Exploratory Data Analysis

- Numerical data description (robust)
- Visualisation [Labs]

→ Understand structure of data

EDA linked to Preprocessing

- Standardisation
  - Centring $[C_n = I_n - \frac{1}{n} 1_n 1_n^T]$
  - Scaling to unit variance

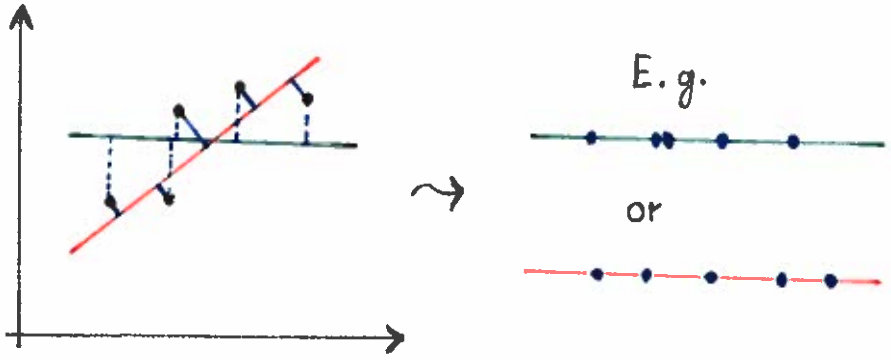
  ] linear operations
- Outlier detection / removal
  - Tukey's fences

How to represent our data?

- to find and exploit structure
- to facilitate further processing

Quiz: Useful data direction

- Given: 2D data
- Represent using just 1D data
- Linear projection



Should we select the direction such that we:

A) maximise the variance

B) minimise the approximation error (squared)

C) approximate the sample covariance matrix as well as possible (in the Frobenius norm)

D) either A, B or C depending on the data struct

E) do something else

F) ... you don't know

Principal Component Analysis

First view: Find directions in data space along which data are maximally variable

↝ "PC directions"

First PC direction: $\underline{w}_1$

First PC: Projected variable $z_1 = \underline{w}_1^T \underline{x}$

where $\underline{x}$: centred random variable

$\text{Var}[z_1] = \text{Var}[\underline{w}_1^T \underline{x}] = \underline{w}_1^T \Sigma \underline{w}_1$ ($\text{Cov}[A\underline{x}+\underline{b}] = A\,\text{Cov}[\underline{x}]\,A^T$)

where $\Sigma = \text{Cov}[\underline{x}]$

↝ Optimisation problem:

$$\max_{\underline{w}_1} \underline{w}_1^T \Sigma \underline{w}_1$$

$$\text{s.t. } \|\underline{w}_1\| = 1$$

EV decomposition: $\Sigma = U \Lambda U^T$

W.l.o.g. $\lambda_1 \geq \lambda_2 \geq \dots \geq \lambda_d \geq 0$

U orth. basis $\Rightarrow \underline{w}_1 = \sum_{i=1}^{d} a_i \underline{u}_i = U\underline{a}$

$$\Rightarrow \underline{w}_1^T \Sigma \underline{w}_1 = \underline{a}^T U^T U \Lambda U^T U \underline{a} = \underline{a}^T \Lambda \underline{a} = \sum_{i=1}^{d} a_i^2 \lambda_i$$

$$\|\underline{w}_1\|^2 = \underline{w}_1^T \underline{w}_1 = \underline{a}^T U^T U \underline{a} = \sum_{i=1}^{d} a_i^2 \overset{!}{=} 1$$

maximise $a_1, \dots, a_d$

Solution: $\underline{a} = (1\,0\,0 \dots 0)^T$

$\Rightarrow \underline{w}_1 = U \begin{pmatrix} 1 \\ 0 \\ \vdots \\ 0 \end{pmatrix} = \underline{u}_1$ i.e. EV with largest eigenv

PC: $z_1 = \underline{w}_1^T \underline{x}$

Properties: $\mathbb{E}[z_1] = \mathbb{E}[\underline{w}_1^T \underline{x}] = \underline{w}_1^T \mathbb{E}[\underline{x}] = 0$

$$\mathrm{Var}[z_1] = \underline{w}_1^T \Sigma \underline{w}_1 = \sum_{i=1}^{d} a_i^2 \lambda_i = \lambda_1$$

For observations $X = (\underline{x}_1\ \underline{x}_2 \dots \underline{x}_n)$:

$(z_1)_i = \underline{w}_1^T \underline{x}_i$ also called PC (scores)

All n first PC scores: $\underline{z}_1^T = \underline{w}_1^T X$

Subsequent PC directions

Maximise variance of $z_2 = \underline{w}_2^T \underline{x}$

but also reveal something new:

Require $\underline{w}_2$ orthogonal to $\underline{w}_1$

→ Optimisation problem:

maximise $\underline{w}_2^T \Sigma \underline{w}_2$

s.t. $\|\underline{w}_2\| = 1$ and $\underline{w}_2^T \underline{w}_1 = 0$

Like before, using the EV decomposition

express $\underline{w}_2 = U\underline{b}$

$$\Rightarrow \underline{w}_2^T \underline{w}_1 = \underline{w}_2^T \underline{u}_1 = \underline{b}^T U^T \underline{u}_1 \overset{\text{orth.}}{=} \underline{b}^T \begin{pmatrix} 1 \\ 0 \\ \vdots \\ 0 \end{pmatrix} = b_1$$

$\rightsquigarrow$ Insert constraint $b_1 = 0$:

$$\underset{b_2,\dots,b_d}{\text{maximise}} \sum_{i=2}^{d} b_i^2 \lambda_i \quad \text{s.t.} \sum_{i=2}^{d} b_i^2 = 1$$

Like before, solution $\underline{b} = (0\,1\,0\,\dots\,0)^T$

$\underline{w}_2 = U\underline{b} = \underline{u}_2$

$\text{Var}[z_2] = \lambda_2$

Also: z_1 and z_2 are uncorrelated:

$\mathbb{E}[z_1 z_2] = 0$

Further PC directions:

$$\underset{\underline{w}_m}{\text{maximise}} \; \underline{w}_m^T \Sigma \underline{w}_m$$

$$\text{s.t. } \|\underline{w}_m\| = 1 \text{ and } \underline{w}_m^T \underline{w}_i = 0, \; i = 1,\dots,m-1$$

$\rightsquigarrow \underline{w}_m = \underline{u}_m, \quad \text{Var}[z_m] = \lambda_m$

Total variance of $z_1, \ldots, z_k$:

(i.e. variance "explained" by k PCs):

$$\sum_{m=1}^{k} \mathrm{Var}[z_i] = \sum_{m=1}^{k} \lambda_m$$

Corresponding normalised measure:

Fraction of variance explained $\dfrac{\sum_{i=1}^{k} \lambda_i}{\sum_{i=1}^{d} \lambda_i}$