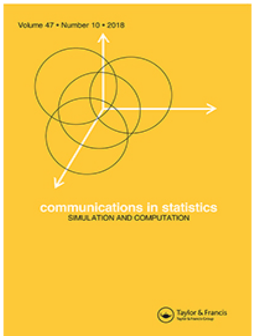

Communications in Statistics - Simulation and Computation

ISSN: 0361-0918 (Print) 1532-4141 (Online) Journal homepage: https://www.tandfonline.com/loi/lssp20

Multidimensional outlier detection and robust estimation using S_n covariance

Sajana O. Kunjunni & Sajesh T. Abraham

COMMUNICATIONS IN STATISTICS - SIMULATION AND COMPUTATION®
https://doi.org/10.1080/03610918.2020.1725820

Multidimensional outlier detection and robust estimation using S_n covariance

Sajana O. Kunjunni and Sajesh T. Abraham

Department of Statistics, St Thomas' College (Autonomous), Thrissur, Kerala, India

ABSTRACT

This article presents a robust method for detecting multiple outliers from multidimensional data using robust Mahalanobis distance. Initial scatter matrix for robust Mahalanobis distance is constructed using a robust estimator of covariance (S_nCov) established from a robust scale estimator S_n and casewise median are chosen to be the location vector. The performance of the proposed method is evaluated using the results of simulated samples. This outlier detection method is compared with some well-known methods available in the current literature. The application of the proposed method in real-life data is also executed in this article.

ARTICLE HISTORY
Received 30 July 2019
Accepted 30 January 2020

KEYWORDS
Location; Mahalanobis distance; Multivariate; Outlier detection; Robust estimation; Scatter

1. Introduction

Application of univariate outlier detection methods in multivariate data may identify unusual observations in individual variables. A multivariate outlier is an inconsistent combination of measurements of more than one variable. An extensive use of univariate method to detect multivariate outliers may not be adequate, since it does not take in to account the relation among variables. To detect multivariate outliers, the distance from the center of mass and covariance structure must be equally considered. Mahalanobis Distance (MD) established by Mahalanobis (1936) is a multivariate measure of distance which consider deviation.

Mean vector and dispersion matrix are the only components of MD. Maximum likelihood estimates of these parameters are sensitive to the presence of outliers in the dataset. Hence, substitution of these estimate in MD is inappropriate for outlier detection. For the purpose of outlier detection a Robust Mahalanobis Distance (RMD) is to be produced by employing robust estimates of the parameters. Various methods have been introduced for robust estimation of location and dispersion of multivariate data in literature. Minimum Volume Ellipsoid (MVE) and Minimum Covariance Determinant (MCD) are introduced by Rousseeuw (1983). MVE is based on the computation of minimum volume ellipsoid containing at least $h = [n/2] + 1$ of the observations of the data, where n is the number of samples. MCD searches for smallest covariance determinant which encompasses at least half of the data points. FAST-MCD was proposed by Rousseeuw and Driessen (1999) as an improved version of MCD. But it still needs

CONTACT Sajana O. Kunjunni sajana.kunjunni@gmail.com Department of Statistics, St Thomas' College (Autonomous), Thrissur, Kerala, India.

substantial time for detection when the number of dimensions are more. Peña and Prieto (2001) and Peña and Prieto (2007) established Kurtosis algorithm. This method consist of maximization and minimization of projection kurtosis coefficients based on some directions generated using stratified random sampling. This procedure also has some limitations in high dimensions and correlated samples. Orthogonalized Gnanadesikan–Kettenring (OGK) estimator introduced by Maronna and Zamar (2002) used a robust covariance matrix defined by Gnanadesikan and Kettenring (1972) which is non-positive semi definite and not an affine-equivariant. This method contains an orthogonalization technique which makes the covariance matrix positive definite and affine-equivariant. Similar type of orthogonalization is adopted by Sajesh and Srinivasan (2012) in the Comedian approach for multivariate outlier detection. In the context of psychological science, Leys et al. (2018) proposed a Robust Variant Mahalanobis Distance (RVMD) method for multivariate outlier detection. The RVMD method constitutes the MCD with two threshold values for detecting multivariate outliers in the data and both act differently for various contamination levels. An extension of median into a multidimensional situation is applied by Sajana and Sajesh (2018) in detecting multivariate data. They proposed a multivariate outlier detection using Spatial Median(SM) and the performances of method is limited to small percentage of contamination in the data. The recommended ratio of n and p is $n > 5p$ for the performance of MCD, to rectify this restriction, Boudt et al. (2019) proposed Minimum Regularized Covariance Determinant(MRCD) method that regularize $h-$ subset based on a predetermined positive definite target matrix.

In this article, a robust distance based approach is proposed using RMD. A multivariate version of the robust S_nCov proposed by Sajana and Sajesh (2019) and variable wise median are used for the computation of RMD. The efficiency of proposed outlier detection method is measured through simulation studies. Robustness properties of this method is tested using theoretical and empirical approaches. Methods which are popularly known for multivariate outlier detection like Comedian, Kurtosis, FAST-MCD, OGK, SM method, RVMD, and MRCD are compared with the proposed method.

2. Multidimensional expansion of S_nCov

Sajana and Sajesh (2019) proposed a bivariate robust covariance estimator S_nCov for covariance parameter based on the robust scale estimator S_n presented by Rousseeuw and Croux (1993). Robust scale estimator S_n of a univariate random variable X is defined as

$$S_n(X) = 1.1926\, \underset{i}{med}\, \underset{j}{med}\, |x_i - x_j|$$

where med stands for low median ($[\frac{n+1}{2}]^{\text{th}}$ order statistic) for outer median and high median ($([\frac{n}{2}] + 1)^{\text{th}}$ order statistic) for inner median and 1.1926 is the consistency factor for normal distributions. And the robust covariance S_nCov of a bivariate random variable (X, Y) is defined as

$$S_nCov(X, Y) = \underset{i}{\text{med}} \left\{ \underset{j \neq i}{\text{med}} \left[(x_i - x_j)(y_i - y_j) \right] \right\} \tag{1}$$

where $1 \leqslant i, j \leqslant n$ and med stands for low median ($[\frac{n+1}{2}]^{\text{th}}$ order statistic). If n is odd inner median will be replaced by $[\frac{n}{2}]^{\text{th}}$ order statistic. It generalize robust scale estimator S_n as it equals S_n^2 when $X = Y$. Moreover S_nCov is symmetric, location invariant and scale equivariant. The multivariate expansion of this bivariate dispersion is defined as follows.

Let $\mathbf{X}$ be a $n \times p$ data matrix with independent observations $\mathbf{x}_i^T = \{x_1, ..., x_n\}$ and columns $\mathbf{X}_j (j = 1, ...p)$ the covariance matrix based on S_nCov is defined as

$$\mathbf{COV}_{S_n}(\mathbf{X}) = (S_nCov(X_i, X_j)), i, j = 1, 2, ..., p \tag{2}$$

Corresponding correlation matrix of $\mathbf{COV}_{S_n}$ denoted by ξ_{S_n} is defined as

$$\xi_{S_n}(\mathbf{X}) = \mathbf{DCOV}_{S_n}(\mathbf{X})\mathbf{D}^T \tag{3}$$

where $\mathbf{D}$ is diagonal matrix with diagonals $1/S_n(x_i), i = 1, ..., p$

Since S_nCov is a robust alternative for classical bivariate covariance, it is possible to state that $\mathbf{COV}_{S_n}$ is a robust alternative to covariance matrix. Basically, this matrix is non-positive semi definite. To solve non-positive semi definiteness, a procedure implemented by Maronna and Zamar (2002) to obtain positive definite and approximately affine equivariant scatter estimates is adopted. To obtain positive definite dispersion matrix and robust estimates, the following steps are applied.

1. Define matrix $\mathbf{E}$ with columns $\mathbf{e}_j$ for $j = 1, ..., p$, where $\mathbf{e}_j$ is the eigenvector corresponding to eigenvalue λ_j of correlation matrix ξ_{S_n}. Hence, ξ_{S_n} can be written as $\xi_{S_n} = \mathbf{E}\boldsymbol{\Lambda}\mathbf{E}^T$, where $\boldsymbol{\Lambda} = \text{diag}(\lambda_1, ..., \lambda_p)$
2. Let $\mathbf{R} = \mathbf{D}^{-1}\mathbf{E}$ and $\mathbf{z}_i = \mathbf{R}^{-1}\mathbf{x}_i$. Then assume that $\mathbf{z}_i^T(i = 1, ..., n)$ and $\mathbf{Z}_j(j = 1, ..., p)$ are rows and columns of orthogonalized matrix $\mathbf{Z}$.
3. The resulting robust estimates for location vector $\mathbf{L}_r(\mathbf{X})$ and scatter matrix $\mathbf{C}_r(\mathbf{X})$ in the following way,

$$\mathbf{m}_r(\mathbf{X}) = \mathbf{R}\boldsymbol{v} \tag{4}$$

$$\mathbf{C}_r(\mathbf{X}) = \mathbf{R}\boldsymbol{\Gamma}\mathbf{R}^T \tag{5}$$

where $\boldsymbol{v} = (\text{med}(\mathbf{Z}_1), ..., \text{med}(\mathbf{Z}_p))^T$ and $\boldsymbol{\Gamma} = \text{diag}(S_n(\mathbf{Z}_1)^2, ..., S_n(\mathbf{Z}_p)^2)$ here, med stands for median and S_n is the robust scale estimate. This process can be iterated to improve estimates by replacing ξ_{S_n} with the form of $\mathbf{C}_r$.

Then squared RMD on the basis of robust estimates is defined as

$$\text{RMD}(\mathbf{x}_i, \mathbf{m}_r, \mathbf{C}_r) = rmd_i = (\mathbf{x}_i - \mathbf{m}_r)^T\mathbf{C}_r^{-1}(\mathbf{x}_i - \mathbf{m}_r), i = 1, ..., n \tag{6}$$

where $\mathbf{m}_r$ and $\mathbf{C}_r$ are defined in (4) and (5), respectively. Decision regarding cutoff value is one of the significant task in outlier detection. To increase the performance of proposed method an adjusted cutoff is considered for different regions of dimensions, i.e.

$$cv = \begin{cases} b\chi^2_{(0.95,p)} & if\ p < 15 \\ \dfrac{\chi^2_{(0.95,p)}\text{med}(rmd_1, ..., rmd_n)}{\chi^2_{(0.5,p)}} & if\ p \geqslant 15 \end{cases} \tag{7}$$

Thus, an $\mathbf{x}_i$ observation is identified as an outlier if $RMD(\mathbf{x}_i, \mathbf{m}_r, \mathbf{C}_r) > cv$. A positive definite and robust estimate can be formulated by a weight function on the basis of

Table 1. RSD comparison.

			$\delta = 5$							
p	λ	γ	S_n method	Comedian	Kurtosis	FAST-MCD	OGK	SM	RVMD	MRCD
5	0.25	0.3	95	95	98	60	81	86	0	0
	0.01	0.2	100	100	99	39	100	100	99	100
		0.3	99	70	99	0	34	94	0	0
	1	0.3	100	100	97	100	83	71	0	0
10	0.25	0.2	100	100	100	41	100	100	57	100
		0.3	100	99	79	0	99	41	0	0
	0.01	0.2	100	100	99	0	100	100	100	100
		0.3	100	83	91	0	38	56	0	0
	1	0.2	100	100	75	100	100	100	100	100
		0.3	97	99	21	99	100	51	0	0
20	0.25	0.2	100	100	90	0	100	75	3	100
		0.3	100	100	3	0	100	0	0	0
	0.01	0.1	100	100	100	0	100	78	42	100
		0.2	100	100	85	0	100	81	0	100
		0.3	88	99	0	0	52	0	0	0
	1	0.1	100	100	49	100	100	52	100	100
		0.2	100	100	1	100	100	58	100	100
		0.3	100	100	0	2	100	0	97	0
			$\delta = 10$							
p	λ	γ	S_n method	Comedian	Kurtosis	FAST-MCD	OGK	SM	RVMD	MRCD
5	0.25	0.3	100	100	100	100	100	100	0	0
	0.01	0.2	100	100	100	100	100	100	100	100
		0.3	100	100	100	0	100	100	0	0
	1	0.3	100	100	100	100	100	100	0	0
10	0.25	0.2	100	100	100	100	100	100	100	100
		0.3	100	100	90	0	100	99	0	0
	0.01	0.2	100	100	100	0	100	100	74	100
		0.3	100	100	92	0	100	98	0	0
	1	0.2	100	100	100	100	100	100	100	100
		0.3	100	100	38	100	100	99	0	0
20	0.25	0.2	100	100	92	0	100	99	1	100
		0.3	100	100	2	0	100	30	0	0
	0.01	0.1	100	100	100	86	100	100	100	100
		0.2	100	100	94	0	100	100	0	100
		0.3	100	100	3	0	100	46	0	0
	1	0.1	100	100	46	100	100	100	100	100
		0.2	100	100	1	100	100	99	100	100
		0.3	100	100	0	2	100	24	97	0

RMD and *cv*. Here, b is a constant which takes value 1 if $p \leq 5$ and 2.5 if $p > 5$. The resulting method of multivariate outlier detection using $\mathbf{COV}_{S_n}$ can be represented as S_n method of outlier detection.

3. Simulation

The effectiveness of proposed S_n method is tested through a series of simulation processes and later experimented with real datasets. Masking and swamping are the two errors occurring while detecting possible outliers. Two aspects of outlier identification are assayed, i.e. rate of successful complete detection of contained outliers which is expressed by rate of successful detection (RSD) and rate of false detection (RFD) indicating rate of detection of inliers as outliers. Sajesh and Srinivasan (2012) presented Comedian method and found out that it is better than Kurtosis, FAST-MCD, and OGK

Table 2. RFD comparison.

			$\delta=5$							
λ	p	γ	S_n method	Comedian	Kurtosis	FAST-MCD	OGK	SM	RVMD	MRCD
0.01	5	0.1	0	4	7	17	15	0	15	15
		0.2	0	5	7	41	10	0	23	5
		0.3	0	4	5	51	10	0	31	25
	10	0.1	0	4	6	22	16	0	60	15
		0.2	0	6	42	44	13	0	68	5
		0.3	0	3	45	54	7	0	70	25
	20	0.1	0	1	10	39	18	0	90	15
		0.2	0	3	40	47	12	0	80	5
		0.3	0	3	40	63	12	5	70	15
0.25	5	0.1	0	3	5	18	16	0	15	15
		0.2	0	2	5	11	11	0	13	5
		0.3	0	2	5	32	7	0	25	25
	10	0.1	0	2	5	24	20	0	61	15
		0.2	0	2	7	36	10	0	67	5
		0.3	0	2	31	40	9	0	62	25
	20	0.1	0	1	9	38	16	0	90	15
		0.2	0	2	13	39	10	0	80	5
		0.3	0	1	39	40	7	0	70	19
1	5	0.1	0	3	6	14	15	0	15	15
		0.2	0	2	6	9	13	0	13	5
		0.3	0	2	6	7	7	0	13	25
	10	0.1	0	2	9	23	16	0	59	15
		0.2	0	2	6	15	13	0	50	5
		0.3	0	2	6	13	7	0	44	25
	20	0.1	0	2	8	28	16	0	90	15
		0.2	0	1	5	18	12	0	80	5
		0.3	0	1	4	27	8	0	70	7

with RSD and RFD. In this article, the proposed method is compared with Comedian and other methods.

To create a data contaminated with outliers, $100(1-\gamma)$ observations are generated from $N(0,\mathbf{I})$ distribution with dimension p for a given level of contamination γ and 100γ observations are replaced by $N(\delta\mathbf{a},\lambda\mathbf{I})$ distribution, where $\mathbf{a}$ represents the vector $(1,\ldots,1)^T$ and $\mathbf{I}$ the identity matrix. The test is undertaken for different choices of dimensions p $(p=5,10,20)$ and contamination level γ $(\gamma=0.1,0.2,0.3)$. For determining the ability to identify minor disparities in data, the experiment is performed for small deviations of δ $(\delta=5,10)$ and λ $(\lambda=0.01,0.25,1)$.

Table 1 shows the RSD values of S_n, Comedian, Kurtosis, FAST-MCD, OGK, SM method, RVMD, and MRCD. Some comparable situation of Comedian method presented by Sajesh and Srinivasan (2012) is chosen to produce this table. The rates from the table shows that S_n method works better than Comedian and Kurtosis apart from two cases $(p=10,\gamma=0.3,\delta=5,\lambda=1$ and $p=20,\gamma=0.3,\delta=5,\lambda=0.01)$. Tables 2 and 3 exhibits maximum RFD values in all combinations, comparison of S_n method with other outlier detection methods for location sifts $\delta=5$ and $\delta=10$, respectively. All the combinations of values explained in simulation part are considered for the maximum RFD estimation. In all the cases, S_n method performed better than rest of the methods with zero RFD.

Table 3. RFD comparison.

						$\delta = 10$				
λ	p	γ	S_n method	Comedian	Kurtosis	FAST-MCD	OGK	SM	RVMD	MRCD
0.01	5	0.1	0	3	9	18	11	0	16	15
		0.2	0	6	7	12	12	0	14	5
		0.3	0	4	5	47	7	0	33	25
	10	0.1	0	2	7	23	19	0	59	15
		0.2	0	5	5	42	11	0	67	5
		0.3	0	5	45	56	9	0	68	2
	20	0.1	0	1	8	38	16	0	90	15
		0.2	0	3	40	47	14	0	80	5
		0.3	0	3	40	61	7	2	70	25
0.25	5	0.1	0	5	7	13	14	0	15	15
		0.2	0	3	6	12	10	0	13	5
		0.3	0	2	5	8	6	0	11	25
	10	0.1	0	2	8	21	17	0	46	15
		0.2	0	2	7	15	13	0	52	5
		0.3	0	4	24	37	7	0	61	6
	20	0.1	0	1	8	28	19	0	90	15
		0.2	0	1	14	39	12	0	80	5
		0.3	0	1	40	41	9	0	70	25
1	5	0.1	0	3	6	19	14	0	15	15
		0.2	0	2	6	10	11	0	13	5
		0.3	0	2	6	7	8	0	11	2
	10	0.1	0	2	6	23	18	0	63	15
		0.2	0	3	6	18	10	0	44	5
		0.3	0	1	7	9	7	0	47	25
	20	0.1	0	2	9	28	14	0	90	15
		0.2	0	1	5	19	11	0	80	5
		0.3	0	1	4	30	7	0	70	25

3.1. Simulation in correlated data

The behavior of S_n method in correlated data is analyzed because of its lack of affine-equivariance. Devlin, Gnanadesikan, and Kettenring (1981) applied a correlation matrix $\mathbf{P}$ of dimension p ($p = 6$) for generating Monte–Carlo data from different distributions. The correlation matrix $\mathbf{P} = ((\rho_{ij}))$ has the form

$$\mathbf{P} = \begin{bmatrix} \mathbf{P}_1 & 0 \\ 0 & \mathbf{P}_2 \end{bmatrix} \text{ where } \mathbf{P}_1 = \begin{bmatrix} 1 & 0.95 & 0.30 \\ 0.95 & 1 & 0.10 \\ 0.30 & 0.10 & 1 \end{bmatrix}, \ \mathbf{P}_2 = \begin{bmatrix} 1 & -0.499 & -0.499 \\ -0.499 & 1 & -0.499 \\ -0.499 & -0.499 & 1 \end{bmatrix}$$

The dimension of correlation matrix is large enough to study the multivariate estimate. Here, the range of correlation is high that helps to investigate the capability of this method to identify the outliers in highly correlated dataset. Asymmetrical datasets of size 100 is generated which includes $100(1-\gamma)$ observation from $N(0, \mathbf{P})$ and 100γ observation from $N(5\mathbf{a}, \mathbf{P})$, where $\mathbf{a} = (1, \ldots, 1)^T$. The RFD values for proposed method in correlated data are presented in Table 4. The results in Table 4 shows that, RFD of S_n method is zero, i.e. the proposed method is free from false detection of inliers as outliers.

3.2. Equivariance

This section discuss about the equivariance property of proposed method by simulated data. Equivariance study is significant to the proposed method as the initial estimate of

Table 4. RFDs of S_n method in correlated samples.

γ	S_n method	Comedian	Kurtosis	FAST-MCD	OGK	SM	RVMD	MRCD
0.1	0	7	10	19	17	2	18	25
0.2	0	4	4	14	11	0	19	25
0.3	0	2	5	7	6	1	25	25

dispersion is not equivariant. Consider a multidimensional random variable $\mathbf{X} = \{\mathbf{x}_1, ...\mathbf{x}_n\}$ with each $\mathbf{x} \in \mathbb{R}^p$. Let $\mathbf{X_A} = \{\mathbf{Ax}_1, ...\mathbf{Ax}_n\}$ where $\mathbf{A}$ is $p \times p$ nonsingular matrix. If the estimates of location and scatter are affine-equivariant, then

$$\boldsymbol{m}_A = \boldsymbol{m}(\boldsymbol{X}_A) = \boldsymbol{Am}(\boldsymbol{X}) \text{ and } \boldsymbol{C}_A = \boldsymbol{C}(\boldsymbol{X}_A) = \boldsymbol{AC}(\boldsymbol{X})\boldsymbol{A}^T$$

The Mahalanobis distance of $\mathbf{X_A}$ from $\mathbf{m_A}$ based on $\mathbf{C_A}$ holds affine-equivariance property if both the location and scatter are affine-equivariant. Maronna and Zamar (2002) generated a random matrices as $\mathbf{A} = \mathbf{TD}$ where $\mathbf{T}$ is a random orthogonal matrix and $\mathbf{D} = \text{diag}(u_1, ..., u_p)$, where u_j's are independent and uniformly distributed in 0, 1.

Simulation of untransformed data has been repeated to investigate the performance of proposed method under transformation. Each data matrix is transformed by multiplying random nonsingular matrix. The proposed method is then applied to the transformed data matrix to detect outlier. The experiment is conducted to different values of p $(p = 5, 10, 20)$ and contamination level γ $(\gamma = 0.1, 0.2, 0.3)$. Table 5 shows simulated results under transformed data and it could be observed that the S_n method is able to detect all the outliers in the dataset, except for some stray situations.

3.3. Breakdown value of S_n method

Maximum proportion of outlier that an estimator can safely tolerate before giving incorrect estimate is termed as breakdown value. Similarly, the breakdown value of an outlier detection method could be defined as the maximum proportion (γ^*) of outliers that the method can precisely identify. Clearly, if $\gamma > \gamma^*$ the method fails to detect majority of the outliers and faultily spot the inliers as outliers or decreases RSDs and increases RFDs. Hence, it is relevant to use RSD and RFD for examining the breakdown value of an outlier detection method.

The experiment to find the breakdown value of S_n method contains generation of symmetrically and asymmetrically distributed contaminations. At first, data of size n is simulated from $N(0, \mathbf{I})$ with dimension p. Then symmetric outliers are inserted by multiplying ith observation with $100i$. For asymmetric contamination ith observation was replaced by $(100i)\mathbf{1}$, where $\mathbf{1} = (1, ...1)$. Different values of p $(p = 10, 30, 50, 80, 100)$ and γ $(\gamma = 10, 20, 30, 40, 48)$ were chosen to determine empirical breakdown value of S_n method. The results for selected sample size $n = 1000$ are presented in Table 6. This empirical experiment shows 100% RSD and 0 RFD.

4. Real dataset

The efficacy of proposed method in real dataset is explained in this section. Bushfire data is considered for studying real data application and it was collected by Campbell

Table 5. RSDs and RFDs of S_n method in transformed data.

			$\delta = 5$		$\delta = 10$	
λ	p	γ	RSD	RFD	RSD	RFD
0.01	5	0.1	100	0	100	0
		0.2	100	0	100	0
		0.3	99	0	100	0
	10	0.1	100	0	100	0
		0.2	100	0	100	0
		0.3	100	0	100	0
	20	0.1	100	0	100	0
		0.2	100	0	100	0
		0.3	100	0	100	0
0.25	5	0.1	100	0	100	0
		0.2	100	0	100	0
		0.3	88	0	100	0
	10	0.1	100	0	100	0
		0.2	100	0	100	0
		0.3	100	0	100	0
	20	0.1	100	0	100	0
		0.2	100	0	100	0
		0.3	84	0	100	0
1	5	0.1	100	0	100	0
		0.2	100	0	100	0
		0.3	100	0	100	0
	10	0.1	100	0	100	0
		0.2	100	0	100	0
		0.3	100	0	100	0
	20	0.1	100	0	100	0
		0.2	100	0	100	0
		0.3	100	0	100	0

Table 6. Empirical results for breakdown value.

		Symmetric		Asymmetric	
p	γ	RSD	RFD	RSD	RFD
10	10	100	0	100	0
	20	100	0	100	0
	30	100	0	100	0
	40	100	0	100	0
	48	100	0	100	0
30	10	100	0	100	0
	20	100	0	100	0
	30	100	0	100	0
	40	100	0	100	0
	48	100	0	100	0
50	10	100	0	100	0
	20	100	0	100	0
	30	100	0	100	0
	40	100	0	100	0
	48	100	0	100	0
80	10	100	0	100	0
	20	100	0	100	0
	30	100	0	100	0
	40	100	0	100	0
	48	100	0	100	0
100	10	100	0	100	0
	20	100	0	100	0
	30	100	0	100	0
	40	100	0	100	0
	48	100	0	100	0

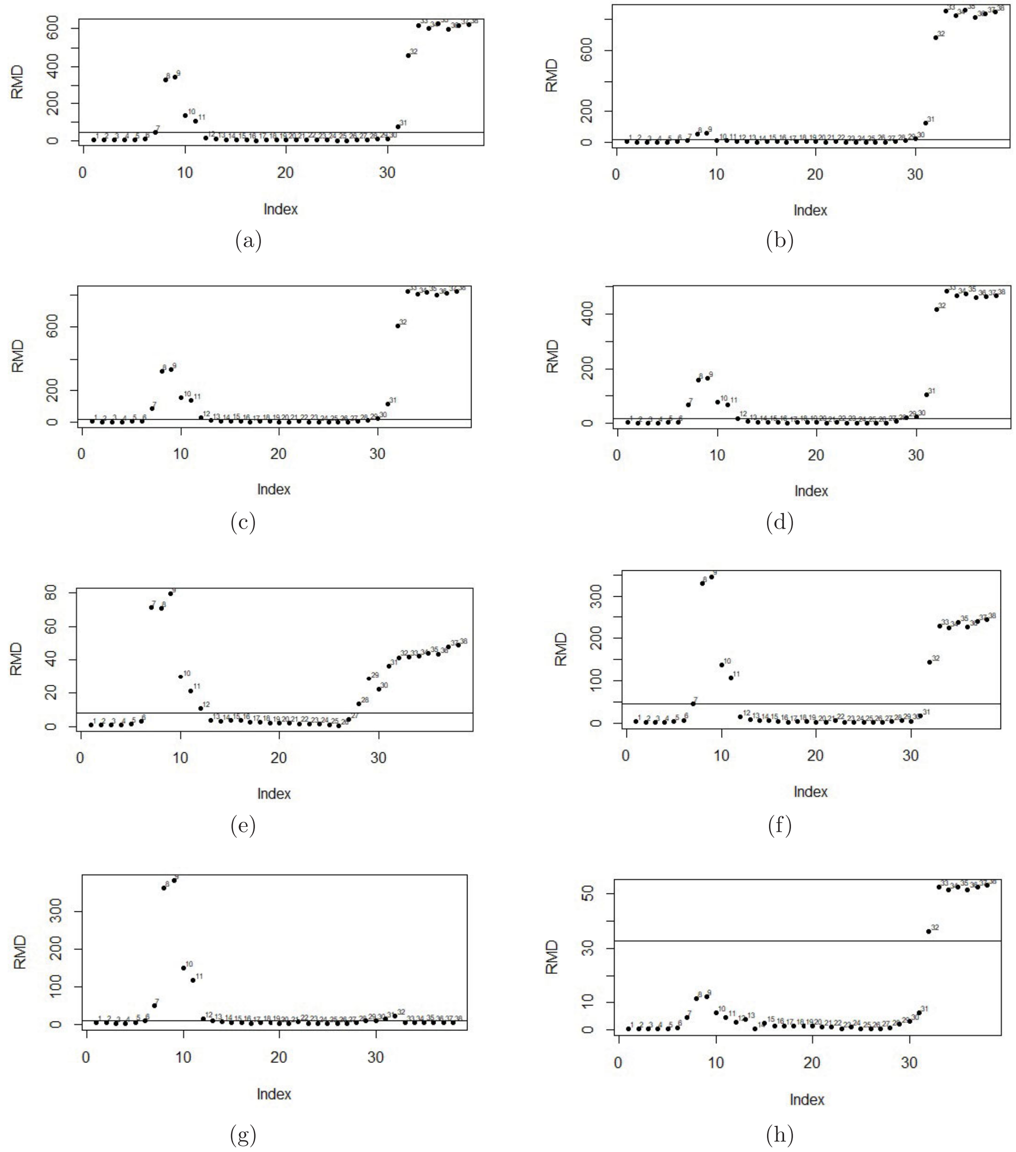

Figure 1. Outlier detection plot for Bushfire data. (a) S_n method, (b) Comedian, (c) Kurtosis, (d) FAST-MCD, (e) OGK, (f) SM, (g) RVMD, (h) MRCD.

(1989) which consist of satellite measurement on five frequency bands each corresponding to 38 pixels. The Bushfire dataset is also openly available at https://vincentarelbundock.github.io/Rdatasets/datasets.html. Maronna and Yohai (1995) analyzed the dataset and concluded that observations 7–11 are outlying and they can be easily identified by various robust methods. But, the observations 32–38 are masked by the first group of outliers and Stahel–Donoho projection estimator implemented by Maronna and Yohai (1995) does not get effected by this error. Outliers in bushfire data are identified using S_n method, Comedian, Kurtosis, FAST-MCD, OGK, SM method, RVMD, and MRCD and the diagnostic plot is presented in Figure 1. From the figure, it can be see that the

S_n method is able to detect observations 7–11 and 31–38 as possible outliers and it has relatively better result with less swamping error. In the case of other methods, Comedian method identified 8–9 and 30–38 as outliers. Kurtosis method observed that sample 30 and FAST-MCD method indicated that observation 29 are additional outliers. According to OGK method, it is found that sample 28 is also a deviated observation. In addition to S_n method, SM method is able to detect the possible outliers. But RVMD and MRCD are only capable of detecting few outliers presented in the dataset.

5. Conclusion

Outlier detection is a significant part of data preprocessing since it could influence the inferences of analysis. An alternative method to detect multivariate outliers on the basis of repeated median covariance matrix is presented through this article. The effectiveness of the method is discussed and compared with well-known methods OGK, MCD, Kurtosis, Comedian, SM method, RVMD, and MRCD.

The simulation study is executed and explained in different possible choices of parameters. Simulation results of RSD and RFD shows that the proposed method performed better than Kurtosis, FAST-MCD, SM method, RVMD, and MRCD. In the case of comparison with comedian and OGK, the proposed method appeared better in RSD measurements except some rare cases. But it outperformed in RFD values. To understand the capability of proposed method in collinear data, highly correlated data is generated in specific dimension. The RFDs presented here reflects low swamping error of S_n method in correlated data. Affine-equvariance property of the method is also tested beacause of lack of equivariance of $\mathbf{COV}_{S_n}$. RSDS and RFDS seems similar in both affinely transformed and untransformed data. Symmetrically and asymmetrically contaminated datasets are generated to estimate the breakdown value of proposed method. The simulation result of breakdown value shows that, the method is robust even under highly contaminated situations. In the real datasets SM method performed equivalent to S_n method, but performance of SM method in simulated datasets are less uncompromisable. The application of proposed method in real dataset reflects its effectiveness of simulated result by detecting possible outliers with low swamping. Hence, S_n method can apply in multivariate datasets for cleansing multiple outliers with minimum errors.

Acknowledgment

The authors are thankful to the reviewer for their valuable comments and efforts toward improving our manuscript.

References

Boudt, K., P. J. Rousseeuw, S. Vanduffel, and T. Verdonck. 2019. The minimum regularized covariance determinant estimator. *Statistics and Computing* 30 (1):113–28. doi:10.1007/s11222-019-09869-x.

Campbell, N. A. 1989. *Robust Bushfire mapping using NOAA AVHRR data: Technical Report.* North Ryde, Australia: CSIRO.

Devlin, S. J., R. Gnanadesikan, and J. R. Kettenring. 1981. Robust estimation of dispersion matrices and principal components. *Journal of the American Statistical Association* 76 (374):354–62. doi:10.1080/01621459.1981.10477654.

Gnanadesikan, R., and J. R. Kettenring. 1972. Robust estimates, residuals and outlier detection with multiresponse data. *Biometrics* 28 (1):81–124. doi:10.2307/2528963.

Leys, C., O. Klein, Y. Dominicy, and C. Ley. 2018. Detecting multivariate outliers: Use a robust variant of the Mahalanobis distance. *Journal of Experimental Social Psychology* 74:150–6. doi: 10.1016/j.jesp.2017.09.011.

Mahalanobis, P. C. 1936. On the generalized distance in statistics. *Proceedings of the National Institute of Sciences* 2: 49–55.

Maronna, R. A., and V. J. Yohai. 1995. The behaviour of the Stahel-Donoho robust multivariate estimator. *Journal of the American Statistical Association* 90 (429):330–41. doi:10.1080/01621459.1995.10476517.

Maronna, R. A., and R. H. Zamar. 2002. Robust estimates of location and dispersion for high-dimensional datasets. *Technometrics* 44 (4):307–17. doi:10.1198/004017002188618509.

Peña, D., and F. J. Prieto. 2001. Multivariate outlier detection and robust covariance matrix estimation. *Technometrics* 43 (3):286–310. doi:10.1198/004017001316975899.

Peña, D., and F. J. Prieto. 2007. Combining random and specific directions for outlier detection and robust estimation in high-dimensional multivariate data. *Journal of Computational and Graphical Statistics* 16 (1):228–54. doi:10.1198/106186007X181236.

Rousseeuw, P. J. 1983. Multivariate estimation with high breakdown point. *Fourth Pannonian Symposium on Mathematical Statistics and Probability*, Bad Tatzmannsdorf, Austria.

Rousseeuw, P. J., and C. Croux. 1993. Alternatives to the median absolute deviation. *Journal of the American Statistical Association* 88 (424):1273–83. doi:10.1080/01621459.1993.10476408.

Rousseeuw, P. J., and K. V. Driessen. 1999. A fast algorithm for the minimum covariance determinant estimator. *Technometrics* 41 (3):212–23. doi:10.1080/00401706.1999.10485670.

Sajana, O. K., and T. A. Sajesh. 2018. Detection of multidimensional outlier using multivariate spatial median. *Journal of Computer and Mathematical Sciences* 9 (12):1875–81. doi:10.29055/jcms/934.

Sajana, O. K., and T. A. Sajesh. 2019. *Sn* covariance. *Communications in Statistics - Theory and Methods*. doi:10.1080/03610926.2019.1628275.

Sajesh, T. A., and M. Srinivasan. 2012. Outlier detection for high dimensional data using the comedian approach. *Journal of Statistical Computation and Simulation* 82 (5):745–57. doi:10.1080/00949655.2011.552504.