Projet : Implémentation du LASSO

Statistiques en grande dimension

Alexis Lucas Cyprien Malgoyre Maël Mauduit Clément Mazzocchi Mathis Metz
Paul Vandercoilden

École nationale des ponts et chaussées
Institut Polytechnique de Paris

26 mai 2025

Plan

Plan

1.1. *Coordinate descent* (descente par coordonnées)

Fonction objectif : $F_\lambda(\beta) = \|Y - \mathbb{X}\beta\|_2 + \lambda\|\beta\|_1$

$$\hat{\beta}_\lambda = \arg\min_\beta F_\lambda(\beta)$$

- Choisir un β intial
- Tant que la convergence n'est pas atteinte :
  - Choisir une coordonnée i
  - Choisir un pas η
  - Mettre à jour $\beta^i \leftarrow \beta^i - \eta \dfrac{\partial F_\lambda}{\partial \beta^i}(\beta)$

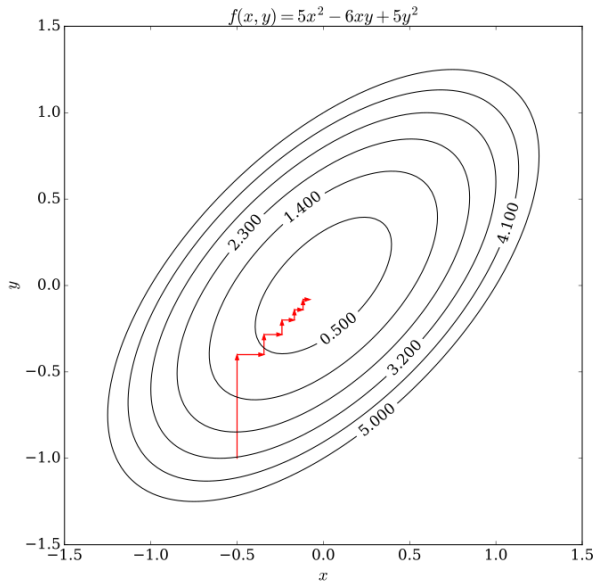


Figure: `https://commons.wikimedia.org/wiki/File:Coordinate_descent.svg`

Efficace en grande dimension et si les colonnes de $\mathbb{X}$ ne sont pas trop corrélées.

1.2. *Cross-validation* (validation croisée)

Optimisation des hyperparamètres (ici λ) :

- Diviser les données en k blocs
- Entraîner le modèle sur $k-1$ blocs et tester sur le bloc restant
- Répéter l'opération pour chaque combinaison
- Choisir le λ qui minimise l'erreur moyenne de prédiction

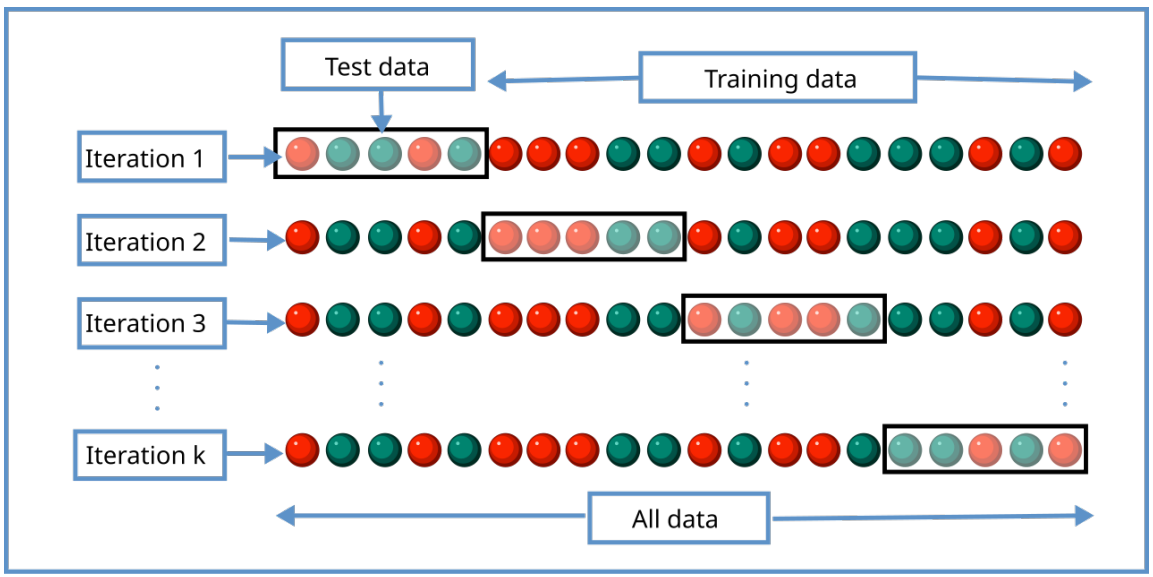


Figure: `https://commons.wikimedia.org/wiki/File:K-fold_cross_validation_EN.svg`

Plan

2.1. Création de l'échantillon

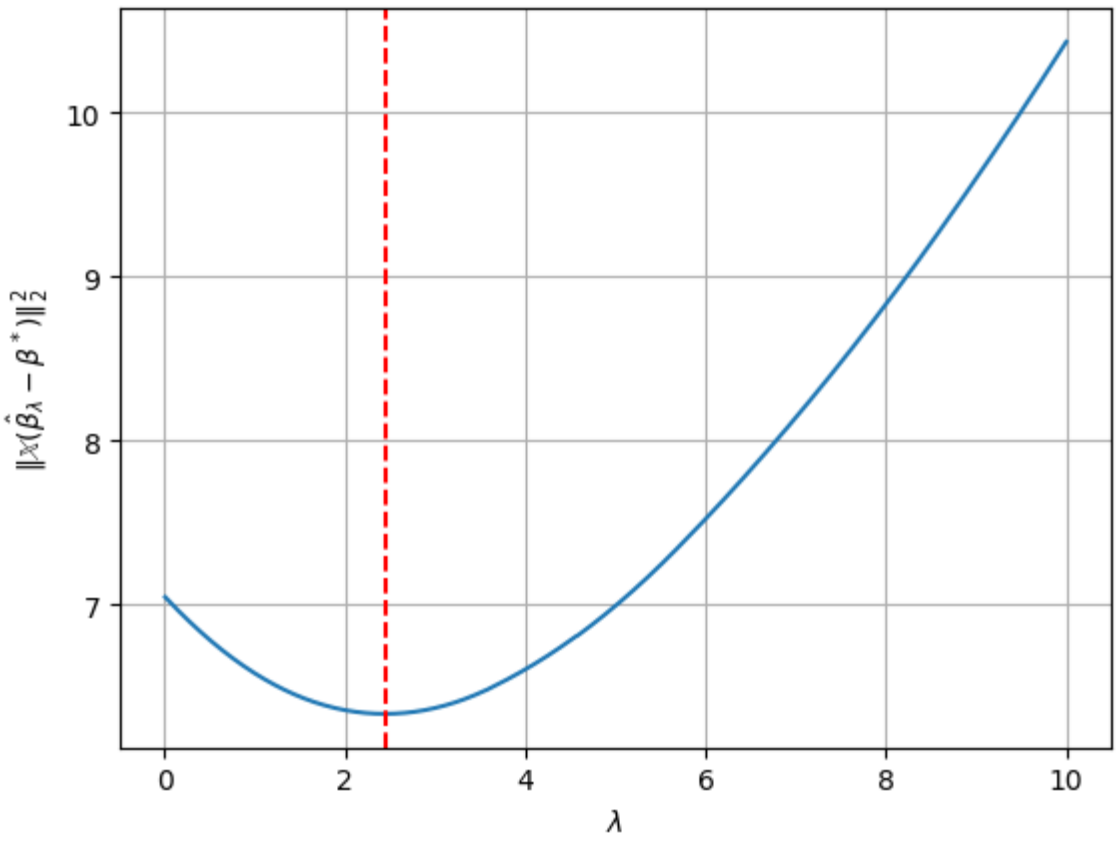
On se donne 20 points (X_i, Y_i) tels que $X_i \sim \mathcal{N}(0, I_8)$, $Y_i = X_i^T \beta^* + \varepsilon_i$, $\varepsilon_i \sim \mathcal{N}(0,1)$ et on suppose dans cet exemple connaître β^*

$$\beta^* = \begin{pmatrix} 5 \\ 3 \\ 0 \\ 0 \\ 1,5 \\ 0 \\ 0 \\ 0 \end{pmatrix}$$

2.2. Recherche de l'hyperparamètre

La fonction à minimiser est

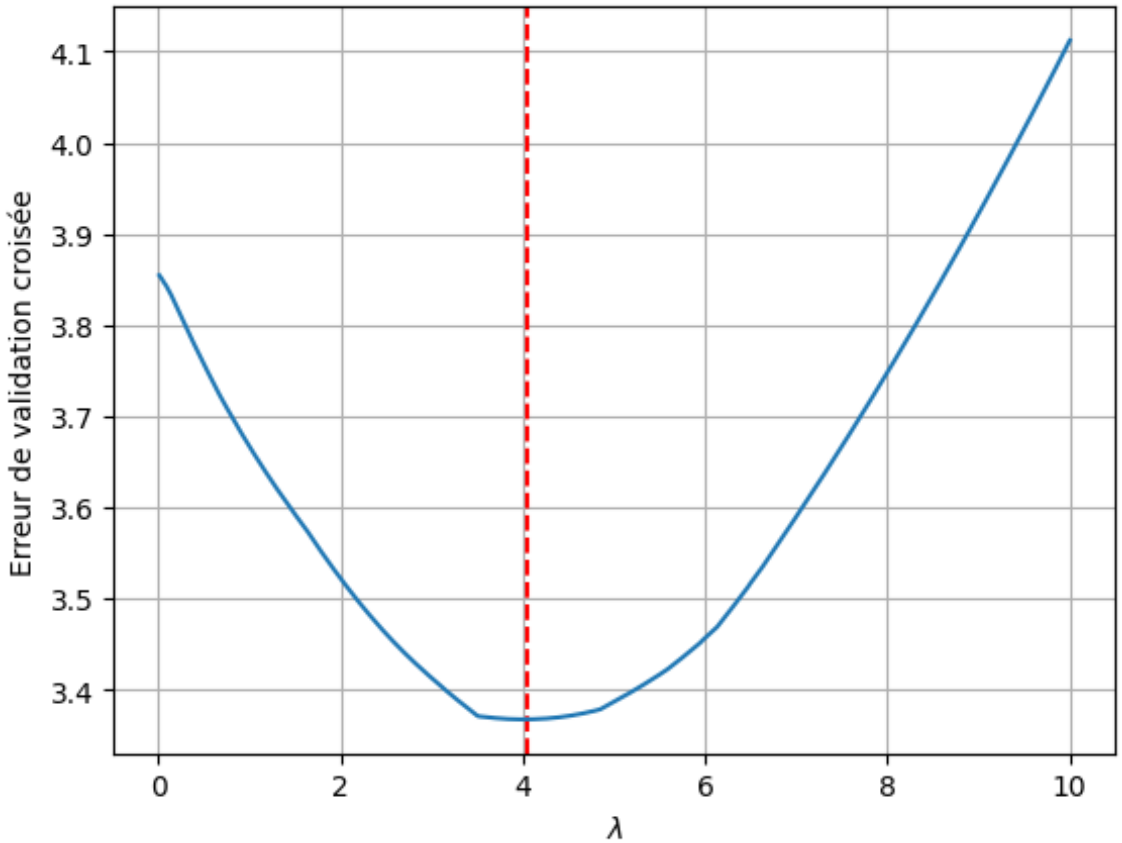
$$\lambda \mapsto \|\mathbb{X}(\hat{\beta}_\lambda - \beta^*)\|_2^2$$



Le lambda optimal obtenu est $\lambda^* = 2,45$ pour une erreur minimale de $6,33$.

2.3. Utilisation de la validation croisée

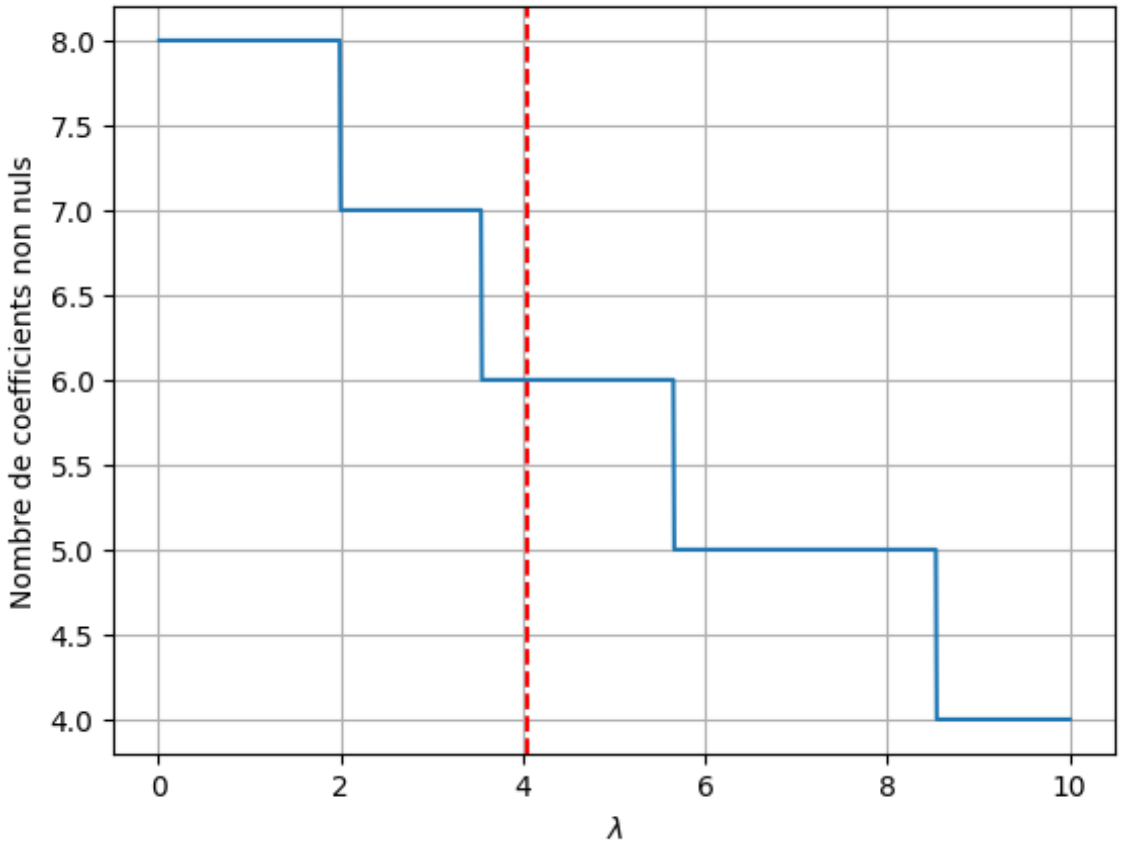
En pratique, β^* est inconnu donc on ne peut pas calculer l'erreur exacte du modèle.



Le lambda optimal obtenu par validation croisée est $\lambda^{\text{cv}} = 4,05$.
En utilisant λ^{cv} dans le LASSO, on obtient une erreur de $6,61$.

2.4. Nombre de paramètres actifs

On détermine alors $\lambda \to \|\hat{\beta}_\lambda\|_0$.



Avec λ^{cv}, on garde 6 composantes actives.

Plan

3.1. Jeu de données

Le *Breast Cancer Wisconsin (Diagnostic) Data Set* est composé de 569 observations de cancers du sein, caractérisées par 30 variables permettant de prédire si la tumeur est maligne ou bénigne.

id ID number	diagnosis The diagnosis of breast tissues (M = malignant, B = benign)	# radius_mean mean of distances from center to points on the perimeter	# texture_mean standard deviation of gray-scale values	# perimeter_mean mean size of the core tumor
8670 … 911m	B 63% M 37%	6.98 … 28.1	9.71 … 39.3	43.8 … 189
842302	M	17.99	10.38	122.8
842517	M	20.57	17.77	132.9
84300903	M	19.69	21.25	130
84348301	M	11.42	20.38	77.58
84358402	M	20.29	14.34	135.1
843786	M	12.45	15.7	82.57
844359	M	18.25	19.98	119.6
84458202	M	13.71	20.83	90.2
844981	M	13	21.82	87.5
84501001	M	12.46	24.04	83.97

Figure: https://doi.org/10.24432/C5DW2B

3.2. LASSO logistique

On souhaite ici faire de la sélection de variables : le LASSO linéaire n'est pas adapté, il faut utiliser la régression logistique adapté pour les problèmes de classification.

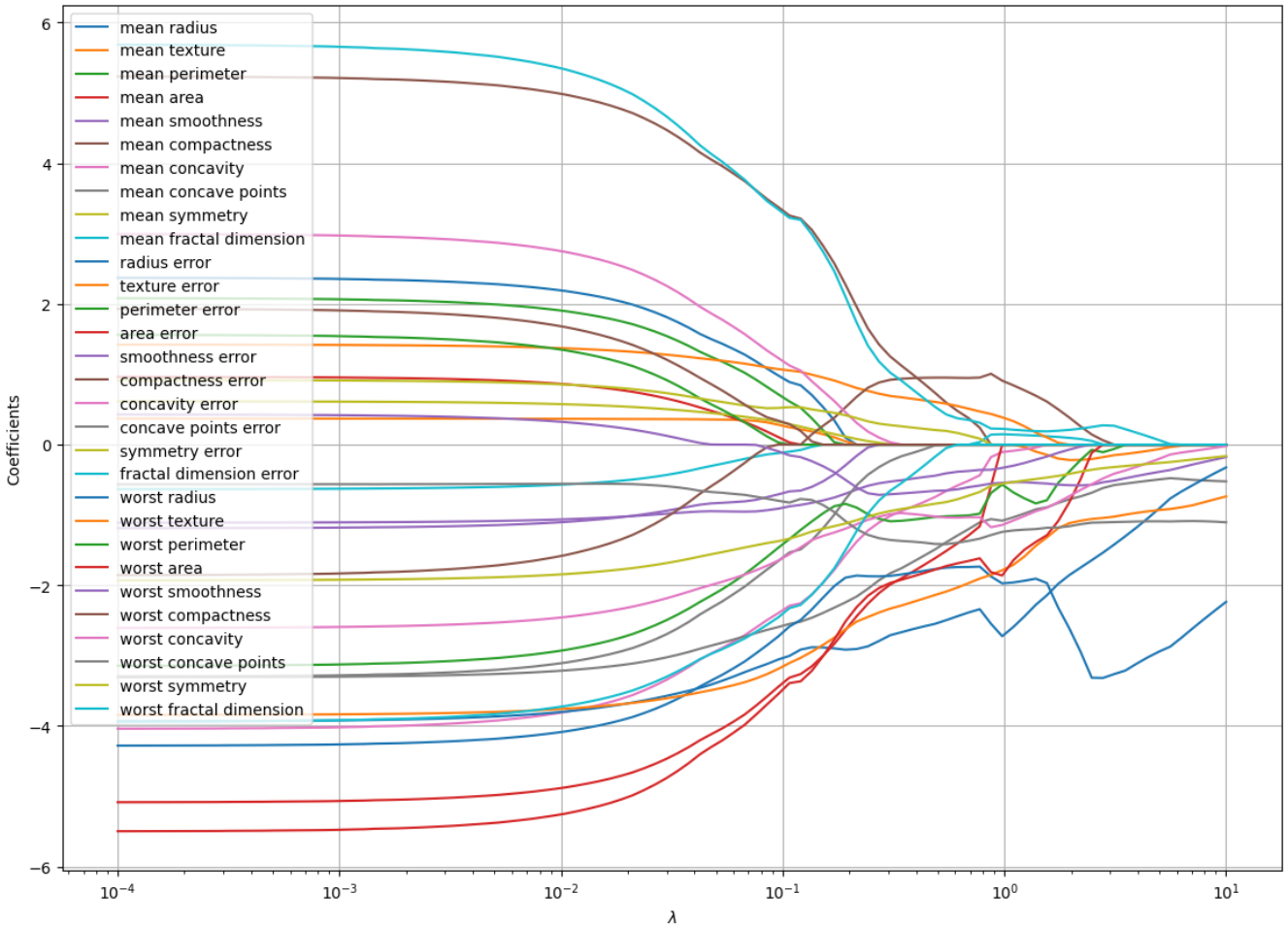
$$\mathbb{P}(Y_i = 1|X_i) = \frac{1}{1 + \exp\left(-X_i^T \beta\right)}$$

On obtient β en maximisant la log-vraisemblance $\mathcal{L}(\beta)$.
Le LASSO logistique permet de pénaliser β pour encourager la sparsité :

$$\hat{\beta}_\lambda = \arg\min_\beta \{-\mathcal{L}(\beta) + \lambda \|\beta\|_1\}$$

3.3. Chemin de régularisation

On trace la fonction $\lambda \mapsto \hat{\beta}_\lambda$ et on indique les variables les plus importantes :



- *worst radius*
- *worst texture*
- *worst concave points*
- *mean concave points*
- *radius error*
- *worst smoothness*
- *worst symmetry*
- *worst concavity*

3.4. Validation croisée

On utilise la validation croisée pour déterminer λ.
Le λ optimal trouvé vaut $0,36$. Il conduit aux variables actives suivantes :

- *worst radius*
- *radius error*
- *worst texture*
- *mean concave points*
- *worst concave points*
- *area error*
- *mean concavity*
- *worst area*
- *mean compactness*
- *fractal dimension error*
- *compactness error*
- *worst concavity*
- *worst symmetry*
- *texture error*
- *worth smoothness*
- *worst perimeter*
- *worst fractal dimension*
- *smoothness error*
- *symmetry error*
- *concave points error*

3.5. Conclusion

Le LASSO logistique indique quelles caractéristiques tumorales sont critiques pour déterminer la malignité de la tumeur.
Ici, seulement 20 variables sur les 30 du jeu de données sont véritablement informatives. Le diagnostic doit donc se concentrer sur ces 20 variables.

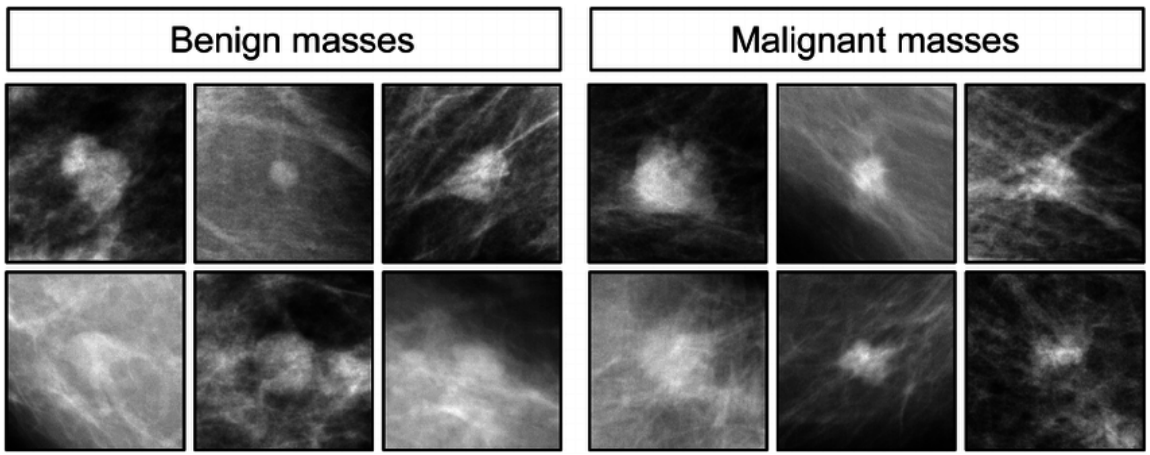


Figure: `https://www.researchgate.net/figure/Examples-of-benign-left-and-malignant-right-masses-in-mammograms-Subsequent-biopsy_fig1_284165798`

Merci pour votre attention.

Avez-vous des questions ?

ÉCOLE NATIONALE DES PONTS ET CHAUSSÉES
IP PARIS