**PRINCIPAL COMPONENT ANALYSIS**

|  |  |
| --- | --- |
| **Tags** | principal components, correlation matrix, covariance matrix, factor loadings,  eigenvalues, eigenvectors, dimensionality reduction, variance,  projection error |

**1. Basics**

**Principal Components Analysis (PCA)** operates as an unsupervised machine learning algorithm.

Its aim is to reduce highly correlated features by transforming the feature covariance matrix into a set of fewer uncorrelated composite variables. Through this process, PCA generates eigenvectors that define the principal components, representing the new uncorrelated composite variables.

Additionally, it produces eigenvalues, offering insights into the proportion of the overall variance in the original data explained by each eigenvector and its corresponding principal component.

**2. Guide: List of inputs and outputs**

**Inputs of PCA**:

1. Dataset of multivariate observations with correlated features.
2. Correlation or Covariance Matrix which serves as a basis for features transformation.

**Outputs of PCA**:

1. Principal Components (PCs): PCA is used to transform highly correlated features of dataset into a few main and uncorrelated composite variables (PCs) that capture the maximum variance in the dataset.

Importantly, PCA involves two key concepts: eigenvectors and eigenvalues.

1. Eigenvectors define new, mutually uncorrelated composite variables that are linear combinations of the original features. Given that it is a vector, it also shows the direction.
2. Eigenvalue is associated with each eigenvector. It provides the proportion of total variance in the initial dataset explained by each eigenvector.

Using the scree plot we could visualize how the the eigenvectors are ordered from highest to lowest according to their eigenvalues. Their usefullness depends on how much of the total variance in the initial data is explained by each of them.

Remember that PCs are linear combinations of the initial dataset, so we need only a few PCs to explain most of the total variance in the initial covariance matrix.

1. Explained variance by each PC out of the total variance in the initial data.

**3.Features of PCA**

**Feature extraction** is used when raw features are transformed into new ones. This is the linear features' transformation performed by mapping from original data.

A black lines with numbers and symbols

Description automatically generated with medium confidence

where 𝑘 (number of PCs)<<𝑁.

The linear transformation is represented by the following equation

𝑦=𝑊𝑋

where 𝑊 are the weights (components loadings), 𝑋 the initial features and 𝑦, the principal components (PCs)

Notably, these new variables 𝑦 are orthogonal (independent) of each other.

**4.Principal components computation**

Unlike many ML algorithms, PCA doesn't have hyperparameters that are tuned in the same way as in a SVM supervised learning or a neural network techniques.

At the same time there are few important considerations to be taken into account in the PCA process.

1. Number of Principal Components (K). We need to decide on the number of PCs to retain, which determine the dimensionality of the transformed data.
2. Centering the Data. The data is centered by calculating the averages of features and shifting around the center of data.
3. Scaling the Data. We need to scale (standardize) data before applying PCA. Standardization involves subtracting the mean of the dataset and dividing it by variance.
4. PCs computation.

* The main method of finding PCs is arranged via the variance-covariance matrix.
* In the equation

𝑌=𝑋𝐴

the rows of the matrix 𝐴 represent the eigenvectors which show how the PCs are oriented, and their composition is referred as loadings to --indicate the weight in PCs' contribution.

* The variance-covariance matrix of the principal components

𝑆𝑌=𝐴𝑆𝑋𝐴𝑇

is a diagonal matrix whose eigenvalues represent the explained variance by the corresponding PCs.

On the figure below you could see the visual representation of first and second principal components of the hypothetical three-dimensional dataset.

* We assume that data has been standardized along the x-, y-, and z-axes with mean 0 and standard deviation 1, and PCA has been applied.
* Projection error represents the vertical distance between each data point and PC1.
* The spread (variation) of the data along PC1 represents the distance between each data point in the parallel direction.
* PC1 is found by selecting the line for which the sum of the projection errors for all data points is minimized and the sum of the spread is maximized. Therefore, PC1 is the unique vector with the largest proportion of the variance in the initial data.
* PC2 is uncorrelated with PC1 and is at right angles to PC1 and captures the next largest proportion of varianceA diagram of a graph

  Description automatically generated

*Source: Kathleen DeRose, CFA, et. al., 2024 "CFA Program. Refresher Reading. Machine Learning"*

**5.Mathematics behind PCA**

1. The eigenvalues of **A** 𝑛×𝑛 matrix are the solution to the characteristic equation

determinant(**𝐀**−λ**𝐈**)=|(**𝐀**−λ**𝐈**)|=0

If λ is an eigenvalue of **𝐀**, then there exists a vector 𝑥⃗  such that

**𝐀**𝑥⃗ =λ𝑥⃗

1. In a 2×2 matrix **𝐀** with eigenvectors 𝑥1→,𝑥2→ and eigenvalues λ1,λ2

**𝐀**𝚽=𝚽Λ

1. After normalizing the eigenvectors to make them orthogonal we get:

**𝐀**=𝚽Λ𝚽𝑇

**6.Advantages: The benefits of using PCA methodology**

**Dimensionality reduction.**

In datasets with numerous features, visual representation and model fitting can become very complex and prone to "noise," as the random influences specific to the dataset contribute to it.

The aim of dimensionality reduction is to explain the dataset with the numerous correlated features by a smaller set of uncorrelated features that still does quite well in explaining the total variance.

Importantly, dimension reduction is useful even when working with relatively low number of features.

Machine learning models are quicker to train and reduce overfitting in a data if provided with lower dimensional datasets.

**7. Disadvantages: The drawbacks of using PCA methodology**

**1.The main drawback of PCA lies in the difficulty of their interpretation by the end user of PCA.**

The PCs are combinations of the dataset’s initial features, and they usually cannot be easily labeled by the analyst, who may perceive PCA as a “black box.”

**2. The trade-off between lower-dimensional, manageable dataset and loss of information.**

For example, when we analyze how many PCs are sufficient to explain the returns of equity index which contains companies of all economic sectors, the initial dataset consists of index prices and more than 2,000 of technical and fundamental features. Multi-collinearity among these features is a problem because so many features or combinations of them tend to have overlaps. Scree plots are helpful to identify the proportion of total variance in the data explained by each principal component, and select the smallest number of principal components which contribute to explaining the desired proportion of the variance (80% and more)

**7. PCA method application**

1. Import corresponding libraries

2. Create dataset

For the purposes of this PCA analysis we will use the diabetes dataset with 10 baseline variables, and thus 10 dimensions.

3. Initialize with normalization

We have performed the (Z-score normalization) which involved rescaling features such that they have the properties of a standard normal distribution with a mean of zero and a standard deviation of one.

It is an important step, as each one of initial variables contributes equally to the analysis.

4. Fit transform with dataframe

PCA has optimized to store maximum variance in the first PC (contains 40.2 % of the variance in total), then in the second (around 14.9 %), in the PC3 (12%), and so on.

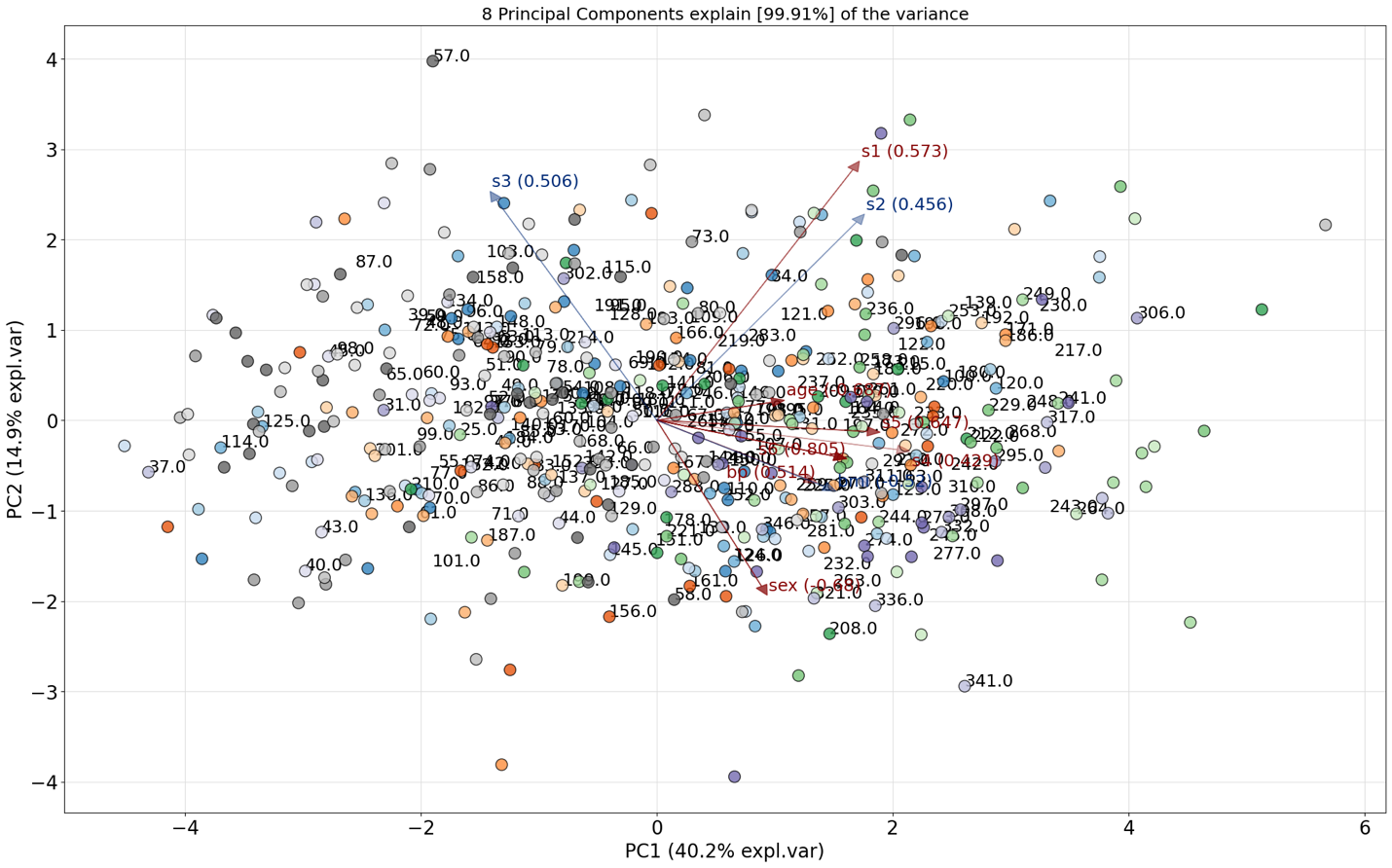
With the top 8 PCs out of 10 we cover 99.91% of all the variance.

A graph with a line and a line

Description automatically generated with medium confidence

After we draw the biplot we see that the first 2 PCs can be used to separate the diabetes classes.

* The arrow's angle shows the contribution of the variable that is seen in
* Large loadings (+ or -) indicate that a variable has a strong relationship to a particular PC.
* The arrow's length describes the strength of the loading.



Now let's examine the loadings in more detail to understand clearly the distribution of the sample given the variable.

The variable s4 (total cholesterol / HDL) has a positive loading of 0.428834 and explains mostly the variance in the first PC1 (it is almost a horizontal line). If we color the samples in the scatter plot based on total cholesterol values, we see a distinction between samples that are left and right side of the scatter plot.

A graph of a graph with blue and white dots

Description automatically generated with medium confidence

Finally, we examine the best performing features.

Their extraction is based on the loadings of the PCs.

We could see that hat most of the variance for the 1st PC is derived from the variable S4 total cholesterol/HDL. For the 2nd component, it is derived by s1 total serum cholesterol, for the PC3 by bp average blood pressure etc.

A screenshot of a computer

Description automatically generated

**References**

* Carol Anne Hargreaves,Chandrika Kadirvel Mani. "The Selection of Winning Stocks Using Principal Component Analysis." *American Journal of Marketing Research*, Vol. 1, No. 3, 2015, pp. 183-188 [https://www.researchgate.net/publication/281042621\_The\_Selection\_of\_Winning\_Stocks\_Using\_Principal\_Component\_Analysis](https://colab.research.google.com/corgiredirector?site=https%3A%2F%2Fwww.researchgate.net%2Fpublication%2F281042621_The_Selection_of_Winning_Stocks_Using_Principal_Component_Analysis)
* Eugene W. Park. "Principal Component Analysis and Hidden Markov Model for Forecasting Stock Returns" *Courant Institute of Mathematical Sciences*, New York University, May, 2023. [https://arxiv.org/pdf/2307.00459.pdf](https://colab.research.google.com/corgiredirector?site=https%3A%2F%2Farxiv.org%2Fpdf%2F2307.00459.pdf)
* PCA datasourse URL: [https://www4.stat.ncsu.edu/~boos/var.select/diabetes.html](https://colab.research.google.com/corgiredirector?site=https%3A%2F%2Fwww4.stat.ncsu.edu%2F%7Eboos%2Fvar.select%2Fdiabetes.html)
* Bradley Efron, Trevor Hastie, Iain Johnstone and Robert Tibshirani (2004) “Least Angle Regression,” Annals of Statistics (with discussion), 2002,407-499.

**Section 3. Technical section**

Unlike many ML algorithms, PCA doesn't have hyperparameters that are tuned in the same way as in a SVM supervised learning or a neural network techniques.

At the same time there are few important considerations to be taken into account in the PCA process.

1. Number of Principal Components (K). We need to decide on the number of PCs to retain, which determine the dimensionality of the transformed data.
2. Centering the Data. The data is centered by calculating the averages of features and shifting around the center of data.
3. Scaling the Data. We need to scale (standardize) data before applying PCA. Standardization involves subtracting the mean of the dataset and dividing it by variance.

**Section 4. Marketing Alpha**

In datasets with numerous features, visual representation and model fitting can become very complex and prone to "noise," as the random influences specific to the dataset contribute to it.

The aim of dimensionality reduction is to explain the dataset with the numerous correlated features by a smaller set of uncorrelated features that still does quite well in explaining the total variance.

Importantly, dimension reduction is useful even when working with relatively low number of features.

Machine learning models are quicker to train and reduce overfitting in a data if provided with lower dimensional datasets.

**Section 5. Learn More**

In the research paper "The Selection of Winning Stocks Using Principal Component Analysis." *the authors have shown that via PCA application they managed to* reduce 24 stock market fundamental variable indicators to 4 variables namely ROI, ROE, Book Value per Share and Revenue per Share.

Importantly, their PCA analysis revealed that the above 4 variables loaded as 2 key important factors 1 and 2 - Management Effectiveness and Common Share Value respectively. The stocks which have high management Effectiveness and common share value could be considered as winning stocks which was demonstrated by authors.

**REGRESSION TREES**

|  |  |
| --- | --- |
| **Tags** | decision trees, predictor space, partitioning predictor space, internal nodes, leaf nodes, prediction via stratification of the feature space, recursive binary splitting, residual sum of squares, rss, greedy algorithm, overfitting, unstable, computational complexity, hyperparameters, criterion, splitter, max depth, min samples split |

**1. Jargons**

1. **Predictor Space**

Predictor space is the space of all possible values of all the independent variables. For example, if we want to predict the average Glassdoor rating of a company, we might use the following independent variables such as the number of employees, salary, job openings etc. The predictor space is the space of all possible values of these independent variables.

1. **Partitioning Predictor Space**

Partitioning predictor space is the process of dividing the predictor space into smaller regions. For example, we might divide the predictor space into two regions: one region where the number of employees is less than 1000 and another region where the number of employees is greater than 1000.

1. **Internal Nodes**

Internal nodes are the nodes of a decision tree that are not leaf nodes. Internal nodes are used to partition the predictor space. For example, we might have an internal node that partitions the predictor space into two regions: one region where the number of employees is less than 1000 and another region where the number of employees is greater than 1000**.**

1. **Leaf Nodes**

Leaf nodes are the nodes of a decision tree that are not internal nodes. Leaf nodes are used to make predictions. For example, we might have a leaf node that predicts the average Glassdoor rating of a company is 4.5, if the number of employees is less than 1000.