MADS-MMS – Mathematics and Multivariate Statistics

Dimensionality Reduction

Prof. Dr. Stephan Doerfel

FACHHOCHSCHULE KIEL
University of Applied Sciences

Moodle (WiSe 24/25)

Outline

Motivation

high dimensional data

Motivation

high dimensional data

- is hard to analyze

Motivation

high dimensional data

- is hard to analyze
- is visualizable only in projections

Motivation

high dimensional data

- is hard to analyze
- is visualizable only in projections
- is expensive to work with (e.g. ML)

Motivation

high dimensional data

- is hard to analyze
- is visualizable only in projections
- is expensive to work with (e.g. ML)
- leads to curse of dimensionality

Motivation

high dimensional data

- is hard to analyze
- is visualizable only in projections
- is expensive to work with (e.g. ML)
- leads to curse of dimensionality
- contains useful information

Motivation

high dimensional data

- is hard to analyze
- is visualizable only in projections
- is expensive to work with (e.g. ML)
- leads to curse of dimensionality
- contains useful information

observations

Motivation

high dimensional data

- is hard to analyze
- is visualizable only in projections
- is expensive to work with (e.g. ML)
- leads to curse of dimensionality
- contains useful information

observations

- high dimensional data is often over-complete (redundant, correlated, or dependent features)

Motivation

high dimensional data

- is hard to analyze
- is visualizable only in projections
- is expensive to work with (e.g. ML)
- leads to curse of dimensionality
- contains useful information

observations

- high dimensional data is often over-complete (redundant, correlated, or dependent features)
- idea of an **intrinsic** dimension – a minimal set of variables that actually describe the data (ideally fully, realistically approximately)

Motivation

high dimensional data

- is hard to analyze
- is visualizable only in projections
- is expensive to work with (e.g. ML)
- leads to curse of dimensionality
- contains useful information

observations

- high dimensional data is often over-complete (redundant, correlated, or dependent features)
- idea of an **intrinsic** dimension – a minimal set of variables that actually describe the data (ideally fully, realistically approximately)

dimensionality reduction

goal: find a small set of variables (not a subset!) that retains most of the information in the data

Chapter Goals

- Understand basic idea of dimensionality reduction
- Apply the steps for principal component analysis (PCA)
- Understand the role of PCA as a preprocessing step in data analysis

Outline

What is Dimensionality Reduction?

- create "new features" from the original ones: **principal components**

What is Dimensionality Reduction?

- create "new features" from the original ones: **principal components**
- new features generate a low-dimensional subspace of the original features space

What is Dimensionality Reduction?

- create “new features” from the original ones: **principal components**
- new features generate a low-dimensional subspace of the original features space
- remove covariance between different variables

What is Dimensionality Reduction?

- create "new features" from the original ones: **principal components**
- new features generate a low-dimensional subspace of the original features space
- remove covariance between different variables
- retain variance of the data

What is Dimensionality Reduction?

- create "new features" from the original ones: **principal components**
- new features generate a low-dimensional subspace of the original features space
- remove covariance between different variables
- retain variance of the data

→ **Principal Component Analysis** (PCA)

Variance

$$\mathrm{var}(x) = \frac{1}{n}\sum_{n=1}^{N}(x_n - \bar{x})^2$$

- a measure for the deviation from the mean

Variance

$$\mathrm{var}(x) = \frac{1}{n}\sum_{n=1}^{N}(x_n - \bar{x})^2$$

- a measure for the deviation from the mean
- a measure for how far the data is spread

Variance

$$\mathrm{var}(x) = \frac{1}{n}\sum_{n=1}^{N}(x_n - \bar{x})^2$$

- a measure for the deviation from the mean
- a measure for how far the data is spread
- goal of PCA: retain variance of data

Covariance

$$cov(x_1, x_2) = \frac{1}{N-1} \sum_{n=1}^{N} (x_{n1} - \bar{x}_1)(x_{n2} - \bar{x}_2)$$

- a measure of common variability of two variables

Covariance

$$cov(x_1, x_2) = \frac{1}{N-1} \sum_{n=1}^{N} (x_{n1} - \bar{x}_1)(x_{n2} - \bar{x}_2)$$

- a measure of common variability of two variables
- multiple variables give rise to a covariance matrix

Covariance

$$cov(x_1, x_2) = \frac{1}{N-1} \sum_{n=1}^{N} (x_{n1} - \bar{x}_1)(x_{n2} - \bar{x}_2)$$

- a measure of common variability of two variables
- multiple variables give rise to a covariance matrix
- high (absolute) covariance values indicate correlation between the two variables

Covariance

$$cov(x_1, x_2) = \frac{1}{N-1} \sum_{n=1}^{N} (x_{n1} - \bar{x}_1)(x_{n2} - \bar{x}_2)$$

- a measure of common variability of two variables
- multiple variables give rise to a covariance matrix
- high (absolute) covariance values indicate correlation between the two variables
- goal of PCA: remove covariance of (different) variables

Outline

Principal Steps of PCA

- Standardize data

Principal Steps of PCA

- Standardize data
- Identify principal components as linear combinations of the the original features

Principal Steps of PCA

- Standardize data
- Identify principal components as linear combinations of the the original features
- Order principal components by variance.

Principal Steps of PCA

- Standardize data
- Identify principal components as linear combinations of the the original features
- Order principal components by variance.
  - First vector is linear combination with highest variance.

Principal Steps of PCA

- Standardize data
- Identify principal components as linear combinations of the the original features
- Order principal components by variance.
  - First vector is linear combination with highest variance.
  - Following vectors have the same property among all vectors that are uncorrelated to the previous principal components.

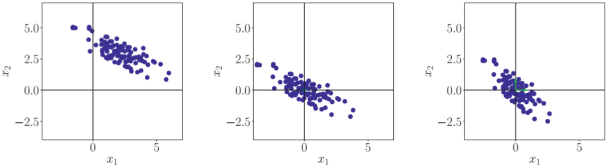
Principal Steps of PCA

- Standardize data
- Identify principal components as linear combinations of the the original features
- Order principal components by variance.
  - First vector is linear combination with highest variance.
  - Following vectors have the same property among all vectors that are uncorrelated to the previous principal components.
- Principal components are orthonormal vectors

Principal Steps of PCA

- Standardize data
- Identify principal components as linear combinations of the the original features
- Order principal components by variance.
  - First vector is linear combination with highest variance.
  - Following vectors have the same property among all vectors that are uncorrelated to the previous principal components.
- Principal components are orthonormal vectors
- Thus, the transformation is a matrix multiplication

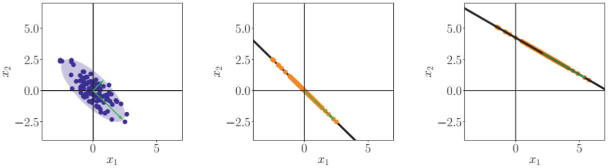
PCA in a Nutshell



(a) Original dataset.

(b) Step 1: Centering by subtracting the mean from each data point.

(c) Step 2: Dividing by the standard deviation to make the data unit free. Data has variance 1 along each axis.



(d) Step 3: Compute eigenvalues and eigenvectors (arrows) of the data covariance matrix (ellipse).

(e) Step 4: Project data onto the principal subspace.

(f) Undo the standardization and move projected data back into the original data space from (a).

Source: [1], p. 304

Transformation Matrix

data

Matrix $X \in \mathbb{R}^{d \times n}$ where d is the number of dimensions (features) and n the number of data instances

Transformation Matrix

data
Matrix $X \in \mathbb{R}^{d \times n}$ where d is the number of dimensions (features) and n the number of data instances

transformation matrix
Matrix $W \in \mathbb{R}^{p \times d}$ where $p \leq d$ is the number of principal components

Transformation Matrix

data

Matrix $X \in \mathbb{R}^{d \times n}$ where d is the number of dimensions (features) and n the number of data instances

transformation matrix

Matrix $W \in \mathbb{R}^{p \times d}$ where $p \leq d$ is the number of principal components

transformation

$\tilde{X} := W \cdot X \in \mathbb{R}^{p \times n}$

Transformations and Loadings

In the transformation, each component is created from the features as a linear combination:

$$PCA : \boldsymbol{x} \mapsto \tilde{\boldsymbol{x}} = W\boldsymbol{x} \quad \text{with} \quad \tilde{\boldsymbol{x}}_i := \sum_{j=1}^{d} w_{ij} x_i.$$

Transformations and Loadings

In the transformation, each component is created from the features as a linear combination:

$$PCA : \boldsymbol{x} \mapsto \tilde{\boldsymbol{x}} = W\boldsymbol{x} \quad \text{with} \quad \tilde{\boldsymbol{x}}_i := \sum_{j=1}^{d} w_{ij} x_i.$$

The **loadings** w_{ij} are the entries of the i-th line of the transformation matrix W.

Transformations and Loadings

In the transformation, each component is created from the features as a linear combination:

$$PCA : \boldsymbol{x} \mapsto \tilde{\boldsymbol{x}} = W\boldsymbol{x} \quad \text{with} \quad \tilde{\boldsymbol{x}}_i := \sum_{j=1}^{d} w_{ij} x_i.$$

The **loadings** w_{ij} are the entries of the i-th line of the transformation matrix W.

- a large positive loading w_{ij} indicates a high influence of feature j on component i in the positive direction

Transformations and Loadings

In the transformation, each component is created from the features as a linear combination:

$$PCA : \boldsymbol{x} \mapsto \tilde{\boldsymbol{x}} = W\boldsymbol{x} \quad \text{with} \quad \tilde{\boldsymbol{x}}_i := \sum_{j=1}^{d} w_{ij} x_i.$$

The **loadings** w_{ij} are the entries of the i-th line of the transformation matrix W.

- a large positive loading w_{ij} indicates a high influence of feature j on component i in the positive direction
- a large negative loading w_{ij} indicates a high influence of feature j on component i in the negative direction

Transformations and Loadings

In the transformation, each component is created from the features as a linear combination:

$$PCA : \boldsymbol{x} \mapsto \tilde{\boldsymbol{x}} = W\boldsymbol{x} \quad \text{with} \quad \tilde{\boldsymbol{x}}_i := \sum_{j=1}^{d} w_{ij} x_i.$$

The **loadings** w_{ij} are the entries of the i-th line of the transformation matrix W.

- a large positive loading w_{ij} indicates a high influence of feature j on component i in the positive direction
- a large negative loading w_{ij} indicates a high influence of feature j on component i in the negative direction

The data and the components can be visualized in combined plots (**biplots**).

Transformations and Loadings

In the transformation, each component is created from the features as a linear combination:

$$PCA : \boldsymbol{x} \mapsto \tilde{\boldsymbol{x}} = W\boldsymbol{x} \quad \text{with} \quad \tilde{\boldsymbol{x}}_i := \sum_{j=1}^{d} w_{ij} x_i .$$

The **loadings** w_{ij} are the entries of the i-th line of the transformation matrix W.

- a large positive loading w_{ij} indicates a high influence of feature j on component i in the positive direction
- a large negative loading w_{ij} indicates a high influence of feature j on component i in the negative direction

The data and the components can be visualized in combined plots (**biplots**).

Notebook 08_1_pca_decathlon

Usecases

- PCA allows visualizing high-dimensional data in low-dimensional spaces
- PCA is an analysis tool: correlated features (biplot)
- PCA is a tool for feature engineering: generates independent features
- PCA can be used a pre-processing step before applying clustering or other ML methods

Exercises 1

Next Steps for your Portfolio

- Run a PCA on your dataset(s).
- Observe and describe the behavior of variance, feature correlation, features and components, ...
- Transform you dataset and select a reasonable number of components.
- Run your previously used clustering approach on the data.
- Visualize results.
- Interpret the new clustering.
- Compare with previous results.

References

M. P. Deisenroth, A. A. Faisal, and C. S. Ong.
Mathematics for Machine Learning.
Cambridge University Press, 2020.