Aalto University
School of Science
Degree Programme of Engineering Physics and Mathematics

Several Sample Location Problem

Mat-2.4108 Independent Research Project in Applied Mathematics

Author:
Sakke Rantala

Instructor & Supervisor:
Prof. Pauliina Ilmonen

June 11, 2015

Contents

1 Introduction

This work studies the location problem of several samples of multivariate data. The purpose is to find a test statistic that can be applied even though the common normality assumptions of multivariate data sets would not hold or the sample sizes are small. Therefore, in addition to the classical multivariate analysis of variance MANOVA, test statistics based on non-parametric score functions known as spatial signs and spatial ranks are considered. The different test statistics are compared using simulated data-sets. The theory is mainly based on the book by Oja [2010] and the focus is on its chapter 11.

The practical frame of reference of this work is to evaluate changes in the quality of a complex process. The process considered in this work is producing energy in a hydro-power cascade. The data of the process is not necessarily normally distributed but can be heavily tailed, due to the nature of the process and due to some technical limitations. The problem of selection and preprocessing of relevant data is discussed shortly utilizing a few simple examples.

As a conclusion of this work, the use of the non-parametric methods are found useful in some cases. Especially the test based on the spatial ranks offers an alternative to the classical MANOVA in the situations where the multinormality assumptions do not exactly hold or the sample sizes are small. The test based on spatial signs was found to be the most robust one, which may lead to losing relevant information regarding to the problem. Still, general rules of application cannot be provided. One drawback of the non-parametric methods presented is that the test statistics are not affine invariant. However, they are location invariant and scalar invariant which is sufficient in numerous practical applications.

2 Notations and definitions of key concepts

Affine map The mapping $f : \mathbb{R}^n \to \mathbb{R}^n$ is called affine if and only if

$$f(\mathbf{Ax} + \mathbf{b}) = \mathbf{A}f(\mathbf{x}) + \mathbf{b},\ \forall \mathbf{A} \in \mathbb{R}^{n \times n},\ \forall \mathbf{b} \in \mathbb{R}^n$$

The linear mapping ($\mathbf{b} = \mathbf{0}$) is a special case of affine mapping.

$\xrightarrow{d}$ The sequence of random variables $X_1, \ldots, X_n$ *converges in distribution* to a random variable X, denoted by $X_n \xrightarrow{d} X$, if and only if their cumulative distribution functions satisfy the following:

$$\lim_{n \to \infty} F(X_n) = F(X)$$

The definition is based on distributions only, and hence, the sample spaces of X_i do not necessarily need to be the same.

$\xrightarrow{p}$ The sequence of random variables $X_1, \ldots, X_n$ defined in the same sample space *converges in probability* to a random variable X, denoted by $X_n \xrightarrow{p} X$, if and only if they satisfy the following $\forall\ \epsilon > 0$:

$$\lim_{n \to \infty} Pr(|X_n - X| > \epsilon) = 0$$

COV The covariance matrix, $\mathbf{COV}(\mathbf{y}) := E[(\mathbf{y} - E[\mathbf{y}])(\mathbf{y} - E[\mathbf{y}])^T]$

F_X *The cumulative distribution function* (cdf) of a random variable X determines the probability of the variable X being less than a real number x. More formally, for a p-variate random variable $\mathbf{X}$, $F : \mathbb{R}^p \to [0, 1]$ is its *cdf* if and only if $F_X(x) = Pr(X < x)$. Note, F_X totally determines the random variable X.

X Data matrix of p-variate observations has the dimensions $(n \times p)$. If the data contains samples of sizes $n_1, \ldots, n_c$, then $n = n_1 + \cdots + n_c$. For more information, see the chapter 4.

:= is a notation to separate *definition* from "equals", denoted by "=". I.e. consider the difference between the definition of a new variable $a := b^c$ and the strict equality $1 = 1$.

ν Calculating a statistic, the *degrees of freedom*, denoted commonly as ν, determines the number of values that are free to vary.

$\mathbf{I}_n$, $\mathbf{I}$ The diagonal elements of $(n \times n)$-sized *identity matrix* equal one, all the others are zero. In case n is obvious, it might be omitted.

p-value relates to statistical tests. It is the probability that the test statistic applied would obtain similar or more extreme results under the assumption of the null hypothesis. Thus, small probabilities give evidence to reject the null hypothesis.

$T(\mathbf{x})$ *Score function* maps the observations into scores that are utilized in analysis.

$\mu(\mathbf{X})$ *Spatial median* generalizes the concept of median in a multivariate case, see the definition 4.3.

$U(\mathbf{x})$ *Spatial sign function* is one example of a score function, see the definition 4.1.

$R(\mathbf{X})$ *Spatial rank function* is one example of a score function, see the definition 4.2.

$tr(\mathbf{A})$ *Trace* of a $(n \times n)$-matrix $\mathbf{A}$ is the sum of its diagonal elements.

$\mathbf{A}^T$ *Transpose* of the matrix $\mathbf{A}$.

$\mathbf{1}_n$, $\mathbf{1}$ is a *vector of ones* of length n. In case n is obvious, it might be omitted.

3 Multivariate Location

Assume that we have a p-variate random variable $\mathbf{x} \in \mathcal{M}(p)$, where $\mathcal{M}(p)$ denotes the set of p-vectors. The cumulative distribution function of $\mathbf{x}$ is $F_{\mathbf{x}}$ that is characterized by the location parameter μ and scatter parameter Ω. We have n observations $\mathbf{x}_1$, $\mathbf{x}_2$, $\ldots$, $\mathbf{x}_n$ from that distribution. These independent observations form the sample data matrix $\mathbf{X} = [\mathbf{x}_1, \ldots, \mathbf{x}_n]^T \in \mathcal{M}(n, p)$, where $\mathcal{M}(n, p)$ is the set of $(n \times p)$-matrices.

Whenever working with sets of data, the interesting questions are: Where is the data? How is it scattered around its location? In this work the focus is on the first question. To relate with it, the definitions of the location functional and its statistic are given in 3.1 and 3.2, respectively.

Definition 3.1. *Location functional:*
If a $(p \times 1)$-vector-valued functional $\mathbf{M}(F_{\mathbf{x}})$ is affine equivariant in the sense that

$$\mathbf{M}(F_{\mathbf{Ax+b}}) = \mathbf{A}M(F_x) + \mathbf{b}$$

for any non-singular $\mathbf{A} \in \mathbb{R}^{p \times p}$ and for any $\mathbf{b} \in \mathbb{R}^p$, then it is called a location functional.

Definition 3.2. *Location statistic:*
The functional $\mathbf{M} : \mathbb{R}^{n\times p} \to \mathbb{R}^p$ *is a location statistic if for any non-singular* $\mathbf{A} \in \mathbb{R}^{p\times p}$ *and for any* $\mathbf{b} \in \mathbb{R}^p$ *the functional is affine equivariant in the following sense:*

$$\mathbf{M}(\mathbf{X}\mathbf{A}^T + \mathbf{1}\mathbf{b}^T) = \mathbf{A}\mathbf{M}(\mathbf{X}) + \mathbf{b}$$

One practical consequence of them being *affine* equivariant is that for example unit conversions can be applied either to the data or to the location statistic, yielding the same results in either case. To provide examples, a common location functional is the mathematical expectation vector $\mathbf{E}(\mathbf{X})$ whereas the sample mean vector $\mathbf{M}(\mathbf{X}) := \bar{\mathbf{X}} := \mathbf{X}^\mathbf{T}\mathbf{1}/N$ is a location statistic satisfying the definition 3.2:

$$\begin{aligned}\mathbf{M}(\mathbf{X}\mathbf{A}^T + \mathbf{1}\mathbf{b}^T) &:= (\mathbf{X}\mathbf{A}^T + \mathbf{1}\mathbf{b}^T)^T\mathbf{1}/N \\ &= \mathbf{A}\mathbf{X}^\mathbf{T}\mathbf{1}/N + \mathbf{b}\mathbf{1}^T\mathbf{1}/N = \mathbf{A}\mathbf{M}(\mathbf{X}) + \mathbf{b}\end{aligned}$$

However, in practise the sample mean $\bar{\mathbf{X}}$ may be quite sensitive in presence of outlying observations. For instance, consider a group of 20 workers with monthly salaries of 2000€ (10 workers), 3000€(9 workers) and 100 000€ (1 boss). Certainly the least paid workers would not be happy with the annual report declaring that the average salary in the firm exceeds 7000€ per worker[1]. In practise, the outliers can be found and removed during the analysis but it gets more difficult as the number of dimensions grows. Going further with the given example, consider another firm of the same size having salaries distributed between 7000€-7700€, yielding the same mean than before. Are the locations of the two datasets equal? This question, though in multivariate case, is discussed in this work.

4 Several Sample Location Problem

In this section, several sample location problem is studied. After agreeing on the notations of the data, the use of the score functions is discussed. Then, the general idea of testing for the location difference is presented. In the end, this general strategy is applied to different score functions.

Let the $p-$vector $\mathbf{x}_{ij}$ denote a single independent observation that is drawn from the distribution F_i. The distribution is characterized by a location

[1]Instead, they might say that the boss is a *mean* guy

parameter μ_i and a scatter parameter Ω. These observation vectors form c independent random samples:

$$\mathbf{X_i} = \begin{bmatrix} \mathbf{x}_{i1}^T \\ \vdots \\ \mathbf{x}_{in_i}^T \end{bmatrix} \in \mathbb{R}^{n_i \times p}, i = 1, \ldots, c$$

Hence, the data matrix $\mathbf{X}$ can be constructed using the samples. The total number of observations is the sum of the sample sizes: $n = n_1 + \cdots + n_c$.

$$\mathbf{X} = \begin{bmatrix} \mathbf{X}_1 \\ \vdots \\ \mathbf{X}_c \end{bmatrix} \in \mathcal{M}(n, p)$$

Note that using $\mathbf{X}_{ij}$ means the j^{th} observation from the sample i. In other words it is the row number $n_1 + \cdots + n_i + j$, not the element in the position $\mathbf{X}(i, j)$.

The main question is whether the centers μ_i of the p-variate samples $1, \ldots, c$ are equal or not. This can be formalized as a null-hypothesis:

$$H_0 : \mu_1 = \cdots = \mu_c \tag{1}$$

As we assumed that for all the cumulative distribution functions F_i the scatter parameter is Ω, the equivalent form of the null-hypothesis would be:

$$H_0 : F_1 = \cdots = F_c \tag{2}$$

This assumption contradicts with the general problem lying behind this work and will be discussed more thoroughly later on.

4.1 Score functions

In general, score functions $\mathbf{T} : \mathbb{R}^p \to \mathbb{R}^p$ are used to construct tests and estimates for the location parameter μ. Three different score functions are discussed. The first one is the identity score $\mathbf{T}(\mathbf{x}) = \mathbf{x}$, which will eventually lead to classical multivariate analysis of variance (MANOVA). After that, the multivariate concepts of spatial signs and spatial ranks are applied. Their functional definitions are given in 4.1 and 4.2, respectively.

Definition 4.1. *Spatial sign function:*
The spatial sign function $\mathbf{U}(\mathbf{x}) : \mathbb{R}^p \to \mathbb{R}^p$ *is a direction vector lying on the unit p-sphere, namely*

$$\mathbf{U}(\mathbf{x}) = \begin{cases} |\mathbf{x}|^{-1}\mathbf{x}, & \mathbf{x} \neq \mathbf{0} \\ \mathbf{0}, & \mathbf{x} = \mathbf{0}, \end{cases}$$

where $|<>|$ *stands for the usual* L_2*-norm (euclidean). For a* $(n \times p)$*-matrix* $\mathbf{X} = \begin{bmatrix} \mathbf{x}_1 & \cdots & \mathbf{x}_n \end{bmatrix}^T$*, where* $\mathbf{x}_i \in \mathbb{R}^{p\times 1}$*, the function* $\mathbf{U}(\mathbf{X}) : \mathbb{R}^{n\times p} \to \mathbb{R}^{n\times p}$ *is defined as follows:*

$$\mathbf{U}(\mathbf{X}) = \begin{bmatrix} \mathbf{U}(\mathbf{x}_1)^T \\ \vdots \\ \mathbf{U}(\mathbf{x}_n)^T \end{bmatrix}$$

Definition 4.2. *Spatial rank function:*
The spatial rank function $\mathbf{R}(\mathbf{X}) : \mathbb{R}^{n\times p} \to \mathbb{R}^{n\times p}$ *is a relative measure of direction and distance from the center of the data, defined as*

$$\mathbf{R}(\mathbf{X}) = \frac{1}{n}\sum_{i=1}^{n} \mathbf{U}(\mathbf{X} - \begin{bmatrix} \mathbf{x}_i^T \\ \vdots \\ \mathbf{x}_i^T \end{bmatrix})$$

The spatial ranks lie in the unit p-ball.

As the spatial rank is a relative measure, it is always $\mathbf{0}$ for only one observation because the one always is relatively at the center. Therefore the spatial rank function is defined for the data matrix. To clarify these concepts, see the figure 1 representing two tailed data sets and corresponding spatial sign and rank functions.

The scores are often centered and/or standardized to perform different tests, or to attain different estimates. One can use either *outer* or *inner* procedure. Note that $\mathbf{T} := [\mathbf{T}_1, \ldots, \mathbf{T}_n]^T := [\mathbf{T}(\mathbf{x_1}), \ldots, \mathbf{T}(\mathbf{x}_n)]^T$ and that average is denoted as "bar": $\bar{\mathbf{T}} = \frac{1}{n}\sum_{i=1}^{n} \mathbf{T}_i$.

1. Outer procedure
   (a) Outer centering of the scores:
   $$\mathbf{T}_i \to \hat{\mathbf{T}}_i = \mathbf{T}_i - \bar{\mathbf{T}}$$
   (b) Outer standardization of the scores:
   $$\mathbf{T}_i \to \hat{\mathbf{T}}_i = \mathbf{COV}(\mathbf{T})^{-1/2}\mathbf{T}_i$$

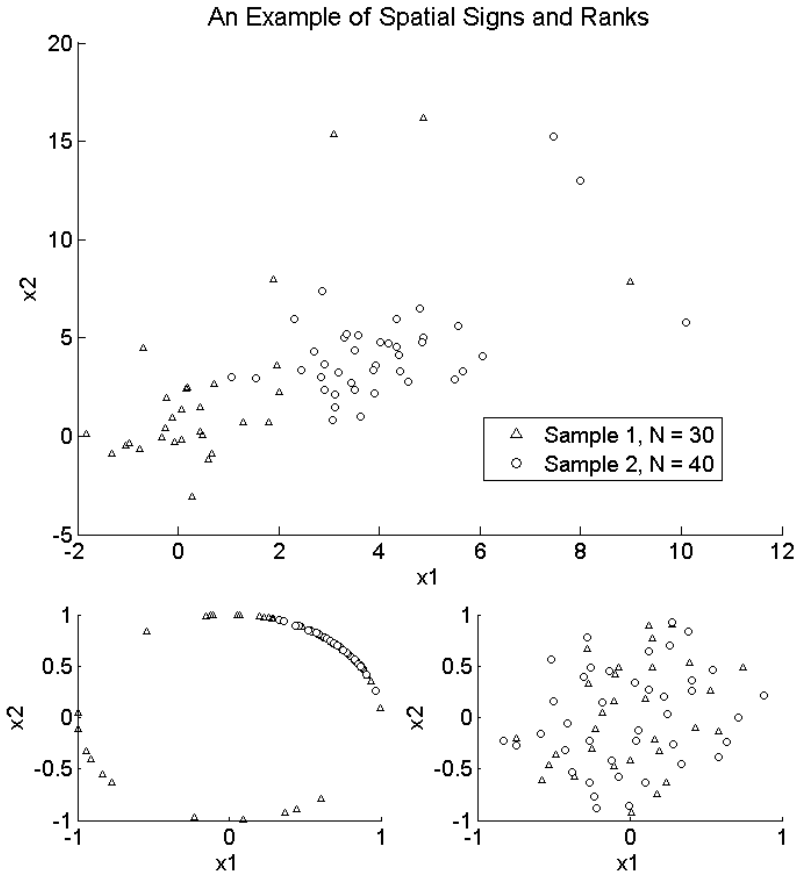

Figure 1: Two bivariate random data samples of 30 and 40 observations. The theoretical locations of the samples are different and both samples include outliers. The corresponding spatial signs and ranks calculated in the smaller figures. The spatial signs show the direction of the observation on the p-sphere whereas the spatial ranks is a relative measure so that all the observations are inside the p-sphere. In this case $p = 2$ and hence the p-sphere is a circle.

(c) The combination of the two:

$$\mathbf{T}_i \to \hat{\mathbf{T}}_i = \mathbf{COV}(\mathbf{T})^{-1/2}(\mathbf{T}_i - \bar{\mathbf{T}})$$

2. Inner procedure

(a) Inner centering of the scores:

$$\mathbf{T}_i \to \hat{\mathbf{T}}_i = \mathbf{T}(\mathbf{y_i} - \mathbf{M})$$

that must yield the $\mathbf{0} \in \mathbb{R}^p$ as an average:

$$\frac{1}{n}\sum_{i=1}^{n} \hat{\mathbf{T}}_i = \mathbf{0}$$

(b) Inner standardization of the scores:

$$\mathbf{T}_i \to \hat{\mathbf{T}}_i = \mathbf{T}(\mathbf{S}^{-1/2}\mathbf{y_i})$$

that must satisfy

$$p \cdot \sum_{i=1}^{n} \hat{\mathbf{T}}_i \hat{\mathbf{T}}_i^T = \mathbf{I_p} \sum_{i=1}^{n} \hat{\mathbf{T}}_i^T \hat{\mathbf{T}}_i$$

(c) Inner centering and standardization combines the two and both of the requirements must be satisfied:

$$\mathbf{T}_i \to \hat{\mathbf{T}}_i = \mathbf{T}(\mathbf{S}^{-1/2}(\mathbf{y_i} - \mathbf{M}))$$

In the case of the spatial sign, an example of the $\mathbf{M}$ to fulfill the condition of the inner centering is the *spatial median*, defined in 4.3.

Definition 4.3. *Spatial median:*
The spatial median $\mu \in \mathbb{R}^p$ of the data matrix $\mathbf{X} \in \mathbb{R}^{n\times p}$ minimizes the L_1-difference from it, namely:

$$\mu = arg \min_{\mu\in\mathbb{R}^p} \sum_{i=1}^{n} |\mathbf{X}_i - \mu^T|$$

One way to obtain the spatial median is the Weiszfeld-method whose iteration step is given in the equation 3. (Kärkkäinen and Äyrämö [2005])

$$\mathbf{u}^{k+1} = \begin{cases} \frac{\sum_{i=1}^{n} \mathbf{x}_i/|\mathbf{u}^k - \mathbf{x}_i|}{\sum_{i=1}^{n} 1/|\mathbf{u}^k - \mathbf{x}_i|}, & \mathbf{u}^k \notin \{\mathbf{x}_1, \ldots, \mathbf{x}_n\} \\ \mathbf{u}^k \end{cases} \tag{3}$$

As the given method fails to converge at times, an other method is used. Gervini [2008] provides a method to evaluate spatial median for the functional data analysis purposes. That method can be applied to this problem by setting the dimensions of a multivariate data set as a time axis with constant discretization. Smoothing along the time-axis is not allowed. The result is then spatial median of all the dimensions. The author of the article provides the Matlab-functions[2] and they are utilized in this work.

4.2 General strategy

In this section the test statistic is derived using inner centering and outer standardization. The first step is to find a shift vector $\mathbf{M}$ such that

$$\frac{1}{c}\sum_{i=1}^{c}\frac{1}{n_i}\sum_{j=1}^{n_i}\mathbf{T}(\mathbf{y}_{ij} - \mathbf{M}) = \mathbf{0} \in \mathbb{R}^p$$

Then the centered score function are defined as

$$\hat{\mathbf{T}}_{ij} = \mathbf{T}(\mathbf{x}_{ij} - \mathbf{M}),\ i = 1, \ldots, c;\ j = 1, \ldots, n_i.$$

Next, the group means are calculated:

$$\hat{\mathbf{T}}_i = \frac{1}{n_i}\sum_{j=1}^{n_i}\hat{\mathbf{T}}_{ij},\ i = 1, \ldots, c.$$

Assume that the general score function $\hat{\mathbf{T}}_{ij}$ is centered and, by the Central Limit theorem that

$$\sqrt{n_i}\frac{1}{n_i}\sum_{j=1}^{n_i}\hat{\mathbf{T}}_{ij} = \sqrt{n_i}\hat{\mathbf{T}}_i \xrightarrow{d} \mathbf{N}_p(\mathbf{0}, \mathbf{B})$$

[2] https://pantherfile.uwm.edu/gervini/www/RFDA.html , accessed 11.5.2015

where $(p \times p)$-covariance matrix $\mathbf{B}$ is defined as $\mathbf{B} := \mathbf{E}(\hat{\mathbf{T}}_{ij}\hat{\mathbf{T}}_{ij}^T)$. Then, $\sqrt{n_i}\mathbf{B}^{-1/2}\hat{\mathbf{T}}_i \xrightarrow{d} \mathbf{N}_p(\mathbf{0}, \mathbf{I})$. This is proven in the general case below:

Statement: For any $\mathbf{B}^{1/2} := \mathbf{A}$ that accomplishes $\mathbf{A}\mathbf{A}^T = \mathbf{B}$, the following holds: $\mathbf{y} \sim \mathbf{N}_p(\mathbf{0}, \mathbf{B}) \Rightarrow \mathbf{B}^{-1/2}\mathbf{y} \sim \mathbf{N}_p(\mathbf{0}, \mathbf{I})$.

Proof: Firstly, any linear combination of multivariate normal random variable is normally distributed. Secondly, the expectation parameter of this new normally distributed variable is:

$$\mathbf{E}(\mathbf{B}^{-1/2}\mathbf{y}) = \mathbf{B}^{-1/2}\mathbf{E}(\mathbf{y}) = \mathbf{B}^{-1/2}\mathbf{0} = \mathbf{0}$$

Thirdly, the covariance matrix of the new variable is, recalling that covariance is an affine equivariant operator:

$$\mathbf{COV}(\mathbf{B}^{-1/2}\mathbf{y}) = \mathbf{B}^{-1/2}\mathbf{COV}(\mathbf{y})(\mathbf{B}^{-1/2})^T = \mathbf{B}^{-1/2}\mathbf{B}(\mathbf{B}^{-1/2})^T$$
$$= \mathbf{B}^{-1/2}\mathbf{B}^{1/2}(\mathbf{B}^{1/2})^T(\mathbf{B}^{-1/2})^T = \mathbf{B}^{-1/2}\mathbf{B}^{1/2}(\mathbf{B}^{-1/2}\mathbf{B}^{1/2})^T = \mathbf{I} \quad \square$$

Next, we define an estimator for $\mathbf{B}$, that is $\hat{\mathbf{B}} := \frac{1}{n}\sum_{i=1}^{c}\sum_{j=1}^{n_i}\hat{\mathbf{T}}_{ij}\hat{\mathbf{T}}_{ij}^T \in \mathbb{R}^{p\times p}$ which *converges in probability* to $\mathbf{B}$. Then, basing on the earlier proof, $\hat{\mathbf{C}}_i := \sqrt{n_i}\hat{\mathbf{B}}^{-1/2}\hat{\mathbf{T}}_i \xrightarrow{d} \mathbf{N}_p(\mathbf{0}, \mathbf{I})$ as well.

If $\hat{\mathbf{C}}_i \xrightarrow{d} \mathbf{N}_p(\mathbf{0}, \mathbf{I})$, then, by the definition of Chi-squared random variable, $\hat{\mathbf{C}}_i^T\hat{\mathbf{C}}_i \xrightarrow{d} \chi^2(p)$.

Now, our test statistic to test whether the location of the samples differ or not, can be formed as a sum of Chi-squared random variables, as presented in the equation 4. In general, the sum of independent Chi-squared random variables follows Chi-squared distribution with degrees of freedom obtained by summing the degrees of freedom of all the independent variables[3]. However p degrees of freedom are lost as the covariance matrix of group means

[3] Proof: Recall that the moment generating function of a random variable X, defined as $M_X(t) = E[e^{tX}]$, if it exists, uniquely determines its distribution. For $X \sim \chi^2(\nu)$

$$M_X(t) = (1-2t)^{-\nu/2}, \; t > \frac{1}{2}$$

Given random variables $X_1, \ldots, X_n$ such that $X_i \sim \chi^2(\nu_i)\ \forall i = 1, \ldots, n$. The random variable Y is then defined as the sum of the previous ones, namely $Y = X_1 + \cdots + X_n$. Its moment generating function is, by the definition and by applying simple calculus:

$$M_Y(t) = E[e^{tY}] = E[e^{t(X_1+\cdots+X_n)}] = E[e^{tX_1} \times \cdots \times e^{tX_n}]$$
$$= E[e^{tX_1}] \times \cdots \times E[e^{tX_n}] = \prod_{i=1}^{n} M_{X_i}(t)$$

$\hat{\mathbf{T}}_i = \sum_{j=1}^{n_i} \hat{\mathbf{T}}_{ij}$ of the centered scores would not be invertible. An intuitive explanation is that the same data cannot be used to estimate both the covariance and its expectation parameter. Hence, only $c-1$ groups can contribute to the degrees of freedom (for details, see e.g. [Oja, 2010, p. 151]). The test statistic gets the following form:

$$Q^2(\mathbf{X}) = \hat{\mathbf{C}}_1^T \hat{\mathbf{C}}_1 + \cdots + \hat{\mathbf{C}}_c^T \hat{\mathbf{C}}_c \xrightarrow[H_0]{d} \chi^2((c-1)p) \tag{4}$$

As the null hypothesis (see the equation 1) assumed that all the centers would be the same, then large values of the test statistic give evidence against it. Using the test statistic derived above, *p-value* can be attained by:

$$\begin{aligned} p = Pr(\chi^2_{(c-1)p} > Q^2(\mathbf{X})) &= 1 - Pr(\chi^2_{(c-1)p} \leq Q^2(\mathbf{X})) \\ &= 1 - F_{\chi^2_{(c-1)p}}(Q^2(\mathbf{X})) \end{aligned}$$

In the end, the test statistic of the equation 4 is derived into a clearer form:

$$\begin{aligned} \hat{\mathbf{C}}_i^T \hat{\mathbf{C}}_i &= (\sqrt{n_i}\hat{\mathbf{B}}^{-1/2}\hat{\mathbf{T}}_i)^T(\sqrt{n_i}\hat{\mathbf{B}}^{-1/2}\hat{\mathbf{T}}_i) \\ &= n_i \hat{\mathbf{T}}_i^T(\hat{\mathbf{B}}^{-1/2})^T\hat{\mathbf{B}}^{-1/2}\hat{\mathbf{T}}_i = \hat{\mathbf{T}}_i^T\hat{\mathbf{B}}^{-1}\hat{\mathbf{T}}_i \end{aligned}$$

Eventually the test statistic is (5):

$$Q^2(\mathbf{X}) = \sum_{i=1}^{c} n_i \hat{\mathbf{T}}_i^t \hat{\mathbf{B}}^{-1} \hat{\mathbf{T}}_i \xrightarrow[H_0]{d} \chi^2((c-1)p) \tag{5}$$

4.2.1 Multivariate analysis of variance (MANOVA)

Multivariate analysis of variance (MANOVA) generalizes the univariate analysis of variance (ANOVA) which tests the null-hypothesis of the equal means of different groups. MANOVA is achieved using the general strategy presented earlier by choosing the identity score function $\mathbf{T}(\mathbf{x}_{ij}) = \mathbf{x}_{ij}$. Then, the sample mean is denoted by

---

As the moment generating functions of X_i are known, the derivation continues as:
$M_Y(t) = (1-2t)^{-\nu_1/2} \times \cdots \times (1-2t)^{-\nu_n/2} = (1-2t)^{-(\nu_1+\cdots+\nu_n)/2}$
Hence, $Y \sim \chi^2(\nu_1 + \cdots + \nu_n)$ □

$$\bar{\mathbf{x}}_i := \frac{1}{n_i}\sum_{j=1}^{n_i}\mathbf{x}_{ij},\ i = 1, \ldots, c$$

and for the total mean

$$\bar{\mathbf{x}} := \frac{1}{c}\sum_{i=1}^{c}\bar{\mathbf{x}}_i.$$

The first step of the general strategy is the centering. Inner and outer centered scores are attained by choosing the shift vector $\mathbf{M}$ to be the total mean vector $\bar{\mathbf{x}}$:

$$\hat{\mathbf{T}}_{ij} := \hat{\mathbf{T}}(\mathbf{x}_{ij}) = \mathbf{x}_{ij} - \bar{\mathbf{x}},\ i = 1, \ldots, c;\ j = 1, \ldots, n_i.$$

The test statistic is then based on

$$\hat{\mathbf{T}}_i := \frac{1}{n_i}\sum_{j=1}^{n_i}\hat{\mathbf{T}}_{ij} = \frac{1}{n_i}\sum_{j=1}^{n_i}(\mathbf{x}_{ij} - \bar{\mathbf{x}})$$

$$= \bar{\mathbf{x}}_i - \bar{\mathbf{x}},\ i = 1, \ldots, c.$$

Hence, the test statistic is given in the equation 6. Under the null hypothesis H_0 the limiting distribution of $Q^2(\mathbf{X})$ is, as derived earlier, $\chi^2((c-1)p)$. Note that the scores are bot inner and outer centered and that the sample covariance matrix $\hat{\mathbf{B}} := \frac{1}{n}\sum_{i=1}^{c}\sum_{j=1}^{n_i}(\mathbf{x}_{ij} - \bar{\mathbf{x}})(\mathbf{x}_{ij} - \bar{\mathbf{x}})^T$ fulfills the conditions of both outer and inner standardization. Therefore the test statistic is affine invariant, say $Q^2(\mathbf{X}\mathbf{H}^T + \mathbf{1}_n\mathbf{b}^T) = Q^2(\mathbf{X})$, where $\mathbf{H}$ is full-rank $(p \times p)$ matrix and $\mathbf{b} \in \mathbb{R}^p$.

$$Q^2(\mathbf{X}) := \sum_{i=1}^{c}(n_i\hat{\mathbf{T}}_i^t\hat{\mathbf{B}}^{-1}\hat{\mathbf{T}}_i) = \sum_{i=1}^{c}(n_i(\bar{\mathbf{x}}_i - \bar{\mathbf{x}})^t\hat{\mathbf{B}}^{-1}(\bar{\mathbf{x}}_i - \bar{\mathbf{x}})) \tag{6}$$

In practise it might be easier to relate a single observation to a group and store that information. Based on this idea, another useful formulation of MANOVA [Oja, 2010, pp. 148-149] utilizes a sample membership matrix $\mathbf{M} \in \mathbb{R}^{n \times c}$ which is defined in the equation 7. Often both the data and the sample membership matrices are available. This method is based on the

decomposition of sums of squares $SS_T = SS_B + SS_W$ meaning that the total error is the sum of the errors between the groups and the errors within the groups. Many test statistics have been derived for this purpose, and two of them are $Q_P^2 = n \cdot tr(SS_B SS_T^2)$ (*Pillai*) and $Q_{LH}^2 = n \cdot tr(SS_B SS_W^{-1})$ (*Lawley-Hotelling*).

$$\mathbf{M}_{ij} = \begin{cases} 1, \text{ the i}^{\text{th}} \text{ observation belongs to the j}^{\text{th}} \text{ sample} \\ 0 \end{cases} \tag{7}$$

The diagonal element of $\mathbf{M}^T\mathbf{M} \in \mathbb{R}^{c \times c}$ determine the sample sizes. The matrix $\mathbf{M}\mathbf{M}^T \in \mathbb{R}^{n \times n}$ might be useful as $(\mathbf{M}\mathbf{M}^T)_{ij}$ equals one if and only if the observations i and j come from the same sample, otherwise it is zero. The projection[4] $\mathbf{P_M X} := \mathbf{M}_n(\mathbf{M}_n^T\mathbf{M}_n)^{-1}\mathbf{M}_n^T\mathbf{X}$ projects the data points to the subspace spanned by the columns of $\mathbf{M}$. As the columns of $\mathbf{M}$ determine the group of the corresponding row (observation), then that projection replaces the observations by the sample means – without caring of the order of the observations. Then, the projection $(\mathbf{I} - \mathbf{P_M})\mathbf{X}$ yields the within-group differences, namely the observation minus the corresponding sample mean.

In the end, the data matrix is centered with a smart move. The projection $\mathbf{1}_n(\mathbf{1}_n^T\mathbf{1}_n)^{-1}\mathbf{1}_n^T\mathbf{X} = \frac{1}{n}\mathbf{1}_n\mathbf{1}_n^T\mathbf{X}$ replaces all the observations $\mathbf{X}$ by the total mean vector. Then, the centered scores can be calculated as $\bar{\mathbf{T}} = (\mathbf{I} - \frac{1}{n}\mathbf{1}_n\mathbf{1}_n^T)\mathbf{X}$, namely the observations minus the total mean. They satisfy the conditions of both outer and inner procedure.

Then, all the projections are applied to these centered scores and we get the decomposition $SS_T = SS_B + SS_W$ into the form:

$$\bar{\mathbf{T}}^T\bar{\mathbf{T}} = \bar{\mathbf{T}}^T\mathbf{P_M}\bar{\mathbf{T}} + \bar{\mathbf{T}}^T(\mathbf{I} - \mathbf{P_M})\bar{\mathbf{T}}$$

Then, the *Pillai Lawley-Hotelling* test statistics can be stated using these notations and they are presented in the equations 8 and 9. Their distributions can often be approximated by $\chi^2((c-1)p)$ but the exact distribution is not known. These test statistics are left out from the section 5 but are presented

[4]If a matrix $\mathbf{A} \in \mathbb{R}^{r \times r}$ satisfies $\mathbf{A}^2 = \mathbf{A}$ (idempotent), then it is called a *projection matrix*. In other words, if a point is already projected, then another projection does not have any effect on it. Moreover, if $\mathbf{A}$ is a projection, then is $\mathbf{I} - \mathbf{A} = (\mathbf{I} - \mathbf{A})^2 = \mathbf{I} - 2\mathbf{A} + \mathbf{A}^2 = \mathbf{I} - 2\mathbf{A} + \mathbf{A} = \mathbf{I} - \mathbf{A}$.

here as a curiosity.

$$Q_P^2 = n \cdot tr(\bar{\mathbf{T}}^T \mathbf{P_M} \bar{\mathbf{T}} (\bar{\mathbf{T}}^T \bar{\mathbf{T}})^{-1}) \tag{8}$$

$$Q_{LH}^2 = n \cdot tr(\bar{\mathbf{T}}^T \mathbf{P_M} \bar{\mathbf{T}} (\bar{\mathbf{T}}^T (\mathbf{I} - \mathbf{P_M}) \bar{\mathbf{T}})^{-1}) \tag{9}$$

4.2.2 The Test Based on Spatial Signs

If the spatial sign function is used as the score function, the inner centered scores are obtained using the spatial median vector $\hat{\mu}$ as the shift vector $\mathbf{M}$:

$$\hat{\mathbf{U}}_{ij} := \hat{\mathbf{U}}(\mathbf{x}_{ij}) = \mathbf{x}_{ij} - \hat{\mu},\ i = 1, \ldots, c;\ j = 1, \ldots, n_i.$$

The test statistic for testing for the difference between the groups is then based on:

$$\hat{\mathbf{U}}_i := \frac{1}{n_i} \sum_{i=1}^{n_i} \hat{\mathbf{U}}_{ij},\ i = 1, \ldots, c$$

Hence, the test statistic is given in the equation 10 and its limiting distribution under the null hypothesis (and under some weak assumptions [Oja, 2010, p. 154]) is again $\chi^2((c-1)p)$. The test statistic is location and scalar invariant: $Q^2(\alpha \mathbf{X} + \mathbf{1}_n \mathbf{b}^T) = Q^2(\mathbf{X})$, where $\alpha \in \mathbb{R}$ and $\mathbf{b} \in \mathbb{R}^p$.

$$Q^2(\mathbf{X}) = \sum_{i=1}^{c} (n_i \hat{\mathbf{U}}_i^t \hat{\mathbf{B}}^{-1} \hat{\mathbf{U}}_i) \tag{10}$$

4.2.3 The Test based on Spatial Ranks

The spatial rank scores

$$\mathbf{R}_{ij} := \mathbf{R}_{\mathbf{X}}(\mathbf{x}_{ij}),\ i = 1, \ldots, c;\ j = 1, \ldots, n_i.$$

are already centered. Then the test test statistic for testing the difference between the groups is based on:

$$\bar{\mathbf{R}}_i := \frac{1}{n_i} \sum_{i=1}^{n_i} \mathbf{R}_{ij},\ i = 1, \ldots, c$$

The test statistic is provided in the equation 11 and its limiting distribution under the null hypothesis (and under some weak assumptions [Oja, 2010, p. 158]) is again $\chi^2((c-1)p)$. The test statistic is location and scalar invariant: $Q^2(\alpha\mathbf{X} + \mathbf{1}_n\mathbf{b}^T) = Q^2(\mathbf{X})$, where $\alpha \in \mathbb{R}$ and $\mathbf{b} \in \mathbb{R}^p$.

$$Q^2(\mathbf{X}) = \sum_{i=1}^{c} (n_i \bar{\mathbf{R}}_i^t \hat{\mathbf{B}}^{-1} \bar{\mathbf{R}}_i) \tag{11}$$

5 Results

The aim of this section is to present and compare different properties of the three test statistics, denoted by A, B and C. The test A is based on the identity score function (MANOVA), B on the spatial signs and C on the spatial ranks. Their formal definitions are given in the equations 6, 10 and 11, respectively.

5.1 Simulations with multivariate normally distributed data with and without anomalies

The idea of an experiment (or a case) is to compare the three different tests A, B and C. For it, we have c random data samples of sizes $n_i,\ i = 1, \ldots, c$. The simulated data is drawn from p-variate normal distribution whose location parameters are μ_i and scatter parameters Σ_i. In most cases the scatter parameter Σ is common for each group. Then, the location parameters are multiplied by a scale factor $b \in [0, 1]$. The smaller the value of b, the closer to each other the theoretical parameters are. The probability of rejecting the H_0 should increase as the value of b increases. To diminish the random effects, the simulation is repeated 1000 times for each b. In all the experiments, b gets 100 distinct values. Hence in each experiment there are $100 \cdot 1000$ random samples evaluated.

The case 1 consists of three samples drawn from bivariate normal distributions with different location parameters. The complete setting is presented in the table 1 and the results, along with those of the case 2, in the figure

i	Sample size, n_i	Location, μ_i	Scatter, Σ
1	50	$\begin{bmatrix}-4 & -4\end{bmatrix}$	$2\mathbf{I}_2$
2	60	$\begin{bmatrix}0 & 0\end{bmatrix}$	$2\mathbf{I}_2$
3	70	$\begin{bmatrix}1 & 1\end{bmatrix}$	$2\mathbf{I}_2$

Table 1: Setting of the case 1: Three samples from bivariate normal distributions whose parameters are given in the table.

i	Sample size, n_i	Location, μ_i	Scatter, Σ
1	80	$\begin{bmatrix}0 & 0\end{bmatrix}$	$2\mathbf{I}_2$
2	20	$\begin{bmatrix}-4 & -4\end{bmatrix}$	$2\mathbf{I}_2$

Table 2: Setting of the case 2: Two samples from bivariate normal distributions whose parameters are given in the table. Moreover, some outliers are added so that the tail is into the direction of the negative quarter of $\mathbb{R}^{2\times 2}$.

3. All the tests A, B and C seem quite similar even though the test B can be distinguished. The setting of the case 2 (see the table 2) is similar to that of the case 1 but this time the group sizes are different and about 20 % of the observations are *outliers*. The outliers do share the same location parameter with the other observations but the scatter parameter is five times larger. Moreover, these outliers are transformed into the negative quarter in order to make the data asymmetrically tailed. One example of the data is given in the figure 2 whereas the general results are given in the figure 3. Clearly, the tests A and C do not reject that many null hypotheses after the contamination of the data.

In the case 3, assume that the location shifts of the groups are into the direction of the minor principal axis of the covariance matrix of the bivariate normal distribution. The setting is given in the table 3 and the general results can be seen in the figure 4. The tests A and C are again quite similar whereas the test B does not indicate the difference that frequently and hence can be distinguished in the figure. One example of that is shown in the figure 5 where the p-value of the test B is remarkably larger than those of A and C.

In the case 4 the tests are applied to four dimensional data and the purpose is to investigate the effect of unequal group-scatter parameters. The complete setting of the case 4 is given in the table 4. Note that the only difference between the case 4a and 4b is the scatter parameter of the largest group, whose theoretical location parameter is almost in the middle of the smaller groups. The results are shown in the figure 6. The tests A and C do change

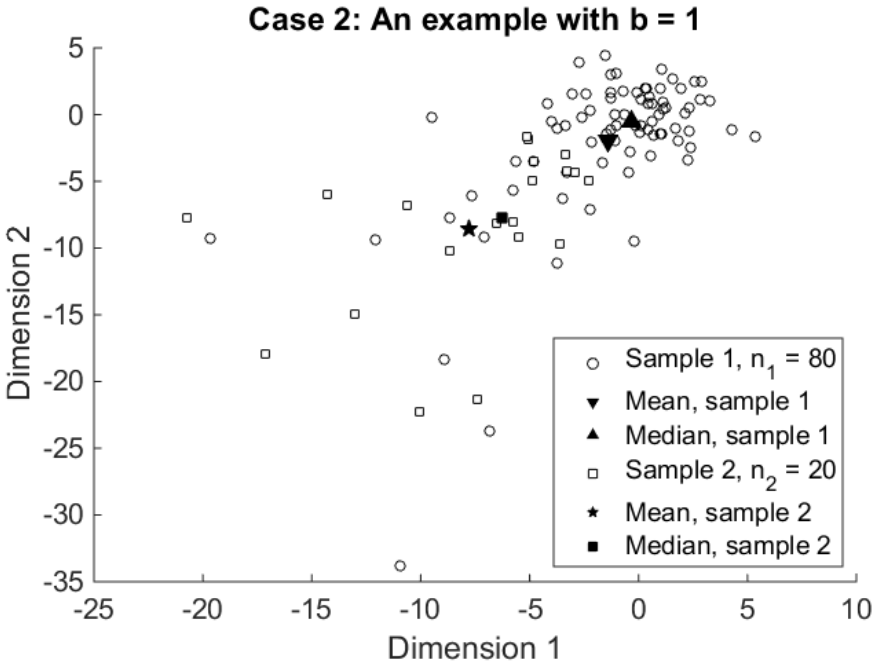

Figure 2: Case 2: One example of the data with the outliers in the negative quarter. In this case the scale factor $b = 1$ and therefore the locations of the groups differ significantly. According to all the tests, there is a lot of evidence to reject the null hypothesis H_0 as all the p-values were less than 10^{-4}.

i	Sample size, n_i	Location, μ_i	Scatter, Σ
1	20	$\begin{bmatrix} 0 & 0 \end{bmatrix}$	$\begin{bmatrix} 17.3954 & 3.4721 \\ 3.4721 & 3.2966 \end{bmatrix}$
2	100	$\begin{bmatrix} 0.8465 & -3.6346 \end{bmatrix}$	
3	20	$\begin{bmatrix} 1.6931 & -7.2692 \end{bmatrix}$	

Table 3: Setting of the case 3: Three samples from bivariate normal distributions whose parameters are given in the table. Note that the covariance matrices are equal and the location shifts are into the direction of the smaller eigen-vector of it.

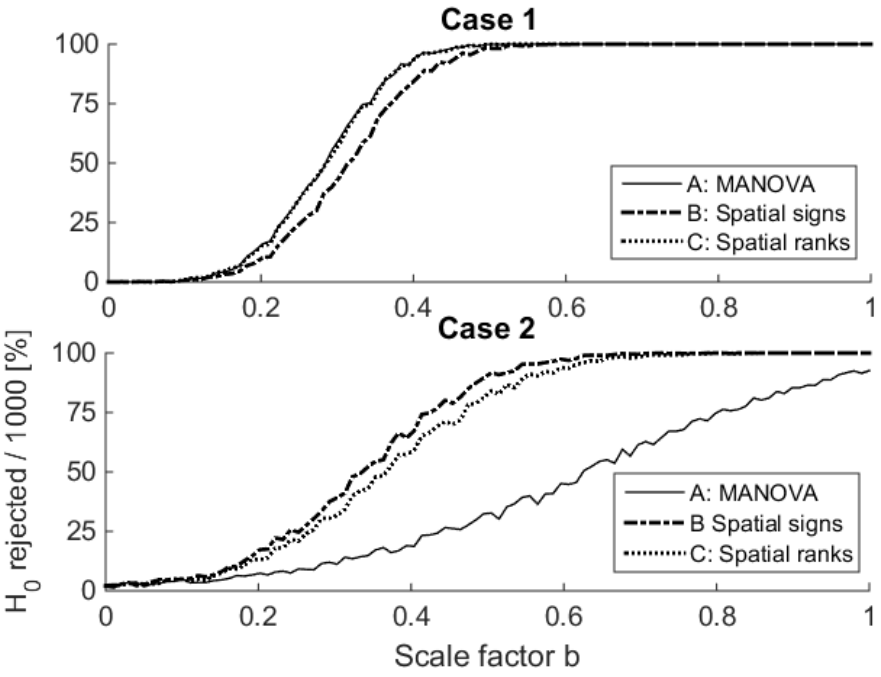

Figure 3: Cases 1 and 2: The percentages of rejecting the null hypothesis H_0 in the cases 1 (upper) and 2 (lower). The threshold value of the rejection was set to $p = 0.05$.

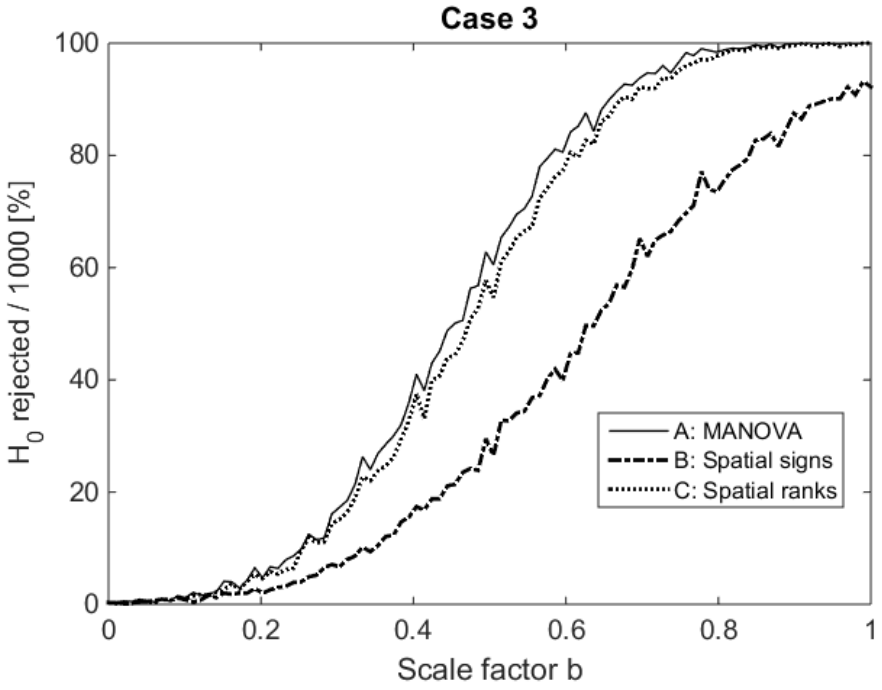

Figure 4: Case 3: The percentages of rejecting the null hypothesis H_0 in 1000 simulated cases for 100 values of the scale factor b.The threshold value of the rejection was set to $p = 0.05$. The test based on spatial signs (B) differs from the other two remarkably.

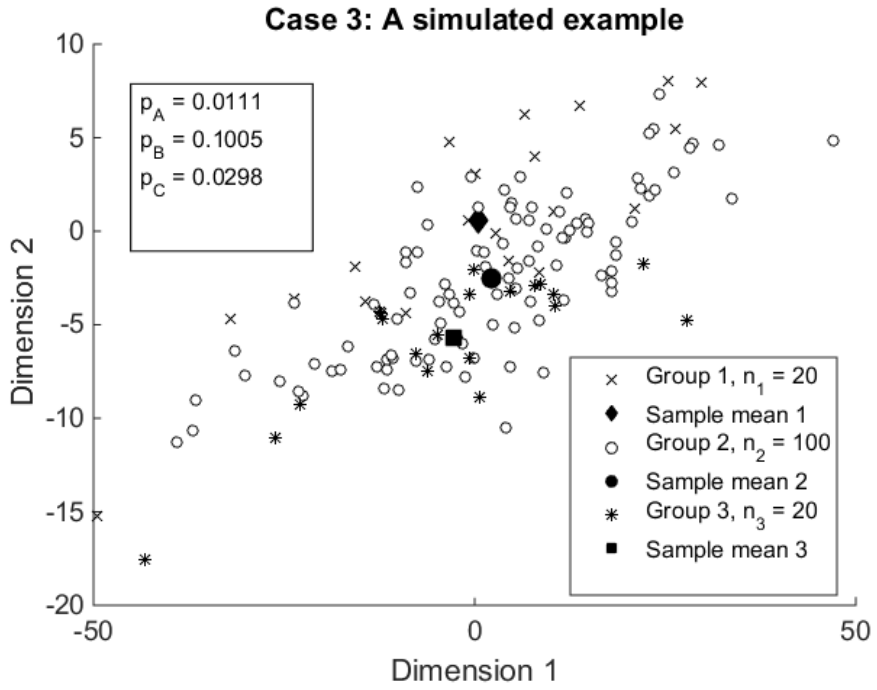

Figure 5: Case 3: An example of simulated data where the spatial sign test does not indicate location shift. The subscripts of the p-values: A := MANOVA, B := Spatial signs and C := Spatial ranks.

Sample, i	Sample size, n_i	Location, μ_i	Scatter, Σ_a	Scatter, Σ_b
1	20	[0,0,0,0]	$2\mathbf{I}_4$	$2\mathbf{I}_4$
2	100	[1,1,2,2]	$2\mathbf{I}_4$	$4\mathbf{I}_4$
3	20	[2,2,4,5]	$2\mathbf{I}_4$	$2\mathbf{I}_4$

Table 4: Setting of the case 4 (variations a and b): Samples from 4-variate normal distribution whose parameters are given in the table.

significantly in the case b whereas the test B remains practically the same in both cases.

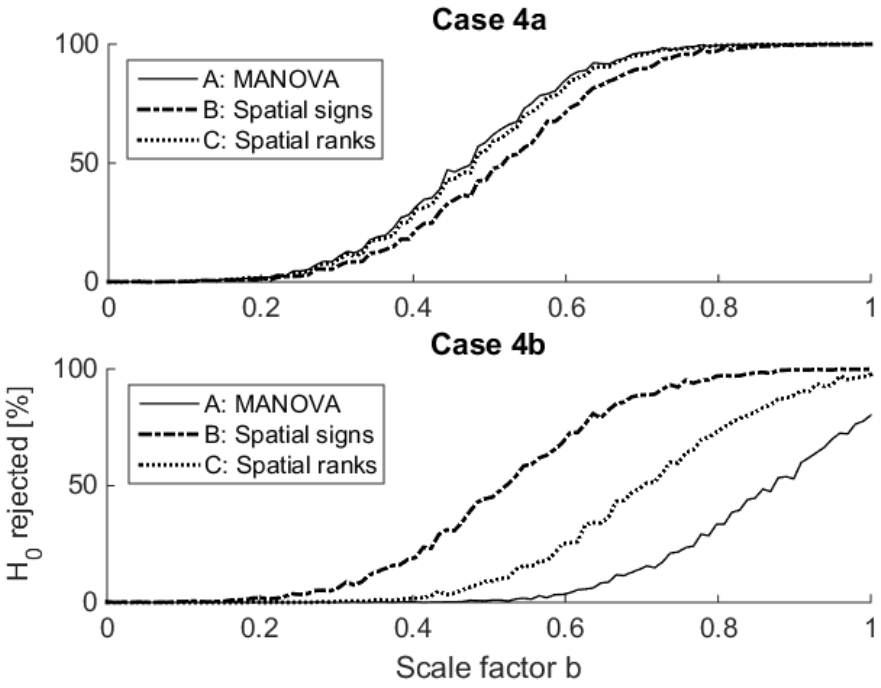

Figure 6: Cases 4a and 4b: The percentages of rejecting the null hypothesis H_0 in 1000 simulated cases for 100 values of the scale factor b.The threshold value of the rejection was set to $p = 0.05$ in both cases.

5.2 Real Data Application

The practical frame of reference of the work is operational quality of a hydropower cascade. The term "operational" means the real and actual actions related to the production. The concept of operational quality is defined here vaguely as "the effect of all the operational decisions on all the aspects it is related to". The principle is that if an operational decision may have an impact on something, then that cannot be left out from the quality considerations. The main aspects of the quality include recreational values, the nature (the erosion of the river benches, the security of the dams, water quality), wearing of the technical components, operational risks and optimality of the production. These aspects are very different by nature. It is, to some extent, possible to measure the optimality at real time whereas the recreational values form a qualitative question of customs and is certainly a matter of opinion. Still, these quantitative values form a remarkable component of the quality and the operational decision making as the river is partly public property and is tightly bound with the lives of the people nearby.

In general, the data of any process is assumed to be in the form of time-series. However, the tests presented in this work do assume scattered data. Hence, relevant time-series are selected and preprocessed to form multivariate samples.

The selection of the data relates to the external factors. For example production conditions like seasons, rains and temperature should be nearly identical in the different samples. However, in practise, the distinction between different production conditions is not that straightforward. For example, the rains in the upper reaches of the river reach the main channel the next day or later.

The preprocessing relates to the concept of the quality. Consider, for example, that one quality objective would be "keeping the levels of the pools high". Then, one possible measure of the quality could be the difference between the minimum level of a day minus the upper restriction. In this case, the preprocessing would be: find the minimum level of a relevant period and calculate the difference to the pool's upper restriction. But, how is the minimum defined? The level that is instantaneously visited, or maybe such minimum level under which the level has been at least half an hour?

How to apply the methods? There are two main ways to conduct the experiments and both of them assumes relevant quality measures. In the first one, some periods of time are chosen to pose as a reference of satisfactory quality

according to the *a priori*-knowledge concerning the process. Then, current process output samples are compared with the corresponding reference sample. The difference in the location of the samples indicate a probable change in process quality. This comparison is conducted using the test statistics derived in this work. The second way is based on similar periods of time possessing some interesting differences like changes of technical components, changes of some core computer systems or changes in operators. Then, these samples are compared using the test statistics. If the tests indicate change in the quality, the samples are taken into thorough consideration. So, the difference is that the first one applies the general knowledge before the tests whereas the latter needs the deep understanding of the process only if some quality deviations have been indicated.

The data of the case Real I is summer time data from the years 2012 – 2014. The reference data is from the years 2012 and 2013 whereas the data under the investigation consists of random days of the summer 2014. The quality measure is the one presented earlier, namely the maximum decline in the level of a pool during a day. The explaining variable is the benefits[5] gained which is measured by the day-average spilling of the water. The data and the results of the three tests are shown in the figure 7. The p-values $p_A = 0.0090$ and $p_C = 0.0076$ are almost equal and less than a fifth of $p_B = 0.0628$ which is the only test not rejecting the null hypothesis.

The case Real II consists of winter time data. During the time of ice cover the changes of downstream levels must be restrained. To provide an example, the quality measure is the day-maximum difference in the downstream water level. The results are shown in the figure 8 and again the test B separates from the others as it is the only one that would reject the null hypothesis ($p_B = 0.0348$). However, the differences between the p-values are quite small.

6 Discussing the methods and the results

The test based on the identity score is affine invariant whereas the non-parametric tests are only location and scalar invariant. However, from practical point of view, it is often sufficient because for example unit conversions are often only a scalar multiplication and/or a location shift. For example,

[5]The main benefits of a hydro-power plant are the power and the real-time adjustment/control to balance production-consumption gap in the power grid. The day-average spilling is not the best possible measure of these benefits but will make a good example in this work.

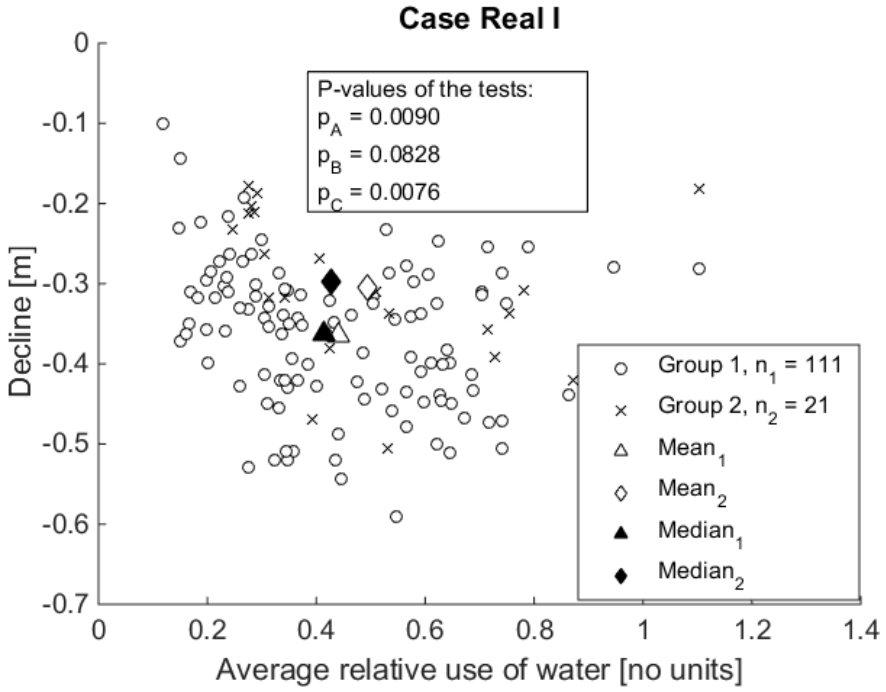

Figure 7: Case Real I: Summertime data from the years 2012 and 2013 form the sample 1 whereas the sample 2 consists of random days from the summer 2014. The y-axis is the maximum decline in the pool's level during the day and the x-axis is a measure of benefits, roughly approximated by the average spilling of the water through the plant. The subscripts of the p-values: A := MANOVA, B := Spatial signs and C := Spatial ranks. The ordinal number of the plant in the cascade: 13.

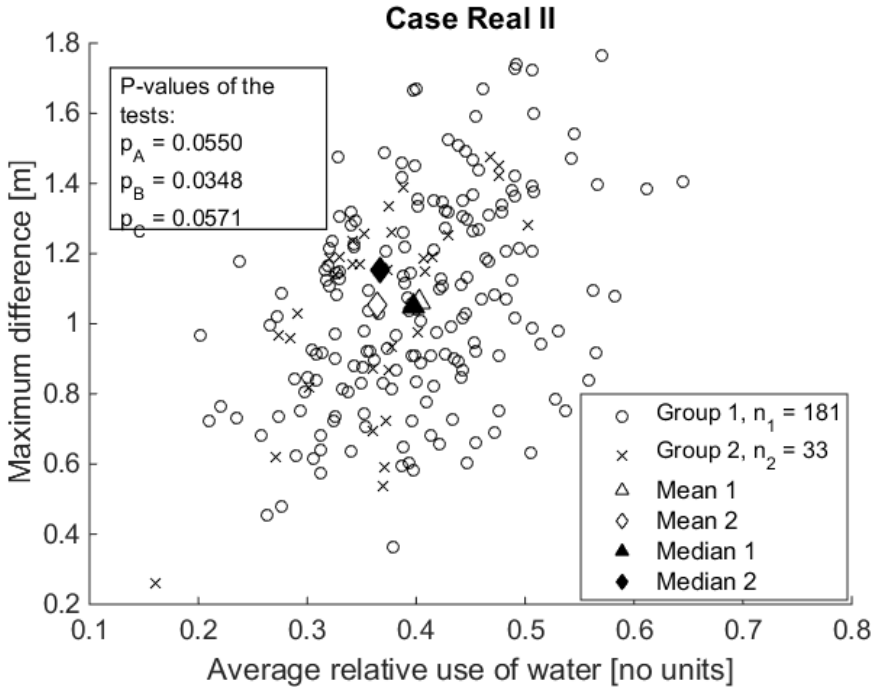

Figure 8: Case Real II: Winter time data from the years 2012 and 2013 form the sample 1 whereas the sample 2 consists of random days from the winter 2014. The y-axis is the maximum difference in the plant's downstream level during a day and the x-axis is a measure of benefits, approximated by the average spilling of the water through the plant. The subscripts of the p-values: A := MANOVA, B := Spatial signs and C := Spatial ranks. The ordinal number of the plant in the cascade: 14.

consider measuring heights of groups of persons and all the measurements are conducted using both inches and meter-scale. In that case, all the test statistics presented would not be affected by the unit change. Oja [2010] provides the affine invariant versions of the non-parametric test statistics as well but they are not presented in this work.

According to the figure 3, in the case 1 the tests based on the identity score function (MANOVA, denoted by A) and on the spatial ranks (C) seem to be quite identical whereas the test based on spatial signs (B) does not reject the null-hypothesis that often. But, when similar data is contaminated with some unsymmetrical outliers in the case 2, the differences between the tests become significant. This time A differs from the others two and it would not reject the null hypothesis in more than 10% of tests even if the theoretical locations differ remarkably. The location shift, though, is into the same direction as is the tail of the data. Hence, the shift of the smaller group is faded behind the outliers of the larger group. The figure 2 shows that the smaller sample is inside the extremes of the larger sample. Also, the outliers pull the sample means into the direction of the tail whereas the sample medians are more robust. In that very case the null hypothesis was rejected by all the three tests.

In the case 3 all the three groups shared the same covariance matrix which produced highly correlated data. The location shift was then into the direction of the smaller main-axis of the ellipse. This appeared to be slightly difficult to the test based on the spatial signs whereas the MANOVA and the test based on spatial ranks were again quite similar

In the case 4, the only difference between its variations a and b is the scatter parameter of the group 2, hence contradicting with the null-hypothesis $F_1 = \cdots = F_c$. In the case $4a$, the p-values of all the tests are more or less similar even though the spatial sign-test can be distinguished. In the case $4b$ the test B seems somewhat identical but the other tests reach the level of 100 %-rejections much later, and A is at significantly lower level.

To revisit the salary example presented in the chapter 3, consider the null hypothesis: the "centrals" of the salaries are equal in the two firms. The tests yield very distinct values of the test statistic and the p-values are $p_A = 1$, $p_B = 1$ and $p_C = 0.0008$. Hence, the tests A and B cannot recognize any difference between the two firms whereas C indicate a "very sure" difference. However, the data is quite odd to be studied utilizing these methods, but it is important to be aware that the classical MANOVA (A) is sensitive to its assumptions, especially when the sample sizes are "too" small.

To sum up the results of the simulations, the tests are mainly similar but there are drastic differences as well. For example, if the data sets follow multivariate normal distributions, then the tests A and C are practically the same but the A is much more sensitive to tailed data. Still, all the tests get affected by the anomalies generated to the data. The test A is the most sensitive to anomalies whereas the test B is the most robust one. On the other hand, the test B does not indicate well the location shift of the case study 3 whose idea was to move theoretical location parameter into the direction of smaller (co)variation. Therefore the test B might be too robust and lose some relevant information.

According to the real cases (I and II), the results of the tests A and C seem to be quite similar. The difference in the p-values in the case I is only $|p_A - p_C| = 0.0014$ and in the case II $|p_A - p_C| = 0.0021$, which both are practically insignificant. The test B differs from the others although the difference in the case II is not significant. The data sets of both cases more or less follow bivariate normal distribution and their marginal distributions are almost normal. Though, the scatter plot (7) has a shape of a banana. Apparently, the change in quality in the case II is much more questionable than the improve in the case I. The test B gave contradicting results and therefore, in these cases, it does not prove to be as reliable as the others.

7 Final Remarks

The purpose of this work was to discuss different test statistics to treat the several sample location problem. The parametric classical MANOVA with two different formulations and two non-parametric tests were considered. The MANOVA is affine invariant whereas the non-parametric tests are only location and scalar invariant which still enables numerous practical applications. The three tests were compared using both simulated data and real data. The practical frame of reference of the work is measuring quality and its changes of a complex process which in this case was producing energy in a hydro-power cascade.

As a conclusion, the tests based on spatial signs and ranks offer an alternative to the classical MANOVA in cases the assumptions of multinormality or equal scatter parameters do not exactly hold. The test based spatial ranks can be used whenever applying the MANOVA. In cases the anomalies of the data are significant, the MANOVA must be discarded and rank-test must be considered thoroughly. The most robust test against the anomalies is the test based on spatial signs but in some cases it might be too robust – leading to loss of relevant information. Still, no kind of guarantee of the methods proposed or general recommendations of their usage can be provided as a result of this work.

From the practical point of view, the relevance of the tools discussed in this work are subordinate to the selection of the data and to the quality measures. In practise, defining the quality measures may be ambiguous because many different values are involved. However, this problem would exist whichever family of methods were applied. For example, functional data analysis would offer inherently time-series based analysis tools but still the definition of the quality measures would involve similar ambiguous phases. Assuming relevant quality measures, the tools discussed in this work are suitable for finding changes in the process quality. In the end, the most important things in this kind of quality analysis are, in the first place, to comprehend the phenomenon under the investigation and, moreover, to know well some applicable tools to conduct the analysis.

References

Daniel Gervini. Robust functional estimation using the median and spherical principal components. *Biometrika 95, pp. 587-600*, 2008.

T. Kärkkäinen and S. Äyrämö. On computation of spatial median for robust data mining. *Evolutionary and Deterministic Methods for Design, Optimization and Control with Applications to Industrial and Societal Problems, EUROGEN*, 2005.

Hannu Oja. *Multivariate Nonparametric Methods With R.* Lecture Notes in Statistics. Springer, 2010.

A Appendix A: Matlab-functions programmed

The most important Matlab-functions programmed for this work are presented here. Some of the functions in this are fundamental to the topic, such as spatial signs whereas some of them are totally dependent on the notation (for example groupMean.m). In the end, there is a script presenting the procedure to conduct an experiment. The dependencies to external functions are shown, including the Matlab-functions that are not part of the Matlab-core installation. Many Matlab's core-functions are utilized even though they are not explicitly listed (i.e. sum, mean, cat, size etc).

```
function U = spatSigns(X)
%SPATSIGNS Calculates the spatial signs of
%data matrix X (n x p).
%   Copyright Sakke Rantala 2015

% L2-norm, sum into the dimension of p. Is zero IFF x_i == 0
norms = sqrt(sum(X.^2,2));
% Every row of X is divided by corresponding norm
U = bsxfun(@rdivide, X, norms);

% If some observation vector (row) was zero, then
% dividing by its norm will produce NaN's:
U(isnan(U)) = 0; % By definition, sign of 0-vector is 0.
end % spatSigns
```

```
function R = spatRanks(X)
%SPATRANK Calculates spatial (centered)
%rank of (n x p)-matrix X
%   R_i = AVE_j(U(y_i - y_j))
%   (c) Sakke Rantala 2015

[n,p] = size(X);

% WAY 1: using for-loops
% R = zeros(size(X));
% for i = 1:n
%     for j = 1:n
%         R(i,:) = R(i,:)  + spatSigns(X(i,:) -X(j,:))/n;
```

```
%     end
% end

% WAY 2: using bsxfun
% Is faster than WAY 1 because spatSigns.m
% can handle the matrices as well as vectors:
R = zeros(size(X));
for j = 1:n
    R = R + spatSigns(bsxfun(@minus, X,X(j,:)));
end
R = R/n; % Averaging

end % spatRanks
```

```
function [med,w,norms] = spatMed(x)
%function [med,w,norms] = spatMed(x)
%
%Spatial Median based on Functional Spatial Median (SpMed.m)
%proposed by Daniel Gervini,
%https://pantherfile.uwm.edu/gervini/www/RFDA.html,
%accessed 11.5.2015
%
% Input variables:
%   x: (n x p) - data matrix. Missing values or NaN values
%                not allowed.
%
% Output variables:
%   med: Spatial median (1 x m vector).
%   w: Weights (n x 1 vector). The spatial median is w'*x.
%   norms: Distances between each curve and
%           the median (n x 1 vector).
%   Note that the median minimizes sum(norms)
%   by definition, and that w = norms.^(-1)/sum(norms.^(-1)).

%% Input check
if any(any(isnan(x)))
    error('NaN-values are not allowed.')
end
[n,p] = size(x);
%% Inner-product matrix
```

```
% By Daniel Gervini, modifications by Sakke Rantala
A = 0.5*(x(:,1:p-1))*x(:,1:p-1)' + ...
    x(:,2:p)*x(:,2:p)';

%% Iterative minimization from sample mean
% By Daniel Gervini, the plotting property by Sakke Rantala
w = ones(n,1)/n;
norms = sqrt(diag(A) + w'*A*w - 2*A*w);
f = sum(norms);
err = 1;
iter = 0;
% Added by Rantala, to save the evolution of the iteration:
% W = [];
while err>1e-7 && iter<100
    iter = iter+1;
    f0 = f;
    if any(norms<eps)
        i0 = find(norms<eps);
        w = zeros(n,1);
        w(i0) = 1/length(i0);
    else
        w = 1./norms;
        w = w/sum(w);
    end
    norms = sqrt(diag(A) + w'*A*w - 2*A*w);
    f = sum(norms);
    err = abs(f/f0-1);
%     W = [W;w'*x];
end
%% Added by Rantala, to plot the evolution of the iteration
% iter
% figure
% plot(1:length(W), W)
%% Output
med = w'*x;

end % spatMed
```

```
function [Q, varargout] = statQ(T, N)
%STATQ Calculates the Q^2-test statistic out of
```

```
%the centered scores T (n x p)-matrix.
%@param N stores the number members in each group.
%   See reference:
%   Oja, Hannu, "Multivariate Nonparametric Methods with R",
%   2010.
%   N = [n_1, n_2, ... , n_c]
%   Note: sum(N) == n and size(T) = [n,p]
%   External functions used:
%   groupMean.m
%   sampleCovariance.m
%   chi2cdf (from the Matlab-statistic toolbox)
%   (C) Sakke Rantala 2015

% The dimensions of the (n x p)-score-matrix:
[n,p] = size(T);

% Calculate the mean of each group
Ti = groupMean(T,N);

% Evaluate the sample covariance matrix
% Using the first c-1 groups:
% NE = cumsum(N);
% B = sampleCovariance(T(1:NE(end-1),:));
% Using all the data:
B = sampleCovariance(T);

% Save the inverse into the memory as it might be
% applied multiple times in
% the for-loop below:
Bin = inv(B);

% Sum up the Q-value:
Q = 0;
for i = numel(N)
    Q = Q + N(i)*Ti(i,:)*Bin*Ti(i,:)';
end

% If p-value requested, return it using
% Matlab-Statistics toolbox function chi2cdf
if nargout > 1
    df = (numel(N) - 1)*p;
```

```
    varargout{1} = 1 - chi2cdf(Q,df); % p-value
end

end % statQ
```

```
function B = sampleCovariance(X)
%SAMPLECOVARIANCE Calculates the sample covariance
%of (n x p) -matrix
%   (c) Sakke Rantala 2015
[n,p] = size(X);
% Estimate for covariance of data, B = E[T_{ij}T_{ij}']
B = zeros(p,p); % Preallocation
for i = 1:n
   B = B + X(i,:)'*X(i,:); % (p x 1) * (1 x p) --> (p x p)
end
B = B/(n); % average
end % sampleCovariance
```

```
function M = groupMean(X, N)
%GROUPMEAN Evaluates (numel(N)) group means
%of the (n x p)-data matrix.
%   N = [n_1, n_2, ... , n_c] stores the group information
%   Note: n == sum(N).
%   (C) Sakke Rantala 2015
[n,p]= size(X); % Dimensions of the input matrix
g = numel(N); % Number of groups
if(n~=sum(N))
    disp('Dimensions do not match.')
    return
end
M = zeros(g,p); % Pre-allocation
NE = cumsum([0,N]);
for i = 1:g
    M(i,:) = mean(X(NE(i)+1:NE(i+1),:),1);
end
end % groupMean
```

```
function sp = groupSpMed(Y,N)
```

```
%GROUPSPMED Evaluates (numel(N)) group spatial medians of
% the (n x p)-data matrix.
%   N = [n_1, n_2, ... , n_c] stores the group information
%   Note: n == sum(N).
%   (C) Sakke Rantala 2015

[n,p]= size(Y);
g = numel(N);

if(n~=sum(N))
    disp('Dimensions do not match.')
    return
end

sp = zeros(g,p); %Pre-allocation
NE = cumsum([0,N]);
for i = 1:g
    sp(i,:) = SpMed(1:p, Y(NE(i)+1:NE(i+1),:));
end

end % groupSpMed
```

```
function M = groupMember(X, N)
%GROUPMEMBER Creates a group membership matrix
%(another formulation of MANOVA)
%   M_ij = 1, if ith observation
%   belongs to the jth group.
%   Assumptions: The groups are in order in @param X
%   and the number of the members of the groups are
%   determined by the vector N = [n_1, n_2, ... , n_c].
%   (c) Sakke Rantala 2015
[n,p] = size(X);
c = numel(N);
M = zeros(n,c); % Pre-allocation
N = cumsum([0,N]);
for i = 1:c
    M(N(i)+1:N(i+1),i) = 1;
end
end % groupMember
```

The Matlab-script below contains only the core information regarding to the experiment conducted, based on the functions presented earlier. For example plotting and saving procedures are mainly omitted.

```
%%% EXPERIMENTS: SEVERAL SAMPLE LOCATION PROBLEM

% *** Preparation for the testing ***
% GROUP SIZES n_i (number of elements define
% the number of groups):
N = [20,100,20]; % USER SET: Groups
p = 2;  % USER SET: Number of dimensions:

% N = [80,20];
NT = cumsum([0,N]);
n = sum(N); % Number of observations
c = numel(N); % Number of groups

% SCATTER PARAMETER (USER SET):
% 1. Neat experiments:
S = 2*eye(p); % % Just for simplicity only p-balls
% 2. For the case 3
% S = 3*(1 - 2*rand(p, p));
% S = S*S'; % To make it positive-definite
% lambda = eig(S);
% [v,l] = eig(S);

% LOCATION PARAMETERS (USER SET):
MU = [0,0,0,0;1,1,2,2;2,2,4,5];

% MU = randi([-5, 5], c, p); % Random locations
% Location determined by the scatter parameter:
% MU = [0,0; 1.5*lambda(1)*(v(:,1))'; 3*lambda(1)*v(:,1)'];

% THE NUMBER OF SIMULATIONS (USER SET)
numT = 100;%50; % different values of the scale factor
numS = 1000;%200; % Simulations/test
% The scale parameter:
b = linspace(0,1,numT);
```

```
% PRE-ALLOCATION
% The value of test statistic for each test
QA = zeros(numT,numS); % TestA
QB = zeros(numT,numS); % Test B
QC = zeros(numT,numS); % Test C
 % p-value of each test
PA = zeros(numT,numS); % Test A
PB = zeros(numT,numS); % Test B
PC = zeros(numT,numS); % Test C

% Counters: how many tests are rejected:
p1 = 0.05; % threshold p1
p2 = 0.01; % threshold p2
p3 = 0.001; % threshold p3

tt = tic;
for i = 1:numT
    % Group center parameters:
    mu = b(i)*MU; % scaled by b(i)
    for j = 1:numS
        % *** GENERATE DATA ***
        % DATA simulation
        X = zeros(sum(N),p);
        for k = 1:numel(N)
            G = bsxfun(@plus, mu(k,:), randn(N(k),p)*S);
% %             Generating unsymmetric outliers
%               (uncomment if needed):
%             a = rand(N(k),1) > 0.8;
%             G(a,:) = -abs( bsxfun(@plus, mu(k,:),...
%                           5*randn(sum(a),p)*S));
            % ADD GROUP TO THE DATA MATRIX
            X(NT(k)+1:NT(k+1),:) = G;
        end
%         X = exp(X); % Log-normal random variables

        % *** TESTING ***

        % TEST A: MANOVA
        Tij = bsxfun(@minus, X, mean(X,1));
        [QA(i,j), PA(i,j)] = statQ(Tij,N);
```

```
        % TEST B: SPATIAL SIGN
        spm = SpMed(1:p, X);
        % By choosing the spatial median as the shift vector:
        shiftedX = bsxfun(@minus, X, spm);
        % We get the inner centered scores (AVE(hatUij) = 0)
        hatUij = spatSigns(shiftedX);
        [QB(i,j), PB(i,j)] = statQ(hatUij, N);

        % TEST C: SPATIAL RANKS
        Rij = spatRanks(X); %
        [QC(i,j), PC(i,j)] = statQ(Rij,N);
    end
    % Timing the procedure
    disp('-----time---------')
    disp(i)
    disp(toc(tt))
end

% *** ANALYSIS ***
% How many times null hypothesis is rejected?
TA1 = PA < p1;
TB1 = PB < p1;
TC1 = PC < p1;

TA2 = PA < p2;
TB2 = PB < p2;
TC2 = PC < p2;

TA3 = PA < p3;
TB3 = PB < p3;
TC3 = PC < p3;

%% An example of PLOTTING
ta = [sum(TA1,2), sum(TA2,2), sum(TA3,2)]/numS*100;
tb = [sum(TB1,2), sum(TB2,2), sum(TB3,2)]/numS*100;
tc = [sum(TC1,2), sum(TC2,2), sum(TC3,2)]/numS*100;

t1 = [sum(TA1,2), sum(TB1,2), sum(TC1,2)]/numS*100;
t2 = [sum(TA2,2), sum(TB2,2), sum(TC2,2)]/numS*100;
t3 = [sum(TA3,2), sum(TB3,2), sum(TC3,2)]/numS*100;
```

```
figure; plot(b,t1') % p1
figure; plot(b,t2') % p2
figure; plot(b,t3') % p3
```