Principal Component Analysis

Peter Juhasz

April 8, 2024

Information

Contact

- name: Peter Juhasz
- email: peter.juhasz@math.au.dk

Agenda

1. Principal Component Analysis: **April 8**
2. t-Distributed Stochastic Neighbor Embedding: **April 15**
3. Uniform Manifold Approximation and Projection: **April 22**

Outline

Introduction

What is PCA?

- statistical dimensionality reduction technique
- identifies patterns in data, simplifies representation

Definition Purpose

- Linear dimensionality reduction
- feature extraction
- noise reduction
- visualization, interpretation

Applications

- image processing, compression
- machine learning
- data analysis
- signal processing

Main Idea

Data Transformation

- goal: find new basis (principal components) to capture maximum variance
- principal components are uncorrelated (orthogonal)

Maximizing Variance

- first principal components capture the most variance of the data with 1 dimension
- subsequent components capture the remaining variance in decreasing order

Uncorrelated Features

- principal components are uncorrelated (orthogonal)
- principal components: new, independent features

Mathematical Foundation

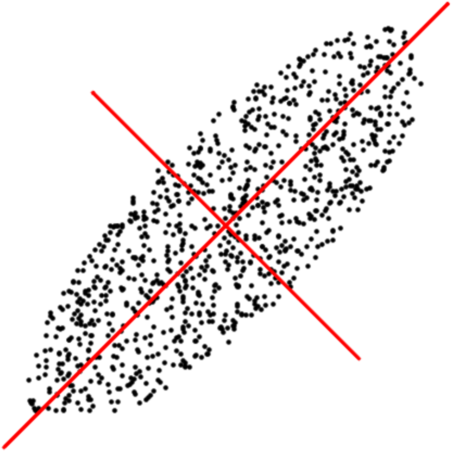
Relationships of Features

- data matrix: $X \in \mathbb{M}_{n \times m}$ (n points, m features)
- correlation structure of data: covariance matrix $\hat{\Sigma} = 1/(n-1)\, X^T X$

Eigenvalue decomposition

- eigenvalue decomposition: $\hat{\Sigma} = Q \Lambda Q^{-1}$
- eigenvectors of $\hat{\Sigma}$ represent directions of maximum variance
- eigenvectors of covariance matrix: principal components
- $\hat{\Sigma}$ has orthogonal eigenvectors

Example



Explained Variance Ratio

Explained Variance Ratio

- How much variance is explained by each principal component?
- How many components should we keep?
- Explained variance ratio: proportion captured variance
  $\mathrm{EVR}_i = |\lambda_i| / \sum_{j=1}^{m} |\lambda_j|$

Dimensionality Reduction

- keep components with high explained variance ratios
- discard components with low ratios

Scree Plot

- visual representation of explained variance ratios
- eigenvalues vs component index
- determine number of components: elbow method

Component Loadings

Definition

- Component loadings: correlation between original features and principal components
- Contribution of each feature to the principal component

Interpretation

- Loading plots: loading values for each feature across different principal components
- Help to understand principal components using original features

Steps of PCA

1. build data matrix
2. standardization
3. covariance matrix
4. eigenvalue decomposition
5. select principal components
6. transform data
7. interpret results:
   plot $\widetilde{X}_{\text{trans}}$ check Q

- $X = \begin{bmatrix} x_1 & \cdots & x_m \end{bmatrix} \in \mathbb{M}_{n \times m}$
- $\widetilde{\mathrm{x}}_\mathrm{i} := (\mathrm{x}_\mathrm{i} - \hat{\mu}(\mathrm{x}_\mathrm{i}))/\hat{\sigma}(\mathrm{x}_\mathrm{i})$
- $\hat{\Sigma} = 1/(n-1)\,\widetilde{X}^T\,\widetilde{X}$
- $\hat{\Sigma} = Q \Lambda Q^{-1}$
- $\mathrm{EVR}_i = |\lambda_i| / \sum_j |\lambda_j|$
- $\widetilde{X}_{\text{trans}} = \widetilde{X}\, Q_{\text{reduced}}$

Connection with Singular Value Decomposition

$$X \in \mathbb{C}^{n\times m} \qquad \Longrightarrow \qquad X = U\,S\,V^T$$

$$U \in \mathbb{C}^{n\times n}, \text{ unitary: } U^T = U^{-1}$$
$$S \in \mathbb{R}^{n\times m}_{\geq 0}, \text{ diagonal}$$
$$V \in \mathbb{C}^{m\times m}, \text{ unitary: } V^T = V^{-1}$$

PCA

- $\hat{\Sigma} = 1/(n-1)\,\widetilde{X}^T\widetilde{X}$
- $\hat{\Sigma} = Q\Lambda Q^{-1}$
- $\widetilde{X}^T\widetilde{X} = (n-1)Q\Lambda Q^{-1}$

SVD

- $\widetilde{X} = U\,S\,V^T$
- $\widetilde{X}^T\widetilde{X} = V\,(S^T S)\,V^T$
- $\widetilde{X}^T\widetilde{X} = V\,S^2\,V^{-1}$

Limitations

Linearity Assumption

- PCA only discovers linear relationships
- Non-linear relationships may be lost

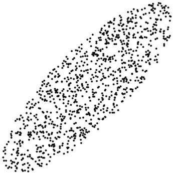
Scaling

- PCA is sensitive to the scaling of the features
- features with larger scales dominate principal components

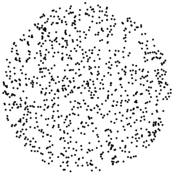
Interpretability

- principal components are hard to interpret
- meaning of components may not be clear

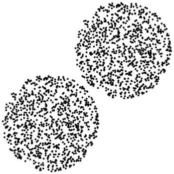
Does PCA Help?



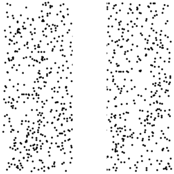
(a) yes



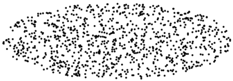
(b) no



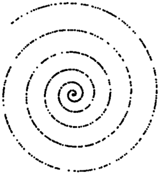
(c) yes



(d) yes



(e) **not always**



(f) no

Exercise

$$X = \begin{bmatrix} 2 & 3 & 4 \\ 3 & 2 & 1 \\ 1 & 5 & 3 \\ 4 & 4 & 5 \\ 5 & 1 & 2 \end{bmatrix} \qquad \widetilde{X} = \sqrt{\tfrac{2}{5}} \begin{bmatrix} -1 & 0 & 1 \\ 0 & -1 & -2 \\ -2 & 2 & 0 \\ 1 & 1 & 2 \\ 2 & -2 & -1 \end{bmatrix}$$

$$\hat{\Sigma} = \tfrac{1}{10} \begin{bmatrix} 10 & -7 & -1 \\ -7 & 10 & 6 \\ -1 & 6 & 10 \end{bmatrix} = Q \Lambda Q^{-1}$$

$$\left| \hat{\Sigma} - \lambda I \right| = \tfrac{1}{10^3} \left(\left(\tfrac{\lambda}{10}\right)^3 - 30 \left(\tfrac{\lambda}{10}\right)^2 + 214 \tfrac{\lambda}{10} - 224 \right) = 0$$

$$\Lambda = \begin{bmatrix} 1.97 & 0 & 0 \\ 0 & 0.90 & 0 \\ 0 & 0 & 0.13 \end{bmatrix} \qquad Q = \begin{bmatrix} -0.54 & 0.65 & -0.53 \\ 0.69 & -0.02 & -0.73 \\ 0.48 & 0.76 & 0.44 \end{bmatrix}$$

Exercise

$$\text{EVR} = [0.66,\ 0.30,\ 0.04] \quad \Longrightarrow \quad \text{first two explains } 96\%$$

$$\widetilde{X}_{\text{trans}} = \begin{bmatrix} 1.02 & 0.11 \\ -1.65 & -1.50 \\ 2.46 & -1.34 \\ 1.11 & 2.15 \\ -2.94 & 0.58 \end{bmatrix} \qquad L = \begin{bmatrix} -0.54 & 0.65 \\ 0.69 & -0.02 \\ 0.48 & 0.76 \end{bmatrix}$$

Steps of PCA in R

- Load dataset
  data(iris) or read.csv()
- Perform PCA
  pca <- prcomp(iris[, -5], scale. = TRUE)
- Interpret / visualize results
  summary(pca)

R Examples

R Examples

Summary

- goal: dimensionality reduction, feature extraction
- eigenvalue decomposition of the covariance matrix: relationship with SVD
- principal components = eigenvectors of the covariance matrx
- basis transformation to principal components

Q & A

Resources

- Hervé Abdi and Lynne J Williams, *Principal component analysis*, Wiley interdisciplinary reviews: computational statistics **2** (2010), no. 4, 433–459.
- Rasmus Bro and Age K Smilde, *Principal component analysis*, Analytical methods **6** (2014), no. 9, 2812–2831.
- Svante Wold, Kim Esbensen, and Paul Geladi, *Principal component analysis*, Chemometrics and intelligent laboratory systems **2** (1987), no. 1-3, 37–52.