International Journal of Scientific Research in Science, Engineering and Technology

Print ISSN - 2395-1990
Online ISSN : 2394-4099

OPEN ACCESS

Available Online at : www.ijsrset.com
doi : https://doi.org/10.32628/IJSRSET

Rank Prediction Using Principal Component Analysis

Vikas Babu Gond

Master of Technology (Computer Science), Saraswati Higher Education and Technology College of Engineering, Babatpur, Varanasi, India

ARTICLEINFO

Article History:

Accepted: 05 April 2024
Published: 15 April 2024

Publication Issue :
Volume 11, Issue 2
March-April-2024
Page Number :
297-301

ABSTRACT

"Rank Prediction Using Feature Selection and Transformation Techniques" seeks to establish the rankings of various countries based on their national security and to rank players from the Indian Premier League (IPL). This is achieved through the implementation Principal Component Analysis (PCA).

Principal Component Analysis (PCA) is a classical multivariate data dimensionality reduction data preprocessing, compression and visualization method suitable for many applications in biology, social sciences and engineering.

A limitation of PCA is the lack of interpretation due to nonzero loading and inconsistencies in high dimensional elements. Principal component analysis (sparse PCA) mainly aims to solve the above problems of PCA. In the last few years, many studies have been prepared on the development and theoretical analysis of sparse principal component analysis. The aim of this article is to conduct a literature review on the latest developments in high dimensional sparse principal component analysis from the perspective of algorithms and statistical theory. We first provide an overview of PCA and sparse PCA. Secondly, different PCA algorithms are divided into several categories, the structures and methods in each category are explained in detail, and the sparse PCA package is given. Considering that the variance of PCA increases with the increase of the index value, the theoretical analysis of sparse PCA was analysed.

Keywords :- Principal Component Analysis, Transformation Techniques, Indian Premier League

I. INTRODUCTION

In recent decades, there has been a significant surge in data generation, emphasizing the growing importance of data and its inherent features. To effectively utilize this vast amount of data, feature extraction becomes crucial. Dimensionality reduction techniques are key methodologies employed to streamline data processing by simplifying the feature space, thus enhancing the utility of large datasets for analytical purposes.

Ranking plays a vital role in various real-world scenarios by enhancing the clarity and understanding of issues, . By ranking nations that are less developed, policymakers can prioritize resource allocation and development strategies to elevate these countries' status. By applying these techniques, we will rank countries based on statistical attributes like population and human resources and cricket batsmen based on metrics such as runs scored and batting averages. This methodology streamlines complex datasets and aids in making informed decisions in sports management and policy development. This approach aims to illustrate how dimensionality reduction can facilitate a more straightforward and practical analysis, providing strategic insights for improving performance and guiding developmental policies.

II. Principal Component Analysis

In Principal Component Analysis [9], Principal Component Analysis (PCA) is a method used to reduce the dimensions of a dataset with multiple variables, enhancing computational efficiency while preserving essential information. The key challenge in PCA is determining the optimal number of principal components, k, that adequately represents the data within a k-dimensional subspace (where k<d, with d being the dataset's original dimensionality). PCA involves calculating the eigenvectors of the data, which form the principal components, and compiling these into a projection matrix. Each eigenvector has a corresponding eigenvalue that indicates its magnitude or importance. When some eigenvalues are considerably larger than others, it suggests that reducing the dataset to a subspace that excludes the less significant eigenvectors could be beneficial. By focusing on the most meaningful dimensions, PCA simplifies the interpretation of complex data, facilitates its visualization, and maximizes the retention of crucial information. This technique is typically used to project data into a new coordinate system where the variation is captured in fewer dimensions, often using the first two principal components for visual clustering in two-dimensional space.

PCA finds utility across various scientific domains, including population genetics, microbiome research, and atmospheric studies, due to its ability to efficiently handle high-dimensional data and aid in the identification of underlying data structures.

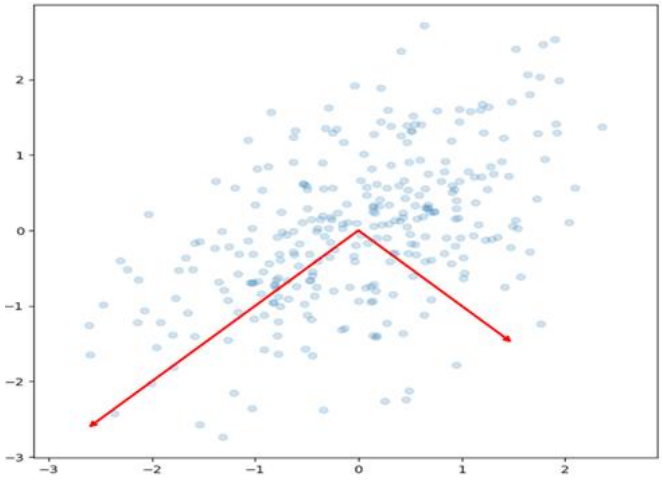

Fig-a PCA of a multivariate Gaussian distribution

Principal Component Analysis (PCA) is applied to a multivariate Gaussian distribution centered at (1,3), with standard deviations of 3 along approximately the (0.866, 0.5) direction and 1 along its orthogonal direction. The vectors are the covariance matrix's eigenvectors, each scaled by the square root of its corresponding eigenvalue and adjusted so their tails originate from the mean.

The principal components of a data set in a real coordinate space form a sequence of p unit vectors, where each vector is the direction of a line that optimally fits the data while being orthogonal to all previous vectors in the sequence. This optimal fit minimizes the mean squared distance perpendicular to the line from the data points. These vectors establish an orthonormal basis, rendering the dimensions of the data linearly uncorrelated.

PCA involves calculating and using these principal components to transform the data basis, often retaining only the most significant components and disregarding

the others. In data analysis, especially where variables are interrelated, PCA simplifies these variables into a smaller set of independent dimensions. It is especially beneficial for reducing the complexity of data by projecting it onto a lower-dimensional space using only the foremost principal components, thus conserving as much variability as possible.

The first principal component is defined as the direction that maximizes the variance of the projected data. Subsequent components are selected orthogonal to the preceding ones and maximize the remaining variance. This process is repeated until all dimensions of variability are accounted for. Utilized widely in exploratory data analysis and predictive modeling, PCA helps in compressing data, enhancing interpretability without significant loss of information.

III. PCA VS LDA

Linear Discriminant Analysis (LDA) and Principal Component Analysis (PCA) are techniques for linear data transformation, each targeting specific data attributes. PCA identifies principal components to maximize data variance, discovering the most expansive directions in the dataset. This attribute of PCA is advantageous for dimensionality reduction, as it projects data onto a new feature space aligned with these principal directions. Conversely, unlike PCA, LDA is tailored to enhance the distinction between multiple classes by incorporating class labels into its calculations. LDA focuses on finding directions that optimally separate classes, which is particularly beneficial for pattern classification tasks, as it projects data into a feature space designed to make classes distinctly separable. Thus, while PCA reorients data to a new coordinate system prioritizing variance irrespective of class structure, LDA modifies the data framework to optimize class differentiation, a key aspect in supervised learning environments

Following are steps of PCA :

Normalize the dataset.

- Compute the Eigenvectors and Eigenvalues from the co-variance or correlation matrix.
- Arrange the eigenvalues in descending order and select the k eigenvectors that correspond to the k largest eigenvalues, where k is the number of dimensions of the new feature space ($k \leqslant d$).
- Form the projection matrix W using the chosen k eigenvectors.
- Apply the projection matrix W to the original dataset X to derive a k-dimensional feature space Y.

Mathematical Background on Principal Component Analysis

PCA aims to reduce the Principal Component Analysis (PCA) is designed to reduce the dimensionality of datasets by transforming observed variables into a smaller number of principal components. These components, also referred to as auxiliary variables, are optimized linear combinations of the original variables that capture most of the variability in the data.

Specifically, consider a random vector X= (X1, X2, ..., Xp) T with p random variables and a covariance matrix Σ Σ. This matrix has eigenvalue-eigenvector pairs (λ 1, e1), (λ 2, e2), ..., (λ p, ep) ordered such that $\lambda 1 \geqslant \lambda 2 \geqslant ... \geqslant \lambda$ p. Each principal component Li is defined as Li = eiTX=ei1X1+ei2X2+...+eipXp for i=1, 2, ..., p, where ei represents the components of eigenvector e.

The principal components are essentially linear combinations of the original variables that maximize variance while remaining uncorrelated with each other. The first principal component, for instance, has the highest variance Var(L1)=λ1, which is greater than the variance of any other linear combination of these variables. The orthogonality of the components (Property ii: Cov(Li,Lj) =0 for I $\neq$ j) ensures that there

is no redundancy in the information captured by different components.

The total variance explained by the PCA is equal to the sum of the eigenvalues (λ1+λ2+...+λp), and the contribution of each component to this total provides a measure of its importance in capturing the underlying data structure. If the first few components account for a significant portion of the variance, they can effectively represent the original dataset with minimal loss of information. To ensure fair representation across variables with different scales, it is standard practice to normalize the data before applying PCA. This prevents components from being overly influenced by variables with larger scales and focuses the analysis on capturing the most significant patterns across all variables.

Data Set Description

The various Terminology, which we have used in this Dissertation are given below to help in understanding the problem in more accurate way.

• Country Dataset Description

Active Reserve Personnel: This category refers to a military organization consisting of civilians who engage in military training and can be mobilized for full-scale warfare or national defense as needed. Typically, these individuals maintain civilian careers and are not active soldiers unless called upon.

Corvettes : Smaller than frigates, corvettes are agile warships typically used for coastal defense and patrol missions.

Reserves of Foreign-Exchange and Gold: This includes the holdings of a country's central bank in foreign currencies and gold, providing a buffer and support for national economic stability.

IPL Dataset Description

Runs: Indicates the total runs a player scored during the IPL 2016 season. A higher run total typically reflects better performance.

Batting Average (Ave): The statistic represents the average number of runs a batsman scores per dismissal. It provides a measure of consistency and skill, with higher averages indicating superior performance. However, this metric can be misleading if a batsman remains not out frequently.

Batting Strike Rate (SR): The Calculated as the number of runs scored per 100 balls faced, this rate measures a batsman's scoring efficiency. Essential in the Twenty20 format, a higher strike rate generally suggests a more aggressive and impactful batting style. Twenty20: However, a high strike rate accompanying a low batting average is not desirable.

IV. RESULTS

The cumulative graphs generated by principal component analysis is shown below. Here the first principal component is contributing around 66 per-cent(figure b) in IPL dataset and around 55 percent(figure c) in country dataset.

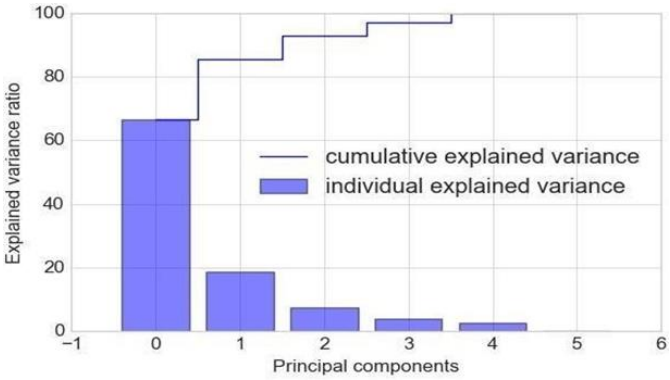

Figure b : Cumulative graph of IPL data set using PCA

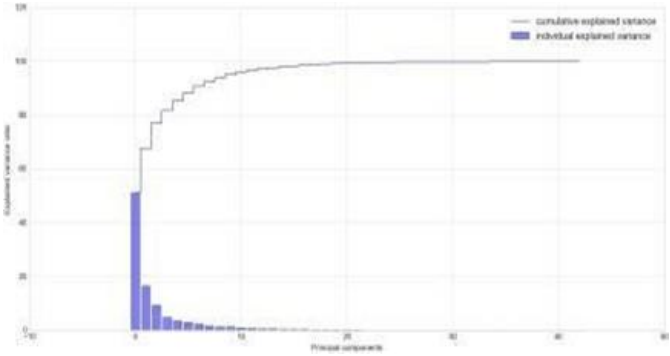

Figure c: Cumulative graph of Country data set using PCA

V. CONCLUSION

The comparison between the ranking of countries by experts on GlobalFirepower [13] and those derived from our algorithm demonstrates significant alignment: eight out of the top ten countries match, although their specific rankings differ. A similar outcome is observed with cricket data, where the rankings generated through our methods align with those listed on Cricbuzz [2], with all top ten entries matching but some positions varying. These findings suggest that our algorithm is highly effective for accurately identifying top-ranked entities, and in many cases, their precise rankings as well.

VI. REFERENCES

[1]. S. Ning-min and L. Jing, “A Literature Survey on High-Dimensional Sparse Principal Component Analysis,” International Journal of Database Theory and Application, vol. 8, no. 6, pp. 57–74, Dec. 2015, doi: 10.14257/ijdta.2015.8.6.06.

[2]. Shang, H.L. A survey of functional principal component analysis. AStA Adv Stat Anal 98, 121–142 (2014). https://doi.org/10.1007/s10182-013-0213-1

[3]. G. O. Young, “Synthetic structure of industrial plastics (Book style with paper title and editor),” in Plastics, 2nd ed. vol. 3, J. Peters, Ed. New York: McGraw-Hill, 1964, pp. 15–64.

[4]. W. K. Chen, Linear Networks and Systems (Book style). Belmont, CA: Wadsworth, 1993, pp. 123–135.

[5]. Parul, Ms., M.Jain and Prof. V.K.Shandliya. “A survey paper on comparative study between Principal Component Analysis (PCA) and Exploratory Factor Analysis (EFA).” International Journal of Managment, IT and Engineering 3 (2013): 415-424.

[6]. Chao Y-S, Wu H-C, Wu C-J and Chen W-C (2018) Principal Component Approximation and Interpretation in Health Survey and Biobank Data. Front. Digit. Humanit. 5:11. doi: 10.3389/fdigh.2018.00011

[7]. Jolliffe Ian T. and Cadima Jorge 2016Principal component analysis: a review and recent developmentsPhil. Trans. R. Soc. A.37420150202 http://doi.org/10.1098/rsta.2015.0202

[8]. S. Sehgal, H. Singh, M. Agarwal, V. Bhasker and Shantanu, "Data analysis using principal component analysis," 2014 International Conference on Medical Imaging, m-Health and Emerging Communication Systems (MedCom), Greater Noida, India, 2014, pp. 45-48, doi: 10.1109/MedCom.2014.7005973.

[9]. Saha, Sumanta, and Sharmistha Bhattacharya. "A Survey: Principal Component Analysis (PCA)." International Jounal of Advance Research in Science and Engineering 6.6 (2017): 1-9.

[10].Alim Samat, Paolo Gamba,Fellow, Sicong Liu,, “Jointly Informative and Manifold Structure Representative Sampling Based Active Learning for Remote Sensing Image Classification,” IEEE Journals, 2017.

[11].Svante Wold, Kim Esbensen, Paul Geladi,Principal component analysis,Chemometrics and Intelligent Laboratory Systems,Volume 2, Issues 1–3,1987,Pages 37-52, ISSN 0169-7439, https://doi.org/10.1016/0169-7439(87)80084-9.