Unveiling Patterns in High Dimensions:

AI/ML Innovations in Multivariate Analysis, Dimension Reduction, and Clustering

Abisa Sinha Adhikary[1]

[1]Department of Statistics
Amity Institute of Applied Sciences
Amity University of Calcutta

AI/ML Applications in Astronomy and Astrophysics,
10th January 2025

Table of Contents

Table of Contents

Introduction

Astronomical data, from any domain, is usually available in forms:

- High dimensions: An increase in dimension reduces the data visibility, increases sparsity and computational complexity.
- Presence of outlying observations and their proper detection.
- Presence of missing values
- Presence of '0' values

- All these issues increases manifold with increase in the data size, and without proper handling of which, it may cause serious problems in the modeling and inference.

Table of Contents

Independent Component Analysis

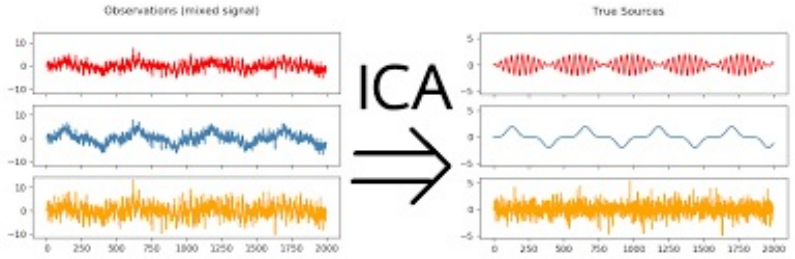
Usual dimension reduction techniques like the PCA have underlying assumption of Gaussian structure of the data in concern. For non-Gaussian data, PCA is not applicable. Even if applied, gives hallucinated results.

Independent Component Analysis

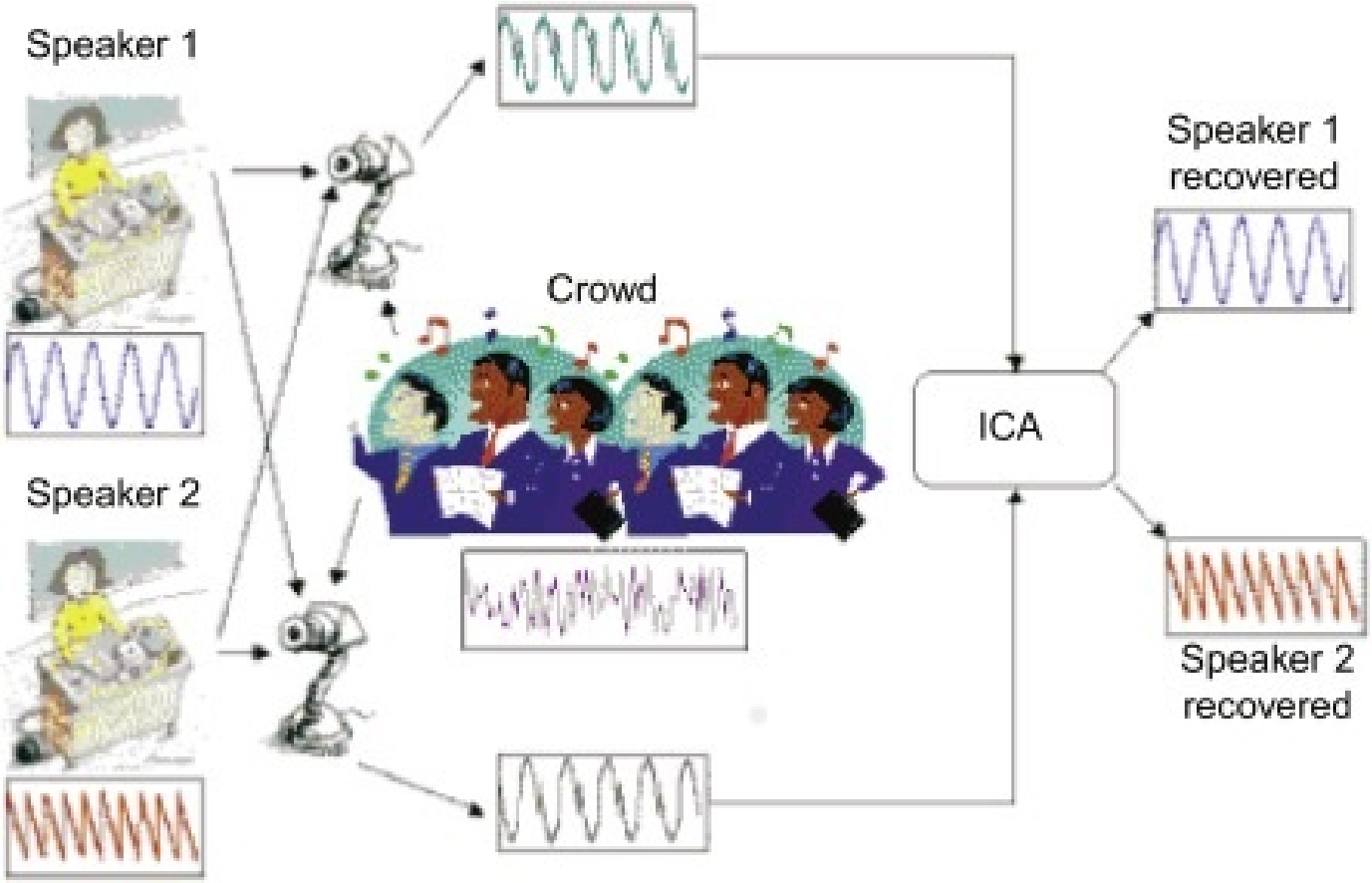
Usual dimension reduction techniques like the PCA have underlying assumption of Gaussian structure of the data in concern. For non-Gaussian data, PCA is not applicable. Even if applied, gives hallucinated results. Independent Component Analysis is a multivariate dimension reduction technique that works on data belonging to a non-Gaussian set-up. It is a method that identifies hidden factors in random variables by separating a mixed signal contaminated with noise into independent components. Moreover, the components, in addition to being uncorrelated, are also mutually independent.

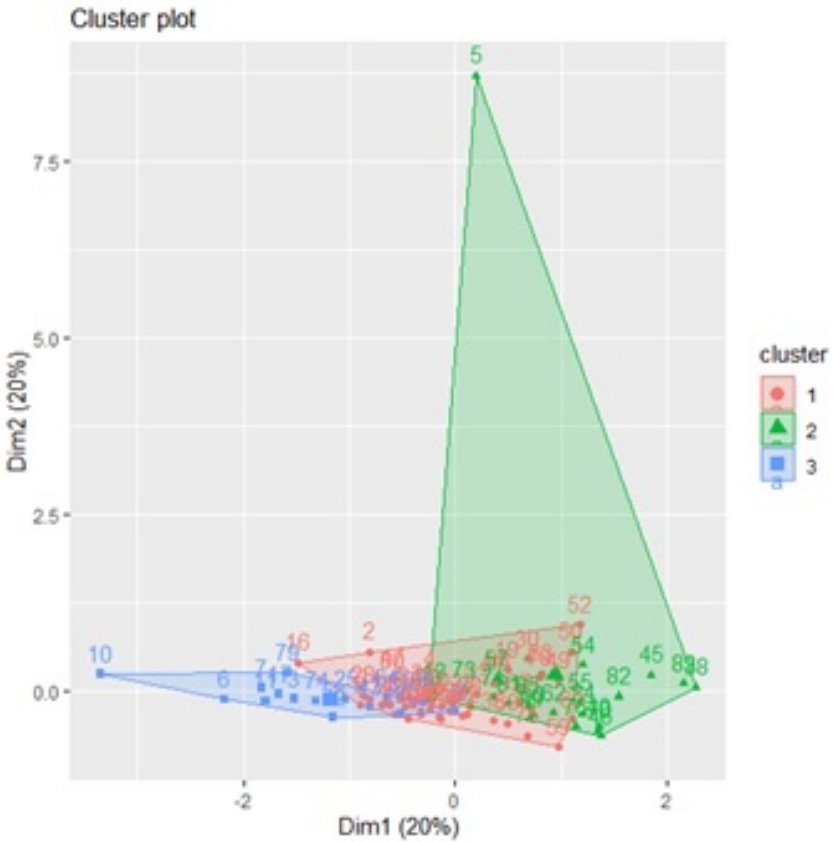
Independent Component Analysis

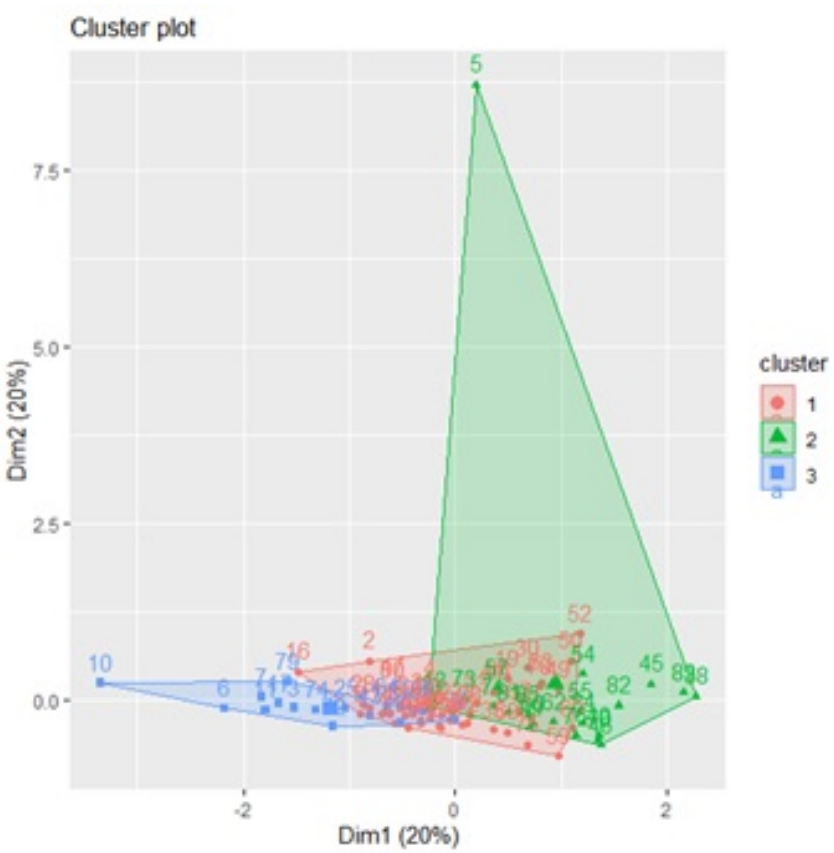
Usual dimension reduction techniques like the PCA have underlying assumption of Gaussian structure of the data in concern. For non-Gaussian data, PCA is not applicable. Even if applied, gives hallucinated results. Independent Component Analysis is a multivariate dimension reduction technique that works on data belonging to a non-Gaussian set-up. It is a method that identifies hidden factors in random variables by separating a mixed signal contaminated with noise into independent components. Moreover, the components, in addition to being uncorrelated, are also mutually independent.

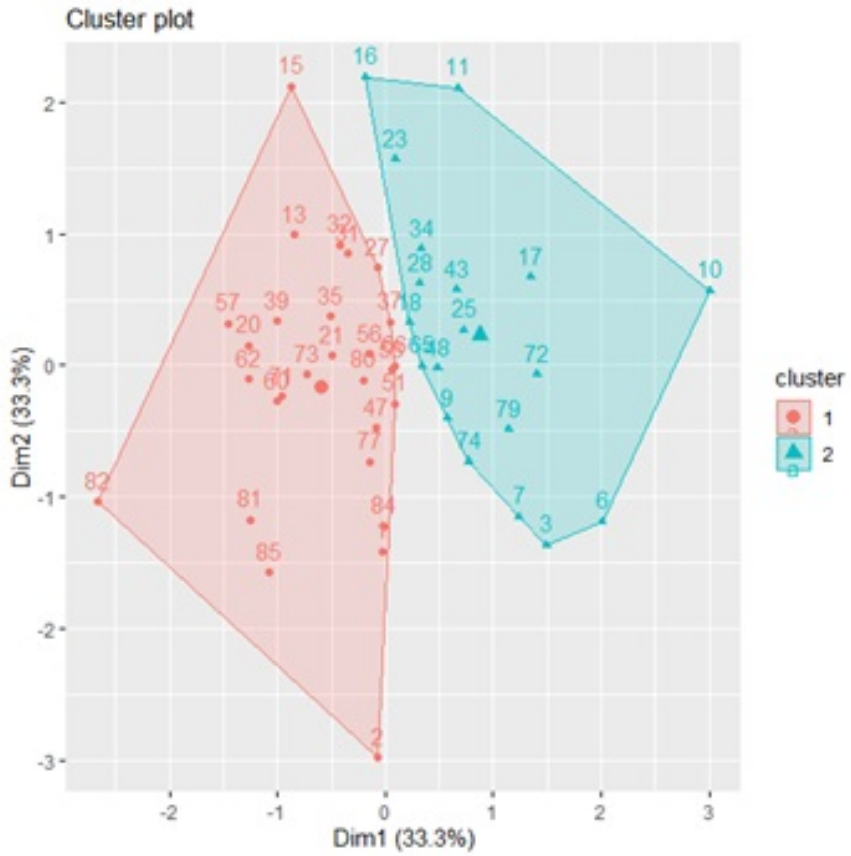


ICA


Speaker 1
Speaker 2
Crowd
ICA
Speaker 1
recovered
Speaker 2
recovered


Cluster plot
Dim2 (20%)
Dim1 (20%)
7.5
5.0
2.5
0.0
-2
0
2
cluster
1
2
3


Cluster plot
Dim2 (20%)
Dim1 (20%)
cluster
1
2
3


Cluster plot
Dim2 (33.3%)
Dim1 (33.3%)
cluster
1
2

ICA performs best when the data comes from a Non-Gaussian setup, BUT if that data is contaminated with outliers then ICA also fails.

ICA performs best when the data comes from a Non-Gaussian setup, BUT if that data is contaminated with outliers then ICA also fails.
What are outliers?

ICA performs best when the data comes from a Non-Gaussian setup, BUT if that data is contaminated with outliers then ICA also fails.
What are outliers?
How can they be detected?

ICA performs best when the data comes from a Non-Gaussian setup, BUT if that data is contaminated with outliers then ICA also fails.
What are outliers?
How can they be detected?
Outlier may be defined as those data point which differs *significantly* from other observations.

ICA performs best when the data comes from a Non-Gaussian setup, BUT if that data is contaminated with outliers then ICA also fails.
What are outliers?
How can they be detected?
Outlier may be defined as those data point which differs *significantly* from other observations.
There are various tools available in Statistics/ ML for outlier detection.

Robust PCA

Without prior detection and even removal of outliers, the Robust PCA is a better option than the conventional PCA or ICA. If the data is represented in form of $n \times p$ matrix with n being the number of objects and p being the original number of variables, $p \geq n$, the ROBPCA method involves both the methods of projection pursuit and robust estimation.

Robust PCA

Without prior detection and even removal of outliers, the Robust PCA is a better option than the conventional PCA or ICA. If the data is represented in form of $n \times p$ matrix with n being the number of objects and p being the original number of variables, $p \geq n$, the ROBPCA method involves both the methods of projection pursuit and robust estimation.
It proceeds in a three-fold way: first, the data is standardized such that the new data belongs to a subspace.

Robust PCA

Without prior detection and even removal of outliers, the Robust PCA is a better option than the conventional PCA or ICA. If the data is represented in form of $n \times p$ matrix with n being the number of objects and p being the original number of variables, $p \geq n$, the ROBPCA method involves both the methods of projection pursuit and robust estimation.
It proceeds in a three-fold way: first, the data is standardized such that the new data belongs to a subspace.
In the next step, the "least outlying" data points are found out by using their covariance matrix to obtain a preliminary subspace.

Robust PCA

Without prior detection and even removal of outliers, the Robust PCA is a better option than the conventional PCA or ICA. If the data is represented in form of $n \times p$ matrix with n being the number of objects and p being the original number of variables, $p \geq n$, the ROBPCA method involves both the methods of projection pursuit and robust estimation.
It proceeds in a three-fold way: first, the data is standardized such that the new data belongs to a subspace.
In the next step, the "least outlying" data points are found out by using their covariance matrix to obtain a preliminary subspace.
In the final step, the scatter matrix of the data points is estimated using the MCD estimator.

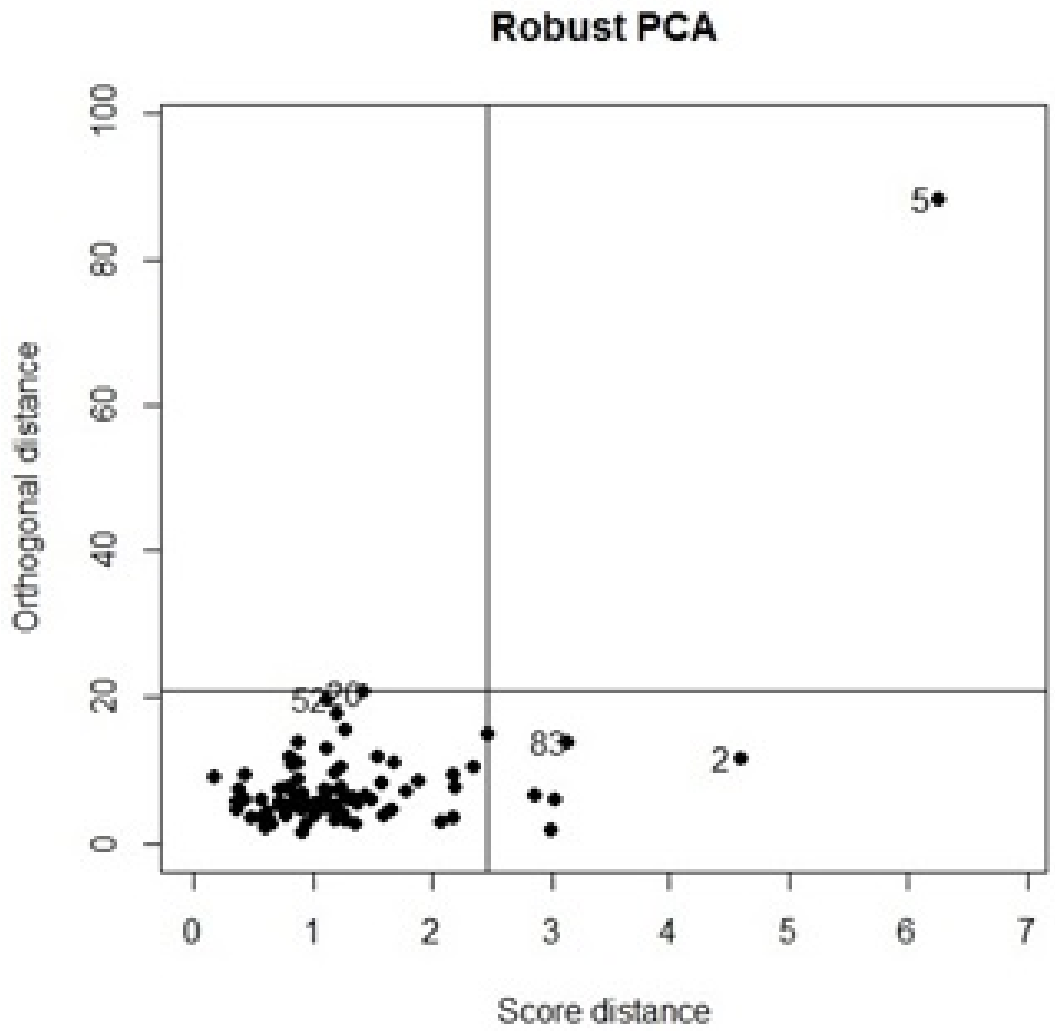
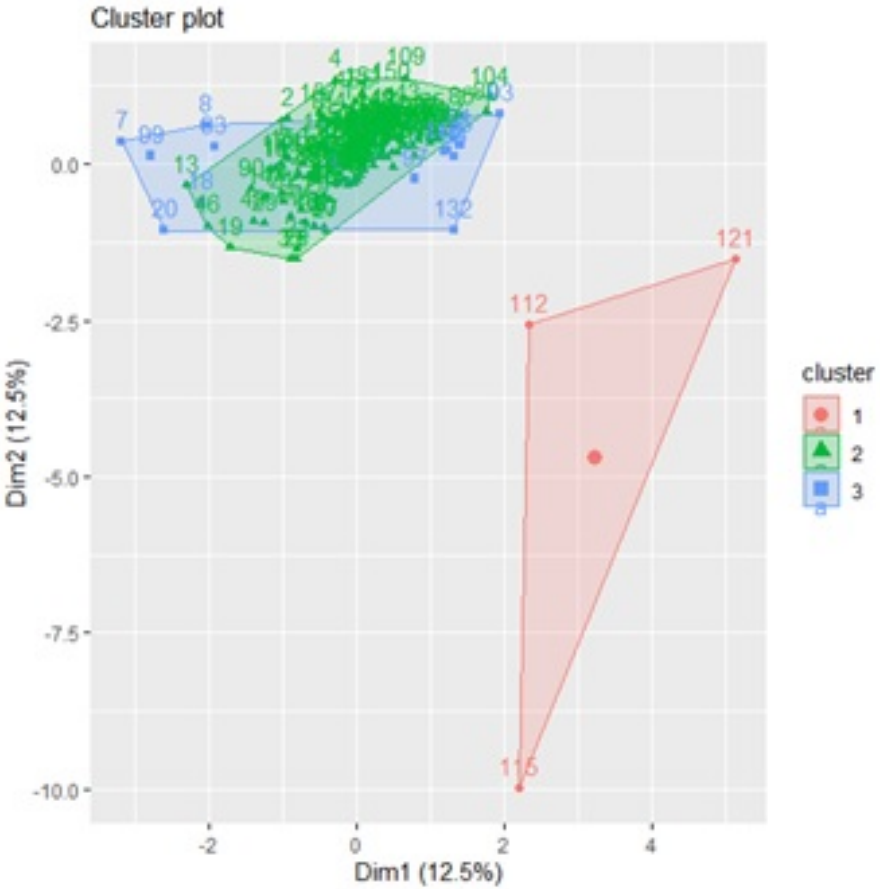
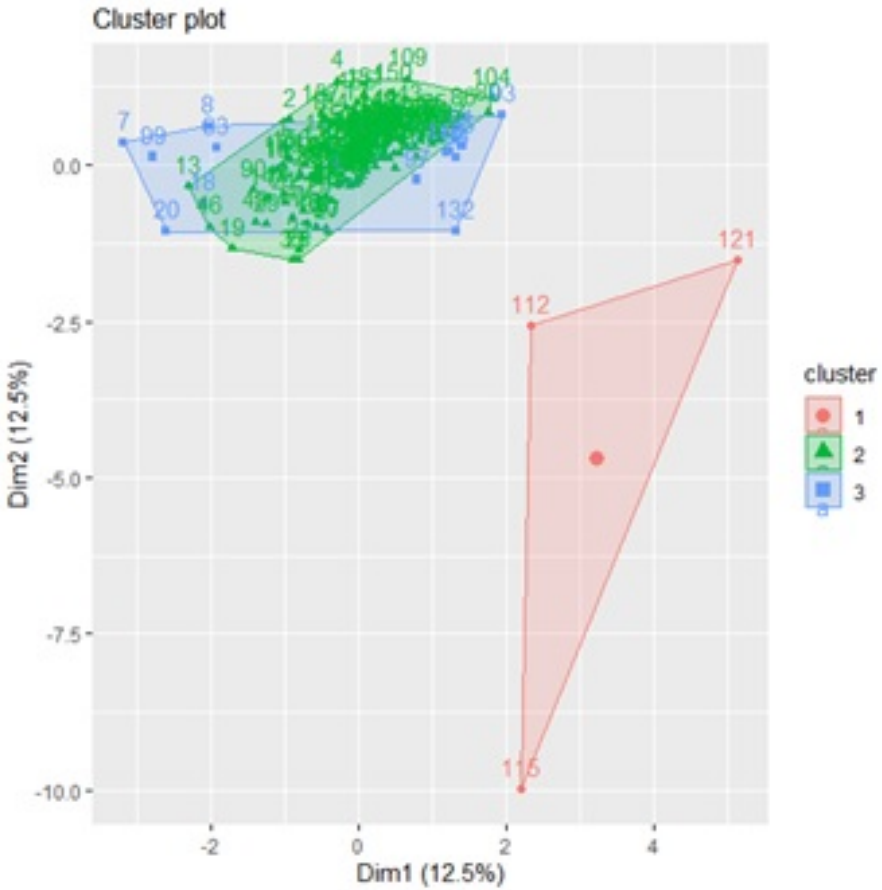
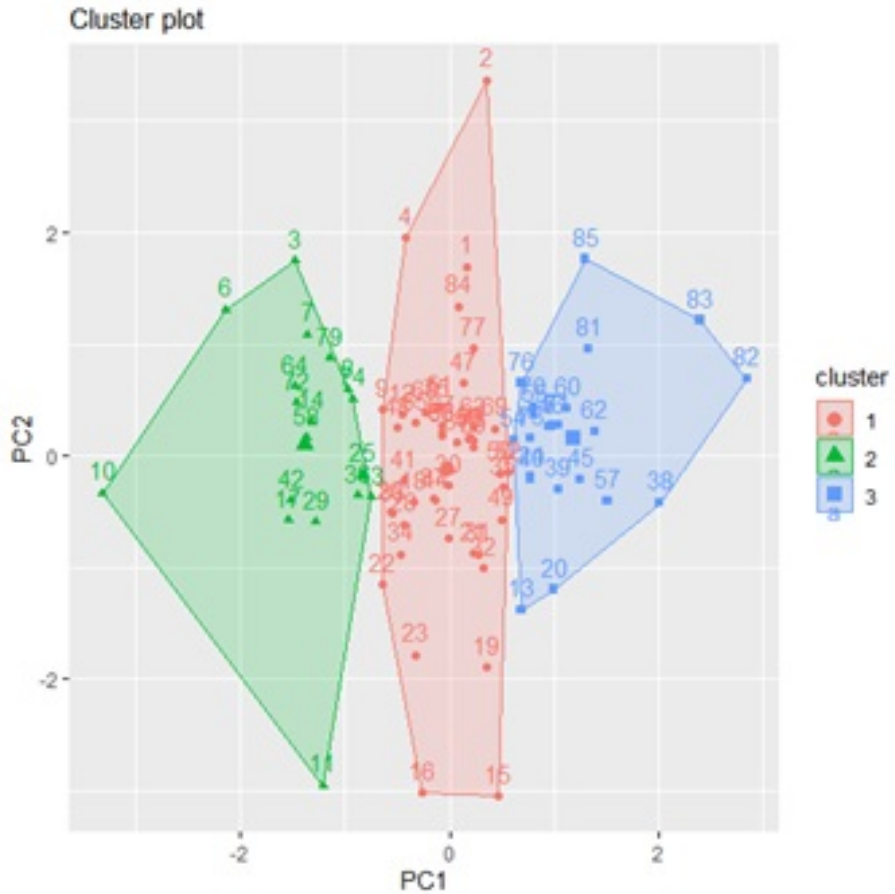


Figure: Diagonistic Plot of ROBPCA


Cluster plot
Dim2 (12.5%)
Dim1 (12.5%)
0.0
-2.5
-5.0
-7.5
-10.0
-2
0
2
4
121
112
115
132
cluster
1
2
3


Cluster plot
Dim2 (12.5%)
Dim1 (12.5%)
cluster
1
2
3


Cluster plot
PC2
PC1
cluster
1
2
3

VBSE

A more recent variation on principal component analysis is called sparse PCA. In particular, sparse weight vectors, or loading that is, weight vectors with a small number of "active" (nonzero) values—are what the SPCA looks for. Because the principal components of this approach are produced as a linear combination of only a few of the original variables, the model's interpret-ability is improved. In a high-dimensional data context, when the number of variables p exceeds the number of observations n, SPCA also prevents over fitting. Overfitting is a ML syndrome when a model is trained to fit a training set too closely loosing its ability to accurately predict a new one.

Table of Contents

- Multivariate Kolmogorov-Smirnov Test

- Multivariate Kolmogorov-Smirnov Test
- Multivariate Anderson-Darling Test

- Multivariate Kolmogorov-Smirnov Test
- Multivariate Anderson-Darling Test
- Multivariate Cramer–von Mises Test

- Multivariate Kolmogorov-Smirnov Test
- Multivariate Anderson-Darling Test
- Multivariate Cramer–von Mises Test
- Multivariate Shapiro-Wilk Test: is the goodness-of-fit test for assessing normality of data which works best on short tailed distributions.

- Multivariate Kolmogorov-Smirnov Test
- Multivariate Anderson-Darling Test
- Multivariate Cramer–von Mises Test
- Multivariate Shapiro-Wilk Test: is the goodness-of-fit test for assessing normality of data which works best on short tailed distributions.
- Multivariate Jarque–Bera Test: is a statistical test that determines if a sample's skewness and kurtosis match a normal distribution.

- Multivariate Kolmogorov-Smirnov Test
- Multivariate Anderson-Darling Test
- Multivariate Cramer–von Mises Test
- Multivariate Shapiro-Wilk Test: is the goodness-of-fit test for assessing normality of data which works best on short tailed distributions.
- Multivariate Jarque–Bera Test: is a statistical test that determines if a sample's skewness and kurtosis match a normal distribution.

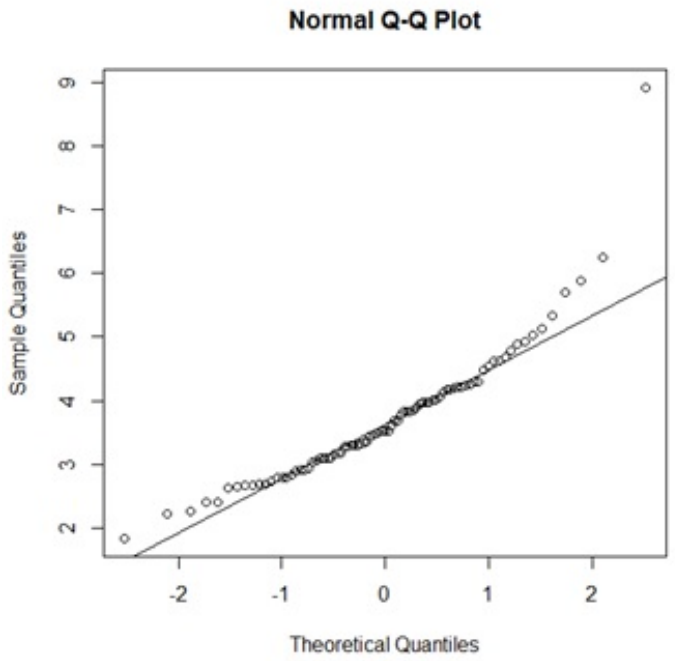


Table of Contents

Multivariate Outlier Detection Methods

- Mahalanobis Distance: based on Mahalanobis $D_i^2 = (X_i - \bar{X})' S_d^{-1} (X_i - \bar{X})$.

Multivariate Outlier Detection Methods

- Mahalanobis Distance: based on Mahalanobis $D_i^2 = (X_i - \bar{X})' S_d^{-1} (X_i - \bar{X})$.
- Jackknife Distance: based on the modified D_i^2 as $J_i^2 = \frac{(n-2)n^2}{(n-1)^3} \times \frac{D_i^2}{1 - \frac{nD_i^2}{(n-1)^2}}$.

Multivariate Outlier Detection Methods

- Mahalanobis Distance: based on Mahalanobis $D_i^2 = (X_i - \bar{X})' S_d^{-1} (X_i - \bar{X})$.
- Jackknife Distance: based on the modified D_i^2 as $J_i^2 = \frac{(n-2)n^2}{(n-1)^3} \times \frac{D_i^2}{1 - \frac{nD_i^2}{(n-1)^2}}$.
- KNN: focuses on measuring the distance of an observation from its neighbors that are most similar. When looking for outliers, K stands for the order of the closest neighbor.

Multivariate Outlier Detection Methods

- Mahalanobis Distance: based on Mahalanobis $D_i^2 = (X_i - \bar{X})' S_d^{-1} (X_i - \bar{X})$.
- Jackknife Distance: based on the modified D_i^2 as $J_i^2 = \frac{(n-2)n^2}{(n-1)^3} \times \frac{D_i^2}{1 - \frac{nD_i^2}{(n-1)^2}}$.
- KNN: focuses on measuring the distance of an observation from its neighbors that are most similar. When looking for outliers, K stands for the order of the closest neighbor.
- Isolation Forest:operates by randomly selecting a feature and a random split point for that feature. It does this recursively until it isolates an outlier or reaches a specified depth in the tree. The algorithm builds multiple isolation trees, also known as iTrees.

Multivariate Outlier Detection Methods

- Mahalanobis Distance: based on Mahalanobis $D_i^2 = (X_i - \bar{X})' S_d^{-1} (X_i - \bar{X})$.
- Jackknife Distance: based on the modified D_i^2 as $J_i^2 = \frac{(n-2)n^2}{(n-1)^3} \times \frac{D_i^2}{1 - \frac{nD_i^2}{(n-1)^2}}$.
- KNN: focuses on measuring the distance of an observation from its neighbors that are most similar. When looking for outliers, K stands for the order of the closest neighbor.
- Isolation Forest:operates by randomly selecting a feature and a random split point for that feature. It does this recursively until it isolates an outlier or reaches a specified depth in the tree. The algorithm builds multiple isolation trees, also known as iTrees.
- Hotelling T^2 Test: based on the assumption that the data follows multivariate Gaussian distribution.

Visualizing outliers

- Scatter plot

Visualizing outliers

- Scatter plot
- Density Plot

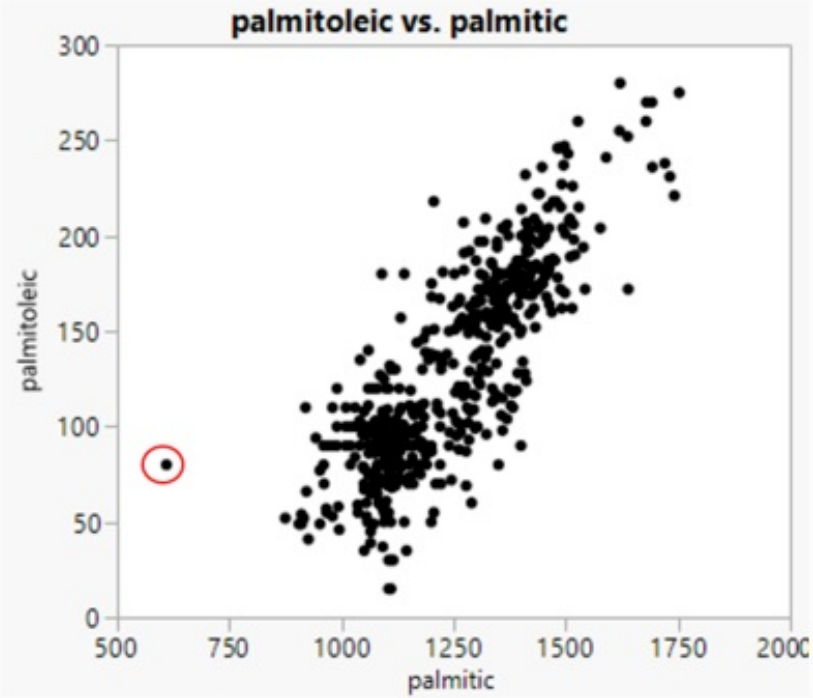
Visualizing outliers

- Scatter plot
- Density Plot
- Box-Plot

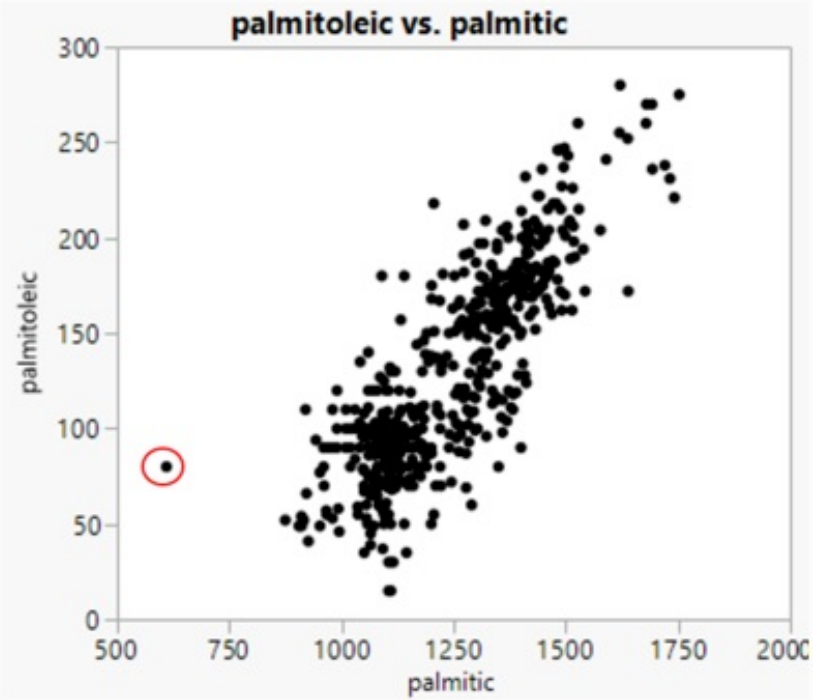
Visualizing outliers

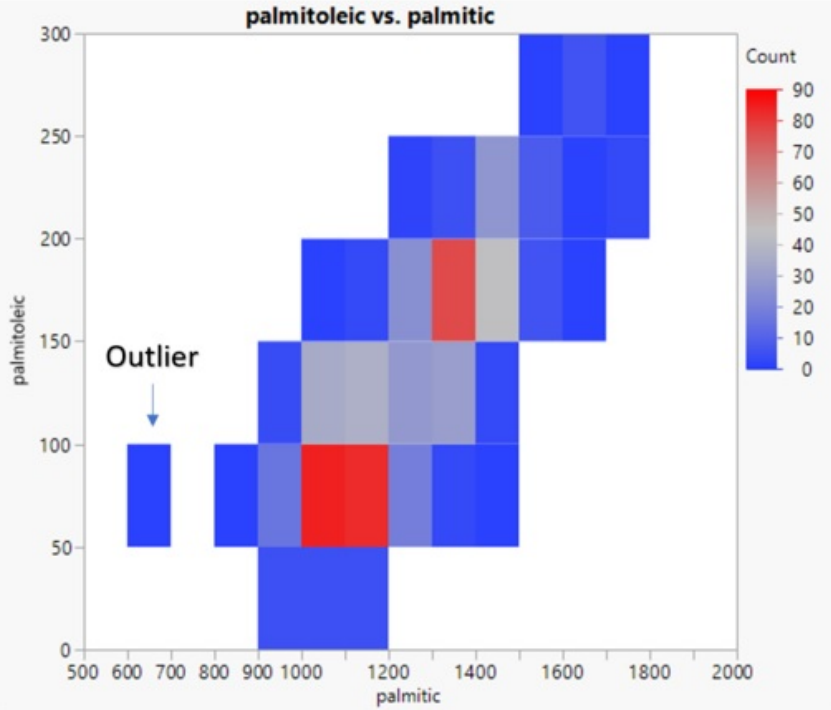
- Scatter plot
- Density Plot
- Box-Plot
- χ^2 Plot: a plot of the ranked ordered values of the Generalized or Mahalanobis distance vs values of the χ^2 statistic. This procedure is based on the covariance matrix of the data. The resulting diagram would be interpreted much as a univariate probability plot where the presence of more than one straight-line segment is taken as evidence of multiple populations, and outliers as individuals or small groups are separated from the remaining data by gaps on the plot.

Scatter Plot and Density Plot

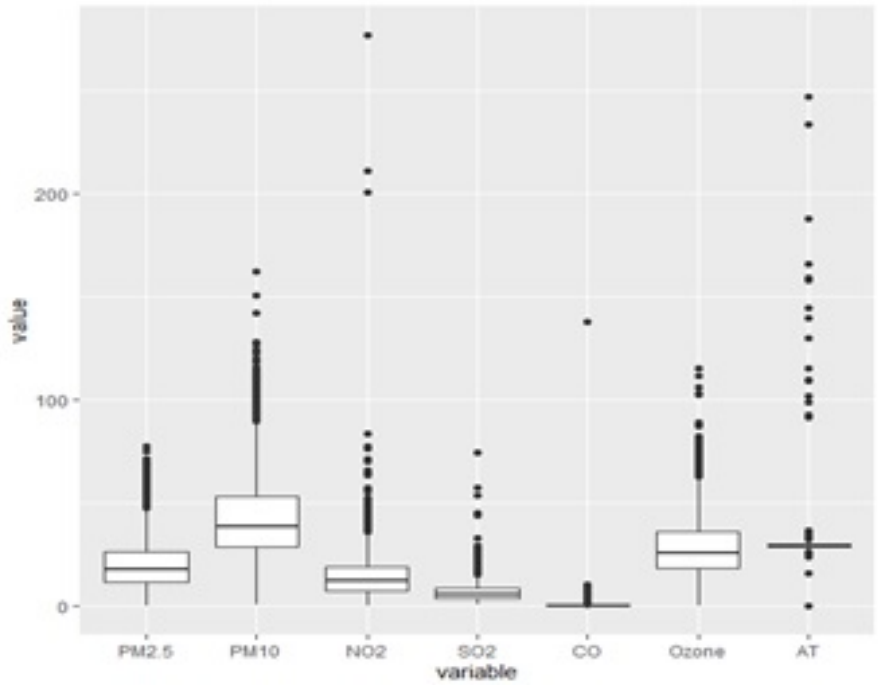


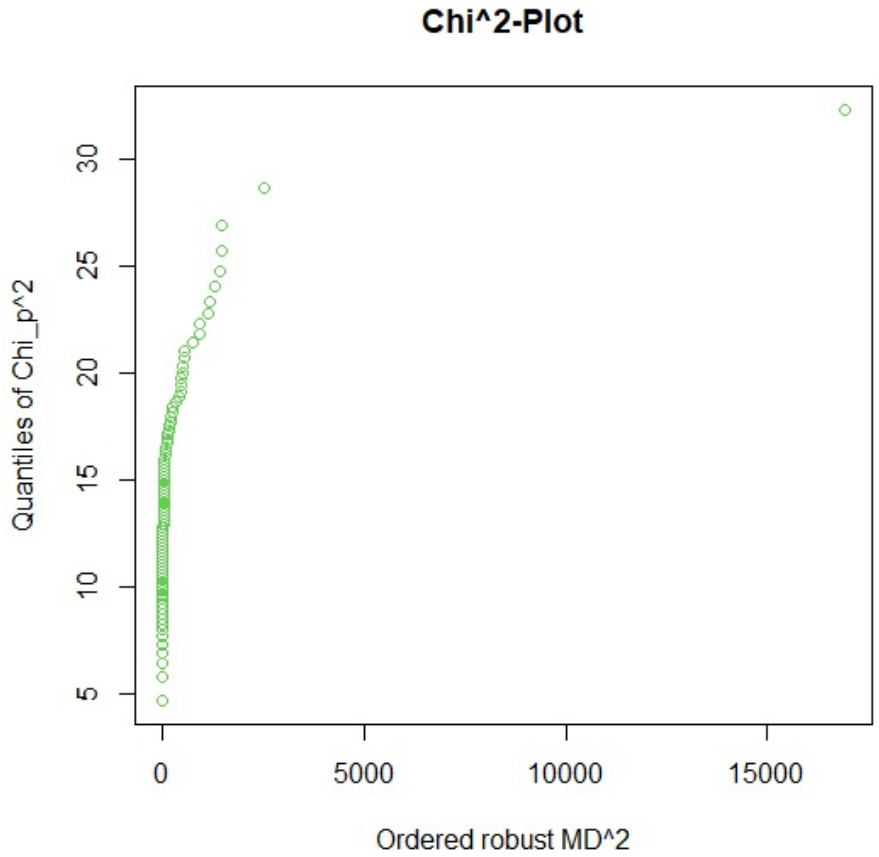
Scatter Plot and Density Plot



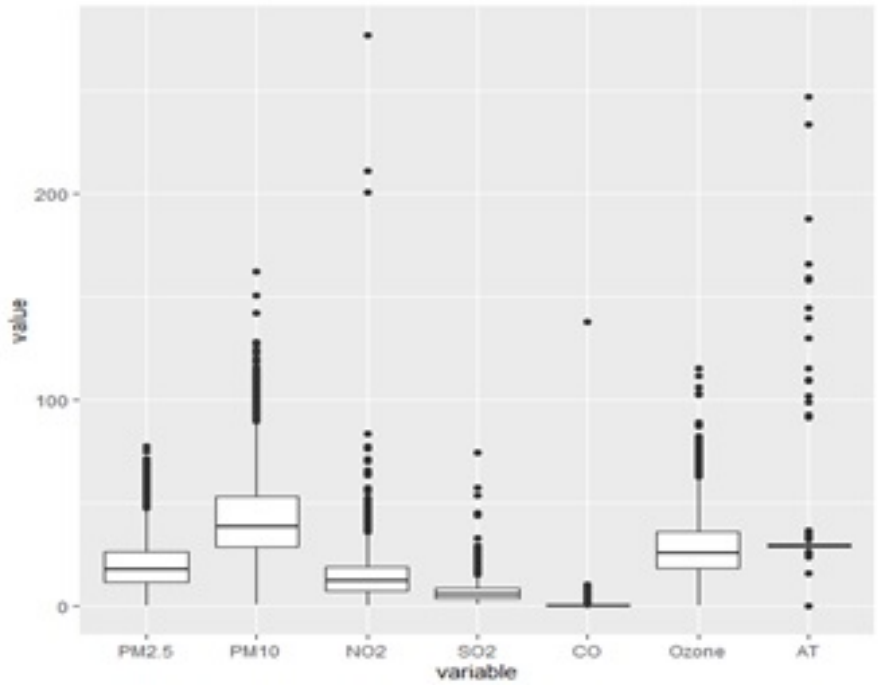


Box-plot and χ^2 Plot





Box-plot and χ^2 Plot



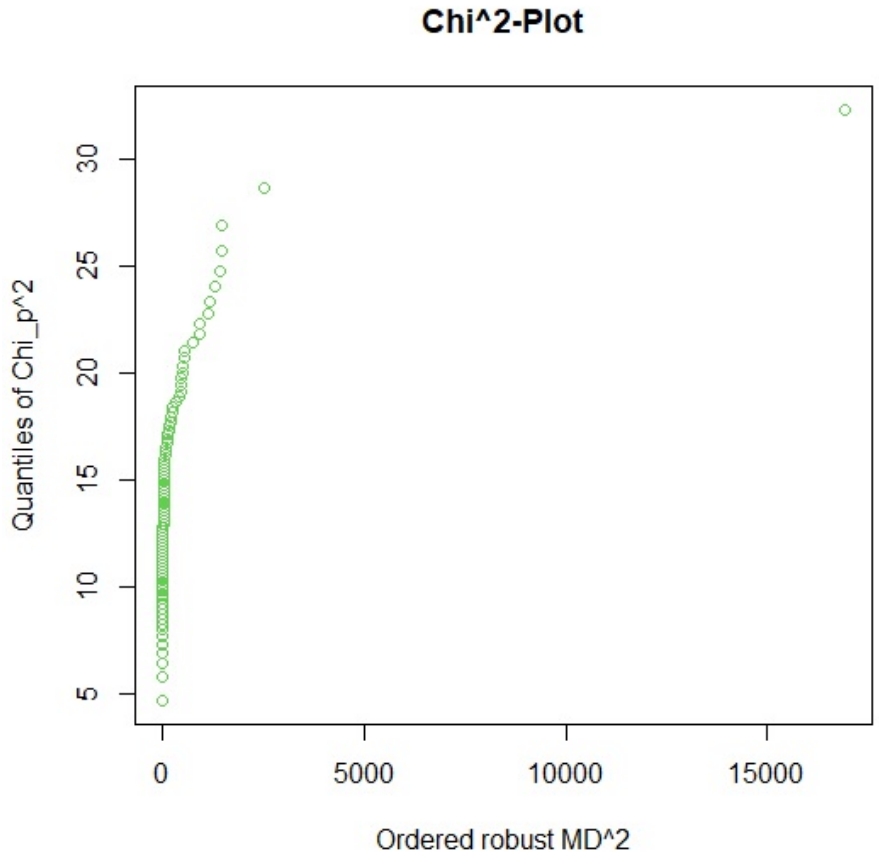


Table of Contents

Fat tailed distributions are characterized by larger values of skewness or kurtosis relative to a Gaussian or Exponential-family distributions. Some common fat-tailed distributions include:

- Cauchy
- Pareto
- Levy

All of these characterize the Non-Gaussian setup. While assuming "Gaussian" or "Gaussian-mixture" models/noise the tail of them must be investigated else once can expect more hallucinating results.

"The best thing about being a Statistician is that you get to play in everyone's backyard".
- Prof. John W. Tukey

Thank You!!!