FOUNDATIONS OF STATISTICAL ANALYSIS & MACHINE LEARNING

Amric Trudel
amric.trudel@epita.fr

COURSE PROGRAM

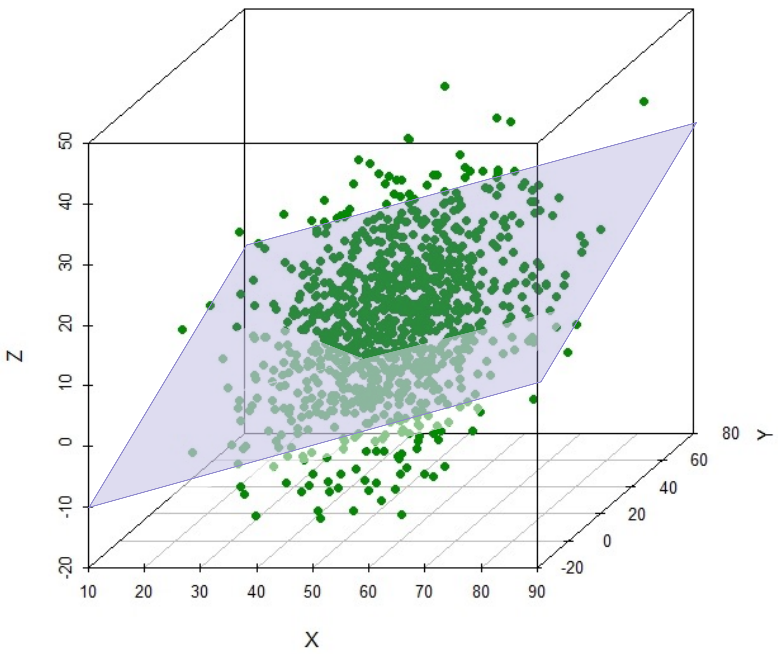
Structure

PREPARATION	Data exploration
	Data preprocessing
REGRESSION	Linear regression with one variable
	Multiple and polynomial regression
CLASSIFICATION	Logistic regression
	Classification model assessment
	k-NN, Decision Tree, SVM
CLUSTERING	k-means, hierarchical clustering
DIMENSIONALITY REDUCTION	**Principal Components Analysis**
ALL NOTIONS	Final assignment

DIMENSIONALITY REDUCTION

Objectives

- Get intuition on the data set
- Better understand the relationship between X and Y
- Limit to a smaller relevant subspace
- Escape the curse of dimensionality
- Speed up training on large datasets
- Visualize decision regions and boundaries on a 2D plane



DIMENSIONALITY REDUCTION

Methods

- **Feature Selection:** selection among the existing features
  - Selected features remain interpretable
  - Risk of losing information with deleted features
  - Features are usually not completely uncorrelated
- **Feature Extraction:** combination of existing features
  - Extracted features are not easily interpretable
  - Insures that the k first extracted features hold the most information

PRINCIPAL COMPONENTS ANALYSIS

Principle

Data set with a large number of interrelated variables

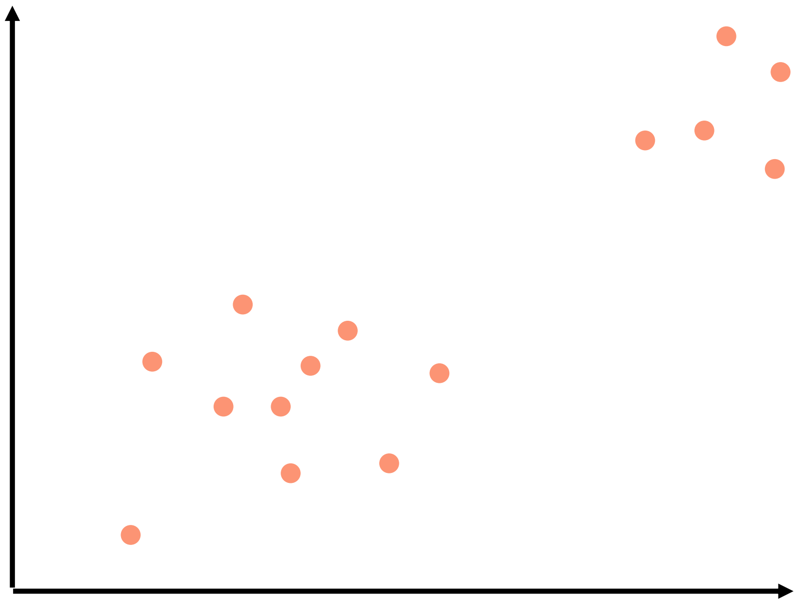
Projection

New set with a small number of uncorrelated variables (called principal components) retaining as much as possible of the variation of the original data set

PRINCIPAL COMPONENTS ANALYSIS

Principle

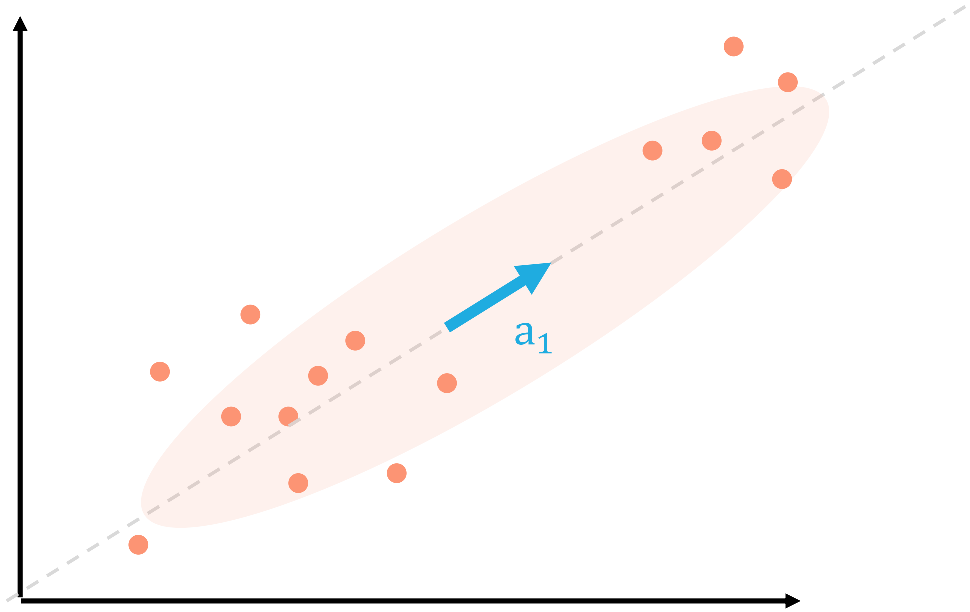
- The data lies in a **space** defined its features



PRINCIPAL COMPONENTS ANALYSIS

Principle

- Find the axis that maximizes the variance of the dataset

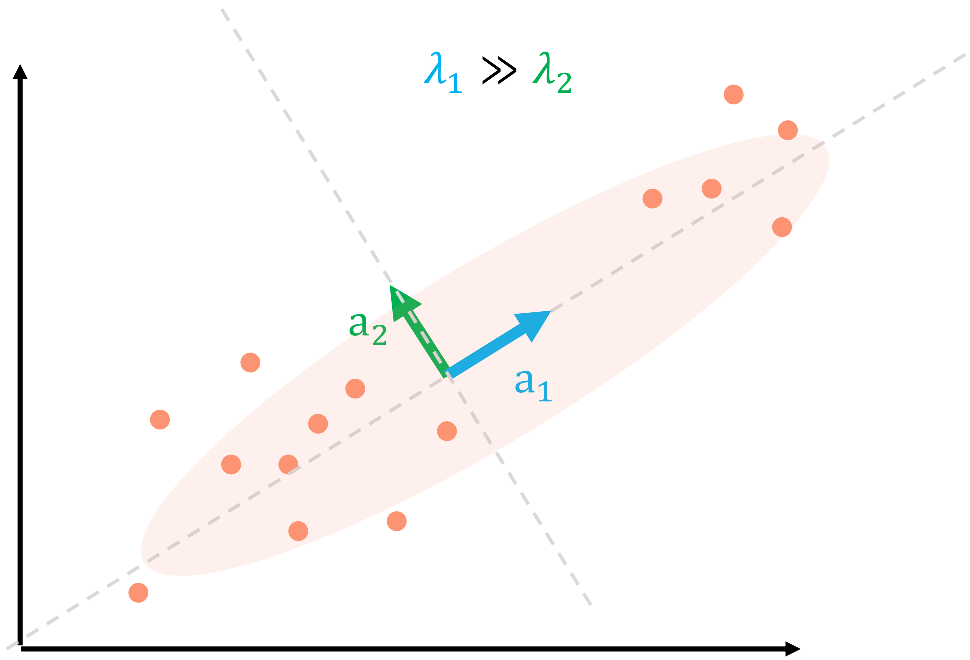


λ_1: variance along $\mathbf{a_1}$

PRINCIPAL COMPONENTS ANALYSIS

Principle

- Find a second axis that is **orthogonal** to the first one and that maximizes the variance of the data set.

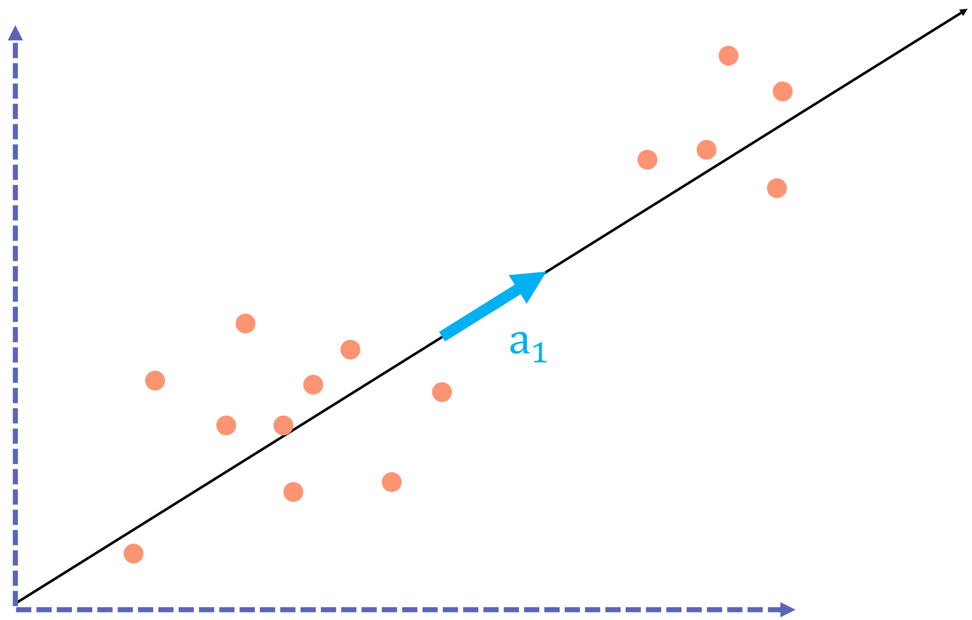


λ_1: variance along $\mathbf{a_1}$

λ_2: variance along $\mathbf{a_2}$

Principle

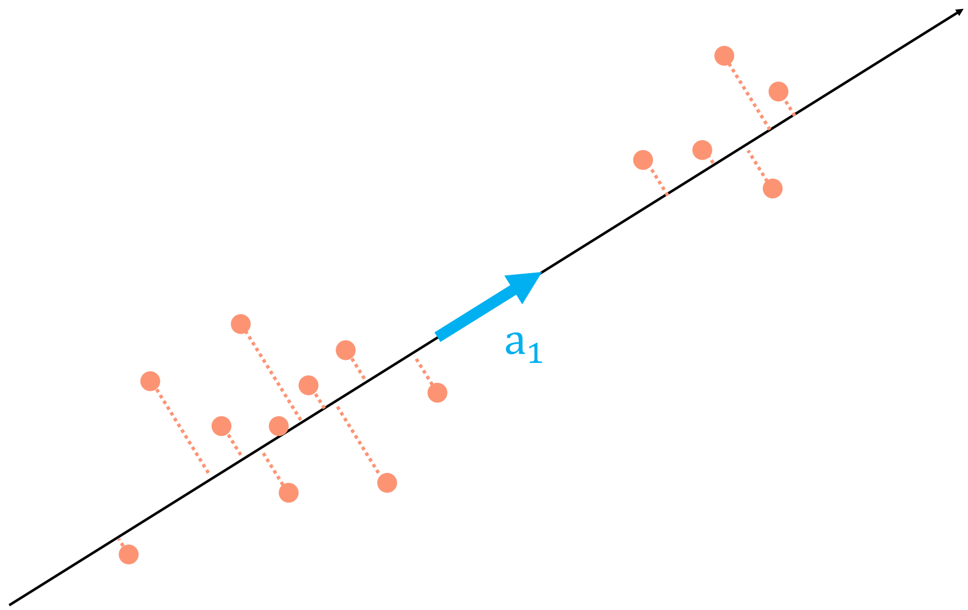
- The derived axes are called **Principal components**. You can select a subset of the principal compents for your dimensionality reduction.



PRINCIPAL COMPONENTS ANALYSIS

Principle

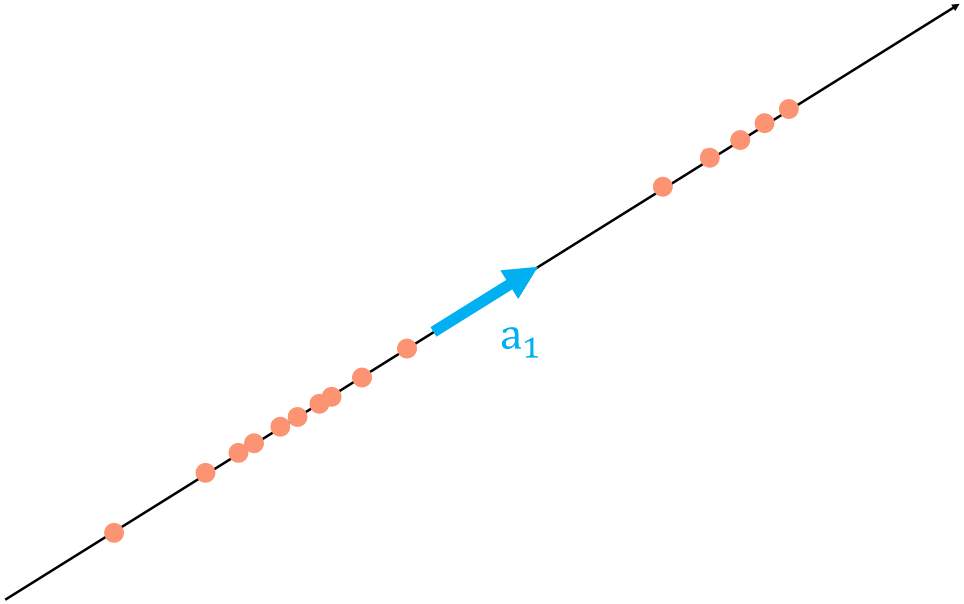
- Project the data in the subspace generated by the principal components you selected



PRINCIPAL COMPONENTS ANALYSIS

Principle

- Project the data in the subspace generated by the principal components you selected



PRINCIPAL COMPONENTS ANALYSIS

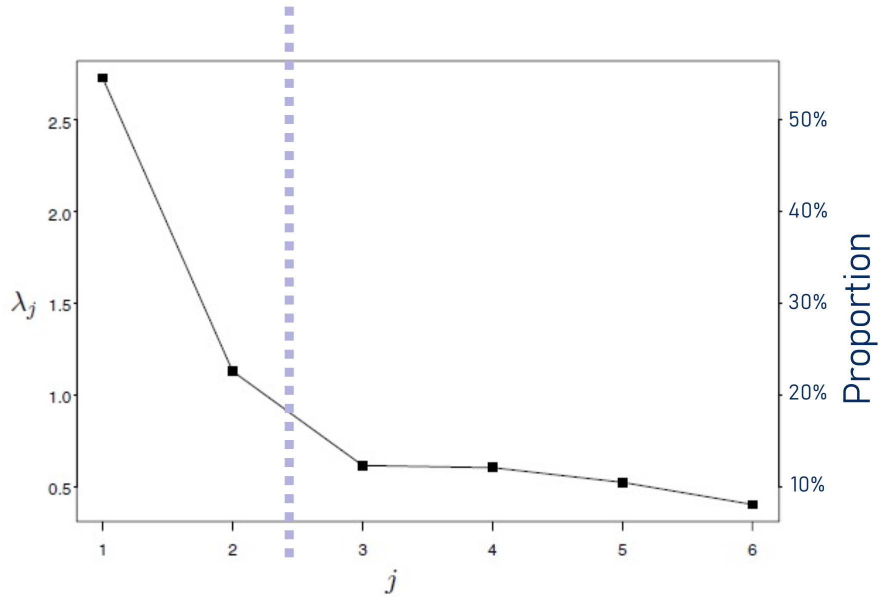
Process

- Compute the **eigenvalues** $(\lambda_i)_{i=1..p}$ and **eigenvectors** $(\mathrm{a}_i)_{i=1..p}$ from the **covariance matrix** of the data set $X_{(p)}$: $\Sigma = \sum_{i=1}^{p} \lambda_i \mathrm{a}_i {\mathrm{a}_i}^T$
- Sort the eigenvalues in descending order: $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_p$. These elements are the coefficients of the **principal components**.
- Select the k eigenvectors that correspond to the k largest eigenvalues: $\Sigma \cong \sum_{i=1}^{k} \lambda_i \mathrm{a}_i {\mathrm{a}_i}^T$
- Construct the **transfer matrix** A from the k selected eigenvectors and use it to project the original data set in the k-dimensional subspace: $\hat{X}_{(k)} = \mathrm{A} X_{(p)}$

PRINCIPAL COMPONENTS ANALYSIS

Process

- Choice of k: based on proportion of variation explained by each principal component $\frac{\lambda_i}{\lambda_1+\lambda_2+\cdots+\lambda_p}$



PRINCIPAL COMPONENTS ANALYSIS

Python implementation

- Training a PCA:

```
from sklearn.decomposition import PCA
pca = PCA(n_components = 2)
pca.fit(X)
```

- Projecting the data along the principal components:

```
X_pca = pca.transform(X)
```

- Getting the sorted eigenvalues ratio and eigenvectors associated to each principal component:

```
pca.explained_variance_ratio_
pca.components_
```

PRINCIPAL COMPONENTS ANALYSIS

Implementation

- Data set: iris data set (characteristics of three species of iris)
- Objectives:
  - Apply a PCA
  - Check the eigenvalues
  - Project the data along the two main components