UNIVERSIDAD EAFIT

On a Combination of Skewness and Kurtosis Matrices for Projection Pursuit Exploratory Cluster Analysis

Esteban Jaramillo
ejaram11@eafit.edu.co
C.C. 1039459465

Asesor: Santiago Ortiz
sortiza2@eafit.edu.co

UNIVERSIDAD EAFIT
Escuela de Ciencias Aplicadas e Ingeniería
Ingeniería Matemática
Medellín
2025

Abstract

Skewness and kurtosis are statistical measures critical for understanding distribution characteristics, particularly in normality testing, clustering, and outlier detection. While kurtosis has been widely explored in the literature, skewness remains underutilized despite its potential for identifying asymmetrical patterns in data. Combining these measures could create a robust tool for exploratory data analysis (EDA). This research proposes a novel approach by developing a convex combination of skewness and kurtosis matrices. Using iterative procedures to maximize or minimize this combination, we aim to construct a matrix serving as a projection index for a projection pursuit algorithm. This matrix can identify clusters and outliers more effectively than either measure alone. To validate the methodology, experiments on artificial datasets and real-world data demonstrate the benefits of this combined approach in detecting non-normal features, evaluating clustering performance, and enhancing outlier detection.

Keywords: Clustering, eigenvalue iteration, robust statistics, third and fourth moments.

1 Introduction

Skewness and kurtosis are two statistical measures presented byPearson (1894, 1905), that describe how data deviate from normality. Skewness quantifies the asymmetry in a given distribution, whereas kurtosis indicates the peakedness or weight relative to the normal distribution. Both measures have been applied in normality testing, clustering, and outlier detection and are essential components of multivariate exploratory data (EDA). Additionally, MacGillivray (1986) and Mardia (1970) developed some methods that expanded the scope of these metrics, allowing the development of more robust statistical models.

Both skewness and kurtosis have an important role in clustering by evaluating the influence of the shape and distance between clusters. Joanes and Gill (1998) worked on comparing different estimators of skewness and kurtosis, demonstrated variability across sample sizes, and presented the importance of statistic selection when analyzing data for clusters. Non-normality in clustering is a well-studied concept; West et al. (1995) presented how deviations from normality affect different statistical techniques, including clustering algorithms that depend on Gaussian assumptions, such as Gaussian Mixture models as discussed by McLachlan and Peel (2000).

Hubert et al. (2017) introduced depth-based measures of skewness and kurtosis, which allowed for a better evaluation of the distribution of a potentially clustered dataset. Hennig (2004) examined how the robustness of maximum likelihood estimators in mixture modes gets influenced by them and reinforced the need for their use when clustering real-world data. Rezaei and Fränti (2016) also highlighted the distribution properties on external cluster validity measures on heavy-tailed distributions and how this impacts cluster compactness and separation.

Two key contributions come from Mardia (1970) and Mardia (1974), who first developed measures for multivariate skewness and kurtosis, establishing the basis of modern normality testing and multivariate clustering. Later, it demonstrates the significance of skewness and kurtosis in clustering applications.

Exploring these statistics allowed us to develop new methodologies for data analysis, such as projections to the maximum and minimum kurtosis vector developed by Peña and Prieto (2001a) who help identify clusters and outlying data in a dataset. Skewness projections help discover asymmetrical anomalies.

For the previously mentioned advances to be possible, the use of Projection pursuit was necessary, introduced by Friedman and Tukey (1973), this methodology identifies meaningful low-dimensional representations of high-dimensional data. Peña and Prieto (2001a,b, 2007), Peña, Prieto, and Viladomat (2010) and Loperfido (2018) have utilized skewness and kurtosis with projection pursuit to reveal clusters, outliers, and asymmetrical patterns, demonstrating their potential as projection indexes. However, these methods are generally applied independently, leaving unexplored opportunities in combining skewness and kurtosis within a unified framework.

Recent studies highlight the potential of kurtosis and skewness for enhancing multivariate data analysis. Methods based on kurtosis, such as those developed by Peña, Prieto, and Viladomat (2010), have been used to identify clusters and outliers, while projections over vectors based on skewness, as demonstrated by Ortiz (2019), excel in detecting asymmetrical anomalies. The iterative procedures behind these techniques, optimizing skewness or kurtosis matrices, highlight their complementary strengths. These complementary strengths motivate the development of a combined framework that integrates these measures to improve clustering, anomaly detection, and the identification of non-normal features.

During data analysis, effectively identifying anomalies, clusters, and non-normal features in multivariate datasets is important for applications in diverse fields. By leveraging the strengths of skewness and kurtosis measures, there is potential to develop new tools that help create a robust framework for identifying anomalies within the data. This work aims to develop a new methodology by mixing projection pursuit and the convex combination of skewness and kurtosis matrices. We aim to advance existing methodologies and enhance the ability to identify relevant data structures. The testing methodology will consist of artificial and real-world datasets exploring robustness across various parameters and configurations.

The document is structured as follows: Section 2 outlines the methodology for constructing the combined skewness-kurtosis matrix. Section 3 presents the results of numerical simulations, and Section 4 discusses experiments conducted on real-world datasets. Finally, in Section 5, there are some conclusions and future research directions.

2 Methodology

This section outlines the methodological framework for constructing a combined skewness and kurtosis matrix and demonstrates how it can serve as a projection index in EDA. The methodology is divided into two parts: the construction of the individual skewness and kurtosis matrices, based on some well-known methods of the literature; and their convex combination. Iterative numerical optimization techniques, inspired by the work of Peña and Prieto (2001a) and Ortiz (2019) form the basis for these computations.

2.1 Kurtosis Matrix

In projection pursuit, it has been shown that the projection direction that minimizes the kurtosis of the projected data, in the case of multivariate normality, converges to the Fisher Linear Discriminant Function (Peña and Prieto, 2000). Thus, given the optimality conditions in terms of discrimination of the minimum kurtosis direction, different methodologies have been developed for either outlier detection or clustering by projection pursuit. In Peña and Prieto (2001b,a), a projection pursuit method is proposed for detecting outliers and

cluster identification, respectively, using the direction that minimize the kurtosis of projected data. Later, Peña and Prieto (2007) proposed an improved version of the method by Peña and Prieto (2001b) for higher dimensions, incorporating random specific Stahel-Donoho directions. Subsequently, in Peña et al. (2010), the directions of extreme kurtosis were proposed for identifying multidimensional clusters.

The proposal of Peña and Prieto (2001b) states an iterative procedure to find these projection directions in which a kurtosis-based matrix, seen as a weighted covariance matrix as well, has been introduced in a new matrix-eigenvalue iteration optimization procedure. Consider a p–dimensional random sample $\boldsymbol{X} = (x_1, \ldots, x_n)$, which comes from a mixture of distributions of the form $\boldsymbol{X} \sim (1-\alpha)\mathcal{F} + \alpha\mathcal{G}$; where $\alpha \in (0, 1/2)$ denotes the contamination proportion, $\mathcal{F}$ denotes the probability distribution of non-outlier observations, and $\mathcal{G}$ denotes the probability distribution of outlier observations. Let the sample mean vector be $\hat{\boldsymbol{\mu}}$ and the sample covariance matrix be $\hat{\boldsymbol{S}}$, calculated for $\boldsymbol{X}$. According to Peña and Prieto (2001b) the computation of either the direction of maximum or minimum kurtosis is as follows:

1. The original data are centered and rescaled as follows:

$$y_i = \hat{\boldsymbol{S}}^{-1/2}(x_i - \hat{\boldsymbol{\mu}}), \quad i = 1, \ldots, n. \tag{1}$$

2. Let $\boldsymbol{w} \in \mathbb{R}^p$ be an unknown unit vector. Then, the direction that maximizes and minimizes the kurtosis of projected data $\boldsymbol{d}_{\max}$ and $\boldsymbol{d}_{\min}$, respectively, are obtained from the following optimization problems:

$$\begin{aligned} \boldsymbol{d}_{\max} &= \underset{\boldsymbol{w}}{\operatorname{argmax}} \ \frac{1}{n}\sum_{i=1}^{n}\left(\boldsymbol{w}^\top y_i\right)^4, \quad \boldsymbol{w}^\top\boldsymbol{w} = 1 \\ \boldsymbol{d}_{\min} &= \underset{\boldsymbol{w}}{\operatorname{argmin}} \ \frac{1}{n}\sum_{i=1}^{n}\left(\boldsymbol{w}^\top y_i\right)^4, \quad \boldsymbol{w}^\top\boldsymbol{w} = 1 \end{aligned} \tag{2}$$

3. Obtain the solution directly from the first-order optimality conditions. The optimality conditions for both optimization problems are where $\lambda \in [0, 0.5)$ represent the contamination proportion:

$$4\sum_{i=1}^{n}\left(\boldsymbol{w}^\top y_i\right)^3 y_i - 2\lambda\boldsymbol{w} = 0, \quad \boldsymbol{w}^\top\boldsymbol{w} = 1. \tag{3}$$

Multiplying by $\boldsymbol{w}$ and replacing the constraint, we obtain the value of λ, obtaining the following equality:

$$\left(\sum_{i=1}^{n}\left(\boldsymbol{w}^\top y_i\right)^3 y_i y_i^\top\right)\boldsymbol{w} = \sum_{i=1}^{n}\left(\boldsymbol{w}^\top y_i\right)^4\boldsymbol{w}, \tag{4}$$

which has the form of an spectral decomposition problem. Thus, it indicates that the optimal $\boldsymbol{w}$ will be a unit eigenvector of the kurtosis matrix:

$$K(\boldsymbol{w}) = \sum_{i=1}^{n} \left(\boldsymbol{w}^\top y_i\right)^3 y_i y_i^\top. \tag{5}$$

Note that the optimal $\boldsymbol{w}$ depends of each kind of optimization problem, i.e., for minimize the optimal solution should be the eigenvector associated with the smallest eigenvalue of $K(\boldsymbol{w})$; the opposite occurs for maximization.

2.2 Skewness Matrix

Similar to the results of Peña and Prieto (2000), Loperfido (2013) showed that in the presence of a mixture of symmetric distributions, the direction that maximizes the third sample moment aligns with the Fisher linear discriminant function. More recently, Loperfido (2018) introduced a novel procedure for computing directions that optimizes sample skewness for exploratory data analysis and preliminary outlier detection. In this way, Ortiz (2019) proposed a multivariate outlier detection methodology that leverages the projection maximizing the third sample moment of the projected data. This technique is based on an eigenvector-based matrix iteration strategy, similar to the proposal provided by Peña and Prieto (2001b). By pointing out directions that enhance skewness, it becomes feasible to identify observations that strongly diverge from the majority, thereby flagging potential outliers.

Similar to the procedure to compute $K(\boldsymbol{w})$ in Equation (5), there exist a skewness matrix, based on a contamination model. Consider a p–dimensional random sample $\boldsymbol{X} = (x_1, \ldots, x_n)$, which comes from a mixture of distributions of the form $\boldsymbol{X} \sim (1-\alpha)\mathcal{F} + \alpha\mathcal{G}$; where $\alpha \in (0, 1/2)$ denotes the contamination proportion, $\mathcal{F}$ denotes the probability distribution of non-outlier observations, and $\mathcal{G}$ denotes the probability distribution of outlier observations. Let the sample mean vector be $\hat{\boldsymbol{\mu}}$ and the sample covariance matrix be $\hat{\boldsymbol{S}}$, calculated for $\boldsymbol{X}$. According to Ortiz (2019) the computation of the direction of maximum skewness is as follows:

1. The original data are centered and rescaled as follows:

$$y_i = \hat{\boldsymbol{S}}^{-1/2}(x_i - \hat{\boldsymbol{\mu}}), \quad i = 1, \ldots, n. \tag{6}$$

2. Let $\boldsymbol{w} \in \mathbb{R}^p$ be an unknown unit vector. Then, the direction that maximizes the skewness of prejected data $\boldsymbol{f}_{\max}$ is obtained from the following optimization problem:

$$\boldsymbol{f}_{\max} = \underset{\boldsymbol{w}}{\operatorname{argmax}} \ \frac{1}{n}\sum_{i=1}^{n} \left(\boldsymbol{w}^\top y_i\right)^3, \quad \boldsymbol{w}^\top \boldsymbol{w} = 1. \tag{7}$$

3. Obtain the solution directly from the first-order optimality conditions as:

$$3\sum_{i=1}^{n} \left(\boldsymbol{w}^\top y_i\right)^2 y_i - 2\lambda \boldsymbol{w} = 0, \quad \boldsymbol{w}^\top \boldsymbol{w} = 1. \tag{8}$$

Multiplying by $\boldsymbol{w}$ and replacing the constraint, we obtain the value of λ, obtaining the following equality:

$$\left(\sum_{i=1}^{n}\left(\boldsymbol{w}^{\top}y_i\right)y_iy_i^{\top}\right)\boldsymbol{w}=\sum_{i=1}^{n}\left(\boldsymbol{w}^{\top}y_i\right)^3\boldsymbol{w}, \tag{9}$$

which has the form of an spectral decomposition problem, and is equivalent to problem in Equation (5). Thus, it indicates that the optimal $\boldsymbol{w}$ will be a unit eigenvector associated with the greatest eigenvalue of the skewness matrix:

$$A(\boldsymbol{w})=\sum_{i=1}^{n}\left(\boldsymbol{w}^{\top}y_i\right)y_iy_i^{\top}. \tag{10}$$

2.2.1 Alternative Skewness Optimization Problem

Ortiz (2019) discussed that the problem presented in Equation (7) could be rewritten, given the fact that maximum skewness projection can not detect multimodality in cases of negative asymmetry; therefore, minimum skewness should be considered instead; to enhance this situation, Ortiz and Becerra (2024) proposed a new optimization problem that considers the direction that optimizes either the absolute third moment or squared third moment of projected data. Thus, problem in Equation (7) can be reformulated by:

$$\begin{aligned}\boldsymbol{f}_{\max}^{[1]}&=\underset{\boldsymbol{w}}{\operatorname{argmax}}\left|\sum_{i=1}^{n}\frac{\left(\boldsymbol{w}^{\top}y_i\right)^3}{n}\right|,\quad \boldsymbol{w}^{\top}\boldsymbol{w}=1.\\ \boldsymbol{f}_{\max}^{[2]}&=\underset{\boldsymbol{w}}{\operatorname{argmax}}\left(\sum_{i=1}^{n}\frac{\left(\boldsymbol{w}^{\top}y_i\right)^3}{n}\right)^2,\quad \boldsymbol{w}^{\top}\boldsymbol{w}=1.\end{aligned} \tag{11}$$

Maximizing the squared skewness rather than standard skewness enhances the sensitivity of the method to asymmetrical multivariate patterns in any direction, allowing for the effective detection of extreme skewness in any asymmetrical clustering structure. Moreover, the maximization of the squared skewness simplifies the optimization process, in comparison with the absolute skewness. Thus, according to these last results, a new matrix, that we denote it as $A^*(\boldsymbol{w})$, arises considering the first-order optimality conditions; nonetheless, we will see in the next subsection how $A^*(\boldsymbol{w})$ will be computed through $A(\boldsymbol{w})$ and, a new matrix, based on a convex mixture of matrices $K(\boldsymbol{w})$ and $A(\boldsymbol{w})$ will be proposed for EDA purposes.

2.3 Third and Fourth Moments Convex Matrix

As we highlighted above, kurtosis and skewness statistics have been widely studied and implemented in statistics, but its combination as well. The weighted linear combination of these statistics, specifically the squared skewness $(m_3(X_j)^2)$ and the squared kurtosis $(m_4(X_j)^2)$ $(j=1,\ldots,p)$, considering weights given by their relative variances as

$m_3(X_j)^2+m_4(X_j)^2/4$, has been widely studied in the literature, see for instance the Jarque-Bera test for univariate normality (Jarque and Bera, 1987). The choice of $1/4$ in $m_4(X_j)^2$, as a measure of departure from normality, was discussed by D'Agostino and Pearson (1973), Pearson et al. (1977), Jones and Sibson (1987), among others.

In addition, regarding the multivariate setting, there have been similar approaches about the combination of skewness and kurtosis. Ortiz et al. (2025) developed a powerful projections pursuit method for outlier identification, that considers the projection pursuit optimization problem $\text{argmax}_{\boldsymbol{w}}\ m_3(\boldsymbol{w}^\top;Y)^2+m_4(\boldsymbol{w}^\top;Y)^2/4$, subjected to $\boldsymbol{w}^\top\boldsymbol{w}=1$. The maximizer of $m_3(\boldsymbol{w}^\top;Y)^2$, which is particularly useful for the asymmetric clusters, the minimizer of $m_4(\boldsymbol{w}^\top;Y)$, which is useful for the symmetric clusters and the linear combination $m_3(\boldsymbol{w}^\top;Y)^2+m_4(\boldsymbol{w}^\top;Y)^2/4$ provides relevant information in both symmetric and asymmetric clustering structures.

In consequence and, according to the theoretical results presented by Ortiz et al. (2025), there exists evidence that suggest that a combination of $K(\boldsymbol{w})$ and $A(\boldsymbol{w})$ (or $A^*(\boldsymbol{w})$) matrices should be useful for EDA, specifically clustering problems. Thus, we proposed a projection pursuit method defined by the convex sum of these matrices, i.e., a weighted linear combination. Using this definition, we will build a new matrix by performing a convex combination of the obtained matrices $K(\boldsymbol{w})$ and $A^*(\boldsymbol{w})$ to then perform a spectral decomposition, in order to obtain a projection direction for multivariate clustering purposes. This new matrix, denoted as $M(\boldsymbol{w})$, is defined as:

$$M(\boldsymbol{w})=(1-\phi)K(\boldsymbol{w})+\phi A^*(\boldsymbol{w}) \tag{12}$$

where $\phi\in(0,1)$ denotes the asymmetry contribution parameter, that controls the amount of skewness in $M(\boldsymbol{w})$. The value of ϕ depends on the multivariate setting, i.e., the number of clusters and their location in the hyper-space.

2.3.1 Computation of $M(\boldsymbol{w})$ the Choice of $\boldsymbol{w}$ and the Combined-Moments Projection Direction

Consider a p–dimensional random sample $\boldsymbol{X}=(x_1,\ldots,x_n)$, which comes from a mixture of distributions of the form $\boldsymbol{X}\sim(1-\alpha)\mathcal{F}+\alpha\mathcal{G}$; where $\alpha\in(0,1/2)$ denotes the contamination proportion, $\mathcal{F}$ denotes the probability distribution of non-outlier observations, and $\mathcal{G}$ denotes the probability distribution of outlier observations. Let $\boldsymbol{w},\boldsymbol{w_1},\boldsymbol{w_2}\in\mathbb{R}^p$ be unknown unit vectors and $\phi\in(0,1)$, denote the following convex matrix $M(\boldsymbol{w})=(1-\phi)K(\boldsymbol{w_1})+\phi A^*(\boldsymbol{w_2})$; then the computation of $M(\boldsymbol{w})$, $K(\boldsymbol{w_1})$ and $A^*(\boldsymbol{w_2})$ proceeds by the following steps:

1. Set $K(\boldsymbol{w_1})$ as initial matrix, $\epsilon>0$, $m\in\mathbb{N}$ and $\phi\in(0,1)$.

2. Set iteration index $l=0$ and compute an initializer unit vector $\boldsymbol{w_1}^{[l]}$. For our case, we choose $\boldsymbol{w_1}^{[l]}$ as the eigenvector associated to the smallest eigenvalue of the covariance matrix of the normalized observations Y, as is proposed by Peña and Prieto (2001b).

3. At iteration $l+1$ compute the spectral decomposition of the matrix $K\left(\boldsymbol{w_1}^{[l]}\right)$ and find the eigenvector corresponding to the smallest eigenvalue, and set it as $\boldsymbol{w_1}^{[l+1]}$.

4. Repeat Step 3. until $\left\|\boldsymbol{w_1}^{[l+1]} - \boldsymbol{w_1}^{[l]}\right\| < \epsilon$ or $l = m$, and set $\boldsymbol{w_1} = \boldsymbol{w_1}^{[l+1]}$.

5. Repeat Steps 2.–4. for $A^*(\boldsymbol{w_2})$ and $\boldsymbol{w_2}$.

   Remark 1: In Step 2. compute the eigenvector corresponding to the greatest eigenvalue.

   Remark 2: In Step 3. $A^*(\boldsymbol{w_2})$ is obtained computing the greatest absolute (or squared) eigenvalue; since, the corresponding eigenvector describes extreme skewness direction.

6. Given ϕ, compute the matrix $M(\boldsymbol{w}) = (1-\phi)K(\boldsymbol{w_1}) + \phi A^*(\boldsymbol{w_2})$ and its eigenvector corresponding to the greatest eigenvalue, i.e., $\boldsymbol{w}$.

7. Compute the projection $Z = \boldsymbol{w}^\top \boldsymbol{X}$.

At this point, the algorithm, that we will denote from now as SKURT model, describes the projection pursuit procedure and the new univariate representation of the multivariate data in $\boldsymbol{X}$. Therefore, for EDA purposes, Y denotes a reduction dimension step, for instance, in cluster analysis, one could apply standard algorithms such as k–means or mean-shifting. In the next section, some clustering applications will be addressed.

3 Numerical Experiments

This section presents the experimental framework used to evaluate the proposed methodology. The experiments are designed to assess the performance of the combined skewness-kurtosis matrix in identifying clusters and outliers across different scenarios. We test the method on both simulated and real-world datasets and evaluate its effectiveness using clustering metrics and confusion matrices.

To evaluate the results of the SKURT model, we base our numerical experiments on a fully dependent p-variant contamination model Alqallaf et al. (2009), defined as follows. Consider a p-dimensional random variable $X = (1-\alpha)X_u + \sum_{j=1}^{2} \frac{\alpha}{2} X_{oj}$, obtained as a mixture of normal distributions. Here, the random variables are defined as $X_u \sim N_p(e_0, I)$, $X_{o1} \sim N_p(\delta e_0, \lambda I)$, $X_{o2} \sim N_p(-\delta e_0, \lambda I)$; where $\delta \neq 0$ represents the distance from the main cluster, and $\lambda > 0$ controls the dispersion of each cluster, I denotes the p–dimensional identity matriz, the contamination percentage is given by $\alpha \in (0, 0.5)$ and e_0 is a p–dimensional vector of zeros.

To explore different contamination scenarios, we introduce modifications to this base model. In the case of single-cluster contamination, we add only one outlier mixture component $X_{o1} \sim N_p(\delta e_0, \lambda I)$. For asymmetric cluster mixtures, we introduce another contamination term $X_{o2} \sim N_p(-\delta e_0, \lambda \Sigma_c)$, where Σ_c represents a randomly generated covariance matrix with an elliptical shape. The evaluated parameters include $\phi \in \{0.1, 0.9\}$, which

represents the proportion of skewness/asymmetry contribution incorporated into the projection matrix, and C, which denotes the number of outlier clusters. The experiments were conducted for the following parameter values: $p \in \{2, 10, 100\}$, $\alpha \in \{0.1, 0.3, 0.5\}$, $\delta \in \{6, 10\}$, $\lambda = 0.3$, $C \in \{1, 2\}$, and $\phi \in \{0, 0.01, 0.02, ..., 0.99, 0.1\}$. We ran a cross validation methodology with 10 folds to test for robustness of the algorithm.

The evaluation of the proposed methods consists of several metrics. The primary metric is the accuracy of the K-means clustering over the projected data, which we calculate as Accuracy = CL/TL, where CL represents the number of correctly labeled samples and TL is the total number of samples. Additionally, we employ several standard clustering evaluation metrics to assess K-means performance: Davies-Bouldin Score (DB-score): Measures cluster similarity by averaging the ratio of within-cluster scatter to between-cluster separation. Lower values indicate better clustering performance (Davies and Bouldin, 1979). Adjusted Rand Index (ARI): Computes the similarity between two clustering results by quantifying the proportion of agreements, adjusted for chance (Hubert and Arabie, 1985). Normalized Mutual Information (NMI): Measures how much information one clustering provides about another, evaluating shared information between clusters (Danilchenko and Iglewicz, 2002). All simulations were performed in Python and the `scikit-learn` library (Pedregosa et al., 2011) to implement K-means clustering and compute the aforementioned evaluation metrics.

Table 1 presents an evaluation where we selected key parameter configurations that best represent the overall results for asymmetric contamination using max skewness and kurtosis protection. The selected parameters include: matrix bias between kurtosis and skewness $\phi \in \{0, 0.33, 0.5, 0.66, 1\}$, number of clusters $C = 2$, cluster separation $\delta = 10$, dispersion factor $\lambda = 0.3$, contamination level $\alpha = 0.5$, and asymmetry cluster $= True, False$, where *True* indicates the presence of an elliptical cluster, and *False* denotes a dataset with only symmetric clusters, here we extracted just some elements from ϕ in a scale to represent the different options to analyze further evolution of ϕ see Figure 2.

ϕ	Method	Dbscore	ARI	NMI	Accuracy
0	kmeans	0.4362	1.0000	1.0000	1.000
	meanshift	0.4316	0.9490	0.9658	0.971
	spectral	0.4408	1.0000	1.0000	1.000
	gmm	0.4097	1.0000	1.0000	1.000
	agglo	0.4354	1.0000	1.0000	1.000
	SKURT	0.9285	0.5727	0.6818	0.802
0 .33	kmeans	0.4419	1.0000	1.0000	1.000
	meanshift	0.4028	0.9496	0.9660	0.971
	spectral	0.4452	1.0000	1.0000	1.000
	gmm	0.4278	1.0000	1.0000	1.000
	agglo	0.4449	1.0000	1.0000	1.000
	SKURT	1.6232	0.5224	0.6273	0.773
0.5	kmeans	0.4339	1.0000	1.0000	1.000
	meanshift	0.4357	0.9490	0.9649	0.970
	spectral	0.4261	1.0000	1.0000	1.000
	gmm	0.4189	1.0000	1.0000	1.000
	agglo	0.4372	1.0000	1.0000	1.000
	SKURT	0.9444	0.6017	0.6884	0.822
0.66	kmeans	0.4415	1.0000	1.0000	1.000
	meanshift	0.3990	0.9510	0.9697	0.973
	spectral	0.4267	1.0000	1.0000	1.000
	gmm	0.4423	1.0000	1.0000	1.000
	agglo	0.4365	1.0000	1.0000	1.000
	SKURT	0.9307	0.6122	0.6901	0.836
1	kmeans	0.4327	1.0000	1.0000	1.000
	meanshift	0.4634	0.8934	0.9296	0.930
	spectral	0.4190	1.0000	1.0000	1.000
	gmm	0.4355	1.0000	1.0000	1.000
	agglo	0.4335	1.0000	1.0000	1.000
	SKURT	1.0108	0.6023	0.6709	0.831

Table 1: Clustering results grouped by ϕ value

In Table 1, we can find evidence of two factors that will be more visible in further graphs. First, we have not reached an accuracy on asymmetrical contamination that matches the performance of traditional datasets in low to high dimensionality. Second, it can start evidencing SKURT increase accuracy as it approaches $\phi = 0.5, 0.66$.

Table 2 presents a general performanc performance of SKURT with a selection of parameters to show how the model performs in different scenarios. The parameters selected try to cover most of the relevant behaviours of this model, if a behaviour is not described in this table it will be described in the subsequent graphs.

p	α	δ	c	θ	**DB Score**	**ARI**	**NMI**	**Accuracy**
				0	1.7431	0.2394	0.1585	0.774
			1	0.5	0.2179	1	1	1
		4		1	0.2482	1	1	1
				0	1.5323	0.1908	0.3329	0.627
	0.1		2	0.5	0.9566	0.3677	0.4845	0.75
				1	1.0467	0.1427	0.3228	0.568
				0	2.2590	0.1707	0.1214	0.722
			1	0.5	0.1001	1	1	1
		10		1	0.1061	1	1	1
				0	1.7811	0.1684	0.3344	0.608
			2	0.5	0.2811	0.8416	0.8750	0.936
2				1	1.4794	0.1452	0.3161	0.576
				0	0.2743	0.9881	0.9806	0.997
				0.5	0.1879	1	1	1
		4		1	0.2121	1	1	1
				0	2.9172	0.4347	0.5391	0.711
			2	0.5	0.3810	0.9096	0.9037	0.971
	0.5			1	2.2402	0.3937	0.5383	0.664
				0	0.1120	1	1	1
			1	0.5	0.0969	1	1	1
		10		1	0.0864	1	1	1
				0	0.6806	0.7275	0.7456	0.902
			2	0.5	0.1175	1	1	1
				1	0.9071	0.4861	0.6433	0.765
				0	1.2022	0.4711	0.3907	0.936
			1	0.5	1.4269	0.3431	0.2778	0.889
				1	1.7742	0.3992	0.3319	0.907
				0	2.2615	0.4987	0.5358	0.817
			2	0.5	2.4591	0.4292	0.4953	0.785
	0.1	4		1	1.2980	0.6672	0.6671	0.903
				0	2.4938	0.2814	0.2246	0.845
			1	0.5	1.2736	0.3686	0.2891	0.920
				1	1.3099	0.3174	0.2549	0.906
				0	1.1449	0.6322	0.6492	0.905
			2	0.5	1.3023	0.6068	0.6277	0.868
50	0.1	10		1	1.2111	0.6914	0.7012	0.891
				0	1.9132	0.1097	0.2307	0.668
			1	0.5	1.8784	0.1137	0.2220	0.654
				1	2.0324	0.1079	0.2096	0.662
				0	1.2611	0.6188	0.6814	0.834
			2	0.5	1.2975	0.6106	0.6973	0.831
		4		1	1.2312	0.6026	0.6913	0.824
				0	1.8993	0.1225	0.2329	0.677
	0.5		1	0.5	1.9261	0.1279	0.2470	0.679
		10		1	1.8786	0.1120	0.2314	0.666
				0	1.0169	0.6184	0.7184	0.835
			2	0.5	0.9141	0.5928	0.7080	0.815
				1	1.0272	0.6355	0.7272	0.846

Table 2: Clustering results across varying parameters (p, α, δ, c, θ). Metrics include Davies-Bouldin Score, Adjusted Rand Index (ARI), Normalized Mutual Information (NMI), and aligned accuracy.

Table 2 presents the same general behavior as the one observed in Table 1; additionally, we can observe the behavior when we have lower and higher dimensionality; initially, we can see that our accuracy in most of the lower dimensionality cases can be higher than, also we dimensionality is low and, we can observe that the impact of a higher delta (distance of the contaminated clusters to the primary dataset) performs slightly better with a higher distance, having 2 clusters in lower dimensionality will also cause a decrease in accuracy. By evaluating higher dimensionality, we can see more uniform performance throughout the entire range of parameters.

Here, we compare each model's general performance by rank. Of all the models, including a generalized comparison of all results.

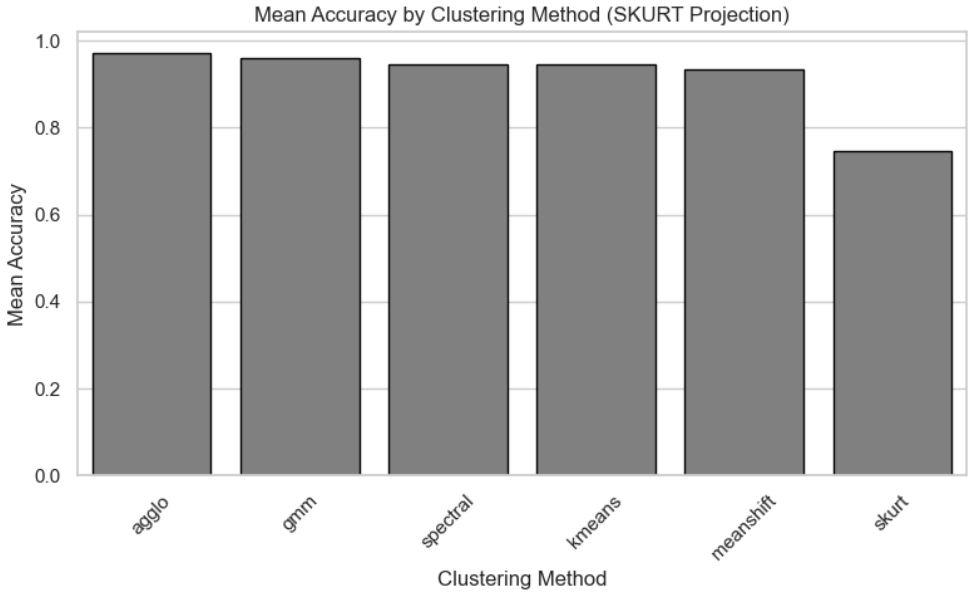

Figure 1: General mean accuracy across all models

In Figure 1 we notice that the average accuracy for SKURT is almost 0.2 accuracy points lower than the rest of the models who are almost reaching 1, we have to take in mind that this graph takes the average of all of the different scenarios, given this we will evaluate how other parameters behave. The following graph evaluates the performance of the Skurt model by desegregating by projection direction and the type of contamination weather it was a symmetrical or asymmetrical contamination.

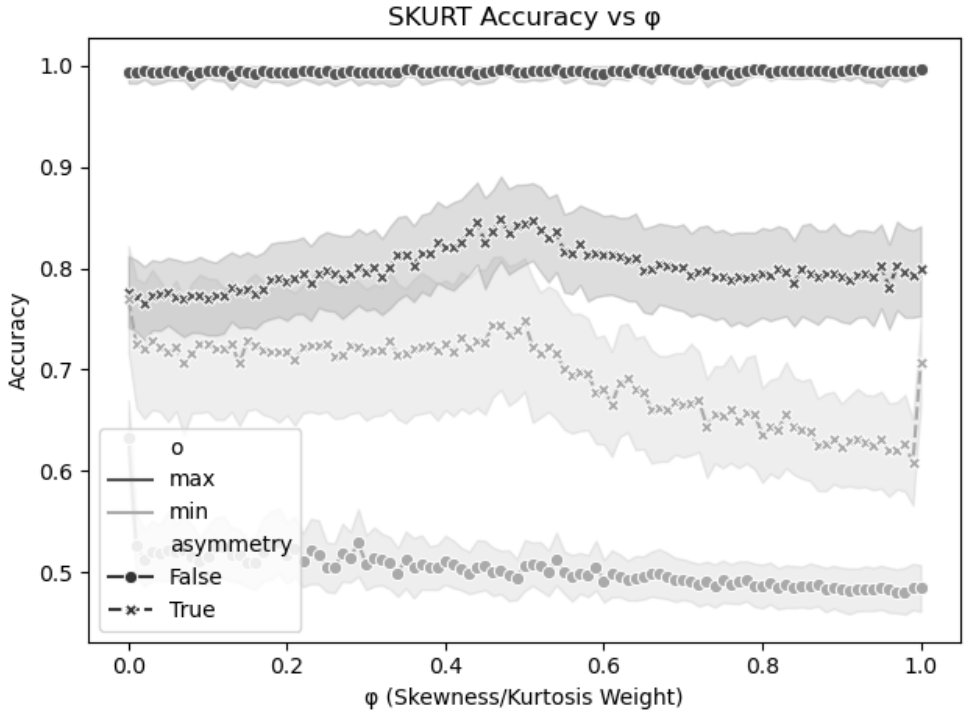

Figure 2: SKURT evaluation with asymmetry projections

When we evaluate SKURT independently, we notice two things. First, when utilizing the maximization model with no asymmetry, the SKURT model performs on par with other models, reaching an average accuracy close to 1. Also, the best-performing direction is the direction that maximizes the Skewness and kurtosis. In contrast, the direction that minimizes Skewness and kurtosis is a direction that performs poorly, affecting the overall performance versus the other models; when evaluating symmetric contamination, we noticed that ϕ does not significantly impact the overall performance. When we evaluate the models with asymmetric contamination, we notice a decrease in the accuracy concerning the use of symmetrical clusters; we also notice that the best performance for clustering is reached when using a skewness/kurtosis weight close to 0.5, meaning that having a mix of Skewness and kurtosis in the SKURT model performs better than simple Skewness and kurtosis Finally when we evaluate the minimization projections when $\phi = 0, 1$ the accuracy increases, 1 when we have asymmetrical clusters and 1 when we have symmetrical clusters.

In Figure 3 and 4 below we will be evaluating the performance of the different models when we have changes in the dimensionality and contamination oof the dataset, since the best performance we have perceived has been with projections to the direction that maximizes we will focus on that direction we will be comparing the average accuracy to the dimension value. we will evaluate these two parameters together as they posses similar behaviours

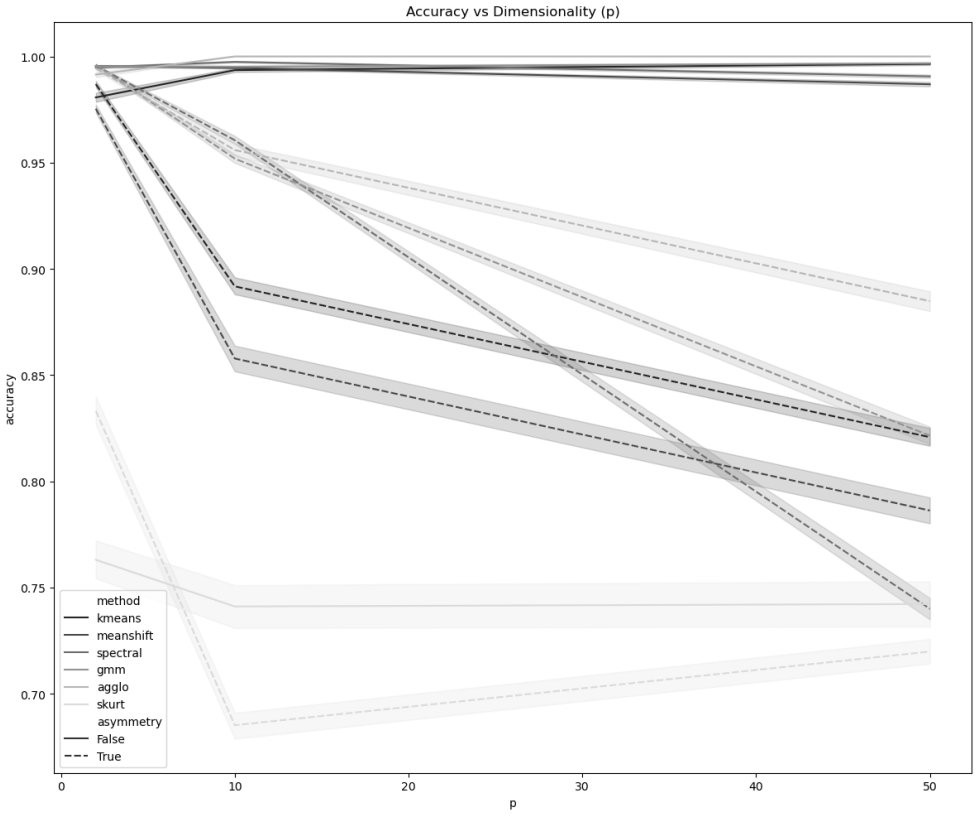

Figure 3: Impact of accuracy Vs Dimensionality

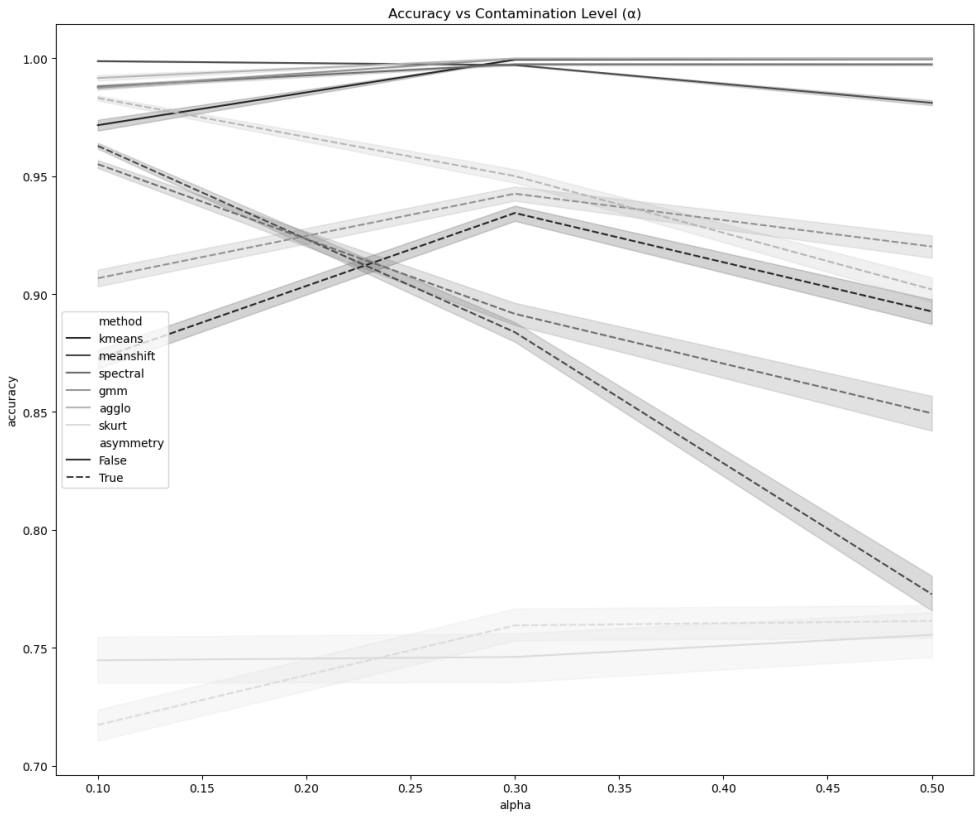

Figure 4: Accuracy VS contamination level

In both Figure 3 and Figure 4, we can observe that the SKURT methodology has certain stability and it is not affected to the increase of either dimension or amount of contaminated data sets, as we oberved in previous figures, we know that SKURT has a lower performance than the other modles, non the less these two images show that as the amount of contamination or dimension decrease the performance of some of the other models such as agglomerative and mean shift, in their performance specially with datasets with high contamination.

4 Test with Real Data

The proposed method and the comparative methods were tested on three real-world datasets of different dimensions to evaluate their effectiveness in clustering analysis. These datasets are the Wine Quality Dataset Cortez et al. (2009), available in the UCI Machine Learning Repository, the Road Accident Survival Dataset Sarder (2021) from Kaggle, and the Weight and Height Dataset Serge (2023) from Kaggle. All three datasets are widely used in machine learning research, as they present complex characteristics and multivariate distributions, making them suitable for evaluating the performance of clustering algorithms in uncovering hidden patterns and natural groupings. The Weight and Height Dataset consists of anthropometric measurements collected from the National Health and Nutrition Examination Survey (NHANES), providing a real-world example of biometric data

distribution, which is valuable for clustering techniques in health and demographic studies.

The alpha parameter, which is the only parameter we can modify, will be Set at 0.1. We will use a matrix with bias for skewness, which has shown better performance through our experiments with this dataset. Table 3 evaluates the Wine Quality Dataset, taken from UCI Cortez et al. (2009). It is used to evaluate the accuracy with which an algorithm is able to cluster or identify outlier data and it is frequently used to test other clustering models. this dataset is composed of multiple characteristics that define a wine's quality such as alcohol level, alkalinity and flavanoids, the dataset contains 13 variables and 178 samples. The Wine Quality Dataset is widely used for developing and validating machine learning models, particularly in clustering tasks where the goal is to group wines with similar chemical compositions and quality scores. The attributes include factors such as acidity, sugar content, alcohol percentage, and pH levels, which are essential for evaluating the quality of the wine. Clustering can help identify natural groupings within the dataset, such as wines of similar quality profiles or production characteristics. The dataset consists of both typical and atypical values, where atypical values can be identified as wine samples with physicochemical properties that deviate significantly from the majority, indicating unique chemical compositions or potential measurement errors. This dataset serves as a valuable resource for testing the effectiveness of the proposed method and other clustering techniques. Our objective is to determine which method best identifies natural groupings in high-dimensional data. Table 3 presents the results obtained by applying different clustering models to this dataset.

SKURT 0.1

	0	1	2
0	46	0	13
1	1	50	20
2	0	19	29

Accuracy: 0.702

SKURT 0.5

	0	1	2
0	46	0	13
1	1	50	20
2	0	19	29

Accuracy: 0.7022

SKURT 0.9

	0	1	2
0	31	27	1
1	8	0	63
2	11	0	37

Accuracy: 0.382

K-Means

	0	1	2
0	27	1	31
1	0	64	7
2	0	37	11

Accuracy: 0.573

Mean Shift

	0	1	2
0	41	2	16
1	0	65	6
2	0	40	8

Accuracy: 0.64044

Spectral Clustering

	0	1	2
0	46	0	13
1	2	45	24
2	0	12	36

Accuracy: 0.7134

Gaussian Mixture

	0	1	2
0	48	0	11
1	1	65	5
2	0	39	9

Accuracy: 0.68539

Agglomerative Clustering

	0	1	2
0	46	0	13
1	2	51	18
2	0	21	27

Accuracy: 0.6966

Table 3: Confusion Matrices for Wine Dataset Cortez et al. (2009) Clustering methodologies Comparison implementing SKURT, K-Means, Aggomerative Clustering, Gaussian Mixture, Mean Shift and Spectral Clustering

In the clustering analysis of the Wine Quality Dataset, the SKURT-based methods demonstrated varying performance depending on the α parameter. For SKURT $\alpha = 0.1$ and SKURT $\alpha = 0.5$, both achieved an accuracy of approximately 70.2%, showing strong clustering capabilities and performing comparably to traditional methods like Spectral Clustering (71.34%) and Gaussian Mixture Models (GMM) (68.54%). These results suggest that lower α values in SKURT maintain the dataset's natural structure, enabling effective separation of wine samples based on their chemical properties. However, at SKURT $\alpha = 0.9$, accuracy dropped significantly to 38.2%, indicating that increasing the skewness-kurtosis transformation too aggressively distorts the data, making it difficult for the model to find meaningful clusters. This decline highlights the sensitivity of SKURT to the choice of α, where higher values can introduce excessive transformations that reduce clustering effectiveness.

Overall, SKURT $\alpha = 0.1$ and SKURT with $\alpha = 0.5$ proved to be viable clustering approaches, yielding accuracy levels similar to other widely used methods. While Spectral

Clustering and GMM remain the top-performing models, the SKURT-based approaches at lower α values demonstrate strong potential in clustering analysis. However, careful tuning of α is essential, as excessive transformations (as seen in SKURT $\alpha = 0.9$) can lead to poor performance and loss of meaningful structure in the data.

The Road Accident Survival Dataset is extensively utilized for developing and validating predictive models in accident risk analysis and can be effectively applied in clustering scenarios to group accidents with similar survival outcomes or contributing factors. The attributes include variables such as accident location, time, age of the victims, and injury severity, which are crucial for assessing survival probabilities. Clustering can reveal hidden patterns, such as groups of accidents with common risk factors or similarities in survival rates, aiding in accident prevention strategies. The dataset contains both common and extreme cases, where extreme cases can be identified as accidents with survival outcomes that significantly deviate from the majority, highlighting unusual accident conditions or data recording errors. This dataset provides a valuable benchmark for evaluating the performance of the proposed clustering method and other unsupervised learning techniques. this dataset provides 6 variables and 101 observations. Our objective is to identify the most effective method for detecting patterns in high-dimensional accident data. Table 4 summarizes the results obtained by applying different clustering models to this dataset.

SKURT 0.1			**SKURT 0.5**			**SKURT 0.9**		
	0	1		0	1		0	1
0	0	98	0	55	43	0	59	39
1	0	101	1	46	55	1	48	53
Accuracy: 0.5075			Accuracy: 0.5527			Accuracy: 0.5628		
K-Means			**Mean Shift**			**Spectral Clustering**		
	0	1		0	1		0	1
0	48	50	0	0	98	0	53	45
1	45	56	1	0	101	1	51	50
Accuracy: 0.5226			Accuracy: 0.5075			Accuracy: 0.5175		
Gaussian Mixture			**Agglomerative Clustering**					
	0	1		0	1			
0	37	61	0	38	60			
1	34	67	1	33	68			
Accuracy: 0.5226			Accuracy: 0.53266					

Table 4: Confusion matrices for car accident dataset clustering methodology comparisons, implementing SKURT, K-Means, Agglomerative Clustering, Gaussian Mixture, Mean Shift, and Spectral Clustering. Data sourced from Sarder (2021).

In the clustering analysis of the Car Accident Dataset, the SKURT-based transformations demonstrated varying levels of effectiveness depending on the α parameter. At SKURT with $\alpha = 0.1$, the model struggled significantly, achieving an accuracy of only 50.75%, which is close to random classification. This result suggests that at $\alpha = 0.1$, the dataset's structure may not be sufficiently transformed to facilitate meaningful clustering. Regarding SKURT with $\alpha = 0.5$, accuracy improved to 55.27%, indicating that a moderate transformation helps better separate clusters, albeit still not reaching optimal performance. The clustering capability was slightly enhanced, suggesting that adjusting the skewness-kurtosis projection improves the separability of accident survival outcomes. SKURT with $\alpha = 0.9$ yielded the highest accuracy among the three SKURT transformations at 56.28%, showing that more aggressive transformations in this dataset contribute to better-defined clusters. However, despite this improvement, the accuracy remains relatively low, suggesting that Car Accident data may not be inherently well-suited for SKURT-based transformations alone.

Comparing SKURT transformations with traditional clustering methods, K-Means (52.26%), Mean Shift (50.75%), and Spectral Clustering (51.75%) performed at similar levels, reinforcing the dataset's challenging nature for clustering. Interestingly, Spectral Agglomerative Clustering (53.27%) and Spectral Gaussian Mixture Models (52.26%) offered marginal im-

provements, but none of the methods achieved high accuracy. These results suggest that the Car Accident Dataset may lack a strong cluster structure, making it difficult for unsupervised methods to effectively group accident survival outcomes. Overall, while SKURT with $\alpha = 0.9$ showed the best performance among the SKURT methods, the general difficulty of clustering this dataset is evident, as all methods struggled to significantly exceed random classification levels. This highlights the complexity and potential limitations of clustering approaches for this particular dataset.

The Weight and Height Dataset from NHANES is widely used for developing and validating predictive models in anthropometric analysis and can be effectively applied in clustering scenarios to group individuals with similar physical characteristics. The dataset includes attributes such as weight, height, age, and gender, which are crucial for identifying population trends and health-related insights. Clustering can uncover hidden patterns, such as groups with common body measurements or deviations from typical growth distributions, aiding in studies related to obesity, nutrition, and overall health assessment. The dataset contains both common and extreme cases, where extreme cases can be identified as individuals with measurements significantly deviating from the majority, highlighting potential outliers or unique physiological conditions. The dataset is composed by 4 variables and 8388 observations. This dataset provides a valuable benchmark for evaluating the performance of various clustering methods and other unsupervised learning techniques. Our objective is to identify the most effective method for detecting patterns in high-dimensional anthropometric data. Table 5 summarizes the results obtained by applying different clustering models to this dataset.

SKURT 0.1

	0	1	2	3
0	0	18	565	1476
1	0	1565	1872	153
2	0	341	1470	806
3	0	0	1	121

Accuracy: 0.37

SKURT 0.5

	0	1	2	3
0	0	25	630	1404
1	0	2328	1145	117
2	0	419	1472	726
3	0	0	1	121

Accuracy: 0.46

SKURT 0.9

	0	1	2	3
0	0	25	630	2505
1	0	1645	1828	117
2	0	0	1	121
3	0	25	630	1404

Accuracy: 0.3867

K-Means

	0	1	2	3
0	1422	10	627	0
1	114	173	1045	678
2	732	298	1587	0
3	120	0	2	0

Accuracy: 0.5677

Mean Shift

	0	1	2	3
0	0	2059	0	0
1	180	3407	3	0
2	0	2617	0	0
3	0	122	0	0

Accuracy: 0.4061

Spectral Clustering

	0	1	2	3
0	1278	0	781	0
1	19	1672	364	1535
2	533	683	1373	28
3	109	0	13	0

Accuracy: 0.5153

Gaussian Mixture

	0	1	2	3
0	1859	0	200	0
1	7	2047	131	1405
2	230	285	2102	0
3	122	0	0	0

Accuracy: 0.7162

Agglomerative Clustering

	0	1	2	3
0	1285	35	739	0
1	47	2284	594	665
2	425	787	1405	0
3	115	0	7	0

Accuracy: 0.5929

Table 5: Confusion matrices for BMI clustering methodology comparisons, implementing SKURT, K-Means, Agglomerative Clustering, Gaussian Mixture, Mean Shift, and Spectral Clustering. Data sourced from Serge (2023).

In the clustering analysis of the BMI dataset, the SKURT transformations exhibited mixed performance across different values of α. At SKURT with $\alpha = 0.1$, the model achieved an accuracy of 37%, indicating that the transformation did not significantly improve cluster separation. The confusion matrix reveals that class misclassification was high, particularly in clusters 2 and 3, where data points were widely spread across multiple categories. Increasing the SKURT parameter to $\alpha = 0.5$ led to a notable improvement in accuracy, reaching 46%. This suggests that a moderate transformation better aligns the dataset's structure with the clustering model, reducing misclassification errors. However, at SKURT $\alpha = 0.9$, the accuracy dropped again to 38.67%, implying that excessive transformation may distort the underlying distribution rather than enhance separability.

Compared to SKURT, traditional clustering methods showed significant variation in performance. K-Means achieved the highest accuracy at 56.77%, demonstrating its effectiveness in partitioning the BMI data into distinct clusters. Spectral Clustering (51.53%) and Agglomerative Clustering (59.29%) also performed reasonably well, indicating that hierarchical and spectral methods were somewhat capable of capturing the dataset's structure. Gaussian Mixture Models (71.62%) yielded the best performance overall, suggesting

that probabilistic approaches best fit the underlying data distribution. Meanwhile, Mean Shift performed the worst at 40.61%, likely due to its sensitivity to the dataset's density variations. Overall, while SKURT $\alpha = 0.5$ provided the best performance among SKURT transformations, it was still outperformed by Gaussian Mixture Models and K-Means. These results suggest that the BMI dataset benefits more from probabilistic and centroid-based clustering approaches rather than skewness-kurtosis transformations, reinforcing the importance of selecting the right clustering technique based on dataset characteristics.

5 Conclusions

The model performs the best with nonsymmetrical contaminations and uses the direction that maximizes SKURT, performing on par with the other models, where ϕ does not significantly affect the performance. What we find interesting comes when we evaluate the performance with asymmetrical clusters for higher dimension and contamination levels, which tell a similar story, as these parameters increase the performance of several models like mean shift, agglomerative clustering, and spectral decomposition decrease more than other models. In contrast, although SKURT initially shows a lower accuracy on lower values of these parameters, as they increase, SKURT does not show degradation in its performance, giving us insight that it can be helpful in specific scenarios. Finally, we have two interesting findings evaluating the performance of SKURT as we change the distribution of the convex combination of Skewness and kurtosis. First, having a balanced combination around $\phi = 0.5$ increases the accuracy with asymmetrical data.

The methodology faces difficulties with high-dimensional data, where overlapping clusters and noise have a significant impact on clustering performance. The convex parameter (ϕ) strongly affects the results. While the skewness bias benefits specific datasets, the choice of ϕ is not universally optimal and requires careful tuning based on the data's characteristics. Performance is still limited in datasets with high asymmetry and contamination despite the advantages of skewness-biased combinations. The combined skewness and kurtosis matrix enhances cluster and outlier detection in exploratory data analysis. In symmetric datasets with well-separated clusters, the method achieves near-perfect clustering metrics. However, its performance declines in high-dimensional data and when clusters significantly overlap. This work demonstrates that combining Skewness and kurtosis with a convex framework is possible; it offers flexibility when exploring data structures and highlights the importance of appropriately tuning the parameter ϕ based on the dataset's characteristics, especially when faced with asymmetric contaminations.

We can further study this topic in several ways. First, we could evaluate how to optimize the ϕ parameter by implementing methodologies such as Bayesian optimization or Stochastich annealing. Second, we could evaluate how to achieve a higher clustering accuracy. Techniques based on Rényi (1960) that use Renyi entropy to maximize or minimize the entropy between clusters, such as the one presented by Jenssen et al. (2003), could be applied to a SKURT projection for a possibly better clustering methodology. Third, we

could identify other types of datasets and contaminations where this projection could be helpful. Fort, test for outlier detection capabilities by applying methodologies like the ones proposed by Peña and Prieto (2001b) where Stahel-Donoho outlying ness is used alongside maximum kurtosis to reduce the randomness of Stahe-donoho framework. The same could be achieved by implementing SKURT directions for outlier detection. Finally, we could experiment by utilizing different clustering algorithms instead as a base with SKURT projection, Such as mean shift algorithms, where we do not have to provide the number of clusters and are less sensitive to the cluster shape.

References

Alqallaf, F. et al. (2009). Propagation of outliers in multivariate data. *The Annals of Statistics*, 37(1):311–331.

Cortez, P., Cerdeira, A., Almeida, F., Matos, T., and Reis, J. (2009). Wine quality dataset. UCI Machine Learning Repository https://archive.ics.uci.edu/dataset/186/wine+quality. Accessed: 2025-02-16.

D'Agostino, R. and Pearson, E. S. (1973). Tests for departure from normality. empirical results for the distributions of b2 and b1. *Biometrika*, 60(3):613–622.

Danilchenko, M. and Iglewicz, B. (2002). Normalized mutual information for clustering. In *Proceedings of the 2002 IEEE International Conference on Data Mining (ICDM)*, pages 25–32. IEEE.

Davies, D. L. and Bouldin, D. W. (1979). A cluster separation measure. *IEEE Transactions on Pattern Analysis and Machine Intelligence. PAMI-1 (2): 224-227.*

Friedman, J. H. and Tukey, J. W. (1973). A projection pursuit algorithm for exploratory data analysis. *Computational Statistics & Data Analysis*, 20:669 – 687.

Hennig, C. (2004). Breakdown points for maximum likelihood estimators of location-scale mixtures. *The Annals of Statistics*, 32(4):1313–1340.

Hubert, L. and Arabie, P. (1985). Comparing partitions. *Journal of Classification*, 2(1):193–218.

Hubert, M., Rousseeuw, P. J., and Segaert, P. (2017). Multivariate and functional skewness and kurtosis based on depth. *Statistical Methods & Applications*, 26:235–256.

Jarque, C. M. and Bera, A. K. (1987). A test for normality of observations and regression residuals. *International Statistical Review / Revue Internationale de Statistique*, 55(2):163.

Jenssen, R., Hild, K. E., Erdogmus, D., Principe, J. C., and Eltoft, T. (2003). Clustering using Rényi's entropy. In *Proceedings of the International Joint Conference on Neural Networks (IJCNN)*, volume 1, pages 523–528. IEEE.

Joanes, D. N. and Gill, C. A. (1998). Comparing measures of sample skewness and kurtosis. *Journal of the Royal Statistical Society: Series D (The Statistician)*, 47(1):183–189.

Jones, M. C. and Sibson, R. (1987). What is projection pursuit? *Journal of the Royal Statistical Society. Series A*, 150:1–37.

Loperfido, N. (2013). Skewness and the linear discriminant function. *Statistics & Probability Letters*, 83(1):93 – 99.

Loperfido, N. (2018). Skewness-based projection pursuit: A computationalapproach. *Computational Statistics and Data Analysis*, 120:42 – 57.

MacGillivray, H. L. (1986). Skewness and asymmetry: Measures and orderings. *The annals of Statistics*, 14:994 – 1011.

Mardia, K. V. (1970). Measures of multivariate skewness and kurtosis with applications. *Biometrika*, 57:519.

Mardia, K. V. (1974). Applications of some measures of multivariate skewness and kurtosis in testing normality and robustness studies. *Sankhyā: The Indian Journal of Statistics*, 36:115 –128.

McLachlan, G. and Peel, D. (2000). *Finite Mixture Models*. John Wiley & Sons.

Ortiz, S. (2019). Multivariate outlier detection and robust estimation using skewness and projections. Master's thesis, Universidad EAFIT, Medellín, Colombia.

Ortiz, S. and Becerra, O. (2024). On a stahel–donoho estimator with skewness-based random projection directions. *Chilean Journal of Statistics*, 15(2):110–125.

Ortiz, S., Laniado, H., Peña, D., and Prieto, F. J. (2025). Dimension reduction for outlier detection in high-dimensional data. *Journal of Multivariate Analysis*. In Press.

Pearson, E. S., D'Agostino, R. B., and Bowman, K. O. (1977). Tests for departure from normality: Comparison of powers. *Biometrika*, 64(2):231 – 246.

Pearson, K. (1894). Contributions to the mathematical theory of evolution. *Philosophical Transactions of the Royal Society of London. A*, 185:71 – 110.

Pearson, K. (1905). *On the general theory of skew correlation and non-linear regression*. Number 14. Dulau and Company.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., and Duchesnay, (2011). Scikit-learn: Machine learning in python. *Journal of Machine Learning Research*, 12:2825–2830.

Peña, D. and Prieto, F. J. (2000). The kurtosis coefficient and the linear discriminant function. *Statistics & Probability Letters*, 49(3):257 — 261.

Peña, D. and Prieto, F. J. (2001a). Cluster identification using projections. *Journal of the American Statistical Association*, 96:1433 – 1445.

Peña, D. and Prieto, F. J. (2001b). Multivariate outlier detection and robust covariance matrix estimation. *Technometrics*, 43:286 – 300.

Peña, D. and Prieto, F. J. (2007). Combining random and specific directions foroutlier detection and robust estimation inhigh-dimensional multivariate data. *Journal of Computational and Graphical Statistics*, 16:228 – 255.

Peña, D., Prieto, F. J., and Viladomat, J. (2010). Eigenvectors of a kurtosis matrix as interesting directions to revealcluster structure. *Journal of Multivariate Analysis*, 101:1995 – 2007.

Rezaei, M. and Fränti, P. (2016). Set matching measures for external cluster validity. *IEEE Transactions on Knowledge and Data Engineering*, 28(8):2173–2186.

Rényi, A. (1960). On measures of information and entropy. In Neyman, J., editor, *Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability*, volume I, pages 547–561, Berkeley, CA, USA. University of California Press.

Sarder, H. (2021). Road accident survival dataset. Kaggle, https://www.kaggle.com/datasets/himelsarder/road-accident-survival-dataset. Accessed: 2025-02-16.

Serge, G. (2023). Weight and height from nhanes. Kaggle, https://www.kaggle.com/datasets/galserge/weight-and-height-from-nhanes. Accessed: 2025-02-16.

West, S. G., Finch, J. F., and Curran, P. J. (1995). Structural equation models with nonnormal variables: Problems and remedies. *Structural Equation Modeling: A Multidisciplinary Journal*, 1(1):56–75.