Mathematics and Statistics 9(5): 653-663, 2021 http://www.hrpub.org
DOI: 10.13189/ms.2021.090505

Robust Multivariate Location Estimation in the Existence of Casewise and Cellwise Outliers

Yik-Siong Pang*, Nor Aishah Ahad, Sharipah Soaad Syed Yahaya

Department of Mathematics & Statistics, School of Quantitative Sciences, Universiti Utara Malaysia, 06010 UUM Sintok, Kedah, Malaysia

Received November 18, 2020; Revised June 16, 2021; Accepted July 5, 2021

Cite This Paper in the following Citation Styles

(a): *[1] Yik-Siong Pang, Nor Aishah Ahad, Sharipah Soaad Syed Yahaya , "Robust Multivariate Location Estimation in the Existence of Casewise and Cellwise Outliers," Mathematics and Statistics, Vol. 9, No. 5, pp. 653 - 663, 2021. DOI: 10.13189/ms.2021.090505.*

(b): *Yik-Siong Pang, Nor Aishah Ahad, Sharipah Soaad Syed Yahaya (2021). Robust Multivariate Location Estimation in the Existence of Casewise and Cellwise Outliers. Mathematics and Statistics, 9(5), 653 - 663. DOI: 10.13189/ms.2021.090505.*

Abstract Multivariate outliers can exist in two forms, casewise and cellwise. Data collection typically contains unknown proportion and types of outliers which can jeopardize the location estimation and affect research findings. In cases where the two coexist in the same data set, traditional distance-based trimmed mean and coordinate-wise trimmed mean are unable to perform well in estimating location measurement. Distance-based trimmed mean suffers from leftover cellwise outliers after the trimming whereas coordinate-wise trimmed mean is affected by extra casewise outliers. Thus, this paper proposes new robust multivariate location estimation known as α-distance-based trimmed median ($\widehat{\mathbf{M}}_{(\mathbf{MSD},\boldsymbol{\alpha})}$) to deal with both types of outliers simultaneously in a data set. Simulated data were used to illustrate the feasibility of the new procedure by comparing with the classical mean, classical median and α-distance-based trimmed mean. Undeniably, the classical mean performed the best when dealing with clean data, but contrarily on contaminated data. Meanwhile, classical median outperformed distance-based trimmed mean when dealing with both casewise and cellwise outliers, but still affected by the combined outliers' effect. Based on the simulation results, the proposed $\widehat{\mathbf{M}}_{(\mathbf{MSD},\boldsymbol{\alpha})}$ yields better location estimation on contaminated data compared to the other three estimators considered in this paper. Thus, the proposed $\widehat{\mathbf{M}}_{(\mathbf{MSD},\boldsymbol{\alpha})}$ can mitigate the issues of outliers and provide a better location estimation.

Keywords Multivariate Outliers, Robust Location Estimation, Distance-Based Trimming, Trimmed Mean, Trimmed Median

1. Introduction

1.1. Multivariate Outliers

Most parametric statistical tools were derived using the population parameters mean ($\boldsymbol{\mu}$) and (co)variance ($\boldsymbol{\Sigma}$), but most of the time, these population parameters are unknown. Consequently, these parameters are usually estimated from the sample mean and sample covariance, denoted as $\overline{\mathbf{X}}$ and $\mathbf{S}$ respectively. However, Hampel [1] mentioned that real data typically comprises of abnormal observations (outliers). The classical mean with 0% breakdown point is highly sensitive to outliers, even with only one outlier could divert the estimation from the supposed location and lead to the defect of the least-square-based (co)variance [2-6]. This situation is even complicated in the context of multivariate data as multivariate outliers are harder to be identified.

In multivariate context, outliers occur in a single form or can be a combination of two abnormalities within a single set of data. These outliers can be categorized as either *casewise outlier*, which refers to one whole observation as a *contaminant* to the data itself such that the observation originates from another distribution, or *cellwise outlier* which refers to *extreme values* exist within each variable of the dataset independently similar to those in univariate context [7-8]. Figure 1 illustrates a condition where both casewise and cellwise outliers coexist in a dataset.

Based on Figure 1, the z_9 and z_{10} are contaminants originated from another distribution but mistakenly included into Z-distribution. Such contaminations can happen due to sampling error or model misspecification [7], while the extreme values within the components can be some random errors. Not until recent decade, researchers began to realize the situation where both casewise and cellwise outliers present simultaneously. Nevertheless, the existing methods particularly dealt with either casewise or cellwise outliers but not both [9,10]. Thus, the main goal of this article is to introduce a more accurate location estimation method even in the presence of either one or both types of outliers.

1.2. Robust Approach - Trimming

Tukey [11] acknowledged the need of a better location estimator dealing with non-normality for Student's *t*-test and suggested the notion of trimming off samples. Trimming refers to a process of removing abnormal observations that caused a sample to be non-normally distributed. In univariate context, the most common trimming procedure for a size n sample is to decide a trimming proportion (α%) and removes $n\alpha$% observations from both end of the ordered statistics.

Generally, multivariate trimming can be done differently via coordinate-wise or distance-based approach. Coordinate-wise trimming (CT) is simple where it treats each of the components of the sample matrix as univariate vector and applies the common trimming procedures. CT is able to retain valid inputs while the outlying inputs were discarded [12]. On the contrary, distance-based approach relies on Mahalanobis Squared Distance (MSD) to determine contaminants for trimming. The common type of distance-based trimming is performed by determining a certain percentile of Chi-Squared distribution to identify multivariate outliers. However, Alloway and Raghavachavari [13] developed a straightforward MSD-based trimmed mean to construct robust Hotelling T^2 control chart by removing two observations with highest MSD values from the sample. One thing to be noted is that, the trimming procedure introduced by [13] was confined to only 2 observations trimming which can be further extended to α-percent distance-based trimmed mean ($\bar{\mathbf{X}}_{(\mathbf{MSD},\boldsymbol{\alpha})}$).

Although the high-distance observations were deemed contaminants and trimmed off, there are possibilities that cellwise outliers may still remain in the multivariate dataset. The remaining data might not be as clean as what we are expecting, for more accurate location estimation. Such situation may cause the application of mean on post-trimming data to be affected by any undetected cellwise outlier. Figure 2 illustrates the post-trimming conditions of a contaminated dataset.

According to Figure 2, the remaining extreme values still exist after employing the distance-based trimming and pose a risk in the estimation of the classical mean ($\bar{\mathbf{Z}}$). In a previous study on Robust Linear Discriminant Analysis [14], the application of median after MSD-based trimming yielded a more consistent performance than of the MSD-based trimmed mean. Therefore, a more reliable location estimator should be applied after the MSD-α%trimming procedure to alleviate the problem faced when using the highly sensitive classical mean.

$$\mathbf{Z} = \begin{bmatrix} z_1 \\ z_2 \\ z_3 \\ z_4 \\ z_5 \\ z_6 \\ z_7 \\ z_8 \\ z_9 \\ z_{10} \end{bmatrix} = \begin{bmatrix} \circ & \circ & \circ & \circ & \times & \circ \\ \circ & \circ & \times & \circ & \circ & \circ \\ \circ & \circ & \circ & \circ & \circ & \circ \\ \circ & \circ & \circ & \times & \circ & \circ \\ \circ & \circ & \circ & \circ & \circ & \circ \\ \circ & \times & \circ & \circ & \circ & \circ \\ \circ & \circ & \circ & \circ & \circ & \circ \\ \times & \circ & \circ & \circ & \circ & \times \\ \triangle & \triangle & \triangle & \triangle & \triangle & \triangle \\ \triangle & \triangle & \triangle & \triangle & \triangle & \triangle \end{bmatrix}_{10\times 6} \qquad \begin{array}{lll} \circ & = & \text{sound data} \\ \times & = & \text{extreme value} \\ \triangle & = & \text{contaminant} \end{array}$$

Figure 1. Illustration of both casewise and cellwise outliers coexists in a dataset.

$$\mathbf{Z}_{\mathbf{trimmed}} = \begin{bmatrix} z_1 \\ z_2 \\ z_3 \\ z_4 \\ z_5 \\ z_6 \\ z_7 \\ z_8 \end{bmatrix} = \begin{bmatrix} \circ & \circ & \circ & \circ & \times & \circ \\ \circ & \circ & \times & \circ & \circ & \circ \\ \circ & \circ & \circ & \circ & \circ & \circ \\ \circ & \circ & \circ & \times & \circ & \circ \\ \circ & \circ & \circ & \circ & \circ & \circ \\ \circ & \times & \circ & \circ & \circ & \circ \\ \circ & \circ & \circ & \circ & \circ & \circ \\ \times & \circ & \circ & \circ & \circ & \times \end{bmatrix}_{8\times 6} \qquad \begin{array}{lll} \circ & = & \text{sound data} \\ \times & = & \text{extreme value} \end{array}$$

Figure 2. Outliers leftover after Contaminant x_9 and x_{10} were trimmed using α-percent distance-based procedure

2. Methodology

2.1. Distance-Based Trimmed Mean ($\bar{\mathbf{X}}_{(\mathbf{MSD},\boldsymbol{\alpha})}$)

As mentioned earlier, the trimming process by [13] is rather straight forward. Observations with high MSD measurement will be trimmed off at α percentage. Supposed that $\mathbf{x}_{ij}$ comes from j-dimensional feature vectors with i=1,...,n^{th} sample vector and j=1,...,d^{th} variable dimension, the classical **MSD** is obtained via following Eq. 1 [15]:-

$$\mathbf{MSD}_{\mathrm{i}} = \left(\mathbf{x}_{ij} - \bar{\mathbf{x}}_j\right)^T \mathbf{S}^{-1}\left(\mathbf{x}_{ij} - \bar{\mathbf{x}}_j\right) \quad (1)$$

where

$\bar{\mathbf{x}}_{\mathrm{j}}$= classical mean vector of n samples of j^{th} dimension

$\mathbf{S}$ = classical covariance matrix of n samples × j dimension

After **MSD** vector containing distance measurement for each of the observation was obtained, the following trimming procedure is carried out:-

Step1: Arrange the **MSD** in ascending order;

Step2: Trim α% observations with the highest MSD;

Step3: Compute the mean for each dimension based on the remaining observations.

2.2. Distance-Based Trimmed Median ($\widehat{\mathbf{M}}_{(\mathbf{MSD},\boldsymbol{\alpha})}$)

Directly employing classical median instead of the classical mean on the original data may seem simple and straightforward, but the existence of casewise outliers may distort the estimation of location measure. Therefore, the distance-based trimming procedure extended from Alloway and Raghavachavari [13] is useful in removing casewise outliers. However, for the cases where both casewise and cellwise outliers coexist, the use of classical mean on the remaining post-trimmed data ($\mathbf{X}_{\mathbf{trimmed}}$) may still be affected by the remaining cellwise outliers in the dataset. Thus, to overcome this shortcoming of the classical mean, [14] employed median as it is more robust location estimation compared to mean.

To ensure that the outliers are well managed, a two-stage outliers filtration location estimator, known as distance-based trimmed median, is proposed. The first stage is to eliminate casewise outlier by using distance-based approach whereas the use of median estimation acts as a secondary outlier filtration for any possible outlier leftovers (casewise and/or cellwise). The same algorithm presented in Section 2.1. is used to compute the median (instead of the mean in Step 3) for each dimension of the trimmed matrix. A simulation study was conducted to compare the performance of the proposed estimator with other chosen estimators for this study.

2.3. Trimming Percentage

Trimming percentages come in various forms. Among the frequently used and suggested percentages of trimming by most researchers are symmetrical trimming of 20% to 25% [16], 20% [17-19], 15% [20], 10% to 15% [21]. For univariate data, the trimming percentage is based on each tail, thus, if the trimming percentage is 10%, the total trimming from left and right tail will be 20% from each dataset. However, when using MSD-α% trimming, the total amount of trimming is the amount of the given $n\alpha$%. Considering trimming percentages from [16-21], the one-tailed distance-based trimming percentage in this study was capped at 40%.

2.4. Simulation Settings

The Tukey-Huber Contamination Model (THCM) [22-23] was commonly employed in robust multivariate analyses to generate contaminated data for simulated study. However, the problem with THCM is that the simulation mechanism generates a two-part distribution, i.e., a "clean" part which is normally distributed data whereas the other part consists of rows of contaminated data (casewise outliers). Thus, in order to simulate both casewise and cellwise outliers, THCM was adapted and modified to randomly include cellwise outliers within the simulated data in this study.

Simulated data, matrix $\mathbf{X}_{\boldsymbol{n}\times\boldsymbol{p}}$, with sample size n and variable dimension p were generated according to Eq. 2 around pre-set Location, L_{p}, and simulated according to the settings presented in Table 1.

$$\mathbf{X}_{\boldsymbol{n}\times\boldsymbol{p}}\text{:}\ \left[(1-\varepsilon_1-\varepsilon_2)nN_p\left(L_p,\mathbf{I}_p\right)+\varepsilon_1 nN_p\left(L_p+\mu_1,\mathbf{I}_p\right)\right]^* + \varepsilon_2 nN_p(L_p+15{,}25\mathbf{I}_p) \quad (2)$$

**where $\varepsilon_1 n$ were randomly selected from clean data for each dimension p independently to perform μ_1 inflation*

Table 1. Simulation s Settings

Contamination Parameters	Controlled Setting
Pre-set Location (L_{p} for p = 1,…,6)	27, 10, 33, 21 42, 38
Data Dimensions (p)	6
Sample Size (n)	40, 100
Cellwise Outliers Proportion (ε_1)	0, 0.1
Cellwise Outliers Inflation (μ_1)	5, 10
Casewise Outliers Proportion (ε_2)	0, 0.2, 0.4
Trimming Percentage (α)	0.2, 0.3, 0.4

To determine the ability of the estimators in handling outliers, the simulation was conducted according to the following procedure:-

Step 1: Generate data according to the settings in Table 1;

Step 2: Compute the location estimators ($\bar{\mathbf{X}}$, $\widehat{\mathbf{M}}$, $\bar{\mathbf{X}}_{(\mathbf{MSD},\boldsymbol{\alpha})}$and $\widehat{\mathbf{M}}_{(\mathbf{MSD},\boldsymbol{\alpha})}$);

Step 3: Find the absolute difference of each location estimators and L_{p};

Step 4: Repeat Step 1 to Step 3 for 10,000 iterations;

Step 5: Compute the average of each location estimations and absolute difference from the 10,000 iterations.

After obtaining the average absolute difference of estimated location and pre-set location, ANOVA test was conducted to confirm significant difference between the estimators' performance and Tukey's HSD test conducted for post-hoc comparisons between the estimators.

3. Results

3.1. Non-contaminated Data ($\varepsilon_1 = 0, \varepsilon_2 = 0$)

Tables 2a and 3a show that classical mean ($\bar{\mathbf{X}}$) is indeed the optimal location estimation with the least average deviation from the pre-set location (L_p), for each dimension, $p = 1, 2, \ldots, 6$, for sample sizes of $n = 40$ and $n = 100$. On the other hand, the performance of classical median ($\hat{\mathbf{M}}$) is comparable to α-distance-based trimmed mean ($\bar{\mathbf{X}}_{(\mathbf{MSD},\boldsymbol{\alpha})}$) with $\alpha = 0.2$ or 0.3 at $n = 40$ (Table 2a) and $n = 100$ (Table 3a). By contrast, the performance of α-distance-based trimmed median ($\hat{\mathbf{M}}_{(\mathbf{MSD},\boldsymbol{\alpha})}$) in the clean data, $n = 40$ and $n = 100$, do not outperform the other estimators regardless of the trimming percentages ($\alpha = 0.2, 0.3, 0.4$).

ANOVA test shows significant difference between the estimators' performance when sample size $n = 40$ ($p = 0.000$) and $n = 100$ ($p = 0.000$). Tukey's HSD test results were obtained and sorted to rank the estimators based on the difference in |T-L| into Tables 2b and 3b. Tukey's HSD test confirms the preliminary results in Tables 2a and 3a, with $\bar{\mathbf{X}}$ being the best fit to the pre-set location L_p whereas $\hat{\mathbf{M}}_{(\mathbf{MSD},0.4)}$ being the least fit to L_p.

Table 2a. Simulation Parameter Settings (n=40, $\varepsilon_1 = 0$, $\varepsilon_2 = 0$)

Location Estimators (T)		Pre-set Location (L_p)					
		27	10	33	21	42	38
$\bar{\mathbf{X}}$	Average	27.0004	10.0013	33.0000	21.0015	41.9993	37.9994
	Ave. Abs. Diff |T-L|	0.1252	0.1266	0.1265	0.1255	0.1256	0.1264
$\hat{\mathbf{M}}$	Average	26.9998	10.0019	32.9975	21.0019	42.0004	37.9977
	Ave. Abs. Diff |T-L|	0.1548	0.1563	0.1563	0.1543	0.1550	0.1558
$\bar{\mathbf{X}}_{(\mathbf{MSD},0.4)}$	Average	27.0004	10.0018	32.9986	20.9997	42.0007	37.9992
	Ave. Abs. Diff |T-L|	0.1625	0.1625	0.1652	0.1614	0.1637	0.1629
$\bar{\mathbf{X}}_{(\mathbf{MSD},0.3)}$	Average	26.9996	10.0002	32.9995	21.0008	41.9995	37.9989
	Ave. Abs. Diff |T-L|	0.1553	0.1561	0.1570	0.1545	0.1558	0.1558
$\bar{\mathbf{X}}_{(\mathbf{MSD},0.2)}$	Average	27.0006	10.0007	32.9996	21.0012	41.9999	37.9989
	Ave. Abs. Diff |T-L|	0.1474	0.1484	0.1482	0.1467	0.1485	0.1469
$\hat{\mathbf{M}}_{(\mathbf{MSD},0.4)}$	Average	26.9993	10.0023	32.9952	21.0014	42.0008	37.9979
	Ave. Abs. Diff |T-L|	0.1927	0.1904	0.1933	0.1911	0.1917	0.1921
$\hat{\mathbf{M}}_{(\mathbf{MSD},0.3)}$	Average	26.9985	10.0008	32.9957	21.0019	42.0002	37.9979
	Ave. Abs. Diff |T-L|	0.1834	0.1818	0.1855	0.1829	0.1846	0.1829
$\hat{\mathbf{M}}_{(\mathbf{MSD},0.2)}$	Average	26.9995	10.0013	32.9970	21.0013	42.0011	37.9978
	Ave. Abs. Diff |T-L|	0.1754	0.1747	0.1760	0.1738	0.1770	0.1742

Table 2b. Estimators Ranking based on Tukey HSD Test (n=40, $\varepsilon_1 = 0$, $\varepsilon_2 = 0$)

Worst Fit							Best Fit
$\hat{\mathbf{M}}_{(\mathbf{MSD},0.4)}$	$\hat{\mathbf{M}}_{(\mathbf{MSD},0.3)}$	$\hat{\mathbf{M}}_{(\mathbf{MSD},0.2)}$	$\bar{\mathbf{X}}_{(\mathbf{MSD},0.4)}$	$\bar{\mathbf{X}}_{(\mathbf{MSD},0.3)}{}^*$	$\hat{\mathbf{M}}^*$	$\bar{\mathbf{X}}_{(\mathbf{MSD},0.2)}$	$\bar{\mathbf{X}}$

Note: All pairwise Tukey HSD results are significant at 0.05 significant level except ${}^*\bar{X}_{(MSD,0.3)}$ and $\hat{M}$ are insignificant with p-value of 0.999

Table 3a. Simulation Parameter Settings (n=100, ε_1 = 0, ε_2 = 0)

Location Estimators (T)		**Pre-set Location (L_p)**					
		27	**10**	**33**	**21**	**42**	**38**
$\bar{X}$	Average	26.9990	9.9998	32.9998	21.0000	42.0002	37.9987
	Ave. Abs. Diff |T-L|	0.0799	0.0797	0.0797	0.0804	0.0791	0.0790
$\hat{M}$	Average	26.9997	9.9998	32.9998	21.0001	41.9991	37.9984
	Ave. Abs. Diff |T-L|	0.0994	0.1001	0.0997	0.1000	0.0991	0.0998
$\bar{X}_{(MSD,0.4)}$	Average	26.9986	9.9993	32.9996	21.0014	42.0008	37.9997
	Ave. Abs. Diff |T-L|	0.1032	0.1034	0.1023	0.1035	0.1025	0.1035
$\bar{X}_{(MSD,0.3)}$	Average	26.9986	10.0001	33.0000	21.0011	42.0001	37.9998
	Ave. Abs. Diff |T-L|	0.0982	0.0989	0.0975	0.0990	0.0976	0.0989
$\bar{X}_{(MSD,0.2)}$	Average	26.9985	9.9999	33.0004	21.0006	41.9996	37.9989
	Ave. Abs. Diff |T-L|	0.0939	0.0940	0.0930	0.0943	0.0928	0.0936
$\hat{M}_{(MSD,0.4)}$	Average	26.9989	9.9989	32.9990	21.0016	41.9998	37.9990
	Ave. Abs. Diff |T-L|	0.1232	0.1242	0.1223	0.1233	0.1225	0.1234
$\hat{M}_{(MSD,0.3)}$	Average	26.9990	10.0004	32.9992	21.0007	41.9993	37.9997
	Ave. Abs. Diff |T-L|	0.1170	0.1184	0.1170	0.1178	0.1169	0.1176
$\hat{M}_{(MSD,0.2)}$	Average	26.9994	10.0004	32.9994	21.0004	41.9985	37.9988
	Ave. Abs. Diff |T-L|	0.1118	0.1130	0.1122	0.1129	0.1123	0.1118

Table 3b. Estimators Ranking based on Tukey HSD Test (n=100, ε_1 = 0, ε_2 = 0)

Worst Fit							Best Fit
$\hat{M}_{(MSD,0.4)}$	$\hat{M}_{(MSD,0.3)}$	$\hat{M}_{(MSD,0.2)}$	$\bar{X}_{(MSD,0.4)}$	$\hat{M}$	$\bar{X}_{(MSD,0.3)}$	$\bar{X}_{(MSD,0.2)}$	$\bar{X}$

Note: All pairwise Tukey HSD results are significant at 0.05 significant level

3.2. Data Contamination at $\varepsilon_1 = 0.1$ and $\varepsilon_2 = 0.2$

3.2.1. Mild Cellwise Outliers (μ_1 = 5)

As cellwise outliers and casewise outliers present in the samples, $\bar{X}$ is inevitably affected. Tables 4a and 5a show that $\bar{X}$ deviates the most from L_p compared to other estimators at both sample sizes, n = 40 and n = 100. On the other hand, $\bar{X}_{(MSD,\alpha)}$ significantly outperforms $\bar{X}$ with better accuracy as trimming percentage (α) increases. Simulation also shows that $\hat{M}$ is better than $\bar{X}$ and $\bar{X}_{(MSD,\alpha)}$ in handling outliers' influence. Meanwhile, $\hat{M}_{(MSD,\alpha)}$ illustrates a better and more stable performance compared to the other estimators.

Similarly, ANOVA test indicates significant difference amongst the estimators' performance when n = 40 (p = *0.000*) and n = 100 (p= *0.000*). The rankings according to Tukey's HSD test determine $\hat{M}_{(MSD,0.2)}$ as the best fit while $\bar{X}$ is the least fit location estimation for both sample sizes n = 40 (Table 4b) and n = 100 (Table 5b).

3.2.2. Severe Cellwise Outliers (μ_1 = 10)

When the cellwise outliers' inflation increases from 5 to 10, $\bar{X}$ further deteriorates as shown in Tables 6a and 7a. The deviation of $\bar{X}_{(MSD,\alpha)}$ began to increase following cellwise outliers' inflation increment. Besides that, $\hat{M}$ maintains with relatively similar results despite the cellwise outliers' inflation increment. Notably, although cellwise outliers' weight increases, $\hat{M}_{(MSD,\alpha)}$'s deviation from L_p decreases with α = 0.3 and α = 0.4, while maintains with α = 0.2.

ANOVA results with p = 0.000 are reported, reported for both sample sizes n = 40 and n = 100. Tables 6b and 7b show that $\bar{X}$ is the least fit estimation for both n = 40 and n = 100. On the contrary, $\hat{M}_{(MSD,\alpha)}$ did not have significant difference (p-value = 1.000) for α = 0.2, 0.3, 0.4 with sample size n = 40 (Table 6b). Similarly, Table 7b (n = 100) also reported no difference between $\hat{M}_{(MSD,0.3)}$ and $\hat{M}_{(MSD,0.4)}$. However, the rankings between the three $\hat{M}_{(MSD,\alpha)}$ estimations against other estimators are shown in the following tables.

Table 4a. Simulation Parameter Settings (n=40, ε_1 = 0.1, μ_1 = 5, ε_2 = 0.2)

Location Estimators (T)		**Pre-set Location (L_p)**					
		27	**10**	**33**	**21**	**42**	**38**
$\bar{\mathbf{X}}$	Average	30.5387	13.4922	36.4967	24.4955	45.5179	41.4825
	Ave. Abs. Diff |T-L|	3.5695	3.5273	3.5300	3.5253	3.5512	3.5185
$\widehat{\mathbf{M}}$	Average	27.3451	10.3439	33.3445	21.3451	42.3444	38.3412
	Ave. Abs. Diff |T-L|	0.3689	0.3661	0.3672	0.3685	0.3654	0.3645
$\bar{\mathbf{X}}_{(\mathbf{MSD},0.4)}$	Average	27.4618	10.4611	33.4630	21.4618	42.4627	38.4546
	Ave. Abs. Diff |T-L|	0.4819	0.4815	0.4825	0.4824	0.4820	0.4737
$\bar{\mathbf{X}}_{(\mathbf{MSD},0.3)}$	Average	27.5377	10.5365	33.5396	21.5369	42.5358	38.5307
	Ave. Abs. Diff |T-L|	0.5423	0.5419	0.5444	0.5421	0.5415	0.5355
$\bar{\mathbf{X}}_{(\mathbf{MSD},0.2)}$	Average	27.6289	10.6290	33.6299	21.6305	42.6284	38.6265
	Ave. Abs. Diff |T-L|	0.6290	0.6291	0.6301	0.6306	0.6284	0.6265
$\widehat{\mathbf{M}}_{(\mathbf{MSD},0.4)}$	Average	27.1557	10.1588	33.1553	21.1550	42.1567	38.1509
	Ave. Abs. Diff |T-L|	0.2537	0.2511	0.2515	0.2530	0.2511	0.2493
$\widehat{\mathbf{M}}_{(\mathbf{MSD},0.3)}$	Average	27.1659	10.1689	33.1680	21.1645	42.1665	38.1620
	Ave. Abs. Diff |T-L|	0.2442	0.2423	0.2441	0.2438	0.2411	0.2398
$\widehat{\mathbf{M}}_{(\mathbf{MSD},0.2)}$	Average	27.1776	10.1806	33.1795	21.1787	42.1794	38.1766
	Ave. Abs. Diff |T-L|	0.2380	0.2376	0.2374	0.2373	0.2360	0.2360

Table 4b. Estimators Ranking based on Tukey HSD Test (n=40, ε_1 = 0.1, μ_1 = 5, ε_2 = 0.2)

Worst Fit							Best Fit
$\bar{\mathbf{X}}$	$\bar{\mathbf{X}}_{(\mathbf{MSD},0.2)}$	$\bar{\mathbf{X}}_{(\mathbf{MSD},0.3)}$	$\bar{\mathbf{X}}_{(\mathbf{MSD},0.4)}$	$\widehat{\mathbf{M}}$	$\widehat{\mathbf{M}}_{(\mathbf{MSD},0.4)}{}^{*}$	$\widehat{\mathbf{M}}_{(\mathbf{MSD},0.3)}{}^{*,**}$	$\widehat{\mathbf{M}}_{(\mathbf{MSD},0.2)}{}^{**}$

Note: All pairwise Tukey HSD results are significant at 0.05 significant level
except $^{*}\widehat{\mathrm{M}}_{(\mathrm{MSD},0.4)}$ and $\widehat{\mathrm{M}}_{(\mathrm{MSD},0.3)}$ are insignificant with p-value of 0.375;
$^{**}\widehat{\mathrm{M}}_{(\mathrm{MSD},0.3)}$ and $\widehat{\mathrm{M}}_{(\mathrm{MSD},0.2)}$ are insignificant with p-value of 0.881.

Table 5a. Simulation Parameter Settings (n=100, ε_1 = 0.1, μ_1 = 5, ε_2 = 0.2)

Location Estimators (T)		**Pre-set Location (L_p)**					
		27	**10**	**33**	**21**	**42**	**38**
$\bar{\mathbf{X}}$	Average	30.5047	13.4883	36.4974	24.4988	45.4872	41.4940
	Ave. Abs. Diff |T-L|	3.5052	3.4886	3.4977	3.4992	3.4876	3.4944
$\widehat{\mathbf{M}}$	Average	27.3417	10.3439	33.3443	21.3439	42.3445	38.3432
	Ave. Abs. Diff |T-L|	0.3441	0.3463	0.3466	0.3464	0.3468	0.3457
$\bar{\mathbf{X}}_{(\mathbf{MSD},0.4)}$	Average	27.4985	10.4965	33.4931	21.4979	42.4984	38.4964
	Ave. Abs. Diff |T-L|	0.5009	0.4990	0.4955	0.5003	0.5011	0.4991
$\bar{\mathbf{X}}_{(\mathbf{MSD},0.3)}$	Average	27.5498	10.5501	33.5485	21.5495	42.5517	38.5509
	Ave. Abs. Diff |T-L|	0.5499	0.5504	0.5487	0.5499	0.5520	0.5512
$\bar{\mathbf{X}}_{(\mathbf{MSD},0.2)}$	Average	27.6243	10.6246	33.6242	21.6252	42.6256	38.6252
	Ave. Abs. Diff |T-L|	0.6243	0.6246	0.6242	0.6252	0.6256	0.6252
$\widehat{\mathbf{M}}_{(\mathbf{MSD},0.4)}$	Average	27.1892	10.1864	33.1870	21.1873	42.1897	38.1875
	Ave. Abs. Diff |T-L|	0.2178	0.2152	0.2172	0.2167	0.2190	0.2160
$\widehat{\mathbf{M}}_{(\mathbf{MSD},0.3)}$	Average	27.1822	10.1808	33.1810	21.1803	42.1835	38.1813
	Ave. Abs. Diff |T-L|	0.2050	0.2031	0.2042	0.2034	0.2060	0.2044
$\widehat{\mathbf{M}}_{(\mathbf{MSD},0.2)}$	Average	27.1788	10.1787	33.1792	21.1791	42.1808	38.1789
	Ave. Abs. Diff |T-L|	0.1957	0.1957	0.1960	0.1957	0.1970	0.1959

Table 5b. Estimators Ranking based on Tukey HSD Test (n=100, ε_1 = 0.1, μ_1 = 5, ε_2 = 0.2)

Worst Fit							Best Fit
$\bar{\mathbf{X}}$	$\bar{\mathbf{X}}_{(\mathbf{MSD},0.2)}$	$\bar{\mathbf{X}}_{(\mathbf{MSD},0.3)}$	$\bar{\mathbf{X}}_{(\mathbf{MSD},0.4)}$	$\widehat{\mathbf{M}}$	$\widehat{\mathbf{M}}_{(\mathbf{MSD},0.4)}$	$\widehat{\mathbf{M}}_{(\mathbf{MSD},0.3)}$	$\widehat{\mathbf{M}}_{(\mathbf{MSD},0.2)}$

Note: All pairwise Tukey HSD results are significant at 0.05 significant level

Table 6a. Simulation Parameter Settings (n=40, ε_1 = 0.1, μ_1 = 10, ε_2 = 0.2)

Location Estimators (T)		Pre-set Location (L_p)					
		27	10	33	21	42	38
$\bar{X}$	Average	31.0387	13.9922	36.9967	24.9955	46.0179	41.9825
	Ave. Abs. Diff |T-L|	4.0538	4.0091	4.0116	4.0094	4.0344	3.9999
$\hat{M}$	Average	27.3452	10.3440	33.3457	21.3443	42.3419	38.3420
	Ave. Abs. Diff |T-L|	0.3694	0.3660	0.3679	0.3691	0.3638	0.3653
$\bar{X}_{(MSD,0.4)}$	Average	27.8204	10.8152	33.8243	21.8228	42.8218	38.8073
	Ave. Abs. Diff |T-L|	0.8408	0.8358	0.8454	0.8437	0.8420	0.8276
$\bar{X}_{(MSD,0.3)}$	Average	28.0346	11.0294	34.0383	22.0397	43.0352	39.0264
	Ave. Abs. Diff |T-L|	1.0379	1.0330	1.0412	1.0429	1.0380	1.0293
$\bar{X}_{(MSD,0.2)}$	Average	28.2664	11.2636	34.2650	22.2639	43.2642	39.2592
	Ave. Abs. Diff |T-L|	1.2664	1.2636	1.2650	1.2639	1.2643	1.2592
$\hat{M}_{(MSD,0.4)}$	Average	27.1296	10.1314	33.1344	21.1326	42.1279	38.1287
	Ave. Abs. Diff |T-L|	0.2398	0.2386	0.2414	0.2418	0.2357	0.2391
$\hat{M}_{(MSD,0.3)}$	Average	27.1549	10.1564	33.1587	21.1573	42.1543	38.1544
	Ave. Abs. Diff |T-L|	0.2381	0.2361	0.2390	0.2396	0.2341	0.2364
$\hat{M}_{(MSD,0.2)}$	Average	27.1789	10.1808	33.1814	21.1790	42.1784	38.1779
	Ave. Abs. Diff |T-L|	0.2398	0.2377	0.2393	0.2398	0.2363	0.2367

Table 6b. Estimators Ranking based on Tukey HSD Test (n=40, ε_1 = 0.1, μ_1 = 10, ε_2 = 0.2)

Worst Fit							Best Fit
$\bar{X}$	$\bar{X}_{(MSD,0.2)}$	$\bar{X}_{(MSD,0.3)}$	$\bar{X}_{(MSD,0.4)}$	$\hat{M}$	$\hat{M}_{(MSD,0.4)}{}^{*}$	$\hat{M}_{(MSD,0.2)}{}^{*,**}$	$\hat{M}_{(MSD,0.3)}{}^{**}$

Note: All pairwise Tukey HSD results are significant at 0.05 significant level except $^{*}\hat{M}_{(MSD,0.4)}$ and $\hat{M}_{(MSD,0.2)}$ are insignificant with p-value of 1.000; $^{**}\hat{M}_{(MSD,0.2)}$ and $\hat{M}_{(MSD,0.3)}$ are insignificant with p-value of 1.000.

Table 7a. Simulation Parameter Settings (n=100, ε_1 = 0.1, μ_1 = 10, ε_2 = 0.2)

Location Estimators (T)		Pre-set Location (L_p)					
		27	10	33	21	42	38
$\bar{X}$	Average	31.0047	13.9883	36.9974	24.9988	45.9872	41.9940
	Ave. Abs. Diff |T-L|	4.0049	3.9883	3.9974	3.9988	3.9872	3.9940
$\hat{M}$	Average	27.3415	10.3438	33.3442	21.3430	42.3446	38.3428
	Ave. Abs. Diff |T-L|	0.3440	0.3464	0.3466	0.3456	0.3471	0.3453
$\bar{X}_{(MSD,0.4)}$	Average	27.7964	10.7883	33.7880	21.7918	42.7905	38.7934
	Ave. Abs. Diff |T-L|	0.8004	0.7926	0.7919	0.7955	0.7944	0.7973
$\bar{X}_{(MSD,0.3)}$	Average	28.0090	11.0062	34.0055	22.0073	43.0058	39.0107
	Ave. Abs. Diff |T-L|	1.0092	1.0064	1.0055	1.0075	1.0058	1.0109
$\bar{X}_{(MSD,0.2)}$	Average	28.2506	11.2505	34.2503	22.2512	43.2519	39.2512
	Ave. Abs. Diff |T-L|	1.2506	1.2505	1.2503	1.2512	1.2519	1.2512
$\hat{M}_{(MSD,0.4)}$	Average	27.1368	10.1367	33.1384	21.1371	42.1394	38.1378
	Ave. Abs. Diff |T-L|	0.1815	0.1804	0.1830	0.1816	0.1843	0.1809
$\hat{M}_{(MSD,0.3)}$	Average	27.1570	10.1574	33.1590	21.1570	42.1582	38.1584
	Ave. Abs. Diff |T-L|	0.1861	0.1865	0.1875	0.1863	0.1875	0.1866
$\hat{M}_{(MSD,0.2)}$	Average	27.1788	10.1794	33.1804	21.1789	42.1812	38.1792
	Ave. Abs. Diff |T-L|	0.1956	0.1965	0.1967	0.1951	0.1970	0.1957

Table 7b. Estimators Ranking based on Tukey HSD Test (n=100, ε_1 = 0.1, μ_1 = 10, ε_2 = 0.2)

Worst Fit							**Best Fit**
$\bar{X}$	$\bar{X}_{(MSD,0.2)}$	$\bar{X}_{(MSD,0.3)}$	$\bar{X}_{(MSD,0.4)}$	$\hat{M}$	$\hat{M}_{(MSD,0.2)}$	$\hat{M}_{(MSD,0.3)}{}^{*}$	$\hat{M}_{(MSD,0.4)}{}^{*}$

Note: All pairwise Tukey HSD results are significant at 0.05 significant level except $^{*}\hat{M}_{(MSD,0.3)}$ and $\hat{M}_{(MSD,0.4)}$ are insignificant with p-value of 0.091.

3.3. Data Contamination at $\varepsilon_1 = 0.1$ and $\varepsilon_2 = 0.4$

3.3.1. Mild Cellwise Outliers ($\mu_1 = 5$)

With ε_1 maintained at 0.1 and ε_2 increased from 0.2 to 0.4, the total outlying proportion in the data added up to 50%. The $\bar{\mathbf{X}}$ estimation is now severely deviated from L_p regardless of sample sizes (Tables 8a and 9a). Obvious deviation also can be observed in $\bar{\mathbf{X}}_{(\mathbf{MSD},\boldsymbol{\alpha})}$ when the outlying proportion is greater than trimming percentages (α = 0.2, 0.3, 0.4), while $\widehat{\mathbf{M}}$ starts showing increase in deviation, as well. Meanwhile, noticeable differences in $\widehat{\mathbf{M}}_{(\mathbf{MSD},\boldsymbol{\alpha})}$ between the trimming percentages can be detected, as higher α will have lower |T-L|.

The ANOVA test and Tukey HSD test for n = 40 ($p=0.000$) and n = 100 ($p=0.000$) have reported significant difference in performance between the estimators. Tables 8b and 9b show identical rankings for estimators, with $\widehat{\mathbf{M}}_{(\mathbf{MSD},\mathbf{0.4})}$ has the best fit while $\bar{\mathbf{X}}$ has the least fit for both sample size n = 40 and n = 100.

3.3.2. Severe Cellwise Outliers ($\mu_1 = 10$)

The most severe outlying condition in the study, $\varepsilon_1 = 0.1$, $\mu_1 = 10$ and $\varepsilon_2 = 0.4$, was simulated and reported into Table 10a and Table 11a for n = 40 and n = 100 respectively.

As expected, $\bar{\mathrm{X}}$ deviates the most from L_p with average deviation of 6.997 for both n = 40 and n = 100 under each dimension. Further deviation also can be seen on $\bar{\mathrm{X}}_{(\mathrm{MSD},\alpha)}$, which corresponds with the cell wise outliers' inflation.

However, $\widehat{\mathbf{M}}$ performance maintains at this level of outliers' proportions, $\varepsilon_1 = 0.1$ and $\varepsilon_2 = 0.4$, despite the cellwise outliers' inflation. Similarly, $\widehat{\mathbf{M}}_{(\mathbf{MSD},\boldsymbol{\alpha})}$'s deviation does not affected much from the inflation, maintaining its low deviation even at higher trimming percentage.

Lastly, ANOVA test again indicates significant difference on estimators' performance with p = 0.000 for n = 40 and n = 100. The ranking of the estimators is similar to the previous section (as shown in Tables 10b and 11b).

Table 8a. Simulation Parameter Settings (n=40, ε_1 = 0.1, μ_1 = 5, ε_2 = 0.4)

Location Estimators (T)		Pre-set Location (L_p)					
		27	10	33	21	42	38
$\bar{\mathbf{X}}$	Average	33.4903	16.4767	39.4676	27.5324	48.4886	44.5046
	Ave. Abs. Diff |T-L|	6.4974	6.4820	6.4746	6.5406	6.4971	6.5120
$\widehat{\mathbf{M}}$	Average	27.7786	10.7759	33.7732	21.7850	42.7767	38.7792
	Ave. Abs. Diff |T-L|	0.7833	0.7820	0.7802	0.7912	0.7846	0.7855
$\bar{\mathbf{X}}_{(\mathbf{MSD},0.4)}$	Average	27.8379	10.8354	33.8372	21.8405	42.8360	38.8402
	Ave. Abs. Diff |T-L|	0.8380	0.8354	0.8372	0.8405	0.8360	0.8402
$\bar{\mathbf{X}}_{(\mathbf{MSD},0.3)}$	Average	29.4462	12.4359	35.4263	23.4202	44.4207	40.4206
	Ave. Abs. Diff |T-L|	2.4840	2.4783	2.4743	2.4666	2.4624	2.4609
$\bar{\mathbf{X}}_{(\mathbf{MSD},0.2)}$	Average	30.9615	13.9400	36.9208	24.9491	45.9403	41.9223
	Ave. Abs. Diff |T-L|	3.9977	3.9781	3.9652	3.9908	3.9777	3.9537
$\widehat{\mathbf{M}}_{(\mathbf{MSD},0.4)}$	Average	27.2522	10.2499	33.2518	21.2533	42.2475	38.2556
	Ave. Abs. Diff |T-L|	0.3062	0.3014	0.3053	0.3058	0.3041	0.3084
$\widehat{\mathbf{M}}_{(\mathbf{MSD},0.3)}$	Average	27.3795	10.3763	33.3789	21.3807	42.3746	38.3832
	Ave. Abs. Diff |T-L|	0.4084	0.4056	0.4086	0.4108	0.4070	0.4131
$\widehat{\mathbf{M}}_{(\mathbf{MSD},0.2)}$	Average	27.5086	10.5052	33.5054	21.5101	42.5034	38.5077
	Ave. Abs. Diff |T-L|	0.5246	0.5230	0.5240	0.5290	0.5221	0.5258

Table 8b. Estimators Ranking based on Tukey HSD Test (n=40, ε_1 = 0.1, μ_1 = 5, ε_2 = 0.4)

Worst Fit							Best Fit
$\bar{\mathbf{X}}$	$\bar{\mathbf{X}}_{(\mathbf{MSD},0.2)}$	$\bar{\mathbf{X}}_{(\mathbf{MSD},0.3)}$	$\bar{\mathbf{X}}_{(\mathbf{MSD},0.4)}$	$\widehat{\mathbf{M}}$	$\widehat{\mathbf{M}}_{(\mathbf{MSD},0.2)}$	$\widehat{\mathbf{M}}_{(\mathbf{MSD},0.3)}$	$\widehat{\mathbf{M}}_{(\mathbf{MSD},0.4)}$

Note: All pairwise Tukey HSD results are significant at 0.05 significant level

Table 9a. Simulation Parameter Settings (n=100, ε_1 = 0.1, μ_1 = 5, ε_2 = 0.4)

Location Estimators (T)		Pre-set Location (L_p)					
		27	10	33	21	42	38
$\bar{X}$	Average	33.5093	16.5053	39.4953	27.4935	48.4974	44.4803
	Ave. Abs. Diff |T-L|	6.5093	6.5053	6.4953	6.4935	6.4975	6.4803
$\hat{M}$	Average	27.7574	10.7575	33.7545	21.7554	42.7574	38.7558
	Ave. Abs. Diff |T-L|	0.7575	0.7577	0.7547	0.7555	0.7576	0.7559
$\bar{X}_{(MSD,0.4)}$	Average	27.8335	10.8350	33.8352	21.8345	42.8362	38.8347
	Ave. Abs. Diff |T-L|	0.8335	0.8350	0.8352	0.8345	0.8362	0.8347
$\bar{X}_{(MSD,0.3)}$	Average	29.3949	12.3975	35.3905	23.3757	44.3889	40.3816
	Ave. Abs. Diff |T-L|	2.3963	2.3988	2.3913	2.3766	2.3899	2.3830
$\bar{X}_{(MSD,0.2)}$	Average	30.8926	13.8929	36.8797	24.8796	45.8869	41.8764
	Ave. Abs. Diff |T-L|	3.8939	3.8936	3.8810	3.8803	3.8872	3.8771
$\hat{M}_{(MSD,0.4)}$	Average	27.2514	10.2519	33.2530	21.2537	42.2513	38.2531
	Ave. Abs. Diff |T-L|	0.2637	0.2637	0.2651	0.2653	0.2640	0.2651
$\hat{M}_{(MSD,0.3)}$	Average	27.3815	10.3822	33.3817	21.3812	42.3803	38.3808
	Ave. Abs. Diff |T-L|	0.3850	0.3855	0.3851	0.3841	0.3841	0.3843
$\hat{M}_{(MSD,0.2)}$	Average	27.5069	10.5078	33.5059	21.5078	42.5077	38.5082
	Ave. Abs. Diff |T-L|	0.5081	0.5089	0.5074	0.5087	0.5090	0.5092

Table 9b. Estimators Ranking based on Tukey HSD Test (n=100, ε_1 = 0.1, μ_1 = 5, ε_2 = 0.4)

Worst Fit							Best Fit
$\bar{X}$	$\bar{X}_{(MSD,0.2)}$	$\bar{X}_{(MSD,0.3)}$	$\bar{X}_{(MSD,0.4)}$	$\hat{M}$	$\hat{M}_{(MSD,0.2)}$	$\hat{M}_{(MSD,0.3)}$	$\hat{M}_{(MSD,0.4)}$

Note: All pairwise Tukey HSD results are significant at 0.05 significant level

Table 10a. Simulation Parameter Settings (n=40, ε_1 = 0.1, μ_1= 10, ε_2 = 0.4)

Location Estimators (T)		Pre-set Location (L_p)					
		27	10	33	21	42	38
$\bar{X}$	Average	33.9903	16.9767	39.9676	28.0324	48.9886	45.0046
	Ave. Abs. Diff |T-L|	6.9937	6.9791	6.9717	7.0367	6.9934	7.0085
$\hat{M}$	Average	27.7774	10.7757	33.7724	21.7897	42.7799	38.7793
	Ave. Abs. Diff |T-L|	0.7832	0.7820	0.7791	0.7964	0.7873	0.7851
$\bar{X}_{(MSD,0.4)}$	Average	28.6833	11.6855	34.6843	22.6871	43.6837	39.6884
	Ave. Abs. Diff |T-L|	1.6833	1.6855	1.6843	1.6871	1.6837	1.6884
$\bar{X}_{(MSD,0.3)}$	Average	30.1593	13.1527	36.1350	24.1264	45.1316	41.1317
	Ave. Abs. Diff |T-L|	3.1674	3.1611	3.1467	3.1389	3.1417	3.1419
$\bar{X}_{(MSD,0.2)}$	Average	31.5740	14.5487	37.5371	25.5597	46.5563	42.5282
	Ave. Abs. Diff |T-L|	4.5904	4.5654	4.5569	4.5769	4.5718	4.5409
$\hat{M}_{(MSD,0.4)}$	Average	27.2517	10.2508	33.2535	21.2574	42.2522	38.2558
	Ave. Abs. Diff |T-L|	0.3058	0.3031	0.3056	0.3094	0.3069	0.3082
$\hat{M}_{(MSD,0.3)}$	Average	27.3775	10.3760	33.3787	21.3820	42.3756	38.3814
	Ave. Abs. Diff |T-L|	0.4069	0.4059	0.4081	0.4117	0.4075	0.4103
$\hat{M}_{(MSD,0.2)}$	Average	27.5069	10.5037	33.5051	21.5097	42.5049	38.5059
	Ave. Abs. Diff |T-L|	0.5245	0.5217	0.5235	0.5281	0.5232	0.5232

Table 10b. Estimators Ranking based on Tukey HSD Test (n=40, ε_1 = 0.1, μ_1 = 10, ε_2 = 0.4)

Worst Fit							Best Fit
$\bar{X}$	$\bar{X}_{(MSD,0.2)}$	$\bar{X}_{(MSD,0.3)}$	$\bar{X}_{(MSD,0.4)}$	$\hat{M}$	$\hat{M}_{(MSD,0.2)}$	$\hat{M}_{(MSD,0.3)}$	$\hat{M}_{(MSD,0.4)}$

Note: All pairwise Tukey HSD results are significant at 0.05 significant level

Table 11a. Simulation Parameter Settings (n=100, ε_1 = 0.1, μ_1 = 10, ε_2 = 0.4)

Location Estimators (T)		Pre-set Location (L_p)					
		27	10	33	21	42	38
$\bar{X}$	Average	34.0093	17.0053	39.9953	27.9935	48.9974	44.9803
	Ave. Abs. Diff |T-L|	7.0093	7.0053	6.9953	6.9935	6.9974	6.9803
$\hat{M}$	Average	27.7555	10.7586	33.7544	21.7555	42.7595	38.7546
	Ave. Abs. Diff |T-L|	0.7556	0.7587	0.7546	0.7556	0.7597	0.7547
$\bar{X}_{(MSD,0.4)}$	Average	28.6711	11.6739	34.6732	22.6734	43.6744	39.6718
	Ave. Abs. Diff |T-L|	1.6711	1.6739	1.6732	1.6734	1.6744	1.6718
$\bar{X}_{(MSD,0.3)}$	Average	30.1152	13.1161	36.1091	24.0942	45.1065	41.1022
	Ave. Abs. Diff |T-L|	3.1153	3.1163	3.1091	3.0942	3.1066	3.1023
$\bar{X}_{(MSD,0.2)}$	Average	31.5109	14.5127	37.4982	25.5011	46.5064	42.4971
	Ave. Abs. Diff |T-L|	4.5113	4.5127	4.4983	4.5014	4.5064	4.4971
$\hat{M}_{(MSD,0.4)}$	Average	27.2502	10.2539	33.2529	21.2541	42.2532	38.2535
	Ave. Abs. Diff |T-L|	0.2627	0.2644	0.2653	0.2658	0.2653	0.2657
$\hat{M}_{(MSD,0.3)}$	Average	27.3801	10.3837	33.3808	21.3819	42.3828	38.3818
	Ave. Abs. Diff |T-L|	0.3838	0.3864	0.3849	0.3850	0.3867	0.3853
$\hat{M}_{(MSD,0.2)}$	Average	27.5042	10.5085	33.5054	21.5082	42.5086	38.5079
	Ave. Abs. Diff |T-L|	0.5054	0.5094	0.5071	0.5093	0.5100	0.5090

Table 11b. Estimators Ranking based on Tukey HSD Test (n=100, ε_1 = 0.1, μ_1 = 10, ε_2 = 0.4)

Worst Fit							Best Fit
$\bar{X}$	$\bar{X}_{(MSD,0.2)}$	$\bar{X}_{(MSD,0.3)}$	$\bar{X}_{(MSD,0.4)}$	$\hat{M}$	$\hat{M}_{(MSD,0.2)}$	$\hat{M}_{(MSD,0.3)}$	$\hat{M}_{(MSD,0.4)}$

Note: All pairwise Tukey HSD results are significant at 0.05 significant level

3.4. Summary

To summarize, the classical mean ($\bar{X}$) is good when applied to clean data, but very susceptible to the existence of outliers. The performance of classical median ($\hat{M}$) is considered moderate as the rankings are consistently in between the other estimators. On the other hand, despite having comparable performance to $\hat{M}$ when applied to clean data, $\bar{X}_{(MSD,\alpha)}$'s performance worsens when dealing with contaminated data. The estimation through $\bar{X}_{(MSD,\alpha)}$ increasingly deviates from L_p as the proportion (ε_2) of casewise outliers increases. The deviation also increases when the cellwise outliers' inflation (μ_1) increases. However, $\hat{M}_{(MSD,\alpha)}$ holds a dominant performance over $\bar{X}$, $\hat{M}$ and $\bar{X}_{(MSD,\alpha)}$ on all contaminated occasions regardless of the trimming percentage applied. Overall, $\hat{M}_{(MSD,0.4)}$ yields best location estimations than the other two trimming percentages (α = 0.2, 0.3).

4. Conclusions

In this paper, the capability of the existing multivariate location estimators in handling situation where both cellwise outliers and casewise outliers coexist is discussed and reviewed. One should take note that when both casewise and cellwise outliers coexist in the dataset, they should be treated separately. Treating the coexistence of these outliers like a single problem could still jeopardize the location measure estimation as the outliers might not be well identified (leftover outliers). Based on the results of simulation study, the proposed distance-based trimmed median, $\hat{M}_{(MSD,\alpha)}$, yields more accurate location estimations despite the leftover outliers compared to the other three chosen estimators in this study.

Pertaining the feasibility of using median over mean after α-distance-based trimming, the simulation study shows that $\hat{M}_{(MSD,\alpha)}$ has consistently lower deviation as compared to $\bar{X}_{(MSD,\alpha)}$. Despite the performance of both $\hat{M}_{(MSD,\alpha)}$ and $\bar{X}_{(MSD,\alpha)}$ getting better as the trimming percentage (α) increases, in cases where α relatively lower than the total outlying proportions in the data, $\hat{M}_{(MSD,\alpha)}$ outperforms $\bar{X}_{(MSD,\alpha)}$ with noticeable differences. When the contaminated data is under-trimmed, the leftover cellwise and casewise outliers can cause $\bar{X}_{(MSD,\alpha)}$ to breakdown. On the other hand, even though trimming percentage equal or greater than the total outlying proportions in the data, some cellwise outliers might be left from trimming due to the cellwise outliers which are randomly spread within the data. Thus, the use of median after α-distance-based trimming is recommended to mitigate the problem.

As a conclusion, the proposed two-stage outlier's filtration location estimator, $\widehat{\mathbf{M}}_{(\mathbf{MSD},\boldsymbol{\alpha})}$, is recommended as a better multivariate location estimator when dealing with unknown outlier's situation.

Acknowledgements

This research was supported by Ministry of Higher Education (MoHE) of Malaysia through Fundamental Research Grant Scheme (FRGS/1/2019/STG06/UUM/02/8).

REFERENCES

[1] F. R. Hampel, "Robust Statistics: A Brief Introduction and Overview", In Seminar für Statistik, Eidgenössische Technische Hochschule, Switzerland, vol. 94, 2001, pp. 1-5.

[2] F. R. Hampel, E. M. Ronchetti, P. J. Rousseeuw and W. A. Stahel, Robust statistics: the approach based on influence functions, Vol. 196, John Wiley & Sons, 2011.

[3] R. J. Pell, "Multiple outlier detection for multivariate calibration using robust statistical techniques", Chemometrics and Intelligent Laboratory Systems, vol. 52, no. 1, 2000, pp. 87-104.

[4] J. Á. Víšek, "Regression with high breakdown point", in Proceedings of ROBUST, Sept.2000, pp. 324-356.

[5] F. R. Hampel, "The breakdown points of the mean combined with some rejection rules", Technometrics, vol 27, no. 2, 1985, pp. 95-107.

[6] P. J. Rousseeuw, "Multivariate estimation with high breakdown point", Mathematical statistics and applications, vol. 8, 1985, pp. 283-297.

[7] D. M. Hawkins, "Outliers", in Encyclopedia of Biostatistics, P. Armitage and T. Colton, Eds. John Wiley & Sons, 2005, pp. 4233-4237.

[8] I. M. Rodrigues and G. Boente, "Multivariate Outliers", in International Encyclopedia of Statistical Science , M. Lovric, Ed. Springer-Verlag Berlin Heidelberg, 2011, pp. 910-912.

[9] C. Agostinelli, A. Leung, V. J. Yohai and R. H. Zamar, "Robust estimation of multivariate location and scatter in the presence of cellwise and casewise contamination", TEST, vol. 24, no. 3, 2015, pp. 441-461.

[10] P. J. Rousseeuw and W. Van den Bossche, "Comments on: Robust estimation of multivariate location and scatter in the presence of cellwise and casewise contamination", TEST, vol. 24, no. 3, 2015, pp. 473-477.

[11] J. W. Tukey, "Some elementary problems of importance to small sample practice", Human Biology, vol. 20, no. 4, 1948, pp. 205-214.

[12] D. K. Srivastava and G. S. Mudholkar, "Trimmed T2: A robust analog of hotelling's T2", Journal of Statistical Planning and Inference, vol. 97, no. 2, 2001, pp. 343-358.

[13] J. Alloway and M. Raghavachari, "Multivariate control charts based on trimmed means", in ASQC Quality Congress Transactions, San Francisco, 1990, pp. 449-453.

[14] Y. S. Pang, N. A. Ahad, S. S. Syed Yahaya and Y. F. Lim, "Robust Linear Discriminant Rule Using Novel Distance-based Trimming Procedure", Jour of Adv Research in Dynamical & Control Systems, vol. 11, 05-Special Issue, 2019, pp. 969-978.

[15] G. McLachlan, "Mahalanobis distance", Resonance, vol. 4, no. 6, 1999, pp. 20-26.

[16] D. M. Rocke, G. W. Downs and A. J. Rocke, "Are robust estimators really necessary?", Technometrics, vol. 24, no. 2, 1982, pp. 95-101.

[17] M. Wu, "Trimmed and winsorized estimators," Ph.D. dissertation, Prob. And Stat. Dept., Michigan State Univ., East Lansing, MI, 2007.

[18] R. R. Wilcox, "Trimming and Winsorization", in Encyclopedia of Biostatistics, New York: John Wiley & Sons, Inc, 2005.

[19] J. L. Rosenberger and M. Gasko, "Comparing location estimators: Trimmed means, medians, and trimean", Understanding robust and exploratory data analysis, 1983, pp. 297-336.

[20] A. R. Othman, H. J. Keselman, A. R. Padmanabhan, R. R. Wilcox and K. Fradette, "Comparing measures of the 'typical' score across treatment groups", British Journal of Mathematical and Statistical Psychology, vol. 57, no. 2, 2004, pp. 215-234.

[21] H. J. Keselman, A. R. Othman, R. R. Wilcox and K. Fradette, "The new and improved two-sample t test", Psychological Science, vol. 15, no. 1, 2004, pp. 47-51.

[22] P. J. Huber, "Robust Estimation of a Location Parameter", The Annals of Mathematical Statistics, vol. 35, no. 1, 1964, pp. 73-101.

[23] J. W. Tukey, "The future of data analysis", The annals of mathematical statistics, vol. 33, no. 1, 1962, pp. 1-67.