $n_R = n_S = 10$ and for $n = 400$ $n_R = n_S = 20$.

4.2 Asymptotics of the two-sample test of location

In this part, we illustrate the behavior of the test statistic from the two-sample test of location described in Section 3.2. We have two random samples $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_{n_1}$ and $\boldsymbol{Y}_1, \ldots, \boldsymbol{Y}_{n_2}$ from a d-dimensional distribution with continuous distribution functions. The null hypothesis of no shift in the samples and the alternative are

$$H_0 : \boldsymbol{\theta} = \mathbf{0} \text{ versus } H_1 : \boldsymbol{\theta} \neq \mathbf{0}.$$

For more details, see Section 3.2.

In Section 3.2.1, we have derived that the test statistic $Q^{(n)}$ (3.5) has asymptotically χ^2_d distribution, where d is the dimension corresponding to the dimension of the two tested samples. We have chosen the following initial parameters for the simulation. We use $n_1 = n_2 = n/2$:

1. $d = 2$, a grid with random ranks and signs separately with factorization $n_1 + n_2 = n = n_R n_S$ chosen in a way that
   - $n = 16, n_R = n_S = 4$,
   - $n = 36, n_R = n_S = 6$,
   - $n = 64, n_R = n_S = 8$,
   - $n = 100, n_R = n_S = 10$,
   - $n = 144, n_R = n_S = 12$,
2. $d = 4$, a grid with Halton signs and random ranks separately with factorization $n_1 + n_2 = n = n_R n_S$ chosen in a way that
   - $n = 16, n_R = n_S = 4$,
   - $n = 36, n_R = n_S = 6$,
   - $n = 64, n_R = n_S = 8$,
   - $n = 100, n_R = n_S = 10$,
   - $n = 144, n_R = n_S = 12$,
   - $n = 196, n_R = n_S = 14$,
   - $n = 400, n_R = n_S = 20$.

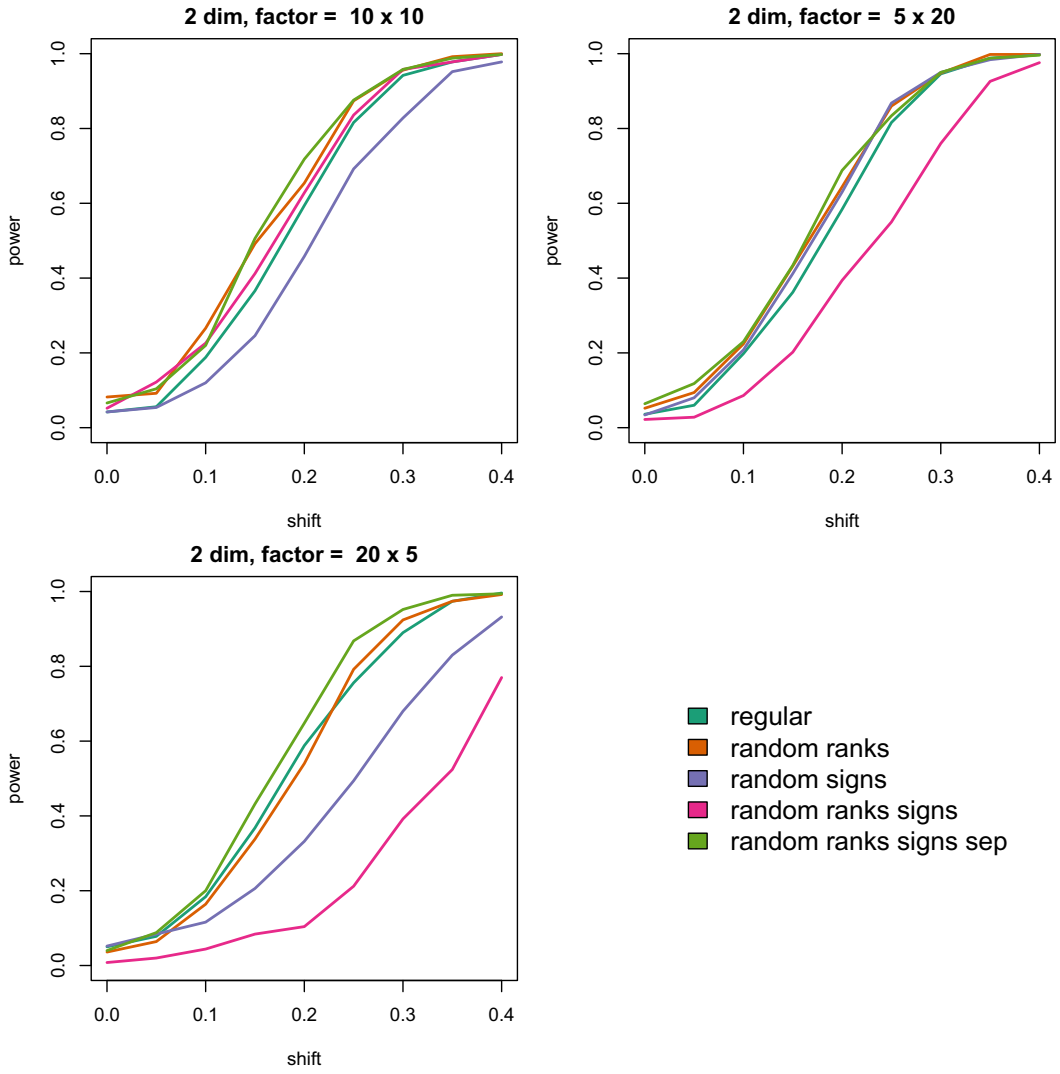

Figure 4.1: Comparison of powers of the one-sample test with randomized signs for different grids and different factorizations computed out of a sample of size 100 in dimension 2.

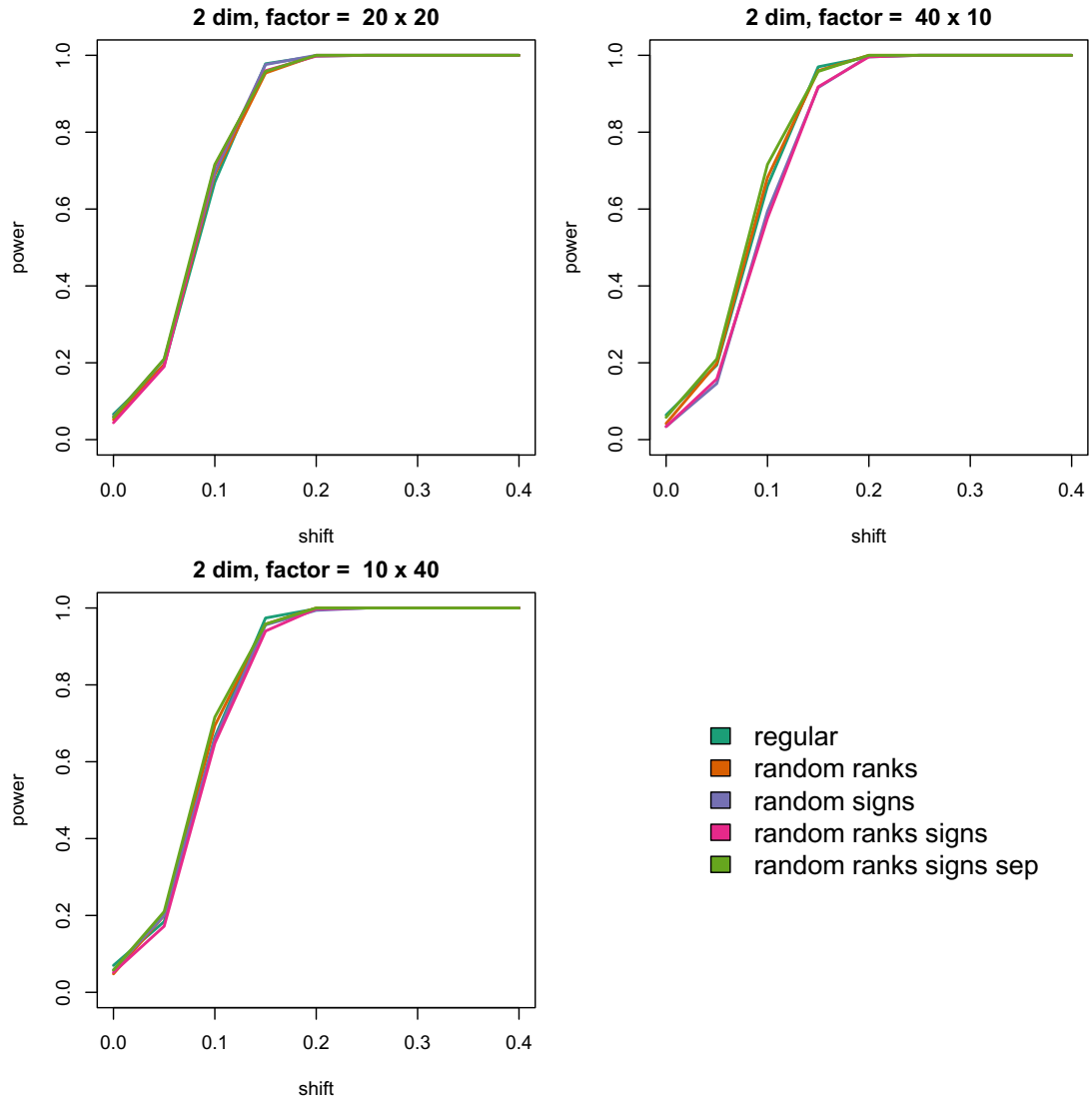

Figure 4.2: Comparison of powers of the one-sample test with randomized signs for different grids and different factorizations computed out of a sample of size 400 in dimension 2.

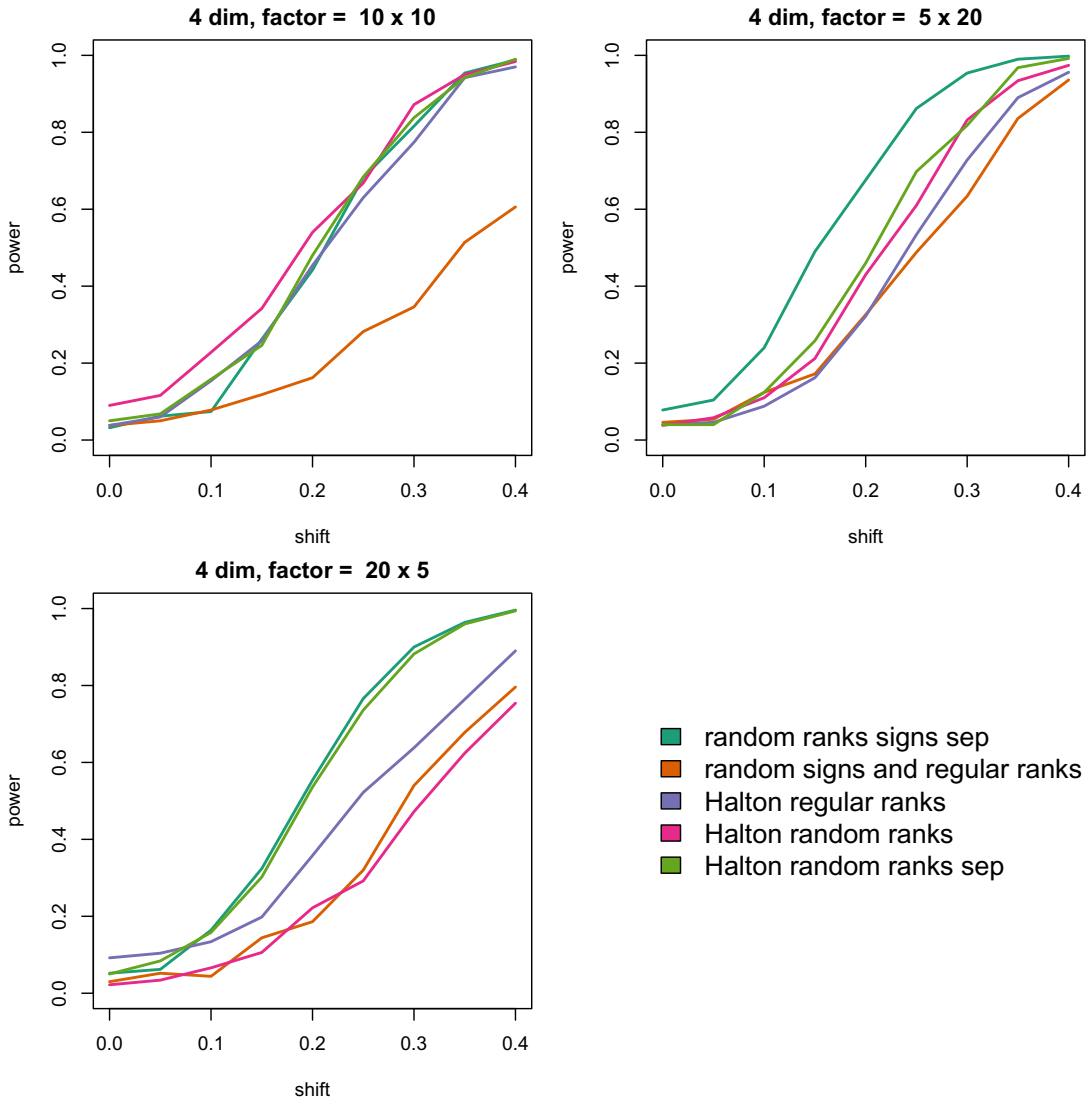

Figure 4.3: Comparison of powers of the one-sample test with randomized signs for different grids and different factorizations computed out of a sample of size 100 in dimension 4.

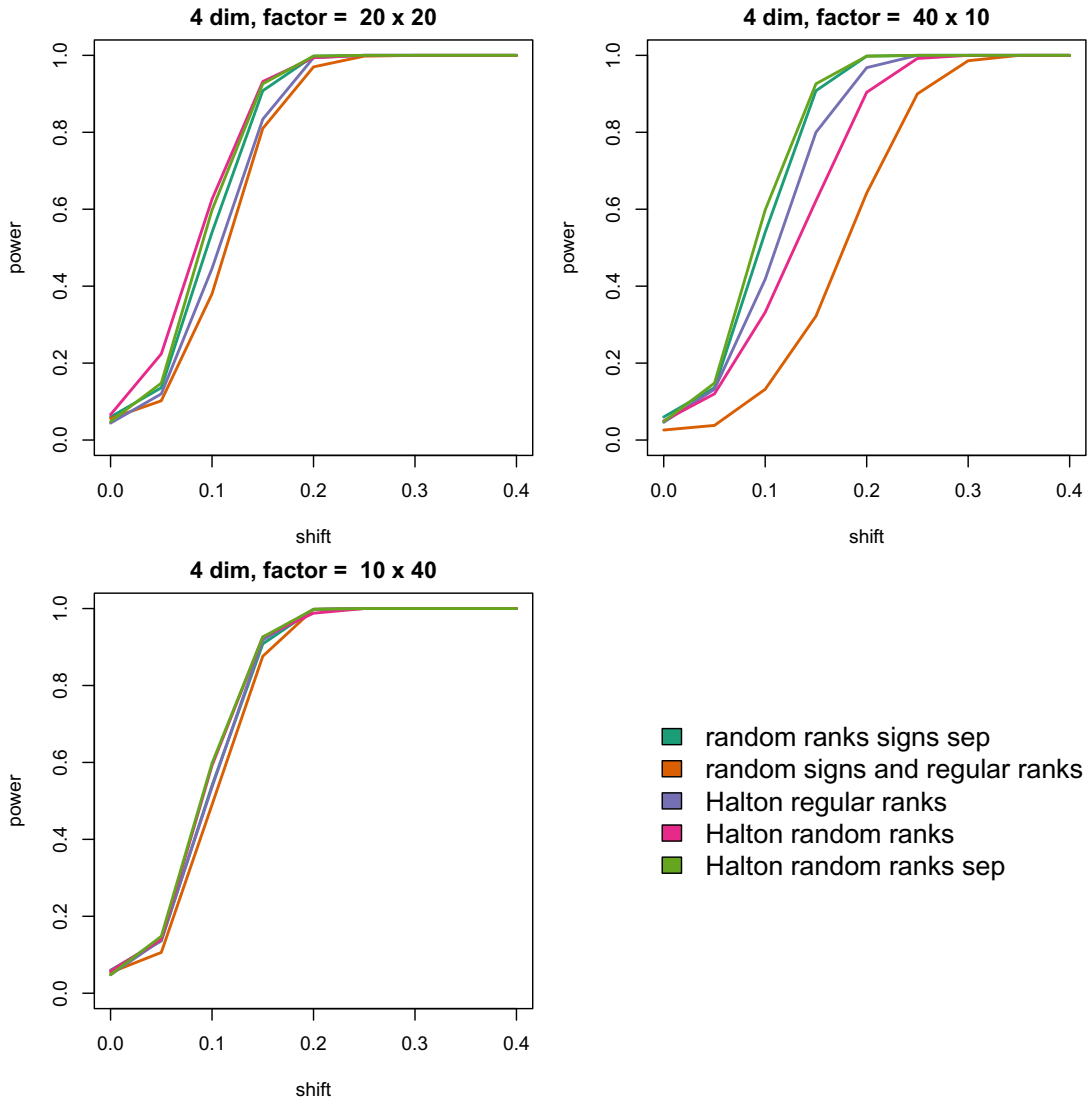

Figure 4.4: Comparison of powers of the one-sample test with randomized signs for different grids and different factorizations computed out of a sample of size 400 in dimension 4.

The test statistic from the two-sample test of location was computed for the previous initial parameters and 500 replications. An empirical distribution function of the test statistic $Q^{(n)}$ is presented for each set of parameters. The results are shown in Figures 4.5 and 4.6.

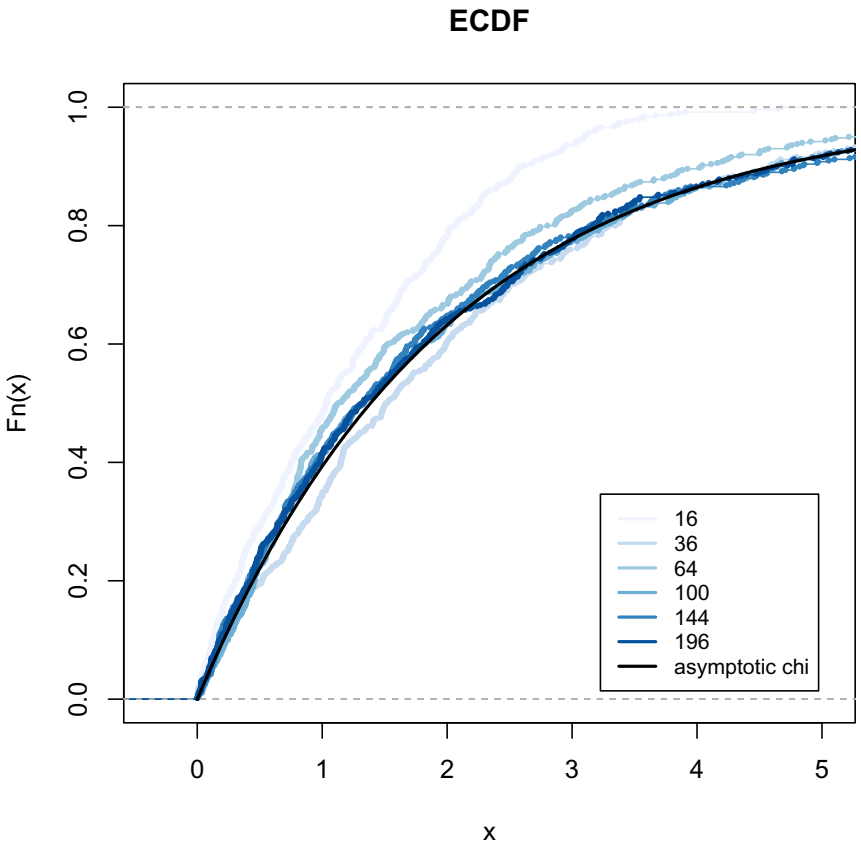

Figure 4.5: The values of the empirical distribution function for the test statistic $Q^{(n)}$ from the two-sample test of location in dimension 2 for different sample sizes. We chose a grid with random ranks and signs separately in $\mathbb{R}^2$, see Section 2.5. The asymptotic χ^2_2 distribution function is added as the black line.

From Figures 4.5 and 4.6, we observe that in dimension 2 for n greater than 100 the asymptotic χ^2_2 distribution seems to approximate the real distribution of the test statistic well enough. In dimension 4, it seems that we need a greater size of samples for the asymptotic distribution to well approximate the real distribution of the test statistic. The simulations indicate that for $n \geq 400$, the approximation by the asymptotic distribution is good enough.

4.3 One-sample test of location

In this part, we try to illustrate the behavior of the proposed one-sample tests of location. From the previous Section 4.1, we chose the factorizations of grid $n = 100 = 10 \times 10 = n_R \times n_S$ and $n = 400 = 20 \times 20 = n_R \times n_S$. We perform the test in dimensions 2 and 4 for the grid with random ranks and signs separately.

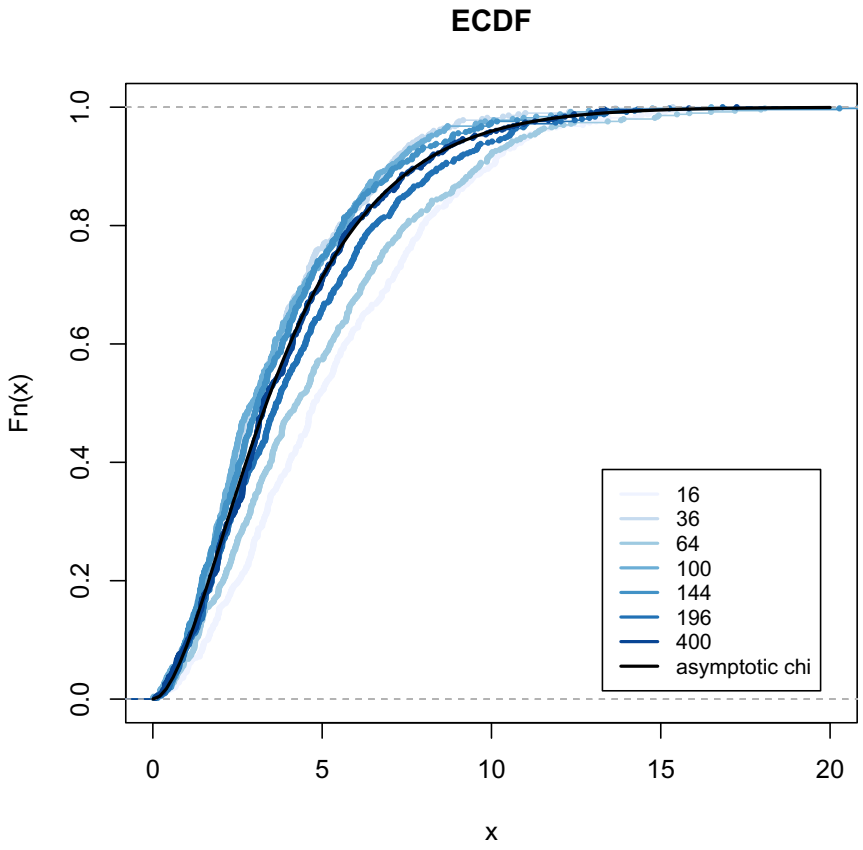

Figure 4.6: The values of the empirical distribution function for the test statistic $Q^{(n)}$ from the two-sample test of location in dimension 4 for different sample sizes. We chose a grid with Halton signs and random ranks separately in $\mathbb{R}^4$, see Section 2.5. The asymptotic χ^2_4 distribution function is added as the black line.

We generate the data from the following distributions:

- mixture of $\mathcal{L}(\boldsymbol{X})$ and $\mathcal{L}(-\boldsymbol{X})$ with the same weights and $\boldsymbol{X}$ with marginals, with exponential distribution with the rate 2, correlation equal to 0.9, and connected through a normal copula,
- multivariate centered normal distribution with identity variance matrix,
- multivariate t distribution with degrees of freedom equal to 1 and identity scale matrix,
- multivariate t distribution with degrees of freedom equal to 2 and identity scale matrix.

All mentioned distributions were generated under the null hypothesis and several alternatives by taking a shift in the form of a d-dimensional vector with all elements shifted by δ, i.e., $s = (\delta, \ldots, \delta)^\top$ for

$$\delta \in \{0, 0.05, 0.1, 0.15, 0.2, 0.25, 0.3, 0.35, 0.4\}.$$

We observed the power of the following tests with the corresponding abbreviations, used in Figures, in parentheses:

- Hotelling's test (hotelling),
- permutation test with added zero (perm0),
- test with randomized signs (rand_signs),
- test with randomized signs under angular symmetry (angular).

For the test with randomized signs under angular symmetry, we chose R from a uniform distribution over $[0, 1]$, same as in Example 3.6. For Hotelling's test, the test with randomized signs, and the test with randomized signs under angular symmetry, we computed 500 replications to get the power of the tests. For the permutation test with added zero from Section 3.3.2, we computed 100 replications, for $n = 100$, we chose $B = 999$ while for $n = 400$, we chose only $B = 199$ because of the high computational complexity. The results are summarized in Figures 4.7, 4.8, 4.9, 4.10.

From the previous Figures 4.7, 4.8, 4.9, 4.10, we observe that for both dimensions and both sample sizes, the power of Hotelling's test for t distribution with 1 degree of freedom is lower compared to the other one-sample tests of location. The other three one-sample tests have similar power, except for the situation in dimension 4 and sample size 400. We would recommend the one-sample test of the location under the angular symmetry for its best performance.

For a normal distribution, the power of the permutation test with added zero is lower in both dimensions and for both sample sizes. Also, the one-sample test under angular symmetry seems to have lower power in all cases. The one-sample test with randomized signs achieves results as good as Hotelling's test despite weaker assumptions.

For t distribution with 2 degrees of freedom, we would recommend the one-sample test with randomized signs or the one-sample test under angular symmetry. The difference in power for Hotelling's test and the permutation test with added zero compared to the one-sample test with randomized signs or the one-sample test under angular symmetry is smaller for sample size 100. For $n = 400$ and dimension 4, the one-sample test under the angular symmetry gets the best results considering the power.

For the mixture distribution, all compared tests perform really well, the differences might be found for sample size 100. In dimension 4 and with sample size 100, the permutation test with added zero seems to fail. The power of the permutation test with added zero also looks not as smooth as the other ones. This might be partly caused by the lower number of chosen permutations ($B = 199$) and the lower number of repetitions (100).

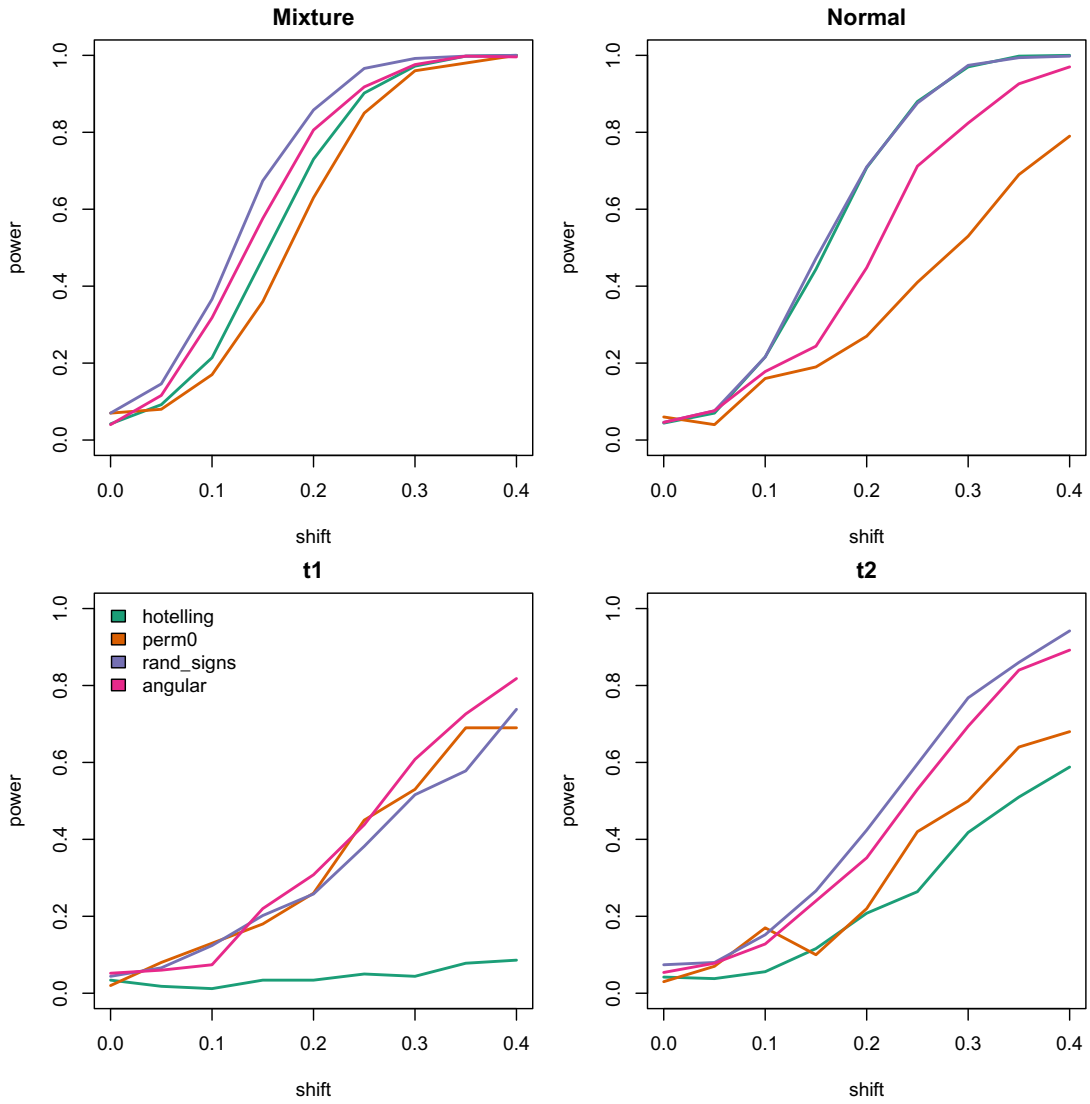

Figure 4.7: Comparison of powers of the one-sample tests of location for different distributions computed out of a sample of size 100 in dimension 2.

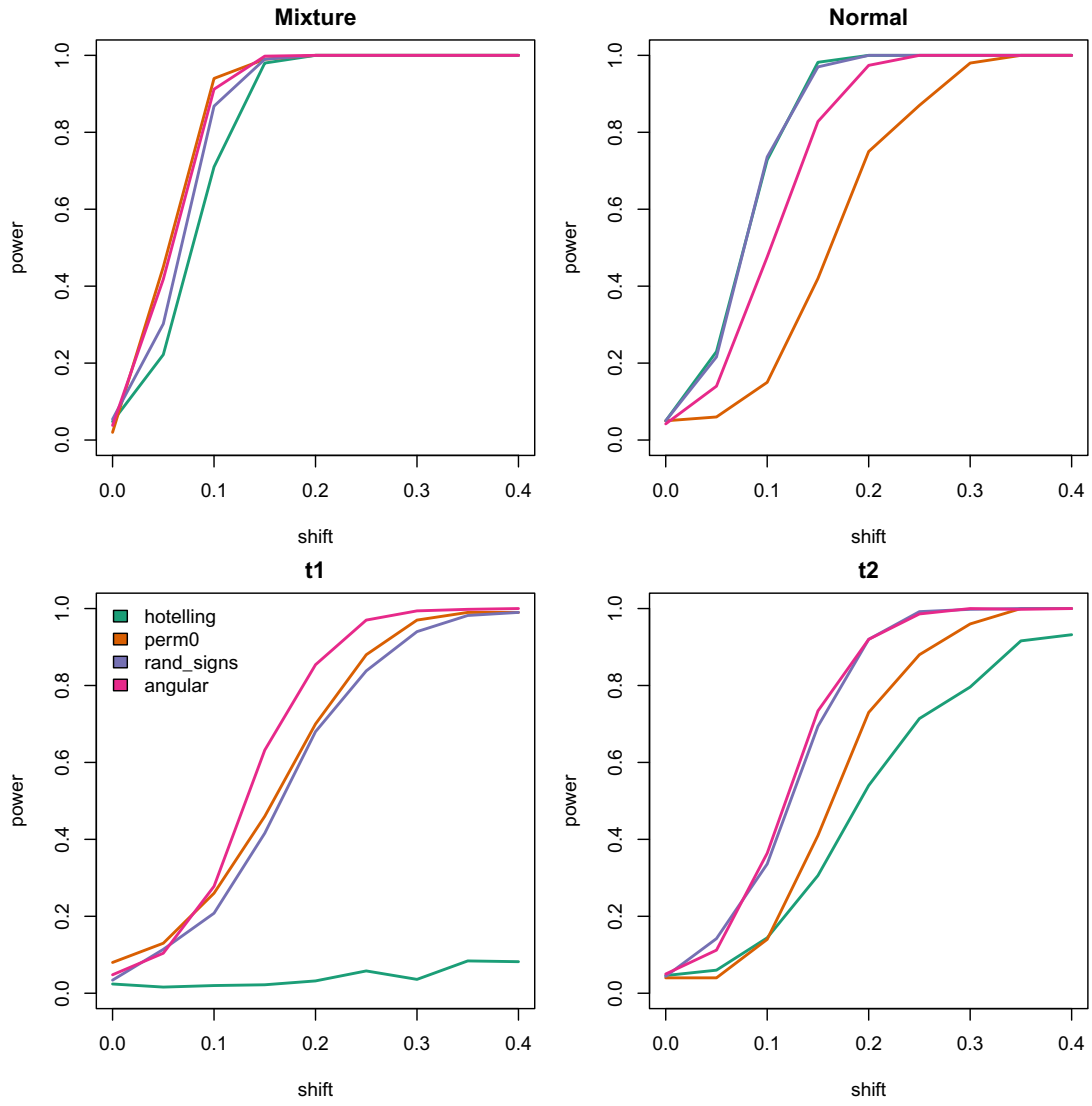

Figure 4.8: Comparison of powers of the one-sample tests of location for different distributions computed out of a sample of size 400 in dimension 2.

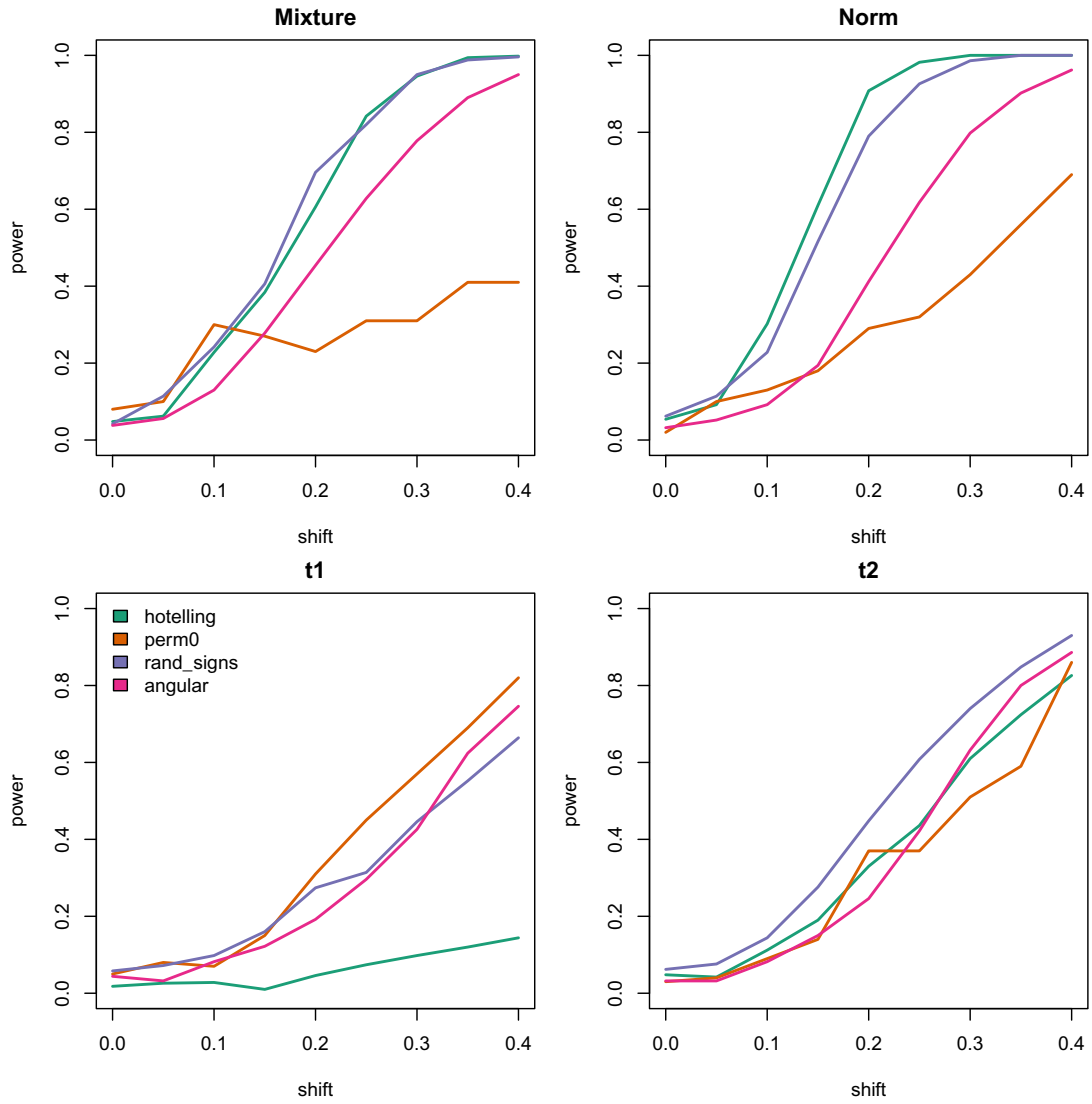

Figure 4.9: Comparison of powers of the one-sample tests of location for different distributions computed out of a sample of size 100 in dimension 4.

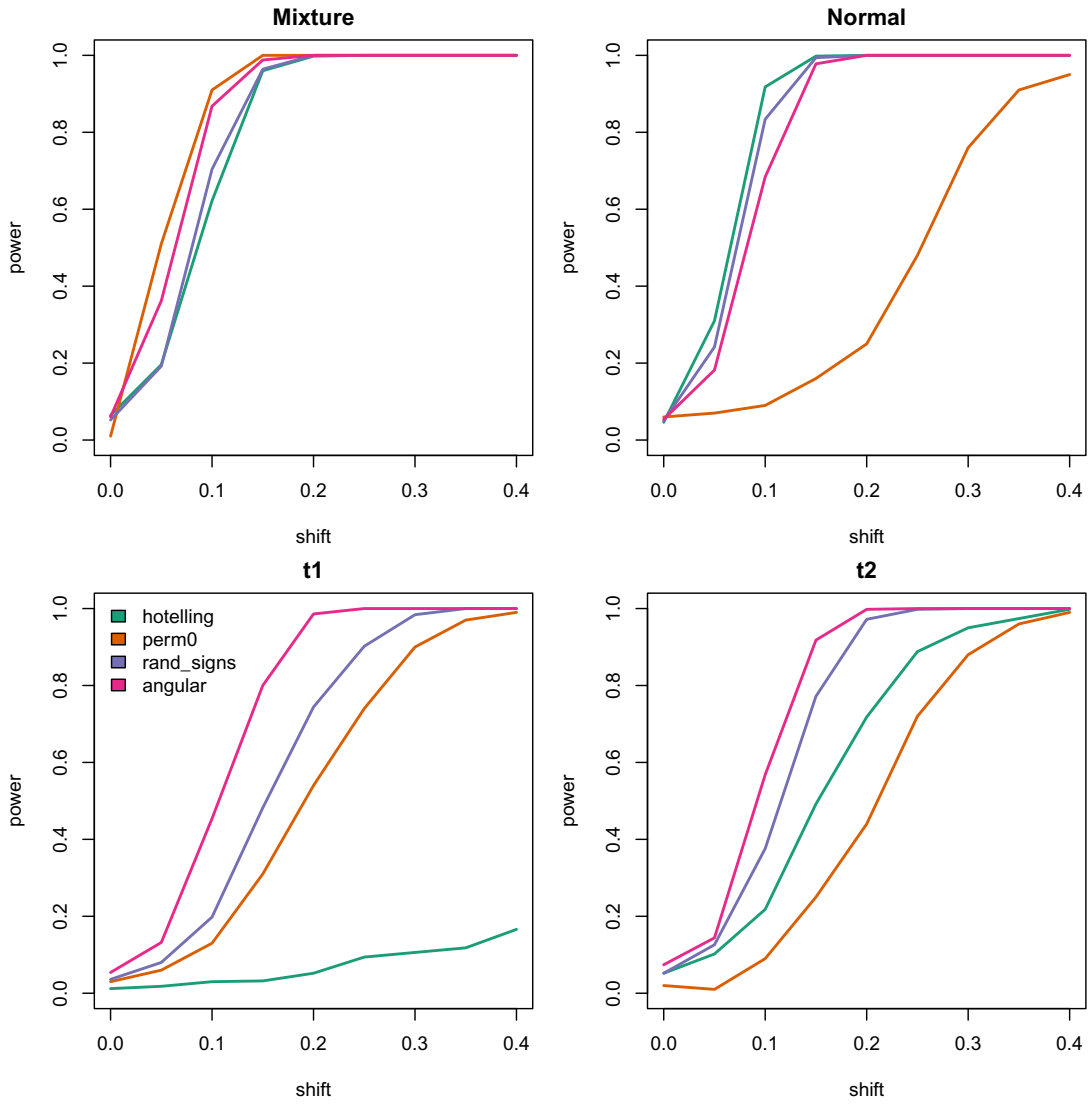

Figure 4.10: Comparison of powers of the one-sample tests of location for different distributions computed out of a sample of size 400 in dimension 4.

4.4 Angular symmetry

This part contains a simulation study for data with angularly symmetric but not centrally symmetric distribution, similarly as in Example 3.6. The angularly symmetric data are created by taking a sample from d-dimensional normal distribution $\mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, where

- for $d = 2$:
  $\boldsymbol{\mu} = (0, 0)^\top$ and
  $$\boldsymbol{\Sigma} = \begin{pmatrix} 1 & 3 \times 0.9 \\ 3 \times 0.9 & 9 \end{pmatrix},$$
  i.e., with correlation 0.9, and variances 1 and 9,
- for $d = 4$: $\boldsymbol{\mu} = (0, 0, 0, 0)^\top$ and
  $$\boldsymbol{\Sigma} = \begin{pmatrix} 1 & 3 \times 0.9 & 1 \times 0.9 & 3 \times 0.9 \\ 3 \times 0.9 & 9 & 3 \times 0.9 & 9 \times 0.9 \\ 1 \times 0.9 & 3 \times 0.9 & 1 & 3 \times 0.9 \\ 3 \times 0.9 & 9 \times 0.9 & 3 \times 0.9 & 9 \end{pmatrix},$$
  i.e., with correlation 0.9, and variances 1, 9, 1 and 9.

Then, we multiplied all observations with the first element greater than 0 by a random variable (for each observation new independent one, see Example 3.6) from a uniform distribution over [0,3] and the rest by a random variable from a uniform distribution over [0,1]. This distribution is not centrally symmetric but angularly symmetric. We generate the data under the null hypothesis and various alternatives by adding a shift in the form of a d-dimensional vector $(\delta, \ldots, \delta)^\top$, where
$$\delta \in \{0, 0.05, 0.1, 0.15, 0.2, 0.25, 0.3, 0.35, 0.4\}.$$
We compare several settings:

- $d = 2$:
  - $n = 100, n_R = n_S = 10$,
  - $n = 400, n_R = n_S = 20$,
- $d = 4$:
  - $n = 100, n_R = n_S = 10$,
  - $n = 400, n_R = n_S = 20$.

For these four settings, we compare the power of the following tests with the corresponding abbreviations in parentheses:

- Hotelling's test (hotelling),
- test with random signs (rand_signs),
- test with random signs under angular symmetry (angular)
  with $R_i, i = 1, \ldots, n$ from the definition of $\widetilde{\boldsymbol{X}}_i$, see Section 3.4, generated from the uniform distribution over [0,1],

- test with random signs under angular symmetry (angular_exp)
  with $R_i, i = 1, \ldots, n$ from the definition of $\widetilde{\boldsymbol{X}}_i$, see Section 3.4, generated from the exponential distribution with a rate equal to 1.

The powers are computed out of 500 replications for all these different types of tests. The results are plotted in Figure 4.11.

The results in Figure 4.11 show that for data from angularly symmetric distribution described in this part of the simulation study, the usage of Hotelling's test or the one-sample test with randomized signs is wrong. Both of these tests have a much higher size of the test under the null hypothesis. For $n = 400$, the power of both tests is 1, no matter the null hypothesis.

This is caused by the fact that in this case, the mean does not correspond to the center of angular symmetry and the data are not centrally symmetric. Therefore, testing the null hypothesis of zero mean or center of central symmetry is completely different from testing the center of angular symmetry. It is important to determine meaningfully what we mean by the location of the sample. Both versions of the one-sample test under the angular symmetry are similar, considering the power for all different settings, and both perform well.

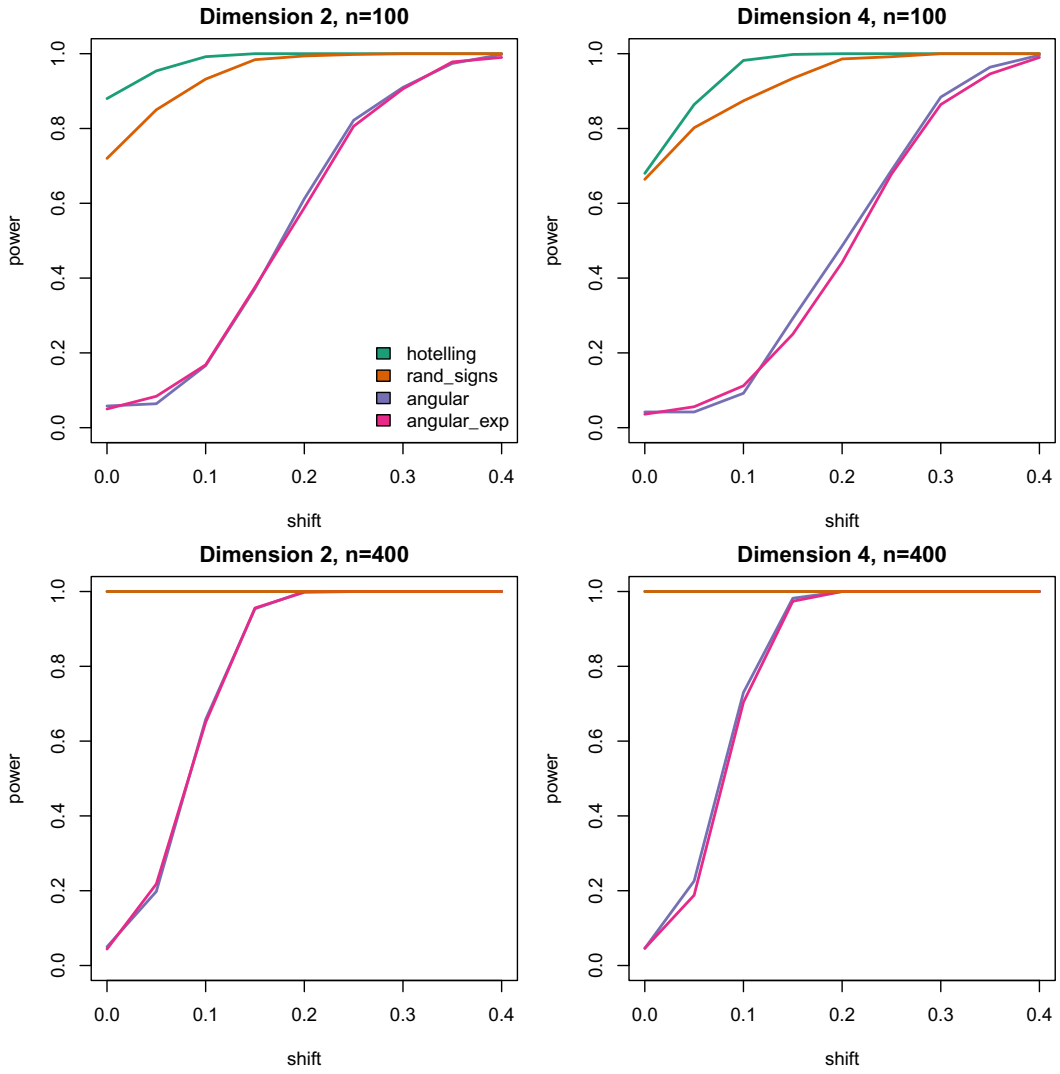

Figure 4.11: Comparison of powers of the one-sample tests of location for angularly but not centrally symmetric distributions computed out of a sample of sizes 100 and 400 in dimensions 2 and 4.

Conclusion

In this thesis, we used the theory of center-outward ranks and signs to introduce and then compare several one-sample tests of location. In the beginning, the theory and main properties of rank-based tests were presented. From the several possible approaches discussed in the introduction, we have chosen the center-outward ranks and signs. The definition was given and its main properties were discussed. The concept of the center-outward ranks and signs is connected with the underlying grid and the measure transportation. We provided several ways how to construct such grids both in 2-dimensional and multidimensional spaces.

In the main part of the thesis, we introduced test statistics based on the previous theory. Their asymptotic normality was shown and used to derive statistical tools for testing. These test statistics were then used for the two-sample test of location. Another possible way to test the same hypothesis was proposed using the permutation test.

The main contribution of this thesis is the proposal of the one-sample tests of location under central and angular symmetry. The first approach is based on the randomized assignment of signs to data, reflection through the origin, and using the two-sample test for the two samples created by allocation of the signs. The next one is adding a zero observation to the sample, transporting the new sample onto the underlying grid, and taking the norm of the empirical distribution function of the zero observation. For evaluating the p-value, the permutation test is used again. All proposed tests are supplemented by illustrative examples.

In the end, we performed a simulation study to compare the presented tests under different settings and using different alternatives. At first, different grids were studied. The results were more different for smaller sample sizes and for higher dimensions. The grid with random ranks and signs taken separately seemed to work fine for both dimensions. For the rest of the simulation, we chose factorization given by square roots. The asymptotics of the two-sample test of location was illustrated for both dimensions, 2 and 4.

The comparison of the one-sample tests of location was added for several distributions with central symmetry as well as for one angularly but not centrally symmetric. For a normal distribution, the one-sample test with randomized signs achieved results as good as Hotelling's test, despite the weaker assumptions. For t distribution with 1 degree of freedom, Hotelling's test failed, but high power was achieved by the one-sample test under angular symmetry as well as the other presented one-sample tests. For data generated from the distribution angularly but not centrally symmetric, the one-sample test under angular symmetry performed well for R_i generated from both the uniform and the exponential distribution. Hotelling's test and the one-sample test with randomized signs failed in this case, probably due to the different null hypotheses. Moreover, the one-sample test under angular symmetry we introduced in this work performed well even in the case of an elliptical distribution with heavy tails.

The one-sample test with added zero performs well for the mixture distribution, which is non-elliptical. Compared to Hotelling's test, it performs better for distributions with heavy tails (t with 1 degree of freedom). On the other hand, we would not recommend it in the case of normal distribution. Also, one must keep in mind that the permutation test with added zero is computationally complex, and it might take time to get the desired p-values.

Bibliography

Anděl, J. (2007), *Základy matematické statistiky*, 2nd edn, Matfyzpress, Praha.

Chernozhukov, V., Galichon, A., Hallin, M. & Henry, M. (2017), 'Monge–Kantorovich depth, quantiles, ranks and signs', *The Annals of Statistics* **45**(1), 223 – 256.

Davison, A. C. & Hinkley, D. V. (1997), *Bootstrap Methods and their Application*, Cambridge Series in Statistical and Probabilistic Mathematics, Cambridge University Press.

Dudley, R. M. (2014), *Uniform Central Limit Theorems*, Cambridge Studies in Advanced Mathematics, 2nd edn, Cambridge University Press.

Fang, K.-T., Kotz, S. & Ng, K. W. (1990), *Symmetric Multivariate and Related Distributions*, Monographs on Statistics and Applied Probability, Springer US.

Fang, K.-T. & Wang, Y. (1994), *Number-theoretic Methods in Statistics*, Monographs on Statistics and Applied Probability 51, Springer US.

Figalli, A. (2018), 'On the continuity of center-outward distribution and quantile functions', *Nonlinear Analysis* **177**, 413–421.

Hallin, M. (2022), 'Measure transportation and statistical decision theory', *Annual Review of Statistics and its Application* **9**, 401–424.

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', *The Annals of Statistics* **49**(2), 1139–1165.

Hallin, M., Hlubinka, D. & Hudecová, Š. (2022), 'Efficient fully distribution-free center-outward rank tests for multiple-output regression and manova', *Journal of the American Statistical Association* pp. 1–17.

Hallin, M., Liu, H. & Verdebout, T. (2022), 'Nonparametric measure-transportation-based methods for directional data', *arXiv* .

Hofert, M. & Lemieux, C. (2020), *qrng: (Randomized) Quasi-Random Number Generators*. R package version 0.0-8.
URL: *https://CRAN.R-project.org/package=qrng*

Hornik, K. (2005), 'A CLUE for CLUster Ensembles', *Journal of Statistical Software* **14**(12).

Hornik, K. (2023), *clue: Cluster Ensembles*. R package version 0.3-64.
URL: *https://CRAN.R-project.org/package=clue*

Hájek, J. & Šidák, Z. (1967), *Theory of Rank Tests*, 1st edn, Academic Press, New York.

Hájek, J., Šidák, Z. & Sen, P. K. (1999), *Theory of Rank Tests*, Probability and Mathematical Statistics, 2nd edn, Academic Press, San Diego.

Li, J. & Liu, R. Y. (2004), 'New Nonparametric Tests of Multivariate Locations and Scales Using Data Depth', *Statistical Science* **19**(4), 686 – 696.

Liu, R. Y. & Singh, K. (1993), 'A quality index based on data depth and multivariate rank tests', *Journal of the American Statistical Association* **88**(421), 252–260.

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

R Core Team (2022), *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria.
URL: *https://www.R-project.org/*

Serfling, R. J. (2006), 'Multivariate symmetry and asymmetry', *Encyclopedia of statistical sciences* **8**, 5338–5345.

Villani, C. (2003), *Topics in Optimal Transportation*, Graduate Studies in Mathematics, American Mathematical Society.