- Le Cor de Gabriel
- Volume de la sphère et coquille
- Le lasso
- Marches aléatoires sur $\mathbb{Z}^d$

Mathieu Ribatet

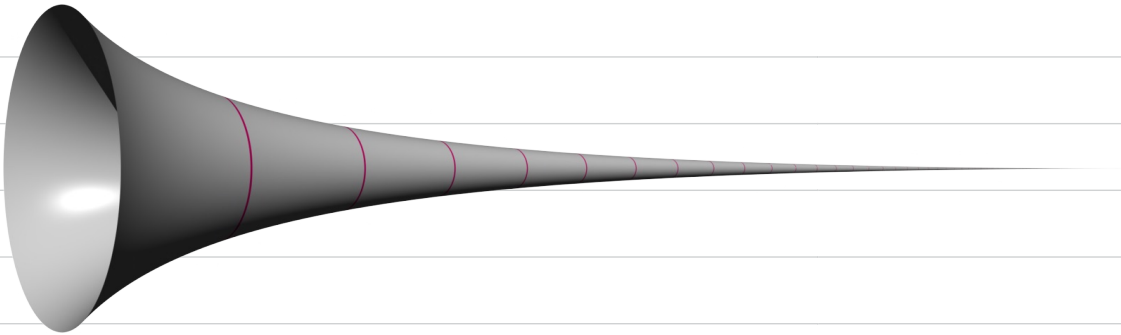
A Nantes depuis Sept. 2019

Montpellier

Stats → Modélisa°

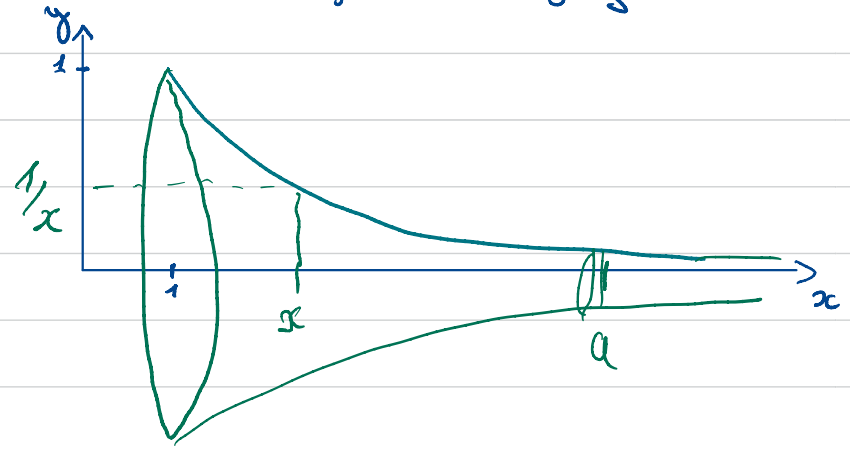
↳ Spatiale, Extrêmes

Le Cor de Gabriel



- Figure inventée par Evangelista Torricelli (XVII ème).
- Référence à l'Archange Gabriel annonçant le jugement dernier

- Obtenue comme un **solide de révolution** via l'équation $y = 1/x$, $x \geq 1$.
- A suscité beaucoup de discussions à l'époque (le calcul intégral n'existait pas encore...)

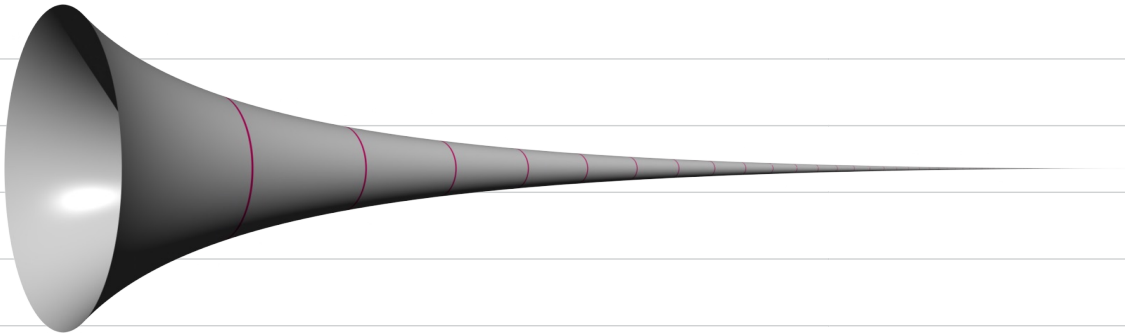


$$V(a) = \int_1^a \pi \left(\frac{1}{x}\right)^2 dx = \pi \left[-\frac{1}{x}\right]_1^a = \pi\left(1 - \frac{1}{a}\right) \longrightarrow \pi \quad a \longrightarrow \infty$$

et pourtant de longueur ∞...

$$S(a) = \int_1^a 2\pi \frac{1}{x}\, dx = 2\pi \left[\ln x\right]_1^a = 2\pi \ln a \longrightarrow \infty \quad , \quad a \longrightarrow \infty$$

Le Cor de Gabriel

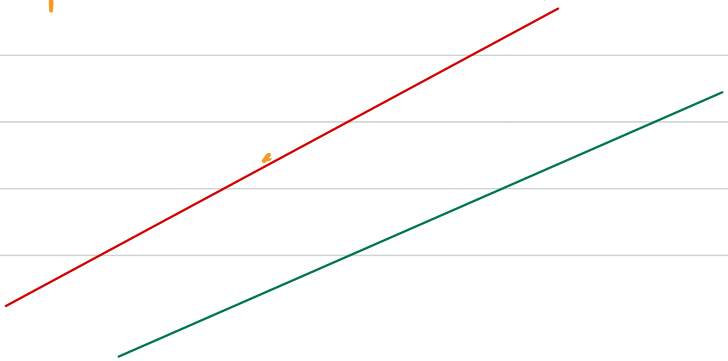


- Figure inventée par Evangelista Torricelli (XVII ème).
- Référence à l'Archange Gabriel annonçant le jugement dernier

- Montre les limites de l'analogie "surface = quantité de peinture". En effet, la couche de peinture aura une certaine épaisseur e et l'on ne pourra "plus peindre" le cor pour $x > 1/e$.

- Peut servir d'ouverture à la géométrie non Euclidienne (ici hyperbolique), i.e, ne supposant pas le postulat des parallèles d'Euclide

"Par un point extérieur à une droite, il passe toujours une parallèle à cette droite, et une seule."



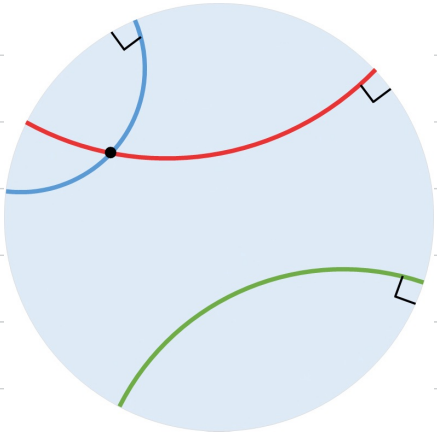


Figure : Disque de Poincarré

Volume de la Sphère

- On s'intéresse à la boule (euclidienne) $S_d(r) = \{x \in \mathbb{R}^d : \|x\|_2 \leq r\}$.
- Son volume vaut :
  - $d = 1$: $V_1(r) = \int_{\mathbb{R}} 1_{\{\|x\|_2 \leq r\}} dx = \int_{-r}^{r} dx = 2r$
  - $d = 2$: $V_2(r) = \int_{\mathbb{R}^2} 1_{\{\|x\|_2 \leq r\}} dx = \pi r^2$
  - $d = 3$: $V_3(r) = \int_{\mathbb{R}^3} 1_{\{\|x\|_2 \leq r\}} dx = \frac{4}{3}\pi r^3$
  - $d = n$: $\boxed{V_n(r) = \frac{\pi^{n/2}}{\Gamma(\frac{n}{2}+1)} r^n,}$

où Γ est la Fonction Gamma ($\Gamma(x+1) = x\Gamma(x)$, $x \geq 0$)

$x! = x \times (x-1)!$

$\Gamma(x+1) = x \times \Gamma(x)$

$\Gamma : x \longmapsto (x-1)!$

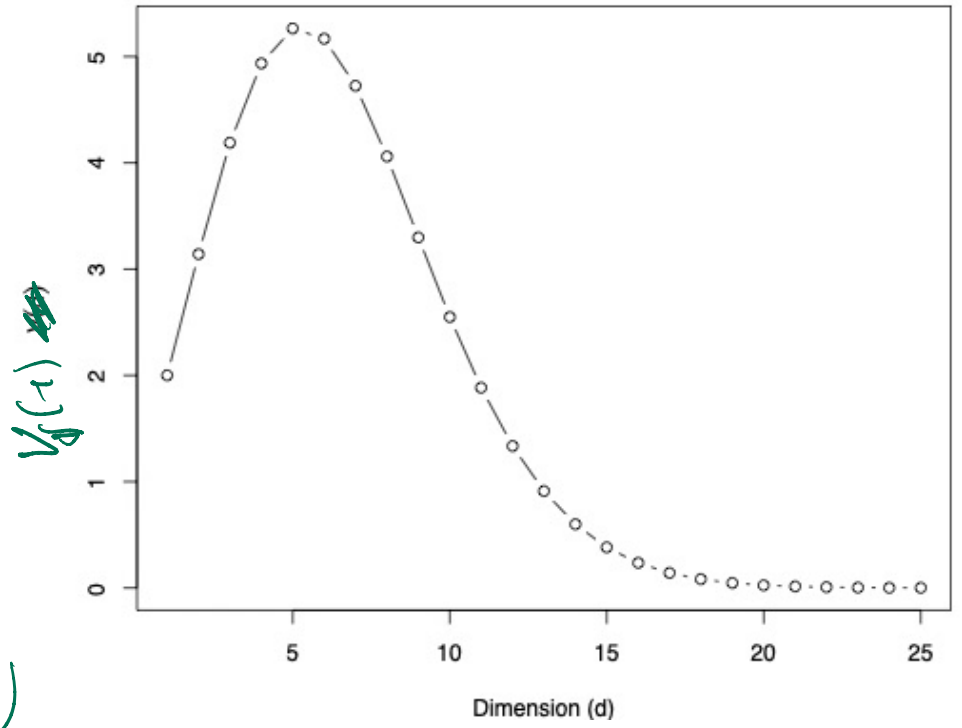


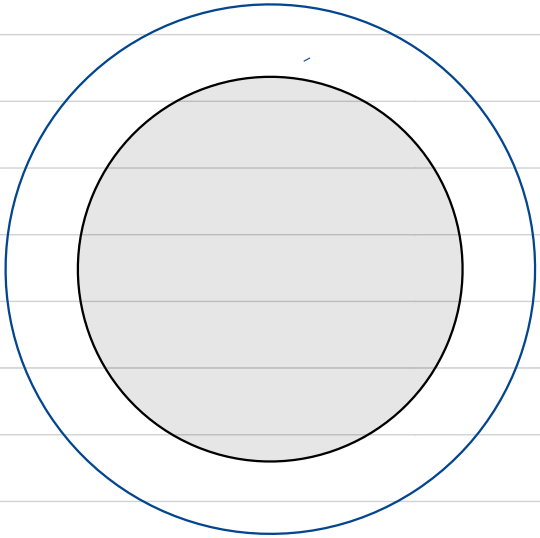
Figure : Volume de l'hyper-sphere unité.

Pour la sphère unité, le volume décroit lorsque $d > 5$!!!

Une Coquille

- Considérons maintenant la coquille

$$C_d(r) = S_d(1) \setminus S_d(r), \quad 0 < r < 1.$$



- Soit $X \sim \text{Uniforme}(S_d(1))$, i.e, $\mathbb{P}(X \in A) = \dfrac{\int_{S_d(1)} \mathbb{1}_{x \in A}\, dx}{V_d(1)}$

$$\mathbb{P}(X \in C_d(r)) = 1 - \mathbb{P}(X \in S_d(r))$$

$$= 1 - \frac{V_d(r)}{V_d(1)}$$

$$= 1 - r^d \xrightarrow{d \to \infty} 1$$

k - plus proches voisins

- Le fait que les espaces de grande dimension soient essentiellement vides est connu sous le terme "fléau de la dimension"

- Par exemple en statistique, une méthode célèbre de régression ou de classification est l'approche des k plus proches voisins

- Le fléau de la dimension nous dit alors que ça ne marchera pas

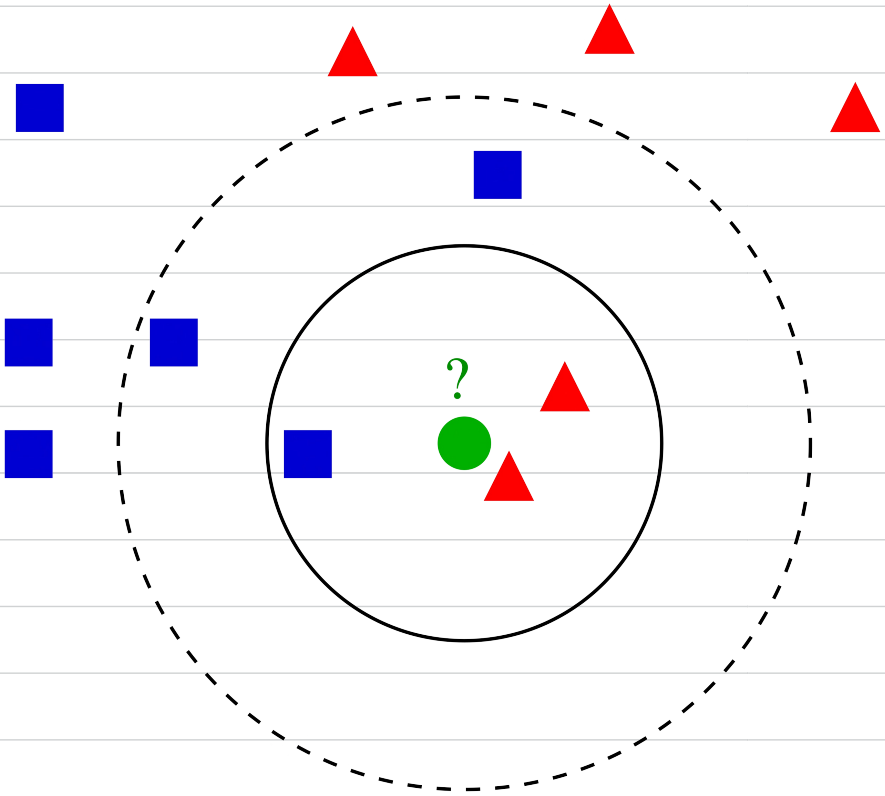


Figure : k plus proches voisins.

classifica

⇒ la statistique en grande dimension a fait de réelles avancées ces dernières années, notamment via la régularisation sparse (Lasso) ou l'utilisation du "kernel trick".

Le Problème de la régression

- On souhaite prédire une quantité d'intérêt Y conditionnellement à certaines caractéristiques X.
  Par exemple, prédire le salaire mensuel Y en fonction de l'âge X_1 et du niveau d'étude X_2.

- Mathématiquement cela revient à déterminer une fonction f_* solution de

$$f_* = \underset{f}{\operatorname{argmin}}\, \mathbb{E}\left[\|Y - f(X)\|_2^2\right]$$

  ← sous la loi de (X, Y)

- C'est un problème généralement trop complexe et l'on supposera souvent une structure particulière sur la classe des fonctions f.

- Un choix très populaire est la classe des fonctions linéaires, i.e, $f(X) = X^T\beta$ où β sont des coefficients à déterminer.

- Une autre difficulté est que l'espérance précédente est inconnue. On s'intéressera au problème approché suivant :

$$\hat{\beta} = \underset{\beta \in \mathbb{R}^p}{\operatorname{argmin}} \sum_{i=1}^{n} \|Y_i - X_i^T\beta\|_2^2 = \|\mathbb{Y} - \mathbb{X}\beta\|_2^2$$

- C'est généralement un problème d'optimisation **sympatique** pour lequel on a :

$$\nabla_\beta \|Y - X\beta\|_2^2 = 0 \iff -2X^T(Y - X\beta) = 0 \iff \beta = (X^TX)^{-1}X^TY$$

- Sympatique sauf si la matrice X^TX n'est pas inversible.

- Ceci arrive par exemple en grande dimension, i.e, lorsque le nombre de colonnes p de X est plus grand que le nombre de lignes n.

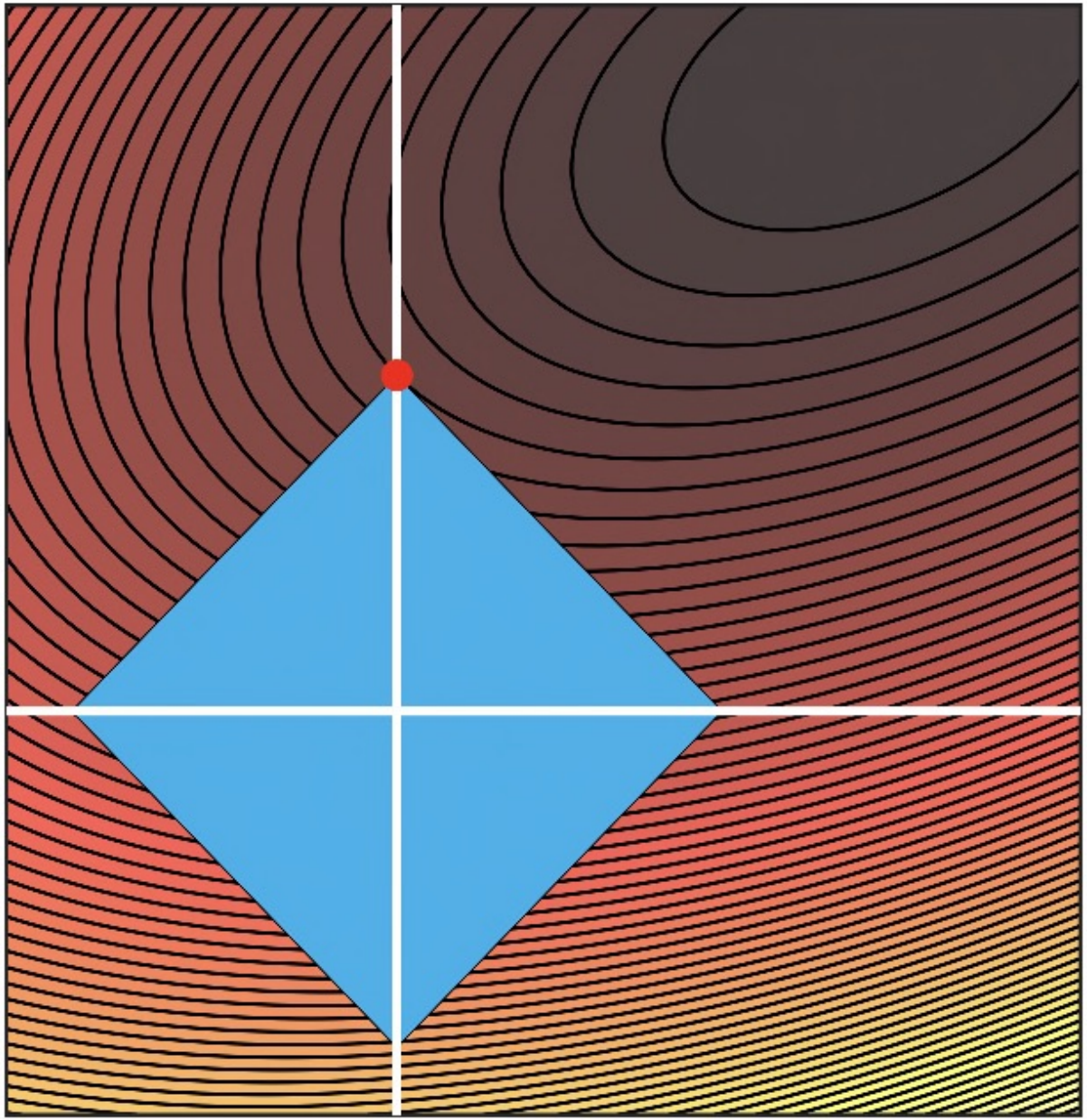
- Le lasso est une technique de régularisation contournant ce problème. Le problème d'optimisation est modifié en ajoutant une pénalisation (de type ℓ_1).

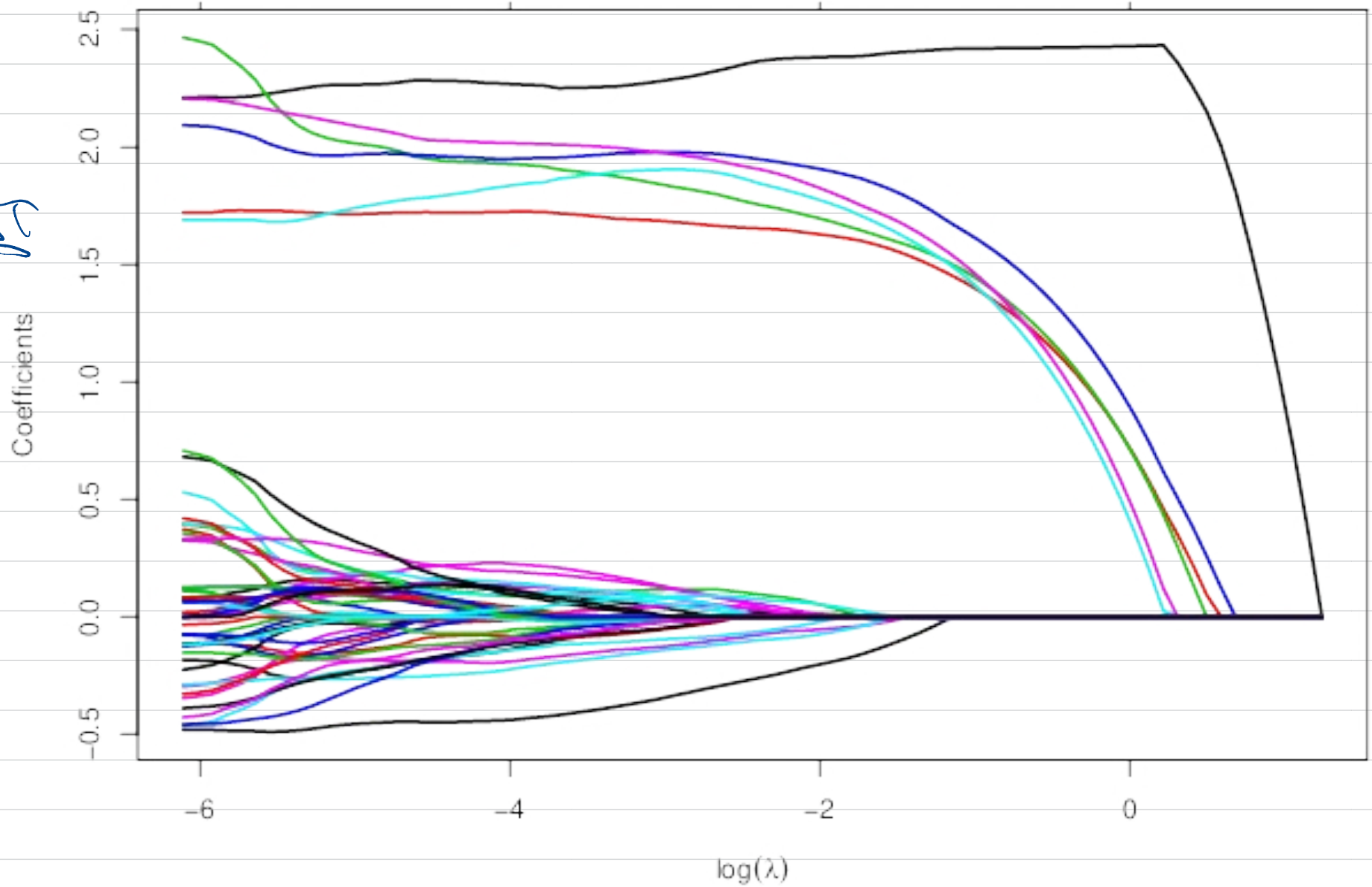
$$\hat{\beta}_\lambda = \underset{\beta \in \mathbb{R}^p}{\arg\min} \|Y - X\beta\|_2^2 + \lambda \|\beta\|_1 \quad , \quad \lambda \geq 0.$$

- La solution $\hat{\beta}_\lambda$ dépend de l'hyperparamètre λ à calibrer.

- Deux cas limites :
  - $\lambda = 0$: régression linéaire classique, i.e, $\hat{\beta}_0 = (X^TX)^{-1}X^TY$
  - $\lambda \to \infty$: solution totalement creuse, i.e, $\hat{\beta}_\infty = 0 \in \mathbb{R}^p$.

- Pas de solutions explicite ($\|\cdot\|_1$ non différentiable) mais des algorithmes très efficaces !!!




Coefficients
2.5
2.0
1.5
1.0
0.5
0.0
−0.5
−6
−4
−2
0
log(λ)

Logement à Boston

- Données classiques en apprentissage statistique
- C'est un problème de régression sur le prix médian d'une maison autour de Boston
- On dispose de $n = 506$ logements et $p = 14$ covariables parmi lesquelles :

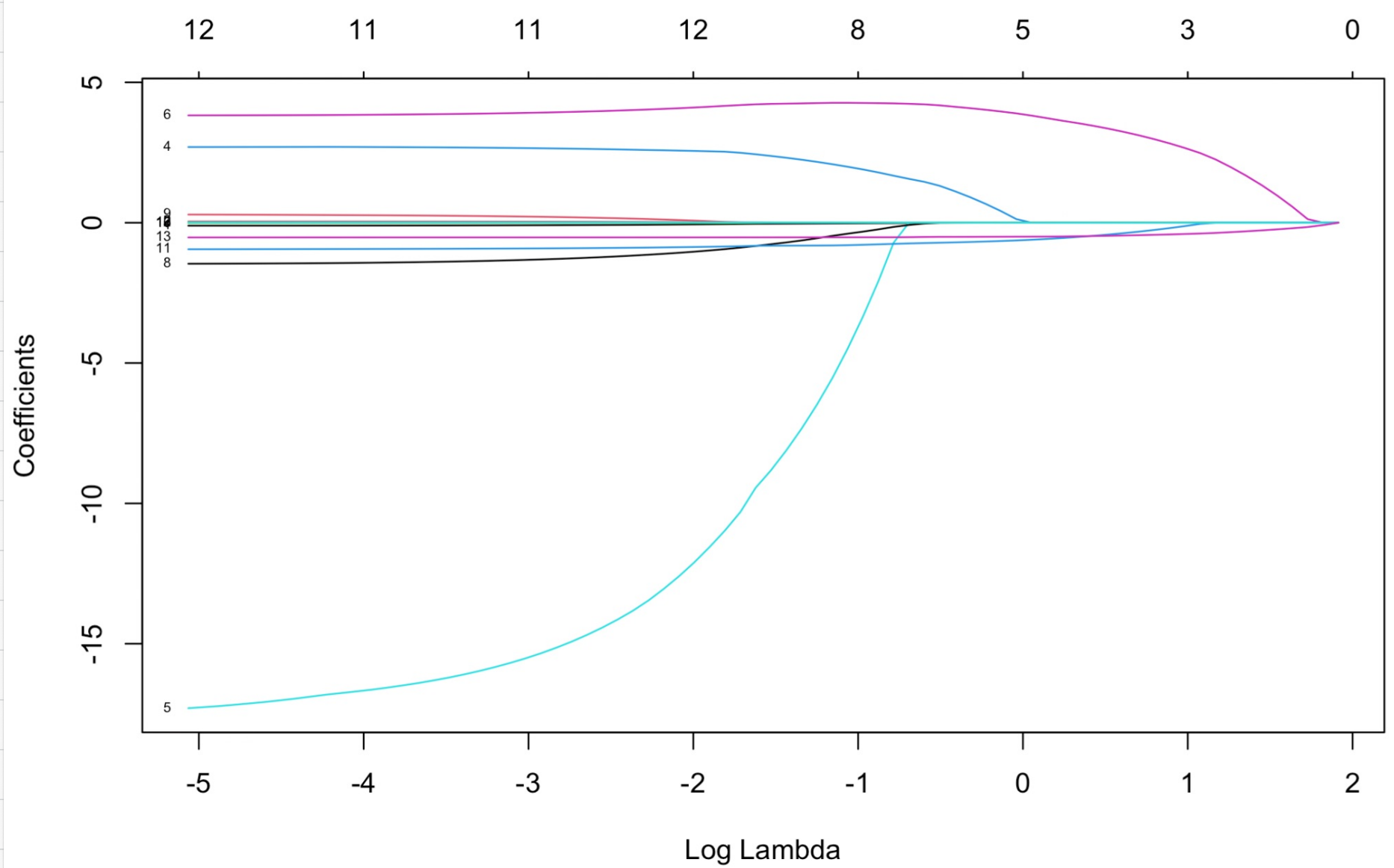
1) le nombre de pièces (rm)
2) le nombre d'élèves par professeur (ptratio)
3) la proportion d'ouvriers et/ou de personnes n'ayant pas été au lycée (lstat)

Console | Terminal × | Jobs ×

R 4.1.0 · ~/

```
> coef(fit, s = 3)
14 x 1 sparse Matrix of class "dgCMatrix"
                    s1
(Intercept) 12.6503687
crim         .
zn           .
indus        .
chas         .
nox          .
rm           2.4271190
age          .
dis          .
rad          .
tax          .
ptratio     -0.0308844
black        .
lstat       -0.3794441
>
```

Plot Zoom


12
11
11
12
8
5
3
0
6
4
9
13
11
8
5
5
0
-5
-10
-15
Coefficients
-5
-4
-3
-2
-1
0
1
2
Log Lambda

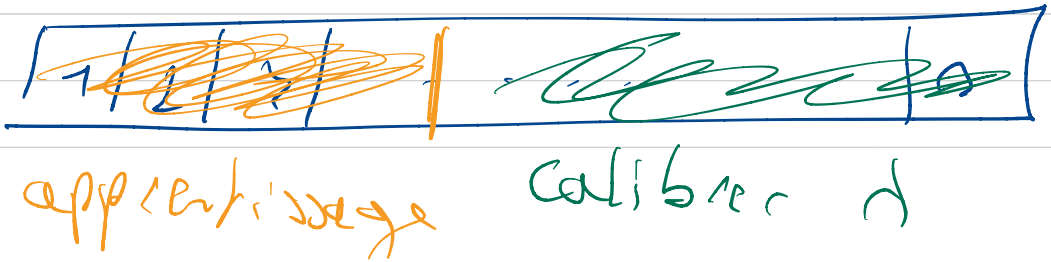
Contrôle de la pénalisation

$$\hat{\beta}_\lambda = \underset{\beta \in \mathbb{R}^p}{\arg\min} \; \|Y - X\beta\|_2^2 + \lambda \|\beta\|_1 \quad , \quad \lambda \geq 0.$$

- Clairement la solution $\hat{\beta}_\lambda$ dépend du choix de λ
- Comment choisir un λ "optimal" ???
- Puisque nous souhaitons prédire Y en fonction de X, il faudrait "choisir" λ comme celui minimisant l'erreur de prédiction

$$\lambda_* = \underset{\lambda > 0}{\arg\min} \; \mathbb{E}\left[(Y_* - X_*^T \hat{\beta}_\lambda)^2 \right] \quad , \quad (X_*, Y_*) \sim (X, Y)$$

- Comme précédemment, l'espérance est sous une loi inconnue et l'on considèrera un problème approché en divisant notre jeu de données initial, i.e,
  - train + validation + ~~test~~
  - validation croisée



Si le temps le permet...

Marches aléatoires sur $\mathbb{Z}$

- On se place en 0. On va tirer successivement et indépendamment une pièce équilibrée. Si "Pile" $\rightarrow +1$, si "Face" $\rightarrow -1$.

- Soit la variable aléatoire N_n = nombre de piles parmi les n tentatives. Clairement $N_n \sim$ et donc $\mathbb{P}(N_n = k) =$ $(0 \leq k \leq n)$

- La position après n lancers est $S_n =$ On en déduit donc que

$$\mathbb{P}(S_n = k) =$$

- On peut aussi s'intéresser au retour en l'origine

$$p_{2n} = \mathbb{P}(S_{2n} = 0) =$$

- De plus, le nombre de passages en 0 $E_\infty = \sum_{k \geq 1} \mathbb{1}_{\{S_k = 0\}}$ vérifie

$$\mathbb{E}(E_\infty) = \sum_{k \geq 1} \mathbb{P}(S_{2k} = 0) = \sum_{k \geq 1} \binom{2k}{k} 2^{-2k} = +\infty$$

$$\left(\text{en effet } \frac{(2k)!}{k!\,k!} \sim \frac{\sqrt{2\pi(2k)}\left(\frac{2k}{e}\right)^{2k}}{\left(\sqrt{2\pi k}\left(\frac{k}{e}\right)^{k}\right)^2} \sim \frac{4^k}{\sqrt{\pi k}} \right)$$

- Le fait que $\mathbb{E}(E_\infty) = \infty$ signifie que l'on repasse une infinité de fois en 0 ; on dit alors que 0 est un état récurrent

- On peut généraliser cette marche aléatoire à $\mathbb{Z}^d$, $d \geq 2$.

- Etonnamment pour $d=2$, 0 est toujours récurrent mais pour $d \geq 3$ ce n'est plus le cas !!!

- On résumera cela en disant que "l'on se perd" dès que $d \geq 3$.

$\Rightarrow$ C'est notamment pour cela que l'algorithme de Metropolis-Hastings "peine un peu" dès lors que $d \geq 3$.