

A Robust Hotelling Test

G. Willems, G. Pison, P.J. Rousseeuw, and S. Van Aelst

Department of Mathematics and Computer Science, University of Antwerp (UIA),
Universiteitsplein 1, 2610 Wilrijk, Belgium.

Received: date / Revised version: 4 october 2001

Abstract Hotelling's T^2 statistic is an important tool for inference about the center of a multivariate normal population. However, hypothesis tests and confidence intervals based on this statistic can be adversely affected by outliers. Therefore, we construct an alternative inference technique based on a statistic which uses the highly robust MCD estimator [9] instead of the classical mean and covariance matrix. Recently, a fast algorithm was constructed to compute the MCD [10]. In our test statistic we use the re-weighted MCD, which has a higher efficiency. The distribution of this new statistic differs from the classical one. Therefore, the key problem is to find a good approximation for this distribution. Similarly to the classical T^2 distribution, we obtain a multiple of a certain F-distribution. A Monte Carlo study shows that this distribution is an accurate approximation of the true distribution. Finally, the power and the robustness of the one-sample test based on our robust T^2 are investigated through simulation.

1 Introduction

Hotelling's T^2 statistic is the standard tool for inference about the center of a multivariate normal distribution. An example is the one-sample T^2 hypothesis test for a sample $X_n = \{\mathbf{x}_1, \dots, \mathbf{x}_n\}$ from $N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ where $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ are unknown. The hypothesis $H_0 : \boldsymbol{\mu} = \boldsymbol{\mu}_0$ is rejected at the level α if

$$T^2 := n(\bar{\mathbf{x}} - \boldsymbol{\mu}_0)' \mathbf{S}^{-1} (\bar{\mathbf{x}} - \boldsymbol{\mu}_0) > \frac{(n-1)p}{(n-p)} F_{p, n-p, 1-\alpha} \quad (1)$$

where $\bar{\mathbf{x}} = \frac{1}{n} \sum_{i=1}^n \mathbf{x}_i$ and $\mathbf{S} = \frac{1}{n-1} \sum_{i=1}^n (\mathbf{x}_i - \bar{\mathbf{x}})(\mathbf{x}_i - \bar{\mathbf{x}})'$ are the mean and covariance of the sample (see e.g. [4, page 227]). Other applications of

the T^2 statistic include simultaneous confidence intervals and two-sample hypothesis tests.

However, these tests and confidence intervals are based on the classical mean and covariance, hence the results can be heavily influenced by outliers. Therefore, we propose to use robust estimators of location and scatter instead of the classical mean and covariance in the expression for T^2 given by (1). In this paper we propose to use the Minimum Covariance Determinant (MCD) estimator of Rousseeuw [9] which is a highly robust estimator of location and scatter. The MCD estimator will be summarized in Section 2.

The distribution of the test statistic based on MCD differs from the classical one. Therefore, the key problem is to find a good approximation for this distribution. In Section 3 we construct an approximate distribution using ideas that are similar to the construction of the classical T^2 distribution. Based on a Monte Carlo study, in Section 4 we construct functions which yield values for the constants that are necessary to obtain the approximate distribution of the robust T^2 statistic. The resulting method gives an accurate approximation as will be shown in Section 5. In Section 6 we compare this approximation for the distribution of the robust T^2 statistic with two approximations that are more intuitively appealing. We show that the latter approximations are much worse than the approximation obtained in Section 4. The power of the resulting one-sample test is investigated in Section 7, and we study its robustness in Section 8 by simulations with contaminated data. Section 9 gives an example of robust inference, and Section 10 concludes.

2 The Minimum Covariance Determinant Estimator

Given n data points in \mathbb{R}^p the MCD is determined by the subset of size $h = \lfloor \gamma n \rfloor$ (where $0.5 \leq \gamma \leq 1$) whose covariance matrix has the smallest determinant. The MCD location estimate T is defined as the mean of that subset, and the MCD scatter estimate C is a multiple of its covariance matrix. The multiplication factor consists of a consistency factor c_γ and a finite-sample correction factor. The consistency factor, given in [2], makes the MCD scatter estimator Fisher-consistent at the normal model and equals $c_\gamma = \gamma / F_{\chi_{p+2}^2}(q_\gamma)$ where $q_\gamma = \chi_{p,\gamma}^2$. The correction factor makes the MCD unbiased at small samples (see [8]).

To increase the efficiency of the MCD we compute the reweighted MCD. The reweighted estimates are the weighted mean

$$T^1 = \frac{\sum_{i=1}^n w_i \mathbf{x}_i}{\sum_{i=1}^n w_i} \quad (2)$$

and weighted covariance

$$C^1 = c_\delta d_{n,p} \frac{\sum_{i=1}^n w_i (\mathbf{x}_i - T^1)(\mathbf{x}_i - T^1)'}{\sum_{i=1}^n w_i} \quad (3)$$

with weights based on the robust distances of the observations. Based on the initial MCD estimates (T, C) the robust distance [11] of an observation \mathbf{x}_i is defined as $d_i(T, C) := \sqrt{(\mathbf{x}_i - T)'C^{-1}(\mathbf{x}_i - T)}$. Observations with robust distance $d_i(T, C)$ below the cutoff value q obtain weight 1, while the other observations obtain weight 0. The factor c_δ equals $c_\delta = (1 - \delta)/F_{\chi_{p+2}^2}(q_\delta)$ where $q_\delta = \chi_{p, 1-\delta}^2$, and makes the reweighted MCD consistent at the normal model. Here, δ is the fraction of observations that obtained weight 0 and have a robust distance smaller than $2q$ (observations that have a robust distance larger than $2q$ are considered to be far outliers). The constant $d_{n,p}$ is a finite-sample correction factor given in [8].

The MCD is an affine equivariant estimator of location and scatter and has a positive breakdown value which depends on γ . The choice $\gamma = 0.5$ yields the maximal breakdown value of 50%. We prefer to use $\gamma = 0.75$ which gives a better efficiency and a breakdown value of 25%, which is more realistic. Moreover, the MCD has a bounded influence function [2] and is asymptotically normal [1]. The reweighted MCD estimators [5, 2, 6] inherit the breakdown value, bounded influence and asymptotic normality of the initial MCD estimators while achieving a higher efficiency. To compute the reweighted MCD we use the FAST-MCD algorithm [10] which makes the MCD available for routine use, even for large data sets.

3 The robust T^2 statistic

To obtain robust inference techniques for the mean of a multivariate normal distribution we construct a new statistic by replacing the classical estimators in Hotelling's T^2 by the reweighted MCD estimators. For a sample of size n from $N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, the distribution of the classical Hotelling T^2 is given by

$$T^2 = n(\bar{\mathbf{x}} - \boldsymbol{\mu})' \mathbf{S}^{-1} (\bar{\mathbf{x}} - \boldsymbol{\mu}) \sim \frac{(n-1)p}{(n-p)} F_{p, n-p}.$$

This distribution follows from three properties of $\bar{\mathbf{x}}$ and \mathbf{S} (see [7]):

- $\bar{\mathbf{x}} \sim N_p(\boldsymbol{\mu}, \frac{1}{n}\boldsymbol{\Sigma})$
- $(n-1)\mathbf{S} \sim W_p(\boldsymbol{\Sigma}, n-1)$ (the Wishart distribution)
- $\bar{\mathbf{x}}$ and \mathbf{S} are independent.

Similarly, based on the reweighted MCD estimates (T^1, C^1) we now define the robust test statistic

$$T_R^2 := n(T^1 - \boldsymbol{\mu})'(C^1)^{-1}(T^1 - \boldsymbol{\mu}). \quad (4)$$

The finite-sample distributions of the reweighted MCD location and scatter are unknown, but it turns out that they have properties similar to the classical estimators. That is, for a sample of size n from $N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ the reweighted MCD estimators approximately satisfy the properties:

1. There exists some κ such that $T^1 \sim N_p(\boldsymbol{\mu}, \kappa \frac{1}{n}\boldsymbol{\Sigma})$.

2. There exist m and c such that $mc^{-1}C^1 \sim W_p(\Sigma, m)$ and $E[C^1] = c\Sigma$.
3. T^1 and C^1 are independent.

Property 2 was formulated in [3]. These properties allow us to obtain an approximate F -distribution for T_R^2 , analogous to the F -distribution of the classical T^2 . From [7, Theorem 3.5.2] we obtain that

$$T_R^2 \approx \kappa c^{-1} \frac{mp}{m-p+1} F_{p, m-p+1} \quad (5)$$

Instead of determining values for the constants κ , c and m (see also [3]) we will use a more direct approach. First we rewrite (5) as

$$T_R^2 \approx dF_{p,q} \quad (6)$$

which has the advantage that only two constants, d and q , have to be determined. The multiplication factor d and q , the degrees of freedom for the denominator of the F -distribution, will be obtained by matching the mean and variance of the distribution given by (6) such that we obtain a very close approximation to the exact distribution of T_R^2 . Since the MCD is affine equivariant, it follows that the T_R^2 statistic is affine invariant. Therefore it suffices to determine values of d and q for samples from the standard Gaussian distribution $N_p(\mathbf{0}, \mathbf{I}_p)$.

From the definition of the F -distribution it follows that

$$E[T_R^2] = d \frac{q}{q-2}$$

$$\text{Var}[T_R^2] = d^2 \frac{2q^2(p+q-2)}{p(q-4)(q-2)^2}$$

By rewriting the two previous equations we obtain the following expressions for the constants d and q :

$$d = E[T_R^2] \frac{q-2}{q} \quad (7)$$

$$q = \left(\frac{\text{Var}[T_R^2] p}{E[T_R^2]^2} - 1 \right)^{-1} (p+2) + 4 \quad (8)$$

Since the mean and variance of the T_R^2 distribution can not be obtained analytically, they will be approximated by simulation. It will be shown that this approach results in a very accurate approximation of the true distribution of T_R^2 . Moreover, we will construct functions which yield values for $E[T_R^2]$ and $\text{Var}[T_R^2]$ for all n and p , so there will be no need for further simulation in practice.

Note that we do know the exact asymptotic distribution of T_R^2 . The asymptotic normality of T^1 implies that

$$\sqrt{n}(T^1 - \boldsymbol{\mu}) \xrightarrow[n \rightarrow \infty]{} N_p(\mathbf{0}, \kappa \Sigma)$$

Table 1 Asymptotic variance of T_j^1 for the standard Gaussian model

p	1	2	3	5	10	20	50
κ	1.258	1.145	1.112	1.085	1.063	1.050	1.040

where κ is the asymptotic variance of a component of T^1 . This asymptotic variance can be derived from the influence function of the reweighted MCD location given in [5]. Table 1 lists values of κ for several dimensions p . It can now easily be shown that

$$T_R^2 = n(T^1 - \boldsymbol{\mu})'(C^1)^{-1}(T^1 - \boldsymbol{\mu}) \xrightarrow[n \rightarrow \infty]{} \kappa \chi_p^2 \quad (9)$$

from which it follows that

$$E[T_R^2] \xrightarrow[n \rightarrow \infty]{} \kappa p \quad \text{and} \quad \text{Var}[T_R^2] \xrightarrow[n \rightarrow \infty]{} \kappa^2 2p \quad (10)$$

These results will be incorporated in the functions that determine finite-sample values of d and q . The simulation study will show that the difference between finite sample values of d and q and their asymptotic counterparts can be quite large.

4 Monte Carlo simulations

From the previous section we know that we need Monte Carlo simulations to obtain values for the constants d and q . For several sample sizes n and dimensions p , we generated $m = 3000$ samples $X_j; j = 1, \dots, m$ from a standard Gaussian distribution. For each sample X_j we computed the reweighted MCD location and scatter estimates $(T_{(j)}^1, C_{(j)}^1)$, and the corresponding value $T_R^{2(j)}$. The mean and variance of these $T_R^{2(j)}$ values are then given by

$$m(T_R^2) := \frac{1}{m} \sum_{j=1}^m T_R^{2(j)} \quad \text{and} \quad s^2(T_R^2) := \frac{1}{m-1} \sum_{j=1}^m (T_R^{2(j)} - m(T_R^2))^2$$

Figure 1 shows some values of $m(T_R^2)$ versus the sample size n for (a) $p = 5$ and (b) $p = 10$ dimensions. The corresponding values of $s^2(T_R^2)$ are shown in Figure 2. These plots show a smooth pattern, hence we determine smooth functions to fit these points. On the plots we also added a horizontal line indicating the asymptotic value which can be obtained from (10) by using the values of κ given in Table 1. For p fixed ($p=1,2,3,4,5,6,7,8$ and 10) we fitted the values of $m(T_R^2)$ using the following regression model:

$$f_p(n) = \kappa p + \frac{\alpha_p}{n^{\beta_p}}$$

The functions obtained in this way are superimposed in Figure 1. Hence, once we have determined the parameter values α_p and β_p for a given dimension p , then for any sample size n the value $f_p(n)$ is an approximation of

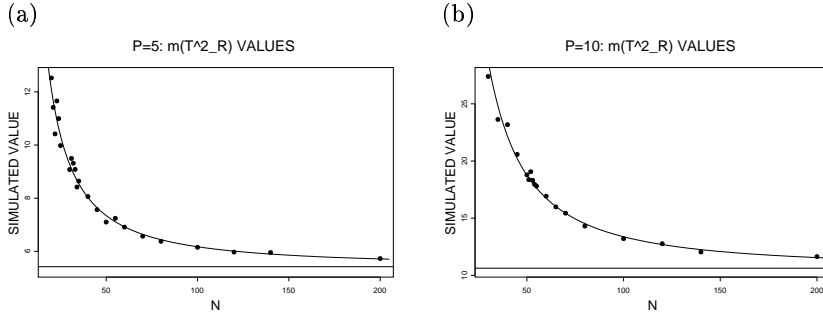


Fig. 1 Simulated values $m(T_R^2)$. (a) $p=5$; (b) $p=10$.

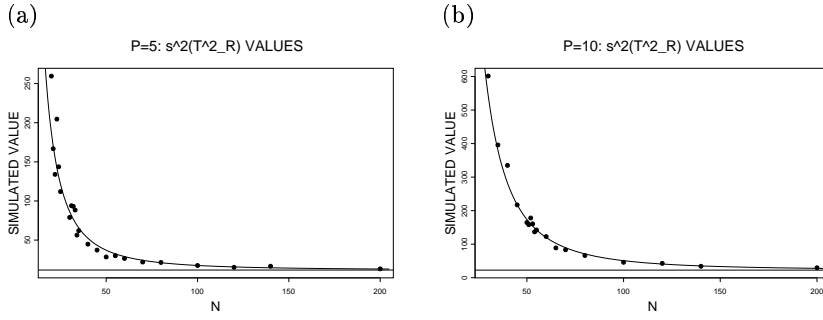


Fig. 2 Simulated values $s^2(T_R^2)$. (a) $p=5$; (b) $p=10$.

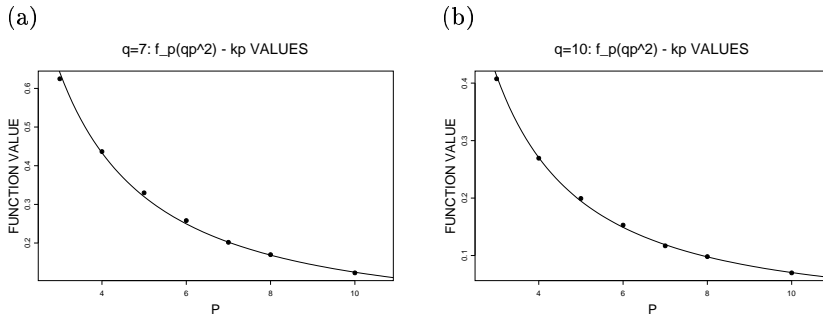


Fig. 3 Regression fit for $f_p(qp^2) - \kappa p$. (a) $q=7$; (b) $q=10$.

$E[T_R^2]$. However, this would require simulations for all dimensions p to determine the corresponding parameter values α_p and β_p . Therefore, for $p \geq 3$ we fitted the values $f_p(qp^2) - \kappa p$ for $q = 7$ and $q = 10$ as a function of the dimension p . Figure 3 shows the values $f_p(qp^2) - \kappa p$ versus the dimension p for (a) $q = 7$ and (b) $q = 10$. We fitted the smooth pattern shown in Figure 3 by using the model

$$g_q(p) = \frac{\gamma_q}{p^{\delta_q}}$$

which yields values for γ_q and δ_q for $q = 7$ and $q = 10$. Finally we obtain the following procedure to determine the value of $E[T_R^2]$ for any sample size n and dimension p :

- If $p = 1$ or $p = 2$ then the value of $E[T_R^2]$ is approximated by $f_1(n)$ and $f_2(n)$ respectively.
- If $p > 2$, then we first solve the following system of equations to obtain the parameter values $\widehat{\alpha}_p$ and $\widehat{\beta}_p$:

$$\frac{\widehat{\alpha}_p}{(7p^2)^{\widehat{\beta}_p}} = \frac{\gamma_7}{p^{\delta_7}}$$

$$\frac{\widehat{\alpha}_p}{(10p^2)^{\widehat{\beta}_p}} = \frac{\gamma_{10}}{p^{\delta_{10}}}.$$

Note that this can be rewritten into a linear system of equations by taking logarithms:

$$\ln(\widehat{\alpha}_p) - \widehat{\beta}_p \ln(7p^2) = \ln\left(\frac{\gamma_7}{p^{\delta_7}}\right)$$

$$\ln(\widehat{\alpha}_p) - \widehat{\beta}_p \ln(10p^2) = \ln\left(\frac{\gamma_{10}}{p^{\delta_{10}}}\right)$$

The value of $E[T_R^2]$ is now approximated by $\widehat{f}_p(n)$ where $\widehat{f}_p(n) := \kappa p + \frac{\widehat{\alpha}_p}{n^{\widehat{\beta}_p}}$.

Similarly a function that yields values of $\text{Var}[T_R^2]$ is derived, now starting by fitting the $s^2(T_R^2)$ values for fixed p by using the model

$$h_p(n) = \kappa^2 2p + \frac{\epsilon_p}{n^{\zeta_p}}$$

The functions obtained in this way are superimposed in Figure 2. Then the values of $h_p(np^2) - \kappa^2 2p$ have been fitted for $p \geq 3$ such that we obtain a procedure similar to the previous one to determine approximations of $\text{Var}[T_R^2]$ for all sample sizes n and dimensions p .

Using these procedures we obtain the functions shown in Figures 4 and 5 for the dimensions $p = 5$ and $p = 10$. We see that these curves are nearly the same as the original ones, which were shown in Figures 1 and 2. Furthermore it is clear that these functions yield good approximations for the actual simulated values of $E[T_R^2]$ and $\text{Var}[T_R^2]$.

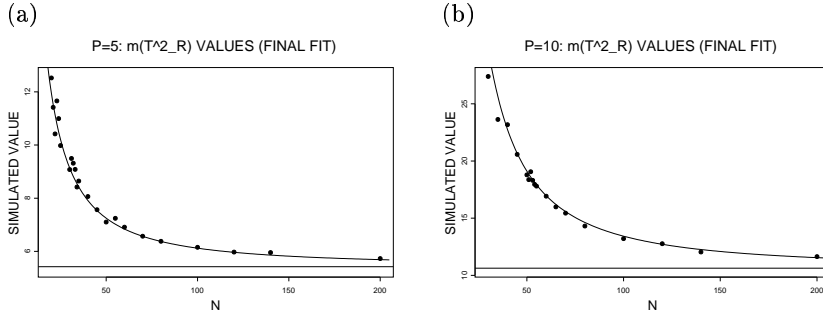


Fig. 4 Simulated values $m(T_R^2)$. (a) $p=5$; (b) $p=10$.

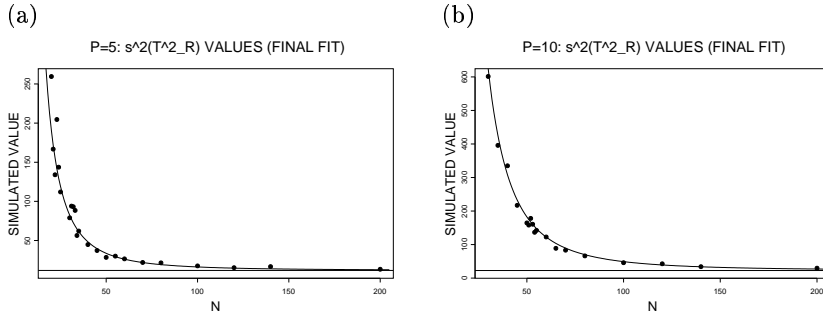


Fig. 5 Simulated values $s^2(T_R^2)$. (a) $p=5$; (b) $p=10$.

5 Approximate distribution of T_R^2

In the previous section we performed a simulation study to obtain estimates for $E[T_R^2]$ and $\text{Var}[T_R^2]$. They are to be used in equations (7) and (8) which in turn yield the appropriate parameter values for the approximate distribution given by (6). Now we verify the accuracy of this approximation. For this purpose, for several sample sizes n and dimensions p we performed simulations with $m = 3000$ data sets $Z_j; j = 1, \dots, 3000$ generated from the standard Gaussian distribution $N_p(\mathbf{0}, \mathbf{I}_p)$. For each data set Z_j we computed the T_R^2 statistic. To compare the empirical distribution of these 3000 T_R^2 statistics with the approximate T_R^2 distribution given by (6), we plotted the square root of the ordered T_R^2 values versus the square root of the quantiles of this approximate distribution. Some of the QQ-plots are shown in Figure 6. The vertical lines in the plots indicate the 95%, 97.5% and 99% quantiles (which are popular choices for the cutoff value of a test) of the approximate T_R^2 distribution. Figures 6a and 6b show that the approximate T_R^2 distribution is excellent for large sample sizes ($n \geq 100$). Also for small samples the approximation is accurate, even in high dimension. For example, for $p = 5$ Figure 6d shows that the approximate T_R^2 distribution is already very accurate for $n = 30$. Hence, we conclude that the approximate T_R^2 distribution obtained by using the values d and q given by the

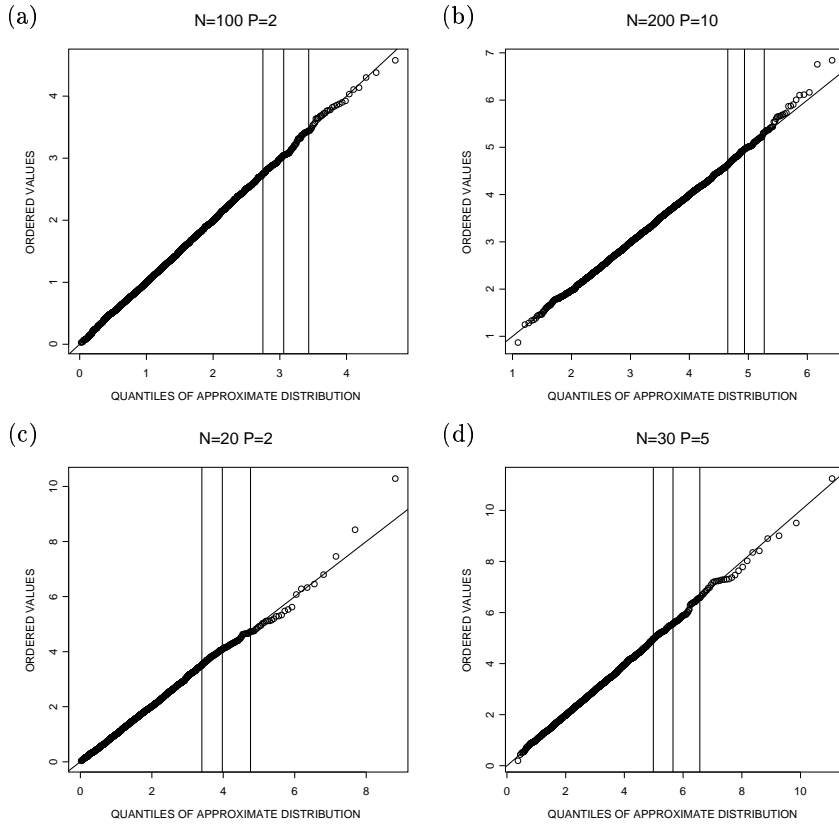


Fig. 6 QQ-plots for T_R^2 ; (a) $p=2$, $n=100$; (b) $p=10$, $n=100$; (c) $p=2$, $n=20$; (d) $p=5$, $n=30$.

functions that we constructed in the previous section, in all cases yields a good approximation to the true finite-sample distribution of T_R^2 .

Our aim is to construct robust inference tools based on the T_R^2 statistic. We mainly focus on the one-sample hypothesis test for the center based on the robust statistic T_R^2 . To further investigate the accuracy of the approximate T_R^2 distribution, we check in our simulations whether the actual percentage of T_R^2 values above the cutoff, given by a quantile of the approximate T_R^2 distribution, corresponds well to the nominal value of the quantile.

For several values of n and p Table 2 shows the actual percentage of T_R^2 values above the 95% quantile of the approximate T_R^2 distribution, while Table 3 contains the results for the 99% quantile.

We clearly see that in all cases the difference between the actual cutoff and the nominal value is very small. Hence, by using this approximate T_R^2 distribution the hypothesis test based on T_R^2 will give good Type I-error probabilities.

Table 2 Percent above the 5% cutoff (F -distribution)

p	n								
	10	20	30	50	70	100	140	200	1000
2	6.6	5.8	5.1	5.2	5.4	5.0	4.9	4.8	4.5
3	7.3	4.7	4.5	4.9	5.0	4.8	5.0	5.0	4.6
5		6.4	4.9	4.8	5.1	5.1	5.1	5.3	5.2
7		3.8	6.5	5.5	5.3	4.9	5.2	5.3	4.5
10			4.8	5.5	4.5	4.4	4.8	5.1	4.8

Table 3 Percent above the 1% cutoff (F -distribution)

p	n								
	10	20	30	50	70	100	140	200	1000
2	1.8	0.9	0.8	1.1	1.0	1.0	0.9	0.8	1.0
3	2.2	1.1	0.9	0.9	1.1	0.9	1.2	1.1	0.9
5		1.8	1.0	0.7	1.1	0.8	1.1	1.0	1.0
7		0.8	1.6	1.0	1.2	0.9	1.1	0.9	0.9
10			0.8	1.0	0.8	0.8	1.1	1.1	0.8

Table 4 Percent above the 5% cutoff (χ^2 -distribution)

p	n								
	10	20	30	50	70	100	140	200	1000
2	19.9	13.9	10.2	8.9	7.6	6.3	6.2	5.4	4.6
3	29.6	17.6	14.9	10.5	8.7	6.8	6.6	5.8	4.8
5		32.2	21.9	14.7	11.4	8.4	7.1	6.7	5.4
7		38.3	36.4	20.8	16.0	10.7	8.5	8.0	4.5
10			51.7	37.3	23.6	16.5	9.2	9.2	5.2

6 Comparison with other approximations

We now compare our approximation of the true finite-sample T_R^2 distribution with two other possible approximations that are more intuitively appealing but will be shown to be less accurate. First, we investigate whether the asymptotic distribution of T_R^2 given by (9) can also be used in the finite-sample case. Table 4 lists the actual percentages of T_R^2 values above the 95% quantile of the χ_p^2 distribution multiplied by κ . We see that indeed the actual percentages converge to the nominal values when n increases. However, for small sample sizes the actual percentages are much larger than the percentages corresponding to the approximate T_R^2 distribution in Table 2.

Another intuitively appealing approach to obtain an approximation for the distribution of the T_R^2 statistic consists of using the classical T^2 distribution that corresponds to the number of observations with weight 1 in the reweighted MCD estimates (2) and (3). Hence, this approach uses the distribution given in (1) but with n replaced by the number of observations

Table 5 Percent above the 5% cutoff (classical approach using the number of observations with weight 1)

p	n								
	10	20	30	50	70	100	140	200	1000
2	11.7	11.7	9.8	9.5	8.9	8.5	8.2	7.7	6.7
3	11.9	11.9	11.8	10.1	9.2	7.8	7.8	7.5	6.4
5		15.2	13.3	11.4	9.9	8.1	7.6	8.0	6.8
7		11.0	17.9	13.4	11.5	9.1	8.1	8.4	5.9
10			17.3	17.3	14.3	10.4	9.1	7.8	5.7

with weight 1. We included this approach in the simulations of Section 5 and the results for the 95% quantile are shown in Table 5. We again see that these percentages are quite different from the ideal value 5.0. Although the results are better than in Table 4, these results are still quite different from the nominal value and much worse than the results in Table 2. Therefore, we conclude that our approximate T_R^2 distribution outperforms both intuitive approaches.

We implemented an S-Plus function that determines for every n and p the approximate T_R^2 distribution given by (6) and which uses the functions derived in Section 4. The function returns a desired quantile of the approximate T_R^2 distribution or the p-value corresponding to the value of the T_R^2 statistic. The S-Plus function is available from our website <http://win-www.uia.ac.be/u/statis/>.

7 Power of the resulting one-sample test

In the previous section we showed that a one-sample test for the center $\boldsymbol{\mu}$ based on the T_R^2 statistic yields reliable significance levels. Moreover, tests based on T_R^2 have the advantage of being robust against outliers in the data. On the other hand it is well known that the classical Hotelling T^2 test is equivalent to the likelihood ratio test for the center in the normal model, and therefore has maximal asymptotic power. To investigate the power of the one-sample hypothesis test based on T_R^2 we simulated $m = 3000$ data sets from the distribution $N_p(\boldsymbol{\mu}, \mathbf{I}_p)$ with $\boldsymbol{\mu} = (\mu_1, \dots, \mu_p)' = (b, \dots, b)'$ and computed the classical T^2 and the T_R^2 statistics. At the 5% significance level we tested the hypothesis $H_0 : \boldsymbol{\mu} = \mathbf{0}$ against the alternative $\boldsymbol{\mu} \neq \mathbf{0}$. Table 6 shows the percentage of tests which rejected the null hypothesis in the simulations for the case $b = 0.2$. We see that the loss in power is acceptable, even for small sample sizes.

8 The T_R^2 statistic in the presence of outliers

We now investigate the robustness of the one-sample hypothesis test for the center based on the T_R^2 test statistic. Therefore, we performed simulations

Table 6 Power for $b=0.2$ in percent, for the 5% cutoff

p		n							
		10	20	30	50	70	100	140	200
2	T_R^2	8.9	11.2	16.4	30.3	44.6	61.6	78.4	91.4
	T^2	8.6	16.7	23.8	32.8	52.9	66.6	84.4	94.5
5	T_R^2		14.3	19.4	43.5	67.9	88.7	97.4	99.8
	T^2		19.2	34.9	62.2	78.5	94.2	99.1	100.0
10	T_R^2			18.0	41.3	74.5	96.9	99.8	100.0
	T^2			43.8	79.6	94.8	99.6	100.0	100.0

Table 7 10% far outliers: Percentage of erroneous rejections of H_0

p		n							
		10	20	30	50	70	100	140	200
2	T_R^2	4.7	5.0	4.4	5.1	5.4	5.9	5.4	5.0
	T^2	2.4	5.4	11.1	40.0	78.8	99.2	100.0	100.0
5	T_R^2		3.9	3.6	4.7	5.0	5.9	6.3	6.7
	T^2		6.1	12.0	23.3	42.2	88.3	100.0	100.0
10	T_R^2			1.3	2.8	3.9	5.3	7.2	6.8
	T^2			10.6	19.2	28.2	57.1	91.4	100.0

with contaminated data sets. For these data sets 90% of the observations were generated from the standard Gaussian distribution $N_p(\mathbf{0}, \mathbf{I}_p)$ and the other 10% were taken from $N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ with $\boldsymbol{\mu} = (5 + p, \dots, 5 + p)'$ and $\boldsymbol{\Sigma} = \text{diag}(0.1, \dots, 0.1)$. Table 7 lists the percentage of tests in the simulations that rejected the hypothesis $H_0 : \boldsymbol{\mu} = \mathbf{0}$ at the 5% level. We see that the actual percentages based on T_R^2 are still close to the nominal values, and are a big improvement compared to the results of the classical T^2 .

9 Example

For an application of robust inference techniques based on T_R^2 , we consider the Philips data (see [10]). The data set consists of 677 diaphragm parts for TV sets, produced by Philips Mecoma, of which nine characteristics were measured. We use six of these characteristics such that we have a data set with $n = 677$ observations and $p = 6$ variables. Earlier analysis of this data ([10]) showed a strongly deviating group of outliers, ranging from index 491 to index 565. Let us denote $\boldsymbol{\mu}_0$ the mean of the majority of the data without the outliers. We performed a one-sample hypothesis test for the hypothesis $H_0 : \boldsymbol{\mu} = \boldsymbol{\mu}_0$. The classical Hotelling test statistic for this data set yields $T^2 = 62.8$ with corresponding p-value smaller than 0.00001. Hence, based on the classical T^2 the null hypothesis is rejected, although we know that

Table 8 95% T^2 and T_R^2 intervals for the components of the mean

Variable	T^2 -interval		T_R^2 -interval		μ_0
	lower	upper	lower	upper	
1	-0.04693	-0.01429	-0.05678	-0.01831	-0.03502
2	-0.05178	-0.02876	-0.05761	-0.03109	-0.04355
3	0.43933	0.44739	0.43456	0.44358	0.44008
4	2.09858	2.11733	2.11713	2.13375	2.12338
5	0.43410	0.44057	0.43167	0.43846	0.43412
6	-0.07718	-0.06817	-0.06916	-0.06240	-0.06589

the center of the majority of the data equals μ_0 . On the other hand, the robust T_R^2 statistic yields $T_R^2 = 5.2$ with corresponding p-value 0.426 hence we accept the null hypothesis in this case, as would be expected from the majority of the data. This clearly illustrates that the T_R^2 statistic based on reweighted MCD inherits the robustness of MCD while the classical Hotelling test statistic based on the mean and covariance is distorted by the outliers.

We now illustrate other robust inference techniques based on the T_R^2 statistic. For example, robust simultaneous confidence intervals for linear combinations of the center can be derived. Similarly to the classical case (see e.g. [4, page 239]) we obtain that

$$\max_l \frac{n(l'T^1 - l'\mu)^2}{l'C^1l} = T_R^2$$

Therefore, robust $100(1 - \alpha)\%$ simultaneous confidence intervals for the linear combinations $l'\mu$ are given by

$$l'T^1 \pm \sqrt{\frac{d}{n} F_{p,q,1-\alpha}} l'C^1l$$

We computed 95% simultaneous confidence intervals for the center of the Philips data. Table 8 shows the classical and robust simultaneous confidence intervals (denoted as T^2 intervals and T_R^2 intervals) for the components of the center of the data as well as the mean μ_0 of the majority of the data. We see that the 4th and the 6th component of μ_0 do not lie in their respective T^2 intervals, while on the other hand every T_R^2 interval contains the corresponding component of μ_0 . We also consider classical and robust Bonferroni simultaneous confidence intervals corresponding to the T^2 and T_R^2 statistics (see [4, page 249]). Table 9 shows the 95% classical and robust Bonferroni intervals for the components of the center of the Philips data. Note that the Bonferroni intervals in Table 9 are shorter than the confidence intervals of Table 8. Moreover, we see that now only the first and second component of μ_0 fall inside the classical Bonferroni interval. On the other hand, the robust Bonferroni intervals do contain the respective components of μ_0 . This clearly illustrates that simultaneous confidence intervals based on the T_R^2 statistic yield robust intervals for the center of a data set.

Table 9 95% Bonferroni intervals for the components of the mean

Variable	Classical Bonf.		Robust Bonf.		μ_0
	lower	upper	lower	upper	
1	-0.04269	-0.01852	-0.05426	-0.02083	-0.03502
2	-0.04879	-0.03175	-0.05588	-0.03283	-0.04355
3	0.44037	0.44635	0.43515	0.44299	0.44008
4	2.10102	2.11490	2.11822	2.13267	2.12338
5	0.43494	0.43973	0.43212	0.43802	0.43412
6	-0.07601	-0.06934	-0.06872	-0.06285	-0.06589

10 Conclusion

In this paper we constructed the T_R^2 statistic as a robust alternative for the classical Hotelling T^2 statistic. The T_R^2 statistic is obtained by replacing the classical mean and covariance in the T^2 statistic by the reweighted MCD location and scatter. We proposed an approximation for the finite-sample distribution of the T_R^2 statistic which allows us to construct robust multivariate tests and confidence intervals for the center of a data set. The approximation is based on matching the mean and variance of a multiple of an F -distribution. Using the results of a Monte Carlo study functions were constructed that, for all sample sizes n and dimensions p , return values for the constants necessary to determine the approximate T_R^2 distribution. Simulations showed that the robust T_R^2 statistic has good power compared to the classical T^2 statistic, and performs much better in the case of contaminated data sets. Finally, an example illustrated how the T_R^2 statistic can be used in practice.

References

1. Butler, R.W., Davies, P.L. and Juhn, M., "Asymptotics for the Minimum Covariance Determinant Estimator," *The Annals of Statistics* **21**, (1993) 1385-1400.
2. Croux, C. and Haesbroeck, G., "Influence and Efficiency of the Minimum Covariance Determinant Scatter Matrix Estimator," *Journal of Multivariate Analysis* **71**, (1999) 161-190.
3. Hardin, J. and Rocke, D.M., "The Distribution of Robust Distances," Technical Report, Univ. of California at Davis.
4. Johnson, R.A. and Wichern, D.W., *Applied Multivariate Statistical Analysis* (Prentice-Hall, Inc., Englewood Cliffs, New Jersey 1988)
5. Lopuhaä, H.P., "Asymptotics of Reweighted Estimators of Multivariate Location and Scatter," *The Annals of Statistics* **27**, (1999) 1638-1665.
6. Lopuhaä, H.P. and Rousseeuw, P.J., "Breakdown Points of Affine Equivariant Estimators of Multivariate Location and Covariance Matrices," *The Annals of Statistics* **19**, (1991) 229-248.
7. Mardia, K.V., Kent J.T. and Bibby J.M., *Multivariate Analysis* (Academic Press Ltd., London 1995)

8. Pison, G., Van Aelst, S. and Willems, G., "Small Sample Corrections for LTS and MCD," Technical Report, Univ.of Antwerp.
9. Rousseeuw, P.J., "Least Median of Squares Regression," *Journal of the American Statistical Association* **79**, (1984) 871-880.
10. Rousseeuw, P.J. and Van Driessen, K., "A Fast Algorithm for the Minimum Covariance Determinant Estimator," *Technometrics* **41**, (1999) 212-223.
11. Rousseeuw, P.J. and Van Zomeren, B.C., "Unmasking Multivariate Outliers and Leverage Points," *Journal of the American Statistical Association* **85**, (1990) 633-651.