Régression linéaire

Pierre Gaillard

12 février 2019

Dans ce cours, nous allons étudier le problème simple mais encore abondamment utilisé de nos jours : la régression[1] linéaire. Ce cours est basé sur les livres Rivoirard et Stoltz, 2012 et Cornillon et Matzner-Løber, 2011.

1 Introduction

Commençons par un exemple de problème pratique. Pour mieux optimiser sa production, un producteur s'intéresse à modéliser la consommation électrique en France en fonction de la température (cf. Figure 1).

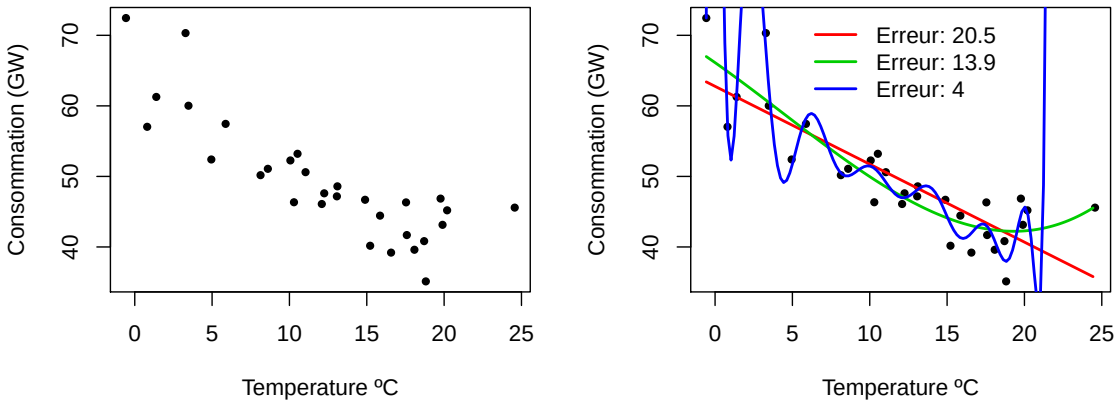

Figure 1 – Consommation électrique française (GW) en fonction de la température (°C). À droite, sont tracés fonctions minimisant l'erreur pour les espace de polynômes de degrés 1 (rouge), 3 (vert) et 30 (bleu)

L'objectif est de trouver une fonction f telle qui explique bien la consommation électrique $(y_i)_{1\leqslant i\leqslant n}$ en fonction de la température $(x_i)_{1\leqslant i\leqslant n}$, soit $y_i \approx f(x_i)$. Pour cela, on peut choisir un espace de fonction $\mathcal{F}$ et résoudre le problème de minimisation du risque empirique :

$$\widehat{f}_n \in \operatorname*{arg\,min}_{f\in\mathcal{F}} \widehat{\mathcal{R}}_n(f) := \operatorname*{arg\,min}_{f\in\mathcal{F}} \frac{1}{n}\sum_{i=1}^{n} \big(y_i - f(x_i)\big)^2 . \tag{1}$$

1. Le mot « régression » aurait été introduit par Galton au XIXe siècle. En modélisant la taille des individus en fonction de celle de leurs pères, Galton a observé un retour (« regression » en anglais) vers la taille moyenne. Les pères plus grands que la moyenne ont tendance a avoir des enfants plus petits qu'eux et vice-versa pour les pères plus petits.

Il faut faire attention au choix de l'espace de fonction pour éviter le sur-apprentissage (cf. Figures 1). Bien que l'erreur quadratique moyenne diminue quand l'espace $\mathcal{F}$ s'agrandit (degrés de polynômes plus grands), l'estimateur $\widehat{f}_n$ perd sont pouvoir prédictif. La question qu'il faut se poser est : $\widehat{f}_n$ sera-t-il performant sur de nouvelles données ?

L'espace des fonctions linéaires de la forme $f : x \mapsto ax + b$ est le plus simple. C'est celui que nous allons étudier dans ce cours.

Formalisation

En apprentissage statistique, on suppose que les observations $(x_i, y_i) \in \mathcal{X} \times \mathcal{Y}$ (ici $\mathcal{X} = \mathbb{R}^p$ et $\mathcal{Y} = \mathbb{R}$) sont les réalisations de variables aléatoires (X_i, Y_i). Le jeu de donnée aléatoire sera souvent noté $D_n := \{(X_i, Y_i), i = 1, \dots, n\}$. Les X_i a valeur dans $\mathcal{X}$ sont les variables d'entrée, les Y_i à valeur dans $\mathcal{T}$ sont les variables sortie. On fera l'hypothèse dans ce cours que les données sont i.i.d.

$$(X_i, Y_i) \overset{i.i.d.}{\sim} \nu\,.$$

La loi ν est inconnue du statisticien, il s'agit de l'apprendre à partir des données D_n. Une règle d'apprentissage $\mathcal{A}$ est une fonction qui associe à des données d'entrainement D_n une fonction de prédiction $\widehat{f}_n$ (le chapeau sur f indique qu'il s'agit d'un estimateur) :

$$\begin{array}{rcl} \mathcal{A} : \cup_{n\in\mathbb{N}} \big(\mathbb{R}^p \times \mathbb{R}\big)^n & \to & \mathbb{R}^{\mathbb{R}^p} \\ D_n & \mapsto & \widehat{f}_n \end{array}\,.$$

La fonction estimée $\widehat{f}_n$ est construite en vue de prédire Y' à partir d'un nouveau X', où $(X', Y') \sim \nu$ est une paire de *données test*, c'est à dire non observée dans les données d'entraînement. La fonction $\widehat{f}_n$ est un estimateur car elle dépend des données D_n et d'aucun paramètre non observé (comme ν). Comme D_n est aléatoire, il s'agit d'une fonction aléatoire.

L'objectif est de trouver construire un estimateur $\widehat{f}_n$ qui prévoit bien des nouvelles données en minimisant le risque :

$$\mathcal{R}(\widehat{f}_n) := \mathbb{E}\Big[\big(Y' - \widehat{f}_n(X')\big)^2 \,\Big|\, D_n\Big] \qquad \text{où} \qquad (X', Y') \sim \nu\,. \tag{Risque}$$

Cependant le statisticien ne peut pas calculer le risque car il ne connaît pas ν. Une méthode courante en apprentissage statistique et intuitive consiste donc à remplacer le risque par le risque empirique

$$\widehat{\mathcal{R}}_n(f) = \frac{1}{n}\sum_{i=1}^{n} \big(Y_i - f(X_i)\big)^2 = \frac{1}{n}\big\|Y - f(X)\big\|^2 \tag{Risque empirique}$$

où $Y = (Y_1, \dots, Y_n) \in \mathbb{R}^n$ et $X = (X_1, \dots, X_n)^\top \in \mathbb{R}^{n\times p}$.

Il faut cependant faire attention au sur-apprentissage (cas où $\widehat{\mathcal{R}}_n(f)$ est beaucoup plus faible que $\mathcal{R}(f)$, cf. Figure 2). Dans ce cours, on va étudier la performance du minimiseur des moindre carré dans le cas du modèle linéaire.

2 Le modèle linéaire

Dans cette partie, on pose un cadre stochastique qui va nous permettre d'analyser la performance de minimiseur du risque empirique dans le cas linéaire. Dans un modèle linéaire on suppose une équation de la forme :

$$Y_i = \beta_0 + \beta_1 X_{i,1} + \cdots + \beta_p X_{i,p} + \varepsilon_i\,, \tag{2}$$

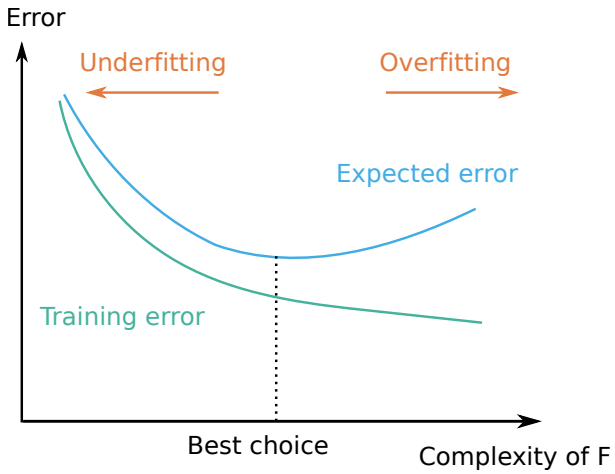

FIGURE 2 – Sur-apprentissage et sous-apprentissage en fonction de la complexité de $\mathcal{F}$. En bleu le vrai risque $\mathcal{R}(f)$, en vert le risque empirique $\widehat{\mathcal{R}}_n(f)$.

où $\varepsilon = (\varepsilon_1, \dots, \varepsilon_n)^\top \in \mathbb{R}^n$ est un vecteur d'erreurs (ou de bruit). Il vient du fait qu'en pratique les observations y_i ne collent jamais complètement à la prévision linéaire. On suppose qu'il s'agit d'un n-échantillon (i.e., les ε_i sont i.i.d.) d'espérance nulle $\mathbb{E}[\varepsilon_i] = 0$. Souvent, il est supposé de loi Gaussienne $\varepsilon_i \sim \mathcal{N}(0, \sigma^2)$. Pour simplifier les notations, nous supposons que le premier vecteur des variables explicatives est le vecteur constant $X_{\cdot,1} = (1, \dots, 1)^\top$ de sorte que le coefficient β_0 de l'équation (2) n'est plus nécessaire. Nous définissons la matrice des variable explicatives

$$X := \begin{bmatrix} 1 & X_{1,2} & \dots & X_{1,p} \\ \vdots & \vdots & \dots & \vdots \\ 1 & X_{n,2} & \dots & X_{n,p} \end{bmatrix}$$

composée des lignes $(1, X_i) \in \mathbb{R}^p$. Nous pouvons réécrire la définition du modèle linéaire (2) de façon matricielle.

Definition 2.1 (Modèle linéaire)**.** *Un modèle linéaire se définit par une équation de la forme :*

$$Y = X\beta + \varepsilon$$

où $Y \in \mathbb{R}^n$ est le vecteur des observations, $X \in \mathbb{R}^{n \times p}$ est la matrice de variables explicatives et $\beta \in \mathbb{R}^p$ le vecteur de coefficients à estimer. On suppose que la matrice X est injective (i.e., $rg(X) = p$) et que le vecteur de bruit $\varepsilon \in \mathbb{R}^n$ est un n-échantillon centré ($\mathbb{E}[\varepsilon] = 0$ et $\mathrm{Var}[\varepsilon] = \sigma^2 I_n$).

Design déterministe vs design aléatoire Dans toute la suite, bien que X puisse être aléatoire, il s'agit en pratique de variables explicatives observées (ou calculées) par le prévisionniste avant d'effectuer une prévision. On supposera donc qu'il s'agit de variables déterministes. Le cas aléatoire peut se retrouver en considérant des espérances conditionnelles.

3 L'estimateur des moindres carrés ordinaire (OLS)

Le jeu de données $(X_i, Y_i) \in \mathbb{R}^p \times \mathbb{R}$ étant observé, l'objectif est de trouver la fonction affine $\widehat{f} : x \mapsto x^\top \widehat{\beta}$ minimisant l'erreur quadratique moyenne (1). On suppose que l'intercepte est inclus

dans les variables explicatives de sorte que $X_{i,1} = 1$. Dans le cas linéaire, le risque empirique s'écrit

$$\mathcal{R}_n(\beta) := \frac{1}{n}\sum_{i=1}^{n}\Big(Y_i - \beta_1 + \sum_{j=2}^{p}\beta_j X_{i,j}\Big)^2 = \frac{1}{n}\big\|Y - X\beta\big\|^2,$$

où $\|\cdot\|$ est la norme euclidienne.

Definition 3.1 (Estimateur des moindres carrés ordinaires (OLS))**.** *On appelle estimateur des moindres carrés ordinaire, le minimiseur $\widehat{\beta} \in \mathbb{R}^p$ du risque empirique :*

$$\widehat{\beta} \in \operatorname*{arg\,min}_{\beta\in\mathbb{R}^p} \widehat{\mathcal{R}}_n(\beta) = \operatorname*{arg\,min}_{\beta\in\mathbb{R}^p} \big\|Y - X\beta\big\|. \tag{3}$$

Proposition 3.1. *Sous les hypothèses et les notations de la Définition 2.1, l'OLS existe et est unique. Il s'écrit $\widehat{\beta} = \big(X^\top X\big)^{-1} X^\top Y$.*

Démonstration. Par définition, l'OLS est le minimiseur du risque empirique $\widehat{\mathcal{R}}_n : \mathbb{R}^p \to \mathbb{R}_+$ définie par : pour tout $\beta \in \mathbb{R}^p$

$$\widehat{\mathcal{R}}_n(\beta) = \frac{1}{n}\big\|Y - X\beta\big\|_2^2 = \frac{1}{n}\big(\beta^\top\big(X^\top X\big)\beta - 2\beta^\top X^\top Y + \|Y\|^2\big).$$

Pour être un minimum, il est nécessaire que le gradient s'annule. Celui-ci s'écrit :

$$\nabla\widehat{\mathcal{R}}_n(\widehat{\beta}) = \frac{1}{n}\big(\widehat{\beta}^\top(X^\top X) + (X^\top X)\widehat{\beta} - 2X^\top Y\big) = \frac{2}{n}\big((X^\top X)\widehat{\beta} - Y^\top X\big).$$

où la dernière inégalité est parce que la matrice $X^\top X \in \mathbb{R}^{p\times p}$ est symétrique. Comme X est de rang p par hypothèse, $X^\top X$ est même symétrique définie positive et inversible (laissé en exercice). On en déduit donc qu'un optimum de $\widehat{\mathcal{R}}_n$ vérifie nécessairement :

$$\widehat{\beta} = \big(X^\top X\big)^{-1} XY.$$

Il reste cependant à vérifier qu'il s'agit bien d'un minimum et donc que la Hessienne est définie positive. Ce qui est le cas car : $\nabla^2\widehat{\mathcal{R}}_n(\widehat{\beta}) = \frac{2}{n}(X^\top X)$. □

Une interprétation géométrique Le modèle linéaire cherche à modéliser le vecteur $Y \in \mathbb{R}^n$ des observations par une combinaison linéaire de la forme $X\beta \in \mathbb{R}^n$. L'image de X est l'espace des solutions, noté $\mathrm{Im}(X) = \{Z \in \mathbb{R}^n : \exists\beta \in \mathbb{R}^p \text{ tq } Z = X\beta\} \subseteq \mathbb{R}^n$. C'est le sous-espace vectoriel de $\mathbb{R}^n$ engendré par les $p < n$ colonnes de la matrice de variables explicatives. Comme $\mathrm{rg}(X) = p$, il est de dimension p. Selon la définition du modèle linéaire, Y est la somme d'un élément de $\mathrm{Im}(X)$ et d'un bruit. En minimisant $\|Y - X\beta\|$ (cf. Définition 3.1), on cherche donc l'élément de $\mathrm{Im}(X)$ le plus proche de Y. Il s'agit du projeté orthogonal de Y sur $\mathrm{Im}(X)$, noté $\widehat{Y}$. Par définition de l'OLS et par la Proposition 3.1, on a :

$$\widehat{Y} \overset{\text{Déf 3.1}}{=} X\widehat{\beta} \overset{\text{Prop. 3.1}}{=} X(X^\top X)^{-1}X^\top Y.$$

En particulier, $P_X := X(X^\top X)^{-1}X^\top$ est la matrice de projection orthogonal sur $\mathrm{Im}(X)$.

On peut montrer que l'OLS a de bonnes propriétés statistiques.

Proposition 3.2. *Sous les hypothèses de la Définition 2.1, l'estimateur $\widehat{\beta}$ est sans biais, $\mathbb{E}\big[\widehat{\beta}\big] = \beta$ et de variance* $\mathrm{Var}(\widehat{\beta}) = \sigma^2(X^\top X)^{-1}$.

Démonstration. On peut calculer son espérance en utilisant $\mathbb{E}[\varepsilon] = 0$:

$$\mathbb{E}[\widehat{\beta}] = \mathbb{E}\big[(X^\top X)^{-1} X^\top Y\big] = \mathbb{E}\big[(X^\top X)^{-1} X^\top X\beta + \cancel{(X^\top X)^{-1} X^\top \varepsilon}\big] = \beta\,.$$

et sa variance en utilisant $\mathrm{Var}(Y) = \mathrm{Var}(\varepsilon) = \sigma^2 I_n$.

$$\mathrm{Var}(\widehat{\beta}) = \mathrm{Var}\big((X^\top X)^{-1} X^\top Y\big) = (X^\top X)^{-1} X^\top \mathrm{Var}(Y) X (X^\top X)^{-1} = \sigma^2 (X^\top X)^{-1}\,.$$

□

On peut même montrer que l'OLS vérifie la propriété de Gauss-Markov. Il est optimal parmi les estimateurs sans biais de β, dans le sens où il a une matrice de variance-covariance minimale (pour une certaine relation d'ordre partiel). Cependant, la Proposition 3.2 est peu pratique pour estimer la fiabilité de $\widehat{\beta}$ en pratique car elle fait intervenir la variance du bruit, qui est inconnue. Il peut également être intéressant de l'estimer aussi. Un estimateur naturel et de regarder l'erreur moyenne

$$\widehat{\sigma}^2_{\mathrm{nat}} = \frac{1}{n}\sum_{i=1}^{n} (y_i - X_{i,\cdot}\widehat{\beta})^2 = \frac{\|Y - X\widehat{\beta}\|^2}{n}\,.$$

Cependant cet estimateur est biaisé. On utilise donc l'estimateur non biaisé (exercice) :

$$\widehat{\sigma}^2 = \frac{\|Y - X\widehat{\beta}\|^2}{n-p}\,.$$

On peut de même estimer le risque associé à l'OLS.

Proposition 3.3. *Soit $(X', Y') \in \mathbb{R}^p \times \mathbb{R}$ satisfaisant le modèle linéaire $Y' = \beta X' + \varepsilon'$, avec ε' bruit centré indépendant de $\varepsilon \in \mathbb{R}^n$ et de variance $\mathrm{Var}(\varepsilon') = \sigma^2$. Alors*

$$\mathbb{E}\big[\mathcal{R}(\widehat{\beta})\big] = \sigma^2(1 + X'^\top \big(X^\top X\big)^{-1} X').$$

Démonstration.

$$\begin{aligned}
\mathbb{E}\big[\mathcal{R}(\widehat{\beta})\big] &:= \mathbb{E}\big[(Y' - X'^\top\widehat{\beta})^2\big] \\
&= \mathbb{E}\big[(X'^\top\widehat{\beta} - \mathbb{E}[X'^\top\widehat{\beta}])^2\big] + \cancel{\mathbb{E}\big[(\mathbb{E}[X'^\top\widehat{\beta}] - \mathbb{E}[Y'])^2\big]} + \mathbb{E}\big[(Y' - \mathbb{E}[Y'])^2\big] \\
&= X'^\top \mathrm{Var}(\widehat{\beta}) X' + \sigma^2 \\
&= \sigma^2(1 + X'^\top \big(X^\top X\big)^{-1} X').
\end{aligned}$$

□

Remarque : X et X' sont supposés déterministes et connus ici. On parle de "deterministic design". On peut facilement étendre les résultats au cas de design aléatoire, dans quel cas X et X' sont aussi aléatoires. Il faut alors expliciter le fait que les espérances $\mathbb{E}[Y]$ et $\mathbb{E}[X^\top\beta]$ dans les calculs ci-dessus sont conditionnelles à X et X'.

Le cas Gaussien Un cas particulier très considéré est celui du bruit Gaussien : $\varepsilon \sim \mathcal{N}(0, \sigma^2 I_n)$. Ce choix vient non seulement du fait qu'il permet de calculer de nombreuses propriétés statistiques supplémentaires sur $\widehat{\beta}$ et de faire des tests (intervalles de confiance, significativité des variables,...). Il est en pratique motivé par le théorème central limite et le fait que le bruit est souvent une addition de nombreux phénomènes non expliqués par la combinaison linéaire des variables explicatives.

Proposition 3.4. *Dans le cas Gaussien, les estimateurs du maximum de vraisemblance de β et σ vérifient respectivement :*

$$\widehat{\beta}_{MV} = \left(X^\top X\right)^{-1} XY \qquad et \qquad \widehat{\sigma}^2_{MV} = \frac{\|Y - X\widehat{\beta}\|^2}{n}\,.$$

On retrouve donc l'estimateur des moindres carrés obtenu par la minimisation du risque empirique. L'estimateur de la variance est biaisé.

Problème non-linéaire : régression polynomiale, spline, à noyau L'hypothèse que les observations Y_i s'explique comme une combinaison des variables explicatives $X_{i,j}$ peut paraître forte. Cependant, on peut appliquer la théorie précédente sur des transformations des variables $x_{i,j}$. Par exemple, en ajoutant les puissance des variables $X^k_{i,j}$ ou leurs produits $X_{i,j}X_{i,j'}$. Cela permet de se comparer aux espaces polynomiaux. Faire une régression linéaire sur des transformation polynomiales des variables revient à faire une régression polynomiale.

Bien sûr d'autres bases (transformations) existent comme des régression sur des bases de splines (polynômes par morceaux avec contraintes sur les bords). C'est le modèle qu'utilise par exemple EDF en opérationnel pour prévoir la consommation électrique en fonction de variables comme l'heure, le jour de la semaine, la température ou la couverture nuageuse. Une autre forme de régression que nous aborderons dans la suite sera la régression à noyau. De façon générale, on peut considérer des transformations $\phi : \mathcal{X} \to \mathbb{R}^d$ avec le modèle

$$Y_i = \phi(X_i)^\top \beta + \varepsilon_i\,,$$

pour un paramètre $\beta \in \mathbb{R}^d$ inconnu.

Dans la Figure 1, on a de cette façon minimisé le risque empirique sur les espaces polynomiaux de degré 1 (modèle linéaire), 3 et 30. On voit qu'il faut faire attention à ne pas considérer des espaces trop grands, au risque que le modèle soit mal posé (X non injective). Inversement, pour que la Proposition 3.3 soit vérifiée, il faut se trouver dans le vrai modèle $Y = \phi(X)\beta + \varepsilon$. Il faut donc s'assurer que $\phi(X)$ contienne suffisamment de descripteurs pour que la dépendance entre Y et $\phi(X)$ soit effectivement linéaire. Autrement on paye un terme de biais supplémentaire.

4 Régularisation

Si X n'est pas injective (i.e., $\mathrm{rg}(X) = p$), la matrice $(X^\top X)$ n'est plus inversible et le problème admet plusieurs solutions qui minimisent le risque empirique. On dit que le problème est mal posé ou non identifiable.

La Proposition 3.3 nous rappelle que le risque augmente proportionnellement à la variance de $\widehat{\beta}$ qui elle-même dépend du conditionnement de la matrice $(X^\top X)^{-1}$. Plus les colonnes de cette dernière risquent d'être dépendantes et moins $\widehat{\beta}$ sera précis. Plusieurs solutions permettent de traiter le cas où $\mathrm{rg}(X) < p$:

- *contrôle explicite de la complexité* en réduisant l'espace des solutions $\mathrm{Im}(X)$. Cela peut se faire en retirant des colonnes de la matrice X jusqu'à ce qu'elle deviennent injective (par exemple, en réduisant le degré des polynômes). On peut aussi poser des contraintes d'identifiabilité de la forme $\beta \in V$ un sous-espace vectoriel de $\mathbb{R}^p$ tel que tout élément $y \in \mathrm{Im}(X)$ a un unique antécédent $\beta \in V$ avec $y = X\beta$. On peut par exemple choisir $V = \mathrm{Ker}(X)^\perp$.
- *contrôle implicite de la complexité* en régularisant le problème de minimisation du risque empirique comme on le voit rapidement ci-après avec les régression Ridge et Lasso.

Régression Ridge La régularisation la plus courante est la *régression Ridge*, où on utilise la norme 2 :

$$\widehat{\beta}_{(\text{ridge})} \in \underset{\beta\in\mathbb{R}^p}{\arg\min} \left\{ \frac{1}{n}\|Y - X\beta\|_2^2 + \lambda\|\beta\|_2^2 \right\} .$$

Le paramètre de régularisation $\lambda > 0$ règle le compromis entre la variance de $\widehat{\beta}$ et son biais. En refaisant les calculs de la Proposition 3.1, on voit que la solution est unique (même si X n'est pas injective) donnée par :

$$\widehat{\beta}_{(\text{ridge})} = \left(X^\top X + n\lambda I_n\right)^{-1} X^\top Y .$$

On voit qu'il n'y a plus le problème d'inversion de $X^\top X$ puisque la régression Ridge revient à remplacer $\left(X^\top X\right)^{-1}$ par $\left(X^\top X + n\lambda I_n\right)^{-1}$ dans la solution de l'OLS. Le calibration du paramètre λ est cependant essentielle en pratique. Elle peut par exemple se faire de façon analytique, par *validation croisée (généralisée)* ou par *l'heuristique de pente* (cf. cours sur la sélection de modèle).

Proposition 4.1 (Biais et variance de $\widehat{\beta}_{(\text{ridge})}$)**.** *L'estimateur Ridge vérifie :* $\widehat{\beta}_{(ridge)} = (X^\top X + \lambda I_p)^{-1} X^\top X\beta$, $\mathbb{E}\big[\widehat{\beta}_{(ridge)}\big] = \beta - \lambda(X^\top X + \lambda I_p)^{-1}\beta$ *et* $\mathrm{Var}\big(\widehat{\beta}_{(ridge)}\big) = \sigma^2 (X^\top X + \lambda I_p)^{-1} X^\top X (X^\top X + \lambda I_p)^{-1}$.

L'estimateur Ridge est donc biaisé au contraire de l'OLS ce qui constitue un handicape. Cependant, sa variance ne fait pas intervenir l'inverse de $X^\top X$ mais de $X^\top X + \lambda I_p$ qui est mieux conditionnée. Il aura donc une plus faible variance. Exercice : calculer le risque de l'estimateur Ridge.

Le Lasso Une deuxième régularisation courante est la régularisation par la norme 1. On parle alors de l'estimateur Lasso.

$$\widehat{\beta}_{(\text{lasso})} \in \underset{\beta\in\mathbb{R}^p}{\arg\min} \left\{ \frac{1}{n}\|Y - X\beta\|_2^2 + \lambda\|\beta\|_1 \right\} .$$

L'estimateur Lasso n'a pas de formule analytique dans le cadre général. Cependant, dans le cas où $X^\top X = I_p$, le Lasso a une solution explicite. L'estimateur Lasso correspond alors à un seuillage doux de $\widehat{\beta}$ la solution des moindres carrées (OLS) : pour tout $j = 1, \dots, p$

$$\widehat{\beta}_{\text{lasso}} = \text{sgn}(\widehat{\beta}) \max\left(0, |\widehat{\beta}| - \lambda\right) .$$

L'estimateur Lasso est en particulier très utilisé quand le nombre de variables explicatives est très grand voir supérieur au nombre d'observations : $p \gg n$. Le caractère "anguleux" de la norme 1 va avoir tendance à tronquer (annuler) de nombreux coefficients de $\widehat{\beta}_{\text{lasso}}$ pour proposer des solutions parcimonieuses. Ceci se voit mieux avec la formulation équivalente du Lasso sous la forme d'un problème de régularisation sous contrainte : pour chaque $\lambda > 0$, il existe $U = U(X) > 0$ tel que :

$$\widehat{\beta}_{(\text{lasso})} \in \underset{\beta\in\mathbb{R}^p:\ \|\beta\|_1 \leqslant U}{\arg\min} \|Y - X\beta\|^2 .$$

Ridge se reformule de la même façon en remplaçant la norme 1 par la norme 2 au carré.

5 Calcul de $\widehat{\beta}$

La formule close $\widehat{\beta} = \left(X^\top X\right)^{-1} X^\top Y$ de l'OLS est utile pour l'analyser. Cependant, la calculer de façon naïve peut s'avérer prohibitif. En particulier quand p est grand, on préfère éviter l'inversion de la matrice $X^\top X$ qui coûte $\mathcal{O}(p^3)$ par la méthode de Gauss-Jordan et qui peut être très instable quand la matrice est mal conditionnée.

Décomposition QR Cependant, pour améliorer la stabilité, on peut utiliser la décomposition QR. Rappelons que $\widehat{\beta}$ est solution de l'équation :

$$X^\top X\widehat{\beta} = X^\top Y\,.$$

On écrit $X \in \mathbb{R}^{n\times p}$ de la forme $X = QR$, où $Q \in \mathbb{R}^{n\times p}$ est une matrice orthogonale (i.e., $QQ^\top = I_n$) et $R \in \mathbb{R}^{p\times p}$ triangulaire supérieure. Les matrices triangulaires supérieures sont très utiles pour résoudre les systèmes linéaires. En substituant dans l'équation précédente, on obtient :

$$\begin{aligned} R^\top(Q^\top Q)R\widehat{\beta} = R^\top Q^\top Y \quad &\Leftrightarrow \quad R^\top R\widehat{\beta} = R^\top Q^\top Y \\ &\Leftarrow \quad R\widehat{\beta} = Q^\top Y\,. \end{aligned}$$

Il ne reste alors plus qu'à résoudre un système linéaire avec une matrice triangulaire supérieure, ce qui est facile.

Descente de gradient. Pour éviter totalement le coût de l'inversion de matrice, une autre solution est la descente de gradient. Qui consiste à résoudre le problème de minimisation pas à pas en s'approchant du minimum grâce à des pas de gradient. On initialise par exemple $\widehat{\beta}_0 = 0$, puis on met à jour :

$$\begin{aligned} \widehat{\beta}_{i+1} &= \widehat{\beta}_i - \eta\nabla\widehat{\mathcal{R}}_n(\widehat{\beta}_i) \\ &= \widehat{\beta}_i - \frac{2\eta}{n}\big((X^\top X)\widehat{\beta}_i - Y^\top X\big)\,, \end{aligned}$$

où $\eta > 0$ est un paramètre d'apprentissage. On voit que si l'algorithme converge, alors il converge vers un point annulant le gradient, donc vers la solution de l'OLS. Pour avoir convergence, il faut que le paramètre η soit bien calibré, mais on verra cela plus en détails dans le cours sur la descente de gradient.

Si le jeu de donné est beaucoup trop grand $n \gg 1$. Il peut également être prohibitif de charger toutes les données pour faire le calcul de $\nabla\widehat{\mathcal{R}}_n(\widehat{\theta}_i)$. La solution courante est alors de faire une descente de gradient stochastique, où on ne fait des pas de gradients que sur des estimés de $\nabla\widehat{\mathcal{R}}_n(\widehat{\theta}_i)$, calculés sur une sous-partie aléatoire des données.

Références

Cornillon, P.-A. et E. Matzner-Løber (2011). "La régression linéaire simple". In : *Régression avec R*, p. 1-28.

Rivoirard, V. et G. Stoltz (2012). *Statistique mathématique en action.*