Journal of Statistics Applications & Probability
An International Journal

http://dx.doi.org/10.18576/jsap/110215

On Compare Robust Wilks' Statistics for the One-Way MANOVA: A Simulation Study and Application

Abdullah A. Ameen[1] *and Osama H. Abbas*[2,*]

[1]Department of Mathematics, College of Science, University of Basrah, Basrah, Iraq
[2]General Directorate of Education, Basrah, Iraq

Received: 26 Jan. 2021, Revised: 12 Apr. 2021, Accepted: 4 May 2021
Published online: 1 May 2022

Abstract: The classical Wilks' statistic is mostly used to test hypotheses in the one-way multivariate analysis of variance (MANOVA), which is highly sensitive to the effects of outliers. The non-robustness of the test statistics based on normal theory has led many authors to examine various options. In 2010, Todorov and Filzmoser proposed a robust Wilks' statistic depends on reweighted minimum covariance determinant estimator (RMCD) and constructed its approximate distribution. In this paper, we presented a robust version of the Wilks' statistics based on reweighted minimum covariance determinant estimator and reweighted minimum volume ellipsoid estimator, and constructed it's another approximate distribution depends on the weights of observations, where the weights are calculated based on Hampel weight function. A comparison was made between the proposed statistics, classical Wilks' statistic, and the robust Wilks' statistic which is proposed by Todorov and Filzmoser. The Monte Carlo studies are used to obtain performance assessment of test statistics in different data sets. Moreover, the results of the type I error rate and the power of test were considered as statistical tools to compare test statistics. The study reveals that, under normally distributed, the type I error rates for the classical and the proposed Wilks' statistics are close to the true significance levels, and the power of the test statistics are so close. In addition, in the case of contaminated distribution, the proposed statistics is the best. A real data are used to further evaluate the proposed robust statistics in this study.

Keywords: One-Way Multivariate Analysis of Variance, Outliers, Robustness, Minimum Covariance Determinant Estimator, Minimum Volume Ellipsoid Estimator, Wilks' Statistic.

1 Introduction

One-way MANOVA deals with testing the null hypothesis H_0 of equal mean vectors of multivariate normal groups. To formalize the hypothesis, let us assume that there are many independent random groups, say $k \geq 2$ groups, for every sample there are n_i multivariate normal observations $\mathbf{y}_{ij},\ i = 1,2,\ldots,k,\ j = 1,2,\ldots,n_i$ of p dimension with mean vector $\boldsymbol{\mu}_i$ and equal covariance matrix $\boldsymbol{\Sigma}$. Then, the null and alternative hypotheses can be written as:

$$H_0 : \boldsymbol{\mu}_1 = \boldsymbol{\mu}_2 = \ldots = \boldsymbol{\mu}_k,$$

$$H_1 : \boldsymbol{\mu}_i \neq \boldsymbol{\mu}_j \, for \, at \, least \, one \, i \neq j.$$

Many of statistics used for testing H_0, one of the most widely used is Wilks' statistic Λ which is defined as (see Rencher, (2002) [1]):

$$\Lambda = \frac{|\boldsymbol{W}|}{|\boldsymbol{W}+\boldsymbol{B}|}, \tag{1}$$

where $\boldsymbol{B}$ and $\boldsymbol{W}$ are the "between" and "within" of $p \times p$ matrices, respectively, having the formulas:

$$\boldsymbol{B} = \sum_{i=1}^{k} n_i(\bar{\mathbf{y}}_{i.} - \bar{\mathbf{y}}_{..})(\bar{\mathbf{y}}_{i.} - \bar{\mathbf{y}}_{..})^t, \tag{2}$$

[*] Corresponding author e-mail: osama.stat82@gmail.com

$$\boldsymbol{W}=\sum_{i=1}^{k}\sum_{j=1}^{n_i}(\mathbf{y}_{ij}-\overline{\mathbf{y}}_{i.})(\mathbf{y}_{ij}-\overline{\mathbf{y}}_{i.})^t, \tag{3}$$

where

$$\overline{\mathbf{y}}_{i.}=\frac{1}{n_i}\sum_{i=1}^{n_i}\mathbf{y}_{ij}, \qquad \overline{\mathbf{y}}_{..}=\frac{1}{n}\sum_{i=1}^{k}\sum_{j=1}^{n_i}\mathbf{y}_{ij},\ and \quad n=\sum_{i=1}^{k}n_i.$$

The hypothesis H_0 is reject if $\Lambda \leqslant \Lambda_{\alpha,p,v_W,v_B}$ where $\Lambda_{\alpha,p,v_W,v_B}$ is the exact critical values for Wilks' table with level of significance α and degrees of freedom p, $v_W=n-k$ and $v_B=k-1$.

Assuming that all the groups are originate from the multivariate normal distribution, many classical statistics are extremely sensitive to the influence of outliers (see [2]). Several statistics have been presented which are robust against possible outliers in the data. In 1985, Nath and Pavur [3] presented an alternative statistic for the one-way MANOVA depend on the rank order of the data. In the one-group, Hotelling's statistic is the basic tool for inference about the mean of a multivariate normal distribution. Willems et al. (2002) [4] introduced a robust Hotelling's statistic depend on the minimum covariance determinant (MCD) estimator. Candan and Aktas (2003) [5] proposed another robust Hotelling's statistic upon minimum volume ellipsoid (MVE) estimator. In 2010, Todorov and Filzmoser [6] introduced a robust Wilks' statistic for the one-way MANOVA depend on MCD estimator. Van Aelst and Willems (2011) [7] used S and MM-estimators to construct a robust Wilks' statistic for testing the hypotheses in the one-way MANOVA.

The effect of outliers on the Wilks' statistic will be explained in the simulation study in Section 5. Therefore, we introduce another alternative robust Wilks' statistics to the classical Wilks' statistic and has approximation differs from those suggested by Todorov and Filzmoser. The MCD and MVE estimators that proposed by Rousseeuw in (1984)[8], and (1985)[9], respectively, are highly robust estimator of location and scatter, for this purpose they are used. To increase efficiency while retaining high robustness, one can apply reweighted steps for MCD estimator (RMCD), and MVE estimator (RMVE) which are summarized in Section 2. The robust Wilks' statistics are reviewed in section 3. In Section 4, we construct the proposed approximations and examine their accuracy. A simulation study is used to evaluate the proposed statistical performance and to compare the different test statistics in different distribution cases in terms of significance level, the power of the test and robustness. Section 5 describes the simulation study and its results. To further evaluate the proposed robust statistics, a real data set are used in Section 6.

2 Robust Estimators

To construct the robust Wilks' statistics, we want to estimate the multivariate parameters of the data set. The MCD and MVE estimators of Rousseeuw (1984)[8], and (1985) [9], respectively, are a highly robust estimators of multivariate location and scatter. MCD estimator looks for a subset of h observations with the lowest determinant of the sample covariance matrix, where the subset size h is selected between half and the full size of sample. The mean observations of the subset h represent the MCD location estimate $\boldsymbol{T}$ and a multiple of its covariance matrix is the MCD scatter estimate $\boldsymbol{C}$. The effective algorithm for calculating the MCD estimates is found in most known statistical software packages such as R, $S|Plus$, SAS, and $Matlab$. The minimum volume ellipsoid (MVE) estimator was the first popular high breakdown point estimator of location and scatter. It was searches for the ellipsoid of minimal volume containing at least half of the points in the data set $\boldsymbol{Y}$ of n observations. The location estimate is defined as the center of this ellipsoid and the covariance estimate is provided by its shape. The effective algorithm for calculating the MVE estimates is found in the statistical software packages R. To increase the efficiency of the MCD and MVE estimators, a reweighted version is used. Several methods have been proposed to estimate the common covariance matrix. The method which was introduced by He and Fung (2000) [10] for S estimates and by Hubert and Van Driessen (2004) [11] for MCD estimates is used. In this method, the observations $\mathbf{y}_{ij}$ are centered and pooled as a single sample $\boldsymbol{Z}=\mathbf{z}_{ij}$ to estimate the covariance matrix. First, it starts by computing the location estimates $\mathbf{t}_i,\ i=1,2,\ldots,k$ for each group as the MCD or MVE location estimates. These group means are swept from the original observations for centralized observations $\mathbf{z}_{ij}=\mathbf{y}_{ij}-\mathbf{t}_i$. Second, the common covariance matrix $\hat{\boldsymbol{\Sigma}}_z$ is estimated as the MCD or MVE covariance matrix of the centered observations $\boldsymbol{Z}$. Finally, the location estimate $\hat{\boldsymbol{\mu}}_z$ of $\boldsymbol{Z}$ is used to adjust the group means $\hat{\boldsymbol{\mu}}_i=\hat{\boldsymbol{\mu}}_z+\boldsymbol{t}_i,\ i=1,2,\ldots,k$.

3 The Robust Wilks' Statistic

Assuming that all groups arise from the multivariate normal distribution, the classical Wilks' statistic is very sensitive to the influence of outliers. Therefore, Nath and Pavur [3] were presented the robust Wilks' statistic based on the ranks of the observations. Also, Todorov and Filzmoser [6] were introduced an alternative proposal for the Wilks' statistic based on RMCD estimator defined as:

$$\Lambda_R = \frac{|\boldsymbol{W}_R|}{|\boldsymbol{B}_R + \boldsymbol{W}_R|}, \tag{4}$$

where $\boldsymbol{B}_R$ and $\boldsymbol{W}_R$ are the weighted "between" and "within" matrices given by:

$$\boldsymbol{B}_R = \sum_{i=1}^{k} w_{i.}(\bar{\mathbf{y}}_{w_{i.}} - \bar{\mathbf{y}}_{w..})(\bar{\mathbf{y}}_{w_{i.}} - \bar{\mathbf{y}}_{w..})^t \tag{5}$$

$$\boldsymbol{W}_R = \sum_{i=1}^{k}\sum_{j=1}^{n_i} w_{ij}(\mathbf{y}_{ij} - \bar{\mathbf{y}}_{w_{i.}})(\mathbf{y}_{ij} - \bar{\mathbf{y}}_{w_{i.}})^t \tag{6}$$

where
$w_{i.} = \sum_{j=1}^{n_i} w_{ij}$, $\bar{\mathbf{y}}_{w_{i.}} = \frac{1}{w_{i.}}\sum_{j=1}^{n_i} w_{ij}\mathbf{y}_{ij}$, $\bar{\mathbf{y}}_{w..} = \sum_{i=1}^{k}\sum_{j=1}^{n_i}\frac{1}{w}w_{ij}\mathbf{y}_{ij}$, and $w = \sum_{i=1}^{k} w_{i.}$
The weights w_{ij} for each observation $\mathbf{y}_{ij}$ computed by the Huber weight function defined as:

$$w_{ij} = \begin{cases} 1 & MD(\mathbf{y}_{ij}) \leq \sqrt{\chi^2_{p,0.975}}, \\ 0 & otherwise. \end{cases}$$

4 The proposed approximation distribution of Wilks' statistic

The distribution of classical Wilks' statistic Λ, which was considered by Anderson (1958) [12] as a ratio of two Wishart distributions, is very complicated. Therefore, Bartlett introduced a good approximation of the Wilks' statistic given by (see [1]):

$$-\left(\nu_E - \frac{1}{2}(p - \nu_H + 1)\right)\ln\Lambda \simeq \chi^2_{p\nu_H}. \tag{7}$$

Todorov and Filzmoser (2007) [6] were assumed for Λ_R the following approximation:

$$L_R = -\ln\Lambda_R \simeq d\,\chi^2_q, \tag{8}$$

where the multiplication factor d and the degrees of freedom q of the χ^2 distribution as

$$d = \frac{E[L_R]}{q}, \ and \quad q = \frac{2E[L_R]^2}{Var[L_R]}.$$

The mean $E[L_R]$ and variance $Var[L_R]$ of the robust Wilks' statistic Λ_R are not possible to obtain analytically. So, they are determined them by simulation after repeated m times as:

$$ave[L_R] = \frac{1}{m}\sum_{i=1}^{m} L_R^{(i)}, \ and \quad var[L_R] = \frac{1}{m-1}\sum_{i=1}^{m}(L_R^{(i)} - ave[L_R])^2.$$

The estimated parameters d and q will be reused to analyze data with the same dimension and number of groups. To perform the robust Wilks' statistic that proposed by Todorov and Filzmoser, it will take a lot of time during simulations to find d and q for approximate distribution. In this the present study, we introduce the following suggestions:

- In order to increase efficiency of RMCD estimator while retaining high robustness, we followed the following approach:
  1. Compute the location estimators $\hat{\boldsymbol{\mu}}_i^0$, $i = 1,2,\ldots,k$ and the common covariance matrix $\hat{\boldsymbol{\Sigma}}_z^0$ based on RMCD estimator.
  2. Compute the weights w_{ij} of the observations $\mathbf{y}_{ij}$ by the Hampel weight function that is defined as (see Campbell, (1980) [13]):

$$w_{ij} = \begin{cases} 1 & MD(\mathbf{y}_{ij}) \leq d_0 \\ d/MD(\mathbf{y}_{ij}) & MD(\mathbf{y}_{ij}) > d_0, \end{cases} \tag{9}$$

where

$$d = d_0 \exp\left(\frac{-1}{2}\left(\frac{MD(\mathbf{y}_{ij}) - d_0}{b_2}\right)^2\right), \quad d_0 = \sqrt{p} + \frac{b_1}{\sqrt{2}}, \quad b_1 = 2, \quad b_2 = 1.25,$$

and the Mahalanobis distances $MD(\mathbf{y}_{ij}) = \sqrt{(\mathbf{y}_{ij} - \hat{\boldsymbol{\mu}}_i^0)^t \hat{\boldsymbol{\Sigma}}_{\mathbf{z}}^{\mathbf{0}\,-1} (\mathbf{y}_{ij} - \hat{\boldsymbol{\mu}}_i^0)}$.

3.For $g = g+1$, compute the weighted location estimators $\hat{\boldsymbol{\mu}}_i^g$ and the weighted common covariance matrix $\hat{\boldsymbol{\Sigma}}_z^g$ as:

$$\hat{\boldsymbol{\mu}}_i^g = \frac{1}{\sum_{j=1}^{n_i} w_{ij}} \sum_{j=1}^{n_i} w_{ij} \mathbf{y}_{ij},$$

$$\hat{\boldsymbol{\Sigma}}_z^g = \frac{1}{\sum_{i=1}^{k} \sum_{j=1}^{n_i} w_{ij} - 1} \sum_{i=1}^{k} \sum_{j=1}^{n_i} w_{ij}^2 (\mathbf{y}_{ij} - \hat{\boldsymbol{\mu}}_i^g)(\mathbf{y}_{ij} - \hat{\boldsymbol{\mu}}_i^g)^t.$$

4.Repeat until the measure of deviation from sphericity, $\phi(\hat{\boldsymbol{\Sigma}}_z^g) = \frac{(tr(\hat{\boldsymbol{\Sigma}}_z^g/p))^p}{det(\hat{\boldsymbol{\Sigma}}_z^g)}$ (see [14]) needs to be as small as possible.

- To perform the robust Wilks' statistic Λ_R that proposed by Todorov and Filzmoser, it will take a lot of time during simulations to find d and q for approximate distribution. Therefore, we will introduce a robust version of Wilks' statistic is similar to Λ_R in (7), namely Λ_{R_1}, but based on RMCD estimator with Hampel weight function and constructed its approximate distribution defined as:

$$-\left(v_{W_R} - \frac{1}{2}(p - v_{B_R} + 1)\right) \ln \Lambda_{R_1} \simeq \chi^2_{p\,v_{B_R}}. \tag{10}$$

To compute the degrees of freedom v_{W_R}, and v_{B_R} for the robust Wilks' statistic Λ_R, the sum of squares $\boldsymbol{B}_R$ and $\boldsymbol{W}_R$ in (5) and (6) can be written as:

$$\boldsymbol{B}_R = \boldsymbol{Y}^t (\boldsymbol{Q}_n - \boldsymbol{P}_n) \boldsymbol{Y},$$

$$\boldsymbol{W}_R = \boldsymbol{Y}^t (\boldsymbol{W}_n - \boldsymbol{Q}_n) \boldsymbol{Y},$$

where $\boldsymbol{Y}$ is the data matrix, $\boldsymbol{Q}_n = diag(\boldsymbol{Q}_{ii})$, $\boldsymbol{Q}_{ii} = [\frac{1}{w_i} w_{ij} w_{ih}]$ is a block diagonal matrix with $k \times k$ blocks of size $n_i \times n_i$, $\boldsymbol{P}_{ii} = [\frac{1}{w} w_{ij} w_{ih}]$ is a block matrix with $k \times k$ blocks of size $n_i \times n_i$, and $\boldsymbol{W}_n = diag(w_{ii})$, $w_n = diag(w_{ij})$, $i = 1, 2, \ldots, k$, $j = 1, 2, \ldots, n_i$, $h = 1, 2, \ldots, n_i$.
So,

$$v_{W_R} = trace(\boldsymbol{W}_n - \boldsymbol{Q}_n) = w - \sum_{i=1}^{k} \frac{v_i}{w_i}, \quad and \quad v_{B_R} = trace(\boldsymbol{Q}_n - \boldsymbol{P}_n) = \sum_{i=1}^{k} \frac{v_i}{w_i} - \frac{\sum_{i=1}^{k} v_i}{w},$$

where $v_i = \sum_{j=1}^{n_i} w_{ij}^2$.

- Similarly to above procedures, a robust Wilks' statistic, namely Λ_{R2}, depends on RMVE estimator with Hampel weight function is introduced.

Now we will investigate the accuracy of the approximation of Λ_{R1}, and Λ_{R2} by computing the robust Wilks' statistics Λ_{R1}, and Λ_{R2} for $m = 3000$ samples from the standard normal distribution and several values of the dimension p, the number of groups k and the sample sizes n_i, $i = 1, 2, \ldots, k$. The distribution of these m statistics will be compared to the approximate distribution of Λ_{R1}, and Λ_{R2} by QQ-plots, some of them are shown in Figures (1), and (2). The usual cutoff values of a test, 95%, 97.5%, and 99% are shown in these plots of vertical lines. One can see from these plots that the approximations are accurate for lower and higher dimensions, large and small sample sizes, and for equal and unequal groups sizes.

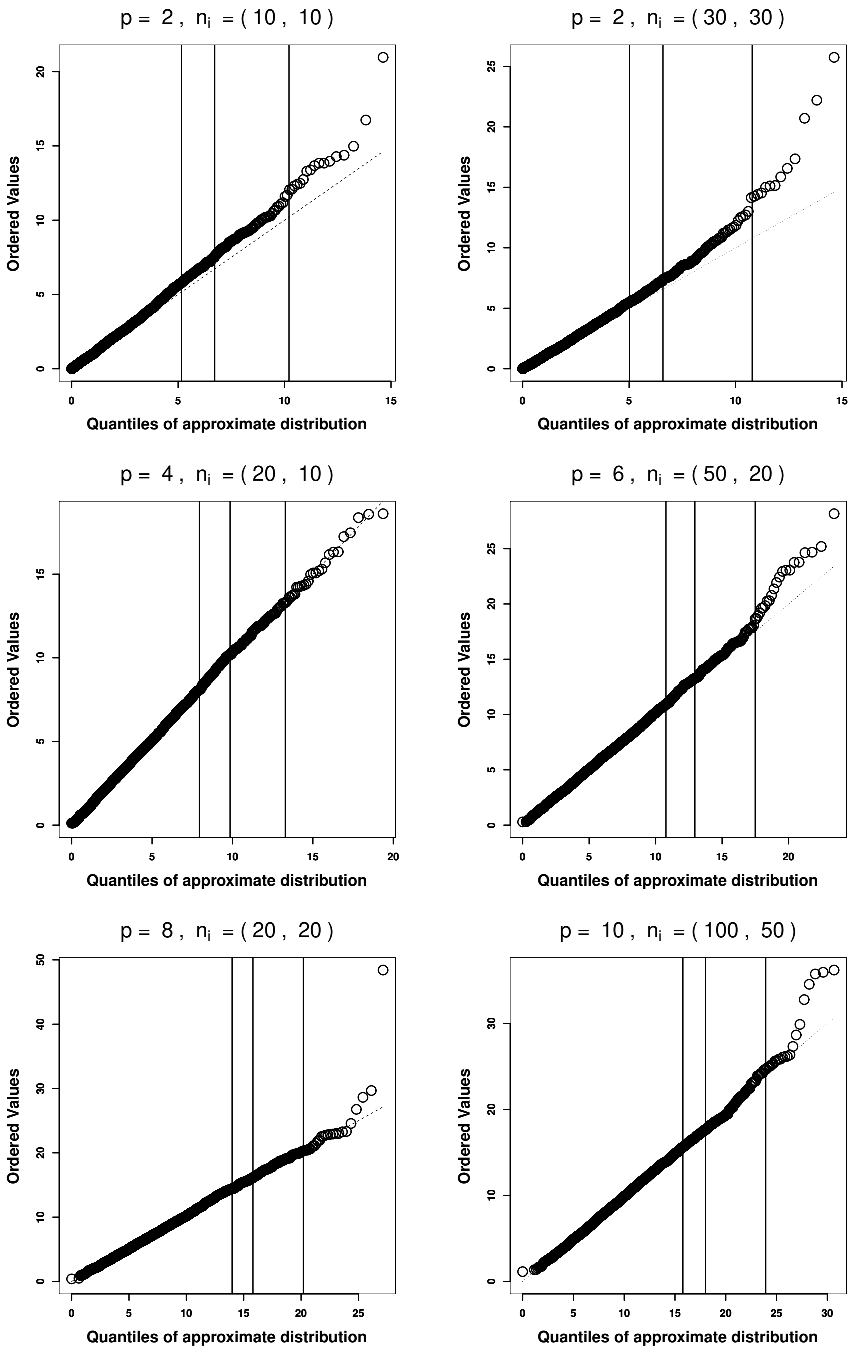

Fig. 1: QQ-plots for the proposed robust Wilks' statistic based on RMCD estimator Λ_{R1} in the case of two groups and several dimension values for p and $n=\sum_{i=1}^{k} n_i$.

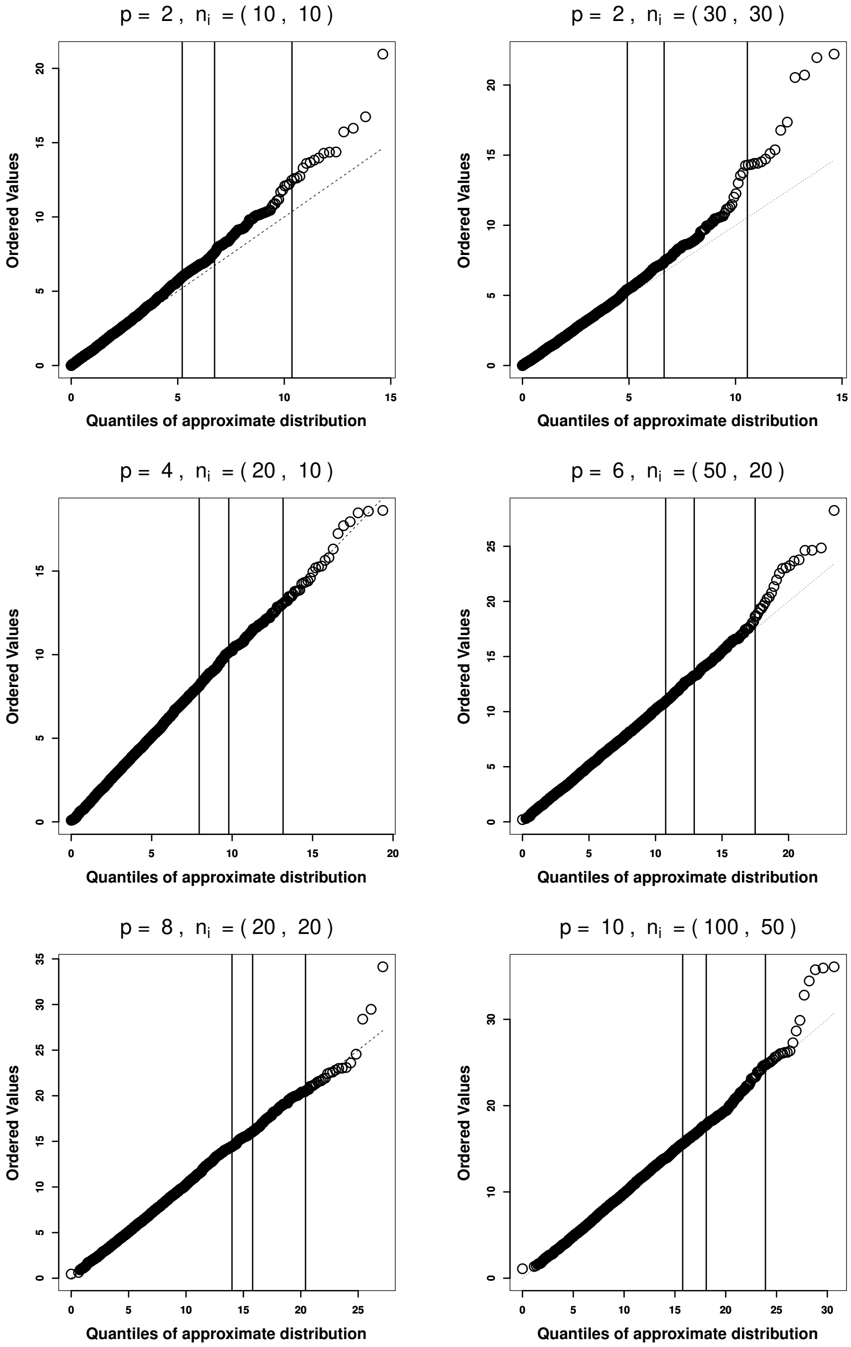

Fig. 2: QQ-plots for the proposed robust Wilks' statistic based on RMVE estimator Λ_{R2} in the case of two groups and several dimension values for p and $n = \sum_{i=1}^{k} n_i$.

5 Monte Carlo Simulation

Monte Carlo study is a good method to assess the statistical performance for the test statistics. The evaluation of the performance of the test statistics includes two measures the type I error rate and the power of the test. In addition, we will investigate the robust statistics behavior in the existence of outliers and compare the results with the classic Wilks' statistic. To study the type I error rate and the power of test of the robust statistics, let us consider number of groups $k = 2,3$, several dimension $p = 2,4,6,8,10$, and sample sizes n_i, $i = 1,2,\ldots,k$. The selected sample sizes are shown in Table (1).

Table 1: Selected group sizes for the simulation study

Two groups (n_1,n_2)	Three groups (n_1,n_2,n_3)
(10, 10)	(10, 10, 10)
(20, 10)	(20, 10, 10)
(20, 20)	(20, 20, 20)
(30, 20)	(30, 20, 10)
(30, 30)	(30, 30, 10)
(50, 20)	(50, 20, 10)
(50, 50)	(50, 50, 20)
(100, 50)	(100, 50, 30)

5.1 Significance level

To compare the type I error rates $\hat{\alpha}$ for the test statistics, we generate the observations from the multivariate normal distribution $\mathbf{y}_{ij} \sim N_p(\mathbf{0},\boldsymbol{I})$ under the null hypothesis $H_0 : \boldsymbol{\mu}_1 = \boldsymbol{\mu}_2 = \ldots = \boldsymbol{\mu}_k = \mathbf{0}$. The classical Wilks' statistic Λ is compared to the Bartlett' χ^2 approximation given in (7), the robust Wilks' statistic Λ_R is compared to the approximation given in (8), and the proposed Wilks' statistics Λ_{R1}, and Λ_{R2} are compared to the approximate distribution given as described in Section 4. This is repeated $m = 3000$ times and then calculate $\hat{\alpha} = L(T)/m$ (where $L(T)$ is the number of times of rejected the test statistic when the hypothesis is true) for the test statistics above. The values $\hat{\alpha}$ are taken as an estimate of the true significance level when the simulated critical values are above the true significance level. The true significance level $\hat{\alpha} = 0.01,\ 0.05,$ and 0.10 with the number of times $m = 3000$, and from the standard error formula of Salter and Fawcett (1989) [15] $\alpha \mp 2\sqrt{\alpha(1-\alpha)/m}$ gives the standard deviation interval about the nominal level as $(0.089, 0.111)$, $(0.042, 0.058)$, and $(0.006, 0.014)$ respectively. In Table (2) , the results of the type I error rates $\hat{\alpha}$ are shown for two groups. It is clear that $\hat{\alpha}$ of the test statistics are very close to the nominal value α (true significance level). We will use the P-value plots proposed by Davidson and McKinnon (1998) [16], which gives a more complete picture of how the test statistics follow the approximate distribution under the null hypothesis in the simulated samples. Figures (3) and (4) show P-value plots of test statistics in three groups $k = 3$ of the multivariate normal distribution, several dimensions p and the sample size $n = \sum_{i=1}^{k} n_i$. It is seen that the test statistics Λ, Λ_{R1}, and Λ_{R2} are close to the 45° line, and the robust Wilks' statistic Λ_R is considerably below the 45° line for small sample sizes.

5.2 Power of test

To compare the power of the test $\hat{\pi}$ for the test statistics we will generate data samples $\mathbf{y}_{ij} \sim N_p(\boldsymbol{\mu}_i,\boldsymbol{I})$ under an alternative hypothesis $(H_1 : \boldsymbol{\mu}_i \neq \boldsymbol{\mu}_j\, for\, at\, least\, one\, i \neq j)$. Also, we will use the same cases of dimensions p, number of groups k, and sample sizes $n_i,\ i = 1,2,\ldots,k$ but each sample has a different mean $\boldsymbol{\mu}_i = (\boldsymbol{\mu}_{i1},\boldsymbol{\mu}_{i2},\ldots,\boldsymbol{\mu}_{ip})^t$. The means of dimensions $p = 2,4,6,8,10$ for the groups $i = 1,2,3$ are selected as:

$$\boldsymbol{\mu}_1 = (0,0,\ldots,0)^t, \quad \boldsymbol{\mu}_2 = (0.5,0,\ldots,0)^t, \quad \boldsymbol{\mu}_3 = (0,0.5,\ldots,0)^t,\ldots, \quad \boldsymbol{\mu}_k = (0,0,\ldots,0.5,0)^t.$$

The power of the test statistics were compared by the resulting size-power curves under alternative hypothesis, as proposed by Davidson and MacKinnon (1998) [16]. The results for the three groups are shown in Figures (5) and (6). It is clearly seen that the size-power curves for the classical statistic Λ, and the proposed statistics Λ_{R1}, and Λ_{R2} are close while the robust Wilks' statistic Λ_R by Todorov and Filzmoser is less.

5.3 Robustness comparisons

Now we will investigate the robustness for the proposed test statistic in the one-way MANOVA. Therefore, we will generate data samples under the null and alternative, and we will contaminate them by adding outliers. The same cases of dimensions p, number of groups k, and sample sizes $i = 1,2,\ldots,k$ will be used.

5.3.1 Significance level

Under the hypothesis $H_0 : \boldsymbol{\mu}_1 = \boldsymbol{\mu}_2 = \ldots = \boldsymbol{\mu}_k$, the data will be generated from the following contamination model: $\mathbf{y}_{ij} \sim (1-\varepsilon)N_p(\mathbf{0},\boldsymbol{I}) + \varepsilon N_p(\boldsymbol{\mu}_i, c\boldsymbol{I})$, where $\varepsilon = 0.1$, $\mu^* = v\sqrt{\chi^2_{p,0.001}}\mathbf{1}^t_p$, $v = 5$, and $c = 0.0625$.

Table 2: Levels of significance of test statistics Λ, Λ_R, Λ_{R1}, and Λ_{R2} for two groups $k = 2$ of multivariate normal distribution, several values of the dimension p and the sample size $n = n_1 + n_2$.

Dimension	Sample Size		Statistic	Significance Level		
				0.01	0.05	0.1
$p=2$	10	10	Λ	0.009	0.046	0.090
			Λ_R	0.011	0.042	0.084
			Λ_{R1}	0.013	0.062	0.110
			Λ_{R2}	0.016	0.064	0.114
	30	30	Λ	0.008	0.055	0.103
			Λ_R	0.008	0.044	0.103
			Λ_{R1}	0.013	0.066	0.118
			Λ_{R2}	0.011	0.067	0.121
$p=4$	20	10	Λ	0.008	0.048	0.098
			Λ_R	0.013	0.050	0.085
			Λ_{R1}	0.009	0.053	0.105
			Λ_{R2}	0.010	0.054	0.106
	30	20	Λ	0.009	0.045	0.106
			Λ_R	0.012	0.043	0.084
			Λ_{R1}	0.011	0.049	0.107
			Λ_{R2}	0.011	0.049	0.108
$p=6$	20	20	Λ	0.014	0.050	0.099
			Λ_R	0.017	0.048	0.094
			Λ_{R1}	0.014	0.052	0.103
			Λ_{R2}	0.014	0.052	0.103
	50	50	Λ	0.009	0.044	0.094
			Λ_R	0.012	0.053	0.100
			Λ_R	0.009	0.045	0.099
			Λ_{R1}	0.008	0.045	0.099
$p=8$	20	20	Λ	0.008	0.052	0.106
			Λ_R	0.014	0.055	0.102
			Λ_{R1}	0.010	0.056	0.111
			Λ_{R2}	0.011	0.056	0.114
	50	50	Λ	0.011	0.056	0.097
			Λ_R	0.013	0.045	0.093
			Λ_{R1}	0.011	0.056	0.100
			Λ_{R2}	0.011	0.055	0.101
$p=10$	30	30	Λ	0.010	0.051	0.096
			Λ_R	0.013	0.043	0.086
			Λ_{R1}	0.010	0.050	0.097
			Λ_{R2}	0.010	0.049	0.097
	100	50	Λ	0.009	0.053	0.104
			Λ_R	0.010	0.046	0.101
			Λ_{R1}	0.009	0.053	0.107
			Λ_{R2}	0.009	0.052	0.106

The P-value plots of the test statistics for three groups are shown in Figures (7) and (8). In these Figures, the P-value plots (actual size) based on the test statistics Λ_{R1} and Λ_{R2} are so close to the 45° line compared to the same of the test statistic Λ_R, while the classical statistic Λ is very bad for all the different cases of dimension p and sample sizes.

5.3.2 Power of test

Under the alternative hypothesis $H_1 : \boldsymbol{\mu}_i \neq \boldsymbol{\mu}_j\, for\, at\, least\, one\, i \neq j$, the data samples will be generated from the following contamination model: $\mathbf{y}_{ij} \sim (1-\varepsilon)N_p(\boldsymbol{\mu}_i, \boldsymbol{I}) + \varepsilon N_p(\boldsymbol{\mu}^*, c\boldsymbol{I})$, where $\boldsymbol{\mu}_i$ are the same mean groups vectors as in Section 5.2, ε, $\boldsymbol{\mu}^*$, and c are take the same values as in Section 5.3.1. The Figures (9) and (10) show the size-power curves of test statistics. It is clearly seen that the proposed robust Wilks' statistics Λ_{R1} and Λ_{R2} are the best compared to the other statistics for all investigated cases of dimension p and sample sizes.

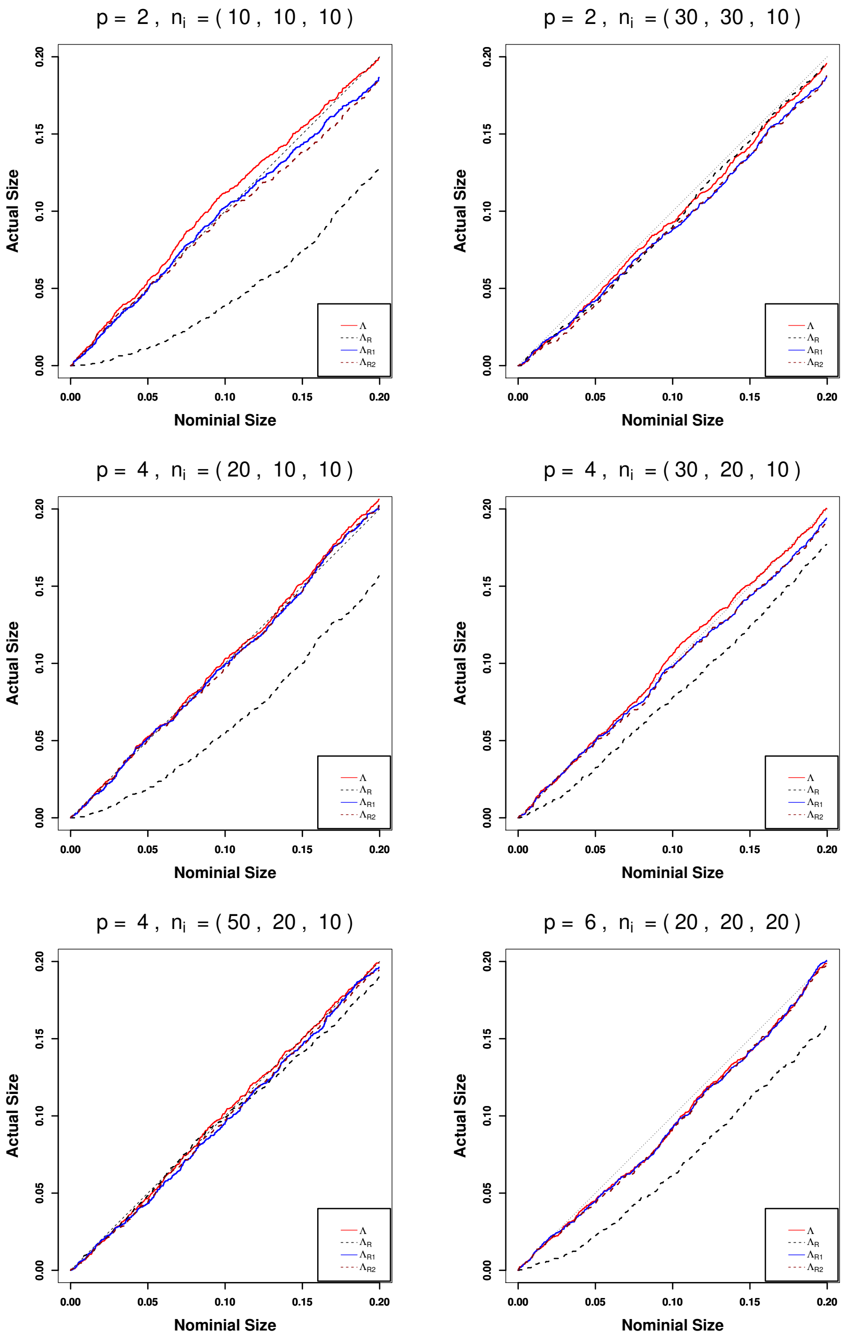

Fig. 3: P-value plots for test statistics Λ (red line), Λ_R (black line), Λ_{R1} (blue line), and Λ_{R2} (dark red line) for three groups $k = 3$ of multivariate normal distribution, several dimensions p and the sample size $n = n_1 + n_2$. The 45° line is given too.

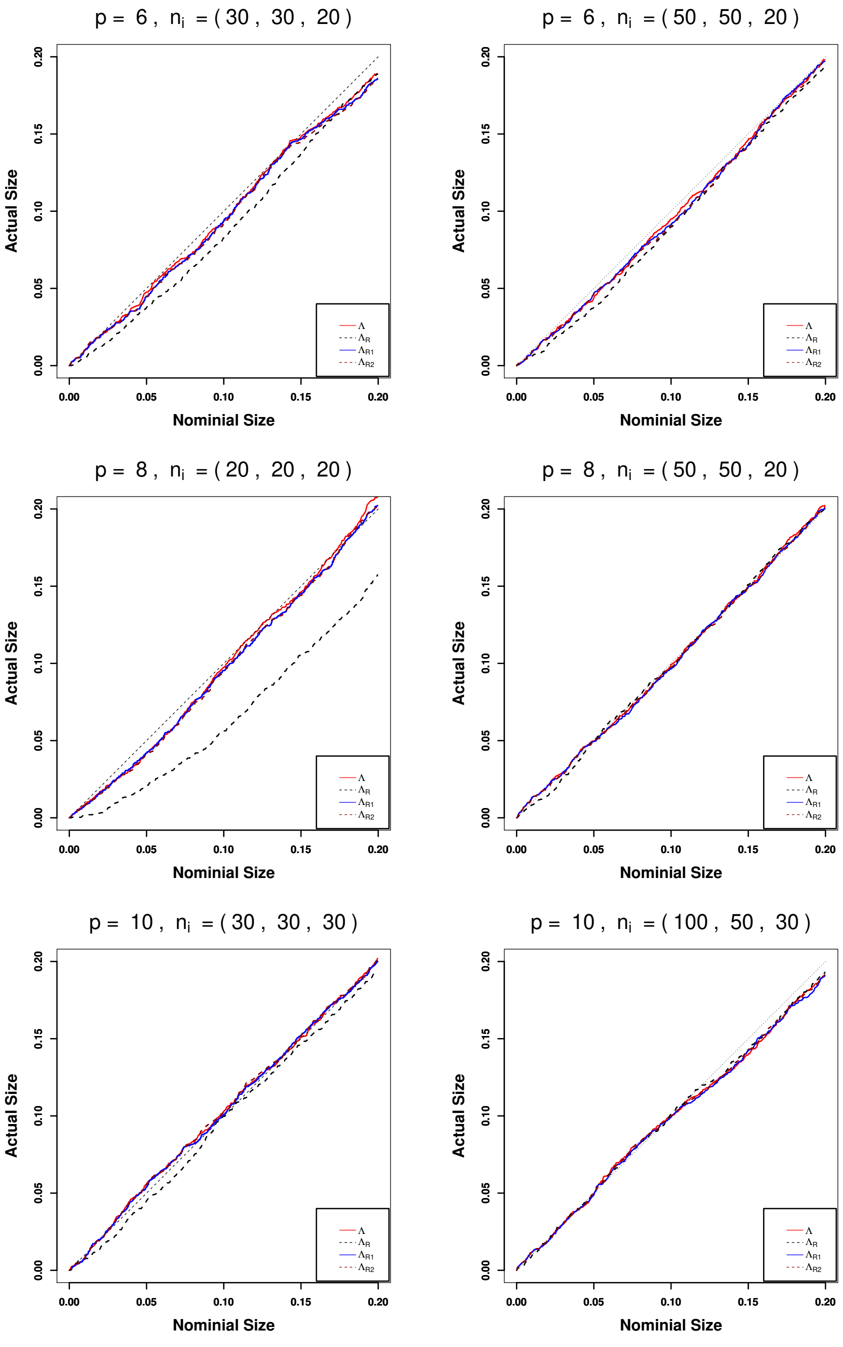

Fig. 4: P-value plots for test statistics Λ (red line), Λ_R (black line), Λ_{R1} (blue line), and Λ_{R2} (dark red line) for three groups $k = 3$ of multivariate normal distribution, several dimensions p and the sample size $n = n_1 + n_2$. The 45° line is given too.

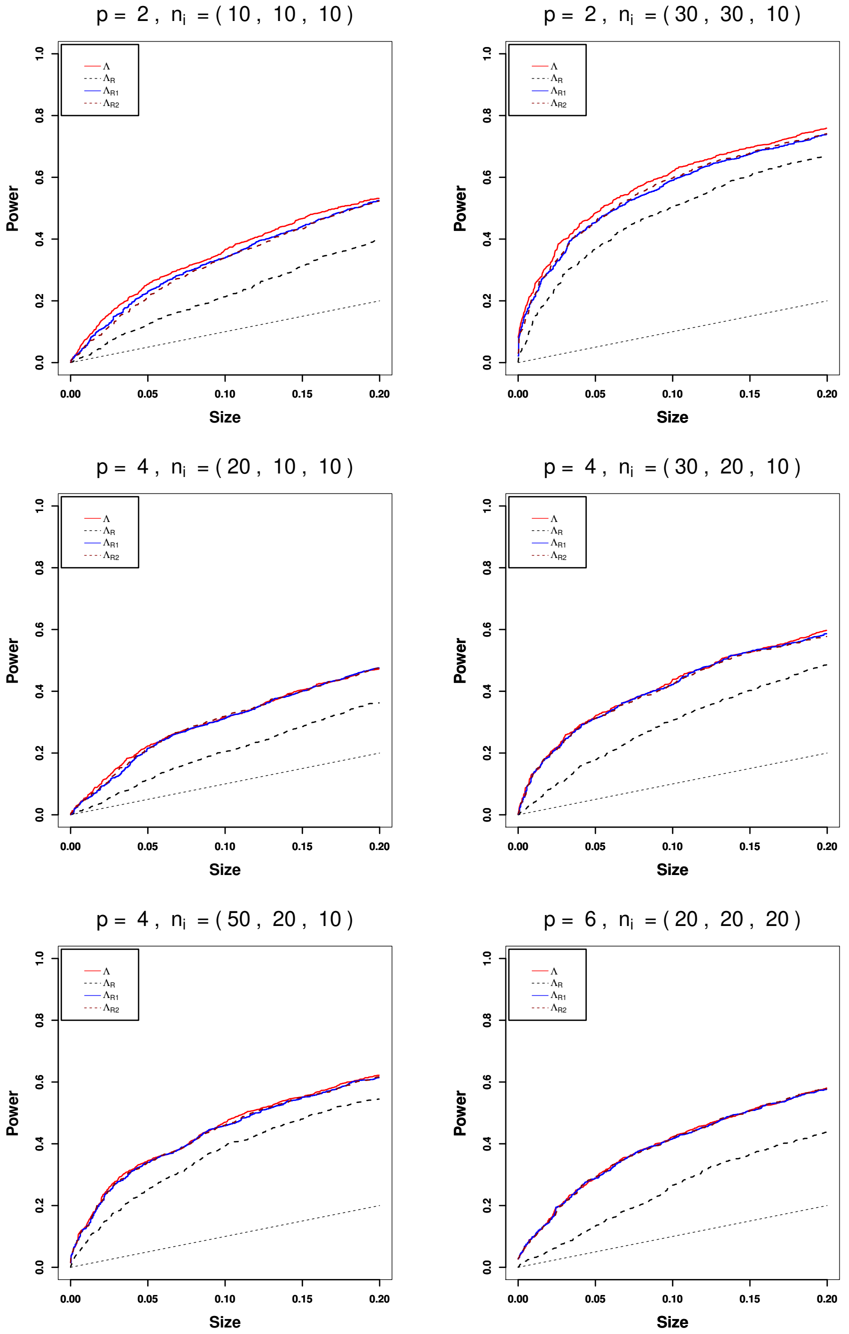

Fig. 5: Size-power curves for test statistics Λ (red line),Λ_R (black line),Λ_{R1} (blue line) , and Λ_{R2} (dark red line) for three groups $k = 3$ of multivariate normal distribution, several dimensions p and the sample size $n = \sum_{i=1}^{k} n_i$. The 45° line is given too.

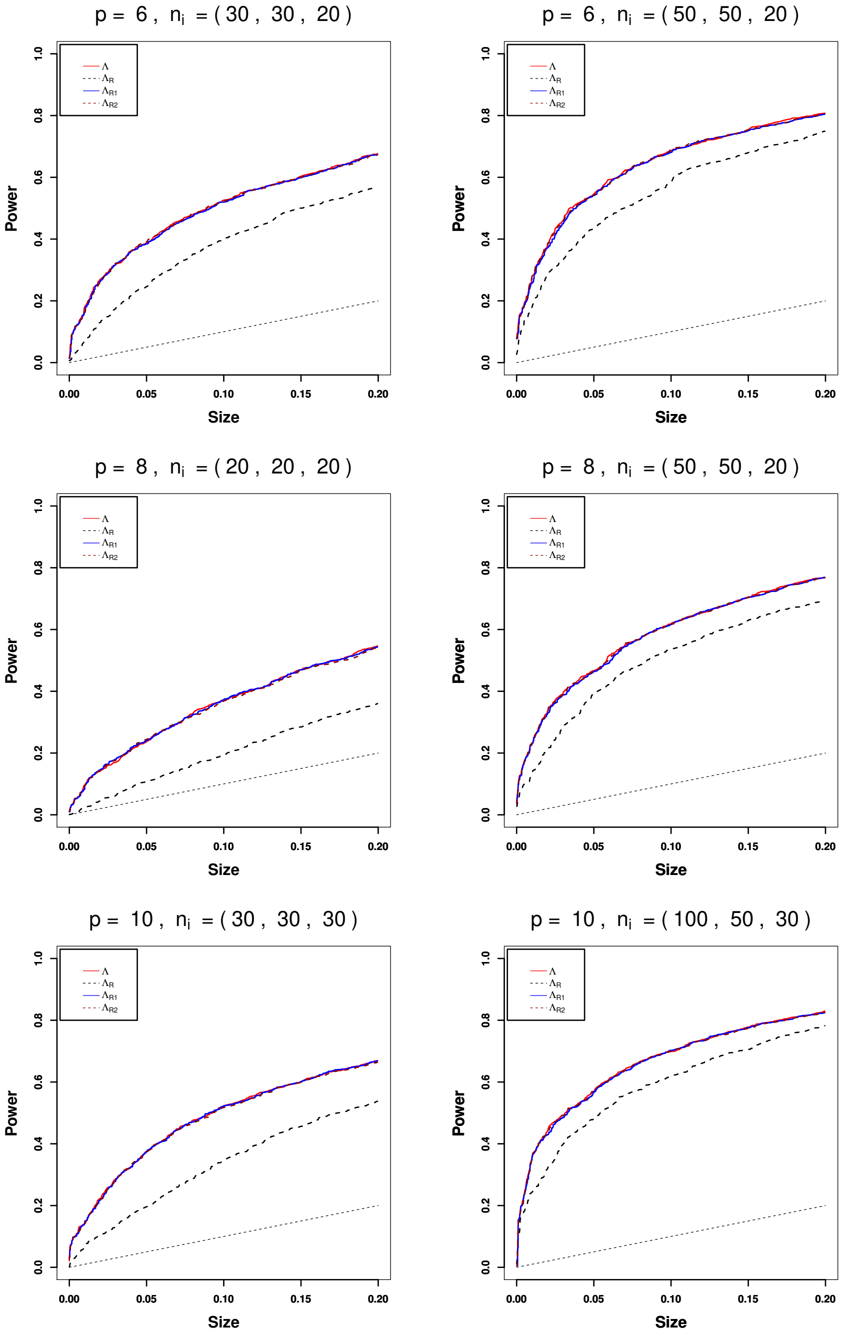

Fig. 6: Size-power curves for test statistics Λ (red line),Λ_R (black line),Λ_{R1} (blue line) , and Λ_{R2} (dark red line) for three groups $k = 3$ of multivariate normal distribution, several dimensions p and the sample size $n = \sum_{i=1}^{k} n_i$. The 45° line is given too.

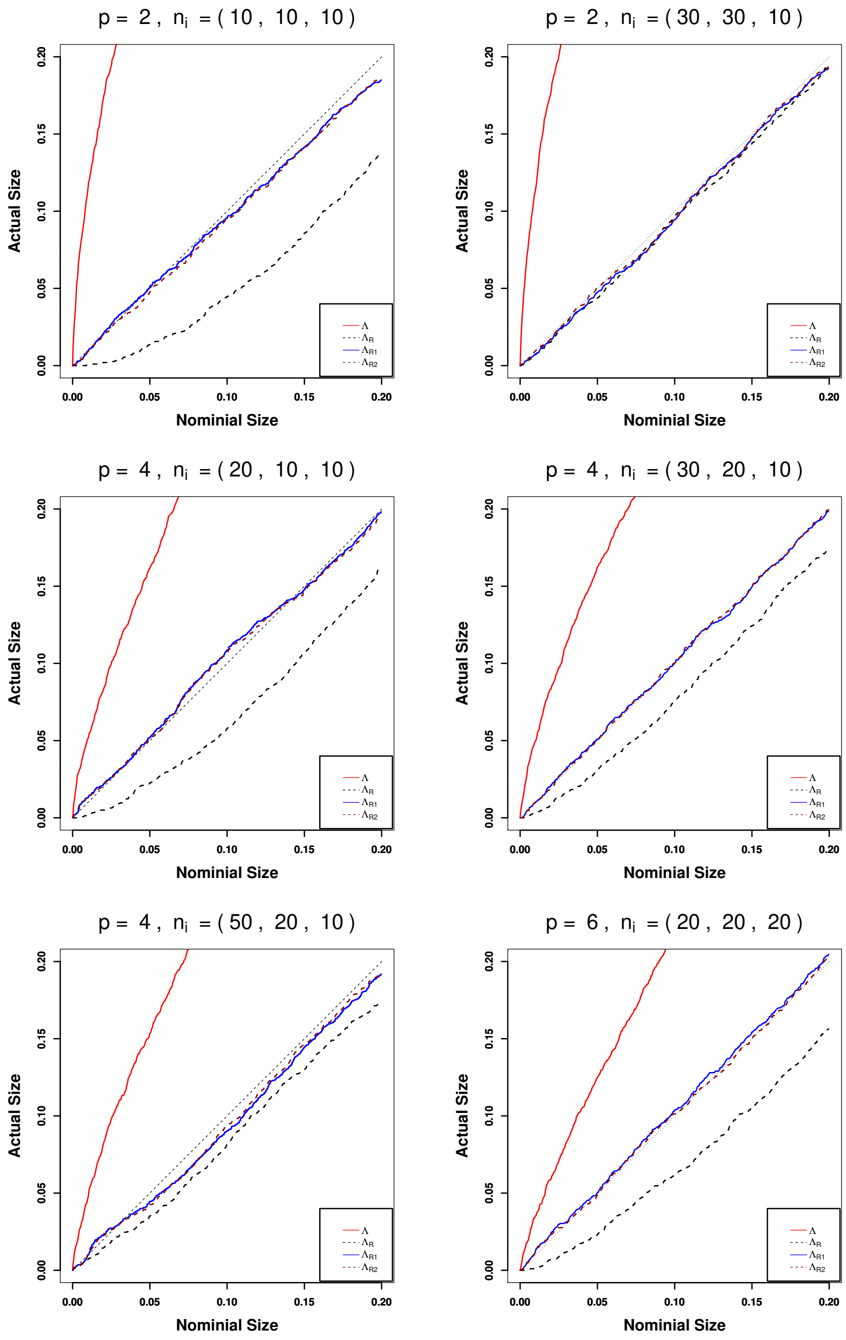

Fig. 7: P-value plots for test statistics Λ (red line), Λ_R (black line), Λ_{R1} (blue line), and Λ_{R2} (dark red line) for three groups $k = 3$ of multivariate contaminated distribution, several dimensions p and the sample size $n = n_1 + n_2$. The 45° line is given too.

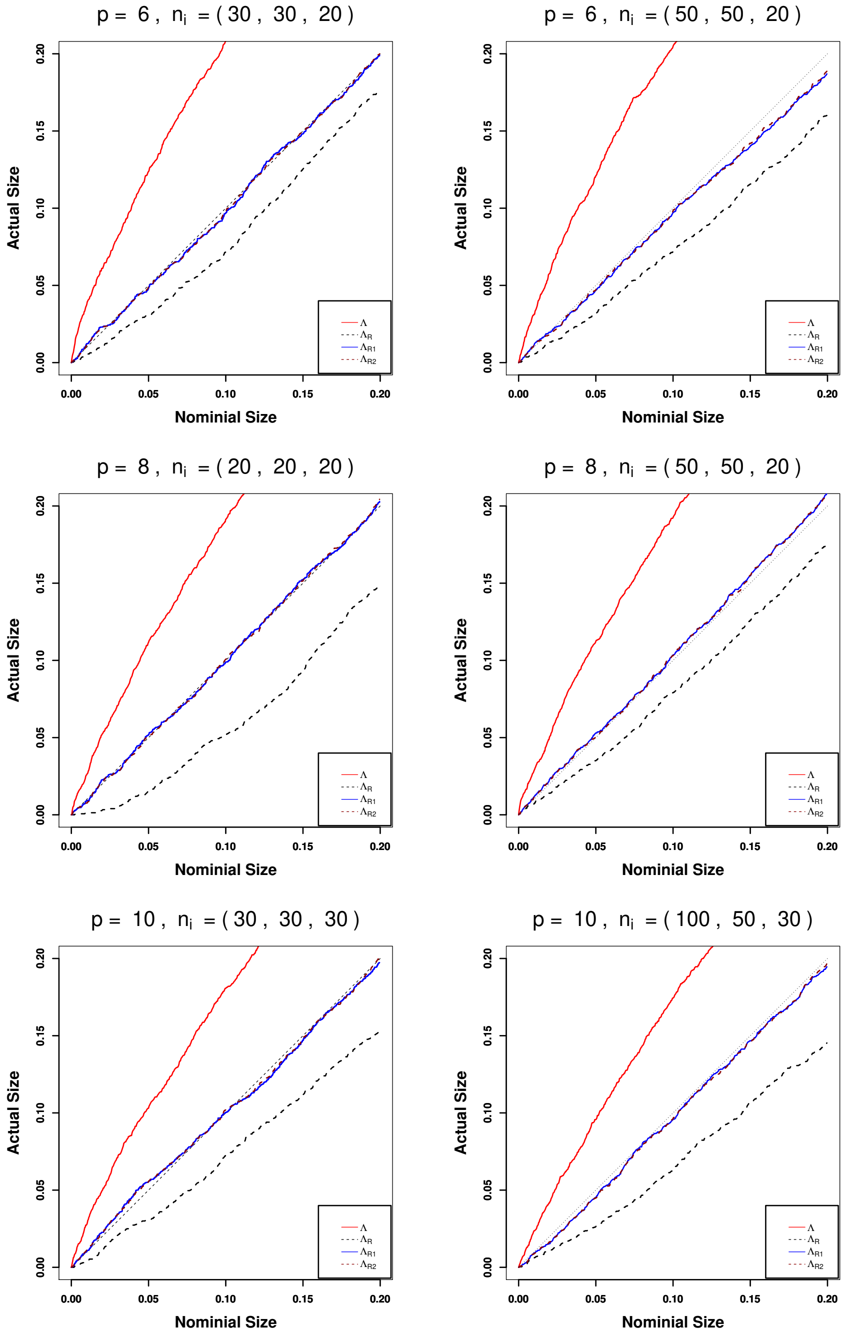

Fig. 8: P-value plots for test statistics Λ (red line), Λ_R (black line), Λ_{R1} (blue line), and Λ_{R2} (dark red line) for three groups $k = 3$ of multivariate contaminated distribution, several dimensions p and the sample size $n = n_1 + n_2$. The 45° line is given too.

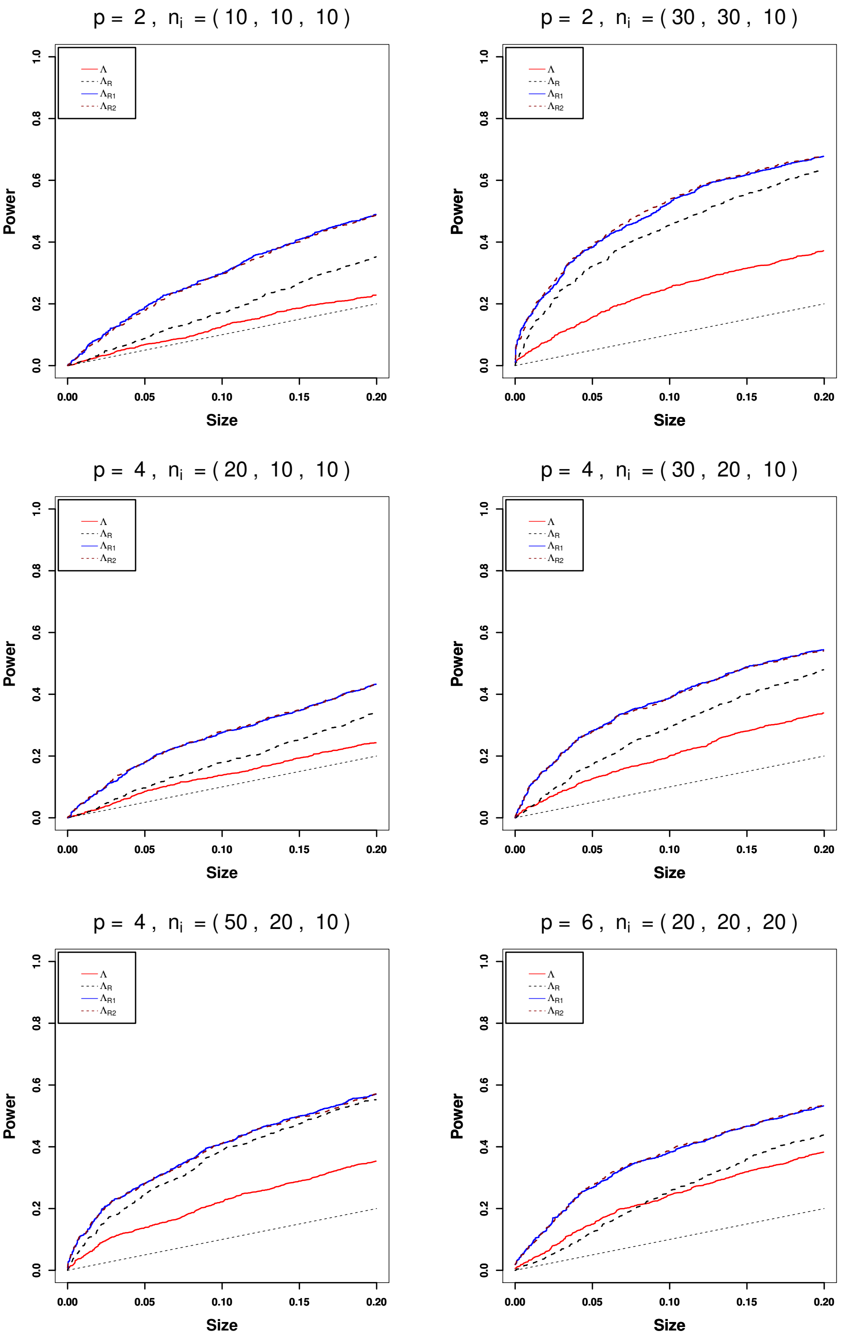

Fig. 9: Size-power curves for test statistics Λ (red line),Λ_R (black line),Λ_{R1} (blue line) , and Λ_{R2} (dark red line) for three groups $k = 3$ of multivariate contaminated distribution, several dimensions p and the sample size $n = \sum_{i=1}^{k} n_i$. The 45° line is given too.

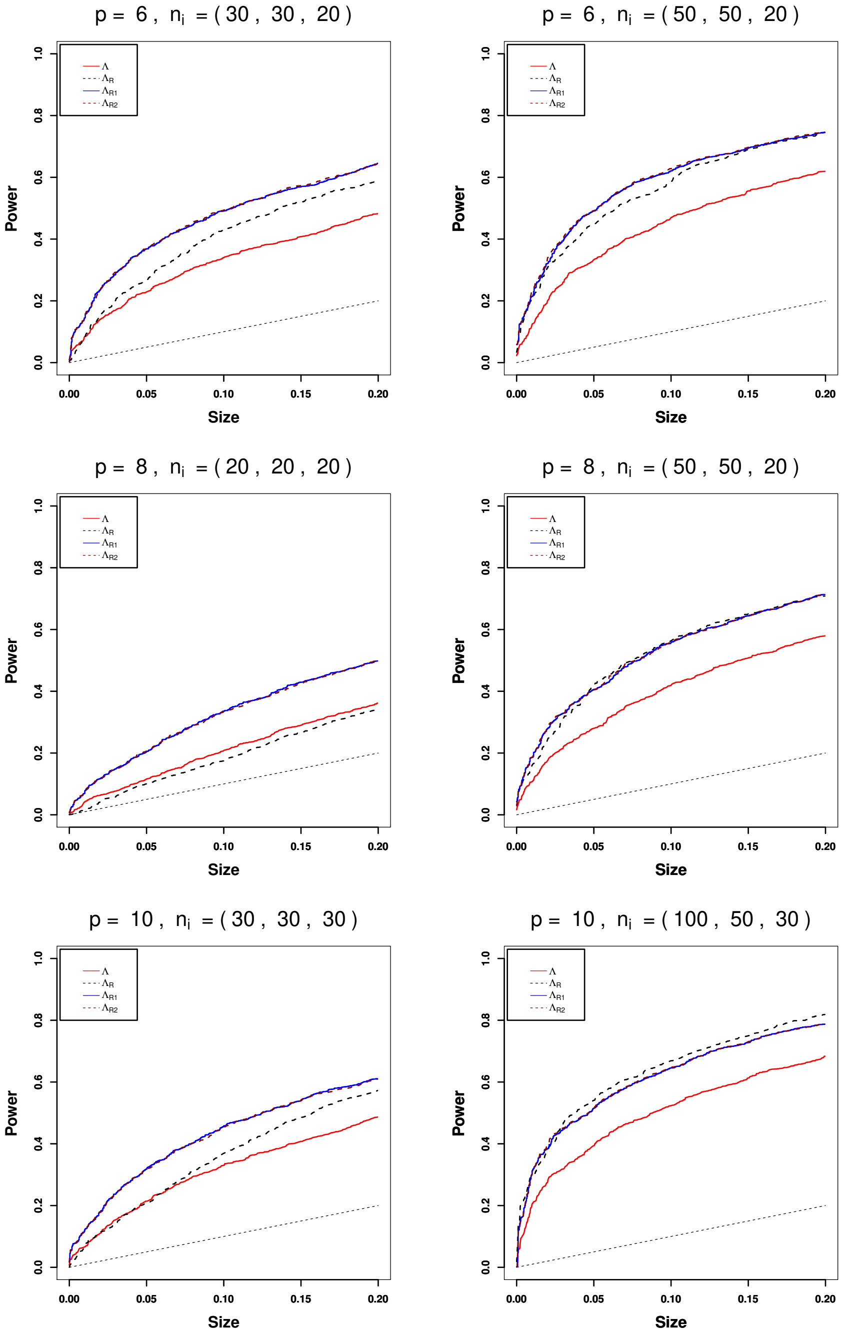

Fig. 10: Size-power curves for test statistics Λ (red line),Λ_R (black line),Λ_{R1} (blue line) , and Λ_{R2} (dark red line) for three groups $k = 3$ of multivariate contaminated distribution, several dimensions p and the sample size $n = \sum_{i=1}^{k} n_i$. The 45° line is given too.

6 Real Data Example

We will illustrate the application of the proposed robust Wilks' statistics using a real data set. The data set includes water data taken from (The Department of Protection and Improvement the Environment in the Southern Region of Iraq) represented by chemical concentrations in the waters of the Shatt Al-Arab River in Iraq over a the year period (2013) in three different stations $SH1$ (Karmat Ali river / water project (25) million unified Basra), $SH2$ (Al-Sanker /Abu Al-Khaseeb District), and $SH3$ (Al-Siba / near the Ceyhan water project). The number of analyzed water data in each station are 18.

For our example, we will only consider four main compounds, namely, Dissolved Oxygen (DO), Nitrate ($NO3$), Calcium (Ca), and Magnesium (Mg) aggregate the percentages of corresponding compounds. The first six observations of the raw data are given in the following Table:

Station	DO	NO_3	Ca	Mg
$SH1$	0.026	0.043	0.533	0.398
$SH1$	0.046	0.101	0.620	0.233
$SH2$	0.015	0.036	0.569	0.380
$SH2$	0.025	0.054	0.700	0.221
$SH3$	0.032	0.034	0.512	0.422
$SH3$	0.029	0.037	0.561	0.398

So, our data matrix consists of $n = 54$ rows and $p = 4$ columns. However, we note that as each row sums up to 1 the observations are compositions being part of the $3-$dimensional simplex (1986) [17].

Most methods from multivariate statistics developed for real valued data are misleading or inapplicable for compositional data. (2013) [18]. Hence, we use the isometric log-ratio (ilr) transformation which is an isometric linear mapping between the p-dimensional simplex and R^{p-1} to obtain a $2-$dimensional data matrix for further analysis. The top panel of Figure 11 shows the scatter plot matrix of the ilr-transformed data together with histograms of each variable. The bottom panel display grouped boxplots for the different factor combination groups.

We now perform a one way MANOVA using the ilr-transformed water data. The hypothesis was tested using the classical Wilks' test statistic Λ, the robust Wilks' of Todorov Λ_R and the proposed robust Wilks' statistics Λ_{R1}, and Λ_{R2}. The P-values of the corresponding statistical testing are given in the Table 3. For the tests based on the classical Wilks' statistic the hypothesis testing cannot be rejected at a significance level of $\alpha = 0.05$, whereas for Λ_R, Λ_{R1}, and Λ_{R2} tests we can reject the hypothesis true for one way MANOVA test.

Table 3: P-values for the classical Wilks' statistic, the robust Wilks' of Todorov and the proposed robust Wilks' statistics.

Statistic	Λ	Λ_R	Λ_{R1}	Λ_{R2}
P-value	0.18923	0.00503	0.00014	0.00017

The results showed that the proposed methods are the best, as it came in accordance with the opinions of specialists in this field, that the different stations have a high impact on the rates of chemical concentrations.

7 Conclusions

In this study, we presented a robust version of the Wilks' statistic based on RMCD estimator Λ_{R1}, and Wilks' statistic based on RMVE estimator Λ_{R2}, and constructed their approximate distributions. The results show that the p-value plots and size-power curves for the proposed robust statistics are close to the classical in case of normal distribution for the data set, while in case of contaminated distribution the proposed robust statistics is the best. Also, the results show the advantage of the proposed robust statistics over the robust Wilks' statistic of Todorov, and Filzmoser especially with small sample sizes.

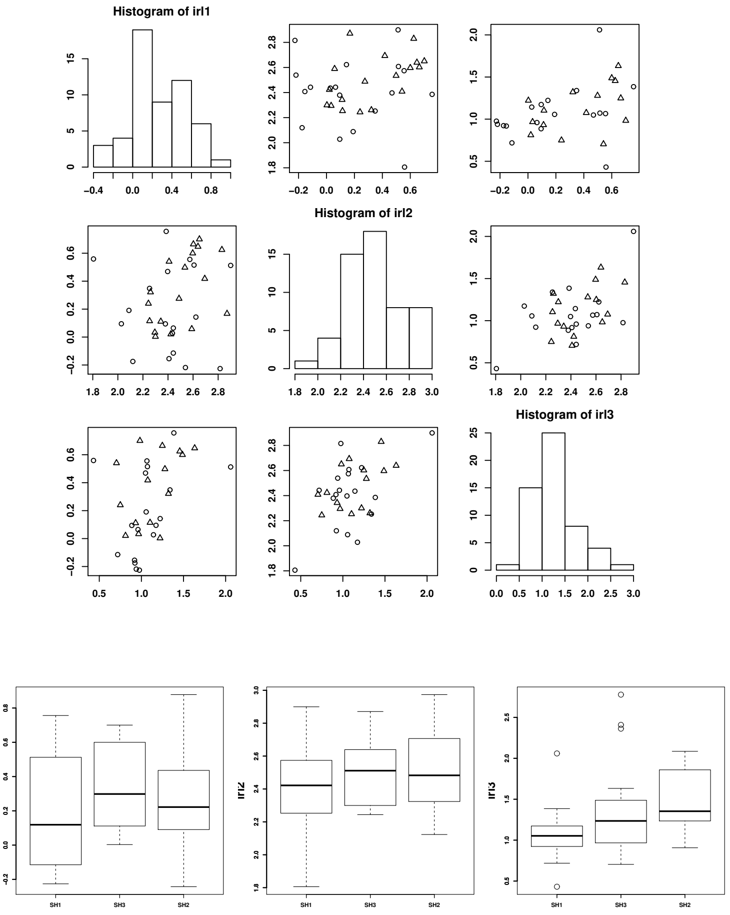

Fig. 11: Isometric log-ratio (ilr) transformed Water data: the top panel shows the scatter plot matrix of the ilr-transformed Water data together with histograms of each variable, the bottom panel display grouped boxplots for the different factor combination groups.

Conflicts of Interests

The authors declare that they have no conflicts of interests

References

[1] Rencher, A. C., Methods of Multivariate Analysis, Second Edition, Brigham Young University, John Wiley and Sons, New York, (2002).

[2] Beak, R. C., Cook, R. D., Outliers, Technometrics, 25, 119-149, (1983).

[3] Nath, R., and Pavur, R., A new statistic in the one-way multivariate analysis of variance, Computational Statistics and Data Analysis, 2(4), 297–315, (1985).

[4] Willems, G., Pison, G., Rousseeuw, P. J., and Van Aelst, S., A robust Hotelling Test, Metrika, 55,125-138,(2002).

[5] Meral Candan, and Serpil Aktas, Hotelling's Statistic Based on Minimum Volume Ellipsoid Estimator, G. U. Journal of Science , 16(4) , 691-695, (2003).

[6] Todorov, V., and Filzmoser, P., Robust Statistic for the One-way MANOVA, Computational Statistics and Data Analysis, 54(1), 37-48, (2010).

[7] Van Aelst, S. and Willems, G., Robust and Efficient One-way MANOVA Tests, Journal of the American Statistical Association, 106-494, (2011).

[8] Rousseeuw, P. J., Least median of squares regression, Journal of the American Statistical Association, 79(388), 871–880, (1984).

[9] Rousseeuw, P. J., Multivariate Estimation with High Breakdown Point", In Mathematical Statistics and Applications, Vol. B, 283-297, (1985).

[10] He, X. and Fung, W. K., High breakdown estimation for multiple populations with applications to discriminant analysis, J. Multivariate Anal, 72, 151–162, (2000).

[11] Hubert, M. and Van Driessen, K., Fast and robust discriminant analysis", Comput. Statist. Data Anal, 5, 301-320, (2004).

[12] Anderson, T., An Introdiction to Multivariate Statistical Analysis, John Wiley and Sons, New York, 1958.

[13] Campbell, N. A., Robust Procedures in Multivariate Analysis I: Robust Covariance Estimation, Applied Statistics, 29, 231–237, (1980).

[14] Croux C, Haesbroeck G., A note on finite-sample efficiencies of estimators for the Minimum Volume Ellipsoid, J Stat Comput Simulation, 72:585–596, (2002).

[15] Salter, K.C., and Fawcett, R.F., A robust and powerful rank test of treatment effects in balanced incomplete block designs, Communications in Partial Differential Equations, 14(4), 807-828, (1989).

[16] Davidson, R., McKinnon, J., Graphical methods for investigating the size and power of hypothesis tests, The Manchester School of Economic & Social Studies, 66 (1), pp. 1-26, (1998).

[17] Aitchison, J., The Statistical Analysis of Compositional Data. Chapman & Hall, London UK, (1986).

[18] van den Boogaart, K. G., Tolosana-Delgado, R., Analyzing Compositional Data with R. Springer, Heidelberg , (2013).