Machine Learning for Big Data: Principal Component Analysis

Lionel Fillatre

fillatre@unice.fr

Examen partiel: jeudi 10 novembre

- 1 heure en TD
- Uniquement sur papier
- Sujet en français
- Questions de cours + exercices à résoudre
- Aucune question Python
- A priori, pas de calculatrice.

Topics

- Introduction
- How does it work?
- PCA
- Practical interpretation
- Conclusion

1 Introduction

PCA Overview

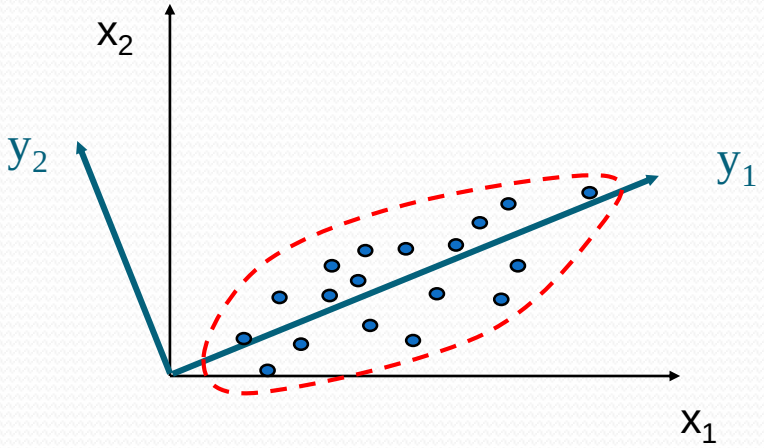
- It is a mathematical tool from applied linear algebra.
- It is a simple, non-parametric method of extracting relevant information from confusing datasets.
- It provides a roadmap for how to reduce a complex data set to a lower dimension.
- We typically have a data matrix of *p* observations on *N* correlated variables $x_1, x_2, \cdots, x_N$.
- PCA looks for a transformation of the *N* variables x_i into *N* new variables y_i that are uncorrelated

Matrix of data $D = (d_{ij})$

	x_1	x_2	...	x_N
Sample 1 d_1	d_{11}	d_{12}		d_{1N}
Sample 2 d_2	d_{21}	d_{22}		d_{2N}
Sample p d_p	d_{p1}	d_{p2}		d_{pN}
	Column 1 d^1	Column 2 d^2		Column N d^N

Dimensionality Reduction

- Find a projection that captures the largest amount of variation in data
- Find the eigenvectors of the covariance matrix, and these eigenvectors define the new space



Eigenvalues and Eigenvectors

- Eigenvectors (for a square $N \times N$ matrix S)

$$\mathbf{S}\mathbf{v} = \lambda\mathbf{v}$$

(right) eigenvector $\mathbf{v} \in \mathbb{R}^N \neq 0$

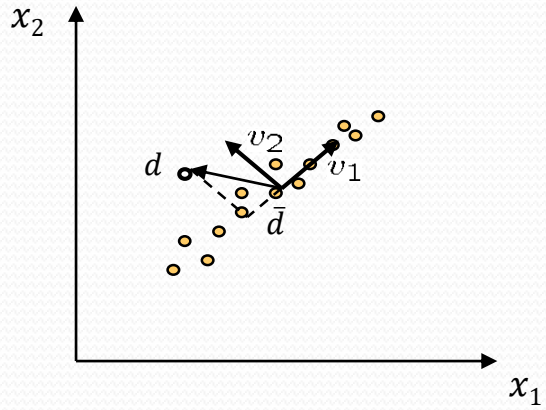
eigenvalue $\lambda \in \mathbb{R}$

- If S is a $N \times N$ symmetric and λ_1 is its largest eigenvalue then

$$\lambda_1 = \sup_{v \in \mathbb{R}^N, \|v\|=1} v^T S v$$

2 How does it work?

Linear Subspaces in case *N=2*



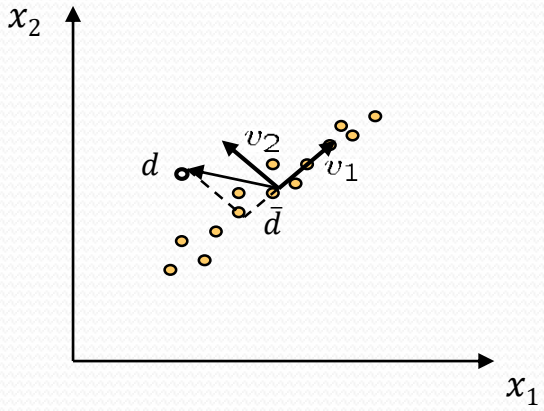
- $\bar{d}$ is the mean of the cloud of points *D*
- Convert d into v_1, v_2 coordinates

$$d \rightarrow ((d - \bar{d})^T v_1, (d - \bar{d})^T v_2)$$

- What does the v_1 coordinate measure?
  - Position along the line
- What does the v_2 coordinate measure?
  - Distance to line

- Suppose the data points are arranged as above
  - Ideal case: fit a line
  - We can represent the points with only their v_1 coordinates since v_2 coordinates are all essentially 0
  - This makes it much cheaper to store and compare points
  - A bigger deal for higher dimensional problems

How to choose linear subspaces?



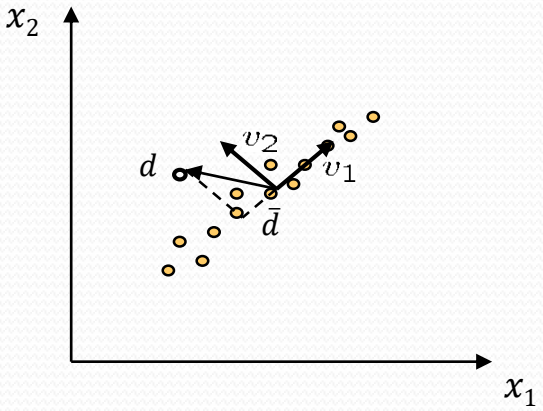
- Consider the variation along direction v among all of the points:

$$\text{var}(v) = \sum_d \left\| (d - \bar{d})^T \, v \right\|^2$$

- What unit vector v_1 maximizes var*(v)*

$$v_1 = \max_v \text{var}(v)$$

Eigenvectors



- Technical derivation:

$$\text{var}(v) = \sum_d \left\| (d - \bar{d})^T v \right\|^2$$

$$\text{var}(v) = \sum_d v^T (d - \bar{d})(d - \bar{d})^T v$$

$$\text{var}(v) = v^T \left(\sum_d (d - \bar{d})(d - \bar{d})^T \right) v$$

$$\text{var}(v) = v^T A v$$

with $A = \sum_d (d - \bar{d})(d - \bar{d})^T$

- Solution: v_1 is eigenvector of A with *largest* eigenvalue
- v_2 is the eigenvector of A with 2th largest eigenvalue
- Etc. in higher dimension until v_N that is the eigenvector of A with smallest eigenvalue.

Higher Dimensions

- Suppose each data point is N-dimensional: same procedure applies:

$$\text{var}(v) = \sum_{d} \left\| (d - \bar{d})^T \, v \right\|^2 = (p-1)\, v^T A v$$

where $A = \frac{1}{p-1} \sum_d (d - \bar{d})(d - \bar{d})^T$ is called the covariance matrix

- The eigenvectors of *A* define a new coordinate system
  - eigenvector with largest eigenvalue captures the most variation among "training" vectors d
  - eigenvector with smallest eigenvalue has least variation

- We can compress the data by only using the top few eigenvectors
  - corresponds to choosing a "linear subspace"
  - represent points on a line, plane, or "hyper-plane"
  - these eigenvectors are known as the **factor axes.**

3 PCA

Transformed variables

- Suppose you have *p* samples of size *N*, so *N* variables $x_1, x_2, \ldots, x_N$
- *N* factor axes associated to *N* eigenvectors $v_i = (v_{i1}, v_{i2}, \cdots, v_{iN})$
- New variables that $y_1, y_2, \ldots, y_N$ are linear combination of the original variables $x_1, \ldots, x_N$:

$$y_i = v_{i1}x_1 + v_{i2}x_2 + \cdots + v_{iN}x_N \text{ for } i=1..N$$

- The new variables y_i are derived in decreasing order of importance (eigenvalues).

Principal Components

- Considering all the input samples $d_j = (d_{j1}, d_{j2}, \cdots, d_{jN})$ for *j=1,…,p*, let

$$c_i = \left(v_i{}^T(d_1 - \bar{d}), v_i{}^T(d_2 - \bar{d}), \ldots, v_i{}^T(d_N - \bar{d})\right) \text{ for } i=1,\ldots,N$$

- c_i is the projection of the samples on the *i*-th factor axis
- c_i is called the *i*-th **principal component** (vector of size *p*).
- The sample d_j is converted into

$$d_j \rightarrow \left((d_j - \bar{d})^T\, v_1, (d_j - \bar{d})^T\, v_2, \cdots, (d_j - \bar{d})^T\, v_N\right)$$

- Hence, each sample can be approximated (**dimension reduction**) by

$$d_j \approx \bar{d} + c_{1j}v_1 + c_{2j}v_2 + \cdots + c_{tj}\, v_t$$

Correlation

- We can show that
  - $\mathrm{var}(c_i) = \lambda_i$ for all *i=1,…,N* (λ_i is the *i*-th eigenvalue).
  - $\mathrm{corr}(c_i, c_k) = 0$ for $i \neq k$.
  - The principal components are not correlated.

- Let the column vector $d^k = (d_{1k}, d_{2k}, \cdots, d_{pk})$ associated to variable x_k. We can show that
  - $\sum_{i=1}^{t} \mathrm{corr}(c_i, d^k)^2 \leq 1$ for all $t \leq N$ and all *k=1,…,N*
  - $\sum_{i=1}^{N} \mathrm{corr}(c_i, d^k)^2 = 1$

Singular Value Decomposition

- PCA is often performed via singular value decomposition, because forming $A = \frac{1}{p-1} D^T D = \frac{1}{p-1} \sum_d (d - \bar{d})(d - \bar{d})^T$ would not be required
- Let $D = U\Sigma V^T$ be the SVD of A.
- The columns of V are the eigenvectors of A.
- The eigenvalues of A are the square of the singular values (up to the scale factor $p - 1$).

4 Practical interpretation

Correlation circle

- From the correlation property of principal components, we deduce that

$$M^k = \left(\text{corr}\left(c_1, d^k\right)^2,\ \text{corr}\left(c_2, d^k\right)^2\right)$$

  is inside the unit circle for all samples d^k.

- Hence the unit circle containing all the points M^k is called the correlation circle.

- Interpretation:
  - If M^j is close to the circle border, it means that the variable x_j is well explained by the principal components c_1 and c_2 .
  - If M^j and M^k are close to the circle border and they are almost « orthogonal », it means that the variables x^j and x^k are almost « not correlated ».

The iris data set

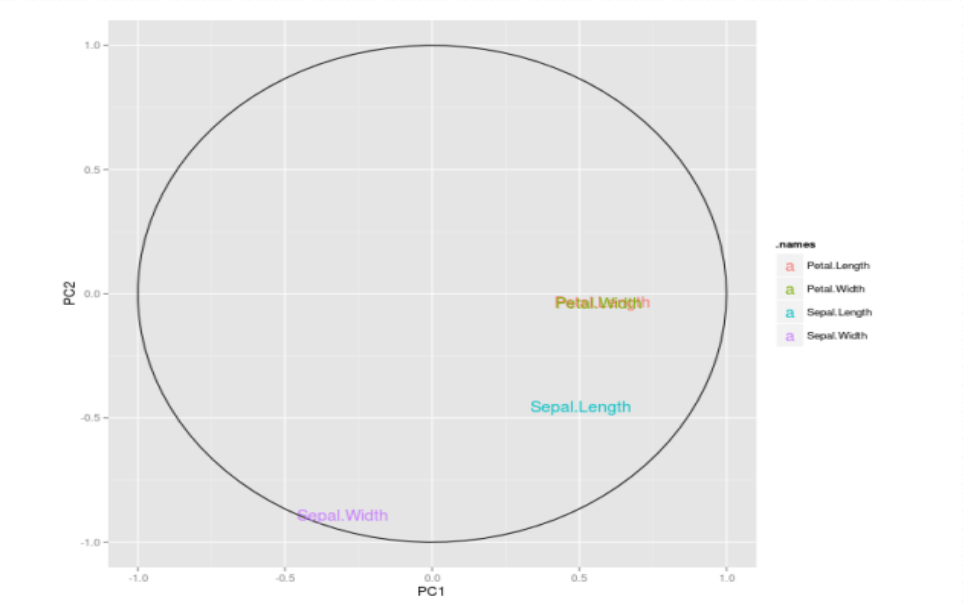
- Three flowers

50 instances of each class, 150 in total

- Attributes
  - Sepal (outermost leaves)
    - length in cm
    - width in cm
  - Petal (innermost leaves)
    - length in cm
    - width in cm
  - Class of flower
    - Iris Setosa
    - Iris Versicolour
    - Iris Virginica

Fower 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

Correlation circle for « IRIS data set »

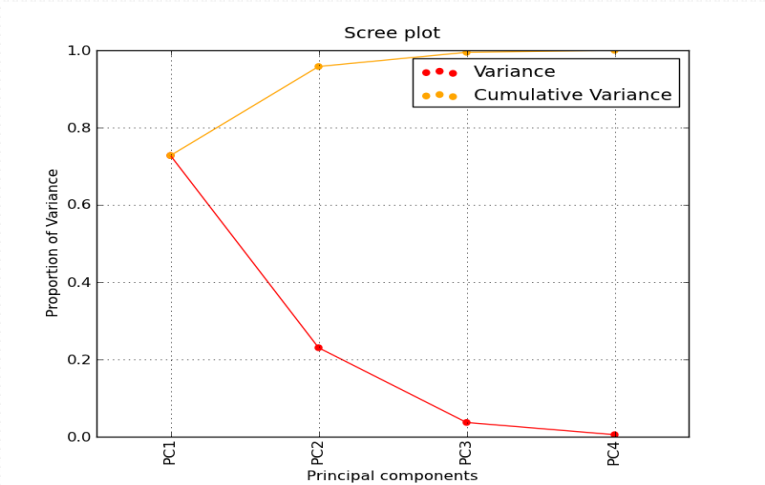


How Many PCs?

- For *N* original dimensions, correlation matrix is *N* x *N*, and has up to *N* eigenvectors. So *N* Principal Components (PC).
- Where does dimensionality reduction come from?
- We only keep *t* PCs.
- Choice based on the explained variance:
  - Total variance = sum of eigenvalues
  - Variance explained = sum of kept eigenvalues
  - Rate of approximation = Variance explained/ Total variance

Screeplot

- Screeplot : show the eigenvalues (so the variance)
- « IRIS dataset »:



Importance of PCA

- In data of high dimensions, where graphical representation is difficult, PCA is a powerful tool for analyzing data and finding patterns in it.
- Data compression is possible using PCA
- The most efficient expression of data is by the use of perpendicular components, as done in PCA.
- Do not forget to center and normalize data for interpretation

Limits of PCA

- How to cut the number of factor axes?
- Interpretation of the factor axes (eigenvectors) can be difficult
- Linear spaces might be not accurate enough
- Extensions: Kernel PCA, etc.

5 Conclusion

Conclusion

- PCA: a very famous method!
- Very useful under suitable conditions and processing
- Data cleaning is very important
- The cornerstone of more advanced methods