Lancaster University
Leipzig

Principal Component Analysis

Jingwen Li

12.08.2024

Contents

1 Definition

2 History

3 Process

4 Implications in data science

5 Alternative technique

Lancaster University Leipzig

lancasterleipzig.de

Lancaster University
Leipzig

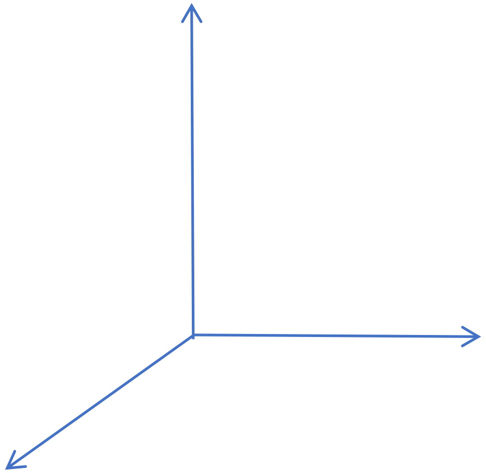
Definition of PCA

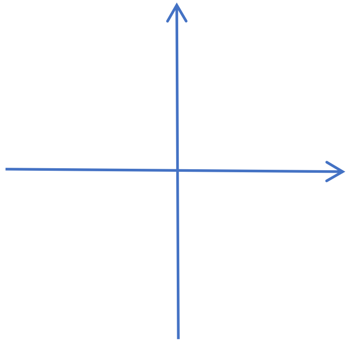
Part.01

lancasterleipzig.de

Definition of PCA

- Principal Component Analysis(PCA) is a statistical technique used for dimensionality reduction, which transforms a high-dimension data set into a low-dimension one while retaining as much of the original information as possible.





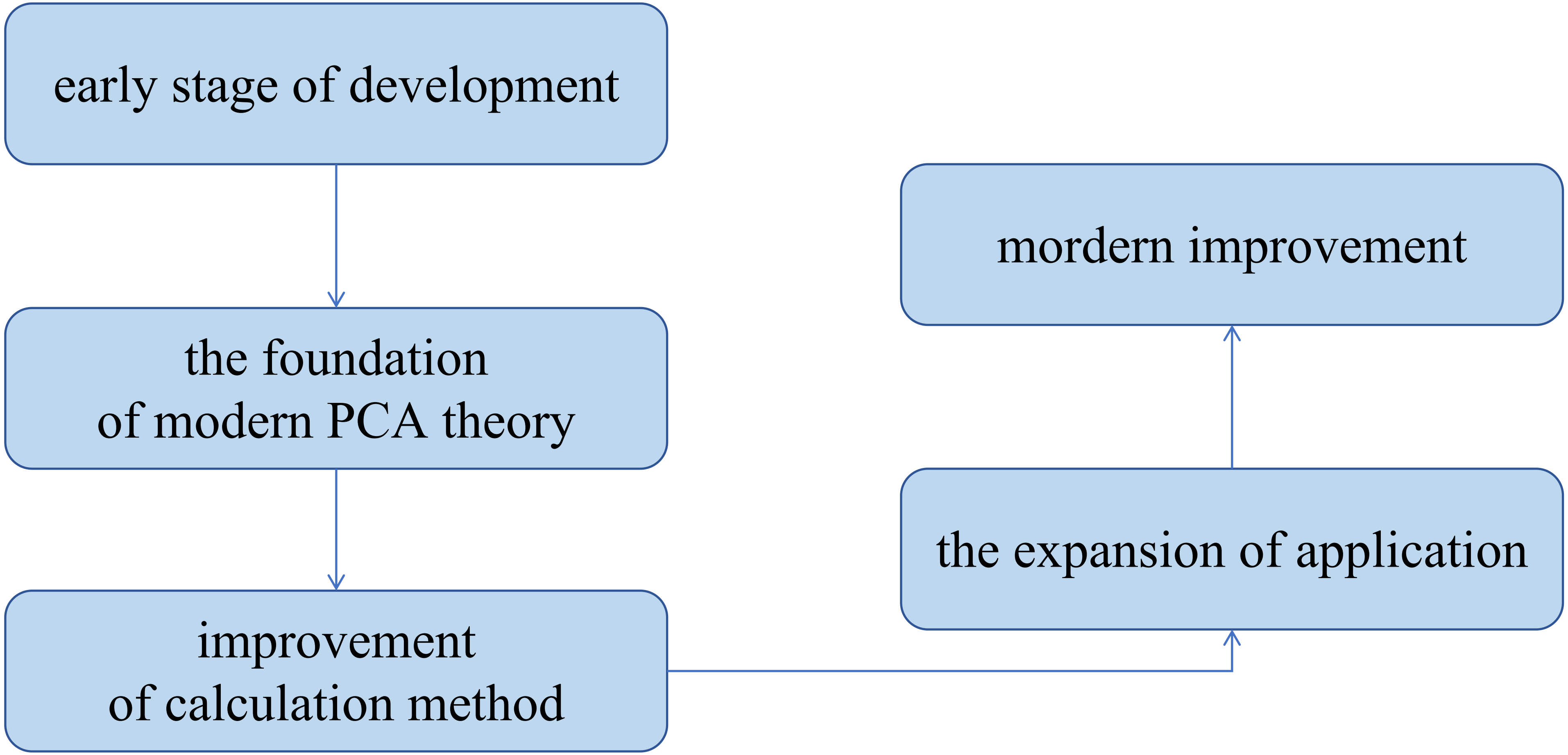
Lancaster University
Leipzig

History of PCA

Part.02

lancasterleipzig.de

History


early stage of development
the foundation of modern PCA theory
improvement of calculation method
the expansion of application
mordern improvement

Process of PCA

Part.03

Process of PCA

- Data standardization

$$\begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1\mathrm{m}} \\ x_{21} & x_{22} & \cdots & x_{2\mathrm{m}} \\ \vdots & \vdots & \ddots & \vdots \\ x_{\mathrm{n}1} & x_{\mathrm{n}2} & \cdots & x_{nm} \end{bmatrix}$$

① $\mu_j = \frac{1}{n}\sum_{i=1}^{n}$

② $\sigma_{\mathrm{j}} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(X_{ij} - \mu_j)^2}$

③ $X_{standardized,ij} = \frac{X_{ij} - \mu_j}{\sigma_j}$

data matrix after data standardized(X_standardized):

	sepal length (cm)	sepal width (cm)	petal length (cm)	petal width (cm)
0	-0.900681	1.019004	-1.340227	-1.315444
1	-1.143017	-0.131979	-1.340227	-1.315444
2	-1.385353	0.328414	-1.397064	-1.315444
3	-1.506521	0.098217	-1.283389	-1.315444
4	-1.021849	1.249201	-1.340227	-1.315444
..	...	...	...	...
145	1.038005	-0.131979	0.819596	1.448832
146	0.553333	-1.282963	0.705921	0.922303
147	0.795669	-0.131979	0.819596	1.053935
148	0.432165	0.788808	0.933271	1.448832
149	0.068662	-0.131979	0.762758	0.790671

[150 rows x 4 columns]

Process of PCA

Lancaster University Leipzig

- Calculate covariance matrix

$$\Sigma = \frac{1}{\text{n - 1}} X^T_{Standardized} X_{Standardized}$$

lancasterleipzig.de

Process of PCA

Lancaster University Leipzig

- Calculate covariance matrix

$$X_{standardized} = \begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1m} \\ x_{21} & x_{22} & \cdots & x_{2m} \\ \vdots & \vdots & \ddots & \vdots \\ x_{n1} & x_{n2} & \cdots & x_{nm} \end{bmatrix} \quad X^{T}_{Standardized} = \begin{bmatrix} x_{11} & x_{21} & \cdots & x_{n1} \\ x_{12} & x_{22} & \cdots & x_{n2} \\ \vdots & \vdots & \ddots & \vdots \\ x_{1m} & x_{2m} & \cdots & x_{nm} \end{bmatrix}$$

$$C = X^{T}_{S\tan dardized} \times X_{S\tan dardized}$$

$$c_{ij} = \sum_{k=1}^{n} x_{ki} \times x_{kj}$$

lancasterleipzig.de

Process of PCA

- Calculate eigenvalues and eigenvectors

$$\Sigma v = \lambda v$$

$$\det(\Sigma - \lambda I) = 0$$

$$(\Sigma - \lambda_i I) v_i = 0$$

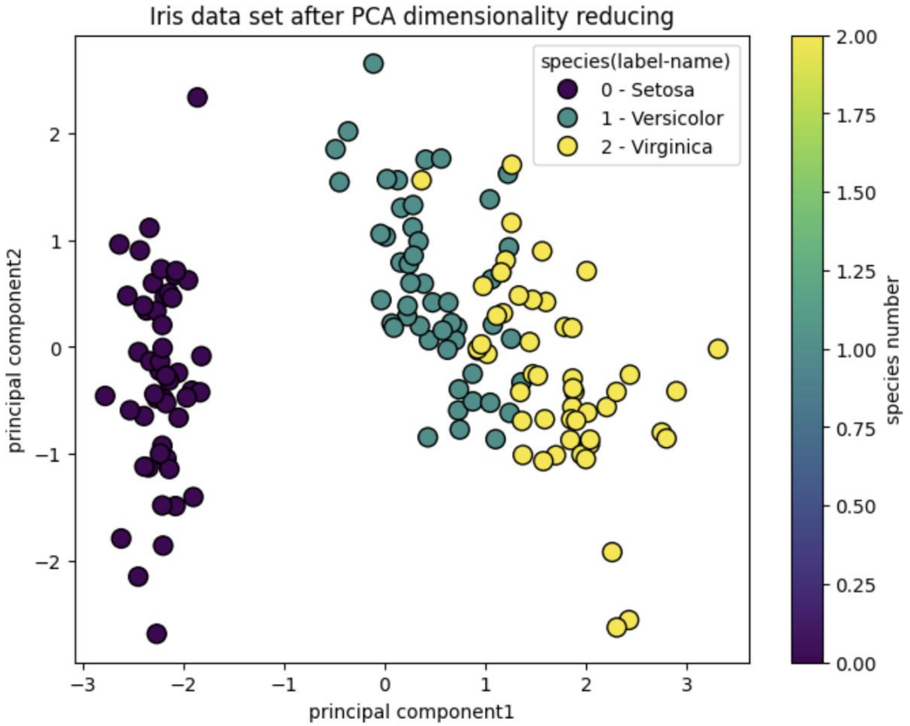
```
 eigenvalues(Eigenvalues):
[2.93808505 0.9201649  0.14774182 0.02085386]

 eigenvectors(Eigenvectors):
                   sepal length (cm)  sepal width (cm)  petal length (cm)  petal width (cm)
sepal length (cm)           0.521066         -0.377418          -0.719566          0.261286
sepal width (cm)           -0.269347         -0.923296           0.244382         -0.123510
petal length (cm)           0.580413         -0.024492           0.142126         -0.801449
petal width (cm)            0.564857         -0.066942           0.634273          0.523597
```

Process of PCA

- Select principal component
- Construct transformation matrix
- Data transformation and dimensionality reduction

$$W = \begin{bmatrix} \mathrm{v}_1 & v_2 & \cdots & v_k \end{bmatrix}$$

$$X' = W^T X_{Standardized}$$



lancasterleipzig.de

Implications in data science

Part.04

Implications in data science

- Dimensionality Reduction
- Noise reduction
- Data visualization
- Feature extraction

Lancaster University
Leipzig

Alternative technique

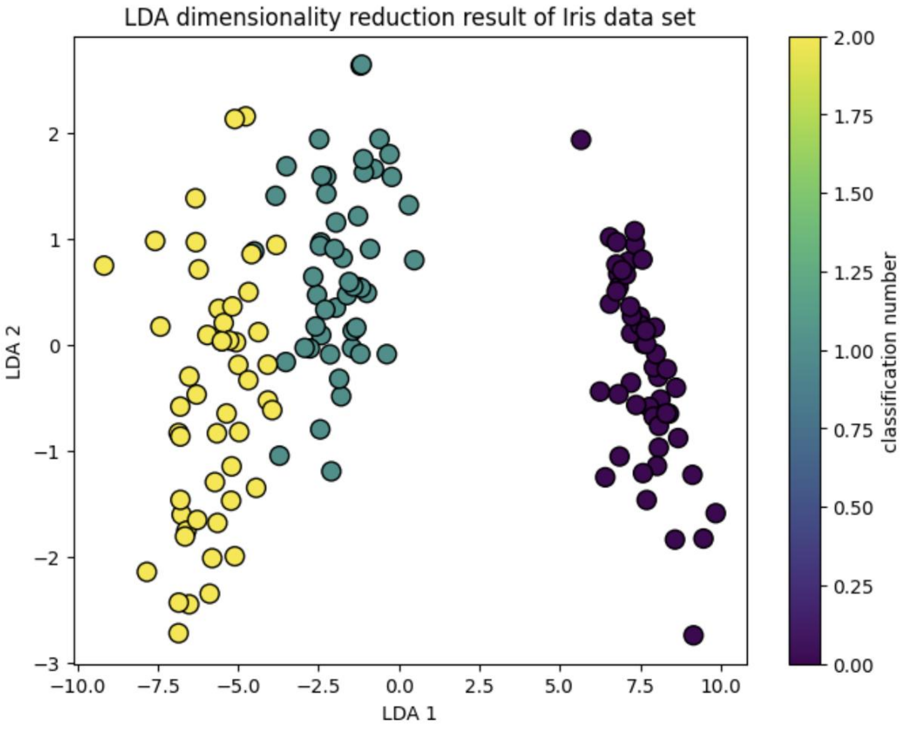
Part.05

lancasterleipzig.de

LDA(Linear Discriminant Analysis)

Lancaster University Leipzig

- Comparison with PCA, LDA is supervised and uses category labels to find directions that best differentiate between different categories. This means that LDA generally performs better than PCA in classification tasks.



lancasterleipzig.de

Lancaster University
Leipzig

Thank you for your time and patience.