Chilean Journal of Statistics
Volume 15, Number 2, December 2024, Pages 110–125
DOI: 10.32372/chjs.15-02-01

STATISTICAL INFERENCE
RESEARCH ARTICLE

On a Stahel-Donoho estimator with skewness-based random projection directions

SANTIAGO ORTIZ[1,*] and OMAR BECERRA[2]

[1]Faculty of Engineering, Universidad de San Buenaventura, Cali, Colombia
[2]School of Applied Sciences and Engineering, Universidad EAFIT, Medellín, Colombia

(Received: 19 April 2024 · Accepted in final form: 04 November 2024)

Abstract

This work introduces a novel version of the Stahel-Donoho multivariate outlier detection procedure, which considers $5p + 1$ specific random directions, where p is the dimensionality of the data, that is, the number of variables in the dataset. These directions are derived by maximizing the squared third sample moment of the projected observations, which then serves as a seed to obtain $5p$ additional directions via a stratified sampling. Compared with the standard Stahel-Donoho estimator and other outlier detection methods, this new version exhibits competitive performance across various high-dimensional datasets and contamination scenarios. By leveraging maximum skewness projection within the Stahel-Donoho framework, the proposed estimator maintains stable results in high dimensions, showing its advantage in efficiently handling complex data structures.

Keywords: Outlier detection · Projection pursuit · Robust statistics · Stratified sampling · Third sample-moment.

Mathematics Subject Classification: Primary 62F35 · Secondary 62H30.

1. INTRODUCTION

The standard Stahel-Donoho (SD) estimator, proposed by Stahel (1981) and Donoho (1982), marks a milestone in robust statistics, introducing a projection pursuit method for multivariate outlier detection and robust estimation (de Paula Alves and Furtado Ferreira, 2020). This estimator computes a fixed number of directions at random, assessing the outlyingness measure in each direction for the projected data. Characterized by its affine equivariance, high breakdown point (Maronna and Yohai, 1995), and asymptotic relative efficiency derived from its influence function (Gervini, 2002), the SD algorithm has laid the groundwork for numerous applications and subsequent developments in multivariate data analysis. Despite its advantages, the widespread use of projection-based methods has been hindered by computational challenges. Several authors have highlighted the importance of efficient optimization routines (Sun, 2006). For instance, semi-robust principal components designed for high-dimensional data (Filzmoser et al., 2008) are an important development. Another key advancement is the use of directions that maximize k–nearest neighbor distances, where k represents the number of neighbors considered. Typically, k is chosen based on the number of variables in the dataset, p say (Kandanaarachchi and Hyndman, 2021).

*Corresponding author. Email: sortiza@usbcali.edu.co (S. Ortiz)

The above-mentioned studies have improved projection pursuit techniques for outlier detection by introducing new approaches related to optimization strategies and criteria. These approaches also include methods that optimize the third and fourth statistical moments of projected data, which further enhance the ability to identify outliers in multivariate datasets.

Well-known kurtosis optimization techniques have been proposed (Peña and Prieto, 2001, 2007; Peña et al., 2010), alongside methods based on projections that maximize the sample skewness coefficient (Loperfido, 2018). Additionally, Van Aelst et al. (2012) proposed a modification of the SD estimator using random directions adapted to a Huberized outlyingness measure. This modification offers a more efficient and effective way to calculate SD directions. Subsequently, Van Aelst (2016) introduced two adaptations: the first one adjusts the calculation of outlyingness, and the second one assigns separate weights to each component of an observation. These adaptations perform well in scenarios with component-wise contamination.

The standard SD algorithm, as proposed by Stahel (1981), relies on a subsampling procedure for a multivariate sample $\boldsymbol{X} \in \mathrm{R}^{n\times p}$, where, as usual, n is the sample size and p the number of variables. Randomly, p points from the sample are selected, and an orthogonal direction to the hyperplane defined by these p points is computed. This procedure is repeated a fixed number of times to generate a subset of projection directions. The number of subsamples required is independent of n, making its computational complexity linear. However, as p increases, the computational complexity grows exponentially (Juan and Prieto, 1995).

To address this computational complexity, Peña and Prieto (2007) proposed a fast algorithm for outlier detection that combines projections on a set of $2p$ deterministic directions, which are extremes of kurtosis, with a set of random directions when the kurtosis direction is not informative. These directions are then used as initial projections for calculating outlyingness. This algorithm not only computes the SD estimator but also provides a robust starting point for iterative estimation.

In high-dimensional problems, Peña and Prieto (2007) investigated kurtosis directions and demonstrated promising results for detecting concentrated outliers. However, the method is less effective when the contamination proportion (α) is approximately 0.3 or when the distributions of non-outlying and outlying data share the same covariance structure. To address this, they proposed the random and specific projections of order one —RASP(1)— algorithm, an improved version of the kurtosis projection algorithm proposed by Peña and Prieto (2001). Unlike the $2p$ directions used previously, RASP(1) focuses on two kurtosis directions complemented with random but specific Stahel-Donoho-type directions (RS-SD). This approach demonstrates robust performance as both p and α increase, particularly in scenarios with concentrated contamination.

The RASP(1) algorithm highlights the importance of further research to evaluate its performance under diverse scenarios and compare it with other methods. Following the logic of RASP(1), RS-SD directions are calculated using a stratified sampling procedure, which enhances the probability of obtaining helpful directions by selecting two random observations, either from non-outlying or outlying points, with probability $\alpha^p + (1-\alpha)^p$.

Loperfido (2013) demonstrated that the direction maximizing the third sample cumulant corresponds to the Fisher linear discriminant function. More recently, Ortiz (2019) proposed an outlier detection procedure that maximizes the absolute third sample moment. Based on this projection direction, the main objective of the present work is to introduce a novel and computationally efficient method for the SD estimator. Instead of generating random directions, this method constructs a set of $5p$ specific directions, including a seed direction that maximizes the squared third sample moment. These directions are derived through stratified sampling of projection directions, enhancing performance in high-dimensional outlier detection while improving computational efficiency.

The article is organized as follows. Section 2 describes the proposed estimator and its construction, based on specific random directions obtained from a skewness-based projection direction seed. In Section 3, we present a simulation study on multivariate outlier detection under various contamination scenarios. In Section 4, the proposed method is applied to real-world datasets. Lastly, Section 5 provides concluding remarks and future research directions.

2. The proposed method

This section introduces the proposed method, which aims to enhance outlier detection through a novel analysis of projections and stratified sampling. The method is anchored by a seed direction determined by the maximum sample squared skewness. Our method offers a robust mechanism for identifying outliers by evaluating univariate outlyingness across a carefully chosen set of projection directions. The mechanism starts by determining the direction that maximizes the sample squared skewness, generating specific basis vectors. These vectors highlight outliers by projecting data onto dimensions where anomalies are most pronounced. In the last step, a weighted outlyingness measure is computed for each projection, forming the basis for outlier identification.

2.1 *Stahel-Donoho estimator*

The SD estimator is a robust method for estimating multivariate location and scatter. It is defined as a weighted mean and covariance matrix, where the weights are based on a measure of outlyingness computed through a penalization function. The outlyingness measure considers the maximum of the one-dimensional projection in which the observation is most outlying, across all possible projection directions. These weights are then used to down-weight the most extreme observations.

Consider the multivariate sample $\boldsymbol{X} = (X_1, \ldots, X_n)$, its observed values $\boldsymbol{x} = (x_1, \ldots, x_n)$, and the set of all p-dimensional unitary projection directions $S_d = \{\boldsymbol{d} \in \mathrm{R}^p \colon \boldsymbol{d}^\top \boldsymbol{d} = 1\}$. The SD outlyingness r of a data point x_i onto a direction $\boldsymbol{d} \in S_d$ is typically computed as the distance between the projected observations $\boldsymbol{d}^\top x_i$ and a univariate location estimate μ, scaled by a univariate scatter estimate σ. Therefore, for any x_i, $r(x_i, \boldsymbol{X}) = r_i$ is defined as

$$r(x_i, \boldsymbol{X}) = \sup_{\boldsymbol{d} \in S_d} \left\{ \frac{|\boldsymbol{d}^\top x_i - \mu(\boldsymbol{d}^\top \boldsymbol{X})|}{\sigma(\boldsymbol{d}^\top \boldsymbol{X})} \right\}, \quad i \in \{1, \ldots, n\}. \tag{2.1}$$

To ensure robustness, μ and σ are often the sample median and median absolute deviation, respectively (Stahel, 1981; Donoho, 1982). Large outlyingness values indicate points that are particularly atypical relative to the rest of the data, while values close to zero indicate that the point is near to the median and, hence, not atypical.

The robust SD estimator for multivariate location and scatter is defined as

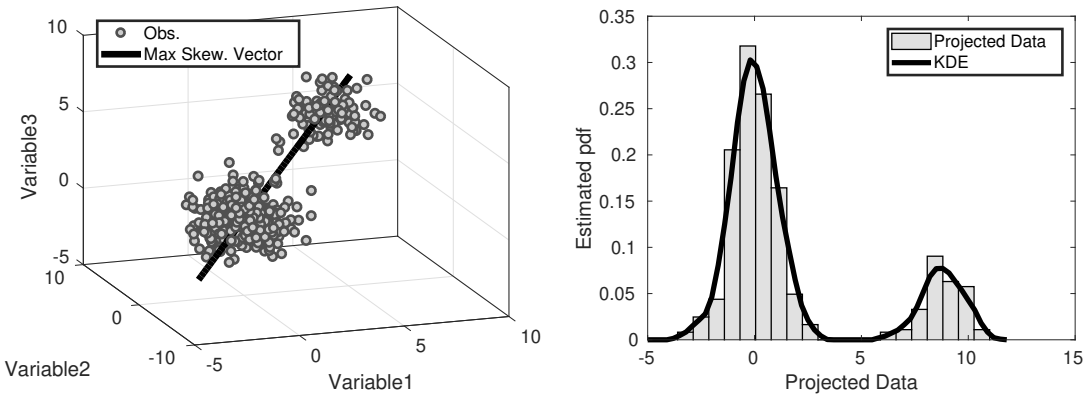
$$\widehat{\boldsymbol{\mu}}_{\mathrm{SD}} = \frac{\sum_{i=1}^n w_i X_i}{\sum_{i=1}^n w_i}, \quad \widehat{\boldsymbol{S}}_{\mathrm{SD}} = \frac{\sum_{i=1}^n w_i \, (X_i - \widehat{\boldsymbol{\mu}}_{\mathrm{SD}})^\top (X_i - \widehat{\boldsymbol{\mu}}_{\mathrm{SD}})}{\sum_{i=1}^n w_i}, \quad i \in \{1, \ldots, n\}, \tag{2.2}$$

where $w_i(r_i) \colon (0, +\infty) \to (0, +\infty)$ penalizes observations with large outlyingness. Various approaches for selecting w_i are described in the literature. A well-known example is the Huber family, which improves outlier detection (Maronna et al., 2006; de Menezes et al., 2021).

2.2 *Sample skewness as a projection pursuit index*

The skewness coefficient, a measure of asymmetry in a probability distribution, has emerged as a relevant statistical index for projection pursuit in the context of multivariate outlier detection. In statistical analysis, skewness provides insights into the distributional characteristics of data and serves as a powerful tool for identifying outliers that deviate from the core distribution. Recent methods have leveraged sample skewness as a mechanism to isolate such anomalies. Loperfido (2013) proposed an estimator based on the singular value decomposition of the third standardized cumulant. This estimator demonstrates that, in the presence of a mixture of symmetric distributions, the direction maximizing the third cumulant aligns with the Fisher linear discriminant function. This echoes the earlier proposal by Peña and Prieto (2000). More recently, Loperfido (2018) introduced a procedure for computing directions that accentuate sample skewness. This approach is particularly advantageous for exploratory data analysis and the preliminary detection of outliers, facilitating a deeper understanding of the underlying structure of the data. Ortiz (2019) developed a multivariate outlier detection method that maximizes the absolute third sample moment of projected data. This technique is rooted in an eigenvector-based matrix iteration strategy similar to that proposed by Peña and Prieto (2001).

Figure 1 provides a graphical representation of the utility of extreme skewness as a statistical index for projection pursuit in outlier detection. By identifying directions that enhance skewness, it becomes feasible to detect observations that diverge from the majority, thereby flagging potential outliers.



(a) 3D scatterplot of the contaminated sample and the direction maximizing extreme skewness.

(b) Kernel density estimate (KDE) of the projected data.

Figure 1. Outlier detection for a 3D simulated dataset. Panel (a) shows a 3D scatterplot of the contaminated sample and the projection direction maximizing the absolute third sample moment. Panel (b) presents the KDE and histogram of the projected data. Adapted from Ortiz (2019).

Consider a p-dimensional contaminated sample $\boldsymbol{Z} = (Z_1, \ldots, Z_n) \in \mathrm{R}^{p\times n}$ and its observed values $\boldsymbol{z} = (z_1, \ldots, z_n) \in \mathrm{R}^{p\times n}$, drawn from the mixture model $\boldsymbol{Z} \sim (1-\alpha)\mathcal{F} + \alpha\mathcal{G}$, where α denotes the contamination proportion, $\mathcal{F}$ is the distribution of non-outlier points, and $\mathcal{G}$ the distribution of outliers. Assuming $\boldsymbol{Z}$ is centrally scaled, let m_3 denote the univariate third sample moment coefficient, computed for the projected data.

We want to find $\boldsymbol{r} \in \mathrm{R}^{p\times 1}$, an unknown unitary p-dimensional vector, by solving

$$\boldsymbol{d}_1 = \underset{\boldsymbol{r}}{\operatorname{argmax}}\{m_3(\boldsymbol{r}^\top \boldsymbol{Z})^2\}, \quad \boldsymbol{r}^\top \boldsymbol{r} = 1, \tag{2.3}$$

which is equivalent to

$$\boldsymbol{d}_1 = \underset{\boldsymbol{r}}{\operatorname{argmax}} \left\{ \left(\sum_{i=1}^{n} \frac{(\boldsymbol{r}^\top z_i)^3}{n} \right)^2 \right\}, \quad \boldsymbol{r}^\top \boldsymbol{r} = 1.$$

To increase sensitivity to asymmetrical contamination, we maximize m_3^2 (instead of m_3 or $|m_3|$). This allows for the effective management of extreme skewness in any direction. Moreover, maximizing m_3^2 simplifies the optimization process, facilitating the identification of a direction that accentuates the skewness of the projected data.

The computation of $\boldsymbol{d}_1$ can be performed numerically using the Newton-Raphson method, for example. However, other optimization methods may also be employed; see, for instance, Loperfido (2015a,b, 2024).

2.3 *Projection direction as a seed for computing random vectors*

According to Peña and Prieto (2007), a direction generated by two points —one from the clean sample and the other from the contamination— can serve as an initial random direction; see Figure 1 for as example.

An important issue is determining how many random directions are required or sufficient to improve outlier detection. For the RS-SD procedure, $10p$ random directions are suggested for $T = n/2p$, where T represents the number of partitions of the projection.

Several authors have discussed the number of directions needed in other statistical methods. Hubert and Van der Veeken (2008) argued that $250p$ random directions are computationally effective for their skewness-adjusted outlyingness proposal. Later, Cuesta-Albertos and Nieto-Reyes (2008) provided empirical evidence suggesting that approximately $5p$ random directions are sufficient to compute a reliable approximation of the random Tukey depth.

In line with this, we propose a modified version of the stratified sampling RS-SD method, introduced by Peña and Prieto (2007), which incorporates the computation of $5p$ random directions. This new method is referred to as the skewness-based random directions (SRD) method. The choice of $5p$ is further supported by empirical simulations we conducted, which demonstrated effective results in outlier detection. The SRD procedure is outlined in Algorithm 2.3.

Algorithm 2.3: SRD —Generation of the set S_p.

1: Compute the initial direction $\boldsymbol{d}_1$ as the solution to the problem stated in Equation (2.3).
2: Define the set $S_p = \{\boldsymbol{d}_1\}$.
3: Project $\boldsymbol{Z}$ onto the direction $\boldsymbol{d}_1$.
4: **for** $l = 1$ to $5p$ **do**
5: Choose $z_a \in \{z_i \in \boldsymbol{Z}\colon \boldsymbol{d}_1^\top z_i \leq \boldsymbol{d}_1^\top \boldsymbol{Z}_{[n/4]}\}$ randomly, where $\boldsymbol{d}_1^\top \boldsymbol{Z}_{[n/4]}$ is an order statistic of $\boldsymbol{d}_1^\top \boldsymbol{Z}$.
6: Select $z_b \in \{z_i \in \boldsymbol{Z}\colon \boldsymbol{d}_1^\top z_i \geq \boldsymbol{d}_1^\top \boldsymbol{Z}_{[3n/4]}\}$ randomly, where $\boldsymbol{d}_1^\top \boldsymbol{Z}_{[3n/4]}$ is an order statistic of $\boldsymbol{d}_1^\top \boldsymbol{Z}$.
7: Calculate the unit direction $\widehat{\boldsymbol{d}}_{(l,z_a,z_b)} = (z_a - z_b)/||z_a - z_b||_2$ defined by points z_a and z_b.
8: Store the direction $\widehat{\boldsymbol{d}}_{(l,z_a,z_b)}$ in the set S_p.
9: **end for**
10: Return the set $S_p = \{\boldsymbol{d}_1, \widehat{\boldsymbol{d}}_{(1,z_a,z_b)}, \ldots, \widehat{\boldsymbol{d}}_{(5p,z_a,z_b)}\}$.

2.4 *Modified SD based on skewness and random specific directions*

Once the SRD procedure is executed and the set S_p is defined, these directions are subsequently used to compute the SD outlyingness measure, incorporating the initial projection seed $\boldsymbol{d}_1$. The SRD algorithm provides the necessary projection directions, which are then utilized by the SD estimator with skewness-based specific directions (SDE-SSD). The SDE-SSD method allow us to estimate location and scatter parameters robustly.

Consider Equation (2.1) and the estimator of location and scale stated in Equation (2.2), where $w_i = w(r_i)$: $\mathrm{R}^+ \to \mathrm{R}^+$ is a weight function that penalizes or down-weights observations with large outlyingness.

Several approaches to selecting w exist in the literature, such as Hard-rejection and Tukey biweight. Following Maronna and Yohai (1995) and Van Aelst et al. (2012), we implement the Huber-type weight function, defined as

$$w(r_i) = 1_{(r_i \le \eta)} + (\eta / r_i)^2 1_{(r_i > \eta)}, \quad i \in \{1, \ldots, n\},$$

where $\eta = \min\{(\chi^2_{(0.5,p)})^{1/2}, 4\}$ and 1_A is the indicator function for the set A. The value of η represents a trade-off between robustness and efficiency. Maronna and Zamar (2002) noted that η should be small to achieve robust estimates in higher dimensions.

For μ and σ stated in Equation (2.1), we use the sample median and Q_n statistic (Rousseeuw and Croux, 1993). The Q_n statistic for a univariate random sample $Y = (Y_1, \ldots, Y_n)$ is defined as

$$Q_n(Y) = b \left\{ \mathrm{abs}(Y_i - Y_j) \colon i < j \right\}_\ell ,$$

where b is a scalar depending on the distribution $\mathcal{H}$ of the random sample Y, and $\{\cdot\}_\ell$ denotes the ℓ-th order statistic.

The median and Q_n statistics were chosen for their robustness against outlier observations. Thus, the modified SDE-SSD outlyingness version $r^*(z_i, \boldsymbol{Z}) = r_i^*$ is expressed as

$$r^*(z_i, \boldsymbol{Z}) = \max_{\boldsymbol{d} \in S_p} \left\{ \frac{|\boldsymbol{d}^\top z_i - \mathrm{med}(\boldsymbol{d}^\top \boldsymbol{Z})|}{Q_n(\boldsymbol{d}^\top \boldsymbol{Z})} \right\}, \quad i \in \{1, \ldots, n\}.$$

Using these outlyingness measures, the SDE-SSD estimators $\widehat{\boldsymbol{\mu}}_{\text{SDE-SSD}}$ and $\widehat{\boldsymbol{S}}_{\text{SDE-SSD}}$ are computed as in Equation (2.2), substituting r_i^* for r_i.

Multivariate outlier identification is based on the robust squared Mahalanobis distance given by

$$\mathrm{MD}^2_{\text{SDE-SSD}}(z_i) = (z_i - \widehat{\boldsymbol{\mu}}_{\text{SDE-SSD}})^\top \widehat{\boldsymbol{S}}^{-1}_{\text{SDE-SSD}} (z_i - \widehat{\boldsymbol{\mu}}_{\text{SDE-SSD}}), \quad i \in \{1, \ldots, n\},$$

and the β quantile (0.975) of a chi-squared distribution with p degrees of freedom. Thus, if $\mathrm{MD}^2_{\text{SDE-SSD}}(z_i) \ge \chi^2_{(\beta,p)}$, then z_i is labeled as a multivariate outlier.

3. Numerical experiments

This section evaluates the proposed SDE-SSD method through numerical experiments designed to assess its effectiveness in outlier detection. By simulating data under various conditions, we rigorously evaluate the performance of SDE-SSD in identifying outliers within multivariate datasets. These experiments also demonstrate the practical applicability of our approach and compare its performance against established methods in the literature.

3.1 *Multivariate outlier detection*

To rigorously assess the robustness of our proposed method, we conducted a comprehensive suite of simulation experiments. We consider a p-dimensional random sample $\boldsymbol{X}$ drawn from a fully dependent contamination model (Alqallaf et al., 2009). This model is defined as a mixture of normal distributions, $\boldsymbol{X} \sim (1-\alpha)\mathrm{N}_p(\boldsymbol{0}, \boldsymbol{I}) + \alpha \mathrm{N}_p(\delta \boldsymbol{e}, \boldsymbol{I})$, where $\delta \neq 0$ is a scalar controlling the distance between the centers of the contaminated and uncontaminated samples, and $\alpha \in (0, 0.5)$ denotes the contamination proportion.

Note that $\boldsymbol{0}$ and $\boldsymbol{e}$ represent the p-dimensional vectors of zeros and ones, respectively, and $\boldsymbol{I}$ denotes the p-dimensional identity matrix. The experimental framework was set with $\alpha \in \{0.1, 0.2, 0.3, 0.4\}$, dimensions $p \in \{5, 10, 30, 50, 100\}$, and displacement magnitudes $\delta \in \{5, 7, 9\}$. For each simulation setting, $n = 10p$ observations were generated, and $m = 100$ random repetitions were conducted.

These experiments compare the performance of the proposed SDE-SSD method with other well-known outlier detection techniques using two metrics: the true positive rate (c) and the false positive rate (f). The six methods selected for comparison are: (i) the SD estimator (Stahel, 1981); (ii) the Huberized SD estimator —SDEH— (Van Aelst et al., 2012); (iii) the skewness-adjusted outlyingness method —SDE-AdjOut— (Hubert and Van der Veeken, 2008); (iv) the minimum covariance determinant —MCD— (Rousseeuw and Van Driessen, 1999); (v) the orthogonalized Gnanadesikan-Kettenring estimator —OGK— (Maronna and Zamar, 2002); and (vi) M-estimators ——Mest (Rocke and Woodruff, 1996).

The M-estimator uses the translated biweight function (t-biweight) with a high breakdown point initial estimate, as defined by Rocke and Woodruff (1996). All experiments were performed in the `R` programming language using the `rrcov` package (Todorov, 2021) for the competing methods. For SDE-AdjOut, we use the `mrfDepth` package (Segaert et al., 2020).

3.2 *Simulation results*

Table 1 presents results of our simulations. The SD and SDE-SSD methods exhibit similar performance in classifying clean data. However, as contamination increases, the SDEH method performance reduces its efficacy. The SDE-AdjOut algorithm shows high performance at low contamination, but its detection performance decreases after 0.2 contamination. Also, this method has a higher computational cost compared to the SD method.

For real data and outlier classification, both the SD and SDEH algorithms experience increased runtime as the multivariate dimension p increases. Higher-dimensional experiments were not conducted for these methods due to numerical instability and computational demands, as noted in their documentation. In contrast, the SDE-SSD algorithm outperforms these methods in terms of both efficiency and classification performance. This makes it a preferable choice for scenarios requiring robust outlier detection and classification.

The results in Table 2 show that SDE-SSD outperforms Mest, OGK, and MCD in classifying outlier data in dimensions 5, 10, and 30. Although the computational time for SDE-SSD is longer than for other methods, it remains relatively short, averaging around one second. The Mest and MCD estimators experience sharp performance declines when faced with contamination levels of 0.2, reducing their classification efficacy. The OGK method begins losing performance at contamination levels exceeding 0.3. Despite longer computation times, SDE-SSD maintains high classification performance, showcasing its robustness in diverse data environments.

For dimensions 50 and 100 presented in Table 3, SDE-SSD does not increase computation times. In comparison, Mest and MCD have longer computation times, while OGK maintains the lowest execution times. Regarding classification performance, Mest and MCD show reduced efficacy in differentiating between real data and outliers. Although OGK performs well in certain scenarios, it suffers performance drops in others. SDE-SSD demonstrates stable performance, maintaining a high level of accuracy compared to other methods.

Table 1. Comparison of outlier detection performances in terms of c and f metrics for SDE-SSD, SD, SDEH, and SDE-AdjOut methods.

p	α	δ	SDE-SSD		SD		SDEH		SDE-AdjOut	
			c	f	c	f	c	f	c	f
5	0.1	5	1.00	0.02	1.00	0.09	1.00	0.05	1.00	0.01
		7	1.00	0.02	1.00	0.09	1.00	0.05	1.00	0.01
		9	1.00	0.02	1.00	0.09	1.00	0.05	1.00	0.01
	0.2	5	1.00	0.02	1.00	0.06	0.97	0.07	1.00	0.02
		7	1.00	0.02	1.00	0.06	0.97	0.06	1.00	0.02
		9	1.00	0.02	1.00	0.06	0.98	0.06	1.00	0.02
	0.3	5	1.00	0.03	1.00	0.02	0.10	0.05	0.00	0.01
		7	1.00	0.02	1.00	0.02	0.16	0.05	0.00	0.01
		9	1.00	0.02	1.00	0.02	0.17	0.05	0.00	0.02
	0.4	5	0.99	0.02	1.00	0.00	0.02	0.06	0.00	0.02
		7	1.00	0.02	1.00	0.09	0.01	0.07	0.00	0.02
		9	1.00	0.02	1.00	0.09	0.00	0.08	0.00	0.02
10	0.1	5	1.00	0.01	1.00	0.08	1.00	0.03	1.00	0.02
		7	1.00	0.01	1.00	0.08	1.00	0.03	1.00	0.02
		9	1.00	0.01	1.00	0.09	1.00	0.03	1.00	0.02
	0.2	5	1.00	0.02	1.00	0.05	0.95	0.05	1.00	0.02
		7	1.00	0.01	1.00	0.05	0.91	0.04	1.00	0.02
		9	1.00	0.01	1.00	0.05	0.92	0.04	1.00	0.02
	0.3	5	1.00	0.02	1.00	0.05	0.02	0.06	0.00	0.03
		7	1.00	0.02	1.00	0.02	0.02	0.05	0.00	0.03
		9	1.00	0.01	1.00	0.02	0.03	0.05	0.00	0.03
	0.4	5	0.96	0.02	1.00	0.01	0.00	0.09	0.00	0.04
		7	1.00	0.01	1.00	0.09	0.00	0.09	0.00	0.04
		9	1.00	0.02	1.00	0.01	0.00	0.09	0.00	0.04

Table 2. Comparison of outlier detection performances in terms of c and f metrics for SDE-SSD, Mest, MCD, and OGK methods.

p	α	δ	SDE-SSD		Mest		MCD		OGK	
			c	f	c	f	c	f	c	f
5	0.1	5	1.00	0.02	1.00	0.08	1.00	0.10	1.00	0.09
		7	1.00	0.02	1.00	0.07	1.00	0.11	1.00	0.08
		9	1.00	0.02	1.00	0.08	1.00	0.09	1.00	0.08
	0.2	5	1.00	0.02	1.00	0.05	1.00	0.07	1.00	0.07
		7	1.00	0.02	1.00	0.06	1.00	0.07	1.00	0.06
		9	1.00	0.02	1.00	0.04	1.00	0.07	1.00	0.07
	0.3	5	1.00	0.03	1.00	0.02	1.00	0.03	0.96	0.05
		7	1.00	0.02	1.00	0.03	1.00	0.04	1.00	0.05
		9	1.00	0.02	1.00	0.03	1.00	0.03	1.00	0.04
	0.4	5	0.99	0.02	0.66	0.13	0.77	0.11	0.03	0.10
		7	1.00	0.02	0.84	0.06	0.98	0.02	0.24	0.07
		9	1.00	0.02	0.94	0.03	0.99	0.01	0.69	0.03
10	0.1	5	1.00	0.01	1.00	0.06	1.00	0.08	1.00	0.08
		7	1.00	0.01	1.00	0.07	1.00	0.08	1.00	0.07
		9	1.00	0.01	1.00	0.07	1.00	0.09	1.00	0.09
	0.2	5	1.00	0.02	1.00	0.04	1.00	0.06	1.00	0.06
		7	1.00	0.01	1.00	0.04	1.00	0.06	1.00	0.06
		9	1.00	0.01	1.00	0.04	1.00	0.07	1.00	0.06
	0.3	5	1.00	0.02	0.92	0.04	0.97	0.06	1.00	0.05
		7	1.00	0.02	1.00	0.02	1.00	0.04	1.00	0.05
		9	1.00	0.01	1.00	0.02	1.00	0.04	1.00	0.04
	0.4	5	0.96	0.02	0.04	0.29	0.03	0.48	0.02	0.11
		7	1.00	0.01	0.15	0.27	0.06	0.47	0.38	0.05
		9	1.00	0.02	0.28	0.22	0.18	0.43	0.91	0.03
30	0.1	5	1.00	0.01	1.00	0.06	1.00	0.09	1.00	0.08
		7	1.00	0.01	1.00	0.06	1.00	0.09	1.00	0.08
		9	1.00	0.01	1.00	0.06	1.00	0.09	1.00	0.08
	0.2	5	1.00	0.01	0.18	0.14	0.02	0.30	1.00	0.06
		7	1.00	0.01	0.34	0.12	0.26	0.24	1.00	0.06
		9	1.00	0.01	0.20	0.14	0.31	0.23	1.00	0.07
	0.3	5	1.00	0.01	0.00	0.23	0.00	0.48	1.00	0.04
		7	1.00	0.01	0.00	0.23	0.00	0.47	1.00	0.04
		9	1.00	0.01	0.00	0.23	0.00	0.47	1.00	0.05
	0.4	5	1.00	0.02	0.00	0.34	0.01	0.66	0.10	0.11
		7	1.00	0.02	0.00	0.35	0.01	0.66	0.88	0.03
		9	1.00	0.02	0.00	0.35	0.01	0.66	1.00	0.02

Table 3. Comparison of outlier detection performances in terms of c and f metrics for the evaluated methods.

p	α	δ	SDE-SSD		Mest		MCD		OGK	
			c	f	c	f	c	f	c	f
50	0.1	5	1.00	0.02	0.78	0.08	0.90	0.12	1.00	0.09
		7	1.00	0.02	0.78	0.08	0.81	0.13	1.00	0.09
		9	1.00	0.02	0.82	0.08	0.93	0.11	1.00	0.08
	0.2	5	1.00	0.02	0.00	0.19	0.00	0.40	1.00	0.06
		7	1.00	0.02	0.00	0.19	0.00	0.40	1.00	0.07
		9	1.00	0.02	0.02	0.18	0.00	0.40	1.00	0.07
	0.3	5	1.00	0.02	0.00	0.26	0.00	0.56	1.00	0.04
		7	1.00	0.02	0.00	0.27	0.00	0.56	1.00	0.05
		9	1.00	0.02	0.00	0.27	0.00	0.56	1.00	0.05
	0.4	5	1.00	0.03	0.00	0.40	0.00	0.76	0.37	0.10
		7	1.00	0.03	0.00	0.40	0.00	0.76	0.93	0.03
		9	1.00	0.03	0.00	0.41	0.00	0.76	0.99	0.03
100	0.1	5	1.00	0.02	0.02	0.17	0.00	0.31	1.00	0.09
		7	1.00	0.02	0.00	0.17	0.00	0.31	1.00	0.08
		9	1.00	0.02	0.01	0.17	0.00	0.31	1.00	0.09
	0.2	5	1.00	0.02	0.00	0.24	0.00	0.51	1.00	0.07
		7	1.00	0.02	0.00	0.24	0.00	0.51	1.00	0.07
		9	1.00	0.02	0.00	0.24	0.00	0.51	1.00	0.07
	0.3	5	1.00	0.02	0.00	0.33	0.00	0.74	1.00	0.04
		7	1.00	0.02	0.00	0.33	0.00	0.74	1.00	0.05
		9	1.00	0.02	0.00	0.34	0.00	0.74	1.00	0.05
	0.4	5	1.00	0.05	0.00	0.47	0.00	0.92	0.70	0.08
		7	1.00	0.05	0.00	0.48	0.00	0.92	1.00	0.02
		9	1.00	0.05	0.00	0.47	0.00	0.92	1.00	0.02

We report the average computation times, denoted as t, for all methods as a function of p, δ, and α in panel plots. Figure 2 presents these plots for each simulation across all methods. In Figure 2(a), we show the computation times for the projection pursuit methods: SDE-SSD, SDEH, SDE-AdjOut, and SD. The SDE-SSD method demonstrates performance comparable to SDE. However, when $\delta = 9$, SD is slightly faster than SDE-SSD.

Figure 2(b) shows the computation times for SDE-SSD, Mest, MCD, and OGK. Once again, the SDE-SSD method shows competitive performance, similar to Mest. In contrast, MCD exhibits the highest computational complexity, while OGK is the fastest method. Overall, the proposed SDE-SSD exhibits strong computational efficiency.

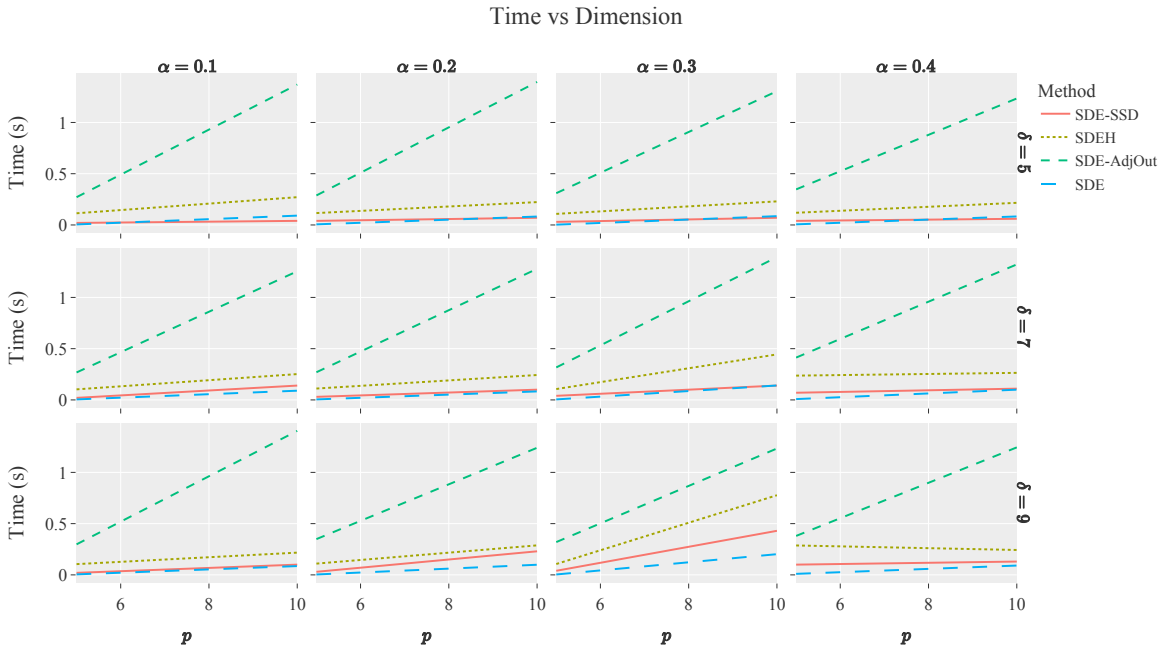
4. Applications with real data

In this section, the proposed method is applied to real-world datasets.

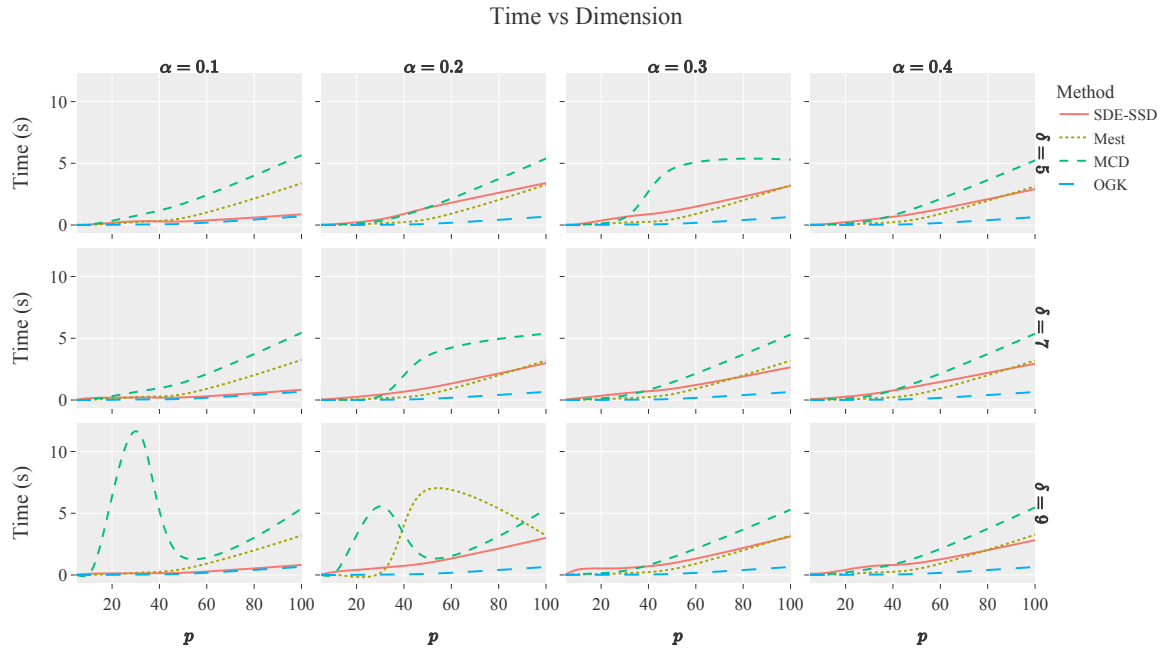
4.1 *Context*

The proposed SDE-SSD method was implemented on three datasets with different sample-space dimensions from various fields to evaluate its efficiency in identifying outliers. In all datasets, the observations labeled as outliers are known. When applying the Mest, MCD, and OGK methods to these datasets, numerical instability was observed, preventing effective evaluation in the high-dimensional dataset. This instability is attributed to an insufficient number of samples to support the model structure in higher dimensions, leading to difficulties in matrix inversion and stable estimate computation.

To compare their performance in outlier classification on these three datasets, we also implemented five other well-known methods from the literature: (i) the minimum regularized covariance determinant (MRCD) (Boudt et al., 2020); (ii) the k-means algorithm (MacQueen, 1967), where k is the number of clusters; (iii) the k-medians algorithm (Jain and Dubes, 1981), which also uses k as the number of clusters; (iv) the standard h-nearest neighbor (h-NN) algorithm (Cover and Hart, 1967), where h is the number of neighbors; and (v) the density-based spatial clustering of applications with noise (DBSCAN) (Ester et al., 1996).



(a) Computation times for the SDE-SSD, SD, SDEH, and SDE-AdjOut methods.



(b) Computation times for the SDE-SSD, Mest, MCD, and OGK methods.

Figure 2. Computation times for the SDE-SSD and competing methods. Panel (a) covers $p \in \{2, 4, 6, 8, 10\}$, while panel (b) covers $p \in \{10, 20, 40, 60, 80, 100\}$.

As mentioned, the performance metrics used were the true positive rate (c), true negative rate ($1 - f$), and corresponding accuracy measure (A).

4.2 *Arrhythmia data*

The Arrhythmia dataset contains 279 attributes, of which 206 are continuous and the remainder are nominal, designed to advance the understanding of cardiac arrhythmia (Guvenir et al., 1998). This extensive dataset primarily differentiates between the presence of cardiac arrhythmia and classifies them into one of 16 distinct categories. Class 01 represents 'normal' electrocardiogram (ECG) results, while Classes 02 through 15 correspond to various types of arrhythmia, and Class 16 encompasses all other unclassified cases.

For arrhythmia detection, we conducted experiments to assess the effectiveness of our method. The results measure the performance of SDE-SSD and compare it with other unsupervised techniques. This evaluation analyzes accuracy balanced across classes and compares the c and $1-f$ metrics of the proposed method with existing approaches in the literature.

Table 4. Classification performance for the arrhythmia, musk, and WDBC datasets using the metrics c, $1-f$, and A for SDE-SSD, MRCD, k-means, k-medians, h-NN, and DBSCAN.

Method	Arrhythmia dataset			Musk dataset			WDBC dataset		
	c	$1-f$	A	c	$1-f$	A	c	$1-f$	A
SDE-SSD	0.84	0.63	0.68	0.98	0.96	0.98	0.70	0.98	0.83
MRCD	0.45	0.00	0.45	0.96	0.96	0.50	0.17	0.98	0.80
k-means	0.53	0.59	0.57	0.02	0.96	0.57	0.37	0.99	0.88
k-medians	0.49	0.56	0.53	0.49	0.56	0.53	0.49	0.56	0.53
DBSCAN	0.54	0.46	0.54	0.00	1.00	0.50	0.40	0.98	0.82
h-NN	0.64	0.94	0.81	0.98	1.00	1.00	0.66	0.96	0.82

The evaluation of methods for arrhythmia detection reveals performance differences, as shown in Table 4, second column. The SDE-SSD method achieves a specificity of 0.63 and a c of 0.84, resulting in an overall accuracy of 0.68. In contrast, methods such as k-means, k-medians, and DBSCAN show more balanced but moderate performance, with A values around 0.57, 0.53, and 0.54, respectively. The MRCD method fails to detect arrhythmia effectively. Meanwhile, the h-NN method demonstrates high proficiency in separating data compared to unsupervised methods, achieving an A value of 0.81.

4.3 *Molecules classification data*

The Musk dataset consists of 102 unique molecules, identified by human experts as comprising 39 musks and 63 non-musks (Chapman and Jain, 1994). We classify new molecules as either musks or non-musks. This task is complicated by the fact that each molecule can adopt numerous conformations due to rotational flexibility in bond structures, resulting in structural variability. To address this diversity, all low-energy conformations were enumerated, producing a total of 6,598 conformations. Each conformation is represented by a 166-feature vector, encapsulating the shape and specific structural arrangement of the molecule in its respective state.

Table 4, third column, presents the results for this dataset. The SDE-SSD method demonstrates exceptional efficacy, achieving an almost perfect accuracy of 0.98, along with correspondingly high true positive rate and specificity values of 0.98 and 0.96, respectively.

The level of precision is comparable to that of h-NN, which also achieves near-perfect classification for this dataset. The SDE-SSD method employs advanced algorithmic mechanisms capable of handling the structural variability inherent in molecular conformations.

In contrast, the k-means and k-medians methods exhibit lower efficacy, with A values of 0.57 and 0.53, respectively. These methods face limitations in accurately classifying molecules, particularly in identifying musks, as reflected by their poor true positive values (0.02 for k-means and 0.49 for k-medians). DBSCAN fails to identify any musk molecules, although it achieves a perfect $1-f$ value, indicating high specificity.

These results suggest that methods such as k-means, k-medians, and DBSCAN struggle with the high dimensionality and dispersed nature of the dataset, which likely hinders their ability to form effective clusters.

4.4 *Wisconsin diagnostic breast cancer data*

The Wisconsin Diagnostic Breast Cancer (WDBC) dataset, developed by the University of Wisconsin and Madison Clinical Sciences Center (Wolberg et al., 1995), contains 569 samples. Each sample is characterized by 30 features derived from a digitized image of a fine needle aspirate of a breast mass. These features quantify cellular properties indicative of tumor severity, facilitating classification into two categories: malignant or benign. The dataset includes 357 benign samples and 212 malignant samples, consistent with the established medical taxonomy for non-cancerous and cancerous tumor cells, respectively.

Table 4, fourth column, presents the results for this application. The SDE-SSD method demonstrates strong proficiency and performs competitively with the k-means and h-NN algorithms.

Meanwhile, methods such as MRCD, k-medians, and DBSCAN achieve respectable A values, with DBSCAN particularly noted for its robustness in clustering-based tasks. Notably, the k-medians method shows lower overall performance and reduced efficacy in the classification of cancerous data compared to other methods.

The SDE-SSD method achieves reasonable performance in the classification of cancerous data, effectively identifying non-cancer samples but including some false negatives. Compared to other approaches, SDE-SSD exhibits competitive and robust performance.

5. Conclusions

In this work, we introduced an approach of the Stahel-Donoho estimator, structured as a finite set of $5p + 1$ directions. The $5p$ directions are estimated through stratified sampling of an informative direction seed, along with the direction that maximizes the squared sample third moment of the projected data. We proved that this approach enables the proposed Stahel-Donoho estimator with skewness-based specific directions to be faster than other Stahel-Donoho variants, without compromising its effectiveness in multivariate outlier detection.

The empirical results indicate that the Stahel-Donoho estimator with skewness-based specific directions performs proficiently in high-dimensional sample spaces and exhibits robust properties, such as a high breakdown point. However, further exploration is necessary to address other contamination scenarios, including asymmetric contaminations with multiple outlying clusters, symmetric contamination, or contamination in non-symmetric multivariate distributions.

The principal advantage of the Stahel-Donoho estimator with skewness-based specific directions lies in the informativeness of the generated directions regarding the core structure of the data. These directions provide insight into the generative process of the sample and are instrumental in identifying the primary generative mechanism. By reducing the number of directions, computational costs are decreased, thereby improving the efficiency of the estimator in outlier detection. Compared to the directions produced by Stahel (1981), those generated by our method offer richer information, leading to a better understanding of the data structure and enhancing the separability between the general sample and any outliers.

It is worth noting that the proposed modification enables the Stahel-Donoho algorithm to function effectively in high-dimensional spaces, addressing a common challenge for conventional projection pursuit methods due to computational constraints. Our approach, which utilizes fewer but more informative directions, preserves high detection efficacy while avoiding computational difficulties in higher dimensions.

Computational experiments revealed that our algorithm performed robustly compared to the studied alternatives, demonstrating consistent performance across various dimensions and levels of contamination. It remained efficient under diverse contamination scenarios, regardless of whether the outliers were close to or distant from the main data cluster, while maintaining practical computation times.

In real-world applications, the Stahel-Donoho estimator with skewness-based specific directions proved effective at classifying data and identifying outliers with high accuracy. It outperformed or matched established methods, even under demanding conditions.

Looking ahead, it is advisable to pursue more sophisticated techniques for direction selection to further enhance the precision and stability of outlier detection. Incorporating additional directional measures, such as kurtosis, may prove beneficial (Peña et al., 2010). Additionally, exploring alternative loss functions that provide greater accuracy is encouraged (Zuo et al., 2004). In addition, the impact of cellwise outliers on multivariate analysis merits closer examination due to their capacity to distort results in extensive datasets (Raymaekers and Rousseeuw, 2025).

Statements

Acknowledgement

The authors thank the Editors and Reviewers for their comments and suggestions that have improved the presentation of the article.

Author contributions

Conceptualization, S.O.; formal analysis, S.O., O.B.; investigation, S.O., O.B.; methodology, S.O.; software, S.O., O.B.; validation, S.O., O.B.; visualization, S.O., O.B.; writing –original draft preparation, S.O., O.B.; writing –review and editing, S.O.; All authors have read and agreed to the published version of the article.

Conflicts of interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this article.

Funding

This research was partially supported by Universidad EAFIT, grant number 954-000002 (S.O.), by Ministerio de Ciencia Tecnología e Innovación de Colombia, projects I) Convocatoria 852 2019 (O.B.) and II) Convocatoria 909-2 2022 (S.O.) and by Grupo Argos project 1216-852-72082 called "Descriptive and predictive analysis of cement and concrete production process" (O.B.).

References

Alqallaf, F., Van Aelst, S., Yohai, V.J., and Zamar, R.H., 2009. Propagation of outliers in multivariate data. The Annals of Statistics, 37, 311–331.

Boudt, K., Rousseeuw, P.J., Vanduffel, S., and Verdonck, T., 2020. The minimum regularized covariance determinant estimator. Statistics and Computing, 30, 113–128.

Chapman, D. and Jain, A., 1994. Musk (Version 2). UCI Machine Learning Repository. Available at https://doi.org/10.24432/C51608.

Cover, T. and Hart, P., 1967. Nearest neighbor pattern classification. IEEE Transactions on Information Theory, 13, 21–27.

Cuesta-Albertos, J.A. and Nieto-Reyes, A., 2008. The random Tukey depth. Computational Statistics and Data Analysis, 52, 4979–4988.

de Menezes, D.Q.F., Prata, D.M., Secchi, A.R., and Pinto, J.C., 2021. A review on robust M-estimators for regression analysis. Computers and Chemical Engineering, 147, 107254.

de Paula Alves, H.J. and Furtado Ferreira, D., 2020. On new robust tests for the multivariate normal mean vector with high-dimensional data and applications. Chilean Journal of Statistics, 11, 117–136

Donoho, D., 1982. Breakdown properties of multivariate location estimators. Technical report, Harvard University, Boston, MA, US.

Ester, M., Kriegel, H.P., Sander, J., and Xu, X., 1996. A density-based algorithm for discovering clusters in large spatial databases with noise. Proceedings of the Second International Conference on Knowledge Discovery and Data Mining, pp. 226–231. AAAI Press, Washington DC, US.

Filzmoser, P., Maronna, R.A., and Werner, M., 2008. Outlier identification in high dimensions. Computational Statistics and Data Analysis, 52, 1694–1711.

Gervini, D., 2002. The influence function of the Stahel-Donoho estimator of multivariate location and scatter. Statistics and Probability Letters, 60, 425–35.

Guvenir, H., Acar, B., Muderrisoglu, H., and Quinlan, R., 1998. Arrhythmia. UCI Machine Learning Repository. Available at https://doi.org/10.24432/C5BS32.

Hubert, M. and Van der Veeken, S., 2008. Outlier detection for skewed data. Journal of Chemometrics, 22, 235–246.

Jain, A.K. and Dubes, R.C., 1981. Algorithms for Clustering Data. Prentice-Hall, New York, NY, US.

Juan, J. and Prieto, F.J., 1995. A subsampling method for the computation of multivariate estimators with high breakdown point. Journal of Computational and Graphical Statistics, 4, 319–334.

Kandanaarachchi, S. and Hyndman, R.J., 2021. Dimension reduction for outlier detection using DOBIN. Journal of Computational and Graphical Statistics, 30, 204–219.

Loperfido, N., 2013. Skewness and the linear discriminant function. Statistics and Probability Letters, 83, 93–99.

Loperfido, N., 2015a. Singular value decomposition of the third multivariate moment. Linear Algebra and its Applications, 473, 202–216.

Loperfido, N., 2015b. Vector-valued skewness for model-based clustering. Statistics and Probability Letters, 99, 230–237.

Loperfido, N., 2018. Skewness-based projection pursuit: A computational approach. Computational Statistics and Data Analysis, 120, 42–57.

Loperfido, N., 2024. Tensor eigenvectors for projection pursuit. TEST, 33, 453–472.

MacQueen, J.B., 1967. Some methods for classification and analysis of multivariate observations. Proceedings of the 5th Berkeley Symposium on Mathematical Statistics and Probability, pp. 281–297. University of California Press, Oakland, CA, US.

Maronna, R.A. and Yohai, V.J., 1995. The behavior of the Stahel-Donoho robust multivariate estimator. Journal of the American Statistical Association, 90, 330–341.

Maronna, R.A. and Zamar, R.H., 2002. Robust estimates of location and dispersion for high-dimensional datasets. Technometrics, 44, 307–317.

Maronna, R.A., Martin, R.D., and Yohai, V.J., 2006. Robust statistics: Theory and methods. Wiley, New York, NY, US.

Ortiz, S., 2019. Multivariate outlier detection and robust estimation using skewness and projections. Master thesis, Universidad EAFIT, Medellín, Colombia.
Peña, D. and Prieto, F.J., 2000. The kurtosis coefficient and the linear discriminant function. Statistics and Probability Letters, 49, 257–261.
Peña, D. and Prieto, F.J., 2001. Multivariate Outlier Detection and Robust Covariance Matrix Estimation. Technometrics, 43, 286–300.
Peña, D. and Prieto, F.J., 2007. Combining random and specific directions for outlier detection and robust estimation in high-dimensional multivariate data. Journal of Computational and Graphical Statistics, 16, 228–254.
Peña, D., Prieto, F.J., and Viladomat, J., 2010. Eigenvectors of a kurtosis matrix as interesting directions to reveal cluster structure. Journal of Multivariate Analysis, 101, 1995–2007.
Raymaekers, J. and Rousseeuw, P.J., 2025. Challenges of cellwise outliers. Econometrics and Statistics, pages in press available at https://doi.org/10.1016/j.ecosta.2024.02.002
Rocke, D.M. and Woodruff, D.L., 1996. Identification of outliers in multivariate data. Journal of the American Statistical Association, 91, 1047–1061.
Rousseeuw, P.J. and Croux, C., 1993. Alternatives to the median absolute deviation. Journal of the American Statistical Association, 88, 1273–1283.
Rousseeuw, P.J. and Van Driessen, K., 1999. A fast algorithm for the minimum covariance determinant estimator. Technometrics, 41, 212–223.
Segaert, P., Hubert, M., Rousseeuw, P., and Raymaekers, J., 2020. mrfDepth: Depth measures in multivariate, regression and functional settings. Available at https://CRAN.R-project.org/package=mrfDepth. R package version 1.0.13.
Stahel, W.A., 1981. Robuste schatzungen: Infinitesimale optimalitat und schatzungen von kovarianzmatrizen. Ph.D. thesis, ETH, Zurich, Switzerland.
Sun, J., 2006. Projection pursuit. In Encyclopedia of Statistical Sciences, Vol. 10. Wiley, New York, NY, US.
Todorov, V., 2021. rrcov: Scalable robust estimators with high breakdown point. Available at https://CRAN.R-project.org/package=rrcov. R package version 1.6-0.
Van Aelst, S., 2016. Stahel-Donoho estimation for high-dimensional data. International Journal of Computer Mathematics, 93, 628–639.
Van Aelst, S., Vandervieren, E., and Willems, G., 2012. A Stahel-Donoho estimator based on huberized outlyingness. Computational Statistics and Data Analysis, 56, 531–542.
Wolberg, W., Mangasarian, O., Street, N., and Street, W., 1995. Breast Cancer Wisconsin (Diagnostic). UCI Machine Learning Repository. Available at https://doi.org/10.24432/C5DW2B.
Zuo, Y., Cui, H., and He, X., 2004. On the Stahel-Donoho estimator and depth-weighted means of multivariate data. The Annals of Statistics, 32, 167–188.