Robust Hotelling's *T*2 statistic based on M-estimator

Robust Hotelling's T^2 statistic based on M-estimator

Mohd Aizat Ahlam Mohamad Mokhtar[1], Nur Syahidah Yusoff [2], and Chuan Zun Liang[3]

[1,2,3]Centre for Mathematical Sciences, College of Computing & Applied Sciences, Universiti Malaysia Pahang, Lebuhraya Tun Razak, 26300, Gambang, Pahang, Malaysia

Email: wnsyahidah@ump.edu.my

Abstract. Hotelling's T^2 statistic is the multivariate generalization of the student's *t*-statistic. Hotelling's T^2 statistics is a method for testing hypotheses about multidimensional means. However, the classical Hotelling's T^2 statistic is very sensitive to the presence of outliers. In order to overcome this limitation, a modification is needed so that Hotelling's T^2 is robust. In this paper, classical Hotelling's T^2 statistic has been modified by substituting mean vector and covariance matrix with a robust estimator. M-estimator has been used for this modification. The performance of modified Hotelling's T^2 statistic has been compared with the classical Hotelling's T^2 statistic and discussed in this paper to illustrate the advantage of modified Hotelling's T^2 statistic towards outliers. The performance of modified Hotelling's T^2 statistic is better than classical Hotelling's T^2 when number of sample, *n* and dimension, *p* is small.

Keywords: Hotelling's T^2 statistic, M-estimator, Robust estimator, Outlier

1. Introduction

The statistic Hotelling's T^2 is one of the multivariate statistical tools which are widely used for testing hypotheses about the mean [1]. It is called Hotelling's T^2 in honor of the one who first obtained its sampling distribution, Harold Hotelling [2]. Hotelling's T^2 is a multivariate generalization of the square of the univariate *t*. Unlike univariate *t*, Hotelling's T^2 examining group differences simultaneously on several dependent variables [3].

There are many situations where Hotelling's T^2 can be applied. For example, Hotelling's T^2 is used to compare mean vectors from two populations. In this study, single-sample Hotelling's T^2 are tested and each of the tested variables represents a characteristic of the populations. Besides that, Hotelling's T^2 are used to compare mean vector under two independent samples, paired comparison and also repeated measurement [2]. The details of these applications can be seen in [2]. Other than that, Hotelling's T^2 are also used for control chart [3].

In this study, Hotelling's T^2 performance has been evaluated. However, Hotelling's T^2 is sensitive to outliers [4], even a single extreme outlier can have a large distorting influence on its performance [5]. Moreover, multiple outliers not only decrease the performance of classical Hotelling's T^2 but also creating "masking effect" [6]. It is known that to calculate mean vector, $\bar{\boldsymbol{x}}$ and

covariance matrix, $\boldsymbol{S}$ every single data has to be used. Hence, outliers in data will affect the outcome of mean vector, $\bar{\boldsymbol{x}}$ and covariance matrix, $\boldsymbol{S}$. In multivariate setting, it is difficult to avoid from these outliers.

In order to overcome this limitation, a robust estimator has been introduced. Many robust estimators have been proposed and performed well when outliers are presented. In this study, M-estimator has been integrated into Hotelling's T^2. M-estimator was first introduced for the estimation of a one-dimensional location parameter by Huber [7]. Later, Maronna successfully defined M-estimator for multivariate location and parameter [8]. The breakdown point of this M-estimator is at most equal to $1/(p+1)$. From the viewpoint of breakdown point, as the dimension increases, M-estimator become more sensitive [9]. By using this estimator, a robust alternative to Hotelling's T^2 has been constructed in order to avoid the negative effect of outliers.

The objective of this study is to evaluate the performances of classical and modified Hotelling's T^2. There are two way of modification, the first one is by substituting covariance matrix, $\boldsymbol{S}$ with M-estimator, $\boldsymbol{S}_M$. the second one is by substituting both mean vector, $\bar{\boldsymbol{X}}$ and covariance matrix, $\boldsymbol{S}$ with M-estimator, $\bar{\boldsymbol{X}}_M$. and $\boldsymbol{S}_M$.

2. Methodology

2.1 Classical Hotelling's T^2

Let $\boldsymbol{X}_1, \boldsymbol{X}_2, \ldots, \boldsymbol{X}_n$ be a random sample from an $N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ population. Then, classical Hotelling's T^2 [2] is as follows

$$T^2 = n(\bar{\boldsymbol{x}} - \boldsymbol{\mu}_0)'\boldsymbol{S}^{-1}(\bar{\boldsymbol{x}} - \boldsymbol{\mu_0}) \tag{1}$$

where,

$\bar{\boldsymbol{x}}$ = sample mean vector
$\boldsymbol{S}^{-1}$ = the inverse of sample covariance matrix
n = number of sample
$\boldsymbol{\mu}_0$ = plausible value for the mean vector

In order to test hypothesis of H_0; $\boldsymbol{\mu} = \boldsymbol{\mu}_0$ and H_1: $\boldsymbol{\mu} \neq \boldsymbol{\mu}_0$. The critical value of (1) is determined by (2)

$$CV_F = \frac{(n-1)p}{(n-p)} F_{p,n-p}(\alpha) \tag{2}$$

where n is number of sample, p is number of dimension, and α is type I error. In this case, if T^2 > critical value, H_0 is rejected which indicate there is a differences in mean vector.

2.2 Modified Hotelling's T^2 based on M-estimator

Let $\boldsymbol{X}_1, \boldsymbol{X}_2, \ldots, \boldsymbol{X}_n$ be a random sample from an $N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ population. Then, the mean and covariance matrix for M-estimator [10] is given by

$$\boldsymbol{X}_M = \sum w_i \boldsymbol{X}_i / n \tag{3}$$

$$\boldsymbol{S}_M = \frac{1}{\tau n} \sum {w_i}^2 (\boldsymbol{X}_i - \bar{\boldsymbol{X}})(\boldsymbol{X}_i - \bar{\boldsymbol{X}})' \tag{4}$$

where, w_i is a function, τ is chosen so that $\boldsymbol{S}$ is an unbiased estimate of the covariance matrix. M-estimator is basically a downweight of a proportion of K observations. Let ϱ^2 be the $1 - K$ quantile of a chi-squared distribution with p degrees of freedom. Let $w_i = 1$ if $d_i \leq \varrho$ and otherwise $w_i = \varrho/d_i$.

$$d_i{}^2 = (\boldsymbol{X}_i - \bar{\boldsymbol{X}})'\boldsymbol{S}^{-1}(\boldsymbol{X}_i - \bar{\boldsymbol{X}}) \tag{5}$$

where d_i is Mahalanobis distance. These updated estimates are used to update the squared Mahalanobis distances, which in turn yields a new updated estimate of the mean and covariance matrix. This process is continued until convergence is achieved.

In this study, classical Hotelling's T^2 have been modified by substituting covariance matrix, $\boldsymbol{S}$ with M-estimator, $\boldsymbol{S}_M$. The modified Hotelling's T^2 is given by.

$$T_M^2 = n(\boldsymbol{x_i} - \boldsymbol{X})'\boldsymbol{S}_M{}^{-1}(\boldsymbol{x_i} - \boldsymbol{X}) \tag{6}$$

where n is a number of sample, $\bar{\boldsymbol{X}}$ is a sample mean vector and $\boldsymbol{S}_M^{-1}$ is the inverse covariance of M-estimator. Although the primary goal is to analyse Equation (6), we also had analyse another modified Hotelling's T^2 Equation (7) where both sample mean vector and covariance matrix have been substituted with M-estimator.

$$T_{ME}^2 = n(\boldsymbol{x_i} - \boldsymbol{X}_M)'\boldsymbol{S}_M{}^{-1}(\boldsymbol{x_i} - \boldsymbol{X}_M) \tag{7}$$

where n is a number of sample, $\bar{\boldsymbol{X}}_M$ is a mean vector of M-estimator and $\boldsymbol{S}_M^{-1}$ is the inverse covariance of M-estimator.

In this study, K has been set at 0.1 in every situation. In simulation design, the amount of contamination can be set according to the researcher interest. However, in real dataset, the amount of contamination is unknown. Hence, in this simulation design, the value of K has been set at 0.1 regardless of the amount of contamination. The value of K has been set at 0.1 as suggested by [10].

There are many robust estimators other than M-estimator. Some of the examples are Minimum Volume Ellipsoid (MVE), Minimum Covariance Determinant (MCD), S-Estimators, Constrained M-estimators and The Stahel-Donoho W-estimator [10].

2.3 Simulation design

2.3.1 Critical value. The distribution of modified Hotelling's T^2 is not known and simulation has been used to determine this distribution for both classical and modified Hotelling's T^2. We generated 5000 data sets from the $N_p(0, \boldsymbol{\Sigma})$ at the value of type 1 error, $\alpha = 0.05$. The value of $\boldsymbol{\mu}$ is set equal to 0 for each dimension, p and $\boldsymbol{\Sigma}$ is set as default. However, for each dimension, p, the amount of deviation, σ are set to be equal to 1. By using these data sets, we calculate T^2 for classical and modified Hotelling's T^2 as given by Equations (1), (6) and (7). The algorithm of M-estimator used is written by Wilcox in *Rallfun-v35*5 source with *MARest* function. The values of w_i and τ is set as default. The function of w_i depend on the value of K where in this study K has been chosen equal to 0.1. The value of τ is chosen so that $\boldsymbol{S}$ is an unbiased estimate. The 95th quantile from the results will be set as CVs. We also use the CVs from Equation (4) and Chi-Squared distribution as comparison. The simulated CVs have been calculated for each n = 30, 50, 100, and 200 and p = 2, 3, and 5. The results can be found in Table 1.

2.3.2 Performance of the Hotelling's T^2. In order to evaluate the performance of classical Hotelling's T^2 and modified Hotelling's T^2, multiple criteria data sets have been generated in simulation. The sample sizes were n = 30, 50, 100 and 200 observations and the number of dimensions were p = 2, 3,

and 5. The amount of contamination were ε = 0, 0.1 and 0.2. For each specific criteria, 1000 data sets have been generated and computed.

The simulation model used in this simulation is contaminated model by using a mixture of normal.

$$(1 - \varepsilon)N_p(\boldsymbol{\mu}_0, \boldsymbol{\Sigma}_0) + \varepsilon N_p(\boldsymbol{\mu}_1, \boldsymbol{\Sigma}_1) \quad (8)$$

where ε is the proportion of outliers, $\boldsymbol{\mu}_0$ and $\boldsymbol{\Sigma}_0$ are the uncontaminated mean vector and covariance matrix and $\boldsymbol{\mu}_1$ and $\boldsymbol{\Sigma}_1$ are the contaminated mean vector and covariance matrix.

$$(1 - \varepsilon)N_p(0, \boldsymbol{\Sigma}_0) + \varepsilon N_p(0, \boldsymbol{\Sigma}_1) \quad (9)$$

The amounts of contamination, ε used are as stated previously which are 0, 0.1 and 0.2. The values of $\boldsymbol{\mu}_0$ and $\boldsymbol{\mu}_1$ used are 0 for each dimension, p. The value of $\boldsymbol{\Sigma}_0$ is set by using standard deviation, $\boldsymbol{\sigma}_0$ for each variable, p, where $\boldsymbol{\sigma}_0$ is set to be 1. On the other hand, the value of $\boldsymbol{\Sigma}_1$ is set by using extreme standard deviation, σ_1 for each variable, p. In this study, σ_1 is set to be 3 and 5 [11].

Using 1000 replications, simulation has been analysed with different value of sample sizes, number of dimensions and amount of contamination. The type I error, α used in this simulation is 0.05. The simulation steps are as follow:

1) Dataset have been generated.
2) The values of T^2, T^2_M and T^2_{ME} have been computed by using Equations (1), (6) and (7).
3) The values of T^2, T^2_M and T^2_{ME} greater than critical value (from Table 1) have been determined.
4) The performance of T^2, T^2_M and T^2_{ME} has been recorded in percentage of Type 1 error.

3. Results and Discussion

The critical values for simulated and usual distribution are presented in table 1. The performances of classical and modified Hotelling's T^2 for few cases are presented in table 2 -4.

Table 1. Simulated and usual critical values

p	n	T^2_C	T^2_M	T^2_{ME}	CV_F	χ^2_p
2	30	7.1522	7.4571	7.4659	6.9186	5.99
	50	6.4259	6.4941	6.6408	6.5129	5.99
	100	6.0823	6.1618	6.2398	6.2431	5.99
	200	6.1446	6.1946	6.2058	6.1107	5.99
3	30	9.4099	9.8009	9.8665	9.5700	7.81
	50	8.1816	8.7471	8.9214	8.7887	7.81
	100	8.0271	8.2616	8.2497	8.2976	7.81
	200	8.1448	8.1042	8.0963	8.0610	7.81
5	30	14.9833	15.559	15.4741	15.0800	11.07
	50	12.8931	13.219	13.3285	13.1756	11.07
	100	11.8932	11.9945	12.0158	12.0363	11.07
	200	11.4768	11.5647	11.6420	11.5318	11.07

Table 2. False alarm rate of classical Hotelling's T^2 and modified Hotelling's T^2 when $p = 2$, and $\alpha = 0.05$.

n	ε	σ	T^2_C	T^2_M	T^2_{ME}
30	0	(1, 1)	0.043	0.038	0.043
	0.1	(3, 3)	0.027	0.095	0.040
		(5, 5)	0.022	0.280	0.042
	0.2	(3, 3)	0.049	0.149	0.066
		(5, 5)	0.033	0.379	0.073
50	0	(1, 1)	0.050	0.053	0.051
	0.1	(3, 3)	0.037	0.120	0.048
		(5, 5)	0.027	0.305	0.057
	0.2	(3, 3)	0.048	0.160	0.073
		(5, 5)	0.047	0.399	0.083
100	0	(1, 1)	0.050	0.052	0.056
	0.1	(3, 3)	0.046	0.136	0.053
		(5, 5)	0.037	0.316	0.056
	0.2	(3, 3)	0.049	0.183	0.066
		(5, 5)	0.041	0.422	0.076
200	0	(1, 1)	0.058	0.062	0.057
	0.1	(3, 3)	0.047	0.140	0.065
		(5, 5)	0.043	0.323	0.069
	0.2	(3, 3)	0.047	0.172	0.058
		(5, 5)	0.047	0.419	0.073

Table 3. False alarm rate of of classical Hotelling's T^2 and modified Hotelling's T^2 when $p = 3$, and $\alpha = 0.05$.

n	ε	σ	T^2_C	T^2_M	T^2_{ME}
30	0	(1, 1, 1)	0.051	0.047	0.046
	0.1	(3, 3, 3)	0.054	0.129	0.063
		(5, 5, 5)	0.039	0.347	0.067
	0.2	(3, 3, 3)	0.044	0.158	0.054
		(5, 5, 5)	0.028	0.454	0.065
50	0	(1, 1, 1)	0.057	0.051	0.050
	0.1	(3, 3, 3)	0.046	0.135	0.059
		(5, 5, 5)	0.035	0.370	0.063
	0.2	(3, 3, 3)	0.049	0.184	0.068
		(5, 5, 5)	0.042	0.509	0.079
100	0	(1, 1, 1)	0.050	0.048	0.053
	0.1	(3, 3, 3)	0.050	0.146	0.075
		(5, 5, 5)	0.041	0.404	0.077
	0.2	(3, 3, 3)	0.045	0.191	0.069
		(5, 5, 5)	0.047	0.536	0.087
200	0	(1, 1, 1)	0.055	0.055	0.057
	0.1	(3, 3, 3)	0.054	0.176	0.072
		(5, 5, 5)	0.054	0.447	0.078
	0.2	(3, 3, 3)	0.052	0.216	0.072
		(5, 5, 5)	0.046	0.537	0.098

Table 4. False alarm rate of of classical Hotelling's T^2 and modified Hotelling's T^2 when p = 5, and α = 0.05.

n	ε	σ	T^2_C	T^2_M	T^2_{ME}
30	0	(1, 1, 1, 1, 1)	0.058	0.053	0.055
	0.1	(3, 3, 3, 3, 3)	0.037	0.116	0.054
		(5, 5, 5, 5, 5)	0.032	0.389	0.059
	0.2	(3, 3, 3, 3, 3)	0.030	0.158	0.061
		(5, 5, 5, 5, 5)	0.017	0.556	0.069
50	0	(1, 1, 1, 1, 1)	0.045	0.040	0.042
	0.1	(3, 3, 3, 3, 3)	0.045	0.169	0.062
		(5, 5, 5, 5, 5)	0.030	0.486	0.067
	0.2	(3, 3, 3, 3, 3)	0.051	0.246	0.082
		(5, 5, 5, 5, 5)	0.028	0.665	0.091
100	0	(1, 1, 1, 1, 1)	0.054	0.058	0.059
	0.1	(3, 3, 3, 3, 3)	0.040	0.199	0.066
		(5, 5, 5, 5, 5)	0.030	0.574	0.080
	0.2	(3, 3, 3, 3, 3)	0.043	0.298	0.082
		(5, 5, 5, 5, 5)	0.035	0.709	0.104
200	0	(1, 1, 1, 1, 1)	0.054	0.054	0.053
	0.1	(3, 3, 3, 3, 3)	0.046	0.196	0.059
		(5, 5, 5, 5, 5)	0.040	0.583	0.068
	0.2	(3, 3, 3, 3, 3)	0.057	0.297	0.090
		(5, 5, 5, 5, 5)	0.056	0.724	0.124

The results from table 1 show that when n grows, the critical values for all the T^2 (robust and non-robust) and CV_F are similar to the χ^2_p. Generally, we can say that T^2_{ME} have the highest critical value followed by T^2_M, T^2_C, CV_F and χ^2_p.

The performance of T^2 is evaluated by comparing the false alarm rate with 0.05. The closer the false alarm rate of T^2 to 0.05, the better the performance of T^2. The result from table 2 - 4 shows that, the performance of T^2_C become lower as the outlier, σ increases. This is the effect of outliers in classical Hotelling's T^2 which is described as sensitive to outliers in introduction section. Classical Hotellling's T^2 have a tendency not to detect a shifted in mean. In hypothesis testing procedure, classical Hotelling's T^2 have a tendency of failing to reject H_o when outliers presented.

The performance of T^2_M, is extremely bad when outliers presented. This modified Hotelling's T^2 is extremely sensitive to outliers, unlike the classical Hotellling's T^2. T^2_M are sensitive in a way of it has extreme tendency of rejecting H_0 when outliers presented.

Similar to T^2_M, T^2_{ME} in general, has a tendency of rejecting H_0 when outlier presented. However, unlike T^2_M, T^2_{ME} performance is much better. From this simulation, T^2_{ME} is a better modified Hotelling's T^2 compared to T^2_M. Thus, the performance of T^2_{ME} will be used to be compared with the performance of classical Hotelling's T^2.

From table 2 - 4, In general, it is shown that, when n = 30, T^2_{ME} outperform T^2_C. However, as n increases, T^2_C perform better than T^2_{ME}. Despite T^2_C perform better as n increases, T^2_{ME} still perform well when p = 2. However, as p increases, T^2_C perform better than T^2_{ME}.

One of the important information here is the false alarm rate of T^2_C tend to decrease as the value of outliers increase as shown in table 2 – 4. On the other hand, the false alarm rate of T^2_{ME} tend to increase as the value of outliers increase as shown in Table 2 - 4. If both T^2_C and T^2_{ME} gives the same results which is reject or fail to reject H_0, the result is a good result. First, it is because the result is consistent and second, the behaviour of its performances, T^2_C tend to underestimate while T^2_{ME} tend to overestimate. However, if the result is different from each other, the simulation results can be used as reference by examine the number of sample and dimension. It is also suggested to use another robust estimator.

The difference between T^2_M and T^2_{ME} is $\bar{\boldsymbol{X}}$ and $\bar{\boldsymbol{X}}_M$ (refer to Equation (6) and (7)), and with this difference T^2_{ME} performed much better than T^2_M. T^2_M needs a modification that T^2_M will be used if the value of $\bar{\boldsymbol{X}}$ is approximately to $\bar{\boldsymbol{X}}_M$. A new method need to be developed to determine whether the value of $\bar{\boldsymbol{X}}$ and $\bar{\boldsymbol{X}}_M$ is close or far.

4. Conclusion

In general, based on those findings, T^2_{ME} outperform T^2_C when n is small. However, when n increases T^2_C outperform T^2_{ME}. When p increases, the performance of T^2_{ME} become lower compared to T^2_C. In conclusion, T^2_{ME} perform better when number of sample, n and dimension, p is small. If number of sample, n or dimension, p is larger, T^2_C is a better choice. As for T^2_M, a modification is needed to increase its performance. It is suggested to evaluate T^2_M performance only if $\bar{\boldsymbol{X}}$ is close to $\bar{\boldsymbol{X}}_M$.

Acknowledgement

The authors would like to thank the Universiti Malaysia Pahang for providing financial support under Internal Research grant RDU1903124.

References

[1] Raykov T and Marcoulides G A 2008 *An Introduction to Applied Multivariate Analysis*
[2] Johnson R and Wichern D 2014 *Applied Multivariate Statistical Analysis*
[3] Alfaro J L and Ortega J F 2009 A comparison of robust alternatives to Hotelling ' s T 2 control chart *J. Appl. Stat.* **36** pp 1385–96
[4] Haddad F 2018 Improvement of The Hotelling's T 2 Charts Using Robust Location Winsorized: One Step M-Estimator (WMOM) **50** pp 97–112
[5] Maronna R A and Martin R D 2006 *Robust Statistics Theory and Methods*
[6] Huber P J 1981 *Robust Statistics* vol 82
[7] Huber P J 1964 Robust Estimation of a Location Parameter *Ann. Math. Stat.* **35** pp 73–101
[8] Maronna R A 1976 Robust M-Estimators of Multivariate Location and Scatter *Ann. Stat.* **4** pp 51– 67
[9] Lopuhaa H P 1989 On the relation between S-estimators and M-estimators of multivariate: location and covariance *Ann. Stat.* **17** pp 1662–1683
[10] Wilcox R 2005 *Introduction to robust estimation and hypothesis testing*
[11] Alfaro J L and Ortega, J F 2008 A Robust Alternative to Hotelling's T 2 Control Chart: Trimmed *Qual. Reliab. Eng. Int.* pp 601–11