Vital Annex: International Journal of Novel Research in Advanced Sciences
ISSN: 2751-756X
Volume 03 Number 03 (2024)
https://journals.innoscie.com/index.php/ijnras

Article

Principal Component Analysis for Feature Extraction

Noor Hasan Fadhil[1], Akmam Majed Mosa[2], Ihsan Sahib Abdulsaheed[3]

1. Computer Center, Al-Qasim Green University, Iraq
* Correspondence: noor_hassan@uoqasim.edu.iq
2. Computer Center, Al-Qasim Green University, Iraq
* Correspondence: akmammajed@uoqasim.edu.iq
3. Teacher in Directorate of Education, Babylon, Ministry of Education, Iraq
* Correspondence: sahibihsan@gmail.com

Abstract: Feature extraction in image processing involves transforming raw pixel data into a more meaningful representation that can be used for various tasks such as image classification, object detection, or image retrieval. The goal is to extract important attributes or characteristics (features) from the image that capture essential information and reduce the dimensionality of the data while preserving its most significant aspects. One of most Common Feature Extraction Techniques One popular The Principal Component Analysis (PCA) method is used to reduce dimensionality and extract features. In various domains, including image processing, finance, and bioinformatics. This paper explores the fundamentals of PCA, its mathematical foundation, and practical applications for feature extraction. We demonstrate how High-dimensional data can be converted into a lower-dimensional space using PCA, while retaining significant information, enhancing computational efficiency, and improving model performance. Using PCA for feature extraction involves transferring, as much as possible, the variance (information) of the initial data with high dimensions placed in an area with lower dimensions. Images are inherently high-dimensional data, with each pixel representing a feature. For example, a 256x256 grayscale image has 65,536 features. Analyzing and processing such high-dimensional data can be computationally intensive and may lead to overfitting in machine learning models. Autism Facial image dataset used in this paper. PCA reduces this dimensionality by identifying the most significant components (principal components) that demonstrate the variation in the image data.

Keywords: PCA, Dimensionality, Robust Technique, Image Processing

Citation: Fadhil, N. H., Mosa, A. M., & Abdulsaheed, I. S. Principal Component Analysis for Feature Extraction. Vital Annex: International Journal of Novel Research in Advanced Sciences 2024, 3(3), 100-103.

Received: 14th Aug 2024
Revised: 21st Aug 2024
Accepted: 28th Aug 2024
Published: 4th Sept 2024

1. Introduction

To identify large-scale data, such image data, feature extraction is required [1]. It is crucial that the features acquired contain all of the input data's information. In the era of big data, handling and analyzing large datasets is a common challenge. High-dimensional data may result to problem such as overfitting, more complex computations, and challenges with data visualization [2].

Dimensionality reduction techniques like PCA help mitigate these challenges by reducing the number of features while preserving essential information. PCA is particularly effective in feature extraction, allowing us to identify the most significant components of the data and discard noise and redundant information [1, 3]. Prior to World War II, Principal Components Analysis (PCA), a multivariate statistics technique, was developed. But it wasn't until the 1960s, during the Natural and Social Sciences' "Quantitative Revolution," that this approach was used more widely [4].

The primary cause of this delay was the extreme complexity of the calculations required by this technique. Principal components and other multivariate statistical

techniques have nearly infinite applications, but this was only made possible with the invention and advancement of computers [2, 5].

At the same time, since the use of principal components in technical problems necessitated perfect accuracy, exact numerical techniques for calculating eigenvalues and eigenvectors, among other things, became necessary [5].

2. Materials and Methods

PCA for Feature Extraction

Data Preparation

To effectively apply PCA, the data needs to be preprocessed and standardized. Standardization guarantees that every feature makes an equal contribution to the analysis, preventing features with larger scales from influencing the outcomes [6].

Applying PCA

Once the data is standardized, PCA can be applied. The number of components to retain can be specified based on the desired amount of variance to preserve [5].

Interpretation and Visualization

The principal components can be analyzed to understand the variance they explain. Visualization techniques, such as scatter plots, can help interpret the transformed data [7].

3. Results and Discussion

Mathematical Foundation of PCA

According to any data projection, the biggest variance is located in the principal component, also known as the first coordinate., followed by the second coordinate, the second greatest variance, and so on. This is how PCA transforms the data into a new coordinate system. The procedures in PCA are [7]:

1. **Standardization**: transforming the data to a variance of 1 and a mean of 0. Before applying PCA, it's essential to standardize the data so that each feature has a mean of 0 and a variance of 1.

   Given a dataset (X) that includes p features and n samples:

$$\mathbf{Xstandardized} = \mathbf{X} - \frac{\boldsymbol{\mu}}{\mathbf{Q}} \ldots\ldots\ldots\ldots \quad \text{Eq (1)}$$

   where μ is each feature's mean, and σ is each feature's standard deviation [6].

2. **Covariance Matrix Computation**: Calculating the covariance matrix to understand how the variables of the data relate to each other.

   The covariance matrix captures the variance and the linear relationships between features.

   For the standardized data matrix Xstandardized (denoted as X for simplicity), the covariance matrix C is given by:

$$\mathbf{C} = \frac{\mathbf{1}}{\mathbf{n-1}} * \boldsymbol{X^T} * \mathbf{X} \ldots\ldots\ldots\ldots \quad \text{Eq (2)}$$

   where X^T is the transpose of X [8].

$$\mathbf{Cv} = \boldsymbol{\lambda}\mathbf{v} \ldots\ldots\ldots \quad \text{Eq (3)}$$

   where λ is an eigenvalue and v is the corresponding eigenvector.

The eigenvalues ($\lambda 1, \lambda 2, \ldots, \lambda p$) represent the amount of variance captured by each principal component [9].

The eigenvectors (v1,v2,…,vp) represent the directions of the principal components.

3. **Eigenvalue and Eigenvector Calculation**: To determine the principal components, the eigenvalues and eigenvectors are derived from the covariance matrix.

Eigenvalues and eigenvectors of the covariance matrix C are computed to identify the principal components.

The eigenvalue equation is

4. **Formation of Principal Components**: Ordering the eigenvectors by eigenvalues in descending order and forming the principal components.

Principal components are formed by projecting the original data onto the eigenvectors.

The main components Z are given by:

$$Z = XV$$

Where V/math {V}V is the matrix of eigenvectors [6, 9].

5. **Projection of Data**: Transforming the original data into the new space defined by the principal components.

The number of principal components to retain can be determined by the explained variance.

The explained variance ratio for each principal component is:

$$\boldsymbol{R} = \frac{\boldsymbol{\lambda i}}{\sum_{j=1}^{p} \boldsymbol{\lambda j}} \ldots\ldots\ldots\ldots \quad \text{Eq (4)}$$

Where R represent Explained Variance Ratio to retain a certain percentage (e.g., 95%) of the variance, select the smallest K such that: $\sum_{j=1}^{k}$ Explained Variance Ratio ≥ 0.95 [10].

Practical Applications

Image Processing

PCA can lower an image's dimensionality during image processing., making storage and computation more efficient while preserving essential features for tasks like face recognition and object detection [11].

Finance

PCA is used in finance to identify underlying factors that influence asset prices, aiding in portfolio management and risk assessment [12].

Bioinformatics

In bioinformatics, PCA helps in analyzing gene expression data by reducing noise and highlighting significant patterns in high-dimensional genomic data [13].

4. Conclusion

PCA is a robust technique It is essential in many domains working with big datasets for feature extraction and dimensionality reduction. Using PCA, high-dimensional data is transformed into a lower-dimensional space. Enhances computational efficiency and

model performance, while retaining critical information. As data continues to grow in complexity and volume, PCA will remain a fundamental tool in the data scientist's arsenal.

By projecting high-dimensional data onto the directions of maximum variance, PCA converts it into a lower-dimensional space. The directions (eigenvectors) of the covariance matrix that correspond to the largest eigenvalues are known as the principal components. PCA reduces the dimensionality of the data while maintaining major information by keeping the principal components that capture the most variance.

This mathematical foundation and example illustrate how PCA works and how it can be applied to feature extraction in various applications, including image processing.

REFERENCES

[1] A. Maćkiewicz and W. Ratajczak, "Principal Components Analysis (PCA)," *Computers & Geosciences*, vol. 19, no. 3, pp. 303-342, 1993.

[2] H. Abdi and L. J. Williams, "Principal Component Analysis," *Wiley Interdisciplinary Reviews: Computational Statistics*, vol. 2, no. 4, pp. 433-459, 2010.

[3] S. P. Mishra, U. Sarkar, S. Taraphder, S. Datta, D. Swain, R. Saikhom, and M. Laishram, "Multivariate Statistical Data Analysis-Principal Component Analysis (PCA)," *International Journal of Livestock Research*, vol. 7, no. 5, pp. 60-78, 2017.

[4] M. Greenacre, P. J. Groenen, T. Hastie, A. I. d'Enza, A. Markos, and E. Tuzhilina, "Principal Component Analysis," *Nature Reviews Methods Primers*, vol. 2, no. 1, p. 100, 2022.

[5] B. M. S. Hasan and A. M. Abdulazeez, "A Review of Principal Component Analysis Algorithm for Dimensionality Reduction," *Journal of Soft Computing and Data Mining*, vol. 2, no. 1, pp. 20-30, 2021.

[6] A. Tharwat, "Principal Component Analysis—A Tutorial," *International Journal of Applied Pattern Recognition*, vol. 3, no. 3, pp. 197-240, 2016.

[7] H. Cardot and D. Degras, "Online Principal Component Analysis in High Dimension: Which Algorithm to Choose?," *International Statistical Review*, vol. 86, no. 1, pp. 29-50, 2018.

[8] S. Naveen, A. Omkar, J. Goyal, and R. Gaikwad, "Analysis of Principal Component Analysis Algorithm for Various Datasets," in *2022 International Conference on Futuristic Technologies (INCOFT)*, 2022, pp. 1-7.

[9] F. Yao, J. Coquery, and K. A. Lê Cao, "Independent Principal Component Analysis for Biologically Meaningful Dimension Reduction of Large Biological Data Sets," *BMC Bioinformatics*, vol. 13, p. 1-15, 2012.

[10] A. Tharwat, "Principal Component Analysis—A Tutorial," *International Journal of Applied Pattern Recognition*, vol. 3, no. 3, pp. 197-240, 2016.

[11] G. R. Naik, Ed., *Advances in Principal Component Analysis: Research and Development*. Springer, 2017.

[12] F. Kherif and A. Latypova, "Principal Component Analysis," in *Machine Learning*, Academic Press, 2020, pp. 209-225.

[13] M. Zhao, Z. Jia, Y. Cai, X. Chen, and D. Gong, "Advanced Variations of Two-Dimensional Principal Component Analysis for Face Recognition," *Neurocomputing*, vol. 452, pp. 653-664, 2021.

[14] J. Deng, K. Wang, D. Wu, X. Lv, C. Li, J. Hao, and W. Chen, "Advanced Principal Component Analysis Method for Phase Reconstruction," *Optics Express*, vol. 23, no. 9, pp. 12222-12231, 2015.

[15] Y. Hao, "Integrated Analysis of Multimodal Single-Cell Data," *Cell*, vol. 184, no. 13, pp. 3573-3587, 2021, doi: 10.1016/j.cell.2021.04.048.

[16] J. Chong, "Using MetaboAnalyst 4.0 for Comprehensive and Integrative Metabolomics Data Analysis," *Current Protocols in Bioinformatics*, vol. 68, no. 1, 2019, doi: 10.1002/cpbi.86.

[17] J. Grove, "Identification of Common Genetic Risk Variants for Autism Spectrum Disorder," *Nature Genetics*, vol. 51, no. 3, pp. 431-444, 2019, doi: 10.1038/s41588-019-0344-8.

[18] R. D. Riley, "Calculating the Sample Size Required for Developing a Clinical Prediction Model," *The BMJ*, vol. 368, 2020, doi: 10.1136/bmj.m441.