Robust Statistic for the One-way MANOVA ⋆

Valentin Todorov [a] and Peter Filzmoser [b]

[a] *United Nations Industrial Development Organization (UNIDO), Vienna International Centre, P.O. Box 300, A-1400 Vienna, Austria*

[b] *Department of Statistics and Probability Theory, Vienna University of Technology, Wiedner Hauptstr. 8-10, 1040 Vienna, Austria*

Abstract

The Wilks' Lambda Statistic (likelihood ratio test, LRT) is a commonly used tool for inference about the mean vectors of several multivariate normal populations. However, it is well known that the Wilks' Lambda statistic which is based on the classical normal theory estimates of generalized dispersions, is extremely sensitive to the influence of outliers. A robust multivariate statistic for the one-way MANOVA based on the Minimum Covariance Determinant (MCD) estimator will be presented. The classical Wilks' Lambda statistic is modified into a robust one through substituting the classical estimates by the highly robust and efficient reweighted MCD estimates. Monte Carlo simulations are used to evaluate the performance of the test statistic under various distributions in terms of the simulated significance levels, its power functions and robustness. The power of the robust and classical statistics is compared using size-power curves, for the construction of which no knowledge about the distribution of the statistics is necessary. As a real data application the mean vectors of an ecogeochemical data set are examined.

Key words: robustness, MCD, Wilks' Lambda, MANOVA

1 Introduction

One-way multivariate analysis of variance (MANOVA) deals with testing the null hypothesis of equal mean vectors across the g considered groups. The setup is similar to that of the one-way univariate analysis of variance (ANOVA)

⋆ The views expressed herein are those of the authors and do not necessarily reflect the views of the United Nations Industrial Development Organization.

Email addresses: v.todorov@unido.org (Valentin Todorov), p.filzmoser@tuwien.ac.at (Peter Filzmoser).

but the intercorrelations of the independent variables are taken into account, i.e. the variables are considered multivariate. Under the classical assumptions that all groups arise from multivariate normal distributions, many test statistics are discussed in the literature, one of the most widely used being the likelihood ratio test. This test statistic is better known as Wilks' Lambda in MANOVA. The Wilks' Lambda is reported as part of the test output in almost all statistical packages. However, this measure which uses the classical normal theory as well as the inference based on it, can be adversely affected by outliers present in the data. The non-robustness of the Wilks' Lambda statistic in the context of variable selection in linear discriminant analysis was demonstrated already in Todorov (2007a). In the one-sample case, where the Hotelling's T^2 statistic is the standard tool for inference about the center of a multivariate normal distribution, a robust version based on the Minimum Covariance Determinant Estimator was proposed by Willems *et al.* (2002).

The effect of outliers on the quality of the hypothesis test based on the classical Wilks' Lambda statistics will be illustrated in the examples in Section 6 and in the simulation study in Section 5. Therefore we propose to use robust estimators instead of the classical ones for computing Wilks' Lambda statistic. For this purpose we will use the Minimum Covariance Determinant (MCD) estimator of Rousseeuw (1985) which is one of many possible robust estimators of multivariate location and scatter (see Hubert *et al.*, 2008, for an overview). Furthermore a fast algorithm for approximating the MCD estimator is available - the FAST MCD of Rousseeuw and Van Driessen (1999). The adaptations of the MCD estimator for estimating the common covariance matrix will be summarized in Section 2. Since the distribution of the robust Wilks' Lambda statistic based on MCD differs from the classical one it is necessary to find a good approximation for this distribution. In Section 3 we construct an approximate distribution based on a Monte Carlo study and examine its accuracy.

Monte Carlo simulations are used to evaluate the performance of the proposed test statistic under various distributions in terms of the simulated significance levels, its power functions and robustness. The power of the robust and classical statistic is compared using size-power curves, for the construction of which no knowledge about the distribution of the statistic is necessary (see Davidson and McKinnon, 1998). Section 5 describes the design of the simulation study and its results and in Section 6 illustrative examples are presented.

The nonrobustness of the normal theory based test statistic has led many authors to search for alternatives. One such example is the rank transformed Wilks' Lambda proposed by Nath and Pavur (1985). This test statistic is outlined briefly in Section 4 and its performance is also evaluated in the Monte Carlo simulation.

2 The robust Wilks' Lambda statistic

Let $\boldsymbol{x}_{1k}, \boldsymbol{x}_{2k}, \ldots, \boldsymbol{x}_{n_k k}$ be n_k independent and identically distributed p-dimensional observations from a continuous p-variate distribution with distribution function $F_k(\boldsymbol{u})$ where $k = 1, 2, \ldots, g$ and the number of groups $g \geq 2$. If all g distributions are exactly the same but only their locations differ we have

$$F_k(\boldsymbol{u}) = F(\boldsymbol{u} - \boldsymbol{\mu}_k)$$

Then the hypothesis we want to test is that all F_k are identical, i.e.

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

against the alternative hypothesis

$$H_a : \boldsymbol{\mu}_i \neq \boldsymbol{\mu}_j \text{ for at least one } i \neq j$$

Under the classical assumptions that all groups arise from multivariate normal distributions, many test statistics are discussed in the literature, one of the most widely used being the likelihood ratio test. This test statistic is better known as Wilks' Lambda in MANOVA. The Wilks' Lambda statistic is the ratio of the within generalized dispersion to the total generalized dispersion. The within generalized dispersion is the determinant of the within-group sums of squares and cross-products matrix $\mathbf{W}$ and the total generalized dispersion is the determinant of the total sums of squares and cross-products matrix $\mathbf{T}$ (see e.g. Johnson and Wichern, 2002, chapter 6, p. 299). This statistic (1) where $det(\mathbf{A})$ means the determinant of $\mathbf{A}$,

$$\Lambda = \frac{det(\mathbf{W})}{det(\mathbf{T})} \tag{1}$$

takes values between zero and one.

In order to obtain a robust procedure with high breakdown point for inference about the means in the one-way MANOVA model we construct a robust version of the Wilks' Lambda statistic by replacing the classical estimators by the reweighted MCD estimators. The Minimum Covariance Determinant (MCD) estimator introduced by Rousseeuw (1985) looks for a subset of h observations with lowest determinant of the sample covariance matrix. The size of the subset h defines the so called trimming proportion and depending on the desired robustness it is chosen between half and the full sample size. The MCD location estimate $\mathbf{M}$ is defined as the mean of that subset and the MCD scatter estimate $\mathbf{C}$ is a multiple of its covariance matrix. The multiplication factor is selected so that $\mathbf{C}$ is consistent at the multivariate normal model and unbiased at small samples - see Pison *et al.* (2002). This estimator is not very efficient at normal models, especially if h is selected so that maximal breakdown point is achieved, but in spite of its low efficiency it is the mostly used robust estimator in practice, mainly because of the existing efficient algorithm for computation as well as the readily available implementations in most of the well known statistical software packages like *R, S-Plus, SAS* and *Matlab.* This was also the main reason for choosing the MCD estimator in the present work. To overcome the low efficiency of the MCD estimator, a reweighed version is used.

We start by finding initial estimates of the group means $\mathbf{m}_k^0$ and the common covariance matrix $\mathbf{C}_0$ based on the reweighted MCD estimate. There are several methods for estimating the common covariance matrix based on a high breakdown point estimator.

The easiest one is to obtain the estimates of the group means and group covariance matrices from the individual groups $(\mathbf{m}_k, \mathbf{C}_k), k = 1, \ldots, g$ and then pool them to yield the common covariance matrix

$$\mathbf{C} = \frac{\sum_{k=1}^{g} n_k \mathbf{C}_k}{\sum_{k=1}^{g} n_k - g} \tag{2}$$

This method, using MVE and MCD estimates, was proposed by Todorov *et al.* (1990) and Todorov *et al.* (1994) and was also used, based on the MVE estimator by Chork and Rousseeuw (1992). Croux and Dehon (2001) applied this procedure for robustifying linear discriminant analysis based on S estimates. A drawback of this method is that the same trimming proportions are applied to all groups which could lead to loss of efficiency if some groups are outlier free.

In the context of discriminant analysis, another method was proposed by He and Fung (2000) for S estimates and later adapted by Hubert and Van Driessen (2004) for MCD estimates. Instead of pooling the group covariance matrices, the observations are centered and pooled to obtain a single sample for which the covariance matrix is estimated. It starts with obtaining the individual group location estimates $\mathbf{t}_k, k = 1, \ldots, g$ as the reweighted MCD location estimates of each group. These group means are swept from the original observations to obtain the centered observations

$$\mathbf{Z} = \{\mathbf{z}_{ik}\}, \qquad \mathbf{z}_{ik} = \mathbf{x}_{ik} - \mathbf{t}_k \tag{3}$$

The common covariance matrix $\mathbf{C}$ is estimated as the reweighted MCD covariance matrix of the centered observations $\mathbf{Z}$. The location estimate $\boldsymbol{\delta}$ of $\mathbf{Z}$ is used to adjust the group means $\mathbf{m}_k$ and thus the final group means are

$$\mathbf{m}_k = \mathbf{t}_k + \boldsymbol{\delta} \tag{4}$$

This process could be iterated until convergence, but since the improvements from such iterations are negligible (see He and Fung, 2000; Hubert and Van Driessen, 2004) we are not going to use it.

The third approach is to modify the algorithm for high breakdown point estimation itself in order to accommodate the pooled sample. He and Fung (2000) modified Ruperts's SURREAL algorithm for S estimation in case of two groups. Hawkins and McLachlan (1997) defined the Minimum Within-group Covariance Determinant estimator (MWCD) which does not apply the same trimming proportion to each group. Unfortunately their estimator is based on the Feasible Solution Algorithm (see Hawkins and McLachlan, 1997, and the references therein), which is extremely time consuming compared to the FAST-MCD algorithm. Hubert and Van Driessen (2004) proposed a modification of this algorithm taking advantage of the FAST-MCD, but it is still necessary to compute the MCD for each individual group. A thorough investigation and comparison of these methods is worth doing, but in this work, for the sake of facilitating the computations we choose the method of pooling the observations.

Using the obtained estimates $\mathbf{m}_k^0$ and $\mathbf{C}_0$ we can calculate the initial robust distances (Rousseeuw and van Zomeren, 1991)

$$RD_{ik}^0 = \sqrt{(\mathbf{x}_{ik} - \mathbf{m}_k^0)^t \mathbf{C}_0^{-1} (\mathbf{x}_{ik} - \mathbf{m}_k^0)}. \tag{5}$$

With these initial robust distances we can define a weight for each observation $\mathbf{x}_{ik}$, $i = 1, \dots, n_k$ and $k = 1, \dots, g$ by setting the weight to 1 if the corresponding robust distance is less or equal to $\sqrt{\chi^2_{p,0.975}}$ and to 0 otherwise, i.e.

$$w_{ik} = \begin{cases} 1 & RD_{ik}^0 \leq \sqrt{\chi^2_{p,0.975}} \\ 0 & \text{otherwise.} \end{cases} \tag{6}$$

With these weights we can calculate the final reweighted estimates, namely the group means $\mathbf{m}_k$, the within-groups sum of squares and cross-products matrix $\mathbf{W}_R$, the between-groups sum of squares and cross-products matrix $\mathbf{B}_R$ and the total sum of squares and cross-products matrix $\mathbf{T}_R = \mathbf{W}_R + \mathbf{B}_R$ which are necessary for constructing the robust Wilks' Lambda statistic Λ_R as defined in equation (1).

$$
\begin{aligned}
\mathbf{m}_k &= (\sum_{i=1}^{n_k} w_{ik}\mathbf{x}_{ik})/\nu_k, \qquad \mathbf{m} = (\sum_{k=1}^{g} \nu_k \mathbf{m}_k)/\nu \\
\mathbf{W}_R &= \sum_{k=1}^{g}\sum_{i=1}^{n_k} w_{ik}(\mathbf{x}_{ik}-\mathbf{m}_k)(\mathbf{x}_{ik}-\mathbf{m}_k)^t \\
\mathbf{B}_R &= \sum_{k=1}^{g} \nu_k(\mathbf{m}_k-\mathbf{m})(\mathbf{m}_k-\mathbf{m})^t \\
\mathbf{T}_R &= \sum_{k=1}^{g}\sum_{i=1}^{n_k} w_{ik}(\mathbf{x}_{ik}-\mathbf{m})(\mathbf{x}_{ik}-\mathbf{m})^t = \mathbf{W}_R + \mathbf{B}_R
\end{aligned}
\tag{7}
$$

where ν_k are the sums of the weights within group k for $k = 1, \ldots, g$ and ν is the total sum of weights:

$$
\nu_k = \sum_{i=1}^{n_k} w_{ik} \quad \text{and} \quad \nu = \sum_{k=1}^{g} \nu_k
$$

Substituting these $\mathbf{W}_R$ and $\mathbf{T}_R$ into equation (1) we obtain a robust version of the test statistic Λ given by

$$
\Lambda_R = \frac{det(\mathbf{W}_R)}{det(\mathbf{T}_R)}. \tag{8}
$$

For computing the MCD and the related estimators the FAST-MCD algorithm of Rousseeuw and Van Driessen (1999) will be used as implemented in the R package *rrcov* - see Todorov (2007b).

3 Approximate distribution of Λ_R

The distribution of Λ is considered by Anderson (1958) as a ratio of two Wishart distributions but it is so complicated that except for some special cases it is hardly usable in practice. One of the most popular approximations is Bartlett's χ^2 approximation given by

$$
-(n-1-(p+g)/2)ln\Lambda \approx \chi^2_{p(g-1)} \tag{9}
$$

where $n = \sum_{k=1}^{g} n_k$. Analogously to this χ^2 approximation of the classical statistic we can assume for Λ_R the following approximation:

$$
L_R = -ln\Lambda_R \approx d\chi^2_q \tag{10}
$$

and then express the multiplication factor d and the degrees of freedom of the χ^2 distribution q through the expectation and variance of L_R

$$
\begin{aligned}
E[L_R] &= dq \\
Var[L_R] &= 2d^2q
\end{aligned} \tag{11}
$$

$$
\begin{aligned}
d &= E[L_R]\frac{1}{q} \\
q &= 2\frac{E[L_R]^2}{Var[L_R]}
\end{aligned}
$$

Since it is not possible to obtain the mean and variance of the robust Wilks' Lambda statistic Λ_R analytically, they will be determined by simulation. These values will be used to approximate the true distribution of Λ_R.

For a given dimension p, number of groups g and sample sizes of each group $n_k, k = 1, \dots, g$, samples $\mathbf{X}^{(j)} = \{\mathbf{x}_1, \dots, \mathbf{x}_n\}$ of size $n = \sum n_k$ from the standard normal distribution will be generated, i.e. $\mathbf{x}_i \sim N(\mathbf{0}, \mathbf{I}_p)$. For each sample the robust Wilks' Lambda statistic Λ_R based on the reweighted MCD will be calculated. After performing $m = 3000$ trials, mean and variance of Λ_R will be obtained as

$$
\begin{aligned}
ave(\Lambda_R) &= \frac{1}{m}\sum_{j=1}^{m}\Lambda_R^{(j)} \\
var(\Lambda_R) &= \frac{1}{m-1}\sum_{j=1}^{m}(\Lambda_R^{(j)} - ave(\Lambda_R))^2
\end{aligned}
$$

Substituting these values into equation (11) we can obtain estimates for the constants d and q which in turn will be used in equation (10) to obtain the approximate distribution of the robust Wilks' Lambda statistic Λ_R.

Now we will investigate the accuracy of this approximation. For several values of the dimension p, the number of groups g and the sample sizes of each group $n_k, k = 1, \dots, g$, $m = 3000$ samples from standard normal distribution will be generated and for each of them Λ_R will be calculated. The empirical distribution of these statistics will be compared to the approximate distribution given by equation (10) by QQ-plots, some of which are shown in Figure 1. The usual cutoff values of a test, the 95%, 97.5% and 99% quantiles are shown in these plots by vertical lines. It is seen from the plots that the approximation is very precise for large and small sample sizes, as well as for large dimensions ($p = 10$), for equal and unequal groups sizes.

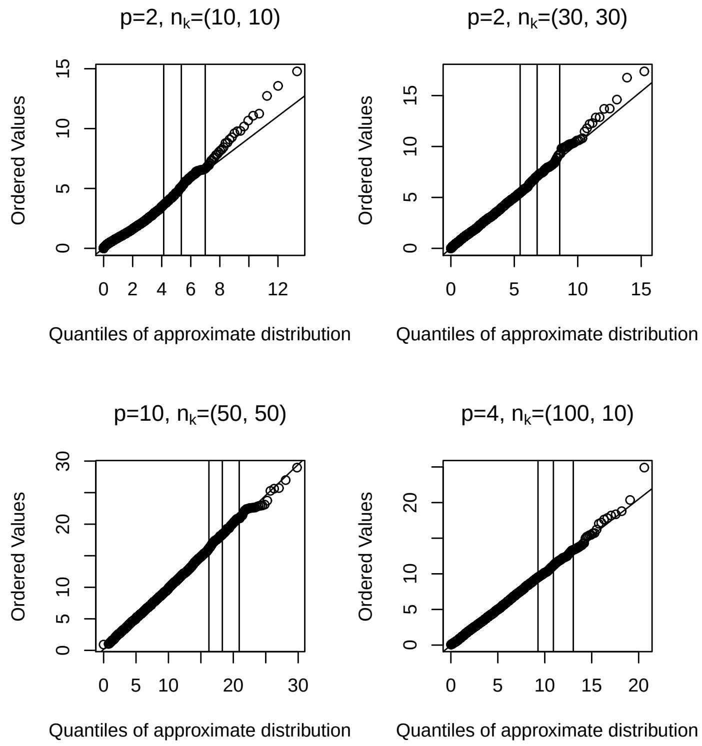

Figure 1. QQ-plots for the robust Wilks' Lambda statistics: $L_R = -ln(\Lambda_R)/d$ for two groups and several values for p and $n = \sum n_k$.

4 Rank transformed Wilks' Lambda statistic

An alternative proposal for the Wilks' Lambda statistic was made by Nath and Pavur (1985) which uses the ranks of the observations. The testing procedures obtained through the so called rank transformation are known to perform better then their classical counterparts especially when contamination is expected in the data. The Wilks' Lambda statistic is calculated on the ranks of the original data and is referred to as rank-transformed statistic. For performing the test the distribution of the statistic is approximated with that of the normal-theory statistic. Further details can be found in the above mentioned reference. In this work the rank transformed Wilks' Lambda statistic will be included in the simulation study and the results will be compared to the MCD based statistic.

5 Monte Carlo simulations

In this section a Monte Carlo study is undertaken to asses the performance of the proposed statistic. The assessment of the performance of any test statistics involves two measures - the attained significance level and the power of the test. Additionally we will investigate the behavior of the robust statistic in the presence of outliers and will compare the results to the classical as well as to the rank transformed Wilks' Lambda statistic.

5.1 Significance Levels

First we study the attained significance level (i.e., type I error rate or size) of the proposed robust test. We will consider several dimensions $p = \{2, 4, 6, 8, 10\}$, numbers of groups $g = \{2, 3\}$ and sample sizes $n_k, k = 1, \dots, g$. Equal as well as unequal group sizes are investigated. The sample sizes for two and three groups are selected as shown in Table 1. Only the cases where $p > 2 * n_k$ for all $k = 1, \dots, g$ were considered, since otherwise the MCD estimate is not computable.

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)
(20,20)	(20,20,20)
(30,30)	(30,30,30)
(50,50)	(50,50,50)
(100,100)	(100,100,100)
(200,200)	(20,20,10)
(20,10)	(30,30,10)
(30,10)	(50,50,20)
(50,10)	(100,50,20)

Under the null hypothesis H_0 in one-way MANOVA we assume that the observations come from identical multivariate distributions, i.e. $H_0 : \boldsymbol{\mu}_1 = \boldsymbol{\mu}_2 = \dots = \boldsymbol{\mu}_g$. Since the considered statistics are affine equivariant, without loss of generality we can assume each location vector $\boldsymbol{\mu}_k$ to be a null vector, i.e. $\boldsymbol{\mu}_k = (0, \dots, 0)^t$ and the covariance matrix to be $\mathbf{I}_p$. Thus we generate $n = \sum_{k=1}^{g} n_k$ p-variate vectors distributed as $N_p(\mathbf{0}, \mathbf{I}_p)$ and calculate the classical statistic Λ, the robust version based on MCD estimates Λ_R and the rank transformed Wilks' Lambda statistic Λ_{rank}. This is repeated $m = 3000$ times and the percentages of values of the test statistics above the appropriate critical value of the corresponding approximate distribution are taken as an estimate of the true significance level. The classical Wilks' Lambda and the rank transformed Wilks' Lambda are compared to the Bartlett approximation given by equation (9) while the MCD Wilks' Lambda is compared to the approximate distribution given in equation (10). Note that the approximate distribution of the robust test was simulated only once, by estimating the parameters d and q from equation (11) via average and variance of L_R (see above). This is a simplification, and a more precise procedure would be an approximation of the distribution for *each* of the 3000 simulated samples. This, however, would not only be computationally very expensive, it would also hardly make a difference to the simplified procedure.

The true significance levels α are taken to be 0.10, 0.05 and 0.01 which together with the number of replications $m = 3000$ yields the two standard deviation intervals around the nominal levels as follows: (0.089, 0.111), (0.042, 0.058), (0.006, 0.014) respectively - here the standard error is $\sqrt{(\alpha(1-\alpha)/3000)}$.

In Table 2 the results for two groups are shown. It is clearly seen that the difference between the actual cutoff and the nominal value is very small, i.e. the approximations are capable to keep the significance levels across all investigated combinations of dimension p and sample sizes. Although the results of the robust test are slightly worse than these of the classical test and the rank transformed test they are satisfactory, having in mind the gained advantage in case of non-normal data. The results for three groups are similar. A complete set of results is available from the authors upon request.

Instead of showing further tables, we will make use of the *P value plots* proposed by Davidson and McKinnon (1998), which give a more complete picture of how the test statistics follow the approximate distribution under the null in the simulated samples. Figure 2 shows plots of the empirical distribution function of the p-values for the classical and robust Wilks' Lambda statistics in the case of two groups, for several values of dimension p and sample size n_k. The corresponding results for three groups are similar. The rank transformed statistic Λ_{rank} is not shown since it closely follows the classical statistic. The most interesting part of the P value plot is the region where the size ranges from zero to 0.2 since in practice a significance level above 20% is never used. Therefore we limited both axis to p-value ≤ 0.2. We expect that the results in the P value plot follow the 45° line since the p-values are distributed uniformly on (0,1) if the distribution of the test statistic is correct. Deviations from this line suggest that the empirical distribution differs from the one used to establish the critical values. It is seen that only for very small sample sizes the robust Wilks' Lambda is considerably below the 45° line. Otherwise the robust test works very well and follows closely the 45° line.

Table 2
Significance levels of the test statistics Λ, Λ_R and Λ_{rank} for multivariate normal distributions in the case of two groups for several values of the dimension p and the sample size $n = n_1 + n_2$ ($m = 3000$ Monte Carlo replications, true significance level α=0.1, 0.05 and 0.01).

			$\alpha = 0.1$			$\alpha = 0.05$			$\alpha = 0.01$		
p	n_1	n_2	Λ	Λ_R	Λ_{rank}	Λ	Λ_R	Λ_{rank}	Λ	Λ_R	Λ_{rank}
2	10	10	0.097	0.073	0.101	0.050	0.039	0.051	0.010	0.008	0.013
2	20	20	0.099	0.086	0.103	0.053	0.043	0.053	0.009	0.015	0.010
2	30	30	0.100	0.099	0.099	0.050	0.050	0.049	0.007	0.012	0.009
2	50	50	0.101	0.088	0.100	0.046	0.042	0.050	0.006	0.007	0.009
2	100	100	0.093	0.087	0.092	0.043	0.046	0.046	0.010	0.008	0.009
2	200	200	0.109	0.116	0.111	0.057	0.059	0.058	0.009	0.013	0.013
2	20	10	0.095	0.079	0.097	0.047	0.041	0.049	0.010	0.011	0.011
2	30	10	0.097	0.080	0.098	0.050	0.042	0.046	0.013	0.008	0.010
2	50	20	0.102	0.100	0.101	0.055	0.051	0.058	0.013	0.011	0.013
2	100	10	0.099	0.090	0.097	0.049	0.046	0.049	0.010	0.010	0.009
4	10	10	0.098	0.109	0.105	0.049	0.062	0.053	0.009	0.012	0.011
4	20	20	0.094	0.083	0.101	0.046	0.041	0.050	0.009	0.013	0.011
4	30	30	0.104	0.084	0.104	0.051	0.043	0.054	0.011	0.011	0.012
4	50	50	0.107	0.091	0.105	0.058	0.048	0.060	0.011	0.010	0.011
4	100	100	0.105	0.118	0.106	0.051	0.063	0.057	0.012	0.013	0.009
4	200	200	0.088	0.088	0.091	0.042	0.047	0.043	0.012	0.012	0.014
4	20	10	0.097	0.084	0.096	0.048	0.044	0.053	0.009	0.018	0.011
4	30	10	0.088	0.066	0.094	0.046	0.034	0.047	0.010	0.011	0.009
4	50	20	0.098	0.090	0.095	0.053	0.047	0.050	0.011	0.013	0.013
4	100	10	0.100	0.106	0.098	0.050	0.055	0.046	0.009	0.011	0.009
6	20	20	0.098	0.082	0.102	0.047	0.047	0.050	0.009	0.016	0.009
6	30	30	0.099	0.090	0.094	0.054	0.048	0.058	0.013	0.014	0.014
6	50	50	0.106	0.095	0.106	0.052	0.047	0.053	0.013	0.011	0.011
6	100	100	0.109	0.103	0.105	0.056	0.051	0.058	0.011	0.011	0.009
6	200	200	0.108	0.100	0.105	0.051	0.046	0.052	0.010	0.008	0.011
6	50	20	0.109	0.100	0.116	0.052	0.050	0.053	0.010	0.013	0.011
8	20	20	0.102	0.089	0.100	0.051	0.051	0.049	0.012	0.014	0.012
8	30	30	0.101	0.087	0.102	0.049	0.048	0.060	0.011	0.013	0.012
8	50	50	0.112	0.095	0.105	0.055	0.049	0.058	0.012	0.011	0.013
8	100	100	0.096	0.097	0.096	0.049	0.047	0.048	0.011	0.008	0.012
8	200	200	0.092	0.093	0.093	0.047	0.049	0.048	0.007	0.009	0.008
8	50	20	0.095	0.105	0.096	0.047	0.058	0.046	0.009	0.019	0.010
10	30	30	0.100	0.085	0.109	0.049	0.049	0.052	0.012	0.014	0.011
10	50	50	0.097	0.097	0.096	0.052	0.048	0.054	0.010	0.010	0.012
10	100	100	0.098	0.101	0.104	0.050	0.048	0.052	0.013	0.013	0.011
10	200	200	0.112	0.111	0.110	0.055	0.057	0.058	0.015	0.016	0.016

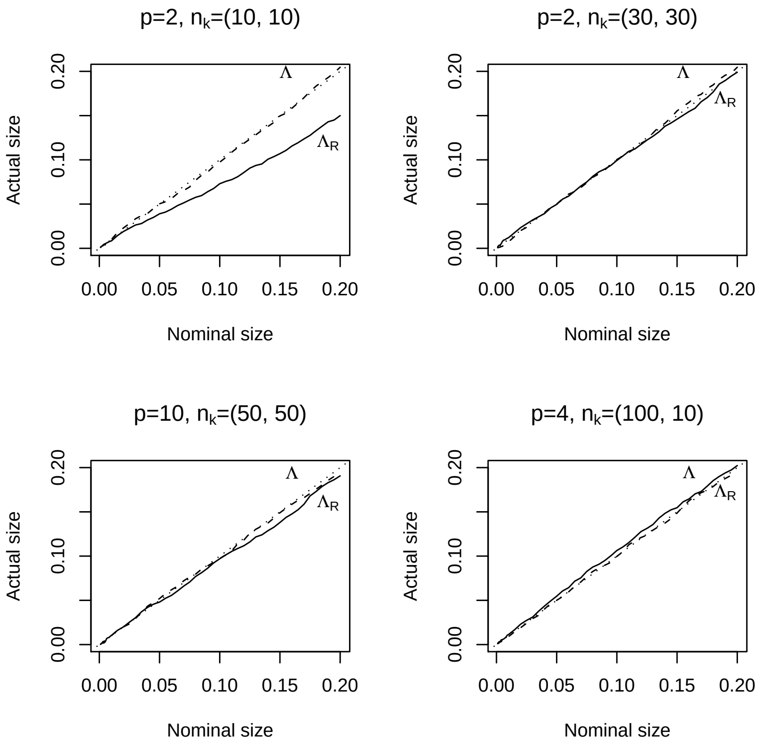

Figure 2. P value plots for the Wilks' Lambda statistic Λ (dashed line) and the robust Wilks' Lambda statistic Λ_R (solid line) for two groups and several values for p and $n = \sum n_k$. The $45°$ line is given too, represented by a dotted line. The results for Λ_{rank} closely follow those of Λ and therefore they are not shown.

5.2 Power comparisons

In order to asses the power of the robust Wilks' Lambda statistics we will generate data under an alternative hypothesis (H_a : not all $\boldsymbol{\mu}_k, k = 1, \ldots, g$ are equal) and will examine the frequency of incorrectly failing to reject H_0 (i.e. the frequency of type II errors). The same combinations of dimensions p, number of groups g and sample sizes $n_k, k = 1, \ldots, g$ as in the experiments for studying the significance levels will be used. There are infinitely many possibilities for selecting H_a but for the purpose of the study we will use the following fixed alternatives: all groups π_k, $k = 1, \ldots, g$, come from multivariate normal distribution with the same spherical covariance matrix $\mathbf{I}_p$; the mean of the first group is the origin, the mean of the second group is at distance $d = 1$ along the first coordinate, the mean of the third group is at distance $d = 1$ along the second coordinate and so on. With this simple model the number of groups can be at most $p + 1$ but this is not a restriction since we will consider only two and three groups. More precisely, the data sets are generated from the following p-dimensional normal distributions, where each group π_k, $k = 1, \ldots, g$, has a different mean $\boldsymbol{\mu}_k$ and all of them have the same covariance matrix $\mathbf{I}_p$,

$$\pi_k \sim N_p(\boldsymbol{\mu}_k, \mathbf{I}_p), \quad k = 1, \ldots, g, \tag{12}$$

with

$$\begin{aligned}
\boldsymbol{\mu}_1 &= (0, 0, \ldots, 0)^t \\
\boldsymbol{\mu}_2 &= (d, 0, \ldots, 0)^t \\
\boldsymbol{\mu}_3 &= (0, d, 0, \ldots, 0)^t \\
&\ldots \\
\boldsymbol{\mu}_g &= (0, \ldots, 0, d, 0, \ldots, 0)^t
\end{aligned}$$

The classical and the robust test statistics are computed and the rejection frequency (out of $m = 3000$ runs) where the statistic exceeds its appropriate critical value is the estimate of the power for the specific configuration.

The power of the two statistics can be visually compared by simulating size–power curves under fixed alternatives, as proposed by Davidson and McKinnon (1998). Constructing size-power plots does not require knowledge of the distribution of the test statistic. For other recent applications of the size-power see Siani and de Peretti (2006) and Gelper and Croux (2007). The size-power curves are simulated in the following way: (i) First $m = 3000$ data sets under the null hypothesis are generated. For each of them the test statistics are computed and the obtained values are sorted in increasing order. The j^{th} value of this ordered sequence is denoted by θ_j. If the critical value is chosen as θ_j then the quantity $s_j = (m - j)/(m + 1)$ equals the size of the test; (ii) After that $m = 3000$ data sets are generated under the fixed alternative hypothesis and for each of them the test statistics are computed. For a certain critical value θ_j, the power of the test f_j is estimated by the fraction of test statistics that exceed θ_j; (iii) The pairs $(s_j, f_j), j = 1, \ldots, m$ representing the power vs. the size of the test are plotted as size–power curves.

The size–power curve should lie above the 45° line, the larger the distance between the curve and the 45° line the better. The most interesting part of the size–power curve is the region where the size ranges from zero to 0.2 since in practice a significance level above 20% is never used. In Figure 3 the size-power curves for several values of the dimension p and the sample sizes n_k in case of two groups are shown. The results for three groups are similar. It is clearly seen that in all of the investigated combinations of dimensions p and sample sizes both curves are far above the 45° line with the line of the robust statistic being slightly below the classical one. Thus the loss of power for the robust statistic is acceptable throughout the investigated range of dimensions and sample sizes. The size–power curves for the rank transformed Wilks' Lambda statistic Λ_{rank} are similar to those of the classical test statistic and therefore not shown.

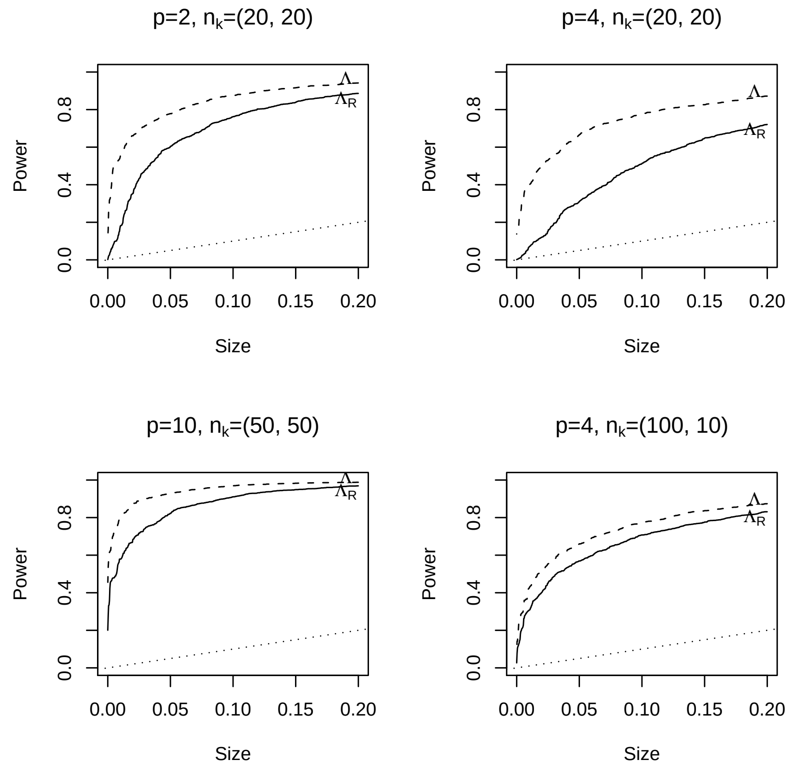

Figure 3. Size-power curves for the Wilks' Lambda statistic Λ (dashed line) and the robust Wilks' Lambda statistic Λ_R (solid line) for two groups and several values for p and $n = \sum n_k$. The 45° line is represented by a dotted line. The results for Λ_{rank} closely follow those of Λ and therefore they are not shown.

5.3 Robustness comparisons

Now we will investigate the robustness of the one-way MANOVA hypothesis test based on the proposed robust version of the Wilks' Lambda statistic Λ_R. For this purpose we will generate data sets under the null hypothesis $H_0 : \boldsymbol{\mu}_1 = \boldsymbol{\mu}_2 = \ldots = \boldsymbol{\mu}_g$ and will contaminate them by adding outliers. More precisely the data will be generated from the following contamination model:

$$\pi_k \sim (1-\varepsilon)N_p(\mathbf{0}, \mathbf{I}_p) + \varepsilon N_p(\hat{\boldsymbol{\mu}}_k, 0.25^2\mathbf{I}_p), \quad k = 1, \ldots, g, \tag{13}$$

$$\hat{\boldsymbol{\mu}}_k = (\nu Q_p, \ldots, \nu Q_p)^t$$
$$Q_p = \sqrt{\chi^2_{p;0.001}/p},$$

where $\varepsilon = 0.1$ and $\nu = 5$. By adding νQ_p to each component of the outliers we guarantee a comparable shift for different dimensions p, see Rocke and Woodruff (1996). The same combinations of dimensions p, numbers of groups g and sample sizes $n_k, k = 1, \ldots, g$, as in the experiments for studying the significance levels will be used.

Again we generate $n = \sum_{k=1}^{g} n_k$ p-variate vectors and calculate the classical statistic Λ, the robust version based on MCD estimates Λ_R and the rank transformed Wilks' Lambda statistic Λ_{rank}. This is repeated $m = 3000$ times and the percentages of values of the test statistics above the appropriate critical value of the corresponding approximate distribution are taken as an estimate of the true significance level. We present the results for three groups with P value plots in Figure 4. Similar results are obtained for two groups.

The difference between the actual cutoff based on Λ_R and the nominal value remains acceptably small for the different combinations of dimension p and sample size $n_k, k = 1, \ldots, g$. Furthermore this difference is much smaller compared to the classical Wilks' Lambda statistic Λ and the rank transformed statistic. Note that a deviation below the 45 line, as often observed for the robust test, refers to a conservative test.

6 Example

We will illustrate the application of the proposed robust statistic with the Oslo transect data (see Reimann *et al.*, 2007, and the references therein). Samples of different plant species were collected along a 120 km transect running through the city of Oslo, Norway, and the concentrations of 25 chemical elements for the sample materials are reported. The factors that influenced the observed element concentrations in the sample materials were investigated. For our example we will consider only the lithology as a factor. This factor has four levels which are listed in Table 3. The last column shows the number of objects in each group.

We select the variables P, K, Zn and Cu that represent elements from the group of the nutrients and expect that the lithology strongly influences the take in of the plants when compared to the effect of the plant species themselves, i.e. we expect that the multivariate group means are significantly different. After removing the observations with missing values we remain with a data matrix of $n = 332$ rows and $p = 4$ columns. Since geochemical data are usually right skewed we log-transform the variables. The left panel of Figure 5 shows the scatter plot matrix of the log–transformed data together with histograms of each variable and the right panel presents the box plots for the different groups. In Figure 6 the scatter plot matrices of each group are presented sepa-

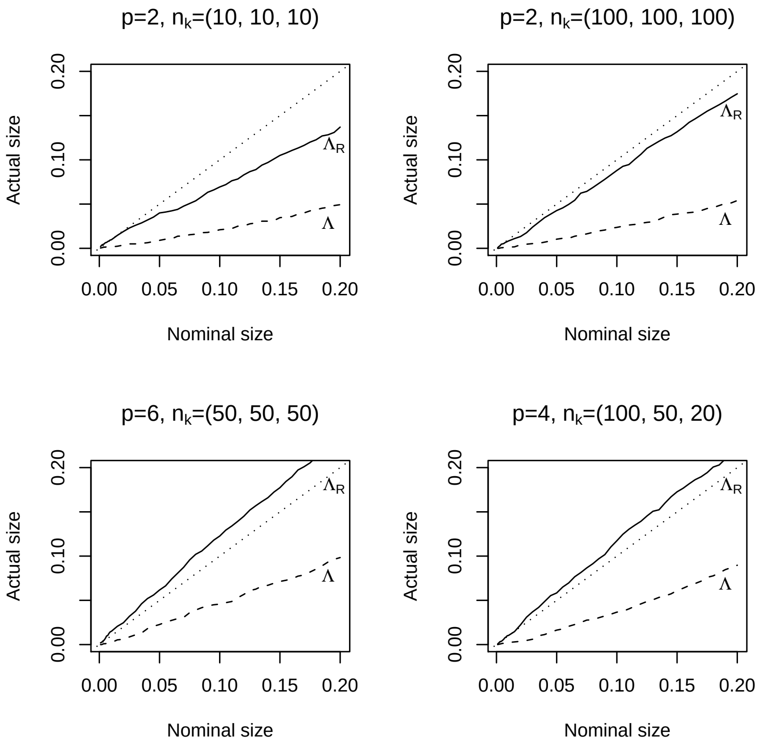

Figure 4. Robustness comparisons - P value plots for Wilks' Lambda statistic Λ (dashed line), robust Wilks' Lambda statistic Λ_R (solid line) and rank transformed Wilks' Lambda statistic Λ_{rank} (long-dashed line) for three groups and several values for p and $n = \sum n_k$. The 45° line is represented by a dotted line.

Table 3
Oslo transect data: Names of the lithology groups. The last column shows the number of objects in each class.

		Lithological group	#
1	CAMSED	Cambro-Silurian sedimentary rocks	98
2	GNEISS_O	Precambrian gneisses - Oslo	89
3	GNEISS_R	Precambrian gneisses - Randsfjord	32
4	MAGM	Magmatic rocks of the Oslo Rift	113

rately. The classical and robust 97.5% tolerance ellipses clearly show that the data are not normally distributed and that outliers are present.

Let us denote the means of the four groups by $\boldsymbol{\mu}_1, \boldsymbol{\mu}_2, \boldsymbol{\mu}_3$ and $\boldsymbol{\mu}_4$ and perform a one-way MANOVA, testing the hypothesis $H_0 : \boldsymbol{\mu}_1 = \boldsymbol{\mu}_2 = \boldsymbol{\mu}_3 = \boldsymbol{\mu}_4$. The classical Wilks' Lambda statistic for this data set yields $\Lambda = 0.9755$ which corresponds to a p-value of 0.7757. This suggests that the hypothesis of equal means cannot be rejected at the 10% significance level. On the other hand

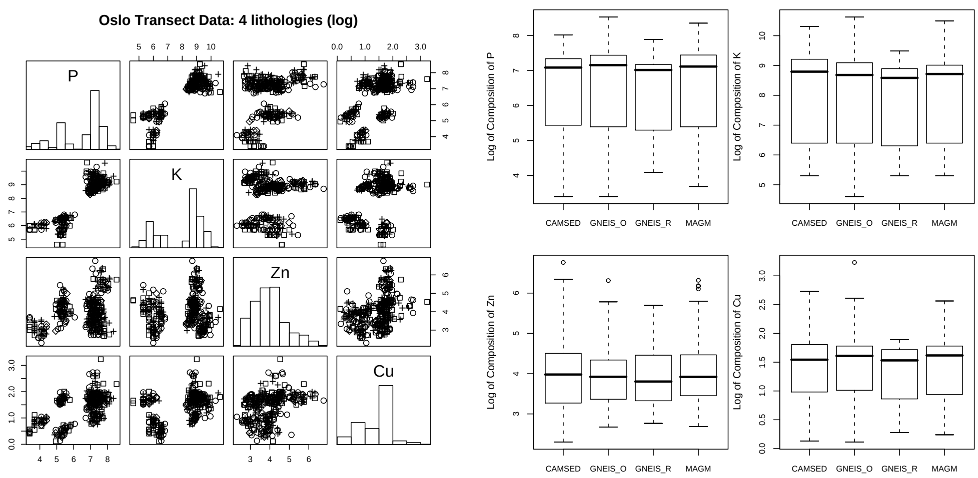

Figure 5. Scatter plot matrix of the log–transformed Oslo Transect Data (left panel) and box plots of the same data (right panel).

the robust Wilks' Lambda statistic yields $\lambda_R = 0.8947$ which corresponds to a p-value of 0.0025 and we can reject the null hypothesis even at the 1% significance level.

This example as well as all computations presented in Section 5 were performed with the package `rrcov` (Todorov, 2007b) in the statistical environment R (R Development Core Team, 2007). The data set from Reimann *et al.* (2007) discussed here (as well as many other data) are provided as example data sets in `rrcov` and can be loaded by the `data` command. We can log-transform the (numerical part of the) data set, choose the desired variables and groups and perform the test using the formula interface of the function `Wilks.test()`. The default method is the classical Wilks' Lambda (`method="c"`).

```
        One-way MANOVA (Bartlett Chi2)
data:  OsloTransect
Wilks' Lambda = 0.9755, Chi2-Value = 8.12, df = 12.00,
    p-value = 0.7757
sample estimates:
               P        K       Zn        Cu
CAMSED  6.456762 8.093613 3.948710 1.421572
GNEIS_O 6.540716 8.003046 3.982276 1.446860
GNEIS_R 6.386443 7.863500 3.934097 1.317420
MAGM    6.517899 7.982188 4.043223 1.401263
```

To perform the robust test we need to specify `method="mcd"`. It will take some time while performing the simulations for finding the multiplication factor and the degrees of freedom for the approximate distribution. These parameters will be returned in the resulting object too and can be reused for analysis of data

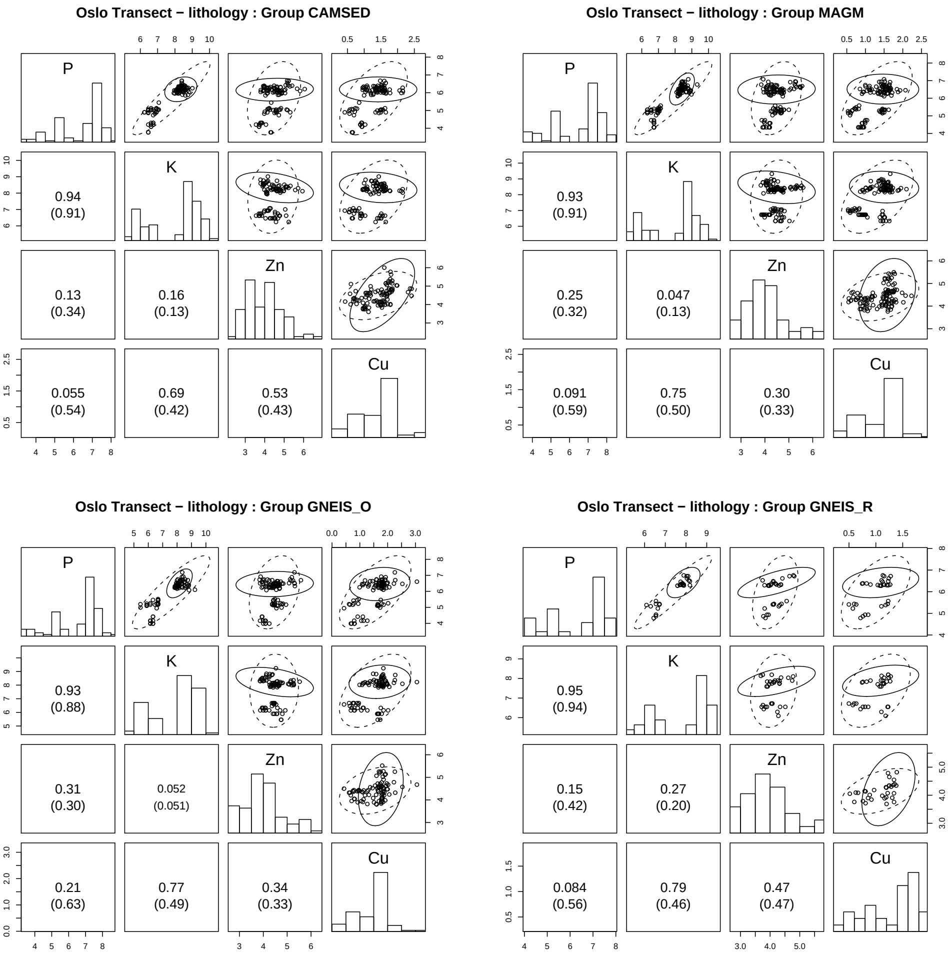

Figure 6. Log-transformed Oslo Transect Data: Scatter plot matrices of each group separately with classical (dashed line) and robust (solid line) 97.5% tolerance ellipses in the upper triangle, classical (in parentheses) and robust correlations in the lower triangle and histograms on the diagonal.

with the same dimension, number of groups and number of observations in each group.

```
        Robust One-way MANOVA (Bartlett Chi2)
data:  OsloTransect
Wilks' Lambda = 0.8947, Chi2-Value = 30.019, df = 11.791,
    p-value = 0.002473
sample estimates:
               P        K       Zn       Cu
CAMSED  7.246640 9.109288 4.092121 1.628208
GNEIS_O 7.400005 9.012718 4.067534 1.672883
GNEIS_R 7.252277 8.875761 4.056851 1.506315
MAGM    7.394440 9.014556 4.153965 1.626010
```

7 Conclusions

A robust version of Wilks' Lambda statistic was introduced by replacing the classical estimates for mean and covariance with robust counterparts. The MCD estimator was chosen for this purpose, and an approximate distribution of the robust test statistic was derived with simulations. In further Monte Carlo studies the significance level and the power of the new test was compared with the classical and the rank transformed Wilks' Lambda test. Various different situations were investigated by changing the dimension, the group sizes, and the number of groups. Only a selection of the results is shown in the paper which, however, are typical representative outcomes. These allow to conclude that the significance level of the robust test is reasonably precise in case of normal distribution but also in case of deviations from normal distribution. In the latter case it turned out that the actual size of the robust test is in general much closer to the nominal size than the classical and the rank transformed Wilk's Lambda test. Furthermore, as indicated by the size-power curves, the robust test does not loose much power compared to the classical and to Wilk's Lambda test.

The new test has been implemented as the function `Wilks.test` in the R package `rrcov`, with the options classical, rank transformed and robust test statistic. Moreover, also the Hotelling T^2 test has been implemented as function `T2.test` in the same package, and it can be used for comparison with two groups.

Acknowledgements

We are thankful to the anonymous referees for their careful review and constructive comments that substantially improved the article.

References

Anderson T. (1958) *An Introduction to Multivariate Statistical Analysis*, John Wiley.

Chork C. and Rousseeuw P.J. (1992) Integrating a high-breakdown option into discriminant analysis in exploration geochemistry, *Journal of Geochemical Exploration*, 43, 191–203.

Croux C. and Dehon C. (2001) Robust linear discriminant analysis using S-estimators, *The Canadian Journal of Statistics*, 29, 473–492.

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

Gelper S. and Croux C. (2007) Multivariate out-of-sample tests for Granger causality, *Computatational Statistics and Data Analysis*, 51, 3319 3329.

Hawkins D.M. and McLachlan G. (1997) High-breakdown linear discriminant analysis, *Journal of the American Statistical Association*, 92, 136–143.

He X. and Fung W. (2000) High- breakdown estimation for multiple populations with applications to discriminant analysis, *Journal of Multivariate Analysis*, 72, 151–162.

Hubert M. and Van Driessen K. (2004) Fast and robust discriminant analysis, *Computational Statistics and Data Analysis*, 45, 301–320.

Hubert M., Rousseeuw P.J. and Van Aelst S. (2008) High-breakdown robust multivariate methods, *Statistical Science*, 23, 92–119.

Johnson R.A. and Wichern D.W. (2002) *Applied Multivariate Statistical Analysis*, Prentice Hall, International, fifth edition.

Nath R. and Pavur R. (1985) A new statistic in the one way multivariate analysis of variance, *Computatational Statistics and Data Analysis*, 2, 297–315.

Pison G., Van Aelst S. and Willems G. (2002) Small sample corrections for LTS and MCD, *Metrika*, 55, 111–123.

R Development Core Team (2007) *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria, ISBN 3-900051-07-0.

Reimann C., Arnoldussen A., Boyd R., Finne T.E., Koller F., Nordgullen O. and Englmair P. (2007) Element contents in leaves of four plant species (birch, mountain ash, fern and spruce) along anthropogenic and geogenic concentration gradients, *Science of the Total Environment*, 377, 416–433.

Rocke D.M. and Woodruff D.L. (1996) Identification of outliers in multivariate data, *Journal of the American Statistical Association*, 91, 1047–1061.

Rousseeuw P.J. (1985) Multivariate estimation with high-breakdown point, in: *Mathematical Statistics and Applications Vol. B*, Grossmann W., Pflug G., Vincze I. and Wertz W., eds., Reidel Publishing, Dordrecht, 283–297.

Rousseeuw P.J. and Van Driessen K. (1999) A fast algorithm for the Minimum Covariance Determinant estimator, *Technometrics*, 41, 212–223.

Rousseeuw P.J. and van Zomeren B.C. (1991) Robust distances: Simulation and cutoff values, in: *Directions in Robust Statistics and Diagnostics, Part II*, W.Stahel and S.Weisberg, eds., Springer Verlag, New York.

Siani C. and de Peretti C. (2006) Analysing the performance of bootstrap neural tests for conditional heteroskedasticity in ARCH-M models, *Computatational Statistics and Data Analysis*, 51, 2442–2460.

Todorov V. (2007a) Robust selection of variables in linear discriminant analysis, *Statistical Methods and Applications*, 15, 395–407, doi:10.1007/s10260-006-0032-6.

Todorov V. (2007b) *rrcov: Scalable Robust Estimators with High-Breakdown Point*, R package version 0.5-02.

Todorov V., Neykov N. and Neytchev P. (1990) Robust selection of variables in the discriminant analysis based on MVE and MCD estimators, in: *Proceedings in Computational statistics, COMPSTAT'90*, Physica Verlag, Heidelberg.

Todorov V., Neykov N. and Neytchev P. (1994) Robust two-group discrimination by bounded influence regression, *Computational Statistics and Data Analysis*, 17, 289–302.

Willems G., Pison G., Rousseeuw P.J and Van Aelst S. (2002) A robust Hotelling test, *Metrika*, 55, 125–138.