Feature Selection in High Dimensional spaces for Drug Discovery(Group 31)

Jatin Pahuja (m24mac002), m24mac002@iitj.ac.in
Prachi Sahu (m24mac005), m24mac005@iitj.ac.in
Sidharth (m24mac013), m24mac013@iitj.ac.in

November 2024

1 Abstract

Analysis of high dimensional data is of great concern to the knowledge developed by drug discovery and development. Many features in such datasets create difficulties in pattern finding; thus, this project applies three major dimensionality reduction techniques - PCA, LDA, and ICA-on the massive dataset of a drug discovery application with a view to capturing the essence of variance with fewer dimensions.

Data on ChEMBL is utilized by the project. The type of data set is categorical as well as numerical type. Following preprocessing techniques, where missing values are handled, the categorical features are encoded, and data is scaled, PCA, LDA, and ICA are applied consecutively in the above order. PCA is used to reduce dimensions by retaining the directions with maximum variance, providing plots of cumulative explained variance in order to determine the number of components required to achieve related thresholds of 90% and 95% explained variance. Finally, LDA is employed to maximize between-class variance; the 'Type' column is used as the class label, with a cumulative variance plot in order to select how many components are required to meet related thresholds. ICA is next used to recover statistically independent components of the pre-scaled data and combines its output with visualized component variances and distributions for a richer understanding of the feature independence structure.

Regarding variance retention and the effectiveness of the techniques in reducing dimensionality, it can be seen that each of the techniques applied here differs: PCA orthogonal projections, LDA's supervised class-separating methodology, and ICA approach. A plot of the cumulative variance along with an analysis of the component distributions makes the project provide a comprehensive comparison of PCA, LDA, and ICA in reducing data complexity for drug discovery. The analysis allows selection of the best technique by properties of dataset

and provides insight into the procedure for the researchers by simplifying data without needing immense information loss.

2 Introduction

High-dimensional data forms a backbone in drug discovery: it carries comprehensive molecular and bioactivity information that can potentially help in identifying drug candidates. Simultaneously, the management of these types of datasets involves inherent difficulties: abundant feature dimensions introduce redundancy, prolong computation time, and complexity obscures underlying patterns. Dimensionality reduction is a powerful tool that handles the challenges by reducing the number of variables, simplifying analysis, and preserving as much critical information as possible.

The three divergent dimension reduction methods at inception-PCA, LDA, and ICA-will be applied to a large dataset from ChEMBL to further dimension reduce the same. The database contains detailed information regarding bioactive molecules and their biological targets, which makes ChEMBL a generally accepted repository in drug discovery. Proper feature reduction is therefore needed on this sizeable dataset for maximum analytical efficiency without such loss of integrity.

The preprocessing steps of getting data ready for dimensionality reduction involve addressing missing values, encoding categorical variables, and normalization to make the values of different features comparable with each other. After these processes, PCA, LDA, and ICA are separately applied on preprocessed data and compared to see if one of them is better suited for the task of getting fewer dimensions without losing meaning in the information. PCA is used to identify components that contribute to the maximum variance; LDA maximizes separability between classes that have been predefined using the ”Type” column as the class label and ICA identifies statistically independent components.

The work done in this project is that of cumulative explained variance, comparing which of these methods allow the greatest simplification of data that are high dimensional within the context of drug discovery. Its result can be used to decide which technique best holds up under ideal balance between simplicity and retaining information, thereby illuminating that would prove useful for overcoming obstacles in drug discovery efforts within data-driven realms and could perhaps prove helpful in drug candidate identification and selection.

3 Project Objective

This project aims to apply and compare three dimensionality reduction techniques—Principal Component Analysis (PCA), Linear Discriminant Analysis (LDA), and Independent Component Analysis (ICA)—on high-dimensional drug discovery data. The focus is on reducing data dimensions while preserving essential variance and class separability. Key objectives include:

1. Reducing data dimensions using PCA, LDA, and ICA.
2. Visualizing the explained variance and component distribution.
3. Evaluating and comparing these techniques to determine which best preserves the data's original structure, targeting 90% and 95% variance retention.

4 Theory

Undoubtedly, data reduction through the mechanism of PCA, LDA, and ICA is one of the important simplifications used primarily by researchers in their data, most importantly in drug discovery. Inter-dependence between these methods is that they improve the preservation of information but at the same time remove redundant features, thus making it much more efficient and interpretable in the analysis of data.

4.1 Principal Component Analysis (PCA)

PCA is an unsupervised technique for maximizing the amount of variance captured by each of the new axes, or principal components, in a dataset. In turn, PCA captures the directions of maximum variance while reducing dimensions by transforming the data into a new set of orthogonal axes. This is very useful when working with datasets whose feature spaces are high-dimensional, with many features not contributing much toward variance.

4.1.1 Mathematical Intuition

PCA computes the covariance matrix of standardised data to compute how features vary together. The eigenvectors of a covariance matrix are its eigenvalues that signify variance and directions for each principal component. Mathematically:

- **Centering and Standardization:** Center the data to have a mean of zero and standardize to unit variance.
- **Covariance Matrix:** The covariance matrix $\Sigma = \frac{1}{n}X^T X$ is calculated; this captures the relationship between any two features.
- **Eigenvalue Decomposition:** It decomposes the covariance matrix to capture eigenvalues or variances by components and eigenvectors, corresponding to the principal directions.

$$\Sigma v = \lambda v$$

- **Selection of Principal Components:** The eigenvectors are sorted on the descending order of their corresponding eigenvalues and the top k components were selected based on cumulative explained variance.

$$Z = XW$$

Applying PCA to the scaled dataset for variance capture. Calculate the cumulative explained variance ratio to know what is minimum number of components retaining 90% and 95% of the total variance of data.

4.2 Linear Discriminant Analysis (LDA)

LDA is a supervised dimension reduction technique. LDA enhances maximization of the separation between previously defined classes, and the other PCA is an unsupervised technique trying to capture directions of maximum variance. In other words, LDA tries to find such a feature subspace in which class separability is maximized. That is why it is extremely useful and effective for all applications, like classification, where keeping differences between categories is essential .

4.2.1 Mathematical Intuition

LDA maximizes the ratio of between-class variance to within-class variance such that classes are as distinct from each other as possible. This means computing

- **1. Mean Vectors:** Compute the mean vector for each class.
- **2. Scatter Matrices:** Compute the within-class scatter matrix S_W (measuring variance within classes) and the between-class scatter matrixb S_B (measuring variance between class means).

$$S_W = \sum_{i=1}^{C} \sum_{x_j \in D_i} (x_j - \mu_i)(x_j - \mu_i)^T$$

$$S_B = \sum_{i=1}^{C} N_i(\mu_i - \mu)(\mu_i - \mu)^T$$

- Objecti
- **3. Optimization:** In order to discover the directions or discriminant vectors with the maximum class separation, we need to solve the eigenvalue problem for
$$S_W^{-1} S_B w = \lambda w$$

LDA uses the column 'Type' as class labels to introduce some components that maximize the separation between classes and calculates the explained variances ratio for each LDA component to verify how each discriminant vector can successfully distinguish the classes.

4.3 Independent Component Analysis (ICA)

Independent Component Analysis (ICA) transforms data into a set of statistically independent components. In contrast, whereas PCA maximizes variance or LDA maximizes class separation, ICA searches for a representation in which each component is statistically independent of the others. ICA is particularly valuable for hidden and non-Gaussian signals in complex datasets.

4.3.1 Mathematical Intuition

ICA assumes that observed data is a linear mixture of independent source signals. It decomposes the data into a product of a mixing matrix and an independent component matrix. The steps include:

- **1. Centering and Whitening:** Data is centered and whitened to remove correlations between features.
- **2. Iterative Optimization:** An algorithm like FastICA is used to maximize the statistical independence of the components, often measured by kurtosis or negentropy.

ICA has transformed the data into showing statistically independent sources. In a way, it computes for every variance of the independent components then cumulative variance analysis to establish information retained by each.

Every variant of the dimensionality reduction techniques has something unique to offer about:

- **LDA** is supervised and uses the knowledge of classes; therefore, it applies very well to tasks in which class differences are most crucial since it maximises separability between labelled groups.
- **ICA** focuses on statistical independence; therefore, it applies very well in applications in which one assumes that the underlying factors are independent, non-Gaussian sources.

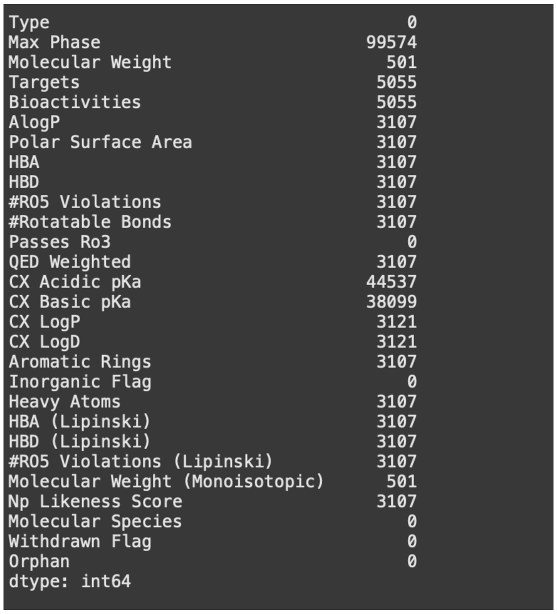
The code applies each method, visualizes the cumulative explained variance, and checks how well each technique reduces the dimensionality while preserving the most important information. It gives an idea about the overall comparison that may serve as a guide in choosing appropriate dimensionality reduction techniques in complex high-dimensional drug discovery data.

5 Methodology

It will be preceded by a few steps of data pre-processing and application of various techniques in dimensionality reduction, with corresponding evaluation. Detailed Methodology:

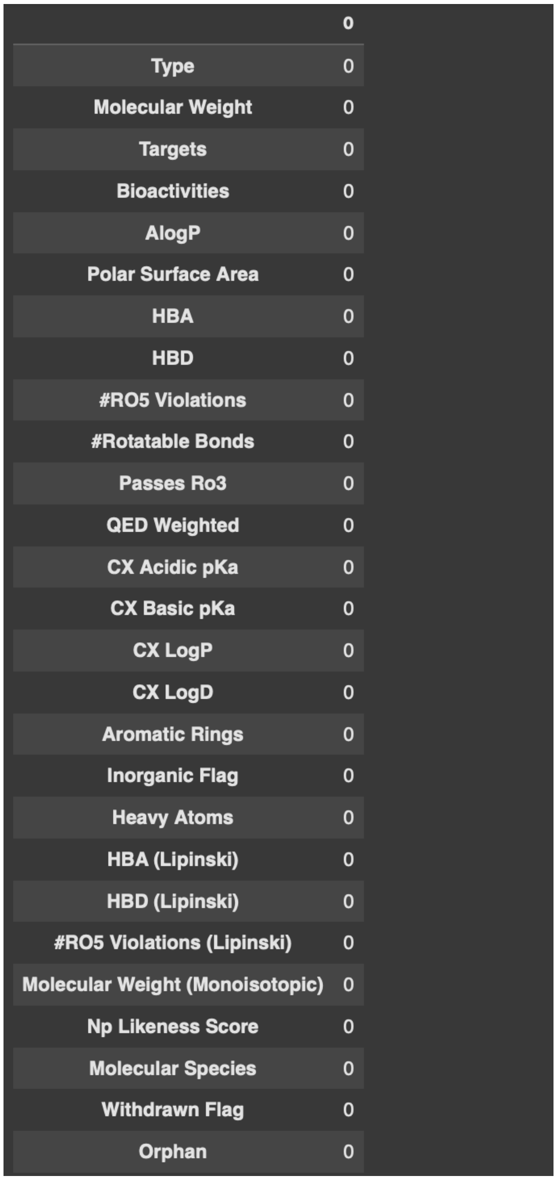
- **1. Loading and Preprocessing Data:** First load the dataset in memory from a CSV file and have a look at its structure: find any missing values, check the type of each column. Identify the columns as categorical and apply label encoding to have them represented numerically. The dataset is shuffled, all rows are ordered randomly, before subsampling for efficiency. The original data set have 20 lakh rows , whereas we have used sample of 1 lakh rows for efficient computation and analysis.

- **2. Handling Missing Values:** It replaces missing numerical values in columns with the mean value for that column. Mean imputation is utilized for both 'Targets' and 'Bioactivities', as those columns have missing values. 'Max Phase' column contains the maximum missing values, therefore, it must be excluded from the dataset not to bias the analysis.



```
Type                                 0
Max Phase                        99574
Molecular Weight                   501
Targets                           5055
Bioactivities                     5055
AlogP                             3107
Polar Surface Area                3107
HBA                               3107
HBD                               3107
#RO5 Violations                   3107
#Rotatable Bonds                  3107
Passes Ro3                           0
QED Weighted                      3107
CX Acidic pKa                    44537
CX Basic pKa                     38099
CX LogP                           3121
CX LogD                           3121
Aromatic Rings                    3107
Inorganic Flag                       0
Heavy Atoms                       3107
HBA (Lipinski)                    3107
HBD (Lipinski)                    3107
#RO5 Violations (Lipinski)        3107
Molecular Weight (Monoisotopic)    501
Np Likeness Score                 3107
Molecular Species                    0
Withdrawn Flag                       0
Orphan                               0
dtype: int64
```

Figure 1: Missing Values before Imputation



	0
Type	0
Molecular Weight	0
Targets	0
Bioactivities	0
AlogP	0
Polar Surface Area	0
HBA	0
HBD	0
#RO5 Violations	0
#Rotatable Bonds	0
Passes Ro3	0
QED Weighted	0
CX Acidic pKa	0
CX Basic pKa	0
CX LogP	0
CX LogD	0
Aromatic Rings	0
Inorganic Flag	0
Heavy Atoms	0
HBA (Lipinski)	0
HBD (Lipinski)	0
#RO5 Violations (Lipinski)	0
Molecular Weight (Monoisotopic)	0
Np Likeness Score	0
Molecular Species	0
Withdrawn Flag	0
Orphan	0

Figure 2: Missing Values after Imputation

- **3. Correlation Matrix:** To understand the relationship of the numerical features among themselves, a correlation matrix is calculated. Correlation matrix is represented as a heatmap to look at the linear correlation between features. Again, it identifies possible redundant or strongly correlated features that may affect the performance of dimension reduction techniques.

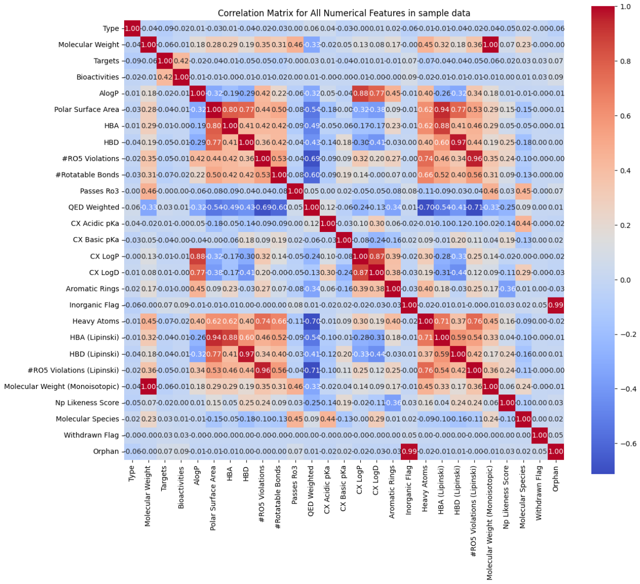


Figure 3: Visualization of correlation matrix after mean imputation

- **4. Feature Scaling:**Because PCA, LDA, and ICA are sensitive to the scale of features, all numerical features of the dataset are standardized using StandardScaler from scikit-learn. That way, the contribution of all features in the analysis is made equal.

- **5. PCA:** Applies PCA in order to reduce the dimensionality of dataset that could retain as much variance. Such number of principal components is specified based on a desired level of variance retention say 90% or 95%. Now, for each component, an explained variance ratio is computed; so, the total amount of variance explained when more and more numbers of components are taken can be plotted and visualized. The number of principal components can then be decided at the threshold where the cumulative variance is greater than 90% and 95%.

- **6. LDA(Liniear Discriminant Analysis):** LDA is applied in order to dimensionally reduce a dataset so that the inter-class separability is maximised. LDA differs from PCA, which considers variance as the only criterion for selection. In LDA, between-class variance is maximised and within-class variance is minimised. Similar to PCA, the ratio of explained variance is calculated for every component and plotted against the components. The number of components for selection is determined by the cut off value of the cumulative variance.

- **7. Independent Component Analysis.** ICA is carried out on the data to obtain the statistically independent components. It can be very helpful for a non-Gaussian data set with mixed sources. Variance for each component of ICA is computed, and the cumulative variance plot is obtained. Also, as in PCA and LDA, numbers of components that explain a certain percentage of the variance is determined.

- **8. Evaluation and Comparison:** Compare the results of the three techniques of dimensionality reduction in a graph with the cumulative explained variance for PCA, LDA, and ICA. Histograms plot the ICA components for visual assessment of the distribution of independent components. Generate a plot showing the combined results: the cumulative explained variance for PCA and LDA together with the variance of components of ICA, so that it becomes easy to compare the techniques directly.

The project concludes by measuring how well one dimensionality reduction technique, say PCA or LDA or ICA, performs better in terms of variance retention and achieving the reduced dimension. With this in view, the ultimate goal of applying the techniques is to understand relationships that would prove to be in the data and thus acts as a basis for further analyzes such as clustering or machine learning model building in drug discovery. This methodology should ensure that the applied dimensionality reduction techniques are used effectively, hence the results evaluated properly with the best fit method to the given dataset.

6 Result

PCA:

- 13 features capture 90% of the data variance.
- Cumulative explained variance rises stepwise with every new feature.
- Good general feature selection, reducing the most variance.

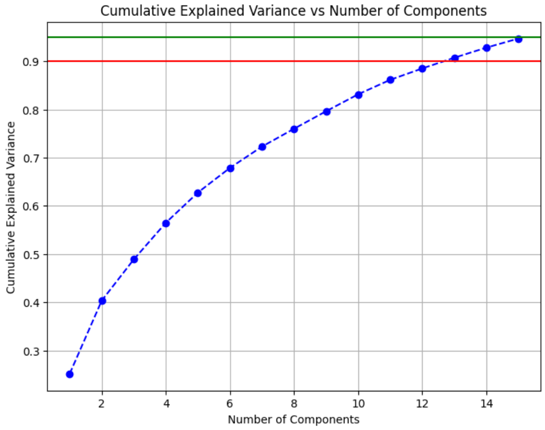


Figure 4: Cumulative explained variance Vs Number of Components for PCA

The cumulative explained variance rises nonlinearly toward approximating 90% after about 12-14 components.This curve can be interpreted as having majorly explained variance early, especially diminishing returns as further components come along.The threshold around the 90% red horizontal line is almost obtained close to 12 or even 13 components and clearly demonstrates that these would cover more or less all variance found in the data.

LDA:

- 1 feature captures 90% of the variance and produces a complete separation of classes.
- LDA maximizes class separability, so LDA is well-adapted to classification.

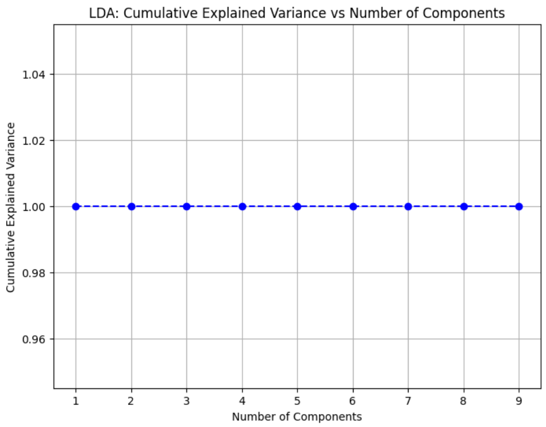


Figure 6: Cumulative explained variance Vs Number of Components for LDA

The cumulative explained variance for LDA is always 100% with every added component.This comes as a consequence of LDA maximizing class separability, not data variance. Since the number of classes minus one sets a limit on the effective components in LDA, the total variance is explained.Since LDA's design has optimality for separation rather than variance per se, all the variance is explained by the first component in this case.

ICA:

- 14 features capture 90% of the variance, and all 15 do 95%.
- Good for applications who need independent source separation even if it is more complicated than PCA or LDA.

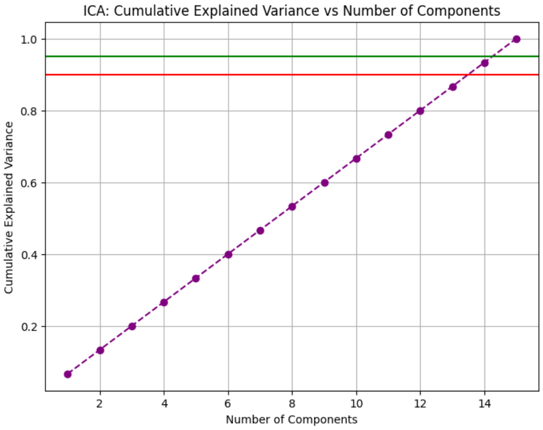


Figure 8: Cumulative explained variance Vs Number of Components for ICA

The graph shows increasing cumulative explained variance with increasing components, almost reaching 100% of the variance explained at the maximum number of components. This suggests that completely explaining data variance requires fully using ICA components.In general, ICA is not trying to maximize the explained variance but is instead focused on separating independent sources. As a result, the cumulative total of explained variance in ICA will not grow like a steep curve in the early runs, typical for PCA or LDA. In general, ICA is not trying to maximize the explained variance but is instead focused on separating independent sources. As a result, the cumulative total of explained variance in ICA will not grow like a steep curve in the early runs, typical for PCA or LDA.

7 Analysis

Dimension reduction techniques used–Independent Component Analysis (ICA), Linear Discriminant Analysis (LDA), and Principal Component Analysis (PCA)–each has its peculiar properties and strengths. This section analyzes in detail the performance of each technique with a conclusion on which works best for this dataset and the problem at hand.

Principal Component Analysis (PCA)

- PCA generates roughly a non-linear form of cumulative explained variance, reaching up to some 90% of the variance in 13 components.
- PCA puts most of the variance into its first few components but doesn't get a lot more beyond that as you go.
- Also, PCA is unsupervised, so it doesn't know anything about any labels for classes and only sees the variance in the data.
- **Best Use Case for PCA** PCA is one of the most effective techniques that could be used to perform general dimensionality reduction when the purpose is reducing the complexity with retaining as much of the variance as possible. It is more useful in unsupervised tasks when it helps in reducing the dimension while losing minimal information.
- **NOTE:** PCA is acceptable for unsupervised learning applications when the goal is to minimize dimensionality as much as possible without a significant information loss in order that the variance of the data is preserved. It is suitable for applications where there is an intrinsic structure within the data, which is captured by fewer components.

Linear Discriminant Analysis (LDA)

- The cumulative variance accounted for by the LDA model is 100% after the first component due to the fact that LDA focuses mainly on maximizing class separability rather than variance in general.
- LDA is designed to work on supervised learning tasks where one wishes to find the components which best separate classes.
- **Best use case :**LDA is very effective for supervised classification tasks, especially when class separability is to be maximized. However, LDA is limited by the number of classes because its maximum number of components is $C - 1$, where C is the number of classes, and it is also used only when there is labeled data. It doesn't apply in unsupervised learning as labels are missing for a class.
- **NOTE :**LDA is perfectly suitable when you're dealing with supervised learning in the direction of class separability. It's also good when you have labeled data and you are actually aiming to reduce dimensionality while trying to retain discriminatory features between classes.

Independent Component Analysis (ICA)

- Cumulative explained variance of ICA increases in a very slow manner so that it reaches nearly 100% explained variance at the maximum numbers of components.
- The purpose of ICA is not to maximize explained variance but to extract statistically independent components from the data.
- **Best use case :** ICA becomes useful when the purpose is to separate independent sources rather than just simply trying to capture the total amount of variance.
- **NOTE :** It is not a great choice if reducing the dimensionality of the data with a retention of as much variance as possible is the prime objective, as its variance explanation increases gradually.
- The ICA is appropriate when input variables come from sources that are independent, especially in signal processing or blind source separation, or where feature independence must be realized in the application.

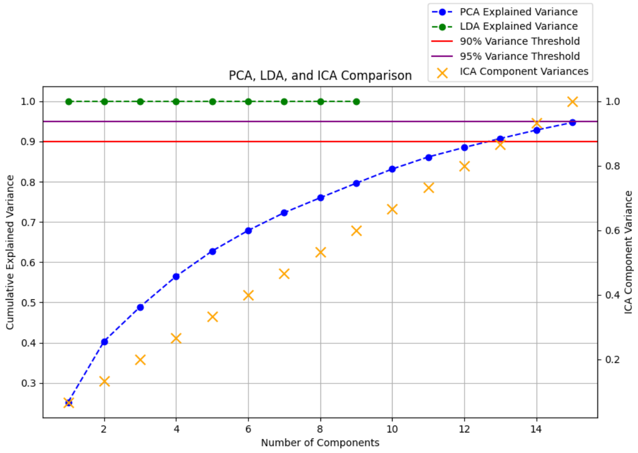


Figure 10: PCA ,LDA, ICA comparison

8 Conclusion

- **For Classification Applications (Supervised Learning):LDA:** The most effective, since it maximizes class separability and even achieves near-perfect class distinction with just one component. It best suits applications where class labels are available, and the improvement in the performance of the classification is mainly the goal.

- **For Unsupervised Dimensionality Reduction (Variance Preservation):PCA** is the best option since it captures most of the variance at the beginning and reduces the dimension wonderfully, but also preserves most of the information that originally existed in the data. The selection of 12-13 components is a very good balance between the reduction of dimensions and variance retention.
- **For Independent Features (Independent Sources or Signal Separation):** When used with independent sources or when signals have mixed, it can separate mixed signals: **ICA** is useful in such scenarios. However, the gradual increase of variance explanation makes it less suitable for standard tasks of dimensionality reduction or classification.

In summary, considering the objectives of this project, PCA is suitable for general reduction of dimensions, whereas LDA is better suited for classification when the data is labelled.

9 References

[1] . European Bioinformatics Institute. (n.d.) ChEMBL Database. Retrieved from https://www.ebi.ac.uk/chembl/g/#search$_{r}esults/all$

[2] . Gyamerah, S., Soori, G. T., Korda, D. R., Tawiah, J. K., Akolgo, E. A., & Dapaah, E. O. (2023). Comparative analysis of feature extraction of high dimensional data reduction using machine learning techniques. ResearchGate.

[3] . Jiang,S.,& Li,M.(2021).”Feature Selection in Drug Discovery Using Hybrid PCA-LDA Models”ResearchGate- Jiang & Li, 2021.

[4] . Cohen, M. X. (2022). Practical linear algebra for data science: From core concepts to applications using Python. O’Reilly Media.