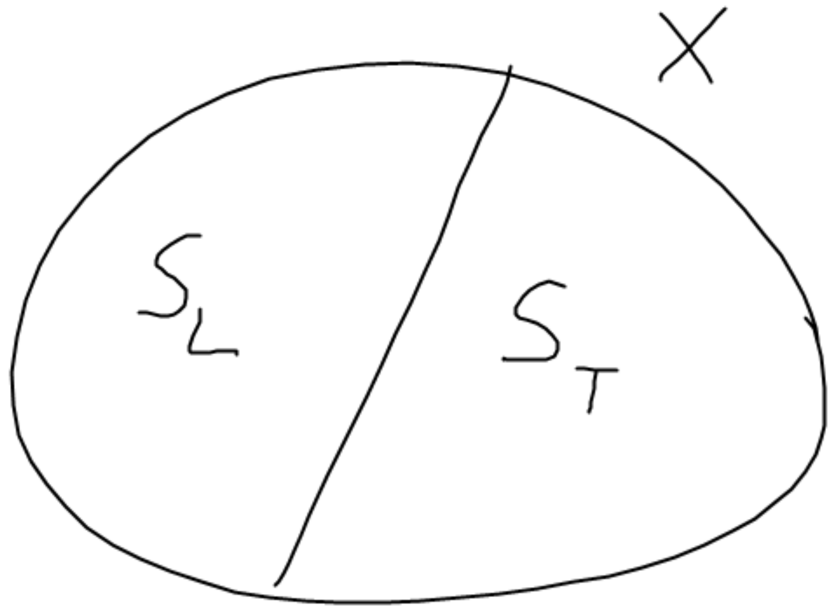
4/ Cross Validation

→ most robust method that can be used to select a model / estimator

→ computationnally intensive

n iid observations $X = (X_1, \dots, X_n)$



you split it into two sets

→ a learning sample / estimation sample / training sample

→ a validation sample / transfer sample / testing sample you also need * a contrast

* d different estimators that may depend on d different models but not necessary

estimator $\longrightarrow$ black box

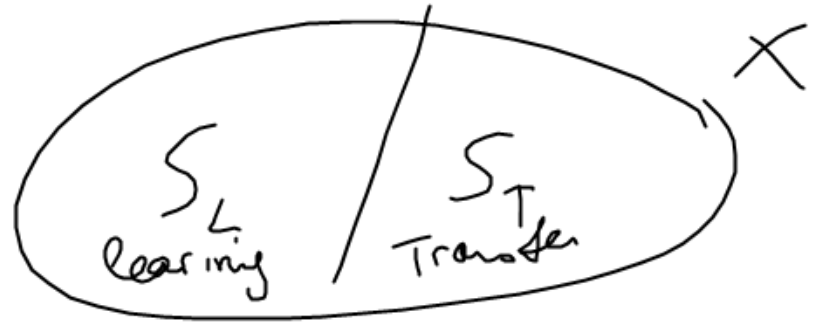
$$\text{data} \longrightarrow \boxed{\hat{f}_j} \longrightarrow \hat{f}_j^{\,\text{data}}$$

estimator of say $\underline{f_0}$.

might be a function or just a param θ_0.

<u>ex</u>: - projection estimators on d different subsets

- density estimators (kernel with different bandwidth)



a) Hold-out

* for each $\hat{f}_j$, you compute them only with S_L

$$\rightarrow \hat{f}_1^{S_L}, \dots, \hat{f}_d^{S_L}$$

* for each of them, you compute $C(\hat{f}_j^{S_L}, S_T)$

where C is a contrast designed for our target f_0

(It means that $\mathbb{E}_{f_0}(C(f, S_T))$ is minimal when $f = f_0$)

the estimator that you choose is given by $\boxed{\hat{j} = \arg\min_{j=1..d} C(\hat{f}_j^{S_L}, S_T)}$

computed with the whole sample

* $\hat{f}_{\hat{j}}(X)$

ex $X_1, \dots, X_n$ iid with density f.

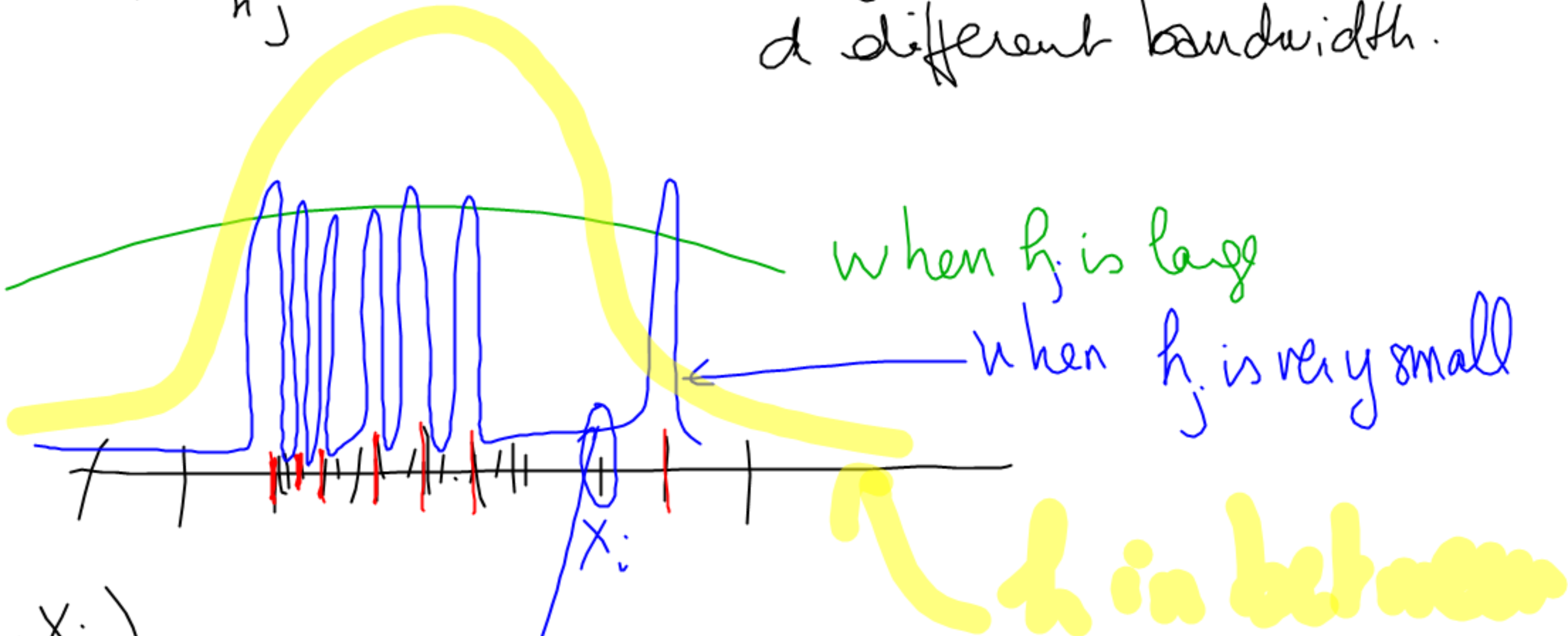
$$\hat{f}_j^X(x) = \frac{1}{n h_j} \sum_{i=1}^{n} K\left(\frac{x - X_i}{h_j}\right)$$

with h_j from $j = 1 \dots d$

d different bandwidth.

$S_L = X_1, \dots, X_{n/2}$

$S_T = X_{n/2+1}, \dots, X_n$



$$\hat{f}_j^{S_L}(x) = \frac{1}{n/2 \, h_j} \sum_{i=1}^{n/2} K\left(\frac{x - X_i}{h_j}\right)$$

$$C(\hat{f}_j^{S_L}, S_T) = -\frac{2}{n/2} \sum_{i=n/2+1}^{n} \hat{f}_j^{S_L}(X_i) + \int \left(\hat{f}_j^{S_L}(x)\right)^2 dx$$

n is even

$$\hat{j} = \underset{j=1\ldots d}{\operatorname{argmin}} \; C\left(\hat{f}_j^{S_L}, S_T\right)$$

$$= \underset{j=1,\ldots,d}{\operatorname{argmin}} \; -\frac{2}{n/2} \sum_{i=n/2+1}^{n} \hat{f}_j^{S_L}(X_i) + \int \left[\hat{f}_j^{S_L}(x)\right]^2 dx$$

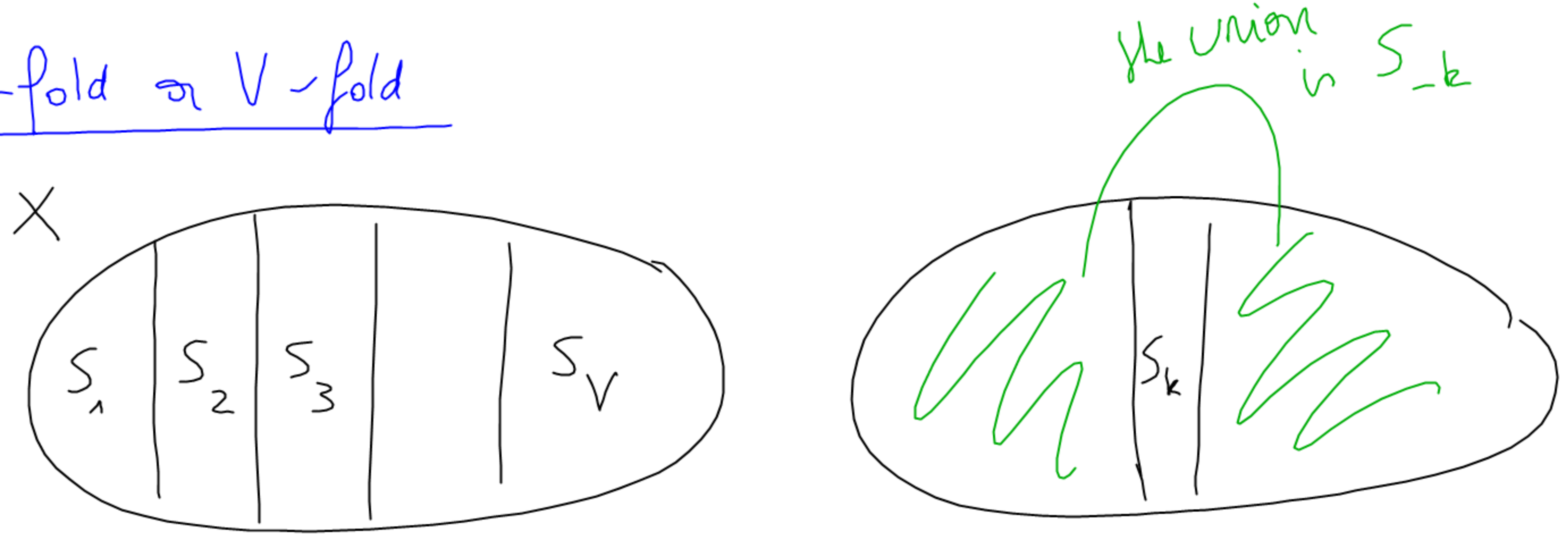
this is in S_T
the transfer sample

The good estimator is therefore $\hat{f}_{\hat{j}}(X)$

I compute it with the whole sample.

I selected the good method by Holdout

b) K-fold or V-fold



S_{-k} = is the sample when we remove S_k , $k = 1..V$

for each k, $\hat{f}_j^{S_{-k}}$ for $j = 1...d$ is computed with S_{-k}.

$\hookrightarrow C(\hat{f}_j^{S_{-k}}, S_k)$

So far it looks like Holdout

$S_L = S_{-k}$; $S_T = S_k$

I select

$$\hat{j} = \underset{j=1\ldots d}{\operatorname{argmin}} \left[\frac{1}{V} \sum_{k=1}^{V} C\left(\hat{f}_j^{S_{-k}}, S_k\right) \right]$$

score for each method j should give an idea of how close method j is to f_0.

here in turn a data is in the learning sample and in the transfer sample → average

Then I use as an estimator of f_0

$$\boxed{\hat{f}_{\hat{j}}(X)}$$

I use the whole sample to compute it.

the method I selected by V-fold cross-validation

V-fold is "an average" of V Hold-out. It is more stable.

V = 5 or 10 are the best choices... (V = n : leave-one-out method)

Other method

you can also use $\displaystyle \frac{1}{V} \sum_{k=1}^{V} \hat{f}^{S_{-k}}_{\hat{\jmath}}$

But this has problems especially because depending on the problem computing averages of estimators do not make sense.

IV What about testing?

Testing and model selection are different in the sense that they do not answer to the same question

$m = 1, \dots, M$ $\quad$ M different models

model selection $\longrightarrow$ you will always get a model $\hat{m}$

$\longrightarrow$ this is the one which is the "best" in the sense of bias/variance equilibrium

$\longrightarrow$ It does not select necessarily the true model!

But only one which is not too far and which has a reasonable nb of parameters given your data.

goodness-of-fit tests:

they are testing H_0: my model m is true

vs H_1: ——— false

e.g. Shapiro and Wilk's test of gaussianity

If you know the models well enough, you can compute one for each model. Do ~~not~~ forget to correct for multiplicity with Bonferroni.

The answer will be
→ this m_1 is plausible e.g. $\mathcal{N}(m, \sigma^2)$
→ this other m_2 is plausible e.g. $\mathcal{E}(\lambda)$.
→ this m_3 is not likely e.g. Uniform ...

→ So this may select no model at all or models that are not compatible!!

When you have $\Delta_1, \dots, \Delta_K$ tests of level α $\longrightarrow$ Bonferroni α/K

then $P(\exists$ one test which wrongly rejects$)$

$$\leq \sum_{k=1}^{K} P(\Delta_i \text{ wrongly rejects})$$

$$\leq \sum_{k=1}^{K} \alpha = K\alpha \longrightarrow \leq \alpha.$$

With $K=15$ $\rightarrow$ one chance over 2 to make a mistake

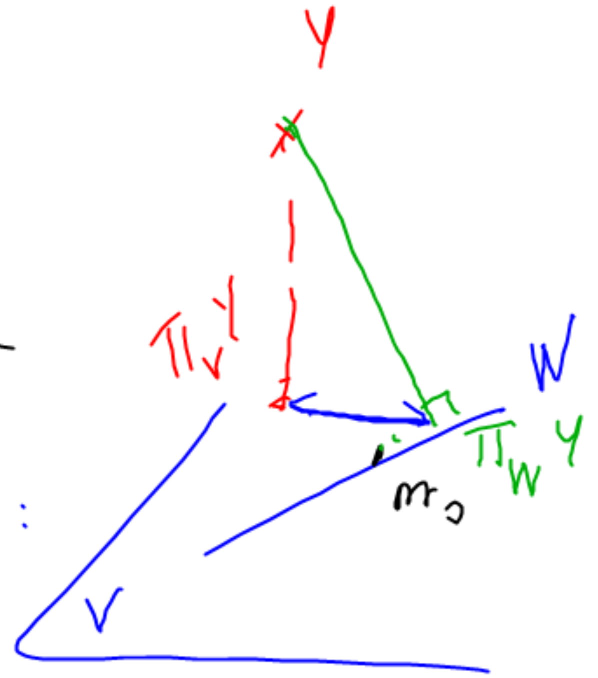
1/ Fisher test

In linear gaussian models.

$$Y = m + \varepsilon \qquad \cdot\; Y \in \mathbb{R}^n, \quad \varepsilon_i \sim \mathcal{N}(0, \sigma^2) \;\; \varepsilon_i \text{ iid}$$

$$\cdot\; m \in V \subsetneq \mathbb{R}^n$$

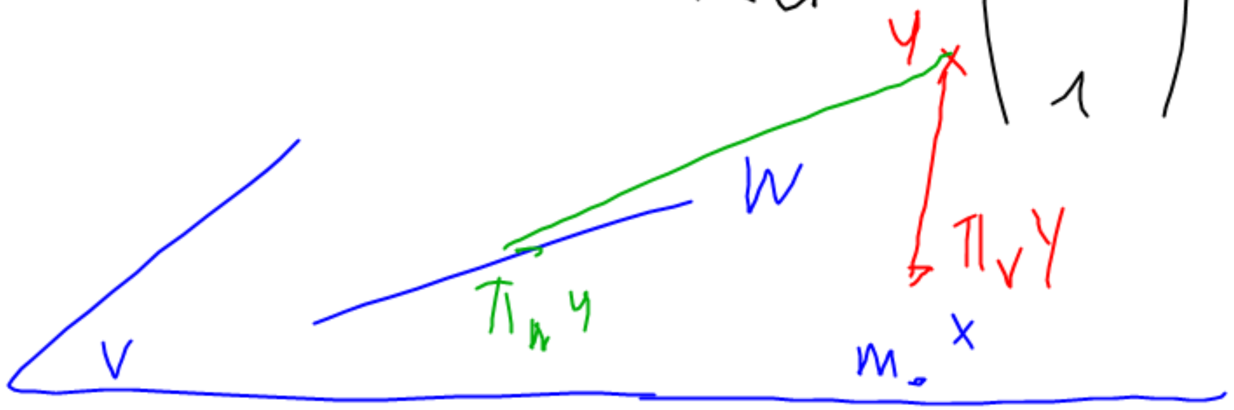
H_0 :



$$H_0 : m \in W \quad \text{vs} \quad H_1 : m \in V \setminus W \quad \text{where } W \subsetneq V.$$

For instance $\quad W = \mathrm{Vect}\begin{pmatrix} 1 \\ \vdots \\ 1 \end{pmatrix} \qquad V = \mathrm{Vect}\left(\begin{pmatrix} 1 \\ \vdots \\ 1 \end{pmatrix}, \begin{pmatrix} \cos u_1 \\ \vdots \\ \cos u_n \end{pmatrix}, \ldots, \begin{pmatrix} \cos(d u_1) \\ \vdots \\ \cos(d u_n) \end{pmatrix}\right)$

H_1



the Fisher test is based on the statistic

$$T = \frac{\|\Pi_V Y - \Pi_W Y\|^2}{\|Y - \Pi_V Y\|^2} \times \frac{n - \dim V}{\dim V - \dim W}$$

under H_0, T obeys a Fisher distribution $\mathcal{F}(\dim V - \dim W, n - \dim V)$

$\longrightarrow$ you reject when T is larger than the corresponding quantile $1-\alpha$.

$\longrightarrow$ you transform that into pvalues.

NB: in R, lm( )

pval one per variable

(——— pval) $\longrightarrow$ this is the pvalue of the Fisher test for $W = \mathrm{Vect}\begin{pmatrix}1\\ \vdots\\ 1\end{pmatrix}$

2] Wilk's theorem and likelihood ratio test

You have two models for your data $X = (X_1, \dots, X_n)$ iid.

model 1 model 2

$$\{P_\theta, \theta \in \Theta_0\} \subset \{P_K, K \in \mathcal{K}\}$$

ex: model 1 is $\{\mathcal{N}(m, 1), \underbrace{m}_{\theta} \text{ is unknown}\}$ $\Theta_0 = \mathbb{R}$.

model 2 is $\{\mathcal{N}(m, \sigma^2), \underbrace{m \text{ and } \sigma^2}_{K = (m, \sigma^2)} \text{ unknown}\}$. $\mathcal{K} = \mathbb{R} \times \mathbb{R}^+$

eg Transfer model

GCM

$$P(\underbrace{\text{object}}_{x} \text{ is put in class } A) = \frac{\sum_{y \in A} S(x,y)}{\sum_{y \text{ are learned}} S(x,y)}$$

$$S(x,y) = \exp(-c\, \underbrace{d(x,y)}_{= \sum_{i=1}^{3} w_i |x_i - y_i|})$$

$i=1 \leftrightarrow$ color

$i=2 \leftrightarrow$ shape

$i=3 \leftrightarrow$ size

model 1

$$\{ P_\theta ,\ \theta \in \Theta_0 \}$$

$$\theta = (c, \omega_1) \qquad c \in \mathbb{R}_+$$

$$\omega_1 \in [0,1] \quad \omega_2 = 1 - \omega_1 \qquad \omega_3 = 0$$

model 2.

$$\{ P_k ,\ k \in \mathbb{K} \}.$$

$$k = (c, \omega_1, \omega_2) \qquad c \in \mathbb{R}_+.$$

$$(\omega_1, \omega_2) \in [0,1]^2 \text{ st } \omega_1 + \omega_2 < 1$$

$$\omega_3 = 1 - \omega_1 - \omega_2$$

$\longrightarrow$ So the test will here answer to the question "is the size important for the categorization?"

The likelihood ratio statistic is given by

$$T = \frac{\overset{\text{model 2}}{\max_{k \in K} L_k(X)}}{\underset{\text{model 1}}{\max_{\theta \in \Theta_0} L_\theta(X)}}$$

$\max_{k \in K} L_k(X) \longrightarrow$ how plausible X is under model 2.

$\max_{\theta \in \Theta_0} L_\theta(X) \longrightarrow$ — " — model 1

if T is large you reject H_0: model 1 holds because model 2 seems more plausible

$$T = \frac{\overset{\text{model 2}}{L_{\hat{k}}(X)}}{\overset{\text{model 1}}{L_{\hat{\theta}}(X)}}$$

where $\hat{k}$ is the MLE in model 2

$\hat{\theta}$ ——————— 1

equivalently you reject when

$$W = 2\left(\underbrace{\ell_{\hat{K}}(X)}_{\text{loglikelihood in model 2}} - \underbrace{\ell_{\hat{\Theta}}(X)}_{\text{log likelihood in model 1}} \right)$$

Wilk's thm says that under H_0 $\quad W \xrightarrow[n \to +\infty]{\text{distrib}} \chi^2(d)$

We reject when W is larger than the corresponding quantile

$\longrightarrow$ pvalues.

d = nb of param in model 2 minus nb of param in model 1

$d = \dim(\text{model } 2) - \dim(\text{model } 1)$

In practice, for intricate models, being sure of the nb of parameters can be intricate

ex you could have parametrized the transfer model with

(c, ω_1, ω_2) in model 1 (forgetting that $w_1 + w_2 = 1$)

(c, w_1, w_2, w_3) —— 2 (———— $w_1 + w_2 + w_3 = 1$)

$$\exp\left(-c\left[\omega_1 d_1(x,y) + \omega_2 d_2(x,y) + \omega_3 d_3(x,y)\right]\right)$$

$$\exp\left(-\left[c_1 d_1(x,y) + c_2 d_2(x,y) + c_3 d_3(x,y)\right)\right)$$

My advise before using this test

→ perform simulation

→ verify that under H_0, pvalues are uniform

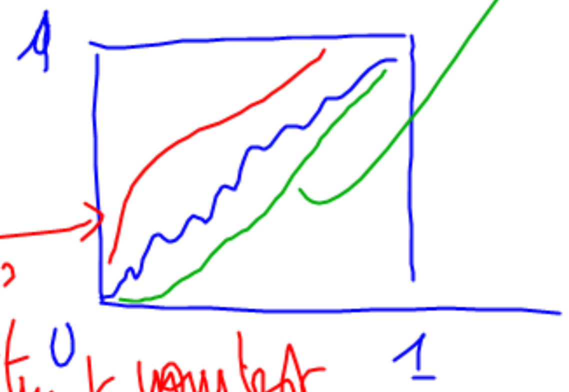
→ ecdf should be diagonal (or under the diagonal) to guarantee the level of the test.

simulate Nsimu times $(X_1 .. X_n)$ under model 1. (under H_0)

Compute each time $W^{(X_1 .. X_n)} \to$ pvalue.

→ Nsimu pvalues → ecdf

you have too much small pval under H_0 you cannot trust your test



you are smaller than uniform → you still control the level of the test

(under H_0 pval are uniform generally)

1 param that you don't know

eg $\quad X_1 \dots X_n$ simulate $\overset{iid}{\text{as}}$ $\mathcal{N}(m, 1)$ $\quad$ model 1

$$W = 2\left(\ell^{\text{model } 2}_{\hat{K}}(X) - \ell^{\text{model } 1}_{\hat{\theta}}(X)\right)$$

2 param unknown

$$\hat{\theta} = \hat{m} = \bar{X}$$

$$\hat{K} = \left(\bar{X}, \overbrace{\frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^2}^{\hat{\sigma}^2}\right) \qquad \left(\text{model 2: } \mathcal{N}(m, \sigma^2)\right)$$

$$\ell_{\hat{K}}(X) = -\frac{\sum_{i=1}^{n}(X_i - \bar{X})^2}{2\hat{\sigma}^2} - \frac{1}{2}\log(2\pi\hat{\sigma}^2) = -\frac{n}{2} - \frac{1}{2}\log\left(\frac{2\pi}{n}\sum_{i=1}^{n}(X_i - \bar{X})^2\right)$$

$$\ell_{\hat{\theta}}(X) = -\frac{\sum_{i=1}^{n}(X_i - \bar{X})^2}{2 \times 1} - \frac{1}{2}\log(2\pi \times 1)$$

(Here I know $\sigma = 1$ in model 1)

3/ Bootstrap

example without bootstrap

$$X \sim \mathcal{E}(\theta_0)$$

I'm interested in the distribution of $\left|\frac{1}{X} - \theta_0\right|$ where $X \sim \mathcal{E}(\theta_0)$

Nsimu $\quad X_1, \dots, X_{Nsim} \sim \mathcal{E}(\theta_0)$

$$T_i = \left|\frac{1}{X_i} - \theta_0\right| \quad \text{for } i = 1, \dots, Nsimu.$$

→ histograms etc to have an idea of the density

→ ecdf ⟶ cdf

→ empirical quantile ⟶ quantile

Why would I need that?

if I observe $X_1 \sim \mathcal{E}(\theta_0)$

then my estimator of θ_0 would be $\dfrac{1}{X_1} =$

and now you want to know $\longrightarrow$ Confidence interval on θ_0

$\longrightarrow$ make tests

$\rightsquigarrow$ you need a distribution for how far is $\hat{\theta}$ from θ_0

$\rightsquigarrow$ But you don't know θ_0, so what would you do?

$\longrightarrow$ Bootstrap But you need n observation to do that

a) parametric bootstrap

ex $X_1, \dots, X_n \sim \mathcal{E}(\theta_0)$ observations

$\longrightarrow$ $\hat{\theta} = \frac{1}{\overline{X}}$

$\longrightarrow$ you would like to know the distribution of $|\hat{\theta} - \theta_0|$ to compute for instance CI.

ε the $1 - \alpha$ quantile of this dist

and my CI would be $[\hat{\theta} \pm \varepsilon]$

$\longrightarrow$ I cannot do that because my dist° depends on θ_0 and I don't know it $\longrightarrow$

$\longrightarrow$ OK so I simulate Nsimu times

typical notation for bootstrap sample

$X_1^*, \dots, X_n^*$ iid $\mathcal{E}(\hat{\theta})$

the one that is computed with the original observations

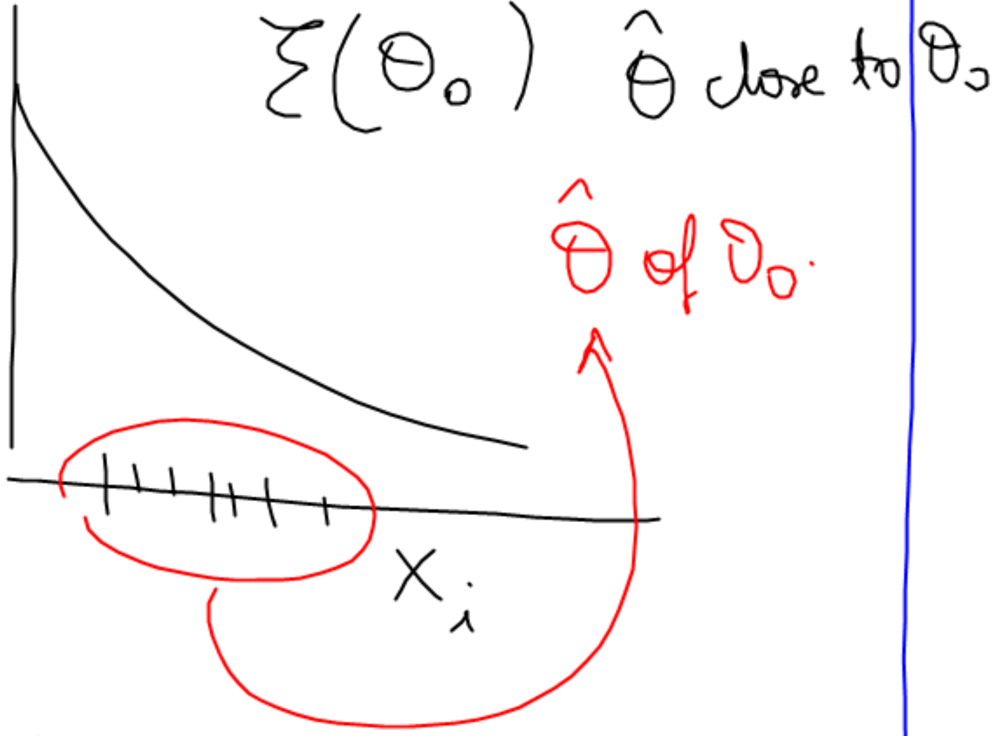
$\longrightarrow$ I compute the bootstrap version of $\hat{\theta}$: $\hat{\theta}_i^*$ the i^{th} simulation

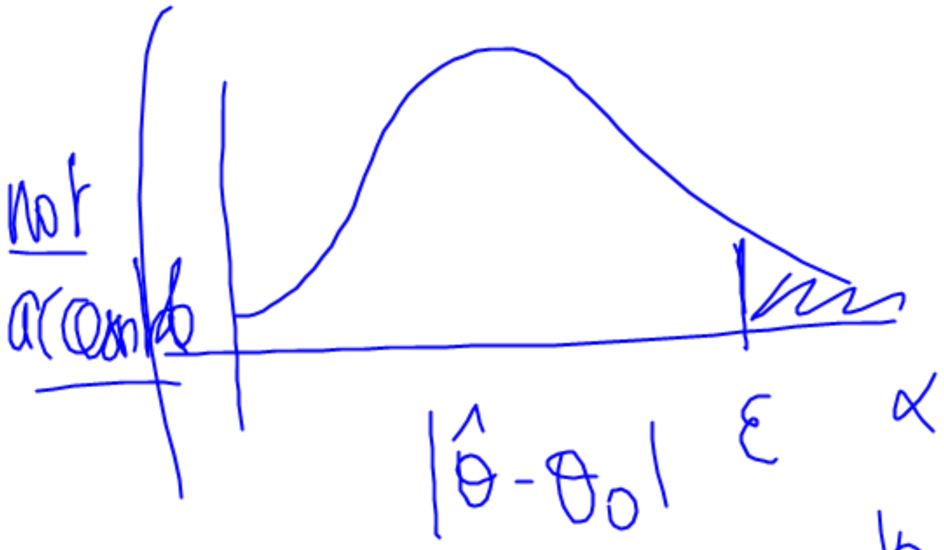
$\longrightarrow$ $T_i^* := |\hat{\theta}_i^* - \hat{\theta}|$ is the surrogate of $|\hat{\theta} - \theta_0|$, that you cannot access

$\longrightarrow$ you can use the Nsimu T_i^* to get cdf, quantiles etc ...

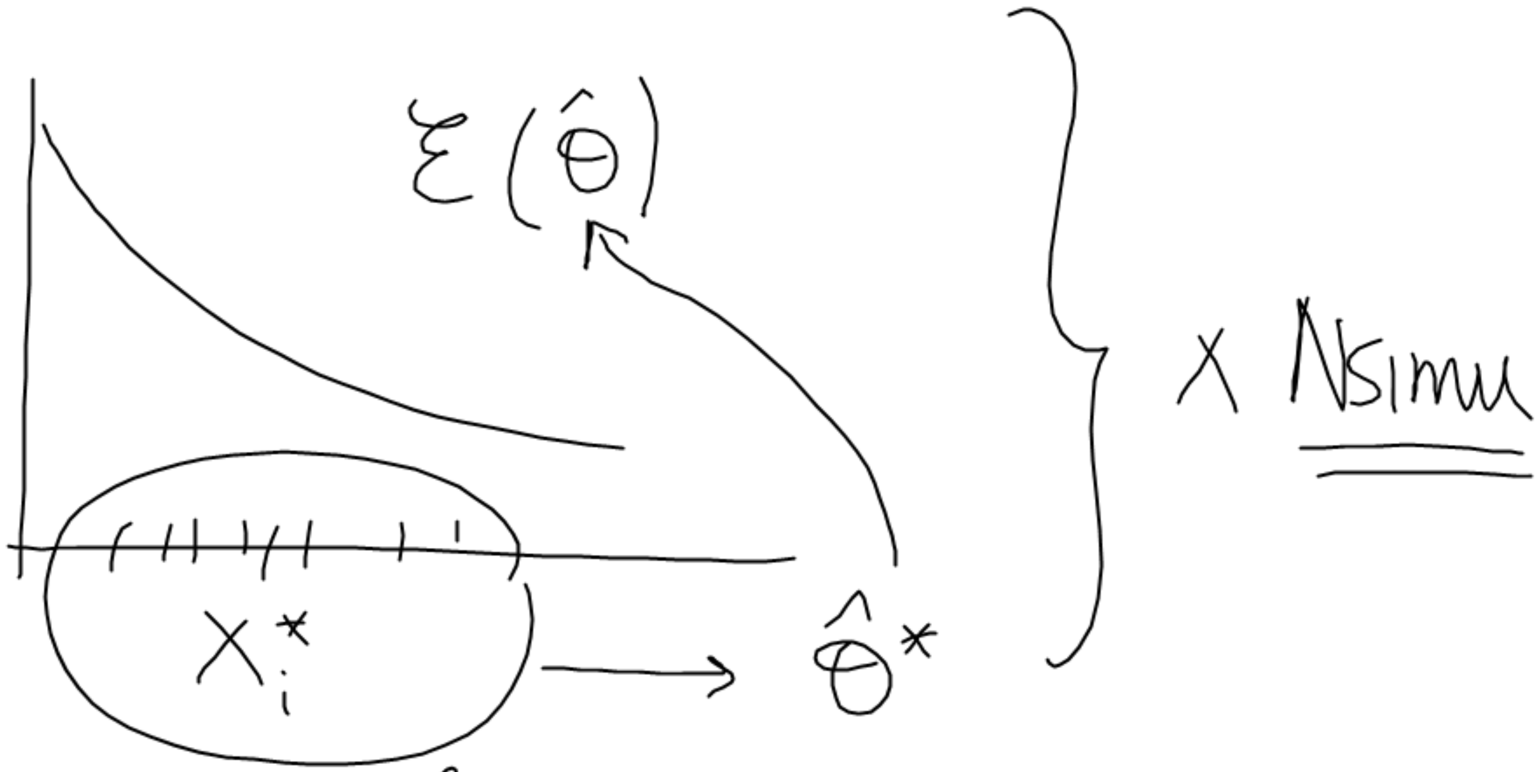
$\longrightarrow$ for instance with the $1-\alpha$ quantile ε^* of this bootstrapped distribution you get the bootstrap confidence interval $[\hat{\theta} \pm \varepsilon^*]$

Original world



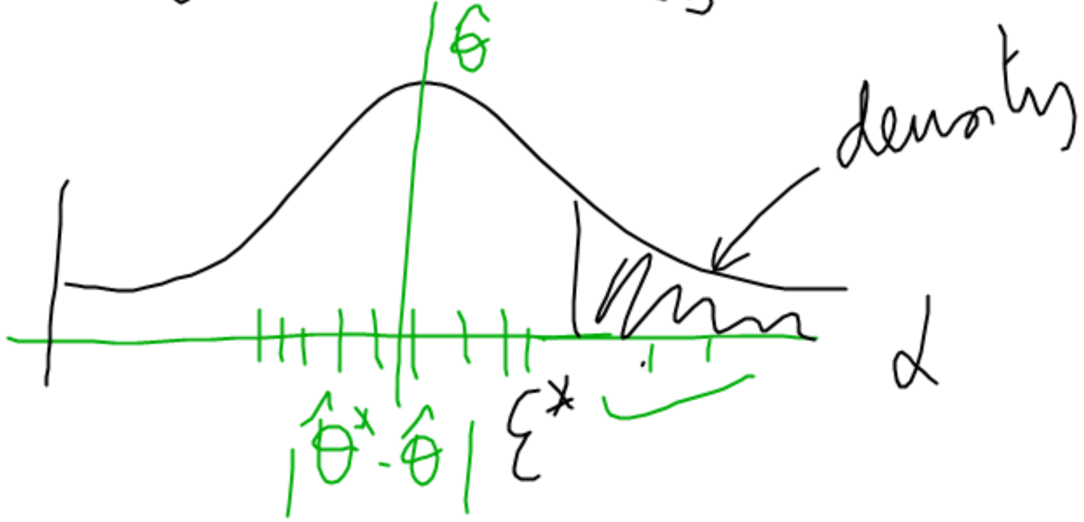


Bootstrap world



you hope that $\hat{\theta}^*$ is close to $\hat{\theta}$ i the same way as $\hat{\theta}$ is close to θ_0.

$\rightarrow$ Nsimu times $\hat{\theta}^*$



But if n is large, ε^* should be close to ε.

In general,

given a model with parameter θ and data $X_1, \dots, X_n$ observed and thought to come from this model with unknown parameter θ_0.

→ propose an estimator $\hat{\theta}$ of θ_0 (MLE, least square, empirical mean, you need $\hat{\theta} \xrightarrow[n \to +\infty]{} \theta_0$)

→ simulate $X_1^*, \dots, X_n^*$ with parameter $\hat{\theta}$. Nsimu times

→ compute each time $\hat{\theta}^* - \hat{\theta}$ (and then up to you to do distance, absolute value etc)

→ arrive at the empirical distribution of the quantity that you want
Bootstrap

→ use this empirical bootstrap dist° as if it was the one of $\hat{\theta} - \theta_0$.

(to build CI, test etc...)

⚠ DO NOT FORGET the centering

$$\boxed{-\theta_0 \longrightarrow -\hat{\theta}.}$$

Theories exist to show that it works but it combines 2 things $N_{simu} \to +\infty$
if n is not big enough, you will pay the fact that $\hat{\theta}$ is far from θ_0. $n \to +\infty$

b) non parametric bootstrap

When you don't have a model, you have at least data and you can always pick again in the sample to create new bootstrap sample.

$\longrightarrow$ bootstrap of the mean

$X_1, \dots, X_n$ iid with unknown mean $m = E(X)$

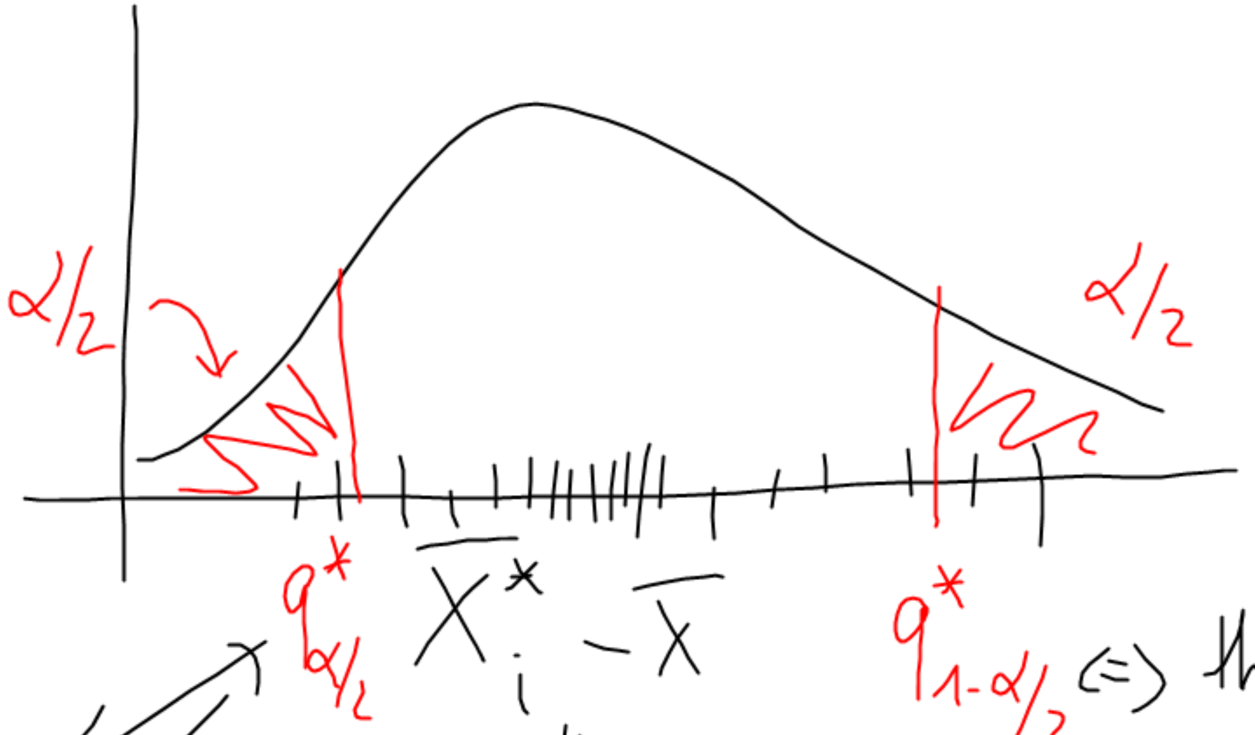
$\longrightarrow$ you can estimate m by $\overline{X}$

$\longrightarrow$ Nsimu times, you pick uniformly at random in $\{X_1, \dots, X_n\}$ (with replacement) to get $X_1^*, \dots, X_n^* \longrightarrow \overline{X^*}$

→ you get Nsim $\bar{X}_i^* - \bar{X}$

to approximate the distribution of $\bar{X} - m$.

for instance



$\bar{X}_i^* - \bar{X} = W_i^*$

the data W_i^* such that a fraction $\alpha/2$ is smaller than that.

$\Leftrightarrow$ the data W_i^* such that a fraction $\alpha/2$ of the data are bigger than that

in R command quantile will do the job.

→ the bootstrap CI at confidence level $1-\alpha$ is $[\bar{X} + q^*_{\alpha/2}, \bar{X} + q^*_{1-\alpha/2}]$

V Other methods with independance

1) Supervised classification

⇝ Deep Learning

2) Unsupervised classification / clustering

* k-means needs k = the nb of clusters
* to estimate k → BIC criterion
  → hierarchical clustering