02450 Introduction to Machine Learning and Data Mining

Week 3: Principal Component Analysis

Bjørn Sand Jensen

18 February 2025

DTU Compute, Technical University of Denmark

Today

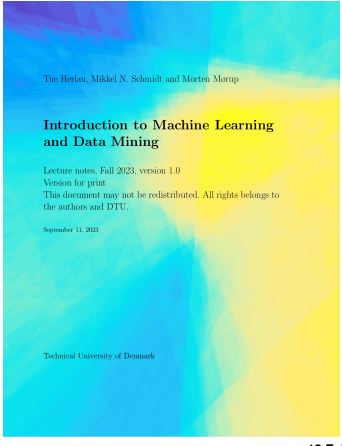
Feedback Groups of the day:

Peter Graugaard, Saeed Mohamud Amin, Carl Victor Cleveland Nielsen, August Daniel Glargaard Mikkelsen, Malthe Brandbyge Petersen, Morten Hedelykke Dietz Fuglsang, Yung-Chun Huang, Pragalva Dhungana, Frederik Raae Johansen, Nishant Sharma, Elias Bech Madsen, Konstantinos Kountras, Sumaya Mahamed Mahamed, Jesper Clement Aaløse, William Hoffmann Hyldig, Frederik Boye Vesth, Mark Øgaard Pedersen, Julia Stephanie Fleay, Magnus Hauge Nielsen, Erik Li, Nurdan Turan, Hannah Espelund Nørrelykke, Rasmus Pultz Niemann, Arti Dinesh Tanna, Maxime Swagel, Virginia Natonek, Manuel Maria Pardo, Eduarda Nágem De Castro Krag, Oliver Christensen, Søren Kostrup Jacobsen, Oliver Magnus Jalving Lystlund, Ana Paula Rodriguez del Moral, Arslan Falak, Anas Mohamed Adan Roble, Noah Zacharias Langergaard Madsen, Kasper Lindhardtsen, Josephine Lund-Kühl, Kasper Mohr Lundqvist, Adam Fouad Rustom, Pernille Taaftegaard Jensen, Sofia Neri, Jakob Moody Vinther, Andreas Stampe Dalgaard

Reading/homework material:

Chapter 3

P3.1, P3.2



Lecture Schedule

1. Introduction
   4 February: C1,C2

Data: Feature extraction, and visualization

2. Summary statistics, similarity and visualization
   11 February: C4,C7
3. **Computational linear algebra and PCA**
   18 February: C3
4. Probability and probability densities
   25 February: C5, C6

Supervised learning: Classification and regression

5. Decision trees and linear regression
   4 March: C8, C9 **(Project 1 due 6 March at 17:00)**
6. Overfitting, cross-validation and Nearest Neighbor
   11 March: C10, C12
7. Performance evaluation, Bayes, and Naive Bayes
   18 March: C11, C13
8. Artificial Neural Networks and Bias/Variance
   25 March: C14, C15
9. AUC and ensemble methods
   1 April: C16, C17

Unsupervised learning: Clustering and density estimation

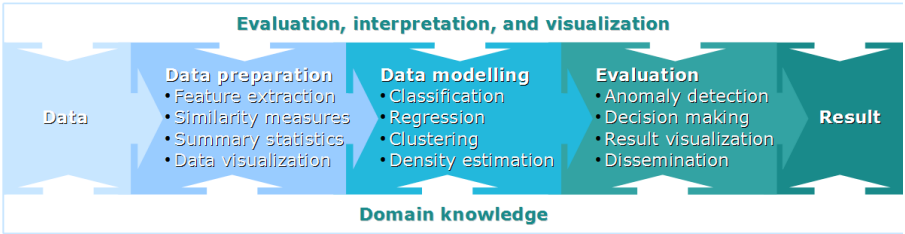
10. K-means and hierarchical clustering
    8 April: C18 **(Project 2 due 10 April at 17:00)**
11. Mixture models and density estimation
    22 April: C19, C20
12. Association mining
    29 April: C21

Recap

13. Recap and discussion of the exam
    6 May: C1-C21

Online help: Piazza
Videos of lectures: `https://panopto.dtu.dk`
Streaming of lectures: Zoom (link on DTU Learn)

Learning Objectives



Learning Objectives

- Understand and apply matrix operations on data matrices
- Understand and apply principal component analysis for data visualization and feature extraction

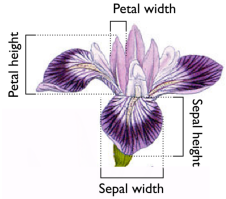
Plan for today:

- Lecture 3 (13:00 – ∼15:00)
  - **Linear algebra revisitted**
  - **Principal Component Analysis**
  - **Applications of PCA**
- Exercises (15:00–17:00)

Practicalities and announcements

- Exercise sessions: Exercises vs homework problems vs projects
- Exam: No printed notes! You can not make the notes on your iPad / computer / touchscreen and print them and bring them to the exam.
- Exam: What is a "non-programmable" calculator?
- Exam date (currently) 28th May 2025. See `https://www.inside.dtu.dk/en/undervisning/regler/regler-for-eksamen/eksamensdatoer`

Recall: Data points as vectors



A data point is a set of **real** numbers represented as a vector

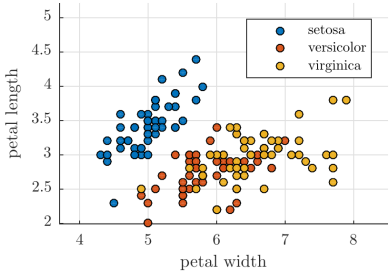
$$\boldsymbol{x} = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_M \end{bmatrix} \in \mathbb{R}^M \quad \text{e.g. } \boldsymbol{x} = \begin{bmatrix} 5.1 \\ 3.5 \\ 1.4 \\ 0.2 \end{bmatrix}$$

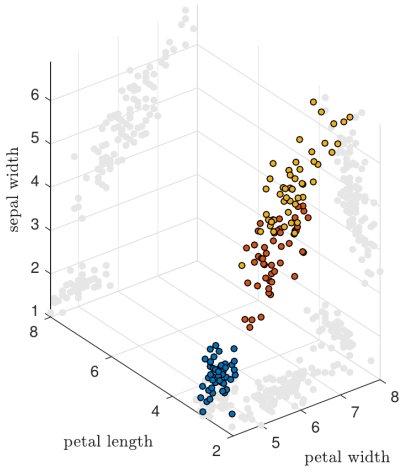
A matrix is used to hold multiple data points (as rows)

$$\boldsymbol{X} = \begin{bmatrix} x_{1,1} & \dots & x_{1,M} \\ \vdots & & \vdots \\ x_{N,1} & \dots & x_{N,M} \end{bmatrix} \in \mathbb{R}^{N \times M}$$

$$\boldsymbol{X} = \begin{bmatrix} \text{Sepal Length} & \text{Sepal Width} & \text{Petal Length} & \text{Petal Width} \\ 5.1 & 3.5 & 1.4 & 0.2 \\ 4.9 & 3.0 & 1.4 & 0.2 \\ 4.7 & 3.2 & 1.3 & 0.2 \\ 4.6 & 3.1 & 1.5 & 0.2 \\ 5.8 & 2.7 & 5.1 & 1.9 \\ \vdots & \vdots & \vdots & \vdots \\ 5.7 & 2.8 & 4.1 & 1.3 \end{bmatrix}$$

Example: Visualizaiton of high-dimensional data





Linear algebra for machine learning

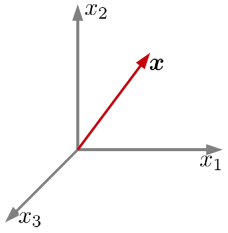
Data representation Data in ML is often represented as vectors and matrices.

Transformations Operations like scaling, rotations, and projections are naturally and efficiently formulated using matrix operations.

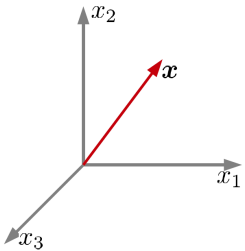
Dimensionality reduction and visualizaiton Data is high-dimensional and sometimes requires reduction in dimensionality to find patterns and extract new features (e.g., as combinations of existing ones). Techniques like PCA use eigenvectors and eigenvalues.

Model formulation Functions are easily and compactly formulated in vector/matrix notation (e.g., linear models $f(\boldsymbol{x}; \boldsymbol{w}) = \boldsymbol{w}^\top \tilde{\boldsymbol{x}}$ and neural networks)

Optimization Modern ML models relies heavily on gradient methods. Gradients are conviently represented as vectors and matrixes (e.g. a Jacobian or Hessian)



Vector spaces



- A M-dimensional vector space is just $\mathbb{R}^M$
- This is the set of all M-dimensional vectors
- A vector space is closed under linear combinations

$$\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_n$$

$$a_1, x_2, \ldots, a_n$$

$$\boldsymbol{x} = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_M \end{bmatrix}$$

Distances and norms

- The (Euclidian) norm of a vector measures it's length (magnitude):

$$\|\boldsymbol{x}\|_2 = \sqrt{\sum_{n=1}^{N} |x_n|^2} = \sqrt{\boldsymbol{x}^\top \boldsymbol{x}}$$

- The Frobenius norm of a matrix measures it's magnitude:

$$\|\boldsymbol{X}\|_F^2 = \sum_{i,j} x_{i,j}^2 = \mathrm{trace}(\boldsymbol{X}\boldsymbol{X}^T) = \mathrm{trace}(\boldsymbol{X}^T\boldsymbol{X})$$

Where trace takes the sum of the diagonal elements, i.e. $\mathrm{trace}(\boldsymbol{A}) = \sum_{i=1}^{N} a_{i,i}$

Scaling, rotation and shear

Rotation:

$$\boldsymbol{R}(\theta) = \begin{bmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{bmatrix}$$

$$\boldsymbol{x}' = \boldsymbol{R}\boldsymbol{x}$$

E.g. $\boldsymbol{R} = \begin{bmatrix} 0.87 & -0.5 \\ 0.5 & 0.87 \end{bmatrix}$

Scaling and stretching:

$$\boldsymbol{S}(\alpha_x, \alpha_y) = \begin{bmatrix} \alpha_x & 0 \\ 0 & \alpha_y \end{bmatrix}$$

$$\boldsymbol{x}' = \boldsymbol{S}\boldsymbol{x}$$

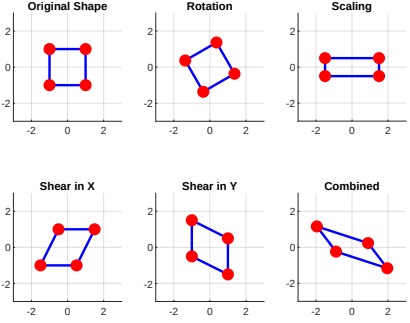
E.g. $\boldsymbol{S} = \begin{bmatrix} 1.5 & 0 \\ 0 & 0.5 \end{bmatrix}$

Shear:

$$\boldsymbol{D}_x(\alpha_x, \alpha_y, \beta) = \begin{bmatrix} \alpha_x & \beta \\ 0 & \alpha_y \end{bmatrix}$$

$$\boldsymbol{x}' = \boldsymbol{D}_x\boldsymbol{x}$$

E.g. $D_x = \begin{bmatrix} 1 & 0.5 \\ 0 & 1 \end{bmatrix}$



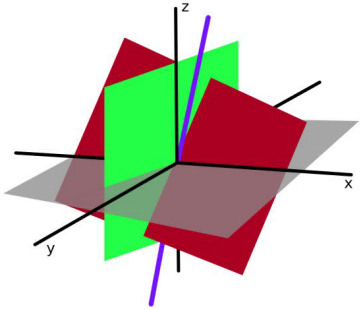
Projection: Subspaces

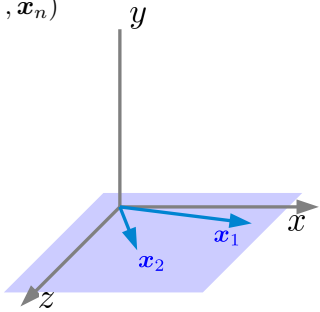
- A **subspace** generalizes the concept of a line/plane
- If we consider n vectors $\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_n$ the span is then all linear combinations

  $$\boldsymbol{z} = a_1\boldsymbol{x}_1 + a_2\boldsymbol{x}_2 + \cdots + a_n\boldsymbol{x}_n$$

  and is said to be a subspace

  $$V = \text{span}(\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_n)$$





Basis of a (sub)space

- Vectors $\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_n$ are said to be **linearly independent** if

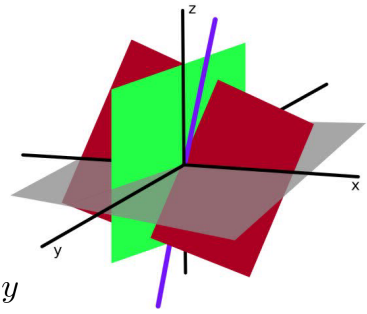
  $\mathbf{0} = a_1\boldsymbol{x}_1 + a_2\boldsymbol{x}_2 + \cdots + a_n\boldsymbol{x}_n$

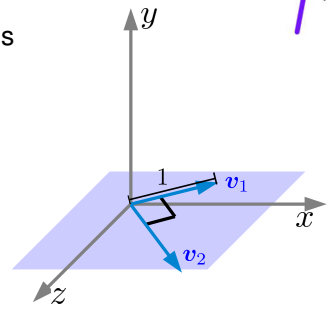
  implies $a_1 = a_2 = \cdots = a_n = 0$
- A **basis** for a vector space, V, are n linearly independent vectors

  $V = \text{span}(\boldsymbol{v}_1, \boldsymbol{v}_2, \ldots, \boldsymbol{v}_n)$
- A basic is orthonormal if the basis is orthogonal and of unit length

  $\boldsymbol{v}_i{}^\top \boldsymbol{v}_j = 0$ for $i \neq j$

  $||\boldsymbol{v}_i|| = 1$



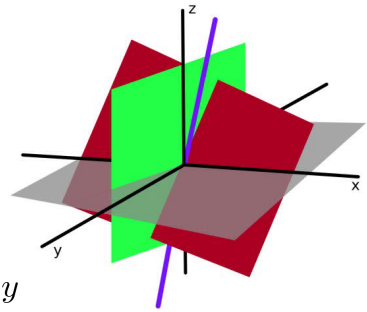


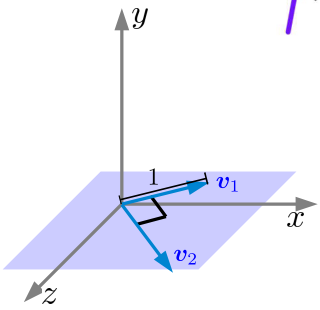
Basis of a (sub)space

- A **basis** for a vector space, V, are n linearly independent vectors
  $\boldsymbol{V} = \text{span}(\boldsymbol{v}_1, \boldsymbol{v}_2, \ldots, \boldsymbol{v}_n)$
- We collect the basis vectors into a matrix

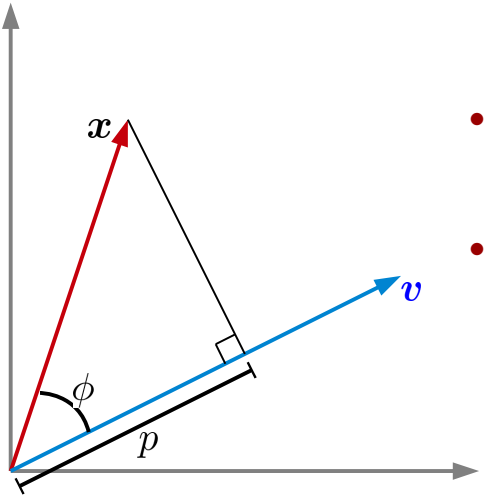
$$\boldsymbol{V} = \begin{bmatrix} | & | & & | \\ \boldsymbol{v}_1 & \boldsymbol{v}_2 & \cdots & \boldsymbol{v}_n \\ | & | & & | \end{bmatrix}$$

- If the basis is orthonormal, the matrix satisfies
  $\boldsymbol{V}^\top \boldsymbol{V} = \boldsymbol{I}$ and $\boldsymbol{V}^\top = \boldsymbol{V}^{-1}$





Projection



- Angle between vectors
  $cos(\phi) = \frac{\boldsymbol{v}^\top \boldsymbol{x}}{\|\boldsymbol{x}\|_2 \|\boldsymbol{v}\|_2}$
- Length of projection
  $p = \|\boldsymbol{x}\|_2 \cos(\phi) = \frac{\boldsymbol{v}^\top \boldsymbol{x}}{\|\boldsymbol{v}\|_2}$
- Projection onto the unit vector
  $p = \boldsymbol{v}^\top \boldsymbol{x}$

Projection onto a subspace

- **Projection onto a subspace**
  - Subspace of dimension n defined by a orthonormal basis matrix $\boldsymbol{V}$.
  - Projection of $\boldsymbol{x}$ (M dimensional) onto $\boldsymbol{V}$ is given by

$$\boldsymbol{b}^\top = \boldsymbol{x}^\top \boldsymbol{V}$$

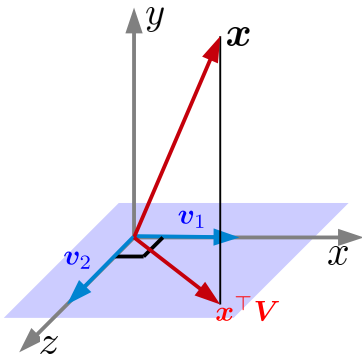
  - 'Reconstruction' can be found as:

$$\boldsymbol{x}' = \boldsymbol{V}\boldsymbol{b}$$

Example I:

$$\boldsymbol{V} = \begin{bmatrix} 1 & 0 \\ 0 & 0 \\ 0 & 1 \end{bmatrix} \qquad \boldsymbol{x} = \begin{bmatrix} x \\ y \\ z \end{bmatrix}$$

$$\boldsymbol{x}^\top \boldsymbol{V} = \begin{bmatrix} x & y & z \end{bmatrix} \begin{bmatrix} 1 & 0 \\ 0 & 0 \\ 0 & 1 \end{bmatrix} = \begin{bmatrix} x & z \end{bmatrix}$$



Projection onto a subspace

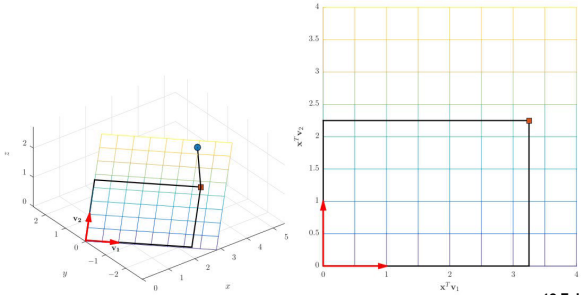
- **Projection onto a subspace**
  - Subspace of dimension n defined by a orthonormal basis matrix $\boldsymbol{V}$.
  - Projection of $\boldsymbol{x}$ (M dimensional) onto $\boldsymbol{V}$ given by

$$\boldsymbol{b}^\top = \boldsymbol{x}^\top \boldsymbol{V}$$

  - 'Reconstruction' can be found as:

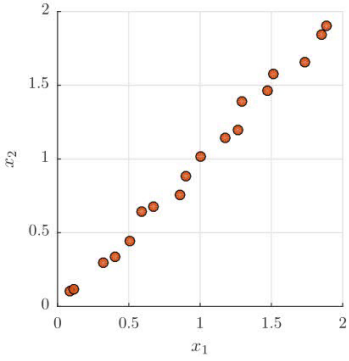
$$\boldsymbol{x}' = \boldsymbol{V}\boldsymbol{b}$$

Example II:



PCA for high-dimensional data

- Much data is high-dimensional
- We want to find a **lower** -dimensional representation of the **high**-dimensional data.



(2 dimensional but really (close to being) 1 dimensional)

Eigenvectors and eigenvalues

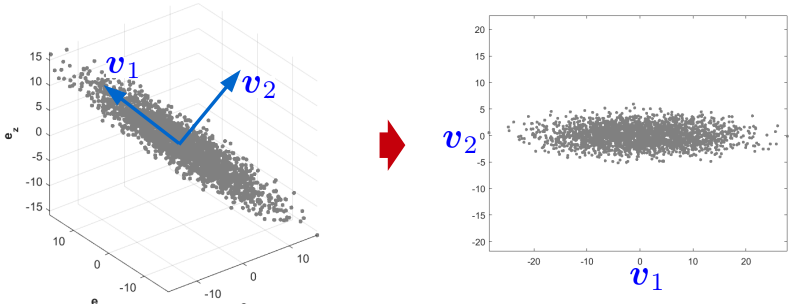
- Suppose

$$\boldsymbol{A}\boldsymbol{v} = \lambda \boldsymbol{v}, \qquad \boldsymbol{A} \text{ is a } N \times N \text{ matrix}$$

  - We say $\boldsymbol{v}$ is an eigenvector with eigenvalue λ
  - If $\boldsymbol{A}$ is symmetric: $\boldsymbol{A} = \boldsymbol{A}^\top$ then $\boldsymbol{A}$ has orthogonal eigenvectors and eigenvalues that are real.

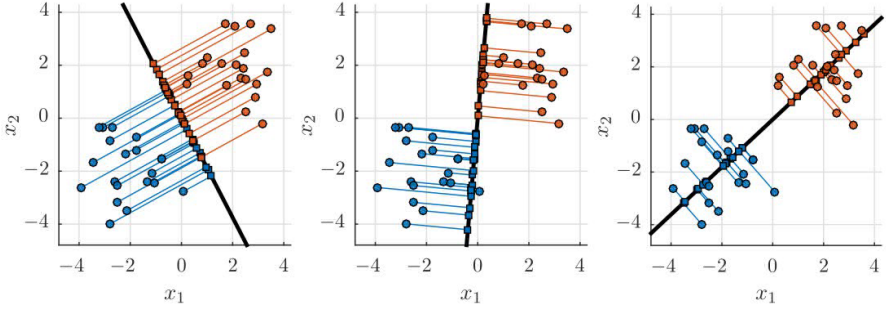
PCA for high-dimensional data

- Much data is high-dimensional
- We can project **high**-dimensional data to a **lower** -dimensional subspace.
- But what is a good projection?



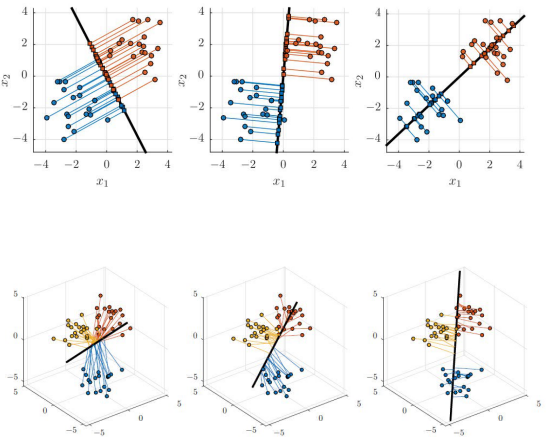
PCA for high-dimensional data

- Much data is high-dimensional
- We can project **high**-dimensional data to a **lower** -dimensional subspace.
- But what is a good projection?
- **Idea**: Select a projection that maximizes the variance of the projected data



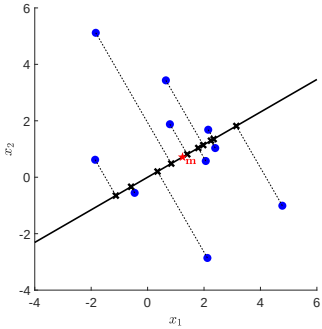
PCA derivation

- Key idea: Find the direction in the M'th dimensional space that captures the most variance.
- Approach:
  1. Measure variance when we project on a given direction given by a unit vector $\boldsymbol{v}$. The projection of a datapoint is $\boldsymbol{x}_i$, is $\boldsymbol{b}_i = \boldsymbol{x}_i^\top \boldsymbol{v}$.
  2. Find the direction $\boldsymbol{v}$ that maximizes the variance of the projected data points.



Determining the variance of a projection

$$
\begin{aligned}
Var[b] &= \frac{1}{N-1}\sum_{i=1}^{N}\left[b_i - \frac{1}{N}\sum_{j}^{N} b_j\right]^2 \\
&= \frac{1}{N-1}\sum_{i=1}^{N}\left[\boldsymbol{x}_i^\top \boldsymbol{v} - \frac{1}{N}\sum_{j}^{N} \boldsymbol{x}_j^\top \boldsymbol{v}\right]^2 \\
&= \frac{1}{N-1}\sum_{i=1}^{N}\left[\left(\boldsymbol{x}_i^\top - \frac{1}{N}\sum_{j}^{N} \boldsymbol{x}_j^\top\right)\boldsymbol{v}\right]^2 \\
&= \frac{1}{N-1}\sum_{i=1}^{N}(\tilde{\boldsymbol{x}}_{\boldsymbol{i}}{}^\top \boldsymbol{v})^2 \qquad \boxed{\tilde{\boldsymbol{x}}_i = \boldsymbol{x}_i - \boldsymbol{m}} \\
&= \frac{1}{N-1}\sum_{i=1}^{N}\boldsymbol{v}^\top \tilde{\boldsymbol{x}}_{\boldsymbol{i}}{}^\top \tilde{\boldsymbol{x}}_{\boldsymbol{i}}\boldsymbol{v} \\
&= \frac{1}{N-1}\boldsymbol{v}^\top \tilde{\boldsymbol{X}}^\top \tilde{\boldsymbol{X}}\boldsymbol{v}
\end{aligned}
$$



PCA derivation: Solving the optimization problem

$$\arg\max_{\boldsymbol{v}} \frac{1}{N-1}\boldsymbol{v}^\top \tilde{\boldsymbol{X}}^\top \tilde{\boldsymbol{X}}\boldsymbol{v} \;\; s.t. \;\; \|\boldsymbol{v}\|^2 = \boldsymbol{v}^\top \boldsymbol{v} = 1$$

- Use tricks from constrained optimization specifically to formulate "the Lagrangian". λ is the Lagrangian multiplier.

$$\arg\max_{\boldsymbol{v},\lambda} = \frac{1}{N-1}\boldsymbol{v}^\top \tilde{\boldsymbol{X}}^\top \tilde{\boldsymbol{X}}\boldsymbol{v} + \lambda(1 - \boldsymbol{v}^\top \boldsymbol{v})$$

- The function we want to maximize is now:

$$\mathcal{L}(\boldsymbol{v},\lambda) = \frac{1}{N-1}\boldsymbol{v}^T \tilde{\boldsymbol{X}}^T \tilde{\boldsymbol{X}}\boldsymbol{v} + \lambda(1 - \boldsymbol{v}^T \boldsymbol{v})$$

- Taking the derivative and equating to 0:

$$\frac{\partial \mathcal{L}}{\partial \boldsymbol{v}} = \frac{1}{N-1} 2\tilde{\boldsymbol{X}}^T \tilde{\boldsymbol{X}}\boldsymbol{v} - 2\lambda\boldsymbol{v} = 0 \qquad \frac{\partial \mathcal{L}}{\partial \lambda} = 1 - \boldsymbol{v}^\top \boldsymbol{v} = 0 \;\; \text{(not very informative)}$$

All-in-all we obtain:

$$\frac{1}{N-1}\tilde{\boldsymbol{X}}^T \tilde{\boldsymbol{X}}\boldsymbol{v} = \lambda\boldsymbol{v}$$

An eigenvalue problem of the form $Ax = \lambda x$.

PCA derivation: What about λ?

$$\frac{\partial \mathcal{L}}{\partial \boldsymbol{v}} = \frac{1}{N-1} 2\tilde{\boldsymbol{X}}^T \tilde{\boldsymbol{X}} \boldsymbol{v} - 2\lambda \boldsymbol{v} = 0 \;\; or \;\; \tilde{\boldsymbol{X}}^T \tilde{\boldsymbol{X}} \boldsymbol{v} = \lambda \boldsymbol{v}$$

$$\begin{aligned} Var[b] &= \frac{1}{N-1} \boldsymbol{v}^\top \tilde{\boldsymbol{X}}^\top \tilde{\boldsymbol{X}} \boldsymbol{v} \quad \text{(we know } \frac{1}{N-1} \tilde{\boldsymbol{X}}^T \tilde{\boldsymbol{X}} \boldsymbol{v} = \lambda \boldsymbol{v}) \\ &= \boldsymbol{v}^\top \lambda \boldsymbol{v} \\ &= \lambda \end{aligned}$$

To find the eigenvector (i.e. the principal component) of $\frac{1}{N-1} \tilde{\boldsymbol{X}}^\top \tilde{\boldsymbol{X}}$ that maximizes the variance, simply select the eigenvector with the largest corresponding eigenvalue λ.

PCA derivation: What is the matrix $\frac{1}{N-1}\tilde{\boldsymbol{X}}^T\tilde{\boldsymbol{X}}$?

Recall the empirical co-variance:

$$\hat{\text{cov}}[\boldsymbol{x}^{(l)}, \boldsymbol{x}^{(k)}] = \frac{1}{N-1}\sum_{i=1}^{N}(x_i^{(l)} - \hat{\mu}_{x^{(l)}})(x_i^{(k)} - \hat{\mu}_{x^{(k)}})$$

And the covariance matrix:

$$\hat{\boldsymbol{S}} = \begin{bmatrix} \hat{\text{cov}}[\boldsymbol{x}^{(1)}, \boldsymbol{x}^{(1)}] & \dots & \hat{\text{cov}}[\boldsymbol{x}^{(1)}, \boldsymbol{x}^{(M)}] \\ \vdots & & \vdots \\ \hat{\text{cov}}[\boldsymbol{x}^{(M)}, \boldsymbol{x}^{(1)}] & \dots & \hat{\text{cov}}[\boldsymbol{x}^{(M)}, \boldsymbol{x}^{(M)}] \end{bmatrix}$$

This can conviniently be computed as:

$$\hat{\boldsymbol{S}} = \frac{1}{N-1}\tilde{\boldsymbol{X}}^T\tilde{\boldsymbol{X}}$$

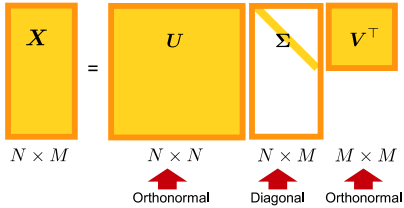
where $\tilde{\boldsymbol{X}}$ is the centered data matrix as usual.

A startegy: Find the eigenvectors and eigenvalues of $\frac{1}{N-1}\tilde{\boldsymbol{X}}^T\tilde{\boldsymbol{X}}$ and pick the eigenvector with the highest eigenvalue...

The Singular Value Decomposition (SVD)

Any matrix can be decomposed as follows:

$$\boldsymbol{X} = \boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^\top$$



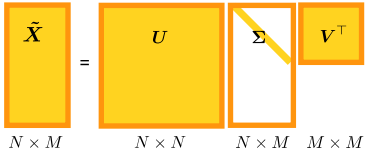
$$\boldsymbol{U} = \begin{bmatrix} \boldsymbol{u}_1, \boldsymbol{u}_2 \dots \boldsymbol{u}_N \end{bmatrix}, \quad \boldsymbol{V} = \begin{bmatrix} | & | & & | \\ \boldsymbol{v}_1 & \boldsymbol{v}_2 & \cdots & \boldsymbol{v}_M \\ | & | & & | \end{bmatrix}, \boldsymbol{\Sigma} = \begin{bmatrix} \sigma_1 & 0 & \dots & 0 \\ 0 & \sigma_2 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & \sigma_M \\ 0 & 0 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & 0 \end{bmatrix}$$

$$\boldsymbol{U}^\top\boldsymbol{U} = \boldsymbol{I}_{N\times N}, \quad \boldsymbol{V}^\top\boldsymbol{V} = \boldsymbol{I}_{M\times M},$$

The Singular Value Decomposition (SVD)

Any matrix can be decomposed as follows (here the centered matrix):

$$\tilde{\boldsymbol{X}} = \boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^\top$$



$$\boldsymbol{\Sigma} = \begin{bmatrix} \sigma_1 & 0 & \dots & 0 \\ 0 & \sigma_2 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & \sigma_M \\ 0 & 0 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & 0 \end{bmatrix}$$

$$\boldsymbol{V}^\top = \begin{bmatrix} - & \boldsymbol{v}_1 & - \\ - & \boldsymbol{v}_2 & - \\ & \vdots & \\ - & \boldsymbol{v}_M & - \end{bmatrix}$$

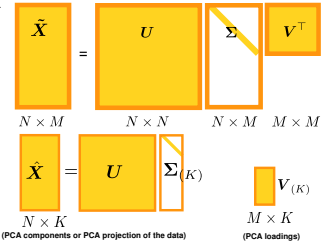
$$\begin{aligned} \frac{1}{N-1}(\tilde{\boldsymbol{X}}^\top \tilde{\boldsymbol{X}})\boldsymbol{v}_i &= \frac{1}{N-1}(\boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^\top)^\top \boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^\top \boldsymbol{v}_i \\ &= \frac{1}{N-1}\boldsymbol{V}\boldsymbol{\Sigma}^\top \boldsymbol{U}^\top \boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^\top \boldsymbol{v}_i \\ &= \frac{1}{N-1}\boldsymbol{V}\boldsymbol{\Sigma}^\top \boldsymbol{\Sigma}\boldsymbol{V}^\top \boldsymbol{v}_i \\ &= \frac{1}{N-1}\boldsymbol{V}\boldsymbol{\Sigma}^\top \boldsymbol{\Sigma}\boldsymbol{e}_i \\ &= \frac{1}{N-1}\sigma_{ii}^2 \boldsymbol{v}_i \end{aligned}$$

$$\boldsymbol{A}\boldsymbol{x} = \lambda \boldsymbol{x}$$

Principal component analysis (PCA)

(Karl Pearson, 1901)

1. Subtract the column-wise mean, $\boldsymbol{m}$, from the data matrix $\boldsymbol{X}$, such that $\tilde{\boldsymbol{x}}_{\boldsymbol{i}} = \boldsymbol{x}_i - \boldsymbol{m}$
   (nominal data can should be represented with the one-out-of-K representation)
2. Apply SVD to the centered matrix, i.e., $\tilde{\boldsymbol{X}} = \boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^\top$
3. Select the K first columns of $\boldsymbol{V}$,
   $\boldsymbol{V}_{(K)} = [\boldsymbol{v}_1 \quad \boldsymbol{v}_2 \quad \dots \quad \boldsymbol{v}_K]$, and first K columns of $\boldsymbol{\Sigma}$



4. Optional: Analyse the eigenvectors / principal components to see which of the original features contribute to the projection and how much.
5. Optional: Project the data onto the V_K basis and visualize the data in the new coordinate system.
6. Optional: Reconstruct any point in the new basis, $\boldsymbol{b}_i$ to the original, non-centered basis using $\boldsymbol{x}'_i = \boldsymbol{V}_{(K)}\boldsymbol{b}_i + \boldsymbol{m}$

PCA: What's really the role of Σ?

$$\tilde{\boldsymbol{X}} = \boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^\top$$

- Entries in the diagonal matrix $\boldsymbol{\Sigma}$ are called **singular values**
  - They are sorted (largest first)
  - Indicate how much variability is explained by the corresponding component
    - 1st component explains most of the variability
    - 2nd component explains most of the remaining variability
    - etc.

$$\boldsymbol{\Sigma} = \begin{bmatrix} \sigma_1 & 0 & \dots & 0 \\ 0 & \sigma_2 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & \sigma_M \\ 0 & 0 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & 0 \end{bmatrix} \quad (N \times M) \quad \sigma_1 \geq \sigma_2 \cdots \geq \sigma_M \geq 0$$

PCA: Explained Variance

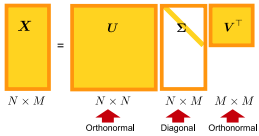
Recall that from SVD: $\tilde{\boldsymbol{X}} = \boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^T$

In the original but centered space, the coordinates of $\tilde{\boldsymbol{X}}$ project onto the first K components are:

$$\boldsymbol{X}' = \boldsymbol{U}\boldsymbol{\Sigma}_{(K)}\boldsymbol{V}^T_{(K)}$$

We can measure how much variance is retained in the reconstruction $\boldsymbol{X}'$:

$$\text{Explained var.} = \frac{\|\boldsymbol{X}'\|_F^2}{\|\tilde{\boldsymbol{X}}\|_F^2}$$



$$\text{cov}(\tilde{\boldsymbol{X}}) \propto \tilde{\boldsymbol{X}}^T\tilde{\boldsymbol{X}}$$

$$\text{trace}(\boldsymbol{A}\boldsymbol{B}) = \text{trace}(\boldsymbol{B}\boldsymbol{A})$$

$$\|\tilde{\boldsymbol{X}}\|_F^2 = \text{trace}(\tilde{\boldsymbol{X}}^T\tilde{\boldsymbol{X}}) = \text{trace}(\tilde{\boldsymbol{X}}\tilde{\boldsymbol{X}}^T)$$

PCA: Explained Variance

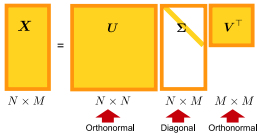
Recall that from SVD: $\tilde{\boldsymbol{X}} = \boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^T$

In the original but centered space, the coordinates of $\tilde{\boldsymbol{X}}$ project onto the first K components are:

$$\boldsymbol{X}' = \boldsymbol{U}\boldsymbol{\Sigma}_{(K)}\boldsymbol{V}^T_{(K)}$$

We can measure how much variance is retained in the reconstruction $\boldsymbol{X}'$:

$$\text{Explained var.} = \frac{\|\boldsymbol{X}'\|_F^2}{\|\tilde{\boldsymbol{X}}\|_F^2} = \frac{\sum_{i=1}^{K}\sigma_i^2}{\sum_{i=1}^{M}\sigma_i^2}$$



$$\text{cov}(\tilde{\boldsymbol{X}}) \propto \tilde{\boldsymbol{X}}^T\tilde{\boldsymbol{X}}$$

$$\text{trace}(\boldsymbol{A}\boldsymbol{B}) = \text{trace}(\boldsymbol{B}\boldsymbol{A})$$

$$\begin{aligned}
\|\tilde{\boldsymbol{X}}\|_F^2 &= \text{trace}(\tilde{\boldsymbol{X}}^T\tilde{\boldsymbol{X}}) = \text{trace}(\tilde{\boldsymbol{X}}\tilde{\boldsymbol{X}}^T) \\
&= \text{trace}(\boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^T(\boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^T)^T) \\
&= \text{trace}(\boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^T\boldsymbol{V}\boldsymbol{\Sigma}^T\boldsymbol{U}^T) \\
&= \text{trace}(\boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{\Sigma}^T\boldsymbol{U}^T) \\
&= \text{trace}(\boldsymbol{U}^T\boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{\Sigma}^T) \\
&= \text{trace}(\boldsymbol{\Sigma}\boldsymbol{\Sigma}^T) = \sum_i \sigma_i^2
\end{aligned}$$

PCA: Explained Variance

Recall that from SVD: $\tilde{\boldsymbol{X}} = \boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^T$

In the original but centered space, the coordinates of $\tilde{\boldsymbol{X}}$ project onto the first K components are:

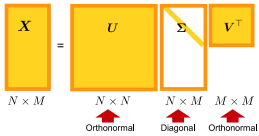
$$\boldsymbol{X}' = \boldsymbol{U}\boldsymbol{\Sigma}_{(K)}\boldsymbol{V}^T_{(K)}$$

We can measure how much variance is retained in the reconstruction $\boldsymbol{X}'$:

$$\text{Explained var.} = \frac{\|\boldsymbol{X}'\|_F^2}{\|\tilde{\boldsymbol{X}}\|_F^2} = \frac{\sum_{i=1}^K \sigma_i^2}{\sum_{i=1}^M \sigma_i^2}$$

Similarly, the fraction of explained variance for the i'th component is

$$\text{Explained var.} = \frac{\sigma_i^2}{\sum_{i=1}^M \sigma_i^2}$$



$$\text{cov}(\tilde{\boldsymbol{X}}) \propto \tilde{\boldsymbol{X}}^T\tilde{\boldsymbol{X}}$$

$$\text{trace}(\boldsymbol{A}\boldsymbol{B}) = \text{trace}(\boldsymbol{B}\boldsymbol{A})$$

$$\begin{aligned}
\|\tilde{\boldsymbol{X}}\|_F^2 &= \text{trace}(\tilde{\boldsymbol{X}}^T\tilde{\boldsymbol{X}}) = \text{trace}(\tilde{\boldsymbol{X}}\tilde{\boldsymbol{X}}^T) \\
&= \text{trace}(\boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^T(\boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^T)^T) \\
&= \text{trace}(\boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^T\boldsymbol{V}\boldsymbol{\Sigma}^T\boldsymbol{U}^T) \\
&= \text{trace}(\boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{\Sigma}^T\boldsymbol{U}^T) \\
&= \text{trace}(\boldsymbol{U}^T\boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{\Sigma}^T) \\
&= \text{trace}(\boldsymbol{\Sigma}\boldsymbol{\Sigma}^T) = \sum_i \sigma_i^2
\end{aligned}$$

Quiz 1: PCA

No.	Attribute description	Abbrev.
x_1	Age (in years)	AGE
x_2	Gender (Female=0, Male=1)	GDR
x_3	Total Bilirubin	TB
x_4	Direct Bilirubin	DB
x_5	Alkaline Phosphotase	AP
x_6	Alamine Aminotransferase	AlA
x_7	Aspartate Aminotransferase	AsA
x_8	Total Protiens	TP
x_9	Albumin	AB
x_{10}	Albumin to Globulin ratio	A/G
y	0=No liver disease, 1=Liver disease	LD

Table 1: Attributes in a study on liver disease among Indians living in the north eastern part of Andhra Pradesh, India. (taken from http://archive.ics.uci.edu/ml/datasets/ILPD +%28Indian+Liver+Patient+Dataset%29). The data has 10 input attributes x_1–x_{10} and one output variable y which defines whether the subject considered has a liver disease ($y = 1$) or not ($y = 0$). x_3–x_9 are non-negative measurements giving the concentrations of various quantities measured in a blood test. x_{10} gives the ratio of Albumin to Globulin in the blood.

A PCA analysis is applied to the standardized data based on the attributes x_1–x_{10}. The squared Frobenius norm of the standardized data matrix $\boldsymbol{X}$ is given by $\|\boldsymbol{X}\|_F^2 = 5780.0$. The first four singular values are $\sigma_1 = 40.1$, $\sigma_2 = 34.2$, $\sigma_3 = 28.1$, and $\sigma_4 = 24.8$, Which of the following statements is *correct*?

A. The first PCA component accounts for more than 35 % of the variation.

B. The second PCA component accounts for more than 30 % of the variation.

C. The first three PCA components account for less than 70 % of the variation in the data.

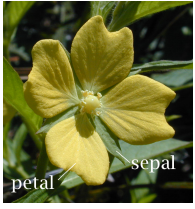
D. The fourth PCA component accounts for less than 10 % of the variation in the data.

E. Don't know.

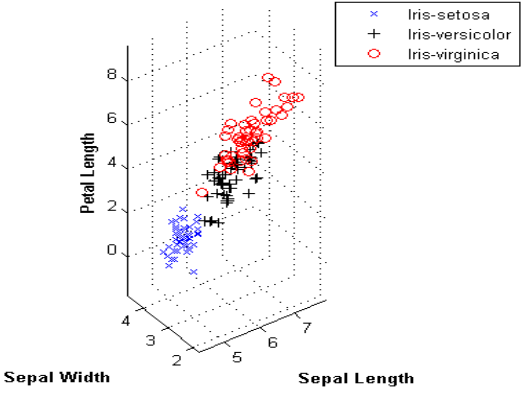
Example: PCA on Fisher's Iris Data

- We will presently consider the first 3 attributes, i.e. Sepal length, Sepal Width and Petal Length.
- Three types of flowers: Iris Setosa, Iris Versicolor, Iris Virginica

Flower ID	Attribute			
	Sepal Length	Sepal Width	Petal Length	Petal Width
1	5.1	3.5	1.4	0.2
2	4.9	3.0	1.4	0.2
3	4.7	3.2	1.3	0.2
4	4.6	3.1	1.5	0.2
.	.	.	.	.
.	.	.	.	.
150	5.9	3.0	5.1	1.8

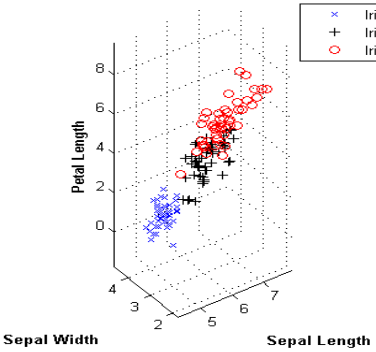


3D scatter plot of Iris Data



What fraction of the total variation in the data will the first principal component account for?

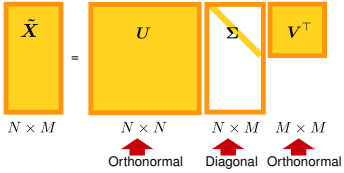
3D scatter plot of Iris Data



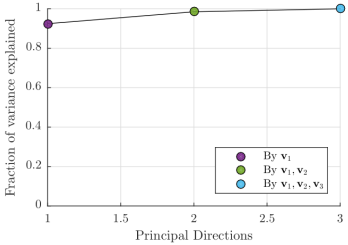
What fraction of the total variation in the data will the first principal component account for?

1) Subtract the mean
2) Apply singular value decomposition (SVD)

$$\tilde{\boldsymbol{X}} = \boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^{\top}$$

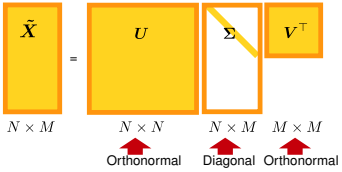


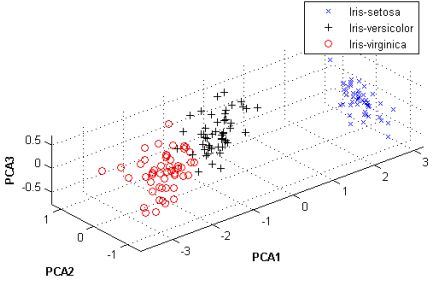
Evaluate the singular values to determine how much of the dynamics is lost when reducing the dimensionality



Visualization of the PCA projections of the data

$$\tilde{\boldsymbol{X}} = \boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^\top$$

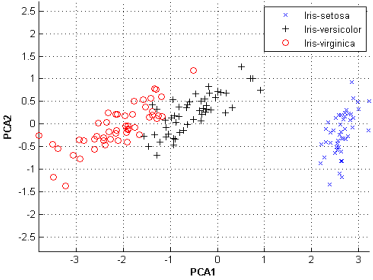




$$PCA1 \quad : \quad \boldsymbol{b}_1 = \tilde{\boldsymbol{X}}\boldsymbol{v}_1 = \boldsymbol{u}_1\sigma_1$$

$$PCA2 \quad : \quad \boldsymbol{b}_2 = \tilde{\boldsymbol{X}}\boldsymbol{v}_2 = \boldsymbol{u}_2\sigma_2$$

$$PCA3 \quad : \quad \boldsymbol{b}_3 = \tilde{\boldsymbol{X}}\boldsymbol{v}_3 = \boldsymbol{u}_3\sigma_3$$



The principal directions V

Sepal Length
Sepal Width
Petal Length

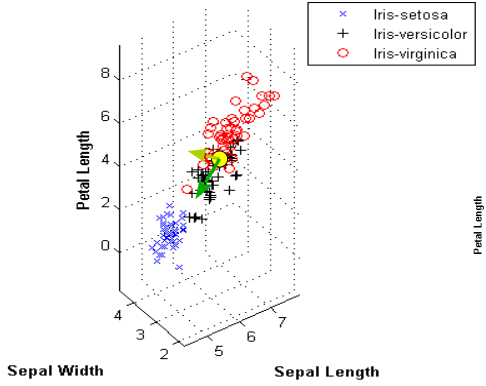
$$\boldsymbol{V} = \begin{bmatrix} -0.39 & -0.64 & -0.66 \\ 0.09 & -0.74 & 0.66 \\ -0.92 & 0.20 & 0.35 \end{bmatrix}$$

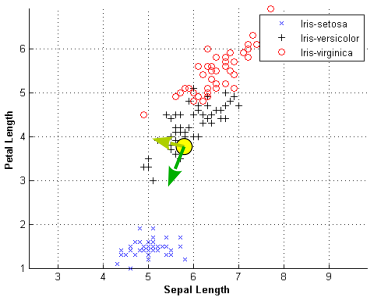
$$\boldsymbol{\mu} = \begin{bmatrix} 5.8 \\ 3.1 \\ 3.8 \end{bmatrix}, \quad \boldsymbol{v}_1 = \begin{bmatrix} -0.39 \\ 0.09 \\ -0.92 \end{bmatrix}, \quad \boldsymbol{v}_2 = \begin{bmatrix} -0.64 \\ -0.74 \\ 0.20 \end{bmatrix}$$

Sepal Length

Sepal Width

Petal Length





Quiz 2: PCA

No.	Attribute description	Abbrev.
x_1	Age (in years)	AGE
x_2	Gender (Female=0, Male=1)	GDR
x_3	Total Bilirubin	TB
x_4	Direct Bilirubin	DB
x_5	Alkaline Phosphotase	AP
x_6	Alamine Aminotransferase	AlA
x_7	Aspartate Aminotransferase	AsA
x_8	Total Protiens	TP
x_9	Albumin	AB
x_{10}	Albumin to Globulin ratio	A/G
y	0=No liver disease, 1=Liver disease	LD

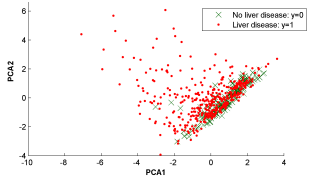


Figure 1: Principal component 1 (PCA1) plotted against principal component 2 (PCA2).

The first and second principal component directions of the liver-dataset are

$$\boldsymbol{v}_1 = \begin{bmatrix} -0.1404 \\ -0.1090 \\ -0.4115 \\ -0.4179 \\ -0.2468 \\ -0.2682 \\ -0.3009 \\ 0.2781 \\ 0.4375 \\ 0.3638 \end{bmatrix}, \quad \boldsymbol{v}_2 = \begin{bmatrix} -0.2859 \\ 0.0130 \\ 0.2510 \\ 0.2622 \\ 0.0525 \\ 0.4162 \\ 0.3927 \\ 0.4197 \\ 0.4323 \\ 0.3052 \end{bmatrix}.$$

In the figure, the data projected onto the first two principal components is plotted, and the colors indicate the presence of liver disease. Which of the following statements is *correct*?

A. Relatively high values of AGE, GDR, TB, DB, AP, AlA, and AsA and low values of TP, AB, and A/G will result in a positive projection onto the first principal component.

B. Relatively low values of the projection onto PCA1 and high values of the projection onto PCA2 indicates the subject does not have a liver disease.

C. PCA2 mainly discriminate between old subjects with low measurements of TB, DB, AlA, AsA, TP, AB, and A/G from young subjects with high values of TB, DB, AlA, AsA, TP, AB, and A/G.

D. The principal component directions are not guaranteed to be orthogonal to each other since the data has been standardized.

E. Don't know.

Example: Visualization of hand written digits

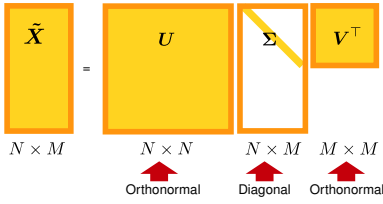
- Data matrix

$$\boldsymbol{X} = \begin{bmatrix} \text{—} & \boldsymbol{x}_1 & \text{—} \\ \text{—} & \boldsymbol{x}_2 & \text{—} \\ & \vdots & \\ \text{—} & \boldsymbol{x}_N & \text{—} \end{bmatrix}$$

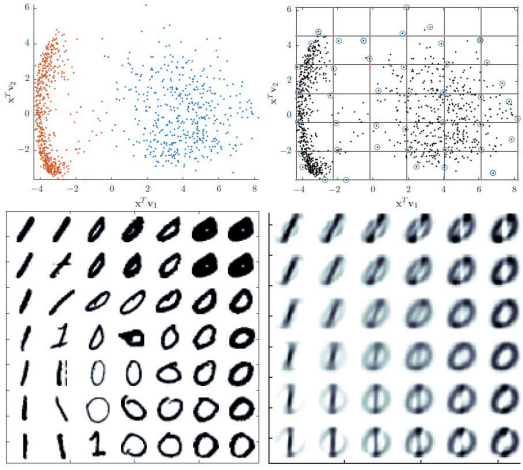
If each image is 28 x 28 pixels then X is a N x 784 matrix 2em

- Principal component analysis

$$\tilde{\boldsymbol{X}} = \boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^\top$$



Visualization of hand written digits

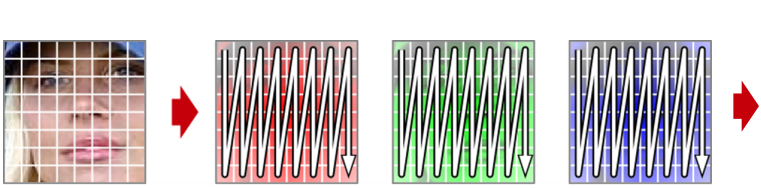


Example: Principal component analysis of faces

- 1000 images, 86 x 86 pixels, 3 RGB intensities

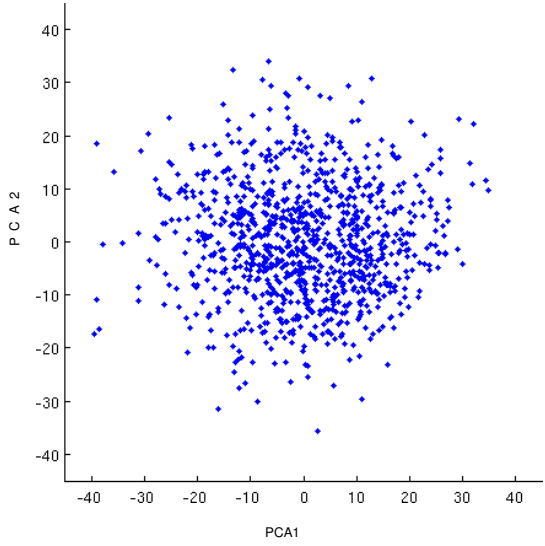
Tamara Berg: "Faces in the wild"

Pre-processing

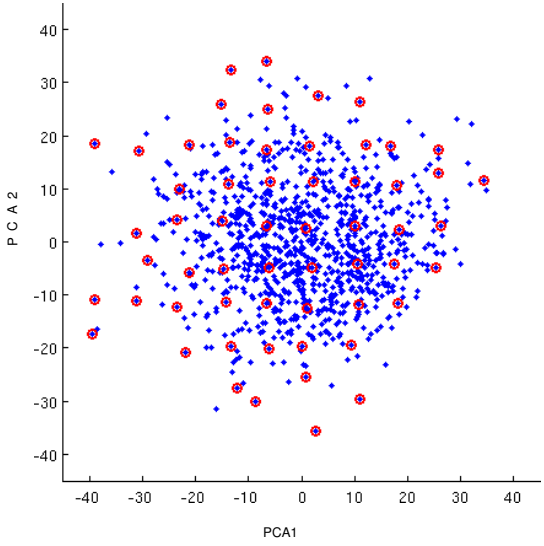


- Concatenate all pixel color values in one long vector
  - 86 x 86 x 3 = 22’188
  - Image is now represented as a 22’188 dimensional vector
- Stack all 1000 images into a big matrix
  - 1000 x 22’188

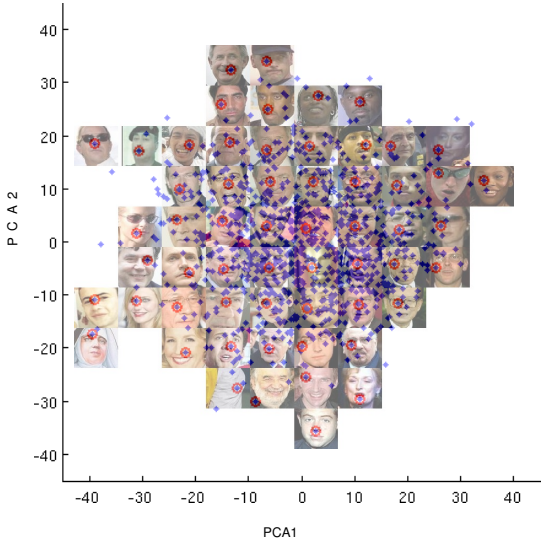
PCA on face images



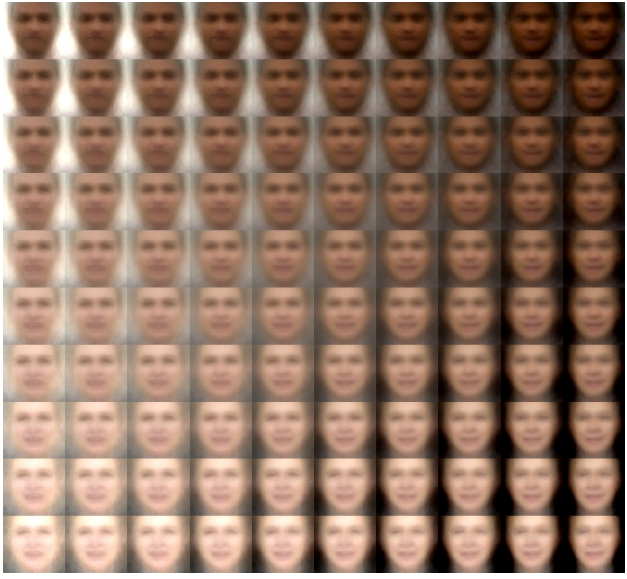
PCA on face images



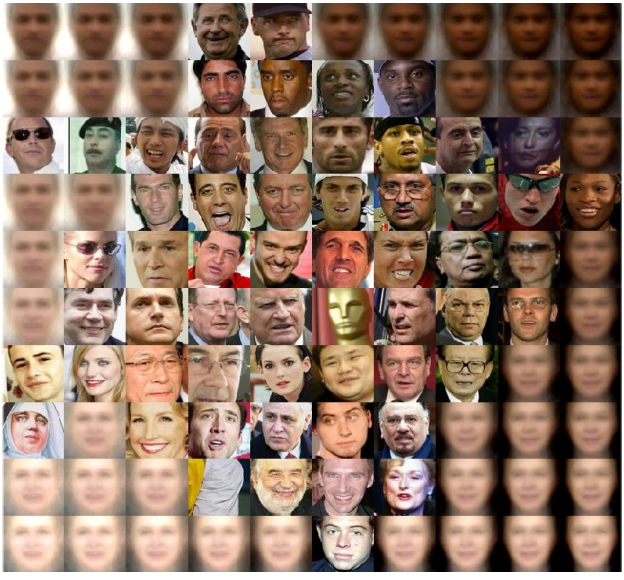
PCA on face images



- What information do the two principal axes capture?

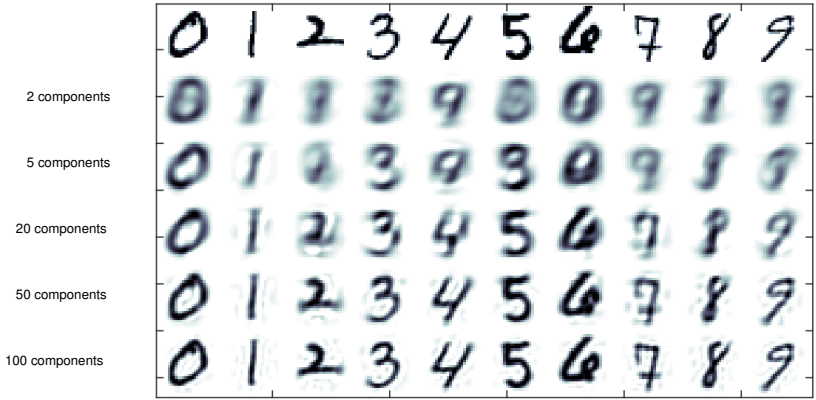


What information do the two principal axes capture?



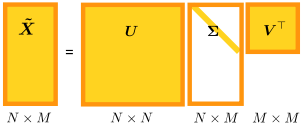
Example: PCA as compression

- Only include a few components: $\hat{\boldsymbol{x}} = \boldsymbol{V}_{(K)}\boldsymbol{b} + \boldsymbol{m} \quad K = \{2, 5, 20, 50, 100\}$

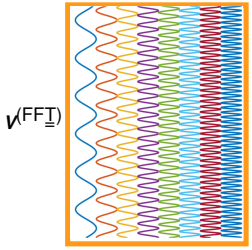


Data and domain driven feature extraction

PCA is an example of a data driven approach for feature extraction.
i.e., we define from data the features extracted in terms of the projections $\boldsymbol{V}^{(PCA)}$ that preserve most of the variance in the data.



The fourier transform is an example of a domain driven approach for feature extraction
i.e.,in the analysis of sound good features are often to use spectral representations. These can be found by projecting the data using the so-called fourier transform matrix $\boldsymbol{V}^{(FFT)}$ where the components are defined as specific frequencies such that the projection of the data onto these frequencies defines the extend to which these frequencies are present in the data.



Resources

DTU Compute Our online PCA demo which highlights key concepts of PCA such as the effect of normalization, variance explained, and much more (http://www2.imm.dtu.dk/courses/02450/DemoPCA.html)

https://arxiv.org A great and more in-depth tutorial on PCA

(https://arxiv.org/abs/1404.1100)

https://www.3blue1brown.com An great, animated recap of linear algebra

(https://www.3blue1brown.com/essence-of-linear-algebra-page/)