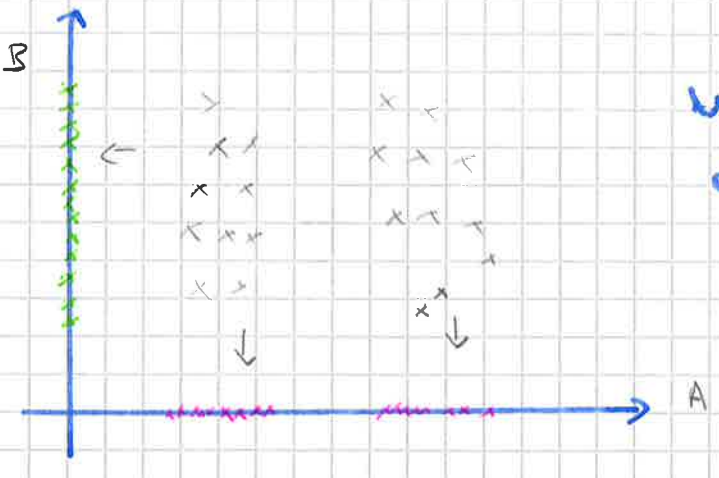
Chapter 8
Principal Component Analysis (PCA)

(This is a case of unsupervised learning)

I) The general problem

We have N objects $(y_1, y_2, \dots, y_n)$. Each of these is a n-dim. vector. If n is large, it is difficult to visualize the data so the usual trick is to do a projection on a D-dim space (we obtain $x_1, x_2, \dots, x_n$).

Ex:



We project on A or B.

We can compute the empirical variance in each one-dim. space. It will be higher for projection A than for projection B.

$$\sigma^2 = \frac{1}{N} \sum_{i=1}^{N} (x_n - \mu_x)^2$$

Projection A preserves the cluster structure (this is why the variance is higher). So variance seems as a

good quantity to maximize when deciding on projection techniques.

II) Principal component analysis

We want to perform a linear projection: each of the projected dimensions is a linear combination of the original dimensions. Suppose we project from $\mathbb{R}^M$ to $\mathbb{R}^D$, we define D vectors $w_1, \dots, w_D$ ($\in \mathbb{R}^M$). The d-th element of the projection is x_{nd}:

$x_{nd} = w_d^T y_n$ (the d-th element of the n-th vector)

Question: choose D and $x_1, \dots, x_D$

PCA chooses w_1 such that the projections $(x_{n_1}, \dots, x_{n_1}, \dots, x_{n_1})$ has a variance as high as possible.
then PCA chooses w_2 such that $(x_{n_2}, \dots, x_{n_2}, \dots, x_{n_2})$ has a variance as high as possible ...

under the constraints: $\begin{cases} \forall i, \|w_i\| = 1 \\ \forall i \neq j, \ w_i^T w_j = 0 \end{cases}$

Simplifying assumption: $\bar{y} := \frac{1}{N} \sum_{i=1}^{N} y_i = 0$

Let us start by projecting into $D = 1$ dimension:
we have $x_n = w^T y_n$. We then get the variance

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

where $\bar{x} = \frac{1}{N} \sum_{n=1}^{N} w^T y_n = w^T \left(\frac{1}{N} \sum_{n=1}^{N} y_n \right) = 0$

so $\sigma_x^2 = \frac{1}{N} \sum_{n=1}^{N} x_n^2 = \frac{1}{N} \sum_{n=1}^{N} (w^T y_n)^2$

$$\sigma_x^2 = \frac{1}{N}\sum_{n=1}^{N} w^T y_n^T y_n w$$

$$= w^T \left(\frac{1}{N}\sum_{i=1}^{N} y_i^T y_i \right) w$$

$$= w^T C w$$

where C is the covariance matrix:

$$C = \frac{1}{N}\sum_{i=1}^{N} (y_n - \bar{y})^T (y_n - \bar{y}) \quad \text{(recall that } \bar{y} = 0\text{)}$$

We want to maximise σ_x^2 under the constraint $w^T w = 1$. We incorporate this in an optimisation through the use of a Lagrangian term (see Chapter 3). We want to find the (w_i, λ) that maximizes:

$$L = w^T C w - \lambda (w^T w - 1)$$

We have $\nabla_w L = 2Cw - 2\lambda w$

↑ you have to know how to justify this

so $Cw = \lambda w$

There are many candidates: all the eigenvectors. We want to one maximizing $w^T C w$. So this is w_1, the eigenvector for the eigenvalue λ_1 (which is the biggest eigenvalue on $1 \dots N$)

When projecting on $D \geq 2$ dimensions:

We order the eigenvalues of C: $\lambda_1, \lambda_2, \dots, \lambda_N$ (such that $|\lambda_1| \geq |\lambda_2| \geq \dots$). Remember that as C is symmetric there exist a basis of eigenvectors: $w_1, w_2, \dots$ $(w_i \leftrightarrow \lambda_i)$ such that $i \neq j \Rightarrow w_i^T w_j = 0$.

For $1 \leq i \leq D$, the variance of the component i of the

projection is $\frac{1}{N}\sum_{n=1}^{N}(w_i^T y_n)^2$ (because $\bar{y}=0$)

$= \sigma_i^2$

So we choose to project on $(w_1, \dots, w_D)$

Remark: The eigen-spectrum (magnitudes of the eigenvalues) gives us some indication of how many interesting features there are in our data.

1) Choosing D

- Use the eigen-spectrum (as said above).
- If we want to visualize the data, we have to restrict ourselves to $D \leq 3$.

2) Limitations of PCA

We are limited to
- data which are real valued
- no missing value in the data.

Exercises: p. 140 - 155 in the python book

I recommend to have a look also at p. 132-139 (rescaling questions).