Journal of Engineering and Applied Sciences 13 (5): 1246-1252, 2018
ISSN: 1816-949X

# Type 1 Error Rate Comparison Between Classical and Modified Box M-Statistic

Shamshuritawati Sharif, Nuraimi Ruslan and Tareq A.M. Atiany
School of Quantitative Sciences, UUM-College of Arts and Sciences,
University Utara Malaysia, 06010 UUM Sintok, Kedah, Malaysia

**Abstract:** Classical Box M-statisticis one of Likelihood Ratio Test (LRT) constructed under the multivariate normality distribution. The performance of classical Box M-statistic by using classical estimators suffers from masking and swamping effects when the outlier occurs in data set. To alleviate the problem, robust estimators are recommended. In this study, a robust Box M-statistic based on a S-estimator, $M_s$ and M-estimator $M_M$ are proposed as the alternative to the classical Box M-statistic. Over the simulation study, the performance comparisonof classical, $M_s$ and MM-statistics are measured using type 1 error rates. From the results, it showed that $M_s$ (Box M-statistic based on S-estimator) has a competitive performance relative to $M_M$ and the classicalstatistic. In summary, $M_s$ can be used for testing the equality of two difference covariance matrices or more when the data contains outlier.

**Key words:** Covariance matrix, type 1 error, S-estimator, M-estimator, multivariate, distribution

## INTRODUCTION

Nowadays, in multivariate setting testing the stability of covariance matrices a serious problem and it has receiving much attention in economic and financial studies. For example, in financial industry Tang (1998) and Lee (2006), in real estate industry Eichholtz (1996). In medical research, Kupek (2002) stated that the covariance structure stability has been used to model the error structure of both observed and latent variables. The covariance structure stability is needed in portfolio optimization to determine the allocation of international real estate securities investments (Yusoff and Djauhari, 2013). Moreover, the covariance structure stability is also used to increase the quality and performance through the entire chain of marketing, expansion, production and sales processes this expected at the delivery a very high quality of product to the customers (Roes and Dorr, 1997).

Tang (1998) stated that the stability of covariance matrices can be examined by testing directly the equality of covariance matrices across time period. For that aim the most popular and widely used test is Box M which is constructed by Box (1949). This test involve determinant of sample covariance matrix, it is difficult to compute for the case of high dimensional data sets (Yusoff and Djauhari, 2012). The Generalized Variance (GV) is multivariate dispersion measure is used for testing the homogeneity of covariance matrices. The role of GV statistic is to test the equality of several independent samples of covariance structure (Sharif *et al*., 2014). However, it has drawback where the measure is imprecise because the two different covariance structures might be acknowledged equal to each other. Moreover, GV statistic needs that the condition of covariance matrix is nonsingular (Djauhari and Salleh, 2011). Due to that, this testis quite cumbersome to compute when the data sets are of high dimension. Djauhari (2007) developed Vector Variance (VV) statistic as a multivariate variability measure to help to solve the singularity problem when dealing with high dimension data set and to overcome the drawback of GV statistic. The computational time for VV statistic is shown to be better compared to GV statistic (Sharif *et al*., 2014).

Among all of those statistics stated above, Box M-statistic is widely practiced and substitute among applied researchers because it can be easily performed using IBM SPSS Software. Consequently, in this research, the focused is assumed on Box M-statistic, since, it is well recognized by applied researchers rather than GV which well known among pure statistical researcher.

The hypothesis used to test the equality of variance-covariance matrices is $H_0$: $\Sigma_1 = \Sigma_2 = ... = \Sigma_m$ versus $H_1$: $\Sigma_i \neq \Sigma_j$ for at least one pair (i, j) where i, j = 1, 2, ..., m. Thus, the M-statistic is derived as follows:

$$M = N\ln\left|\bar{S}\right| - \sum_{i=1}^{m} n_i \ln\left|S_i\right| \qquad (1)$$

**Corresponding Author:** Shamshuritawati Sharif, School of Quantitative Sciences, UUM-College of Arts and Sciences, University Utara Malaysia, 06010 UUM Sintok, Kedah, Malaysia

Where:

$\bar{S}=\frac{1}{N}\sum_{i=1}^{m}n_i s_i$ = The pooled sample variance-covariance matrix
$S_i$ = The variance-covariance matrix calculated from the sample i
m = The number of subgroup where the stability of matrices is hypothesized
N = $n_1+n_2+, ..., + n_m$
$n_i$ = ith sample size

Box (1949) recommended that under $H_0$, the statistical test can be approximated either by $\chi^2$ distribution or F distribution. According to Mardia *et al.* (1979), the $\chi^2$ approximation will be adequate to be used in any practical determinations. Moreover, $\chi^2$ approximation is good if the number of sample sizes, n is >20. Consequently, the statistical test will be rejected at significance level, $\alpha$ if M/b exceeds $\chi^2_{a,\,v}$ where:

$$b=\frac{1}{1\text{-}a};\ a=\frac{\left(2p^2+3p\text{-}1\right)}{6\left(p+1\right)\left(m\text{-}1\right)}\left(\sum_{i=1}^{m}\frac{1}{n_i}\text{-}\frac{1}{N}\right)$$

The (1-$\alpha$)th of Chi-squared distribution with degrees of freedom:

$$\nu=\frac{1}{2}p\ (p+1)(m\text{-}1)$$

where, p is the number of variables. Box M-statistic is improved Likelihood Ratio Test (LRT) constructed under the multivariate normality distribution (Box, 1949). Also, this test is constrained under another two assumptions which are the sample covariance matrices are independent and the sample size, n must be larger than the number of variables, p (Sharif, 2013).

Nevertheless, in practice, data that meet the assumption of normality is difficult to be found. Fail to fulfil the assumption of normality can distort type 1 error rates (Yusof *et al.*, 2013) and make distributional behaviour totally fails (Aslam and Rocke, 2005). Consequently, this statistic is also highly sensitive to the existence of outliers which can cause unacceptable results. Thus, a common recommendation is to use nonparametric test or performing simple transformation.

However, non-parametric test is less powerful if it compared to parametric test. Non-parametric test is call for large sample size to reject the null hypothesis and its computation is tedious and laborious (Daniel, 1990). Otherwise, simple transformation is one way to overcome the problem of outliers. Nevertheless, as identified by Wilcox (2005), simple transformations are failed to treat the outliers with efficiently. As a result, the outliers still exist and minimize the statistical power when applying simple transformations.

Moreover, there are another two alternative methods that can be used in to reduce the effect of outliers. The first method is to calculate the classical estimator after eliminating outliers from the data. The second method is by using robust estimator to replace classical estimator in decreasing the influence of outliers (Yahaya *et al.*, 2011; Yuan and Bentler, 2001). The robust method is aim to produce reliable parameters estimate, related tests and confidence intervals, even though data follow a given distribution correctly but conversely, only approximately in the sense would be qualified (Maronna *et al.*, 2006). Furthermore, it is vital tools in analysing data that are including a contaminated observation (Muthukrishnan and Ravi, 2016). It can be used to identify outliers and to deliver resistant results in the existence of outliers. Thus, robust method attempts to deliver a good result and therefore, would be interested in this study.

There are a lot of multivariate robust estimators of location and scatters widely have been used in previous research. The major researchs started by Huber (1964) who introduced M-estimator and followed by S-estimator (Rousseeuw and Yohai, 1984). These two estimators received much attention by several researchers. M-estimator is shown to be reliable and asymptotically normal under their assumptions (Maronna, 1976). This estimator has a good local robustness properties where it contains of good efficiencies and good bound on the influence function at underlying distribution (Sirkia *et al.*, 2007). S-estimator is suggested for the purpose of minimizing the determinant of variance-covariance matrix. The benefits of S-estimator contain fast computation (Kondo *et al.*, 2012) can accomplish an efficiency up to 33% (Croux *et al.*, 1994) and highly resistant to outliers which is an able to produce the same values as the usual analysis when there is no outliers (Aslam and Rocke, 2005).

Entertainingly, these two estimators have high Breakdown Point (BP) of approximately 50% (Lopuhaa, 1989). Actually, they have the same asymptotic properties (Onur and Cetin, 2011), the calculation is easier compared to other robust estimators (Jeng, 2010; Salibian-Barrera and Yohai, 2006). Accordingly, to accomplish a very high robust quality, M-estimator and S-estimator will be examined for the substitution of the sample covariance matrix.

## MATERIALS AND METHODS

**Overview of modified Box M-statistic:** In this study, a discussion on robust Box M-statistic based on M and S-estimators are presented to accomplished the modification proses.

**Robust Box Mstatistic based on M-estimator:** M-estimator was the early robust estimator for sample mean vector and sample covariance matrix of location and scatter parameters. Based on the ideas of Huber (1964), the univariate M-estimators act as minimizers of objective functions, meanwhile, Maronna (1976) presented that multivariate M-estimators as the results of $\hat{\mu}$ and a positive definite symmetric matrix, $\hat{\Sigma}$ of:

$$\sum_{i=1}^{n} u_1(d_i)(x_i - \hat{u}) = 0$$

$$\sum_{1=1}^{n} u_2(d_i)^2 (x_i - \hat{u})(x_i - \hat{u}) = \hat{\Sigma}$$

Where:

$u_i$ = Weight functions where i = 1, 2
$d_i$ = Distance of $\sqrt{(x_i-\mu)^t \sum^{-1} (x_i-\mu)}$
$\hat{\Sigma}$ = Weighted variance-covariance matrix

The robustness properties of M-estimator are existence, consistency, asymptotic normality, uniqueness, BP and influence function. Maronna (1976) showed that M-estimator is consistent, asymptotically normal and the BP is <1/p+1 where p denotes the number of variable.

In order to improve robust Box M-statistic denoted by $M_M$, the covariance of M estimator, $S_{M(i)}$ where, i = 1, 2, ..., m is substituted into Eq. 1. Thus, the statistic is now turn into the next equation:

$$M_M = N \ln \left|\bar{S}_M\right| - \sum_{i=1}^{m} n_i \ln \left|S_{M(i)}\right| \quad (2)$$

Where:

$$\bar{S}_M = \frac{1}{N} \sum_{i=1}^{m} n_i S_{M(i)}$$

the pooled sample covariance matrix of M-estimator

**Robust M-statistic based on S-estimator:** Let $\rho: \Re^+ \rightarrow \Re^H$ be a continuously differentiable, symmetric, non decreasing function which has $\rho(0)$ and is constant at $\rho(x) = \rho(c)$ for all $x \geq c$. The data set of n observation in $\Re^p$ the S-estimator $(\tilde{\mu}, \tilde{\Sigma})$ is defined to minimize $|\tilde{\Sigma}|$ as follows:

$$n^{-1} \sum_{i} \rho(d_i) = b_0$$

where, $d_i = \sqrt{(x_i-\mu)^t \sum^{-1} (x_i-\mu)}$. Meanwhile, the constant of $b_0$ is the predictable value from $\rho(d)$. Specifically:

$$b_0 = \frac{\nu\chi^2(\nu+2;c^2)}{2} - \frac{\nu(\nu+2)\chi^2(\nu+4;c^2)}{2c^2} + \frac{\nu(\nu+2)(\nu+4)\chi^2(\nu+6;c^2)}{6c^4} + \frac{c^2}{6}\left[1-\chi^2(\nu,c^2)\right]$$

where, $\chi^2(\nu;\ c^2)$ is denoted as the cumulative distribution for a $\chi^2$ variable on $\nu$ degrees of freedom.

Additional, the choice of c depends on the preferred BP for the estimate (Aslam and Rocke, 2005; Campbell *et al*., 1995; Rousseeuw and Yohai, 1984). Aslam and Rocke (2005) have presented translated biweight (t-biweight) which delivers the lowest sensitivity to outliers for a given BP. The t-biweight function (a, b) is as follows:

$$\rho(d_i) = \begin{cases} \frac{d_i^2}{2} - \frac{d_i^4}{2c^2} + \frac{d_i^6}{6c^4} |di| \leq c \\ \frac{c^2}{6} |d_i| \geq c \end{cases}$$

High BP and equally good behaviour with uncontaminated data sets are the basic elements for a good robust estimator (Mili and Coakley, 1996; Aelst and Willems, 2005). According to Lopuhaa (1989), the BP (known as the percentage of outliers) in the sample that an estimator can dealing with is approximately 50%. However, Sakata and White (1998) stated that the BP of an estimator may take different values depending on the contaminated data at which it is evaluated.

Let $M_S$ represents as the robust Box M-statistic based on S-estimator. The covariance of S-estimator, $S_{S(i)}$ where, i = 1, 2, ..., m is used and introduced into Eq. 1. Thus, the statistic for $M_S$ is as follows:

$$M_S = N \ln \left|\bar{S}_S\right| - \sum_{i=1}^{m} n_i \ln \left|S_{S(i)}\right| \quad (3)$$

Where:

$$\bar{S}_S = \frac{1}{N} \sum_{i=1}^{m} n_i S_{S(i)}$$

the pooled sample covariance matrix of S-estimator.

## RESULTS AND DISCUSSION

**Evaluation of the proposed Box M-statistic:** The performance of robust estimator isassessed using type 1 error ($\alpha$) rate. The type 1 error also known as false alarm rate is used to compare the performance of $M_M$ and $M_S$. This error can be denotes as the probability that the null hypothesis is rejected when it is true. Generally, the significance levels are set comparatively low at 0.01, 0.05 or 0.10. The smaller the value of $\alpha$, the more confidence that $H_0$ is really false when it has been identified. The type 1 error value becomes larger if the procedure is unstable due to increasing in variability. Inflated false alarm rate can lead to unnecessary procedure modifications and loss of confidence in the control chart as a monitoring tool (Chang and Bai, 2001). Hence, a method which can control the false alarm rate at the

Table 1: Type 1 error rate for variable, p = 3

| | n = 10 | | | n = 20 | | | n = 30 | | | n = 40 | | | n = 50 | | | n = 100 | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| ε/μ | M | $M_M$ | $M_S$ | M | $M_M$ | $M_S$ | M | $M_M$ | $M_S$ | M | $M_M$ | $M_S$ | M | $M_M$ | $M_S$ | M | $M_M$ | $M_S$ |
| **0** | | | | | | | | | | | | | | | | | | |
| 0 | 0.0454 | 0.0459 | 0.0499 | 0.0443 | 0.0478 | 0.05610 | 0.0475 | 0.0471 | 0.0499 | 0.0438 | 0.0527 | 0.0473 | 0.0496 | 0.0492 | 0.0457 | 0.0478 | 0.0455 | 0.04990 |
| 3 | 0.0438 | 0.1594 | 0.0493 | 0.0505 | 0.2394 | 0.05340 | 0.0458 | 0.2164 | 0.0493 | 0.0471 | 0.2076 | 0.0473 | 0.0486 | 0.2394 | 0.0302 | 0.0294 | 0.1084 | 0.05090 |
| 5 | 0.0497 | 0.1799 | 0.0511 | 0.0510 | 0.2449 | 0.05590 | 0.0477 | 0.2395 | 0.0462 | 0.0459 | 0.2255 | 0.0473 | 0.0459 | 0.2449 | 0.0262 | 0.0259 | 0.1502 | 0.04840 |
| **0.05** | | | | | | | | | | | | | | | | | | |
| 3 | 0.1373 | 0.2502 | 0.0635 | 0.0354 | 0.3376 | 0.04590 | 0.0353 | 0.2764 | 0.0596 | 0.0363 | 0.2612 | 0.0468 | 0.0353 | 0.3498 | 0.0670 | 0.3376 | 0.2728 | 0.03260 |
| 5 | 0.1397 | 0.2646 | 0.1083 | 0.0361 | 0.3369 | 0.04324 | 0.0367 | 0.2995 | 0.0408 | 0.0343 | 0.2832 | 0.0494 | 0.0317 | 0.4590 | 0.0378 | 0.3369 | 0.2598 | 0.03398 |
| **0.1** | | | | | | | | | | | | | | | | | | |
| 3 | 0.0382 | 0.2801 | 0.0458 | 0.3780 | 0.2564 | 0.04640 | 0.0368 | 0.2964 | 0.0459 | 0.0396 | 0.2891 | 0.0272 | 0.0299 | 0.5682 | 0.0425 | 0.2564 | 0.2460 | 0.04310 |
| 5 | 0.0350 | 0.2988 | 0.053 | 0.3687 | 0.2827 | 0.04890 | 0.3670 | 0.3827 | 0.0457 | 0.3589 | 0.3572 | 0.0498 | 0.0322 | 0.6774 | 0.0584 | 0.2827 | 0.2659 | 0.03900 |
| **0.15** | | | | | | | | | | | | | | | | | | |
| 3 | 0.0363 | 0.2955 | 0.0430 | 0.0378 | 0.2964 | 0.04620 | 0.3250 | 0.3376 | 0.0454 | 0.3641 | 0.3342 | 0.0595 | 0.0305 | 0.7866 | 0.0284 | 0.2964 | 0.2527 | 0.03010 |
| 5 | 0.0371 | 0.3076 | 0.0457 | 0.0366 | 0.2995 | 0.04710 | 0.0348 | 0.3369 | 0.0446 | 0.0359 | 0.3312 | 0.0379 | 0.0319 | 0.8958 | 0.0299 | 0.2995 | 0.2766 | 0.02950 |
| **0.2** | | | | | | | | | | | | | | | | | | |
| 3 | 0.0370 | 0.2967 | 0.0422 | 0.0398 | 0.2121 | 0.04930 | 0.0342 | 0.3598 | 0.0437 | 0.0332 | 0.3601 | 0.0369 | 0.0284 | 0.5230 | 0.0293 | 0.2121 | 0.2609 | 0.03640 |
| 5 | 0.0386 | 0.2817 | 0.0437 | 0.0365 | 0.2096 | 0.05030 | 0.0348 | 0.3477 | 0.0419 | 0.0324 | 0.305 | 0.0386 | 0.0293 | 0.2310 | 0.0363 | 0.2096 | 0.2784 | 0.03806 |

Shaded  region indicate type 1 error within [0.025, 0.075]

Table 2: Type 1 error rate for variable, p = 5

| | n = 10 | | | n = 20 | | | n = 30 | | | n = 40 | | | n = 50 | | | n = 100 | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| ε/μ | M | $M_M$ | $M_S$ | M | $M_M$ | $M_S$ | M | $M_M$ | $M_S$ | M | $M_M$ | $M_S$ | M | $M_M$ | $M_S$ | M | $M_M$ | $M_S$ |
| **0** | | | | | | | | | | | | | | | | | | |
| 0 | 0.0454 | 0.0527 | 0.0499 | 0.0443 | 0.0471 | 0.0561 | 0.0475 | 0.0459 | 0.0499 | 0.0438 | 0.0455 | 0.0473 | 0.0489 | 0.0423 | 0.0539 | 0.0505 | 0.0510 | 0.0525 |
| 3 | 0.0438 | 0.2076 | 0.0493 | 0.0505 | 0.2164 | 0.0534 | 0.0458 | 0.1594 | 0.0493 | 0.0471 | 0.1084 | 0.0473 | 0.0467 | 0.2612 | 0.0497 | 0.4660 | 0.3376 | 0.0433 |
| 5 | 0.0497 | 0.2255 | 0.0511 | 0.0510 | 0.2395 | 0.0559 | 0.0477 | 0.1799 | 0.0462 | 0.0459 | 0.1502 | 0.0473 | 0.0475 | 0.2832 | 0.0514 | 0.4670 | 0.3369 | 0.0481 |
| **0.05** | | | | | | | | | | | | | | | | | | |
| 3 | 0.0373 | 0.2612 | 0.0635 | 0.0354 | 0.2764 | 0.1039 | 0.0353 | 0.2502 | 0.0967 | 0.0363 | 0.2728 | 0.0468 | 0.0417 | 0.3342 | 0.1071 | 0.0349 | 0.2564 | 0.1342 |
| 5 | 0.0397 | 0.2832 | 0.1083 | 0.0361 | 0.2995 | 0.3324 | 0.0367 | 0.2646 | 0.4082 | 0.0343 | 0.2598 | 0.0948 | 0.0377 | 0.3312 | 0.4578 | 0.0364 | 0.2827 | 0.5006 |
| **0.1** | | | | | | | | | | | | | | | | | | |
| 3 | 0.0382 | 0.2891 | 0.0458 | 0.0378 | 0.2964 | 0.0464 | 0.0368 | 0.2801 | 0.0459 | 0.0396 | 0.2460 | 0.2727 | 0.0374 | 0.3601 | 0.0431 | 0.0358 | 0.2964 | 0.0529 |
| 5 | 0.0350 | 0.3572 | 0.0530 | 0.0360 | 0.3827 | 0.0489 | 0.0367 | 0.2988 | 0.0457 | 0.0358 | 0.2659 | 0.0498 | 0.0369 | 0.3050 | 0.0500 | 0.0353 | 0.2995 | 0.0607 |
| **0.15** | | | | | | | | | | | | | | | | | | |
| 3 | 0.0363 | 0.3342 | 0.0430 | 0.0378 | 0.3376 | 0.0462 | 0.0325 | 0.2955 | 0.0454 | 0.0364 | 0.2527 | 0.0954 | 0.0385 | 0.2527 | 0.0434 | 0.0360 | 0.2801 | 0.0374 |
| 5 | 0.0371 | 0.3312 | 0.0457 | 0.0366 | 0.3369 | 0.0471 | 0.0348 | 0.3076 | 0.0446 | 0.0359 | 0.2766 | 0.0379 | 0.0345 | 0.2766 | 0.0411 | 0.0347 | 0.2988 | 0.0386 |
| **0.2** | | | | | | | | | | | | | | | | | | |
| 3 | 0.0370 | 0.3601 | 0.0422 | 0.0398 | 0.3598 | 0.0493 | 0.0342 | 0.2967 | 0.0437 | 0.0332 | 0.2609 | 0.0369 | 0.0371 | 0.2609 | 0.0456 | 0.0349 | 0.2955 | 0.0397 |
| 5 | 0.0386 | 0.3050 | 0.0437 | 0.0365 | 0.3477 | 0.0503 | 0.0348 | 0.2817 | 0.0419 | 0.0324 | 0.2784 | 0.0386 | 0.0349 | 0.2784 | 0.0407 | 0.0334 | 0.3076 | 0.0357 |

Shaded region indicate type 1 error within [0.025, 0.075]

desired level is necessary to be examined. In this research, the simulation is performed using MATLAB 7.8.0 (R2009a) with 10000 repetitions at significance level, $\alpha = 0.05$ and the contaminated data ranging from $\xi = 0$, 5, 10, 15 and 20%. The data set consists of different number of variables which are small ($p = 3$ and 5), medium $p = 10$ and 15) and large $p = 20$ and 30) as well as different size of sample, n = 5, 10, 20, 30, 40, 50 and 100.

If the observations of $M_M/b$ and $M_S/b$, respectively are rejected by the critical value, it would be considering as outliers and recommended to remove from the data set.

To examine the effect of outliers on the statistic performance we have considered a contaminated model used Alfaro and Ortega (2009) as follows:

$$(1\text{-}\xi)MVN_p(0, I_p)+\xi MVN_p(\mu, I_p) \tag{4}$$

Where:

$\xi$ = The proportion of contamination data
$\mu$ = The shift in mean
$I_p$ = The identity matrix

For the purpose of comparison and checking on the robust level the The Bradley's criterion of robustness is used as a reference. We can consider a procedure robust if its empirical value of type 1 error is between $0.5\alpha \leq \hat{\alpha} \leq 1.5\alpha$. Furthermore, the closer the value to $\alpha$, the more robust is the statistic. Therefore, when the significance level is set at $\alpha = 0.05$, a statistic is robust when the type 1 error is lies between 0.025 and 0.075. Otherwise, a statistic is considered to be non-robust. The values that closest to the significance level and within the 0.025 and 0.075 are shaded in the tables.

Table 1-6 recorded the type I error for each condition are arranged based on the ascending number of variables, namely small ($p = 3$ and 5), medium ($p = 10$ and 15) and large number variables ($p = 20$ and 30) with $\alpha = 0.05$. The first column in each table displays the percentage of outliers ($\xi$) and followed by shift of mean ($\mu$). The following three columns record the type 1 error rates of the M-statistic, MM-statistic and $M_S$ statistics investigated in this study. This situation is repeated for different sample sizes.

In Table 1 and 2, there are 198 conditions involved in assessing the robustness of statistic for small number of variables ($p = 3$ and 5). There are 50 out of 198 condition of Box M-statistic, 65 condition of $M_S$ statistic and 6 of $M_M$ fall within the robust interval. However, when $p = 5$,

Table 3: Type I error rate for variable, p = 10

| | n = 20 | | | n = 30 | | | n = 40 | | | n = 50 | | | n = 100 | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| ε/μ | M | $M_M$ | $M_s$ | M | $M_M$ | $M_s$ | M | $M_M$ | $M_s$ | M | $M_M$ | $M_s$ | M | $M_M$ | $M_s$ |
| **0** | | | | | | | | | | | | | | | |
| 0 | 0.0443 | 0.0471 | 0.0561 | 0.04750 | 0.0459 | 0.0498 | 0.0438 | 0.0455 | 0.0473 | 0.0489 | 0.0500 | 0.0539 | 0.0505 | 0.0550 | 0.0525 |
| 3 | 0.0505 | 0.0560 | 0.0521 | 0.04090 | 0.0594 | 0.0478 | 0.0471 | 0.1178 | 0.0469 | 0.0477 | 0.2076 | 0.0497 | 0.0460 | 0.2907 | 0.0430 |
| 5 | 0.0521 | 0.0356 | 0.0560 | 0.04547 | 0.0799 | 0.0462 | 0.0465 | 0.1502 | 0.0425 | 0.0493 | 0.2255 | 0.0524 | 0.0407 | 0.2898 | 0.0443 |
| **0.05** | | | | | | | | | | | | | | | |
| 3 | 0.0394 | 0.2764 | 0.1039 | 0.1343 | 0.2502 | 0.0977 | 0.0366 | 0.2754 | 0.0488 | 0.0421 | 0.2612 | 0.0307 | 0.0354 | 0.3472 | 0.0322 |
| 5 | 0.0321 | 0.2995 | 0.3324 | 0.1567 | 0.2646 | 0.4802 | 0.0344 | 0.2598 | 0.0548 | 0.0377 | 0.2732 | 0.0408 | 0.0364 | 0.3242 | 0.0562 |
| **0.1** | | | | | | | | | | | | | | | |
| 3 | 0.0378 | 0.2964 | 0.0464 | 0.0378 | 0.2871 | 0.0450 | 0.0366 | 0.2136 | 0.0772 | 0.0790 | 0.2891 | 0.0402 | 0.1358 | 0.3112 | 0.0529 |
| 5 | 0.0343 | 0.3898 | 0.0489 | 0.0393 | 0.2988 | 0.0475 | 0.0358 | 0.2659 | 0.0498 | 0.0769 | 0.3562 | 0.0550 | 0.1353 | 0.3109 | 0.0671 |
| **0.15** | | | | | | | | | | | | | | | |
| 3 | 0.3718 | 0.3466 | 0.0462 | 0.0925 | 0.2655 | 0.0455 | 0.0964 | 0.2527 | 0.0540 | 0.0850 | 0.3980 | 0.0452 | 0.3612 | 0.3452 | 0.0374 |
| 5 | 0.3666 | 0.3139 | 0.0471 | 0.0948 | 0.3706 | 0.0485 | 0.0859 | 0.2766 | 0.0380 | 0.0845 | 0.2289 | 0.0401 | 0.3475 | 0.2987 | 0.0586 |
| **0.2** | | | | | | | | | | | | | | | |
| 3 | 0.1398 | 0.3598 | 0.0493 | 0.1378 | 0.2967 | 0.0463 | 0.0932 | 0.2690 | 0.0369 | 0.371 | 0.3210 | 0.0455 | 0.3492 | 0.2760 | 0.0399 |
| 5 | 0.1365 | 0.3747 | 0.0503 | 0.1348 | 0.2817 | 0.0420 | 0.0924 | 0.2584 | 0.0386 | 0.349 | 0.3789 | 0.0471 | 0.3341 | 0.3219 | 0.0557 |

Table 4: Type 1 error rate for variable, p = 15

| | n = 20 | | | n = 30 | | | n = 40 | | | n = 50 | | | n = 100 | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| ε/μ | M | $M_M$ | $M_s$ | M | $M_M$ | $M_s$ | M | $M_M$ | $M_s$ | M | $M_M$ | $M_s$ | M | $M_M$ | $M_s$ |
| **0** | | | | | | | | | | | | | | | |
| 0 | 0.0443 | 0.0471 | 0.0561 | 0.0475 | 0.0459 | 0.0499 | 0.0438 | 0.0455 | 0.0473 | 0.0489 | 0.0527 | 0.0539 | 0.0505 | 0.0527 | 0.0525 |
| 3 | 0.0505 | 0.064 | 0.0534 | 0.0458 | 0.1594 | 0.0493 | 0.0471 | 0.1084 | 0.0429 | 0.0467 | 0.3076 | 0.0497 | 0.0466 | 0.2976 | 0.0433 |
| 5 | 0.051 | 0.0395 | 0.0559 | 0.0477 | 0.1799 | 0.0462 | 0.0459 | 0.1502 | 0.0431 | 0.0475 | 0.3255 | 0.0514 | 0.0467 | 0.2455 | 0.0481 |
| **0.05** | | | | | | | | | | | | | | | |
| 3 | 0.0354 | 0.2764 | 0.1039 | 0.0353 | 0.2502 | 0.0967 | 0.0363 | 0.2728 | 0.0468 | 0.0417 | 0.3612 | 0.047 | 0.0349 | 0.2612 | 0.0342 |
| 5 | 0.0361 | 0.2995 | 0.3324 | 0.0367 | 0.2646 | 0.4082 | 0.0343 | 0.2598 | 0.0481 | 0.0377 | 0.3832 | 0.0478 | 0.0364 | 0.2732 | 0.0506 |
| **0.1** | | | | | | | | | | | | | | | |
| 3 | 0.0378 | 0.2964 | 0.1464 | 0.0368 | 0.2801 | 0.0459 | 0.0396 | 0.246 | 0.0727 | 0.1374 | 0.2891 | 0.0431 | 0.3518 | 0.2898 | 0.0529 |
| 5 | 0.036 | 0.3827 | 0.1489 | 0.0367 | 0.2988 | 0.0457 | 0.0358 | 0.2659 | 0.0498 | 0.1369 | 0.3572 | 0.0500 | 0.3536 | 0.3472 | 0.0607 |
| **0.15** | | | | | | | | | | | | | | | |
| 3 | 0.0378 | 0.3376 | 0.0462 | 0.0325 | 0.2955 | 0.0454 | 0.3641 | 0.2527 | 0.054 | 0.2385 | 0.3342 | 0.0434 | 0.136 | 0.3242 | 0.0374 |
| 5 | 0.0366 | 0.3369 | 0.0471 | 0.0348 | 0.3076 | 0.0446 | 0.3591 | 0.2766 | 0.0379 | 0.2345 | 0.3312 | 0.0411 | 0.1347 | 0.3112 | 0.0386 |
| **0.2** | | | | | | | | | | | | | | | |
| 3 | 0.0398 | 0.3598 | 0.0493 | 0.0342 | 0.2967 | 0.0437 | 0.3322 | 0.2609 | 0.0369 | 0.1371 | 0.3601 | 0.0456 | 0.2349 | 0.3109 | 0.0397 |
| 5 | 0.0365 | 0.3477 | 0.0503 | 0.0348 | 0.2817 | 0.0419 | 0.3242 | 0.2784 | 0.0386 | 0.1349 | 0.315 | 0.0407 | 0.2334 | 0.3452 | 0.0357 |

Table 5: Type 1 error rate for variable, p = 20

| | n = 40 | | | n = 50 | | | n = 100 | | |
|---|---|---|---|---|---|---|---|---|---|
| ε/μ | M | $M_M$ | $M_s$ | M | $M_M$ | $M_s$ | M | $M_M$ | $M_s$ |
| **0** | | | | | | | | | |
| 0 | 0.0438 | 0.0455 | 0.0473 | 0.0489 | 0.0566 | 0.0539 | 0.0505 | 0.0521 | 0.0525 |
| 3 | 0.0471 | 0.1084 | 0.0473 | 0.0467 | 0.2728 | 0.0497 | 0.0466 | 0.2189 | 0.0423 |
| 5 | 0.0459 | 0.1502 | 0.0473 | 0.0475 | 0.2598 | 0.0514 | 0.0467 | 0.2276 | 0.0481 |
| **0.05** | | | | | | | | | |
| 3 | 0.0633 | 0.2728 | 0.0468 | 0.0417 | 0.2967 | 0.0397 | 0.0349 | 0.2177 | 0.0342 |
| 5 | 0.2431 | 0.2598 | 0.0948 | 0.3377 | 0.2817 | 0.0457 | 0.0364 | 0.2955 | 0.0506 |
| **0.1** | | | | | | | | | |
| 3 | 0.3963 | 0.2460 | 0.2727 | 0.3742 | 0.2801 | 0.0431 | 0.0358 | 0.3076 | 0.0529 |
| 5 | 0.3581 | 0.2659 | 0.0498 | 0.0369 | 0.2988 | 0.0500 | 0.0353 | 0.2967 | 0.0607 |
| **0.15** | | | | | | | | | |
| 3 | 0.0364 | 0.2527 | 0.0954 | 0.0385 | 0.2609 | 0.0434 | 0.2360 | 0.2817 | 0.0347 |
| 5 | 0.0359 | 0.2766 | 0.0379 | 0.0345 | 0.2784 | 0.0411 | 0.2347 | 0.2385 | 0.0386 |
| **0.2** | | | | | | | | | |
| 3 | 0.0332 | 0.2609 | 0.0369 | 0.0371 | 0.0967 | 0.0456 | 0.3490 | 0.2345 | 0.0377 |
| 5 | 0.0324 | 0.2784 | 0.0386 | 0.0349 | 0.4082 | 0.0407 | 0.3534 | 0.1371 | 0.0357 |

Shaded region indicate type 1 error within [0.025, 0.075]

Table 6: Type 1 error rate for variable, p = 30

| | n = 40 | | | n = 50 | | | n = 100 | | |
|---|---|---|---|---|---|---|---|---|---|
| ε/μ | M | $M_M$ | $M_s$ | M | $M_M$ | $M_s$ | M | $M_M$ | $M_s$ |
| **0** | | | | | | | | | |
| 0 | 0.0524 | 0.0500 | 0.0499 | 0.0506 | 0.0499 | 0.0525 | 0.0456 | 0.0478 | 0.0525 |
| 3 | 0.0600 | 0.1928 | 0.0493 | 0.0496 | 0.1084 | 0.0333 | 0.0486 | 0.2659 | 0.0413 |
| 5 | 0.0586 | 0.1429 | 0.0462 | 0.0498 | 0.1502 | 0.0481 | 0.0422 | 0.2527 | 0.0461 |
| **0.05** | | | | | | | | | |
| 3 | 0.0552 | 0.3995 | 0.0967 | 0.0484 | 0.2728 | 0.0342 | 0.0384 | 0.2766 | 0.0342 |
| 5 | 0.0530 | 0.3843 | 0.4082 | 0.0496 | 0.2598 | 0.0426 | 0.0414 | 0.2609 | 0.0406 |
| **0.1** | | | | | | | | | |
| 3 | 0.5945 | 0.2298 | 0.0459 | 0.4887 | 0.2460 | 0.0498 | 0.4827 | 0.2784 | 0.0529 |
| 5 | 0.5241 | 0.2054 | 0.0457 | 0.4912 | 0.2659 | 0.0607 | 0.4312 | 0.3995 | 0.0547 |

Table 5: Continue

|  | n = 40 |  |  | n = 50 |  |  | n = 100 |  |  |
|---|---|---|---|---|---|---|---|---|---|
| ε/μ | M | $M_M$ | $M_s$ | M | $M_M$ | $M_s$ | M | $M_M$ | $M_s$ |
| **0.15** |  |  |  |  |  |  |  |  |  |
| 3 | 0.5478 | 0.2706 | 0.0454 | 0.4685 | 0.2527 | 0.0374 | 0.4786 | 0.3843 | 0.0374 |
| 5 | 0.5223 | 0.2634 | 0.0446 | 0.4845 | 0.2766 | 0.0386 | 0.4702 | 0.2298 | 0.0386 |
| **0.2** |  |  |  |  |  |  |  |  |  |
| 3 | 0.5165 | 0.2209 | 0.0437 | 0.4642 | 0.2609 | 0.0487 | 0.4068 | 0.2967 | 0.0397 |
| 5 | 0.529 | 0.1256 | 0.0419 | 0.4761 | 0.2784 | 0.0657 | 0.4721 | 0.2817 | 0.0357 |

Shaded region indicate type 1 error within [0.025, 0.075]

there are 65 conditions of Box M-statistic, 55 condition of $M_S$ statistic and 6 of $M_M$ fall within the robust interval. As a conclusion, the type 1 error rate for small number of variables show that $M_S$ statistic is robust when p = 3 while box M-statistic is robust when p = 5.

In Table 3 and 4, there are 165 conditions involved in evaluating the robustness of statistics for medium number of variables (p = 10 and 15). There are 33 out of 165 conditions of Box M-statistic, 51 conditions of $M_S$ statistic and 8 of $M_S$ that fall within the robust interval. Meanwhile, for $M_M$ M-statistic has 35 conditions that fall within the robust interval compared to $M_S$ statistic has only 49 conditions and 8 of $M_M$ fall within the robust interval. Thus, we concluded that $M_S$ statisticis more robust by compared to all other statistics.

When p = 20 and 30, Box M-statistic has 13 and 15 out of 99 conditions that fall within robust interval, respectively. Meanwhile, for both p, all the conditions 33 for $M_S$ statistic are fall within the robust interval. For $M_M$ has 5 and 3 fall within the robust interval. Therefore, we summarized that $M_S$ statistic is a powerful robustness in large number of variables.

## CONCLUSION

Box M-statistic is known as a test for testing two or several covariance matrices, under conditions of non-normality, this test is known to under perform. Other test statistics are suggested to produce active methods regardless of the conditions. In this research, we proposed other procedures to the box M-statistic by using a robust estimator known as the S-estimator for scatter matrix. The S-estimator has the properties such as the affine equivariant and a high BP and has a better calculation. The performance of the suggested robust test by using the S-estimator ($M_S$ and by using the M-estimator ($M_M$) was compared with the Box M-statistic in terms of the type 1 error rate. The result study showed that $M_S$ statistic performs well in terms of controlling type 1 errors.

## ACKNOWLEDGEMENTS

The researchers gratefully acknowledge University Utara Malaysia. This research is sponsored under the Research Generation University Grants SO number 13409. They also thank the reviewers for their helpful comments and suggestions.

## REFERENCES

Aelst, S.V. and G. Willems, 2005. Multivariatr regression s-estimators for robust estimation and inference. Stat. Sin., 15: 981-1001.

Alfaro, J.L. and J.F. Ortega, 2009. A comparison of robust alternatives to Hotelling's $T^2$ control chart. J. Applied Stat., 36: 1385-1396.

Aslam, S. and D.M. Rocke, 2005. A robust testing procedure for the equality of covariance matrices. Comput. Stat. Data Anal., 49: 863-874.

Box, G.E., 1949. A general distribution theory for a class of likelihood criteria. Biometrika, 36: 317-346.

Campbell, N.A., H.P. Lopuhaa and P.J. Rousseeuw, 1995. On the calculation of a robust S-estimator of a covariance matrix. Statist. Med., 17: 2685-2695.

Chang, Y.S. and D.S. Bai, 2001. Control charts for positively-skewed populations with weighted standard deviations. Qual. Reliab. Eng. Intl., 17: 397-406.

Croux, C., P.J. Rousseeuw and O. Hossjer, 1994. Generalized S-estimators. J. Am. Statist. Assoc., 89: 1271-1281.

Daniel, W.W., 1990. Applied Nonparametric Statistic. University of Michigan, Ann Arbor, Michigan, ISBN:9780534919764, Pages: 635.

Djauhari, A.M. and R.M. Salleh, 2011. Robust hotelling's T2 control charting in spike production process. Proceedings of the 2011 International Seminar on the Application of Science and Mathematics (ISASM'11), November 1-3, 2011, Putra World Trade Centre, Kuala Lumpur, Malaysia, pp: 1-8.

Djauhari, M.A., 2007. A measure of multivariate data concentration. J. Appl. Prob. Stat., 2: 139-155.

Eichholtz, P.M., 1996. Does international diversification work better for real estate than for stocks and bonds?. Financial Anal. J., 52: 56-62.

Huber, P.J., 1964. Robust estimation of location parameter. Ann. Math. Statist., 35: 73-101.

Jeng, J.C., 2010. Adaptive process monitoring using efficient recursive PCA and moving window PCA algorithms. J. Taiwan Inst. Chem. Eng., 41: 475-481.

Kondo, Y., M. Salibian-Barrera and R. Zamar, 2012. A robust and sparse K-means clustering algorithm. Mach. Learn., 1: 1-20.

Kupek, E., 2002. Bias and heteroscedastic memory error in self-reported health behavior: An investigation using covariance structure analysis. BMC. Med. Res. Method., 2: 1-14.

Lee, S., 2006. The stability of the co-movements between real estate returns in the UK. J. Property Investment Finance, 24: 434-442.

Lopuhaa, H.P., 1989. On the relation between S-estimators and M-estimators of multivariate location and covariance. Ann. Stat., 17: 1662-1683.

Mardia, K.V., J.T. Kent and J.M. Bibby, 1979. Multivariate Analysis. Probability and Mathematical Statistics. Academic Press, London, ISBN-10: 0124712509, pp: 536.

Maronna, R.A., 1976. M-estimates of multivariate location and scatter. Ann. Stat., 4: 51-67.

Maronna, R.A., R.D. Martin and V. Yohai, 2006. Robust Statistics: Theory and Methods. John Wiley & Sons, Hoboken, New Jersey, ISBN:9780470010921, Pages: 436.

Mili, L. and C.W. Coakley, 1996. Robust estimation in structured linear regression. Ann. Stat., 24: 2593-2607.

Muthukrishnan, R. and J. Ravi, 2016. A robust regression scale of residual estimator: SSAC. Intl. J. Appl. Eng. Res., 11: 5086-5090.

Onur, T.O.K.A. and M. Cetin, 2011. The comparing of s-estimator and m-estimators in linear regression. Gazi Univ. J. Sci., 24: 747-752.

Roes, K.C. and D. Dorr, 1997. Implementing statistical process control in service processes. Int. J. Q. Sci., 2: 149-166.

Rousseeuw, P. and V. Yohai, 1984. Robust Regression by Means of S-Estimators. In: Robust and Nonlinear Time Series Analysis, Franke, J., W. Hardle and D. Martin (Eds.). Springer, New York, USA., ISBN:978-0-387-96102-6, pp: 256-272.

Sakata, S. and H. White, 1998. High breakdown point conditional dispersion estimation with application to S and P 500 daily returns volatility. Econometrica, 66: 529-567.

Salibian-Barrera, M. and V.J. Yohai, 2006. A fast algorithm for S-regression estimates. J. Comput. Graphical Stat., 15: 414-427.

Sharif, S., 2013. A new statistic to the theory of correlation stability testing in finacial market. Ph.D Thesis, University of Technology, Johor Bahru, Malaysia.

Sharif, S., W.N.S.W. Yusoff, Z. Omar and S. Ismail, 2014. Computational efficiency of generalized variance and vector variance. Proceedings of the 2014 Conference on American Institute of Physics Vol. 1635, December 15, 2014, AIP, Maryland, USA., pp: 906-911.

Sirkia, S., S. Taskinen and H. Oja, 2007. Symmetrised M-estimators of multivariate scatter. J. Multivariate Anal., 98: 1611-1629.

Tang, G.Y.N., 1998. The intertemporal stability of the covariance and correlation matrices of Hong Kong stock returns. Applied Financial Econ., 8: 359-365.

Wilcox, R.R., 2005. Introduction to Robust Estimation and Hypothesis Testing. 2nd Edn., Elsevier, Amsterdam, Netherlands, ISBN:0-12-751542-9, Pages: 587.

Yahaya, S.S.S., H. Ali and Z. Omar, 2011. An alternative hotelling $T^2$ control chart based on Minimum Vector Variance (MVV). Mod. Applied Sci., 5: 132-151.

Yuan, K.H. and P.M. Bentler, 2001. Effect of outliers on estimators and tests in covariance structure analysis. Br. J. Math. Stat. Psychol., 54: 161-175.

Yusof, Z.M., N.H. Harun, S.S.S. Yahaya and S. Abdullah, 2013. Modified parametric bootstrap: A robust alternative to classical test. Proceeding of the World Conference on Integration of Knowledge (WCIK'13), November 25-26, 2013, Langkawi International College, Langkawi, Malaysia, ISBN:978-967-11768-2-5, pp: 1-7.

Yusoff, N.S. and M.A. Djauhari, 2012. A correlation network approach to analyze the influence of geothermal environment on porcelain insulator. Intl. J. Appl. Eng. Res., 7: 263-275.

Yusoff, N.S. and M.A. Djauhari, 2013. A statistical test for the stability of covariance structure. J. Technol., 63: 81-83.