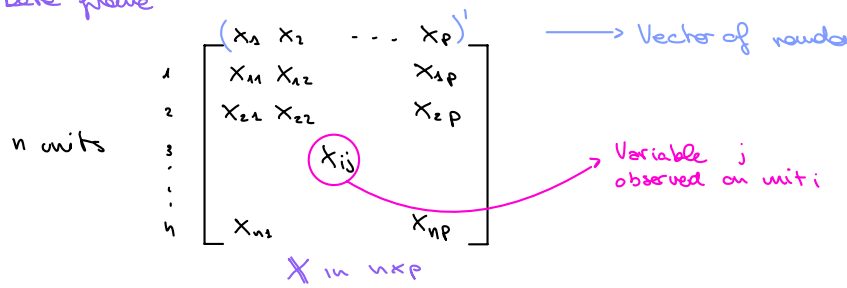
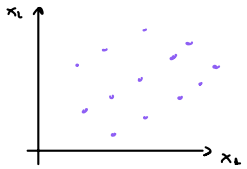
Data frame



Row (unit) perspective

$$\underline{x}_1' = (x_{11}, \dots, x_{1p})$$
$$\underline{x}_2' = (x_{21}, \dots, x_{2p}) \quad \Big\} \in \mathbb{R}^p$$
$$\underline{x}_n' = (x_{n1}, \dots, x_{np})$$



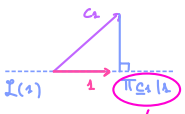
Not the recived data, but the pre organized data

Column perspective

$$\underline{c}_1 = \begin{pmatrix} x_{11} \\ x_{21} \\ \vdots \\ x_{n1} \end{pmatrix} \in \mathbb{R}^n$$

These numbers are all talking about the same information
1 variable for n statistical units (we have information about mutual variability)

Having dependance between 2 variables we can create a prediction.



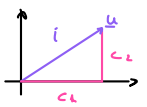
There is a linear subspace which is constant hence there is NOT variability.
All component on the vector are the same for ex $\mathcal{L}(1) = \text{span}(1) = \{\beta \cdot 1 : \beta \in \mathbb{R}\}$

→ This is interesting cause we want to known how to approximate our data with a constant the best way

Euclidian spaces

$\underline{u}, \underline{v} \in \mathbb{R}^n$

$$\|\underline{u}\| = \sqrt{\sum_{i=1}^n u_i^2} = \sqrt{\underline{u}'\underline{u}} = \langle \underline{u}, \underline{u} \rangle$$



$$i^2 = c_1^2 + c_2^2$$

$$\cos\theta = \frac{\langle \underline{u}, \underline{v} \rangle}{\|\underline{u}\| \|\underline{v}\|} = \frac{\underline{u}'\underline{v}}{\sqrt{\underline{u}'\underline{u}}\sqrt{\underline{v}'\underline{v}}}$$

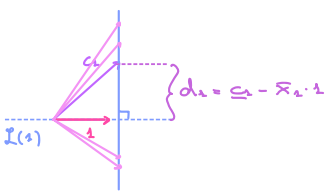
$$\pi_{\underline{u}|\underline{v}} = \|\underline{u}\| \frac{\underline{u}'}{\|\underline{u}\|\|\underline{v}\|} \cdot \frac{\underline{v}}{\|\underline{v}\|} = \frac{\underline{u}'\underline{v}}{\underline{v}'\underline{v}} \underline{v} = \boxed{\frac{\underline{v}\underline{v}'}{\underline{v}'\underline{v}}} \cdot \underline{u}$$

→ orth proj operator on $\mathcal{L}(\underline{v})$

$$\pi_{\underline{c}_1|1} = \frac{1\,1'}{1'1}\,\underline{c}_1 = \frac{1'\underline{c}_1}{1'1} \cdot 1 = \frac{1}{n}\sum_{i=1}^n x_{i1} = \bar{x}_1 \cdot 1$$

The mean

Different sample distribution can have the same mean because they have the same orthogonal projection, but they have different distance from it.



$$\underline{d}_1 = \begin{bmatrix} c_{11} - \bar{x}_1 \\ c_{21} - \bar{x}_1 \\ \vdots \\ c_{n1} - \bar{x}_1 \end{bmatrix}$$

$$\boxed{\|\underline{d}_1\| = \sqrt{\sum_{i=1}^n (x_{i1} - \bar{x}_1)^2}}$$

SAMPLE STANDARD DEVIATION

$$\sqrt{s_{11}} = \sqrt{\frac{1}{n-1}\sum_{i=1}^n (x_{i1} - \bar{x}_1)^2} \qquad \|d_1\| \propto \sqrt{s_{11}} = \sqrt{n-1}\sqrt{s_{11}}$$

$$\boxed{s_{ii} = \frac{1}{n-1}\sum_{i=1}^n (x_{ii} - \bar{x}_i)^2}$$

SAMPLE VARIANCE

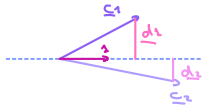
Tchebychev's inequality

$$\text{Freq}\left[\bar{x}_1 - k\sqrt{s_{11}} < x_1 < \bar{x}_1 + k\sqrt{s_{11}}\right] \geq 1 - \frac{1}{k^2} \qquad k > 0$$

$$\text{Freq}\left[x_1 \in \left[\bar{x}_1 - 2\sqrt{s_{11}} < x_1 < \bar{x}_1 + 2\sqrt{s_{11}}\right]\right] \geq 75\%$$

$\underline{c}_1 = \begin{pmatrix} x_{11} \\ x_{21} \\ \vdots \\ x_{n1} \end{pmatrix} \in \mathbb{R}^n \longrightarrow \bar{x}_1 \cdot \mathbb{1}$, $\underline{d}_1 = \underline{c}_1 - \bar{x}_1 \cdot \mathbb{1}$

$\underline{c}_2 = \begin{pmatrix} x_{12} \\ x_{22} \\ \vdots \\ x_{n2} \end{pmatrix} \in \mathbb{R}^n \longrightarrow \bar{x}_2 \cdot \mathbb{1}$, $\underline{d}_2 = \underline{c}_2 - \bar{x}_2 \cdot \mathbb{1}$



$\mathbb{R}^n$

Limiting cases

$\theta_{12} = 0$  $\underline{d}_2 \in \text{span}(\underline{d}_1)$ $\quad \underline{d}_2 = \beta \underline{d}_1$

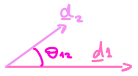
$x_{i2} - \bar{x}_2 = \beta(x_{i1} - \bar{x}_1) \qquad i = 1, 2, \dots, n$

$x_{i2} = \beta x_{i1} + (\bar{x}_2 - \beta \bar{x}_1)$

all information of one in the other

$\theta_{12} = \frac{\pi}{2}$ 

no information one in the other

$0 < \theta_{12} < \frac{\pi}{2}$ 

$$\cos\theta_{12} = \frac{\langle \underline{d}_1, \underline{d}_2 \rangle}{\|\underline{d}_1\| \, \|\underline{d}_2\|} = \frac{\underline{d}_1' \underline{d}_2}{\sqrt{n-1}\sqrt{s_{11}}\sqrt{n-1}\sqrt{s_{22}}} = \frac{\sum_{i=1}^n (x_{i1} - \bar{x}_1)(x_{i2} - \bar{x}_2)}{(n-1)\sqrt{s_{11}}\sqrt{s_{22}}} = \frac{s_{12}}{\sqrt{s_{11}}\sqrt{s_{22}}} = \text{Corr}(x_1, x_2) = \rho_{12}$$

$$\text{cov}(x_1, x_2) = \frac{1}{n-1} \sum_{i=1}^n (x_{i1} - \bar{x}_1)(x_{i2} - \bar{x}_2) = s_{12}$$

$\boxed{0 \le \rho_{12} \le 1}$ (because we have a cosine, hence the correlation is 1 because we have 0 as angle)

Data frame

$\begin{pmatrix} \bar{x}_1 \\ \vdots \\ \bar{x}_p \end{pmatrix} = \bar{\underline{x}}$ vector of means

$s_{jk} = \frac{1}{n-1} \sum_{i=1}^n (x_{ij} - \bar{x}_j)(x_{ik} - \bar{x}_k)$ sample covariance if $j \neq k$

$s_{jj} = \text{Var}(x_j)$ sample variance

X $n \times p$ $\Longrightarrow$ $S = \begin{bmatrix} s_{11} & s_{12} & \dots & s_{1p} \\ & \ddots & & \vdots \\ & & & s_{pp} \end{bmatrix}$ $p \times p$

$\rho_{jk} = \frac{s_{jk}}{\sqrt{s_{jj}}\sqrt{s_{kk}}}$ $\quad j, k = 1, \dots, p$ $\quad \rho = \begin{bmatrix} 1 & \rho_{12} & \dots & \rho_{1p} \\ & \ddots & & \vdots \\ & & & 1 \end{bmatrix}$

Remark (standardization)

$X_1 \xrightarrow{\text{std}} \frac{x_1 - \bar{x}_1}{\sqrt{s_{11}}} = z_1$ → how far you are from the mean

z_1 $\quad \text{Var}(z_1) = 1$ $\quad \mathbb{E}[z_1] = 0$

Stat learning (?)

$X^* = X \cdot Y$ → Target variable

1. Supervised learning
2. Unsupervised learning

Fundamental ideas about Stat Learning

Supervised setting $\quad \underline{x} = (x_1, \dots, x_p)' \in \mathbb{R}^p$ $\quad Y \in \mathbb{R}$

Observed: on n units $\Longrightarrow$ X data frame

Goal: explain the variability of Y as a function of $\underline{x}$

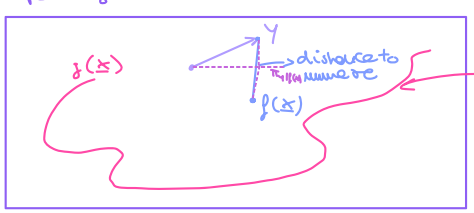
$Y \rightarrow$ Random variable

$\underline{x} \rightarrow$ Random vector

? $f: \mathbb{R}^p \rightarrow \mathbb{R}$ s.t. $f(\underline{x})$ is a "prediction" of Y

What f? "The best"

Space of all r.v. in $\mathbb{R}$ (W s.t. $\mathbb{E}[W^2] < +\infty$)



sub space of r.v.'s measurable w.r.t. $\underline{x}$

2 random vectors, hence we have to minimize it for each unit that we will ever possibly observe

find $f: \mathbb{R}^p \rightarrow \mathbb{R}$ s.t. $|Y - f(\underline{x})|^2$

minimizing not this BUT $\mathbb{E}[|Y - f(\underline{x})|^2]$

$\pi_{Y|\sigma(X)} = \mathbb{E}[Y|X]$ REGRESSION FUNCTION (can have any form, don't confuse it with linear)

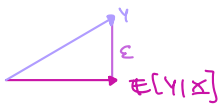
$$\mathbb{E}[(Y-f(X))^2] = \mathbb{E}[(Y-\mathbb{E}[Y|X]+\mathbb{E}[Y|X]-f(X))^2] =$$
$$= \mathbb{E}[(Y-\mathbb{E}[Y|X])^2] + \mathbb{E}[(\mathbb{E}[Y|X]-f(X))^2] + 2\mathbb{E}[(Y-\mathbb{E}[Y|X])(\mathbb{E}[Y|X]-f(X))]$$

W, Z r.v. $\quad \mathbb{E}[W] = \mathbb{E}[\mathbb{E}[W|Z]]$

$$(*) = \mathbb{E}[(\ )(\ )] = \mathbb{E}[\mathbb{E}[(Y-\mathbb{E}[Y|X])(\mathbb{E}[Y|X]-f(X))\,|\,X]] =$$
$$= \mathbb{E}[(\mathbb{E}[Y|X]-f(X))\cdot \underbrace{\mathbb{E}[Y-\mathbb{E}[Y|X]\,|\,X]}_{=0}]$$

$$\boxed{\mathbb{E}[(Y-f(X))^2] = \mathbb{E}[(Y-\mathbb{E}[Y|X])^2] + \mathbb{E}[(\mathbb{E}[Y|X]-f(X))^2]}$$

Hence $\boxed{f(X) = \underset{g(X)}{\arg\min}\, \mathbb{E}[(Y-\mathbb{E}[Y|X])^2] = \mathbb{E}[Y|X]}$ This holds for everything in L^2



General model: $Y = \mathbb{E}[Y|X] + \varepsilon$

$Y = f(X) + \varepsilon$ where $f(X) = \mathbb{E}[Y|X]$

$\mathbb{E}[Y] = \mathbb{E}[f(X)+\varepsilon] = \mathbb{E}[\mathbb{E}[Y|X]] + \mathbb{E}[\varepsilon] = \mathbb{E}[Y] + \mathbb{E}[\varepsilon]$

$\rightarrow \boxed{\mathbb{E}[\varepsilon] = 0}$

General model

$Y = f(X) + \varepsilon$

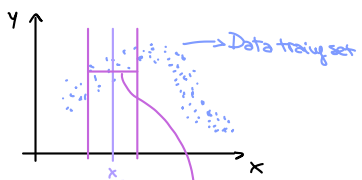
$f(x) = \mathbb{E}[Y|X]$

ε r.v. s.t. $\varepsilon \perp \sigma(X)$ $\quad \mathbb{E}[\varepsilon] = 0$

Learn from data (and from the engineering knowledge)

$\Rightarrow$ data $\xrightarrow{\text{learn}}$ $\hat{f}$ $(= \widehat{\mathbb{E}[Y|X]})$

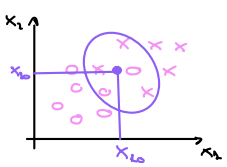
Do we really need a model for f?



For $x \in \mathbb{R}$

let $N_x = \{K \ x_i : x_i \text{ traing set, close to } x\}$

$\hat{f}(x) = \frac{1}{K}\sum_{x_i \in N_x} y_i$



× male ○ female

• K-mean $\quad \hat{f}((x_{2,0}, x_{1,0})) = x$

Local methods for obtaing $\hat{f}$ are not (always) good!

Curse of dimensionality

An idea

$S_1(p)$ universe of radius 1 in $\mathbb{R}^p$

i.e. sphere centered in 0, with radius 1 in $\mathbb{R}^p$

Friends are uniformally distributed in $S_1(p)$. How far to go to meet 10% of friends.

$p=1$  $\quad ?r \longrightarrow \frac{2r}{2} = 0,1 \quad r = 0,1$

$p=2$  $\quad \frac{\pi r^2}{\pi 1^2} \longrightarrow r^2 = 0,1 \quad r = 0,31$

$p=3$  $\quad \frac{\frac{4}{3}\pi r^3}{\frac{4}{3}\pi 1^3} \longrightarrow r^3 = 0,1 \quad r = 0,46$

$p=100 \quad r^{100} = 0,1 \quad r = 0,97$

To extend the concept we have to have higher data to get information on higher dimension. Basically the space is too big and near one there are not people like me, because of dimension.

We have to reduce the dimensionality of the problem:

1. Reducing p (in a smart way, we have to still have information) → PCA, ICA ...
2. Reduce the dim of f → Structured models

Struct model: Find $\hat{f} \in \{ f_\theta \quad \theta \in \mathbb{R}^k \}$ — if θ is known $\Rightarrow$ f_θ is known

$X \to \hat{\theta} \implies \hat{f} = f_{\hat{\theta}}$

EXAMPLE

$f_\theta(X) = \beta_0 + \beta_1 X_1 + \dots + \beta_p X_p$ $\qquad X = (X_1, \dots, X_p)'$

$\theta = (\beta_0, \beta_1, \dots, \beta_p)'$ → this has NOT to be linear, we linearize it in this form

Suppose that:

- obtained the data
- estimated $\hat{f}$ (approx. of $\mathbb{E}[Y|X]$)

$\implies$ Ready to solve my problem

X_0 new features for a new unit $\implies$ predict Y_0 by means of $\hat{f}(X_0)$

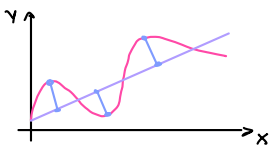
$(Y_0 - \hat{f}(X_0))^2$ $\qquad Y = f(X) + \varepsilon$ $\qquad Y_0 = f(X_0) + \varepsilon$

But given the same input the outputs haven't to be the same value

$$\mathbb{E}_{|X}[(Y_0 - \hat{f}(X_0))^2] = \mathbb{E}_{|X}[(f(X_0) + \varepsilon_0 + \hat{f}(X_0))^2] = \mathbb{E}_X[(f(X_0) - \hat{f}(X_0))^2] + \mathbb{E}_{|X}[\varepsilon_0] + \mathbb{E}_X[\varepsilon_0(f(X_0) - \hat{f}(X_0))]$$
$$\overset{!}{=} (f(X_0) - \hat{f}(X_0))^2 + \mathrm{Var}(\varepsilon_0)$$

Generalization problem:

$$\mathbb{E}_{|X}[(Y_0 - \hat{f}(X_0))^2] = \underbrace{(f(X_0) - \hat{f}(X_0))^2}_{\text{reducible error}} + \underbrace{\mathrm{Var}(\varepsilon_0)}_{\text{irreducible}}$$



If we try to reduce too much we get overfitting → basically our data set, hence we don't have any new information nor anything useful.

Expected General. Error:

$$\mathbb{E}[\mathbb{E}_{|X}[(Y_0 - \hat{f}(X_0))^2]] = \mathbb{E}[(f(X_0) - \hat{f}(X_0))^2] + \mathrm{Var}(\varepsilon_0) = (f(X_0) - \mathbb{E}[\hat{f}(X_0)])^2 + \mathbb{E}[(\hat{f}(X_0) - \mathbb{E}[\hat{f}(X_0)])^2] + \mathrm{Var}(\varepsilon)$$
$$\overset{!}{=} \mathrm{bias}^2 + \mathrm{Var}(\hat{f}(X_0)) + \mathrm{Var}(\varepsilon_0)$$

$$\mathbb{E}[\mathbb{E}_{|X}[(Y_0 - \hat{f}(X_0))^2] = \mathrm{bias}^2 + \mathrm{Var}(\hat{f}(X_0)) + \mathrm{Var}(\varepsilon_0)$$