Principal Component Analysis
Epoch IIT Hyderabad

Chakka Surya Saketh
AI22BTECH11005

Introduction

Principal component analysis is a technique for feature extraction — so it combines our input variables in a specific way, then we can drop the ”least important” variables while still retaining the most valuable parts of all of the variables! As an added benefit, each of the ”new” variables after PCA are all independent of one another. This is a benefit because the assumptions of a linear model require our independent variables to be independent of one another.

The impact of having more dimensions in the model, which is nothing but having multicollinearity in the data can lead to overfitting, and this exposes the model to have variance errors, that is the model may fail to perform or predict for new unseen data. PCA also helps to reduce this dependency or the redundancy between the independent dimensions.

Underlying math

Eigenvectors represent the new set of axes of the Principal component space and also the Eigenvalues carry the information of the amount of variance that each eigenvector has. So to scale back the dimensions of the dataset we are going to choose those Eigenvectors that have more variance and discard those with less variance.

Covariance and Variance are a measure of the “spread” of a set of points around their center of mass(mean).

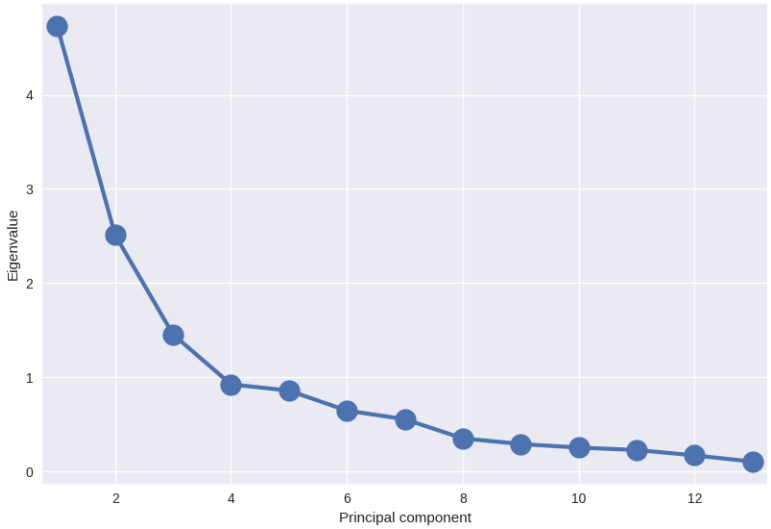
$$cov_{x,y} = \frac{\sum(x_i - x_{mean})(y_i - y_{mean})}{N-1}$$

where N is the number of samples of x. Covariance matrix represents the covariance between more than 2 dimensions as a matrix. This matrix is symmetric about the diagonal.

Workflow of the algorithm

1) Take the matrix of independent variables X and, for each column, subtract the mean of that column from each entry and divide by the variance of each column so that we standardise each column. Call the new matrix Z.
2) Take the transpose of the matrix Z to get Z^T. Multiply these matrices to get the covariance matrix of Z.
3) Eigen decomposition of ZZ^T gives PDP^{-1}, where P is a matrix of eigen vectors and D is a diagonal matrix with corresponding eigen values in the diagonal and zeroes everywhere else. The eigenvalues on the diagonal of D will be associated with the corresponding column in P — that is, the first element of D is λ_1 and the corresponding eigenvector is the first column of P.
4) Take the eigen values and sort them in decreasing order and also sort the matrix P accordingly. This sorted matrix is called P^*. Note that the eigen vectors are independent of one another.
5) Calculate $Z^* = ZP^*$, where the new matrix Z^* is a centred/standardised version of X.Each column of Z^* is also independent of each other.

Because each eigenvalue is roughly the importance of its corresponding eigenvector, the proportion of variance explained is the sum of the eigenvalues of the features you kept divided by the sum of the eigenvalues of all features. The proportion of variance explained by including only principal component 1 is $\frac{\lambda_1}{\lambda_1+\lambda_2+....\lambda_p}$ Once we've dropped the transformed variables we want to drop, we're done performing PCA.



Note on PCA

- PCA removes correlated features.
- It reduces the dimensionality and improves the visualisation of the data.
- It reduces overfitting.
- PCA is highly affected by outliers
- The features after PCA are note as readable as the original features.

Conclusion

In conclusion, Principal Component Analysis (PCA) offers a powerful approach for dimensionality reduction and data simplification. By uncovering essential patterns and reducing noise, PCA aids in feature selection, visualization, and data compression. While effective for many applications, careful consideration of information loss and assumptions is essential. PCA's ability to transform complex data into interpretable components continues to make it a valuable tool in various domains, enhancing data analysis and machine learning endeavours.