

Régression ridge

Séance 5 – Pénaliser pour mieux prédire

Christophe Ambroise 

UEVE, UMR CNRS 8071

12 octobre 2026

1

Régression

2

Classification vs régression

Classification

Y à valeurs finies, coût 0-1, règle optimale $\arg \max P(Y | X)$.

...

Régression

Y **continue** : coût quadratique, règle optimale :

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

3

Le modèle linéaire

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}, \quad \boldsymbol{\varepsilon} \sim \mathcal{N}(0, \sigma^2 \mathbf{I}_n).$$

Moindres carrés : critère et solution.

$$\hat{\boldsymbol{\beta}}_{\text{MCO}} = \arg \min_{\boldsymbol{\beta}} \|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|^2 = (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{y}.$$

4

Jeu de données exemple : **diabetes** de scikit-learn

Jeu **diabète** : 442 patients, 10 variables, progression de la maladie à un an.

<u>age, sex, bmi, bp</u>	données cliniques
<u>s1 ... s6</u>	six dosages sanguins

Variables **centrées-réduites** ; réponse continue, moyenne 152, écart-type 77.

5

Limites des moindres carrés

6

La variance de l'estimateur

$$\text{Var}(\hat{\beta}_{\text{MCO}}) = \sigma^2 (\mathbf{X}^\top \mathbf{X})^{-1}$$

Diagonalisation $\mathbf{X}^\top \mathbf{X} = \sum_i d_i^2 v_i v_i^\top$:

$$\text{Var}(v_i^\top \hat{\beta}) = \frac{\sigma^2}{d_i^2}.$$

Conditionnement $\kappa = (d_{\max}/d_{\min})^2$: sur le diabète, $\kappa = 470$.

Prédiction en $\mathbf{x}$: variance $\sigma^2 \sum_i (\mathbf{x}^\top v_i)^2 / d_i^2$.

Une valeur singulière proche de zéro : variance explosive dans **une** direction — prédiction instable.

7

Trois configurations

- **Colinéarité** : deux variables, presque la même mesure.
- **p proche de n** : trop peu d'observations par paramètre.
- **$p > n$** : $\mathbf{X}^\top \mathbf{X}$ non inversible ; $\hat{\beta}_{\text{MCO}}$ **inexistant**.

Troisième cas : la norme en génomique, imagerie, traitement du signal.

8

Sur le diabète

Six dosages sanguins fortement liés : **s1**, **s2** corrélés à **0,90**.

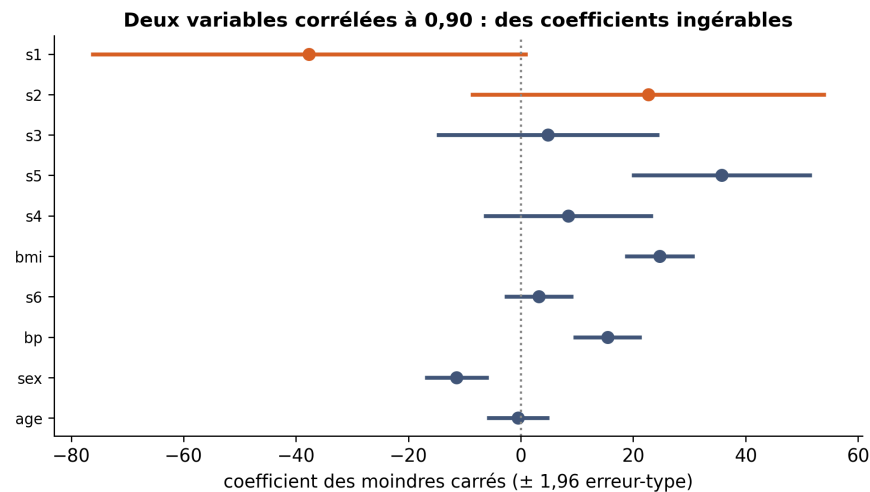


Figure 1

9

Lecture de la figure

Coefficients : **s1** à **-37,7**, **s2** à **+22,7** ; erreurs-types 19,8 et 16,1.

Ni l'un ni l'autre significatif ($|t| < 2$) ; **bmi** largement significatif ($t = 7,8$), coefficient plus petit.

Deux gros coefficients de signes opposés, presque annulés ; le modèle compense l'un par l'autre. **Signe** non déterminé par les données.

10

Retour sur la stratégie de régularisation

$$f^* = \arg \min_{f \in \mathcal{F}} \widehat{R}(f, D_n) + H(f)$$

$\mathcal{F}$ non restreint ; complexité rendue **coûteuse**.

Ridge : réalisation de la stratégie de régularisation pour le modèle linéaire. Autre réalisation en classification : élagage des arbres.

11

La régression ridge

12

Le critère

$$\widehat{\beta}_\lambda = \arg \min_{\beta} \left\{ \|\mathbf{y} - \mathbf{X}\beta\|^2 + \lambda \|\beta\|^2 \right\}, \quad \lambda \geq 0.$$

Un terme ajouté : **norme au carré** des coefficients. Un hyperparamètre : λ .

13

La solution

Gradient annulé :

$$-2\mathbf{X}^\top (\mathbf{y} - \mathbf{X}\beta) + 2\lambda\beta = 0 \iff (\mathbf{X}^\top \mathbf{X} + \lambda \mathbf{I})\beta = \mathbf{X}^\top \mathbf{y}.$$

$$\widehat{\beta}_\lambda = (\mathbf{X}^\top \mathbf{X} + \lambda \mathbf{I}_p)^{-1} \mathbf{X}^\top \mathbf{y}.$$

Différence unique avec les moindres carrés : **arête** ajoutée à la diagonale. D'où le nom.

14

Toujours inversible

Pour tout $\lambda > 0$ et **toute** matrice $\mathbf{X}$, la matrice $\mathbf{X}^\top \mathbf{X} + \lambda \mathbf{I}$ est définie positive, donc inversible.

Preuve. $\mathbf{X}^\top \mathbf{X}$ est semi-définie positive, de valeurs propres $d_i^2 \geq 0$. Celles de $\mathbf{X}^\top \mathbf{X} + \lambda \mathbf{I}$ valent $d_i^2 + \lambda > 0$. $\square$

Solution ridge **même pour $p > n$** , là où les moindres carrés n'en ont aucune. Non un raffinement : un changement de domaine d'existence.

15

L'intercept ne se pénalise pas

$$\min_{\beta_0, \boldsymbol{\beta}} \|\mathbf{y} - \beta_0 \mathbf{1}_n - \mathbf{X}\boldsymbol{\beta}\|^2 + \lambda \|\boldsymbol{\beta}\|^2$$

$\mathbf{X}$ centrée : solution séparée.

$$\hat{\beta}_0 = \bar{y}, \quad \hat{\boldsymbol{\beta}}_\lambda = (\mathbf{X}^\top \mathbf{X} + \lambda \mathbf{I})^{-1} \mathbf{X}^\top \tilde{\mathbf{y}}, \quad \tilde{\mathbf{y}} = \mathbf{y} - \bar{y} \mathbf{1}_n.$$

Pénaliser $\boldsymbol{\beta}_0$: prédiction dépendante de l'origine de $\mathbf{y}$. β_0 toujours exclu.

16

Échelle des variables

Pénalité $\|\boldsymbol{\beta}\|^2$: même traitement de toutes les coordonnées.

Centimètres $\rightarrow$ mètres : coefficient multiplié par 100, pénalité par 10 000.

Ridge : sens sur variables **standardisées** seulement. Même règle qu'en séance 3 pour les k -PPV, raison différente : là la distance, ici la pénalité.

17

Les deux extrémités

$\lambda \rightarrow 0$

$$\hat{\boldsymbol{\beta}}_\lambda \rightarrow \hat{\boldsymbol{\beta}}_{\text{MCO}}$$

Aucune contrainte.

$\lambda \rightarrow \infty$

$$\hat{\boldsymbol{\beta}}_\lambda \rightarrow \mathbf{0}$$

Le modèle prédit $\bar{y}$ partout.

Entre les deux, un continuum de modèles : le k des voisins, le α des arbres, autre habit.

18

Le chemin de régularisation

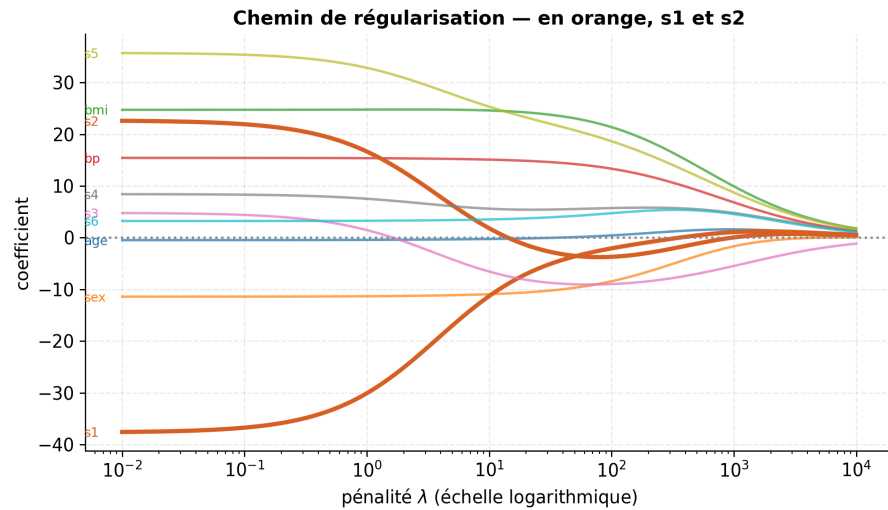


Figure 2

À gauche, **s1** à **-38**, **s2** à **+23** ; dès $\lambda \approx 10$, retour près de zéro, **bmi** presque inchangé. Ridge : cible prioritaire, les coefficients non déterminés par les données.

La forme contrainte

Problème pénalisé $\equiv$ problème sous contrainte :

$$\min_{\beta} \|y - X\beta\|^2 \quad \text{sous} \quad \|\beta\|^2 \leq t,$$

correspondance décroissante $\lambda \leftrightarrow t$.

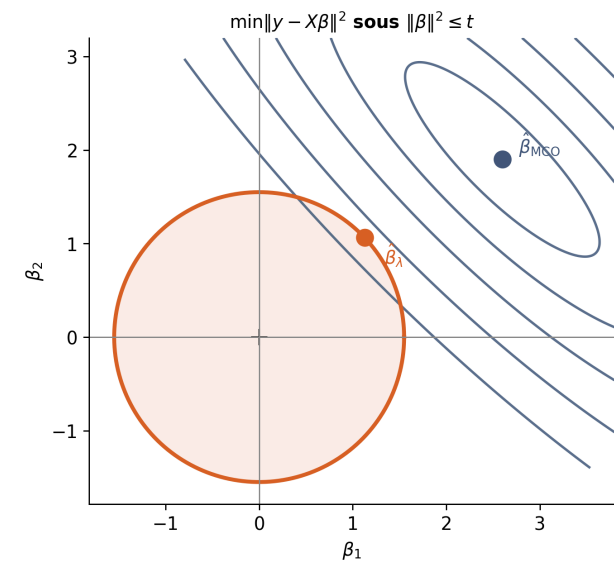


Figure 3

Ce que dit la théorie

21

Biais et variance, explicitement

Notation : $\mathbf{W}_\lambda = (\mathbf{X}^\top \mathbf{X} + \lambda \mathbf{I})^{-1}$.

$$\begin{aligned}\text{Biais}(\hat{\beta}_\lambda) &= -\lambda \mathbf{W}_\lambda \beta, \\ \text{Var}(\hat{\beta}_\lambda) &= \sigma^2 \mathbf{W}_\lambda \mathbf{X}^\top \mathbf{X} \mathbf{W}_\lambda.\end{aligned}$$

Biais **nul en $\lambda = 0$** , croissant avec λ . Variance totale $\sum_j \text{Var}(\hat{\beta}_j)$ **décroissante**.

Deux monotonies opposées : compromis forcé. Biais faible, variance forte $\rightarrow \lambda$ petit. Biais fort, variance faible $\rightarrow \lambda$ grand.

22

Le cas orthogonal

Hypothèse $\mathbf{X}^\top \mathbf{X} = n \mathbf{I}_p$:

$$\hat{\beta}_\lambda = \frac{1}{n + \lambda} \mathbf{X}^\top \mathbf{y} = \frac{n}{n + \lambda} \hat{\beta}_{\text{MCO}}.$$

Chaque coefficient **multiplié par un facteur** < 1 : contraction à l'état pur.

23

La MSE, en fermé

Avec $\text{MSE}(\hat{\beta}) = \frac{1}{n} \mathbb{E} \|\mathbf{X} \hat{\beta} - \mathbf{X} \beta\|^2$:

$$\text{MSE}(\hat{\beta}_\lambda) = \underbrace{\frac{\lambda^2 \|\beta\|^2}{(n + \lambda)^2}}_{\text{biais}^2} + \underbrace{\frac{n p \sigma^2}{(n + \lambda)^2}}_{\text{variance}}.$$

λ de 0 à ∞ : biais² de 0 à $\|\beta\|^2$, variance de $p\sigma^2/n$ (moindres carrés) à 0.

24

Le λ optimal

Dérivée annulée :

$$\lambda^* = \frac{p \sigma^2}{\sum_{j=1}^p \beta_j^2} = \frac{p \sigma^2}{\|\beta\|^2}.$$

Strictement positif pour $\beta \neq 0$.

Croissant avec le bruit et le nombre de variables, décroissant avec la force du signal. Chaque terme à sa place.

25

La MSE, en image

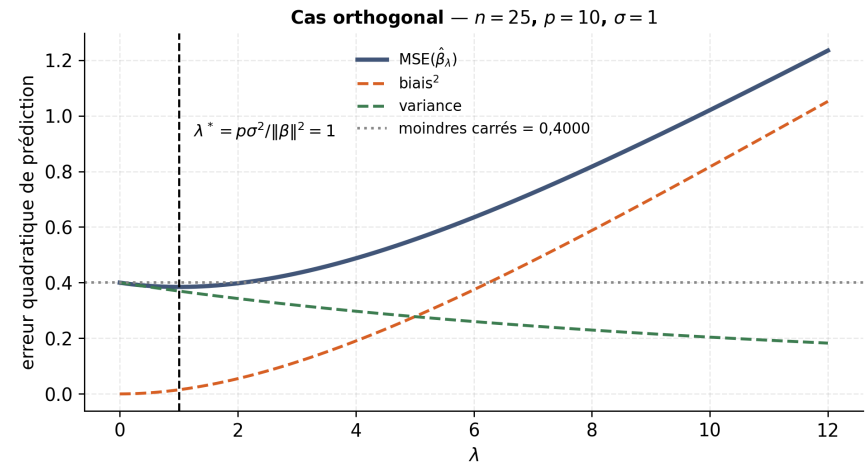


Figure 4

26

Le théorème d'existence

Théorème

Pour tout $\beta \neq 0$, il existe $\lambda^* > 0$ tel que

$$MSE(\hat{\beta}_{\lambda^*}) < MSE(\hat{\beta}_{\text{MCO}}).$$

Idée. $M(\lambda) = MSE(\hat{\beta}_\lambda)$ dérivable, $M'(0) < 0$: variance décroissante d'ordre λ , biais croissant d'ordre λ^2 . Biais incapable de rattraper la variance près de zéro.

27

Portée du théorème

Modèle linéaire **exactement vrai**, bruit exactement gaussien : l'estimateur non biaisé n'est **pas** le meilleur.

Un peu de biais volontaire : gain de variance supérieur au coût — au moins au départ.

Contre-intuitif : fondement de toute la régularisation moderne.

28

Limites du théorème

- **Quel** λ ? λ^* dépend de β et σ , tous deux inconnus.
- **Tout** $\lambda > 0$: aucune garantie d'amélioration.
- **Taille** du gain : parfois négligeable.

Théorème d'existence : garantit quelque chose à chercher, ne le trouve pas.

29

Ridge par la SVD

30

Une seule factorisation

Cent valeurs de λ : cent inversions $p \times p$. La SVD évite tout :

$$\mathbf{X} = \mathbf{U}\mathbf{D}\mathbf{V}^\top, \quad \mathbf{D} = \text{diag}(d_1, \dots, d_p).$$

$$\hat{\beta}_\lambda = \mathbf{V} \text{diag}\left(\frac{d_i}{d_i^2 + \lambda}\right) \mathbf{U}^\top \tilde{\mathbf{y}}.$$

31

Ridge est un filtre

Base des directions principales : chaque coordonnée **séparée**, multipliée par

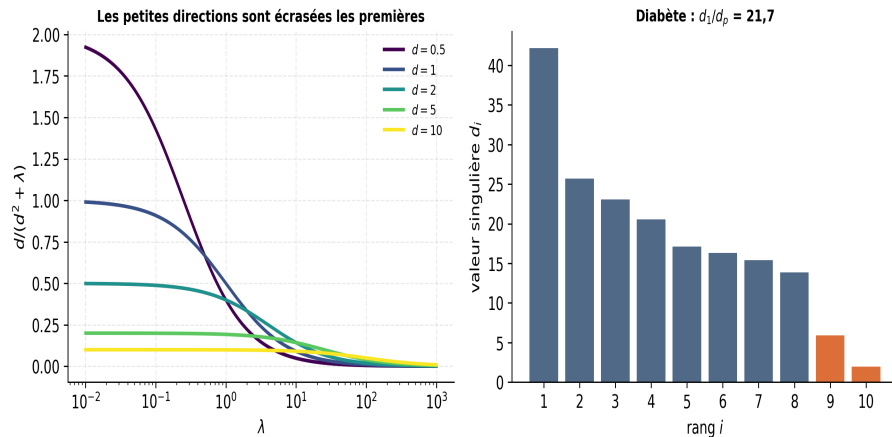
$$\frac{d_i}{d_i^2 + \lambda} \quad \text{au lieu de} \quad \frac{1}{d_i} \quad (\text{moindres carrés}).$$

d_i grand : facteur proche de $1/d_i$, direction intacte.

d_i petit : $1/d_i$ explosif ; $d_i/(d_i^2 + \lambda) \approx d_i/\lambda \rightarrow 0$.

32

Le filtre, en image



33

Sur le diabète

Valeurs singulières : $d_1 = 42,2$ à $d_{10} = 1,9$, rapport **21,7** ;
conditionnement de $\mathbf{X}^\top \mathbf{X}$: 470.

Deux dernières directions : colinéarité **s1–s2**, exactement celles que ridge écrase en premier.

Ridge : action sur les **directions**, non les **variables**. Même changement de base que l'ACP (séance 6) – amortir, non projeter.

34

Choisir λ

35

Degrés de liberté effectifs

Ridge, estimateur **linéaire** : $\hat{\mathbf{y}} = \mathbf{H}_\lambda \mathbf{y}$ avec
 $\mathbf{H}_\lambda = \mathbf{X}(\mathbf{X}^\top \mathbf{X} + \lambda \mathbf{I})^{-1} \mathbf{X}^\top$.

$$\text{df}(\lambda) = \text{tr}(\mathbf{H}_\lambda) = \sum_{i=1}^p \frac{d_i^2}{d_i^2 + \lambda}.$$

$\text{df}(0) = p$, $\text{df}(\infty) = 0$. Complexité **continue**, là où les arbres comptent des feuilles entières.

36

La complexité, en image

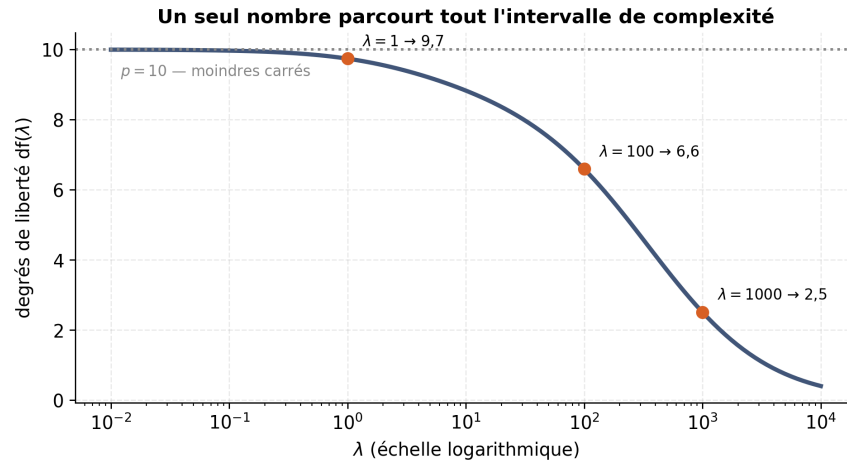


Figure 6

37

Trois façons de choisir

Critères d'information

$$AIC = n \log(\text{SCR}) + 2 \text{ df}$$

$$BIC = n \log(\text{SCR}) + \text{df} \log n$$

Validation croisée

K plis, méthode de la séance 1.
Aucune hypothèse.

Et la **validation croisée généralisée**, proche du laisser-un-de-côté, sans réajustement :

$$\text{GCV} = \frac{1}{n} \sum_i \left(\frac{y_i - \hat{y}_i}{1 - \text{tr}(\mathbf{H}_\lambda)/n} \right)^2.$$

38

La courbe de validation

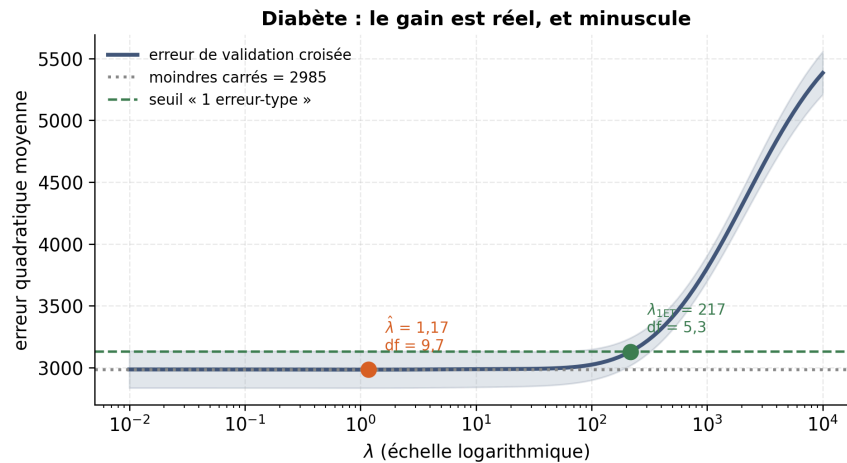


Figure 7

39

Le résultat sur le diabète

Diabète : $\hat{\lambda} = 1,17$, erreur **2 984** contre **2 985** pour les moindres carrés.

Gain : **0,05 %**. Théorème vérifié — et inutile sur ce jeu.

Règle du **1 erreur-type** (séance 4) : $\lambda = 217$, soit $\text{df} = 5,3$ — deux fois moins de complexité, 5 % d'erreur en plus.

40

Raison du faible gain

$n = 442$ pour $p = 10$: quarante-quatre observations par paramètre. Variance des moindres carrés déjà faible : presque rien à racheter.

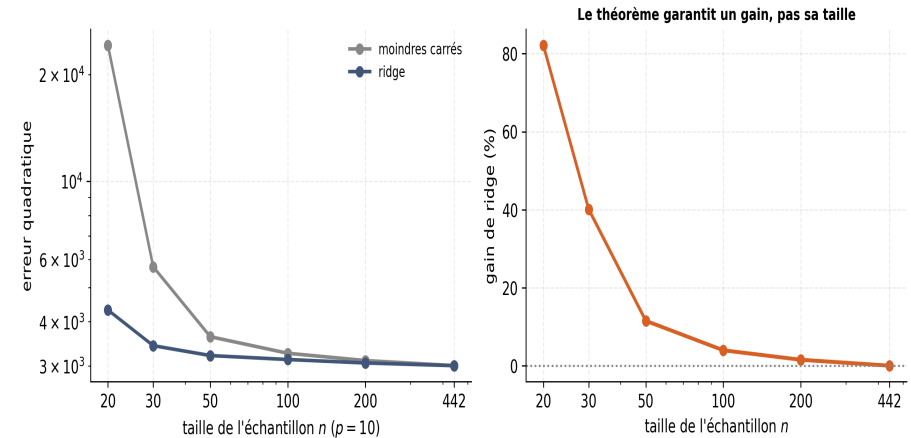


Figure 8

Le vrai régime de ridge

$n = 20$: erreur divisée par **cinq et demi** ($24\,200 \rightarrow 4\,300$).

n	20	30	50	100	200	442
gain	82 %	40 %	12 %	4 %	2 %	0 %

Bénéfice de ridge : fonction de p/n . Beaucoup de données, peu de variables : inutile. L'inverse : indispensable.

En scikit-learn

```
1 from sklearn.linear_model import Ridge, RidgeCV
2 from sklearn.pipeline import make_pipeline
3 from sklearn.preprocessing import StandardScaler
4
5 modele = make_pipeline(StandardScaler(), Ridge(alpha=1.0))
6
7 # alpha choisi par validation croisée sur une grille
8 grille = np.logspace(-2, 4, 100)
9 modele = make_pipeline(StandardScaler(), RidgeCV(alphas=grille, cv=10))
10 modele.fit(X, y)
11 modele[-1].alpha_
```

alpha : le λ du cours. Standardisation **dans** le pipeline : leçon de la séance 3 valable ici aussi.

Limites et suite

44

Origine de la limite

Pénalité $\|\beta\|^2$ **dérivable en zéro** : gradient nul, rien ne pousse un coefficient à s'y arrêter.

Losange à la place de la boule ronde — norme ℓ_1 :

$$\text{lasso : } \|\mathbf{y} - \mathbf{X}\beta\|^2 + \lambda\|\beta\|_1.$$

Coins du losange sur les axes : la solution y tombe, zéros exacts.
Sujet naturel de la suite.

46

Ridge ne sélectionne pas

Chemin de régularisation : tous les coefficients contractés, aucun à zéro pour un λ fini.

Mille variables dont dix utiles : ridge renvoie mille coefficients petits, non dix coefficients et 990 zéros.

Ridge **régularise**. Ne **sélectionne** pas.

45

Le compromis, pour la quatrième fois

Séance	Modèle	Hyperparamètre	De simple à complexe
1	polynômes	degré	↗
3	k -PPV	k	↘
4	arbres	α	↘
5	ridge	λ	↘

Quatre méthodes, un seul schéma : un nombre règle la complexité, la validation croisée le choisit, le jeu de test sert une fois.

47

En bref

1. Variance des moindres carrés explosive en σ^2/d_i^2 : colinéarité, p proche de n , ou $p > n$.
2. Ridge : terme $\lambda \|\beta\|^2$ ajouté ; solution **toujours** existante, même quand celle des moindres carrés n'existe pas.
3. Base de la SVD : ridge, **filtre** qui écrase les directions instables.
4. **Toujours** un $\lambda > 0$ meilleur que les moindres carrés ; gain dépendant de p/n .
5. $\text{df}(\lambda)$: mesure de la complexité ; la validation croisée choisit λ ; ridge ne sélectionne aucune variable.