Austrian Journal of Statistics
February 2020, Volume 49, 1–17.
http://www.ajs.or.at/
doi:10.17713/ajs.v49i2.872

Comparison of Multivariate Outlier Detection Methods for Nearly Elliptical Distributions

Kazumi Wada **Mariko Kawano** **Hiroe Tsubaki**
National Statistics Center (NSTAC)

Abstract

In this paper, the performance of outlier detection methods has been evaluated with symmetrically distributed datasets. We choose four estimators, viz. modified Stahel-Donoho (MSD) estimators, blocked adaptive computationally efficient outlier nominators, minimum covariance determinant estimator obtained by a fast algorithm, and nearest-neighbour variance estimator, which are known for their good performance with elliptically distributed data, for practical applications in national survey data processing. We adopt the data model of multivariate skew-t distribution, of which only the direction of the main axis is skewed and contaminated with outliers following another probability distribution for evaluation. We conducted Monte Carlo simulation under the data distribution to compare the performance of outlier detection. We also explore the applicability of the selected methods for several accounting items in small and medium enterprise survey data. Accordingly, it was found that the MSD estimators are the most suitable.

Keywords: robust estimation, location and scatter, MSD, BACON, Fast-MCD, NNVE, R.

1. Objective

In this paper, we discuss the performance of multivariate outlier detection methods applied on asymmetric data. Asymmetric multivariate t-distribution is adopted for the population probability distribution for evaluation. The targeted multivariate probability model of no-outlying data is unimodal and asymmetric. Outliers here are defined to be the observations which cannot be regarded as following the identical distribution with non-outlying data in the same dataset. We also examine the practical implementation of the proposed methods in the data processing of national survey statistics.

A univariate outlier detection approach like the box plot is commonly used during production of traditional statistical tables in national statistical offices (NSOs), since it is easy to process and understand. On the other hand, multivariate methods have apparently not been widely used yet. A primary reason is that multivariate methods are often computationally burdensome, and it is more difficult to evaluate the outcome. Furthermore, the outliers detected by multivariate methods may vary with different methods, or even with a same methods using different random seeds. Meanwhile, the need for a multivariate method is increasing as microdata have become important in addition to statistical tables. While values in tables are aggregated and relations between variables become less visible, microdata preserve such relations. This is what makes microdata useful to users. Therefore, it is important to remove outliers in microdata, which affect the relations between variables. Removing outliers prevent erroneous estimations, diminishes the risk of disclosure and also helps to obtain statistical tables in good quality. A multivariate method is necessary for detecting outliers based on the relations between variables.

In recent decades, more sophisticated and computationally burdensome multivariate methods have been proposed, along with the development of robust statistics and information technology. A comprehensive research project on the editing and imputation of official statistics, known as EUREDIT, was launched in 2000. It was funded by Eurostat and involved professionals of European national statistics offices and academics. Several modern multivariate outlier detection methods, such as modified Stahel-Donoho (MSD) estimators (Stahel (1981), Donoho (1982)), the blocked adaptive computationally efficient outlier nominators (BACON) proposed by Billor, Hadi, and Velleman (2000), minimum covariance determinant estimator obtained by a fast algorithm (Fast-MCD) by Rousseeuw and Driessen (1999), and Kosinski methods Kosinski (1998) were examined by Béguin and Hulliger (2003). The results suggest that there is no panacea for multivariate outlier detection. Each method has its own characteristics and it applies to suitable data. Wada (2004) compared BACON, neatest-neighbour variance estimator (NNVE) by Wang and Raftery (2002), and Fast-MCD estimator using asymmetrically contaminated normal and skew-t data in addition to some famous datasets for outlier detection, such as Hertzsprung-Russell (Rousseeuw and Leroy 1987), Bushfire (Campbell 1989), Ionosphere from the UCI Machine Learning Repository, to find the difference among these methods. The results show that all those methods are tolerant to skewness. BACON is basically superior to the other two methods and tends to maintain its robustness with skewed data. Fast-MCD essentially follows BACON but, compared to BACON, tends to maintain its robustness with heavy-tailed data. NNVE is suitable for data with contamination of larger variance.

Since the MSD estimators were unavailable in R unlike other methods mentioned above, Wada (2010) implemented the MSD estimators based on Franklin and Brodeur (1997) together with the improved estimators suggested by Béguin and Hulliger (2003). The results confirmed that the suggestions improve performance of the estimators, while they suffer from the *curse of dimensionality* as one of the suggestions is to increase number of orthogonal bases for projection exponentially along with p. Therefore, the improved MSD function can compute up to 11 variables with a size of 100 observations on a 32-bit PC. Wada and Tsubaki (2013) made the function parallelised and confirmed with up to 20 variables that the parallelised MSD function has performance equivalent to that of single core MSD. The parallelised MSD can cope with larger datasets with more than 20 variables. However, we use the single-core version of MSD here since our targeted datasets described in section 4 has six variables. The single-core version is also computationally more efficient than the parallelised one.

In section 2, we explain the features of the four selected methods, then describe random data for evaluation and summarise the results in section 3. The Monte Carlo simulation shows that, with skewed data, MSD is better than the other methods. Further, an attempt at implementation of a multivariate outlier detection method in official statistics is described in section 4. It is supposed to be used for removing outlying observations from possible donors in each imputation class for ratio hot deck imputation with edit constraints. Transformation before outlier detection is also discussed. We confirmed that MSD has a good performance with skewed data and can cope with datasets having six variables in the implementation of the targeted survey data processing as described in section 5.

2. Multivariate outlier detection methods for evaluation

We chose the following four methods with high breakdown points, high efficiency, and affine or scale equivariance and currently available in R. All of them estimate a robust mean vector and covariance matrix, although the methodologies vary. Outliers are then judged by the Mahalanobis distance based on the estimated mean vector and the covariance matrix.

2.1. MSD estimators

The MSD estimators are a combination of the Stahel-Donoho (SD) estimators (Stahel (1981), Donoho (1982)) and projection pursuit (PP) (Patak 1990). The estimators achieve orthogonal equivariance and sufficient robustness due to their high breakdown points.

For estimating a robust mean vector and covariance matrix, the SD estimators project observations onto randomly generated orthogonal bases. As multivariate data are translated into one-dimensional data by projection, the problem is reduced to the estimation of a one-dimensional location and a scale. Possible outliers are determined to be those observations for which the following function exceeds a given threshold, and these are then downweighted: $|x_i - median|$/MAD, where MAD is the median absolute deviation, and x_i is a one-dimensional projected observation.

The MSD estimators use the SD estimators as the initial robust mean vector and covariance matrix. These are then used for a principal component analysis of the PP results; the principal components are regarded as ”interesting” directions in which to find outliers. Projection to the principal components also eliminates any correlation between the variables. Possible outliers are downweighted in the same manner as the SD estimators. The final mean vector and covariance matrix are then derived. Outliers are determined by their Mahalanobis distance.

Franklin and Brodeur (1997) describe how the MSD estimators are adopted for the Annual Wholesale and Retail Trade Survey (AWRTS) of Statistics Canada. Béguin and Hulliger (2003) then suggest a few improvements including skipping data centering in the implementation of Franklin and Brodeur (1997).

Wada (2010) implemented both the original and improved MSD estimators and confirmed that the suggestions made by Béguin and Hulliger (2003) improves the performance of the original estimators adopted by Statistics Canada. However, the application of the improved version is limited to datasets with a small number of variables as described in the previous section. Wada and Tsubaki (2013) enabled the estimator to be applicable to larger datasets by parallelisation and confirmed that the parallelised function maintains equivalent performance by a simulation study using the random variables following a multivariate normal distribution with asymmetric contamination. The code is provided together with the not-parallelised version suggested by Béguin and Hulliger (2003) in a public repository (`https://github.com/kazwd2008/MSD/` and `https://github.com/kazwd2008/MSD.parallel/`) for further evaluation.

We adopted the *msd* function implemented by Wada (2010) which is available at `https://github.com/kazwd2008/MSD`. It is the improved MSD estimators suggested by Béguin and Hulliger (2003). The function returns a robust mean vector and a covariance matrix. And then observations are flagged as potential outliers if their Mahalanobis distances exceed the significance level of 0.999 based on the F distribution with p and $(n - p)$ degrees of freedom based on Franklin and Brodeur (1997).

2.2. BACON

Billor *et al.* (2000) proposed BACON, which is named after Francis Bacon. BACON has a high breakdown point and is nearly affine equivariant. It is based on an earlier idea (Hadi (1992); Hadi and Simonoff (1994, 1997)) that finds a good subset of the data without outliers, and then increases it by adding tested observations. BACON improved Hadi’s method (Hadi and Simonoff 1994) by adding multiple observations at each step, which is computationally much more efficient.

The algorithm is shown in Billor *et al.* (2000), and there are two ways to select the initial subset. Version 1 adopts nonrobust, but affine equivariant Mahalanobis distances, and version

2 uses Euclidean distances from coordinate-wise medians, which are robust but not affine equivariant. The default setting is the latter, and we used it in our evaluation. Since the consecutive iteration process for increasing the subset is robust and affine equivariant, the breakdown point of the entire procedure is approximately 40%. Cédric Béguin developed an original S-PLUS function named *BEM* and published the code in Béguin and Hulliger (2003). Béguin and Hulliger (2003) compared *BEM* with Kosinski methods (Kosinski 1998), and found the result to be Version 2, Kosiniski and Version 1 in an decreasing order of performance.

Masato Okamoto of the Ministry of Internal Affairs ported *BEM* to R, and it was also used by Wada (2004) and Wada and Tsubaki (2013). The details of the modifications from the original S-PLUS code are available at `https://github.com/kazwd2008/BEM/`.

We use the *BEM* function ported by Okamoto with its default settings, which is based on Béguin and Hulliger (2003) and described in Wada and Tsubaki (2013). The size of the initial subset is $3 \times p$, where p is the number of variables and the initial subset is selected by Version 2. *BEM* estimates a robust mean vector and covariance matrix, and the outliers are judged by their Mahalanobis distances with a significance level of 0.99 based on the chi square distribution with p degree of freedom.

2.3. Fast-MCD estimator

The MCD estimator is proposed by Rousseeuw (1984). Rousseeuw (1984) also introduced the minimum volume ellipsoid (MVE) estimator, which has an equivalent high breakdown point with MCD. MCD is statistically more efficient (Rousseeuw and Driessen 1999) because of its asymptotic normality and has a higher convergence rate (Davies 1992) than MVE. In addition, MCD is affine equivariant.

The idea of MCD is to find h observations (with $n/2 \leq h \leq n$) whose classical covariance matrix has the lowest determinant from size n of the targeted dataset. Then its location estimator is the average of these h observations, and the scatter is their covariance matrix. The problem of the algorithm is how to find the h observations and the usage of MCD had been limited to small datasets, such as a few hundreds observations with a few variables, due to its heavy computation, until Rousseeuw and Driessen (1999) introduced a fast algorithm to overcome this problem.

The basic ideas behind the fast algorithm is the procedure to generate initial estimates, which consists of the C-step, selective iteration, and nested extensions. Please see Rousseeuw and Driessen (1999) and Todorov and Filzmoser (2009) for the detailed algorithm. Pison, Van Aelst, and Willems (2002) proposed the finite sample bias correction.

We adopted *covMcd* function in *rrcov* package distributed by CRAN (`http://cran.r-project.org`). The function was prepared by Valentin Todorov based on the code of S-plus for the Fast-MCD algorithm implemented by Rousseeuw and Driessen. The algorithm is presented in Rousseeuw and Driessen (1999) and the finite sample correction step by Pison *et al.* (2002) is added. We used it with the default settings; i.e., the size of the subset is $h = (n + p + 1)/2$ so that the MCD estimator has a highest breakdown point. The number of subsets used for initial estimates is 500.

2.4. NNVE

NNVE was proposed by Wang and Raftery (2002). It uses the standardized Euclidean distance from an observation and its k-th nearest neighbour as a measure of the outlyingness of that observation. The idea of NNVE is based on the nearest-neighbour cleaning (NNC) procedure introduced by Byers and Raftery (1998), which regards the data as a mixture of two different gamma distributions (Wang and Raftery 2002). One distribution consists of correct data and the other consists of outliers. An expectation-maximization (EM) algorithm is used to estimate the mean vectors and covariance matrices of these distributions and the proportions of the mixture. NNVE is scale equivariant, but not affine equivariant. NNC outperformed MVE, especially when the proportion of outliers was very large. However, the NNC underestimated the covariance of correct data when there were no outliers. NNVE overcomes this weak point of NNC by adding some artificial outliers. It also performs well when the correct data do not follow an elliptically symmetric distribution.

The NNVE function *cov.nnve* in the package *covRobust* developed by Wang and Raftery (2002) is distributed by CRAN (`http://cran.r-project.org`). The function *cov.nnve* evaluates each observation to determine whether it is an outlier. We used the default settings, and therefore, 12 was adopted as the number of nearest neighbours.

3. Evaluation with random data

The purpose of the comparison in this section is to figure out which methods perform well with skewed and long tailed data by Monte Carlo simulation.

3.1. Random datasets

Peña and Prieto (2001) used datasets consisting of random variables following a multivariate normal distribution with asymmetric contamination, since many outlier detection procedures often have difficulty coping with this model:

$$(1-\alpha)N_p(0,\mathbf{R}) + \alpha N_p(\delta\mathbf{e}_1, \lambda\mathbf{I}), \tag{1}$$

where the first and second terms represent normal data and outliers, respectively; α is the fraction of contamination (i.e. the rate of outliers); p is the number of variables; $\mathbf{R}$ is the correlation matrix in which all the correlations between variables have the same value r; δ denotes the distance between the normal data and the outliers; $\mathbf{e}_1$ is the first unit vector; and λ is the variance of the outliers. Wada and Tsubaki (2013) also adopt the model (1) and evaluate their MSD estimators together with NNVE and BACON. The results show BACON is the most efficient and has better performance with datasets without correlation between variables; however, MSD performs better with datasets having high correlation. Further, BACON is computationally the most efficient; NNVE takes more time with a larger data size, while Fast-MCD, and especially MSD increases processing time when the number of variables increases.

Wada (2004) extended the model (1) from an asymmetrically contaminated normal distribution to a skew-t distribution as follows:

$$(1-\alpha)ST_p(0,\mathbf{R},\eta\mathbf{e}_1, Df) + \alpha N_p(\delta\mathbf{e}_1, \lambda\mathbf{I}), \tag{2}$$

where ST_p is the p-dimensional skew-t distribution, η is the skewness of the first axis, and Df is the number of degrees of freedom. This type of dataset has longer tails than the data from a normal distribution, and it is asymmetric even if there are no outliers.

The multivariate skew-t distribution can be regarded as the distribution of the sample mean vector for random samples from a mixture of multinormal populations. The expected value vector of the mixture of multinormal populations has a joint multinormal distribution and the covariance matrix has an independent joint Wishart distribution (Ando and Kaufman 1965).

We also adopt the model (2), since most survey data have asymmetric distribution. To make a distribution symmetry, a transformation such as Box-Cox power transformation (Box and Cox 1964) is commonly used. However, the Box-Cox transformation is not outlier robust, and moreover, different transformation among variables in a dataset alters their correlations. We add up to 40% outliers in this study. Other values that we used for the parameters are shown in Table 1.

Table 1: Settings for the contaminated skew-t datasets

Parameter	Explanation	Values in the simulations
α	Rate of outliers	$0, 0.1, 0.2, 0.3, 0.4$
r	Correlations between variables	$0.4, 0.8$
p	Number of variables	10
δ	Distance between the normal data and the outliers	$10, 100$
λ	Variance of the outliers	$1, 5$
η	Skewness of the first axis	$0, 1, 5, 10$
Df	Degrees of freedom of the t-distribution	$2, 10, \infty$

3.2. Results of the random datasets

The results for the datasets obtained using model (2) are shown as Table 2, with a coloured bar added to each cell of the ”Total rate” to clarify its magnitude. Please note that model (2) with $\eta = 0$ and $Df = \infty$ actually corresponds to the model (1).

As an overall tendency, MSD appears to be better than BACON, followed by Fast-MCD and NNVE. BACON could be better when there is considerable contamination. A decrease in the degree of freedom strongly affects all four methods, while an increase in skewness does not have a significant effect. BACON is the most affected by the degree of freedom, and NNVE is the least.

Regarding the evaluation by the random dataset shown in Table 1, MSD appears to be the most promising candidate among the four methods for a wide variety of situations. The next section describes the further evaluation of these methods conducted using an actual survey data.

4. Application to a survey data

In this section, we attempt the practical application of the multivariate outlier detection methods in the survey data processing of the Unincorporated Enterprise Survey in Japan. MSD is the most promising candidate in terms of outlier detection capability for skewed and long tailed multivariate data; however, it is a kind of brute force algorithm that is computationally expensive. We compare MSD and BACON, which is the most computationally efficient method among the four methods, and consider the data transformation prior to the outlier detection process. In addition, We measure their processing time with a larger dataset.

4.1. Unincorporated Enterprise Survey

The Unincorporated Enterprise Survey is one of the fundamental statistical surveys based on the Statistical Act in Japan. This survey aims to clarify the actual conditions of business management in unincorporated establishments engaged in manufacturing, wholesale and retail trade, accommodations and food services, or providing services, and to obtain basic data on the trends in business and for the promotion of small and medium-sized enterprises.

The survey covers about $4,000$ establishments, and questionnaires are distributed and collected by statistical enumerators until year 2018. A major change is planned in the survey from 2019. The coverage will be extended to broader industries, and the number of samples will be ten times larger. Questionnaires will be sent to sampled establishments via mail.

The response rate of the survey is almost 100% at present, and there is no formal imputation process in the current survey data processing system. However, the response rate of a mail survey is typically expected to be lower compared to that collected by enumerators. A study on imputation for the renewed survey is currently under way to prepare for the major change after 2019. These variables have edit constraints that the sum of the total expenses [**06**] and ending inventory [**09**] is equivalent to the sum of beginning inventory [**07**], purchases [**08**], and operating expenses [**10**], for an example. Ratio hot-deck imputation (Memobust project team 2014, e.g.) is a promising method of preparing figures that satisfies the above mentioned relation between the variables. For ratio hot-deck imputation, the nearest-neighbour of the observation with missing values in the same imputation class becomes the donor observation. The corresponding values in the donor observation to the missing recipient variables are imputed after necessary adjustments. The multivariate outlier detection step is considered prior to donor selection. The detected outlying observations in each imputation class are excluded from donor candidates so that extreme values are not used to impute missing recipient observations. This helps in obtaining good complete datasets, which leads to sound statistics.

We chose the enterprise accounting items with correlations as shown in Table 3 for the experiment in this section. The number of variables is supposed to be equal or more than the maximum number in practical application of outlier detection to clean hot-deck donor candidates in the course of survey data processing from 2019. Their fundamental statistics and correlations regarding manufacturing industry are shown in Table 4 and 5, respectively. Sales [**05**] and Total expenses [**06**] tend to have the biggest figures and highly correlated with other variables except for beginning inventory [**07**] and Ending inventory [**09**]. Those two variables are highly correlated to each other; however, both of them have lower correlations with other variables. As the skewness and kurtosis of a normal distribution are zero, all these variables have very high skewness and kurtosis.

Table 3: Target variables

No.	Variables
05	Sales
06	Total expenses
07	Beginning inventory (Inventory as of last December 31)
08	Purchases
09	Ending inventory (Inventory as of last December 31 before last)
10	Operating expenses

Table 4: Fundamental statistics of the manufacturing industry

No.	Variables	Min.	Q1	Median	Mean	Q3	Max	Skewness	Kurtosis
05	Sales	185	5,262	11,364	20,353	22,467	761,461	11.18	154.67
06	Total expenses	67	3,452	7,930	17,428	18,886	760,180	11.56	164.37
07	Beginning inventory	1	100	305	1,875	1,022	134,000	12.06	155.87
08	Purchases	5	958	2,911	8,185	6,041	498,602	11.87	167.00
09	Ending inventory	1	100	326	1,875	1,032	140,100	12.18	158.51
10	Operating expenses	50	1,930	4,623	9,243	11,261	261,578	9.03	116.44

Table 5: Correlations of the variables in the manufacturing industry

Variables			Pearson						Spearman			
No.	**05**	**06**	**07**	**08**	**09**	**10**	**05**	**06**	**07**	**08**	**09**	**10**
05	1.00	0.99	0.79	0.98	0.78	0.94	1.00	0.96	0.44	0.83	0.43	0.90
06	0.99	1.00	0.80	0.98	0.78	0.95	0.83	0.85	0.49	1.00	0.48	0.68
07	0.79	0.80	1.00	0.81	1.00	0.75	0.44	0.47	1.00	0.49	0.95	0.41
08	0.98	0.98	0.81	1.00	0.79	0.88	0.43	0.45	0.95	0.48	1.00	0.39
09	0.78	0.78	1.00	0.79	1.00	0.73	0.90	0.95	0.41	0.68	0.39	1.00
10	0.94	0.95	0.75	0.88	0.73	1.00	0.96	1.00	0.47	0.85	0.45	0.95

4.2. Data transformation and settings of outlier detection

As the target variables are highly skewed enterprise accounting items, it is evident that they require transformation prior to the outlier detection step to make the data close to an elliptical distribution. Figure 1 shows the box plots of the transformed data for the manufacturing industry as an example with no transformation, base 10 log, square root, and forth root transformation. The lambda values of the Box-Cox transformations are listed in Table 6. The suitable transformation depends on each variable. These variables are larger than zero, and they are less likely to have extremely small values. Therefore, the suitable transformation could be either the square root or the fourth root transformation. They are also better than the log transformation as they provide the advantage of being able to treat zero data.

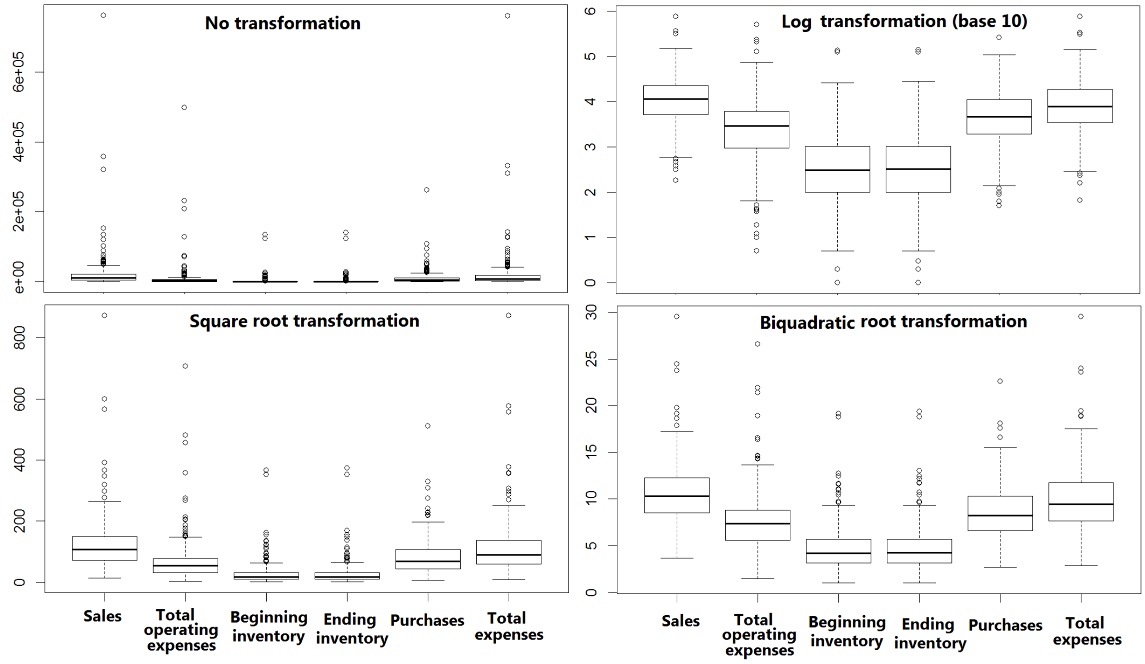

Figure 1: Box plots after transformation of the manufacturing industry

Table 6: Lambda of the Box-Cox transformations in the manufacturing industry

Variables	**05**	**06**	**07**	**08**	**09**	**10**
Lambda	0.321	0.336	0.152	0.151	0.246	0.317

The *BEM* function is used for BACON, and *msd* of the single core version is used for MSD. Their default settings are adopted except for the significance level of BACON, which is 0.999 as same as MSD. The evaluation is conducted on a virtual PC with a 3GHz Xeon processor and 2GB memory available on R.

4.3. Results obtained with the survey data

The number of outliers detected in the manufacturing industry is shown in Table 7. The result suggests that the forth root transformation is the best among the others as it detects fewer outliers than others. MSD still tends to detect fewer outliers compared to BACON with a significance level of 0.999; however, the detected outliers exceed 7%. Thus, further adjustment of the thresholds used to determine outliers may be necessary. Figure 2 shows the scatter plot matrix with the outliers detected by MSD coloured in red. The upper triangular matrix shows the outliers detected using the square root transformation, and the lower triangular matrix shows those detected using the forth square root transformation. Figure 3 is another scatter plot matrix of the same data and same outliers; however, the matrix is displayed with respective transformations. It is apparent that the forth square root transformation makes the data distribution closer to the multivariate normal compared to the square root transformation. Therefore, Figure 3 endorses that the less the elliptical model fits data, the more outliers dtected as well as Figure 1 of the univariate case. A stronger transformation tends to detect the outliers in smaller figures.

Table 7: Number of outliers detected with transformation in the manufacturing industry

Method	Transformation for outlier detection	Manufacturing No.	%
All data (greater than zero)		390	–
	Square root	47	12.05%
MSD	Biquadratic root	28	7.18%
	Log (base 10)	41	10.51%
	Square root	63	16.15%
BACON	Biquadratic root	49	12.23%
	Log (base 10)	53	13.59%

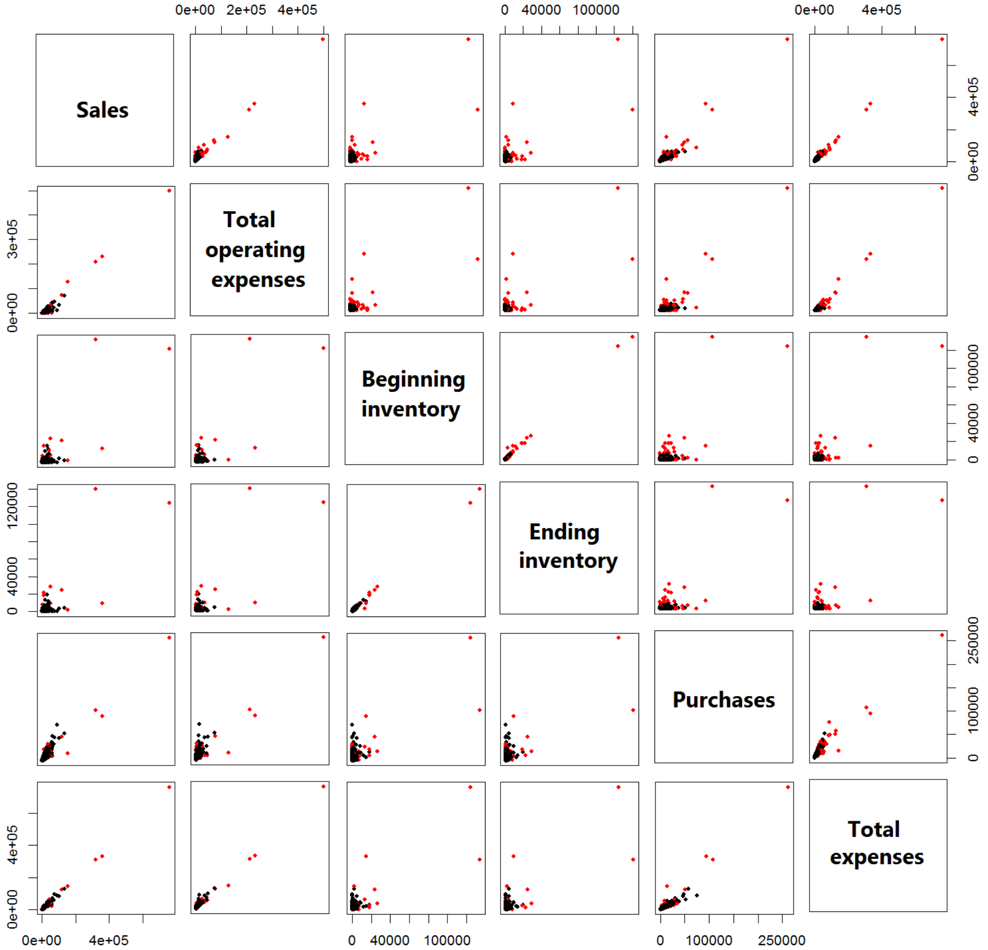

Figure 2: Outliers detected by MSD estimators in the manufacturing industry (without transformation)

Notes: Outliers shown in red in the upper triangular matrix are detected using square root transformation, and those in the lower triangular matrix are detected using biquadratic root transformation. The scatter plots are shown without any transformation.

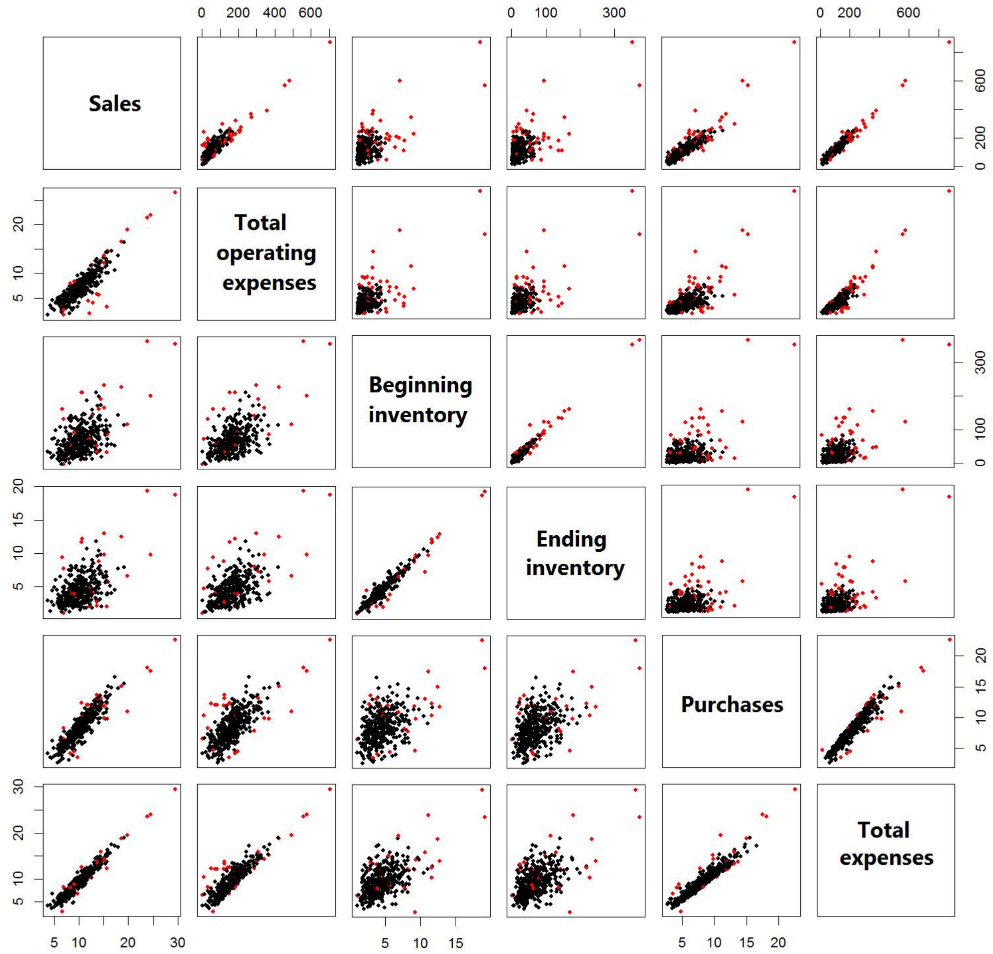

Figure 3: Outliers detected by MSD estimators in the manufacturing industry (with square root biquadratic root transformation)

Notes: Outliers shown in red in the upper triangular matrix are detected using square root transformation, and those in the lower triangular matrix are detected using biquadratic root transformation as same as Figure 2. The scatter plots are also shown with respective transformations.

We also made another experiment with the same dataset of the manufacturing industry to confirm the effect of cleaning donor candidates. The dataset contains no missing data and consists of 390 observations. We randomly extracted 20% of the observations, imputed them using the remaining 80% of the donor candidates, and evaluated their sum of deviations between true and imputed values to see if removing outliers from the donor candidates reduce the deviations.

Let the six variables of the ith observation, $y_{i,\mathbf{05}}, \ldots, y_{i,\mathbf{10}}$, and suppose the observation i is selected in the 20% to be imputed. The nearest neighbour observation m regarding the variable $y_{i,\mathbf{05}}$ is selected as a donor to impute $y_{i,\mathbf{06}}, \ldots, y_{i,\mathbf{10}}$. The ratio hot deck for the observation i is given by:

$$\hat{y}_{i,j} = y_{i,\mathbf{05}}/y_{m,\mathbf{05}} \times y_{m,j}, (j = \mathbf{06}, \ldots, \mathbf{10}).$$

MSD is used for cleaning donor candidates with the default settings. Sum of the deviations between true and imputed values are compared when donor candidates are cleaned and not cleaned. The results of Table 7 show the donor cleaning helps to improve imputation.

Table 8: Sum of the absolute deviation between imputed and true values

Variable	No cleaning	After outlier removal	Reduced rate
06	173,794	172,275	99%
07	101,144	90,082	89%
08	103,717	93,627	90%
09	191,600	183,674	96%
10	164,239	160,500	98%

Lastly, we measured processing time with the entire survey data from 2002 to 2017. After removing missing data and the data less than or equal to zero, $44,537$ observations with 6 variables were obtained. The processing times of MSD and BACON are approximately 45 seconds and 5 seconds, respectively. It is apparent that MSD is the least computationally efficient among the compared four methods in this paper. However, a processing time of less than one minute for an imputation class is not a major problem for the application of this particular survey.

5. Conclusion and future work

In this study, we examined four multivariate outlier detection methods with attractive features and found that the MSD estimators exhibit relatively good performance with skewed and heavy tailed datasets. With a survey data, MSD successfully improved ratio hot-deck imputation. In addition, we confirmed that this computationally expensive method is practically applicable because a dataset containing more than $45,000$ observations of six variables can be processed in less than one minute.

There are also a few issues to be tackled before the practical implementation of the proposed method. These issues are, for example, examining an appropriate size of the imputation class and adjustment of thresholds used to determine outliers, as cleaning of donor candidates reduces a substantial size of imputation class. While cleaning of donor candidates improve estimation, reducing the size of donor candidates may degrades estimation. We also need to consider further about an appropriate data transformation since influential outliers in terms of compiling statistics tend to have larger values, although it is obvious that the fourth square root transformation is better than the square root in terms of the elliptical model of the outlier detection methods. Further work is necessary.

References

Ando A, Kaufman GM (1965). "Bayesian Analysis of the Independent Multinormal Process — Neither Mean Nor Precision Known." *Journal of the American Statistical Association*, **60**(309), 347–358. https://doi.org/10.1080/01621459.1965.10480797.

Béguin C, Hulliger B (2003). "Robust Multivariate Outlier Detection and Imputation with Incomplete Survey Data." *EUREDIT Deliverable, D4/5.2.1/2 Part C*. URL https://www.cs.york.ac.uk/euredit/.

Billor N, Hadi AS, Velleman PF (2000). "BACON: Blocked Adaptive Computationally Efficient Outlier Nominators." *Computational Statistics & Data Analysis*, **34**(3), 279–298.

Box GEP, Cox DR (1964). "An Analysis of Transformations." *Journal of the Royal Statistical Society. Series B (Methodological)*, pp. 211–252.

Byers S, Raftery AE (1998). "Nearest-neighbor Clutter Removal for Estimating Features in Spatial Point Processes." *Journal of the American Statistical Association*, **93**(442), 577–584.

Campbell NA (1989). “Bushfire Mappping Using Noaa Ayhrr Data.” *Technical report*, Technical report, CSIRO.

Davies L (1992). “The Asymptotics of Rousseeuw’s Minimum Volume Ellipsoid Estimator.” *The Annals of Statistics*, **20**(4), 1828–1843.

Donoho DL (1982). “Breakdown Properties of Multivariate Location Estimators.” *Technical report*, Technical report, Harvard University, Boston.

Franklin S, Brodeur M (1997). “A Practical Application of a Robust Multivariate Outlier Detection Method.” In *Proceedings of the Survey Research Methods Section, American Statistical Association*, pp. 186–191.

Hadi AS (1992). “Identifying Multiple Outliers in Multivariate Data.” *Journal of the Royal Statistical Society. Series B (Methodological)*, **65**(3), 761–771.

Hadi AS, Simonoff JS (1994). “Improving the Estimation and Outlier Identification Properties of the Least Median of Squares and Minimum Volume Ellipsoid Estimators.” *Parisankhyan Sammikkha*, **1**, 61–70.

Hadi AS, Simonoff JS (1997). “A More Robust Outlier Identifier for Regression Data.” *Bulletin of the International Statistical Institute*, **14**, 281–282.

Kosinski AS (1998). “A Procedure for the Detection of Multivariate Outliers.” *Computational statistics & data analysis*, **29**(2), 145–161.

Memobust project team (2014). “Imputation under Edit Constraints.” In *The Memobust Handbook on Methodology of Modern Business Statistics*. Eurostat.

Patak Z (1990). *Robust Principal Component Analysis via Projection Pursuit*. M. Sc. thesis, University of British Columbia, Canada.

Peña D, Prieto FJ (2001). “Multivariate Outlier Detection and Robust Covariance Matrix Estimation.” *Technometrics*, **43**(3), 286–310.

Pison G, Van Aelst S, Willems G (2002). “Small Sample Corrections for LTS and MCD.” *Metrika*, **55**, 111–123.

Rousseeuw PJ (1984). “Least Median of Squares Regression.” *Journal of the American statistical association*, **79**(388), 871–880.

Rousseeuw PJ, Driessen KV (1999). “A Fast Algorithm for the Minimum Covariance Determinant Estimator.” *Technometrics*, **41**(3), 212–223.

Rousseuw PJ, Leroy AM (1987). *Robust Regression and Outlier Detection*. Wiley, New York. ISBN ISBN 978-0471488552.

Stahel WA (1981). “Breakdown of Covariance Estimators.” *Research report*, E.T.H. Zurich.

Todorov V, Filzmoser P (2009). “An Object-Oriented Framework for Robust Multivariate Analysis.” *Journal of Statistical Software*, **32**(3), 1–47. URL http://www.jstatsoft.org/v32/i03/.

Wada K (2004). “Comparison of Multivariate Outlier Detection Methods (in Japanese).” In *Proceedings of the 2004 Japannese Joint Statistical Meeting*, pp. 95–96.

Wada K (2010). “Detection of Multivariate Outliers: Modified Stahel-Donoho Estimators (in Japanese).” *Research Memoir of Official Statistics*, **67**, 89–157. URL http://www.stat.go.jp/training/2kenkyu/pdf/ihou/67/wada1.pdf.

Wada K, Tsubaki H (2013). "Parallel Computation of Modified Stahel-Donoho Estimators for Multivariate Outlier Detection." In *Cloud Computing and Big Data (CloudCom-Asia), 2013 International Conference on*, pp. 304–311. IEEE.

Wang N, Raftery AE (2002). "Nearest-neighbor Variance Estimation (NNVE) Robust Covariance Estimation via Nearest-neighbor Cleaning." *Journal of the American Statistical Association*, **97**(460), 994–1019.

Affiliation:

Kazumi Wada
National Statistics Center (NSTAC)
19-1, Wakamatsu-cho, Shinjyuku-ku, Tokyo, 162-8668, Japan
E-mail: kazwd2008@gmail.com

Mariko Kawano
National Statistics Center (NSTAC)
19-1, Wakamatsu-cho, Shinjyuku-ku, Tokyo, 162-8668, Japan
E-mail: mkawano@nstac.go.jp

Hiroe Tsubaki
National Statistics Center (NSTAC)
19-1, Wakamatsu-cho, Shinjyuku-ku, Tokyo, 162-8668, Japan
E-mail: htsubaki@nstac.go.jp

Austrian Journal of Statistics http://www.ajs.or.at/
published by the Austrian Society of Statistics http://www.osg.or.at/
Volume 49 *Submitted:* 2018-09-24
February 2020 *Accepted:* 2019-02-14

Table 2: Comparison with contaminated skew-t datasets (Correlation : 0.4)

DF	SD	Dist-ance	Outliers	Skewness 0												Skewness 01											
				MSD			BEM			MCD			NNVE			MSD			BEM			MCD			NNVE		
		δ	α	False positive	Leakage	Total rate	False positive	Leakage	Total rate	False positive	Leakage	Total rate	False positive	Leakage	Total rate	False positive	Leakage	Total rate	False positive	Leakage	Total rate	False positive	Leakage	Total rate	False positive	Leakage	Total rate
2	-	-	0	22%	-	78%	38%	-	62%	38%	-	62%	19%	-	81%	29%	-	71%	56%	-	44%	45%	-	55%	9%	-	91%
	1	10	0.1	20%	0%	82%	42%	0%	62%	37%	0%	67%	26%	0%	77%	17%	0%	85%	42%	0%	62%	33%	0%	70%	10%	100%	81%
			0.2	13%	50%	80%	53%	0%	58%	30%	0%	76%	11%	100%	71%	20%	55%	73%	40%	0%	68%	31%	0%	75%	15%	100%	68%
			0.3	19%	100%	57%	31%	0%	78%	31%	83%	53%	19%	100%	57%	16%	97%	60%	43%	0%	70%	36%	70%	54%	19%	100%	57%
			0.4	22%	100%	47%	23%	0%	86%	32%	95%	43%	13%	100%	52%	15%	100%	51%	30%	0%	82%	32%	90%	45%	15%	100%	51%
		100	0.1	20%	0%	82%	41%	0%	63%	39%	0%	65%	6%	0%	95%	16%	0%	86%	40%	0%	64%	37%	0%	67%	9%	0%	92%
			0.2	10%	0%	92%	54%	0%	57%	29%	0%	77%	11%	100%	71%	19%	0%	85%	36%	0%	71%	31%	0%	75%	16%	100%	67%
			0.3	19%	0%	87%	31%	0%	78%	21%	0%	85%	24%	100%	53%	14%	0%	90%	49%	0%	66%	21%	0%	85%	16%	100%	59%
			0.4	18%	0%	89%	23%	0%	86%	33%	88%	45%	27%	100%	44%	15%	0%	91%	30%	0%	82%	30%	88%	47%	17%	100%	50%
	5	10	0.1	21%	0%	81%	41%	0%	63%	39%	0%	65%	18%	0%	84%	16%	0%	86%	42%	0%	62%	37%	0%	67%	13%	0%	88%
			0.2	10%	0%	92%	55%	0%	56%	30%	0%	76%	11%	0%	91%	16%	0%	87%	36%	0%	71%	31%	0%	75%	3%	85%	81%
			0.3	16%	0%	89%	31%	0%	78%	21%	0%	85%	11%	0%	92%	11%	0%	92%	49%	0%	66%	21%	0%	85%	16%	0%	89%
			0.4	7%	0%	96%	23%	0%	86%	8%	0%	95%	5%	88%	62%	13%	0%	92%	30%	0%	82%	8%	0%	95%	3%	90%	62%
		100	0.1	20%	0%	82%	41%	0%	63%	39%	0%	65%	6%	0%	95%	16%	0%	86%	40%	0%	64%	39%	0%	65%	10%	0%	91%
			0.2	10%	0%	92%	53%	0%	58%	30%	0%	76%	11%	0%	91%	16%	0%	87%	40%	0%	68%	31%	0%	75%	9%	0%	93%
			0.3	16%	0%	89%	31%	0%	78%	21%	0%	85%	17%	0%	88%	11%	0%	92%	49%	0%	66%	21%	0%	85%	16%	0%	89%
			0.4	8%	0%	95%	23%	0%	86%	8%	0%	95%	5%	73%	68%	12%	0%	93%	30%	0%	82%	8%	0%	95%	3%	73%	69%
10	-	-	0	1%	-	99%	8%	-	92%	15%	-	85%	34%	-	66%	2%	-	98%	4%	-	96%	17%	-	83%	28%	-	72%
	1	10	0.1	4%	0%	96%	7%	0%	94%	13%	0%	88%	16%	0%	86%	0%	0%	100%	0%	0%	100%	10%	0%	91%	7%	0%	94%
			0.2	0%	0%	100%	0%	0%	100%	6%	0%	95%	20%	0%	84%	0%	0%	100%	1%	0%	99%	5%	0%	96%	25%	0%	80%
			0.3	0%	90%	73%	7%	0%	95%	9%	0%	94%	16%	70%	68%	1%	97%	70%	9%	0%	94%	9%	0%	94%	23%	60%	66%
			0.4	2%	100%	59%	7%	0%	96%	3%	0%	98%	15%	85%	57%	0%	100%	60%	2%	0%	99%	5%	83%	64%	35%	25%	69%
		100	0.1	4%	0%	96%	7%	0%	94%	12%	0%	89%	1%	0%	99%	0%	0%	100%	0%	0%	100%	9%	0%	92%	1%	0%	99%
			0.2	0%	0%	100%	0%	0%	100%	3%	0%	98%	28%	0%	78%	0%	0%	100%	1%	0%	99%	6%	0%	95%	35%	0%	72%
			0.3	0%	0%	100%	7%	0%	95%	7%	0%	95%	3%	100%	68%	3%	0%	98%	9%	0%	94%	9%	0%	94%	6%	100%	66%
			0.4	3%	0%	98%	7%	0%	96%	3%	0%	98%	13%	95%	54%	2%	0%	99%	2%	0%	99%	3%	0%	99%	3%	98%	59%
	5	10	0.1	4%	0%	96%	7%	0%	94%	13%	0%	88%	7%	0%	94%	0%	0%	100%	0%	0%	100%	12%	0%	89%	2%	0%	98%
			0.2	0%	0%	100%	0%	0%	100%	4%	0%	97%	4%	0%	97%	0%	0%	100%	1%	0%	99%	5%	0%	96%	0%	0%	100%
			0.3	0%	0%	100%	7%	0%	95%	9%	0%	94%	3%	0%	98%	0%	0%	100%	9%	0%	94%	9%	0%	94%	1%	0%	99%
			0.4	2%	0%	99%	7%	0%	96%	3%	0%	98%	2%	0%	99%	0%	0%	100%	2%	0%	99%	2%	0%	99%	3%	0%	98%
		100	0.1	4%	0%	96%	7%	0%	94%	13%	0%	88%	0%	0%	100%	0%	0%	100%	0%	0%	100%	8%	0%	93%	1%	0%	99%
			0.2	0%	0%	100%	0%	0%	100%	4%	0%	97%	1%	0%	99%	0%	0%	100%	1%	0%	99%	4%	0%	97%	0%	0%	100%
			0.3	0%	0%	100%	7%	0%	95%	6%	0%	96%	3%	0%	98%	0%	0%	100%	9%	0%	94%	9%	0%	94%	4%	0%	97%
			0.4	2%	0%	99%	7%	0%	96%	3%	0%	98%	8%	0%	95%	0%	0%	100%	2%	0%	99%	2%	0%	99%	3%	0%	98%
Inf	-	-	0	1%	-	99%	1%	-	99%	5%	-	95%	38%	-	62%	0%	-	100%	0%	-	100%	1%	-	99%	29%	-	71%
	1	10	0.1	0%	0%	100%	0%	0%	100%	4%	0%	96%	3%	0%	97%	0%	0%	100%	0%	0%	100%	3%	0%	97%	2%	0%	98%
			0.2	0%	0%	100%	0%	0%	100%	1%	0%	99%	8%	0%	94%	0%	0%	100%	4%	0%	97%	8%	0%	94%	13%	0%	90%
			0.3	0%	87%	74%	0%	0%	100%	0%	0%	100%	3%	70%	77%	0%	87%	74%	0%	0%	100%	1%	0%	99%	9%	63%	75%
			0.4	0%	100%	60%	2%	0%	99%	5%	78%	66%	2%	45%	81%	0%	100%	60%	2%	0%	99%	0%	0%	100%	12%	35%	79%
		100	0.1	0%	0%	100%	0%	0%	100%	7%	0%	94%	1%	0%	99%	0%	0%	100%	0%	0%	100%	1%	0%	99%	0%	0%	100%
			0.2	0%	0%	100%	0%	0%	100%	0%	0%	100%	18%	0%	86%	0%	0%	100%	4%	0%	97%	4%	0%	97%	16%	0%	87%
			0.3	0%	0%	100%	0%	0%	100%	0%	0%	100%	23%	13%	80%	0%	0%	100%	0%	0%	100%	1%	0%	99%	24%	10%	80%
			0.4	0%	0%	100%	2%	0%	99%	3%	83%	65%	0%	100%	60%	0%	0%	100%	2%	0%	99%	0%	0%	100%	0%	100%	60%
	5	10	0.1	0%	0%	100%	0%	0%	100%	7%	0%	94%	1%	0%	99%	0%	0%	100%	0%	0%	100%	4%	0%	96%	0%	0%	100%
			0.2	0%	0%	100%	0%	0%	100%	1%	0%	99%	0%	0%	100%	0%	0%	100%	4%	0%	97%	5%	0%	96%	3%	0%	98%
			0.3	0%	0%	100%	0%	0%	100%	0%	0%	100%	0%	0%	100%	0%	0%	100%	0%	0%	100%	1%	0%	99%	0%	0%	100%
			0.4	0%	0%	100%	2%	0%	99%	2%	0%	99%	0%	0%	100%	0%	0%	100%	2%	0%	99%	0%	0%	100%	2%	0%	99%
		100	0.1	0%	0%	100%	0%	0%	100%	4%	0%	96%	1%	0%	99%	0%	0%	100%	0%	0%	100%	1%	0%	99%	0%	0%	100%
			0.2	0%	0%	100%	0%	0%	100%	4%	0%	97%	0%	0%	100%	0%	0%	100%	4%	0%	97%	4%	0%	97%	1%	0%	99%
			0.3	0%	0%	100%	0%	0%	100%	0%	0%	100%	0%	0%	100%	0%	0%	100%	0%	0%	100%	1%	0%	99%	0%	0%	100%
			0.4	0%	0%	100%	2%	0%	99%	2%	0%	99%	0%	0%	100%	0%	0%	100%	2%	0%	99%	0%	0%	100%	0%	0%	100%

Table 2: Comparison with contaminated skew-t datasets (Correlation : 0.4)

DF	SD	Dist-ance	Outliers	Skewness 05												Skewness 10											
				MSD			BEM			MCD			NNVE			MSD			BEM			MCD			NNVE		
		δ	α	False positive	Leakage	Total rate	False positive	Leakage	Total rate	False positive	Leakage	Total rate	False positive	Leakage	Total rate	False positive	Leakage	Total rate	False positive	Leakage	Total rate	False positive	Leakage	Total rate	False positive	Leakage	Total rate
2	-	-	0	22%	-	78%	34%	-	66%	34%	-	66%	7%	-	93%	14%	-	86%	34%	-	66%	37%	-	63%	13%	-	87%
	1	10	0.1	14%	0%	87%	40%	0%	64%	34%	0%	69%	24%	0%	78%	14%	0%	87%	39%	0%	65%	36%	0%	68%	14%	0%	87%
			0.2	20%	100%	64%	49%	0%	61%	30%	0%	76%	21%	100%	63%	11%	30%	85%	30%	0%	76%	29%	0%	77%	6%	100%	75%
			0.3	11%	100%	62%	29%	0%	80%	27%	93%	53%	19%	100%	57%	13%	100%	61%	40%	0%	72%	37%	90%	47%	23%	100%	54%
			0.4	12%	100%	53%	22%	0%	87%	22%	93%	50%	17%	100%	50%	22%	100%	47%	45%	0%	73%	35%	80%	47%	25%	100%	45%
		100	0.1	13%	0%	88%	36%	0%	68%	34%	0%	69%	8%	0%	93%	14%	0%	87%	40%	0%	64%	29%	0%	74%	8%	0%	93%
			0.2	18%	0%	86%	51%	0%	59%	31%	0%	75%	24%	100%	61%	11%	0%	91%	30%	0%	76%	28%	0%	78%	8%	100%	74%
			0.3	10%	0%	93%	29%	0%	80%	21%	0%	85%	17%	100%	58%	14%	0%	90%	40%	0%	72%	21%	0%	85%	14%	100%	60%
			0.4	12%	0%	93%	23%	0%	86%	8%	0%	95%	17%	100%	50%	22%	0%	87%	45%	0%	73%	37%	85%	44%	25%	100%	45%
	5	10	0.1	14%	0%	87%	40%	0%	64%	34%	0%	69%	21%	0%	81%	14%	0%	87%	40%	0%	64%	32%	0%	71%	14%	0%	87%
			0.2	15%	0%	88%	53%	0%	58%	31%	0%	75%	16%	0%	87%	11%	0%	91%	29%	0%	77%	28%	0%	78%	9%	0%	93%
			0.3	10%	0%	93%	30%	0%	79%	21%	0%	85%	16%	0%	89%	10%	0%	93%	40%	0%	72%	21%	0%	85%	13%	0%	91%
			0.4	7%	0%	96%	22%	0%	87%	8%	0%	95%	15%	0%	91%	12%	0%	93%	45%	0%	73%	8%	0%	95%	7%	75%	66%
		100	0.1	13%	0%	88%	40%	0%	64%	34%	0%	69%	6%	0%	95%	14%	0%	87%	40%	0%	64%	32%	0%	71%	7%	0%	94%
			0.2	14%	0%	89%	51%	0%	59%	31%	0%	75%	16%	0%	87%	9%	0%	93%	30%	0%	76%	28%	0%	78%	8%	0%	94%
			0.3	7%	0%	95%	31%	0%	78%	14%	0%	90%	14%	0%	90%	7%	0%	95%	40%	0%	72%	21%	0%	85%	13%	0%	91%
			0.4	5%	0%	97%	22%	0%	87%	8%	0%	95%	15%	0%	91%	12%	0%	93%	45%	0%	73%	8%	0%	95%	5%	35%	83%
10	-	-	0	2%	-	98%	2%	-	98%	14%	-	86%	26%	-	74%	4%	-	96%	7%	-	93%	16%	-	84%	23%	-	77%
	1	10	0.1	4%	0%	96%	7%	0%	94%	20%	0%	82%	7%	0%	94%	2%	0%	98%	13%	0%	88%	18%	0%	84%	7%	0%	94%
			0.2	1%	0%	99%	9%	0%	93%	10%	0%	92%	15%	0%	88%	0%	0%	100%	9%	0%	93%	15%	0%	88%	23%	0%	82%
			0.3	3%	70%	77%	10%	0%	93%	9%	0%	94%	21%	30%	76%	1%	63%	80%	3%	0%	98%	4%	0%	97%	50%	0%	65%
			0.4	0%	100%	60%	12%	0%	93%	7%	0%	96%	20%	40%	72%	0%	100%	60%	3%	0%	98%	0%	0%	100%	17%	38%	75%
		100	0.1	4%	0%	96%	7%	0%	94%	18%	0%	84%	2%	0%	98%	1%	0%	99%	13%	0%	88%	18%	0%	84%	4%	0%	96%
			0.2	1%	0%	99%	9%	0%	93%	9%	0%	93%	15%	0%	88%	0%	0%	100%	9%	0%	93%	13%	0%	90%	28%	0%	78%
			0.3	3%	0%	98%	10%	0%	93%	10%	0%	93%	21%	10%	82%	1%	0%	99%	3%	0%	98%	6%	0%	96%	21%	10%	82%
			0.4	2%	0%	99%	13%	0%	92%	7%	0%	96%	5%	100%	57%	0%	0%	100%	3%	0%	98%	2%	0%	99%	2%	100%	59%
	5	10	0.1	4%	0%	96%	7%	0%	94%	18%	0%	84%	7%	0%	94%	1%	0%	99%	13%	0%	88%	18%	0%	84%	4%	0%	96%
			0.2	1%	0%	99%	9%	0%	93%	13%	0%	90%	3%	0%	98%	0%	0%	100%	9%	0%	93%	13%	0%	90%	3%	0%	98%
			0.3	1%	0%	99%	9%	0%	94%	9%	0%	94%	4%	0%	97%	0%	0%	100%	3%	0%	98%	9%	0%	94%	0%	0%	100%
			0.4	0%	0%	100%	12%	0%	93%	8%	0%	95%	5%	0%	97%	0%	0%	100%	3%	0%	98%	2%	0%	99%	2%	0%	99%
		100	0.1	4%	0%	96%	7%	0%	94%	20%	0%	82%	2%	0%	98%	1%	0%	99%	13%	0%	88%	18%	0%	84%	2%	0%	98%
			0.2	1%	0%	99%	9%	0%	93%	8%	0%	94%	1%	0%	99%	0%	0%	100%	9%	0%	93%	13%	0%	90%	3%	0%	98%
			0.3	1%	0%	99%	9%	0%	94%	9%	0%	94%	6%	0%	96%	0%	0%	100%	3%	0%	98%	9%	0%	94%	1%	0%	99%
			0.4	0%	0%	100%	13%	0%	92%	8%	0%	95%	5%	0%	97%	0%	0%	100%	3%	0%	98%	2%	0%	99%	2%	0%	99%
Inf	-	-	0	0%	-	100%	0%	-	100%	4%	-	96%	26%	-	74%	0%	-	100%	0%	-	100%	4%	-	96%	36%	-	64%
	1	10	0.1	0%	0%	100%	0%	0%	100%	3%	0%	97%	3%	0%	97%	0%	0%	100%	1%	0%	99%	1%	0%	99%	2%	0%	98%
			0.2	0%	0%	100%	0%	0%	100%	1%	0%	99%	8%	0%	94%	0%	0%	100%	4%	0%	97%	5%	0%	96%	11%	0%	91%
			0.3	0%	10%	97%	1%	0%	99%	1%	0%	99%	16%	7%	87%	0%	70%	79%	3%	0%	98%	3%	0%	98%	9%	30%	85%
			0.4	0%	100%	60%	0%	0%	100%	2%	0%	99%	3%	50%	78%	0%	100%	60%	0%	0%	100%	2%	0%	99%	17%	23%	81%
		100	0.1	0%	0%	100%	0%	0%	100%	4%	0%	96%	0%	0%	100%	0%	0%	100%	1%	0%	99%	3%	0%	97%	0%	0%	100%
			0.2	0%	0%	100%	0%	0%	100%	1%	0%	99%	10%	0%	92%	0%	0%	100%	4%	0%	97%	5%	0%	96%	11%	0%	91%
			0.3	0%	0%	100%	1%	0%	99%	3%	0%	98%	16%	10%	86%	0%	0%	100%	3%	0%	98%	3%	0%	98%	20%	7%	84%
			0.4	0%	0%	100%	0%	0%	100%	2%	0%	99%	0%	98%	61%	0%	0%	100%	0%	0%	100%	2%	0%	99%	33%	8%	77%
	5	10	0.1	0%	0%	100%	0%	0%	100%	6%	0%	95%	0%	0%	100%	0%	0%	100%	1%	0%	99%	3%	0%	97%	1%	0%	99%
			0.2	0%	0%	100%	0%	0%	100%	1%	0%	99%	3%	0%	98%	0%	0%	100%	4%	0%	97%	4%	0%	97%	0%	0%	100%
			0.3	0%	0%	100%	1%	0%	99%	3%	0%	98%	0%	0%	100%	0%	0%	100%	3%	0%	98%	3%	0%	98%	1%	0%	99%
			0.4	0%	0%	100%	0%	0%	100%	2%	0%	99%	2%	0%	99%	0%	0%	100%	0%	0%	100%	2%	0%	99%	0%	0%	100%
		100	0.1	0%	0%	100%	0%	0%	100%	7%	0%	94%	0%	0%	100%	0%	0%	100%	1%	0%	99%	2%	0%	98%	0%	0%	100%
			0.2	0%	0%	100%	0%	0%	100%	1%	0%	99%	1%	0%	99%	0%	0%	100%	4%	0%	97%	5%	0%	96%	0%	0%	100%
			0.3	0%	0%	100%	1%	0%	99%	3%	0%	98%	0%	0%	100%	0%	0%	100%	3%	0%	98%	3%	0%	98%	0%	0%	100%
			0.4	0%	0%	100%	0%	0%	100%	2%	0%	99%	2%	0%	99%	0%	0%	100%	0%	0%	100%	2%	0%	99%	0%	0%	100%

Table 2: Comparison with contaminated skew-t datasets (Correlation : 0.8)

DF	SD	Distance	Outliers	Skewness 0												Skewness 01											
				MSD			BEM			MCD			NNVE			MSD			BEM			MCD			NNVE		
		δ	α	False positive	Leakage	Total rate	False positive	Leakage	Total rate	False positive	Leakage	Total rate	False positive	Leakage	Total rate	False positive	Leakage	Total rate	False positive	Leakage	Total rate	False positive	Leakage	Total rate	False positive	Leakage	Total rate
2	-	-	0	22%	-	78%	33%	-	67%	35%	-	65%	21%	-	79%	9%	-	91%	22%	-	78%	31%	-	69%	14%	-	86%
	1	10	0.1	18%	0%	84%	33%	0%	70%	33%	0%	70%	18%	0%	84%	19%	0%	83%	37%	0%	67%	34%	0%	69%	23%	0%	79%
			0.2	15%	0%	88%	33%	0%	74%	31%	0%	75%	11%	95%	72%	15%	0%	88%	33%	0%	74%	30%	0%	76%	26%	0%	79%
			0.3	11%	0%	92%	21%	0%	85%	21%	0%	85%	26%	67%	62%	16%	23%	82%	29%	0%	80%	21%	0%	85%	20%	93%	58%
			0.4	10%	95%	56%	27%	0%	84%	25%	33%	72%	18%	100%	49%	8%	75%	65%	28%	0%	83%	20%	18%	81%	13%	100%	52%
		100	0.1	17%	0%	85%	33%	0%	70%	33%	0%	70%	11%	0%	90%	19%	0%	83%	37%	0%	67%	34%	0%	69%	3%	0%	97%
			0.2	13%	0%	90%	33%	0%	74%	31%	0%	75%	13%	100%	70%	13%	0%	90%	33%	0%	74%	29%	0%	77%	13%	100%	70%
			0.3	10%	0%	93%	21%	0%	85%	21%	0%	85%	16%	100%	59%	13%	0%	91%	29%	0%	80%	21%	0%	85%	16%	100%	59%
			0.4	15%	0%	91%	27%	0%	84%	10%	0%	94%	25%	100%	45%	12%	0%	93%	30%	0%	82%	22%	20%	79%	15%	93%	54%
	5	10	0.1	19%	0%	83%	33%	0%	70%	33%	0%	70%	18%	0%	84%	19%	0%	83%	37%	0%	67%	36%	0%	68%	23%	0%	79%
			0.2	14%	0%	89%	31%	0%	75%	31%	0%	75%	11%	0%	91%	14%	0%	89%	33%	0%	74%	29%	0%	77%	5%	0%	96%
			0.3	11%	0%	92%	21%	0%	85%	21%	0%	85%	16%	0%	89%	13%	0%	91%	29%	0%	80%	21%	0%	85%	14%	0%	90%
			0.4	10%	0%	94%	27%	0%	84%	8%	0%	95%	13%	0%	92%	5%	0%	97%	30%	0%	82%	8%	0%	95%	15%	0%	91%
		100	0.1	17%	0%	85%	33%	0%	70%	33%	0%	70%	9%	0%	92%	18%	0%	84%	34%	0%	69%	36%	0%	68%	1%	0%	99%
			0.2	13%	0%	90%	34%	0%	73%	31%	0%	75%	11%	0%	91%	13%	0%	90%	33%	0%	74%	28%	0%	78%	4%	0%	97%
			0.3	9%	0%	94%	21%	0%	85%	21%	0%	85%	16%	0%	89%	11%	0%	92%	29%	0%	80%	21%	0%	85%	14%	0%	90%
			0.4	10%	0%	94%	27%	0%	84%	8%	0%	95%	17%	0%	90%	3%	0%	98%	28%	0%	83%	8%	0%	95%	13%	0%	92%
10	-	-	0	3%	-	97%	6%	-	94%	12%	-	88%	6%	-	94%	2%	-	98%	3%	-	97%	18%	-	82%	22%	-	78%
	1	10	0.1	2%	0%	98%	3%	0%	97%	19%	0%	83%	11%	0%	90%	1%	0%	99%	4%	0%	96%	10%	0%	91%	8%	0%	93%
			0.2	1%	0%	99%	4%	0%	97%	13%	0%	90%	13%	0%	90%	0%	0%	100%	1%	0%	99%	13%	0%	90%	14%	0%	89%
			0.3	0%	0%	100%	4%	0%	97%	9%	0%	94%	17%	0%	88%	0%	0%	100%	7%	0%	95%	10%	0%	93%	21%	0%	85%
			0.4	0%	40%	84%	2%	0%	99%	0%	0%	100%	15%	0%	91%	2%	50%	79%	5%	0%	97%	7%	0%	96%	12%	0%	93%
		100	0.1	2%	0%	98%	3%	0%	97%	20%	0%	82%	0%	0%	100%	1%	0%	99%	4%	0%	96%	10%	0%	91%	3%	0%	97%
			0.2	1%	0%	99%	5%	0%	96%	10%	0%	92%	15%	0%	88%	0%	0%	100%	1%	0%	99%	11%	0%	91%	11%	0%	91%
			0.3	0%	0%	100%	6%	0%	96%	9%	0%	94%	19%	0%	87%	0%	0%	100%	7%	0%	95%	9%	0%	94%	17%	0%	88%
			0.4	0%	0%	100%	2%	0%	99%	0%	0%	100%	7%	88%	61%	2%	0%	99%	5%	0%	97%	7%	0%	96%	12%	0%	93%
	5	10	0.1	2%	0%	98%	3%	0%	97%	19%	0%	83%	6%	0%	95%	1%	0%	99%	4%	0%	96%	7%	0%	94%	8%	0%	93%
			0.2	1%	0%	99%	4%	0%	97%	10%	0%	92%	10%	0%	92%	0%	0%	100%	3%	0%	98%	10%	0%	92%	6%	0%	95%
			0.3	0%	0%	100%	4%	0%	97%	9%	0%	94%	7%	0%	95%	0%	0%	100%	7%	0%	95%	9%	0%	94%	1%	0%	99%
			0.4	0%	0%	100%	2%	0%	99%	3%	0%	98%	7%	0%	96%	2%	0%	99%	10%	0%	94%	7%	0%	96%	5%	0%	97%
		100	0.1	1%	0%	99%	3%	0%	97%	23%	0%	79%	4%	0%	96%	1%	0%	99%	4%	0%	96%	10%	0%	91%	4%	0%	96%
			0.2	1%	0%	99%	5%	0%	96%	11%	0%	91%	4%	0%	97%	0%	0%	100%	3%	0%	98%	11%	0%	91%	4%	0%	97%
			0.3	0%	0%	100%	4%	0%	97%	9%	0%	94%	7%	0%	95%	0%	0%	100%	7%	0%	95%	10%	0%	93%	4%	0%	97%
			0.4	0%	0%	100%	2%	0%	99%	3%	0%	98%	10%	0%	94%	0%	0%	100%	12%	0%	93%	7%	0%	96%	3%	0%	98%
Inf	-	-	0	0%	-	100%	0%	-	100%	1%	-	99%	8%	-	92%	0%	-	100%	0%	-	100%	3%	-	97%	13%	-	87%
	1	10	0.1	0%	0%	100%	0%	0%	100%	3%	0%	97%	1%	0%	99%	0%	0%	100%	0%	0%	100%	3%	0%	97%	3%	0%	97%
			0.2	0%	0%	100%	0%	0%	100%	1%	0%	99%	14%	0%	89%	0%	0%	100%	1%	0%	99%	8%	0%	94%	13%	0%	90%
			0.3	0%	0%	100%	0%	0%	100%	1%	0%	99%	13%	0%	91%	0%	0%	100%	0%	0%	100%	0%	0%	100%	9%	0%	94%
			0.4	0%	43%	83%	2%	0%	99%	2%	0%	99%	10%	0%	94%	0%	28%	89%	2%	0%	99%	2%	0%	99%	10%	0%	94%
		100	0.1	0%	0%	100%	0%	0%	100%	2%	0%	98%	0%	0%	100%	0%	0%	100%	0%	0%	100%	2%	0%	98%	0%	0%	100%
			0.2	0%	0%	100%	0%	0%	100%	1%	0%	99%	15%	0%	88%	0%	0%	100%	1%	0%	99%	6%	0%	95%	10%	0%	92%
			0.3	0%	0%	100%	0%	0%	100%	1%	0%	99%	9%	0%	94%	0%	0%	100%	0%	0%	100%	0%	0%	100%	9%	0%	94%
			0.4	0%	0%	100%	3%	0%	98%	2%	0%	99%	7%	0%	96%	0%	0%	100%	2%	0%	99%	2%	0%	99%	10%	0%	94%
	5	10	0.1	0%	0%	100%	0%	0%	100%	1%	0%	99%	0%	0%	100%	0%	0%	100%	0%	0%	100%	2%	0%	98%	3%	0%	97%
			0.2	0%	0%	100%	0%	0%	100%	1%	0%	99%	4%	0%	97%	0%	0%	100%	1%	0%	99%	1%	0%	99%	3%	0%	98%
			0.3	0%	0%	100%	0%	0%	100%	1%	0%	99%	1%	0%	99%	0%	0%	100%	0%	0%	100%	0%	0%	100%	7%	0%	95%
			0.4	0%	0%	100%	2%	0%	99%	2%	0%	99%	3%	0%	98%	0%	0%	100%	0%	0%	100%	2%	0%	99%	2%	0%	99%
		100	0.1	0%	0%	100%	0%	0%	100%	2%	0%	98%	0%	0%	100%	0%	0%	100%	0%	0%	100%	1%	0%	99%	0%	0%	100%
			0.2	0%	0%	100%	0%	0%	100%	1%	0%	99%	4%	0%	97%	0%	0%	100%	1%	0%	99%	4%	0%	97%	3%	0%	98%
			0.3	0%	0%	100%	0%	0%	100%	1%	0%	99%	1%	0%	99%	0%	0%	100%	0%	0%	100%	0%	0%	100%	3%	0%	98%
			0.4	0%	0%	100%	3%	0%	98%	2%	0%	99%	0%	0%	100%	0%	0%	100%	0%	0%	100%	2%	0%	99%	0%	0%	100%

Table 2: Comparison with contaminated skew-t datasets (Correlation : 0.8)

DF	SD	Dist-ance	Outliers	Skewness 05												Skewness 10											
				MSD			BEM			MCD			NNVE			MSD			BEM			MCD			NNVE		
		δ	α	False positive	Leakage	Total rate	False positive	Leakage	Total rate	False positive	Leakage	Total rate	False positive	Leakage	Total rate	False positive	Leakage	Total rate	False positive	Leakage	Total rate	False positive	Leakage	Total rate	False positive	Leakage	Total rate
2	-	-	0	19%	-	81%	44%	-	56%	38%	-	62%	10%	-	90%	28%	-	72%	39%	-	61%	44%	-	56%	10%	-	90%
	1	10	0.1	18%	0%	84%	32%	0%	71%	33%	0%	70%	17%	0%	85%	13%	0%	88%	40%	0%	64%	38%	0%	66%	21%	0%	81%
			0.2	20%	0%	84%	36%	0%	71%	31%	0%	75%	16%	100%	67%	20%	0%	84%	38%	0%	70%	31%	0%	75%	8%	100%	74%
			0.3	11%	3%	91%	37%	0%	74%	21%	0%	85%	9%	100%	64%	14%	0%	90%	34%	0%	76%	21%	0%	85%	14%	97%	61%
			0.4	8%	95%	57%	22%	0%	87%	25%	35%	71%	8%	100%	55%	7%	73%	67%	33%	0%	80%	17%	20%	82%	7%	98%	57%
		100	0.1	17%	0%	85%	32%	0%	71%	36%	0%	68%	4%	0%	96%	11%	0%	90%	40%	0%	64%	39%	0%	65%	9%	0%	92%
			0.2	19%	0%	85%	36%	0%	71%	31%	0%	75%	18%	100%	66%	18%	0%	86%	36%	0%	71%	30%	0%	76%	9%	100%	73%
			0.3	14%	0%	90%	37%	0%	74%	21%	0%	85%	7%	100%	65%	13%	0%	91%	34%	0%	76%	21%	0%	85%	13%	100%	61%
			0.4	10%	0%	94%	22%	0%	87%	8%	0%	95%	8%	100%	55%	8%	0%	95%	18%	0%	89%	8%	0%	95%	8%	93%	58%
	5	10	0.1	18%	0%	84%	28%	0%	75%	32%	0%	71%	18%	0%	84%	12%	0%	89%	40%	0%	64%	38%	0%	66%	16%	0%	86%
			0.2	18%	0%	86%	35%	0%	72%	31%	0%	75%	18%	0%	86%	15%	0%	88%	35%	0%	72%	31%	0%	75%	11%	0%	91%
			0.3	10%	0%	93%	37%	0%	74%	21%	0%	85%	6%	0%	96%	13%	0%	91%	34%	0%	76%	21%	0%	85%	11%	0%	92%
			0.4	7%	0%	96%	22%	0%	87%	8%	0%	95%	3%	83%	65%	7%	0%	96%	33%	0%	80%	8%	0%	95%	8%	0%	95%
		100	0.1	16%	0%	86%	31%	0%	72%	32%	0%	71%	4%	0%	96%	12%	0%	89%	42%	0%	62%	38%	0%	66%	9%	0%	92%
			0.2	18%	0%	86%	36%	0%	71%	31%	0%	75%	18%	0%	86%	14%	0%	89%	38%	0%	70%	31%	0%	75%	13%	0%	90%
			0.3	9%	0%	94%	37%	0%	74%	21%	0%	85%	6%	0%	96%	11%	0%	92%	34%	0%	76%	21%	0%	85%	14%	0%	90%
			0.4	7%	0%	96%	22%	0%	87%	8%	0%	95%	3%	73%	69%	7%	0%	96%	37%	0%	78%	8%	0%	95%	8%	0%	95%
10	-	-	0	5%	-	95%	8%	-	92%	17%	-	83%	16%	-	84%	2%	-	98%	5%	-	95%	10%	-	90%	10%	-	90%
	1	10	0.1	1%	0%	99%	6%	0%	95%	16%	0%	86%	6%	0%	95%	3%	0%	97%	16%	0%	86%	20%	0%	82%	4%	0%	96%
			0.2	1%	0%	99%	5%	0%	96%	6%	0%	95%	9%	0%	93%	3%	0%	98%	11%	0%	91%	14%	0%	89%	20%	0%	84%
			0.3	4%	0%	97%	10%	0%	93%	13%	0%	91%	20%	0%	86%	3%	0%	98%	14%	0%	90%	13%	0%	91%	21%	0%	85%
			0.4	0%	43%	83%	8%	0%	95%	8%	0%	95%	10%	0%	94%	2%	23%	90%	8%	0%	95%	7%	0%	96%	25%	0%	85%
		100	0.1	1%	0%	99%	6%	0%	95%	13%	0%	88%	0%	0%	100%	1%	0%	99%	16%	0%	86%	17%	0%	85%	2%	0%	98%
			0.2	0%	0%	100%	5%	0%	96%	6%	0%	95%	19%	0%	85%	3%	0%	98%	11%	0%	91%	15%	0%	88%	21%	0%	83%
			0.3	4%	0%	97%	10%	0%	93%	14%	0%	90%	19%	0%	87%	3%	0%	98%	14%	0%	90%	14%	0%	90%	16%	0%	89%
			0.4	2%	0%	99%	8%	0%	95%	8%	0%	95%	13%	0%	92%	2%	0%	99%	8%	0%	95%	7%	0%	96%	25%	0%	85%
	5	10	0.1	1%	0%	99%	6%	0%	95%	13%	0%	88%	4%	0%	96%	1%	0%	99%	10%	0%	91%	18%	0%	84%	3%	0%	97%
			0.2	1%	0%	99%	5%	0%	96%	8%	0%	94%	6%	0%	95%	3%	0%	98%	11%	0%	91%	15%	0%	88%	3%	0%	98%
			0.3	4%	0%	97%	10%	0%	93%	13%	0%	91%	0%	0%	100%	3%	0%	98%	10%	0%	93%	14%	0%	90%	4%	0%	97%
			0.4	0%	0%	100%	7%	0%	96%	8%	0%	95%	5%	0%	97%	0%	0%	100%	8%	0%	95%	7%	0%	96%	3%	0%	98%
		100	0.1	1%	0%	99%	6%	0%	95%	11%	0%	90%	0%	0%	100%	1%	0%	99%	16%	0%	86%	18%	0%	84%	2%	0%	98%
			0.2	1%	0%	99%	5%	0%	96%	5%	0%	96%	3%	0%	98%	1%	0%	99%	11%	0%	91%	14%	0%	89%	3%	0%	98%
			0.3	4%	0%	97%	10%	0%	93%	13%	0%	91%	4%	0%	97%	3%	0%	98%	10%	0%	93%	14%	0%	90%	9%	0%	94%
			0.4	0%	0%	100%	7%	0%	96%	8%	0%	95%	3%	0%	98%	0%	0%	100%	8%	0%	95%	7%	0%	96%	3%	0%	98%
Inf	-	-	0	0%	-	100%	0%	-	100%	4%	-	96%	26%	-	74%	0%	-	100%	0%	-	100%	7%	-	93%	16%	-	84%
	1	10	0.1	0%	0%	100%	1%	0%	99%	3%	0%	97%	8%	0%	93%	1%	0%	99%	4%	0%	96%	8%	0%	93%	4%	0%	96%
			0.2	0%	0%	100%	1%	0%	99%	3%	0%	98%	9%	0%	93%	0%	0%	100%	0%	0%	100%	3%	0%	98%	5%	0%	96%
			0.3	0%	0%	100%	0%	0%	100%	1%	0%	99%	9%	0%	94%	0%	0%	100%	0%	0%	100%	1%	0%	99%	9%	0%	94%
			0.4	0%	23%	91%	0%	0%	100%	0%	0%	100%	8%	0%	95%	0%	28%	89%	0%	0%	100%	0%	0%	100%	8%	0%	95%
		100	0.1	0%	0%	100%	1%	0%	99%	3%	0%	97%	1%	0%	99%	1%	0%	99%	4%	0%	96%	6%	0%	95%	0%	0%	100%
			0.2	0%	0%	100%	1%	0%	99%	8%	0%	94%	10%	0%	92%	0%	0%	100%	0%	0%	100%	5%	0%	96%	6%	0%	95%
			0.3	0%	0%	100%	0%	0%	100%	1%	0%	99%	10%	0%	93%	0%	0%	100%	0%	0%	100%	1%	0%	99%	9%	0%	94%
			0.4	0%	0%	100%	0%	0%	100%	0%	0%	100%	10%	0%	94%	0%	0%	100%	0%	0%	100%	0%	0%	100%	10%	0%	94%
	5	10	0.1	0%	0%	100%	1%	0%	99%	3%	0%	97%	2%	0%	98%	1%	0%	99%	4%	0%	96%	8%	0%	93%	3%	0%	97%
			0.2	0%	0%	100%	1%	0%	99%	6%	0%	95%	1%	0%	99%	0%	0%	100%	0%	0%	100%	4%	0%	97%	1%	0%	99%
			0.3	0%	0%	100%	0%	0%	100%	1%	0%	99%	1%	0%	99%	0%	0%	100%	1%	0%	99%	1%	0%	99%	0%	0%	100%
			0.4	0%	0%	100%	0%	0%	100%	2%	0%	99%	0%	0%	100%	0%	0%	100%	0%	0%	100%	0%	0%	100%	3%	0%	98%
		100	0.1	0%	0%	100%	1%	0%	99%	4%	0%	96%	1%	0%	99%	1%	0%	99%	4%	0%	96%	6%	0%	95%	0%	0%	100%
			0.2	0%	0%	100%	1%	0%	99%	6%	0%	95%	3%	0%	98%	0%	0%	100%	0%	0%	100%	4%	0%	97%	0%	0%	100%
			0.3	0%	0%	100%	0%	0%	100%	1%	0%	99%	1%	0%	99%	0%	0%	100%	1%	0%	99%	1%	0%	99%	3%	0%	98%
			0.4	0%	0%	100%	0%	0%	100%	2%	0%	99%	0%	0%	100%	0%	0%	100%	0%	0%	100%	0%	0%	100%	3%	0%	98%