



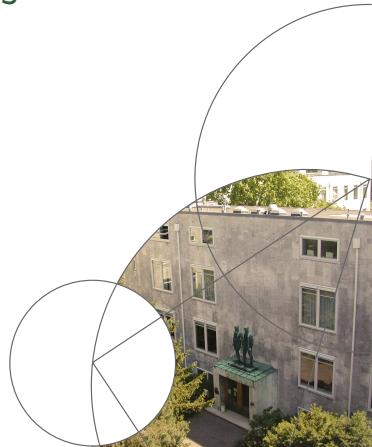
Faculty of Science



Principal Component Analysis

Machine Learning

Christian Igel
Department of Computer Science



Basic questions

- Principal component analysis (PCA, Karhunen-Loève transform) is arguably the most fundamental *unsupervised learning* algorithm.
- It is frequently used for (linear) reduction, (lossy) data compression, feature extraction, and data visualization.
- Let's consider $S = \{\mathbf{x}_1, \dots, \mathbf{x}_N\}$ with $\mathbf{x}_1, \dots, \mathbf{x}_N \in \mathcal{X} = \mathbb{R}^d$.
- We ask how
 - to reduce the length of the description to $k < d$ variables such that as much information as possible is preserved?
 - many dimensions are needed to capture a certain percentage of the variability of the data?
 - to visualize the data in two or three dimensions preserving as much of its variability as possible?



Outline

- ① Warmup: Basis and Coordinate System
- ② More Warmup: Matrix Decomposition
- ③ Principal Component Analysis
- ④ Summary



Outline

- ➊ Warmup: Basis and Coordinate System
- ➋ More Warmup: Matrix Decomposition
- ➌ Principal Component Analysis
- ➍ Summary



Basis vectors

- *Basis*: Set of linearly independent *basis vectors* \mathbf{u}_1, \dots that, in a linear combination, can represent every vector in given vector space
- *Orthonormal basis*: Basis vectors are orthogonal (i.e., $\mathbf{u}_i^\top \mathbf{u}_j = 0$ for $i \neq j$) and have unit length,
 $\|\mathbf{u}_i\| = \sqrt{\mathbf{u}_i^\top \mathbf{u}_1} = 1$
- Example: $\mathbf{x} \in \mathbb{R}^2$, orthonormal basis

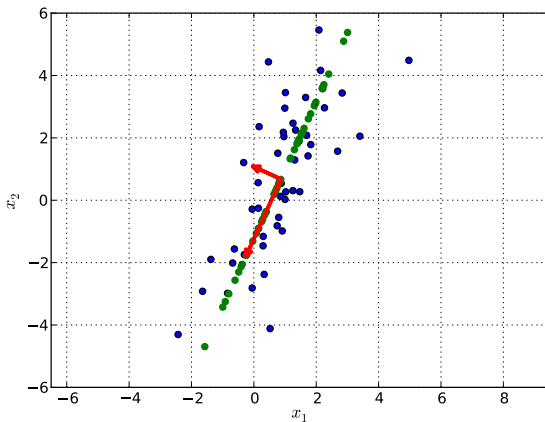
$$\begin{pmatrix} 1 \\ 0 \end{pmatrix}, \begin{pmatrix} 0 \\ 1 \end{pmatrix},$$

$$\mathbf{x} = (1, 0)^\top x_1 + (0, 1)^\top x_2$$

- Example can be read as “way from $\mathbf{0}$ to \mathbf{x} : First go x_1 units in the direction $(1, 0)^\top$ and then x_2 units in direction $(0, 1)^\top$ ”



2-dimensional example



Changing basis

- Changing basis means “expressing the way using two other, non-parallel directions \mathbf{u}_1 and \mathbf{u}_2 : First go z_1 units in the direction \mathbf{u}_1 and then z_2 units in direction \mathbf{u}_2 ” :

$$\mathbf{x} = \sum_{i=1}^d z_i \mathbf{u}_i = \sum_{i=1}^d (\mathbf{x}^\top \mathbf{u}_i) \mathbf{u}_i$$

- A lower dimensional representation uses “fewer directions” and a (different) origin/starting point.



Orthogonal matrix

- Gather basis vectors in $d \times d$ matrix \mathbf{U} such that the columns of \mathbf{U} correspond to the basis vectors.
- i th column of \mathbf{U} is given by \mathbf{u}_i .
- Square matrix composed of an orthonormal basis is an *orthogonal* matrix having the property $\mathbf{U}^\top = \mathbf{U}^{-1}$.
- Define $d \times k$ matrix \mathbf{U}_k as the first k basis vectors.
- i th column of \mathbf{U}_k is given by \mathbf{u}_i and the i th row of \mathbf{U}^\top by \mathbf{u}_i^\top .



Outline

- ① Warmup: Basis and Coordinate System
- ② More Warmup: Matrix Decomposition
- ③ Principal Component Analysis
- ④ Summary



Eigenvectors and eigenvalues

- For an eigenvector $\mathbf{u}_i \in \mathbb{R}^d$ of the $d \times d$ matrix \mathbf{S} it holds by definition:

$$\mathbf{S}\mathbf{u}_i = \lambda_i \mathbf{u}_i$$

λ_i is the corresponding eigenvalue.

- We consider only consider eigenvectors of unit length and assume that the eigenvectors are sorted according to $i < j \Rightarrow \lambda_i \geq \lambda_j$.



Eigen decomposition of real symmetric matrix

Any *real symmetric* $d \times d$ matrix \mathbf{M} can be decomposed as:

$$\mathbf{M} = \mathbf{Q}\mathbf{\Lambda}\mathbf{Q}^\top$$

with

- $\mathbf{Q} \in \mathbb{R}^{d \times d}$ being orthogonal,
- $\mathbf{Q}^\top = \mathbf{Q}^{-1}$,
- $\mathbf{\Lambda} \in \mathbb{R}^{d \times d} = \text{diag}(\lambda_1, \dots, \lambda_d)$ being diagonal, with $\lambda_1 \geq \lambda_2 \geq \dots \geq \lambda_d$ being the eigenvalues of \mathbf{M} , and
- the columns of \mathbf{Q} being the corresponding eigenvectors.



Singular value decomposition (SVD)

Any real $N \times d$ matrix X can be decomposed (let's assume $N \geq d$) as:

$$X = U\Gamma V^T$$

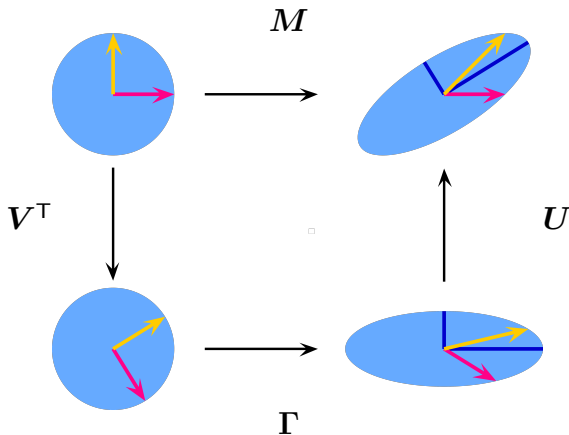
with

- $U \in \mathbb{R}^{N \times d}$ having orthonormal columns,
- $\Gamma \in \mathbb{R}^{d \times d} = \text{diag}(\gamma_1, \dots, \gamma_d)$ being diagonal with $\gamma_1 \geq \gamma_2 \geq \dots \geq \gamma_d$
- $V \in \mathbb{R}^{d \times d}$ being orthogonal.

The columns of V are the *right singular vectors* of X .



SVD example



drawing modified from Wikimedia Commons



SVD and eigen decomposition

- When M is positive semi-definite, its eigen decomposition is also a SVD.
- The right singular vectors of X are the eigenvectors of $X^T X$.
- $X^T X$ is positive (semi-) definite (i.e., all eigenvalues are non-negative), and the singular values of X are the square roots of the eigenvalues of $X^T X$.



Outline

- ① Warmup: Basis and Coordinate System
- ② More Warmup: Matrix Decomposition
- ③ Principal Component Analysis
- ④ Summary



Basic questions

- Principal component analysis (PCA, Karhunen-Loève transform) is arguably the most fundamental *unsupervised learning* algorithm.
- It is frequently used for (linear) reduction, (lossy) data compression, feature extraction, and data visualization.
- Let's consider $S = \{\mathbf{x}_1, \dots, \mathbf{x}_N\}$ with $\mathbf{x}_1, \dots, \mathbf{x}_N \in \mathcal{X} = \mathbb{R}^d$.
- We ask how
 - to reduce the length of the description to $k < d$ variables such that as much information as possible is preserved?
 - many dimensions are needed to capture a certain percentage of the variability of the data?
 - to visualize the data in two or three dimensions preserving as much of its variability as possible?



Example: Cambridge face database



Basic idea

- Find k -dimensional affine linear model $f : \mathbb{R}^k \rightarrow \mathbb{R}^d$ of the d -dimensional data representing S as accurately as possible:

$$f(\mathbf{z}) = \mathbf{b} + \mathbf{U}_k \mathbf{z} \text{ ,}$$

where $\mathbf{z} \in \mathbb{R}^k$, $\mathbf{b} \in \mathbb{R}^d$, and $\mathbf{U}_k \in \mathbb{R}^{d \times k}$.

- Vectors $\mathbf{u}_1, \dots, \mathbf{u}_k \in \mathbb{R}^d$ are the columns of \mathbf{U}_k and we require them to be pairwise orthogonal and of unit length.
- Model represents data in \mathbb{R}^d by k -dimensional parameters \mathbf{z} .



Reconstruction error

- We measure model quality by sum-of-squares *reconstruction error*

$$J = \frac{1}{N} \sum_{i=1}^N \|\mathbf{x}_i - f(\mathbf{z}_i)\|^2 .$$

- Formal goal of PCA

$$\min_{\mathbf{b}, \mathbf{U}_k, \{\mathbf{z}_1, \dots, \mathbf{z}_N\}} \frac{1}{N} \sum_{i=1}^N \|\mathbf{x}_i - f(\mathbf{z}_i)\|^2$$

subject to the constraints on \mathbf{U}_k .



Solution

- The choice for \mathbf{b} is the **empirical mean**:

$$\bar{\mathbf{x}} = \frac{1}{N} \sum_{i=1}^N \mathbf{x}_i$$

- Data points should be encoded by

$$\mathbf{z}_i = \mathbf{U}_k^T (\mathbf{x}_i - \bar{\mathbf{x}}) .$$

- k columns of \mathbf{U}_k should correspond to the **first k eigenvectors of the *data covariance matrix* or *empirical covariance matrix***

$$\mathbf{S} = \frac{1}{N} \sum_{i=1}^N (\mathbf{x}_i - \bar{\mathbf{x}})(\mathbf{x}_i - \bar{\mathbf{x}})^T .$$



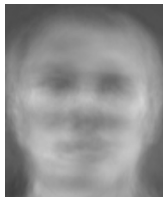
Mean face



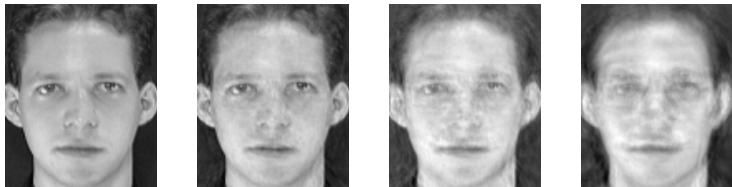
$$\bar{\mathbf{x}} = \frac{1}{N} \sum_{i=1}^N \mathbf{x}_i$$



Eigenfaces



Reconstruction using eigenfaces



using all, 300, 200, 100 eigenfaces



PCA algorithm

Algorithm 1: dimensionality reduction using PCA

Input: data $S = \{\mathbf{x}_1, \dots, \mathbf{x}_N\}$, number of dimensions k

1 compute the empirical mean $\bar{\mathbf{x}} = \frac{1}{N} \sum_{i=1}^N \mathbf{x}_i$

2 compute empirical covariance matrix

$$\mathbf{S} = \frac{1}{N} \sum_{i=1}^N (\mathbf{x}_i - \bar{\mathbf{x}})(\mathbf{x}_i - \bar{\mathbf{x}})^\top$$

3 compute the $d \times k$ matrix \mathbf{U}_k composed of the first k eigenvectors of \mathbf{S} , where the eigenvectors are ordered by decreasing eigenvalue

4 compute $\mathbf{z}_i = \mathbf{U}_k^\top (\mathbf{x}_i - \bar{\mathbf{x}})$ for $i = 1, \dots, N$

Output: mean $\bar{\mathbf{x}}$, principal components \mathbf{U}_k , projected data $\{\mathbf{z}_1, \dots, \mathbf{z}_N\}$, model $f(\mathbf{z}) = \bar{\mathbf{x}} + \mathbf{U}_k \mathbf{z}$



Eckart and Young theorem

The (squared) Frobenius norm of an $N \times d$ matrix \mathbf{X} is

$$\|\mathbf{X}\|_F^2 = \sum_{n=1}^N \sum_{i=1}^d x_{ij}^2 \ .$$

The $N \times d$ matrix $\tilde{\mathbf{X}}$ with rank k approximating the $N \times d$ matrix \mathbf{X} best in terms of the Frobenius norm $\|\mathbf{X} - \tilde{\mathbf{X}}\|_F^2$ is

$$\tilde{\mathbf{X}} = \mathbf{X} \mathbf{V}_k \mathbf{V}_k^\top \ ,$$

where \mathbf{V}_k is the matrix of top- k right singular vectors of \mathbf{X} .



Maximizing variance

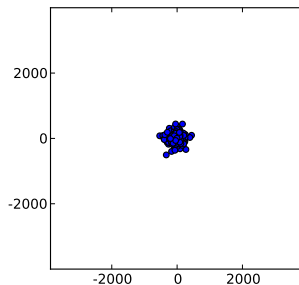
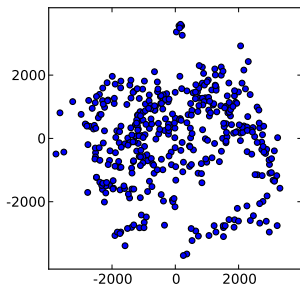
- We measure the variability of S by the trace
 $\text{trace}\{S\} = \sum_{i=1}^d s_{ii} = \sum_{i=1}^d \lambda_i = \sum_{j=1}^d \mathbf{u}_j^\top S \mathbf{u}_j$. (Sum of the eigenvalues of a diagonalizable matrix is equal to its trace.)
- Accordingly the variability of z_1, \dots, z_N is

$$\frac{1}{N} \sum_{i=1}^N [\mathbf{U}_k^\top \mathbf{x}_i - \mathbf{U}_k^\top \bar{\mathbf{x}}] [\mathbf{U}_k^\top \mathbf{x}_i - \mathbf{U}_k^\top \bar{\mathbf{x}}]^\top = \sum_{j=1}^k \mathbf{u}_j^\top S \mathbf{u}_j .$$

- Minimizing reconstruction error corresponds to maximizing variance.



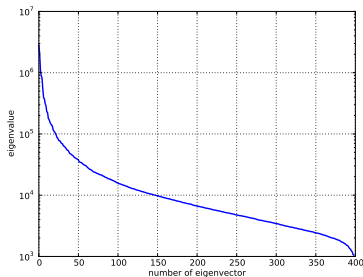
Variance of faces



first two eigenvalues vs. eigenvalues 99 and 100



“Explained variance”



We have:

$$\text{trace}\{\mathbf{S}\} = \sum_{j=1}^d \mathbf{u}_j^T \mathbf{S} \mathbf{u}_j = \sum_{i=1}^d \lambda_i$$

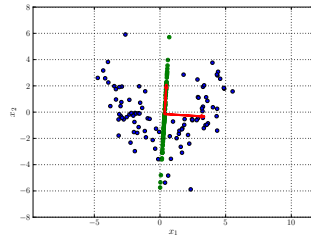
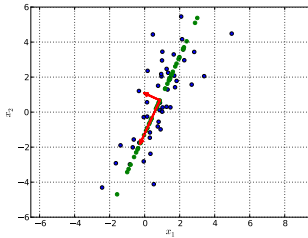
The quotient

$$\frac{\sum_{i=1}^k \lambda_i}{\sum_{i=1}^d \lambda_i}$$

measures the fraction of variance “explained” by the first k principal components.



Linear and non-linear example



Outline

- ① Warmup: Basis and Coordinate System
- ② More Warmup: Matrix Decomposition
- ③ Principal Component Analysis
- ④ Summary



Summary

Principal component analysis (PCA) is frequently used for

- dimensionality reduction,
- noise reduction, and
- visualization.

It

- is an affine linear model of the data,
- is a change of basis, new basis vectors are orthogonal,
- minimizes reconstruction error, and
- maximizes variance.

