Université Paris Dauphine
Année 2014-2015
M1 MIDO Apprentissage statistique et grande dimension

Enoncés des projets (par binôme ou individuels)

- Le projet s'effectue seul ou en binôme. **On ne peut choisir qu'un seul sujet !** [1]
- Les rapports doivent être rendus à la scolarité du M1 au plus tard le **Vendredi 12 juin 2015 impérativement.** [2]
- Chaque projet constitue un ensemble de questions de difficultés variées. Certains projets insistent sur des aspects théoriques, d'autres sur des aspects plus appliqués. Il n'est pas nécessaire de répondre à toutes les questions pour obtenir une bonne note !
- Le format est libre : on rendra un document manuscrit ou tapuscrit (au choix) sans limite inférieure ni supérieure du nombre de pages contenant 3 parties **clairement identifiées** :
  a) Une première partie contenant une introduction où l'on présente la problématique abordée dans le sujet et le fruit de recherches éventuelles (via internet ou une bibliothèque par exemple) pour replacer la problématique dans un cadre d'applications.
  b) Une seconde partie plus standard où l'on rédige les réponses aux différentes questions (sans obligation de répondre à toutes les questions).
  c) Une troisième partie, qui comporte la mise en oeuvre numérique du projet, dans un format libre. On utilisera le logiciel de son choix, et on illustrera numériquement (en particulier à l'aide de graphiques) le ou les phénomènes étudiés dans le projet. On incluera les codes numériques développés.
  d) Il est tout à fait accepté (et même encouragé) de s'éloigner du texte initial ou de n'en traiter qu'une partie si l'on souhaite explorer différents développements possibles inspirés du texte.

1. Seul un sujet sera noté par étudiant ou par binôme.
2. Aucun projet ne sera accepté au delà de cette date !

Université Paris Dauphine Année 2014-2015 Apprentissage statistique et grande dimension

---

Projet 1 : Estimation d'une fonction de régression

Le probléme de prédiction ou d'explication d'une variable Y à l'aide d'une autre variable X est souvent rencontré en pratique. La fonction qui fournit la meilleure prévision (en moyenne quadratique) de Y en fonction de X est l'espérance conditionnelle

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

Cette fonction est appelée fonction de régression et son estimation à partir de n copies indépendantes du couple (X, Y) est un problème fondamental en statistique.

Considérons le cas où $X \in \mathbb{R}^d$ et $Y \in \mathbb{R}$. Si l'on ne connaît pas de forme paramétrique spécifique pour la fonction f (par exemple, fonction linéaire ou polynôme trigonométrique de degré 2), alors les méthodes d'estimation classiques (moindres carrés, maximum de vraisemblance, etc) ne peuvent pas être utilisées directement. On parle alors de probléme d'estimation non-paramétrique. L'objet de ce travail personnel est d'étudier une méthode d'estimation non-paramétrique et de l'illustrer sur des jeux de données simulées.

1 Estimateur par projection

Supposons que la variable explicative X suit la loi uniforme sur $[0, 1]^d$ et que $\{(X_i, Y_i),\ 1 \leq i \leq n\}$ sont n copies indépendantes de (X, Y). De plus, on suppose que la fonction de régression f appartient à $L^2([0, 1]^d)$. Alors, pour toute base orthonormée $\varphi_1, \varphi_2, \ldots$ de $L^2([0, 1]^d)$, on a

$$f = \sum_{j=1}^{\infty} \vartheta_j \varphi_j,$$

où la convergence a lieu dans L^2, avec des coefficients $\vartheta_j = \langle f, \varphi_j \rangle = \int_{[0,1]^d} f\varphi_j$ vérifiant $\sum_{j=1}^{\infty} \vartheta_j^2 < \infty$. Cela implique que $\vartheta_j \to 0$ lorsque $j \to \infty$.

L'idée de l'estimateur par projection consiste donc à remplacer f par une approximation

$$f_{N,\vartheta}(x) = \sum_{j=1}^{N} \vartheta_j \varphi_j(x), \qquad \forall x \in \mathbb{R}^d,$$

et d'estimer le paramètre fini-dimensionnel $\vartheta = (\vartheta_1, \ldots, \vartheta_N)'$ par la méthode classique des moindres carrés. Le choix du niveau de troncature est un point important et il sera fait en fonction des données. Soit $\mathbf{\Phi}_N$ la matrice $n \times N$ dont la $j^{\text{éme}}$ colonne est $\Phi_{\bullet j} = (\varphi_j(X_1), \ldots, \varphi_j(X_n))'$ pour $j = 1, \ldots, N$. On suppose par la suite que $\mathbf{\Phi}_N' \mathbf{\Phi}_N$ est une matrice définie strictement positive.

a) Calculer l'estimateur des moindres carrés $\hat{\vartheta}_{n,N}$ du paramètre ϑ dans le modèle approché $Y_i = f_{N,\vartheta}(X_i) + U_i$ et en déduire un estimateur $\hat{f}_{n,N}(x)$ de $f(x)$.

b) Soit $\boldsymbol{Y} = (Y_1, \ldots, Y_n)'$. Prouver que $(\hat{f}_{n,N}(X_1), \ldots, \hat{f}_{n,N}(X_n))' = \boldsymbol{A}_N \boldsymbol{Y}$ où $\boldsymbol{A}_N = \mathbf{\Phi}_N (\mathbf{\Phi}_N' \mathbf{\Phi}_N)^{-1} \mathbf{\Phi}_N'$ est un projecteur orthogonal sur le sous-espace vectoriel de $\mathbb{R}^n$ engendré par les colonnes de la matrice $\mathbf{\Phi}_N$.

c) Montrer que lorsque $n \to \infty$, la matrice $\frac{1}{n} \mathbf{\Phi}_N' \mathbf{\Phi}_N$ converge vers la matrice identité. Vérifier qu'en remplaçant $\mathbf{\Phi}_N' \mathbf{\Phi}_N$ par l'approximation $n I_{N \times N}$ dans la définition de $\hat{f}_{n,N}(x)$, on obtient l'estimateur

$$\tilde{f}_{n,N}(x) = \sum_{j=1}^{N} \tilde{\vartheta}_j \varphi_j(x), \qquad \tilde{\vartheta}_j = \frac{1}{n} \sum_{i=1}^{n} Y_i \varphi_j(X_i).$$

d) Montrer que $\tilde{\vartheta}_j$ est l'estimateur par la méthode des moments du paramètre ϑ_j.

e) On suppose maintenant que

$$Y_i = f(X_i) + U_i$$

où les variables U_i sont iid indépendantes de $\{X_i\}_{i=1,\ldots,n}$. On suppose de plus que la variance $\sigma^2 = \mathbf{E}[U_i^2]$ existe et est connue. Calculer le biais $b_{n,N}(x)$ de l'estimateur $\tilde{f}_{n,N}(x)$. Comment se comporte-t-il lorsque N augmente ?

f) Pour toute fonction $h \in L^2([0,1]^d)$, on note $\|h\| = [\int_{[0,1]^d} h^2(x)\, dx]^{1/2}$. Montrer que le risque quadratique intégré $R(\tilde{f}_{n,N}, f) = \mathbb{E}[\|\tilde{f}_{n,N} - f\|^2]$ est borné par $\sum_{j=N+1}^{\infty} \vartheta_j^2 + N(\|f\|_\infty^2 + \sigma^2)/n$, où

$$\|f\|_\infty = \sup_x |f(x)|.$$

Comment choisiriez-vous le paramètre N si vous connaissiez la fonction f ?

g) Supposons maintenant que f est bornée par M et l'on connaît un entier $k > 0$ et un réel $L > 0$ tels que $\sum_{j=1}^{\infty} j^{2k}\vartheta_j^2 \leq L$. Prouver que $\sum_{j>N} \vartheta_j^2 \leq LN^{-2k}$ et en déduire une majoration du risque $R(\tilde{f}_{n,N}, f)$. Explicitez la valeur de N (en fonction de n, k, L, M et σ) qui minimise ce majorant de $R(\tilde{f}_{n,N}, f)$.

h) On suppose maintenant que pour un entier naturel $N_0 < n$, le vecteur $(f(X_1), \ldots, f(X_n))'$ appartient à l'espace vectoriel engendré par les vecteurs $\{(\varphi_j(X_1), \ldots, \varphi_j(X_n))';\ 1 \leq j \leq N_0\}$. Montrer que $\hat{\sigma}^2_{N_0} = \frac{1}{n-N_0}\|(I_{n\times n} - A_{N_0})\boldsymbol{Y}\|^2$ est un estimateur sans biais de σ^2.

2 Simulations

On considére le cas unidimensionnel ($d = 1$) et choisit comme base orthonormée de $L^2([0,1])$ la base trigonométrique : $\varphi_1(x) \equiv 1$ et

$$\varphi_j(x) = \begin{cases} \sqrt{2}\cos(2k\pi x), & \text{si } k = (j+1)/2 \in \mathbb{Z}, \\ \sqrt{2}\sin(2k\pi x), & \text{si } k = j/2 \in \mathbb{Z}, \end{cases}, \qquad j = 1, 2, \ldots.$$

On veut vérifier que la méthode de sélection automatique du niveau de troncature donne des résultats satisfaisants. Pour cela :

- ▷ Poser $n = 100$ et générer n variables iid $X_1, \ldots, X_n$ de loi uniforme sur $[0,1]$.
- ▷ Choisir $f(x) = (x^2 2^{(x-1)} - (x-0.5)^3)\sin(10x)$, $\sigma = 0.2$ et calculer le vecteur $\boldsymbol{Y} = (f(X_1), \ldots, f(X_n))' + \sigma\boldsymbol{\xi}$ où $\boldsymbol{\xi}$ est un vecteur gausien $\mathcal{N}(0, I_{n\times n})$.
- ▷ Tracer le nuage des points (X_i, Y_i), $i = 1, \ldots, n$ et, dans le même repère orthogonal la courbe de la fonction f.
- ▷ Pour $N = 5, 10, 15, 20, \ldots, 50$, calculer l'estimateur $\tilde{f}_{n,N}$ et tracer sa courbe superposée de la courbe de f et du nuage des points $\{(X_i, Y_i)\}$. Déterminer visuellement la valeur de N qui correspond au meilleur estimateur.
- ▷ Calculer l'estimateur $\hat{\sigma}^2_{N_0}$ pour $N_0 = 50$ et déterminer

$$\hat{N} = \arg\min_{N=1,\ldots,50} \Big(\|(I_{n\times n} - A_N)\boldsymbol{Y}\|^2 - (n-2N)\hat{\sigma}^2_{N_0} \Big).$$

Cette valeur de $\hat{N}$, est-elle significativement différente de la valeur "optimale" déterminée dans la question précédente ?

- ▷ Tracer la courbe de l'estimateur $\tilde{f}_{n,\hat{N}}$ superposée de la courbe de f.
- ▷ Répéter cette expérience 100 fois ; on obtient ainsi les valeurs $\hat{N}_1, \ldots, \hat{N}_{100}$. Pour avoir une idée de la répartition de ces valeurs, on pourra tracer l'histogramme de $\hat{N}_1, \ldots, \hat{N}_{100}$.

(auteur du texte : A. Dalalyan).

Projet 2 : Estimation minimax dans le modèle gaussien

Supposons que l'on dispose d'une observation $X = (X_1, \dots, X_n) \in \mathbb{R}^n$ dont la loi appartient à une famille de lois de probabilité $(\mathbb{P}_\vartheta, \vartheta \in \Theta)$ sur $\mathbb{R}^n$, où $\Theta \subseteq \mathbb{R}^n$ est un ensemble donné. Soit $\hat{\vartheta} : \mathbb{R}^n \to \mathbb{R}^n$ une fonction borélienne qui définit l'estimateur $\hat{\vartheta} = \hat{\vartheta}(X)$ de ϑ. Notons $R(\hat{\vartheta}, \vartheta)$ le risque quadratique de l'estimateur $\hat{\vartheta}$ au point $\vartheta \in \Theta$:

$$R(\hat{\vartheta}, \vartheta) \triangleq \mathbb{E}_\vartheta \Big[\frac{1}{n} \big\| \hat{\vartheta}(X) - \vartheta \big\|^2 \Big],$$

où $\mathbb{E}_\vartheta$ désigne l'espérance par rapport à $\mathbb{P}_\vartheta$ et $\| \cdot \|$ est la norme euclidienne dans $\mathbb{R}^n$. Une question qui se pose alors est de trouver un estimateur optimal par rapport à ce critère de risque. Comme il n'existe pas d'estimateur $\hat{\vartheta}$ ayant le risque $R(\hat{\vartheta}, \vartheta)$ minimal pour tout ϑ (on pourra démontrer ce résultat négatif), la définition valide de l'optimalité repose sur le passage du risque ponctuel $R(\hat{\vartheta}, \vartheta)$ au risque maximal

$$\mathcal{R}^*(\hat{\vartheta}) = \sup_{\vartheta \in \Theta} R(\hat{\vartheta}, \vartheta).$$

On dit que l'estimateur $\tilde{\vartheta}$ est *optimal au sens minimax sur* Θ (ou, pour abréger, $\tilde{\vartheta}$ est un *estimateur minimax sur* Θ), si

$$\mathcal{R}^*(\tilde{\vartheta}) = \min_{\hat{\vartheta}} \mathcal{R}^*(\hat{\vartheta}) = \min_{\hat{\vartheta}} \sup_{\vartheta \in \Theta} R(\hat{\vartheta}, \vartheta),$$

où $\min_{\hat{\vartheta}}$ désigne le minimum sur tous les estimateurs. La valeur

$$\min_{\hat{\vartheta}} \sup_{\vartheta \in \Theta} R(\hat{\vartheta}, \vartheta)$$

s'appelle *risque minimax sur* Θ.

Bien évidemment, la forme de l'estimateur minimax dépend de l'ensemble Θ. L'objectif de ce projet est de trouver les estimateurs minimax (ou asymptotiquement minimax quand $n \to \infty$) pour quelques exemples importants de Θ dans le modèle gaussien basique :

$$X_i = \vartheta_i + \xi_i, \quad i = 1, \dots, n, \tag{1}$$

où $\vartheta_1, \ldots, \vartheta_n$ sont les coordonnées de ϑ et $\xi_1, \ldots, \xi_n$ sont des variables gaussiennes i.i.d. de moyenne 0 et de variance $\sigma^2 > 0$.

L'analyse de l'optimalité au sens minimax s'appuie sur la notion de risque de Bayes. Le risque de Bayes de l'estimateur $\hat{\vartheta}$ est défini par

$$\mathcal{R}^B(\hat{\vartheta}) = \int_\Theta R(\hat{\vartheta}, \vartheta)\, \pi(d\vartheta),$$

où π est une mesure de probabilité sur Θ appelée loi a priori de ϑ. Il est utile de noter que

$$\mathcal{R}^*(\hat{\vartheta}) \geq \mathcal{R}^B(\hat{\vartheta}) \tag{2}$$

pour tout estimateur $\hat{\vartheta}$ et toute loi a priori π. L'estimateur $\hat{\vartheta}^B$ qui fournit le minimum du risque de Bayes parmi tous les estimateurs s'appelle estimateur de Bayes.

Question 1. On s'intéressera d'abord à la forme de l'estimateur de Bayes pour une famille $(\mathbb{P}_\vartheta, \vartheta \in \Theta)$ générale. Pour abréger les notations, on peut considérer ϑ comme une variable aléatoire de loi π, $\mathbb{P}_\vartheta$ comme la loi conditionnelle de X sachant ϑ et le risque de Bayes comme l'espérance de $\|\hat{\vartheta}(X) - \vartheta\|^2/n$ par rapport à la loi jointe de (X, ϑ). Montrez que l'on peut alors écrire l'estimateur de Bayes sous la forme : $\hat{\vartheta}^B = \mathbb{E}(\vartheta|X) = (\mathbb{E}(\vartheta_1|X), \ldots, \mathbb{E}(\vartheta_n|X))$, i.e., $\hat{\vartheta}^B$ est l'espérance de la loi conditionnelle de ϑ sachant X, dite "loi a posteriori" de ϑ.

Nous allons supposer dans la suite que $\mathbb{P}_\vartheta$ est engendré par les observations gaussiennes (1). Considérons d'abord le cas où il n'y a aucune contrainte sur ϑ, i.e., $\Theta = \mathbb{R}^n$.

Question 2. Soit $\Theta = \mathbb{R}^n$ et soit π la loi gaussienne sur $\mathbb{R}^n$ de moyenne 0 et de matrice de covariance diagonale, avec les éléments diagonaux $\sigma_i^2 > 0$, $i = 1, \ldots, n$. Explicitez l'estimateur de Bayes $\hat{\vartheta}^B$ ainsi que la valeur minimale du risque de Bayes $\mathcal{R}^B(\hat{\vartheta}^B)$.

Question 3. Montrez que l'estimateur $\tilde{\vartheta} = X$ est minimax sur $\Theta = \mathbb{R}^n$. On cherchera d'abord le risque $\mathcal{R}^*(X)$, puis on le comparera avec la valeur minimale du risque de Bayes calculée dans la question précédente pour $\sigma_i^2 = A$, $\forall\ i$.

Question 4. Considérons maintenant l'ensemble des paramètres

$$\Theta = \Theta_0 \triangleq \Big\{\vartheta \in \mathbb{R}^n :\ \vartheta_1 = \vartheta_2 = \cdots = \vartheta_n\Big\}.$$

Dans ce cas, le modèle (1) devient le modèle de n-échantillon de la loi $\mathcal{N}(a, \sigma^2)$, où $a \in \mathbb{R}$ est le seul paramètre inconnu ($\vartheta_i = a$, $i = 1, \dots, n$). Montrez que $\bar{X}_n$, la moyenne arithmétique des X_i, est un estimateur minimax de a par rapport au risque quadratique usuel sur $\mathbb{R}$. Par conséquent, $\tilde{\vartheta} = (\bar{X}_n, \dots, \bar{X}_n)$ est un estimateur minimax sur Θ_0 pour le modèle (1).

Finalement, considérons l'ensemble des paramètres qui est une boule euclidienne dans $\mathbb{R}^n$:

$$\Theta = \Theta(Q) \triangleq \Big\{\vartheta \in \mathbb{R}^n : \ \frac{1}{n}\|\vartheta\|^2 \le Q\Big\},$$

où $Q > 0$ est une constante donnée. Il s'avère que l'estimateur X n'est pas minimax sur $\Theta(Q)$. De plus, la forme de l'estimateur minimax sur $\Theta(Q)$ n'est connue que pour des valeurs particulières de Q. Par contre, il est possible de construire un estimateur qui est minimax parmi les estimateurs linéaires et asymptotiquement minimax parmi tous les estimateurs, au sens qui sera précisé dans la suite.

Question 5. Introduisons une famille des estimateurs linéaires ($\hat{\vartheta}(\lambda), \lambda \in \mathbb{R}$) définie par $\hat{\vartheta}_i(\lambda) = \lambda X_i$, $i = 1, \dots, n$, où $\hat{\vartheta}_i(\lambda)$ est la $i$ème coordonnée de $\hat{\vartheta}(\lambda)$. Explicitez *l'estimateur minimax linéaire* sur $\Theta(Q)$, i.e., l'estimateur $\hat{\vartheta}(\lambda^*)$ tel que

$$\sup_{\vartheta \in \Theta(Q)} R(\hat{\vartheta}(\lambda^*), \vartheta) = \min_{\lambda \in \mathbb{R}} \sup_{\vartheta \in \Theta(Q)} R(\hat{\vartheta}(\lambda), \vartheta) \triangleq \mathcal{R}^*_{\text{Lin}}(Q)$$

et montrez que le *risque minimax linéaire* $\mathcal{R}^*_{\text{Lin}}(Q)$ vaut $\frac{Q\sigma^2}{Q+\sigma^2}$.

Question 6. Montrez que, pour tout estimateur $\hat{\vartheta}$, il existe un estimateur $\hat{\vartheta}'$ à valeurs dans $\Theta(Q)$, tel que $R(\hat{\vartheta}, \vartheta) \ge R(\hat{\vartheta}', \vartheta)$, $\forall \vartheta \in \Theta(Q)$. En déduire qu'il suffit de considérer le risque minimax pour les estimateurs à valeurs dans $\Theta(Q)$:

$$\inf_{\hat{\vartheta}} \sup_{\vartheta \in \Theta(Q)} R(\hat{\vartheta}, \vartheta) = \inf_{\hat{\vartheta} \in \Theta(Q)} \sup_{\vartheta \in \Theta(Q)} R(\hat{\vartheta}, \vartheta).$$

Question 7. Montrez la minoration asymptotique du risque minimax :

$$\liminf_{n \to \infty} \inf_{\hat{\vartheta}} \sup_{\vartheta \in \Theta(Q)} R(\hat{\vartheta}, \vartheta) \ge \frac{Q\sigma^2}{Q + \sigma^2}\,.$$

Indication : utilisez la Question 6 et la minoration par le risque de Bayes avec la loi a priori π définie dans la Question 1 et $\sigma_i^2 = \delta Q$, $\delta \in]0, 1[$, $i = 1, \dots, n$.

Le point délicat est que le support de cette loi n'est pas égal à $\Theta(Q)$, donc on ne peut pas appliquer (2) directement.

Question 8. Déduire de ce qui précède que l'estimateur minimax linéaire $\hat{\vartheta}(\lambda^*)$ est aussi *asymptotiquement* minimax sur $\Theta(Q)$ parmi *tous* les estimateurs en ce sens que

$$\lim_{n\to\infty} \frac{\sup_{\vartheta\in\Theta(Q)} R(\hat{\vartheta}(\lambda^*),\vartheta)}{\inf_{\hat{\vartheta}} \sup_{\vartheta\in\Theta(Q)} R(\hat{\vartheta},\vartheta)} = 1.$$

Un grand défaut de l'estimateur minimax linéaire $\hat{\vartheta}(\lambda^*)$ est ce qu'il dépend du rayon Q qui est généralement inconnu dans la pratique. Cependant, de façon remarquable, il existe des estimateurs qui ne dependent pas de Q et qui sont asymptotiquement minimax sur $\Theta(Q)$ parmi tous les estimateurs *simultanément pour tous* $Q > 0$. De tels estimateurs sont appelés *adaptatifs* sur l'échelle des ensembles $\{\Theta(Q), Q > 0\}$. Notre objectif est maintenant de mettre en évidence le fait que l'estimateur de Stein

$$\hat{\vartheta}_S = \left(1 - \frac{n\sigma^2}{\|X\|^2}\right) X$$

est adaptatif.

Question 9. A l'aide du Lemme de Stein, montrez que pour tout $\vartheta \in \mathbb{R}^n$,

$$\mathbb{E}_\vartheta \|\hat{\vartheta}_S - \vartheta\|^2 \le n\sigma^2 - \beta \mathbb{E}_\vartheta \left(\|X\|^{-2}\right) \tag{3}$$

avec la constante $\beta > 0$ que l'on précisera. Transformez (3) en :

$$\mathbb{E}_\vartheta \|\hat{\vartheta}_S - \vartheta\|^2 \le \frac{n\sigma^2\|\vartheta\|^2 + 4n\sigma^4}{\|\vartheta\|^2 + n\sigma^2}\,. \tag{4}$$

Déduisez de (4) et de ce qui précède que l'estimateur de Stein est adaptatif sur l'échelle $\{\Theta(Q), Q > 0\}$:

$$\forall\ Q > 0\ :\quad \lim_{n\to\infty} \frac{\sup_{\vartheta\in\Theta(Q)} R(\hat{\vartheta}_S,\vartheta)}{\inf_{\hat{\vartheta}} \sup_{\vartheta\in\Theta(Q)} R(\hat{\vartheta},\vartheta)} = 1.$$

L'estimateur $\hat{\vartheta}_S$ est-il minimax ou asymptotiquement minimax sur $\Theta = \mathbb{R}^n$?

(auteur du texte : A. Tsybakov)

Université Paris Dauphine Année 2014-2015 Apprentissage statistique et grande dimension

---

Projet 3 : Facteurs prédictifs du diabète par Lasso et Elastic-Net

---

L'objectif de ce projet est d'analyser les facteurs prédictifs du diabète à partir de données physiologiques et sérologiques de $n = 442$ patients souffrant du diabète. La variable y reflète la progression de la maladie et les $p = 64$ variables explicatives $x^{(1)}, \ldots, x^{(64)}$ décrivent l'âge, le sexe, l'indice de masse corporel, diverses mesures sérologique, etc. L'objectif est double :

(a) parvenir à prédire y à partir des différentes mesures $x^{(1)}, \ldots, x^{(64)}$,

(b) sélectionner les variables $x^{(j)}$ influentes pour prédire y.

Les données sont à télécharger

Y : `http://www.cmap.polytechnique.fr/~giraud/MAP433/Y.txt`

X : `http://www.cmap.polytechnique.fr/~giraud/MAP433/X.txt`

Le modèle

Nous allons considérer le modèle linéaire :

$$y_i = \beta_1 x_i^{(1)} + \ldots, \beta_{64} x_i^{(64)} + \varepsilon_i, \quad i = 1, \ldots, n.$$

En notant Y le vecteur d'entrées y_i, β le vecteur d'entrées β_j et X la matrice $n \times p$ de lignes $X[i,] = [x_i^{(1)}, \ldots, x_i^{(64)}]$, le modèle précédent est équivalent à $Y = X\beta + \varepsilon$.

1 Partie préliminaire

N1. Les variables sont-elles centrées ? réduites ? Les variables explicatives sont-elle corrélées ?

T1. On note $\hat{\beta}_{OLS}$ l'estimateur de β obtenu en minimisant $\hat{\beta}_{OLS} = \text{argmin}_\beta \|\text{Y} - \text{X}\beta\|^2$. Montrez que $\hat{\beta}_{OLS} = (X^T X)^{-1} X^T Y$.

N2. Calculez $\hat{\beta}_{OLS}$ numériquement. Quelles variables semblent les plus importantes ?

2 Estimateur Lasso

L'estimateur Lasso est obtenu comme solution du problème de minimisation :

$$\hat{\beta}^\lambda = \operatorname*{argmin}_\beta F_\lambda(\beta) \quad \text{où} \quad F_\lambda(\beta) = \frac{1}{2}\|Y - X\beta\|^2 + \lambda \sum_{j=1}^{p} |\beta_j| \quad \text{pour } \lambda > 0. \tag{1}$$

On notera X_j la j-ième colonne de X et on supposera pour simplifier que $X_j^T X_j = 1$ pour $j = 1, \ldots, p$.

2.1 Propriétés élémentaires

T2. Quelle propriété caractéristique possède la fonction F_λ ? Supposons que le rang de X est égal à p. La fonction F_λ atteint-elle son minimum ? est-il unique ?

T3. Montrer que

$$\frac{\partial}{\partial \beta_j} F_\lambda(\beta) = -X_j^T(Y - X\beta) + \lambda \frac{\beta_j}{|\beta_j|} \quad \text{pour tout } \beta_j \neq 0.$$

T4. En déduire une valeur $\lambda_{\max}$ telle que si $\hat{\beta}^\lambda = 0$ alors $\lambda \geqslant \lambda_{\max}$. Réciproque (facultatif) ?

Dans les deux questions suivantes, nous supposerons que $X^T X = I_p$ où I_p est la matrice identité.

T5. Que vaut $\hat{\beta}^\lambda$ dans ce cas ? Quelles variables sont sélectionnées (variables j telles que $\hat{\beta}_j^\lambda \neq 0$) ?

T6. Supposons que les ε_i sont i.i.d. de loi $\mathcal{N}(0, \sigma^2)$. En utilisant la propriété $\mathbf{P}(\xi \geqslant x) \leqslant e^{-x^2/2}$ pour ξ de loi $\mathcal{N}(0,1)$ et $x > 0$, montrez que la probabilité que l'ensemble des variables sélectionnées par $\hat{\beta}\sqrt{\alpha\sigma^2 \log(p)}$ ne soit pas inclus dans $\{j : \beta_j \neq 0\}$ est inférieure à $2p^{-(\alpha/2-1)}$ pour $\alpha > 2$.

2.2 Calcul de l'estimateur Lasso

Dans le cas général où $X^T X \neq I_p$, il n'y a pas de formule explicite pour $\hat{\beta}^\lambda$. On peut cependant calculer efficacement $\hat{\beta}^\lambda$ avec un algorithme très simple.

T7. Soit $\beta \in \mathbf{R}^p$ et $\beta^{(j)}$ défini par $\beta_k^{(j)} = \beta_k$ si $k \neq j$ et

$$\beta_j^{(j)} = R_j \left(1 - \frac{\lambda}{|R_j|}\right)_+ \quad \text{avec} \quad R_j = X_j^T \left(Y - \sum_{k \neq j} \beta_k X_k\right).$$

Montrer que $F_\lambda(\beta^{(j)}) \leqslant F_\lambda(\beta)$ avec inégalité stricte si $\beta^{(j)} \neq \beta$.

T8. En déduire un algorithme de minimisation numérique de F_λ. Vaut-il mieux implémenter cet algorithme avec un langage compilé ou un langage interprété ? Quelle est la nature du langage que vous avez utilisé ?

N3. Calculez $\hat{\beta}^\lambda$ pour tout $\lambda \in \Lambda = \{k\lambda_{\max}/10^3 : k = 1, \ldots, 10^3\}$. On pourra procéder comme suit : on commencera par les plus grandes valeurs de λ et pour calculer $\hat{\beta}^{k\lambda_{\max}/10^3}$ on initialisera l'algorithme avec $\hat{\beta}^{(k+1)\lambda_{\max}/10^3}$ (cela permet un net gain en temps de calcul).

N4. Tracez pour chaque j la valeur de $\hat{\beta}_j^\lambda$ en fonction de λ (vous pouvez superposer les courbes sur un même graphique à l'aide de différentes couleurs). Qu'observez-vous ?

3 Cross-Validation

L'objectif de cette partie est de sélectionner la "meilleure" valeur $\lambda \in \Lambda$ pour prédire y à l'aide de $\sum_j \hat{\beta}_j^\lambda x^{(j)}$. Plus précisément, si $y_{new}, x_{new}^{(j)}$ sont de nouvelles observations, on aimerait choisir le λ_* qui donne en moyenne le plus petit résidu $(y_{new} - \sum_j \hat{\beta}_j^\lambda x_{new}^{(j)})^2$. Ce λ_* est inconnu, mais on peut essayer de l'estimer à l'aide de la K-fold cross-validation.

Le principe est le suivant : pour $k = 1, \ldots, K$ on note $I_k = \{1 + (k-1)n/K, \ldots, kn/K\}$ et $I_{-k} = \{1, \ldots, n\} \setminus I_k$. On calcule les estimateurs Lasso $\hat{\beta}^{\lambda:k}$ en se restreignant aux observations pour les individus d'indice i dans I_{-k}. Autrement dit, en notant $X[I_{-k},]$ la matrice obtenue en ne conservant que les lignes d'indice dans I_{-k}, l'estimateur $\hat{\beta}^{\lambda:k}$ est solution de (1) avec Y remplacé par $Y[I_{-k}]$ et X remplacé par $X[I_{-k},]$. La performance de l'estimateur $\hat{\beta}^\lambda$ est alors estimée par

$$\mathcal{R}(\lambda) = \frac{1}{K} \sum_{k=1}^{K} \left\| Y[I_k] - X[I_k,] \hat{\beta}^{\lambda:k} \right\|^2, \quad \text{pour } \lambda \in \Lambda$$

et l'estimateur cross-validé est défini par $\hat{\beta}^{CV} = \hat{\beta}^{\hat{\lambda}}$ où $\hat{\lambda}$ est un minimiseur de $\mathcal{R}(\lambda)$ sur Λ.

N5. Calculez l'estimateur $\hat{\beta}^{CV}$ pour $K = 13$. Que vaut $\mathcal{R}(\hat{\lambda})$?

N6. Quelles sont les variables sélectionnées ? Comparez ce résultat à celui obtenu avec $\hat{\beta}_{OLS}$.

4 Elastic-Net

Lorsque les variables $x^{(j)}$ présentent de fortes corrélations, il est souhaitable de modifier un peu l'estimateur Lasso. Par exemple, on peut modifier le problème de minimisation comme suit (Elastic-net) : pour $\lambda > 0, \mu \geqslant 0$

$$\hat{\beta}^{\lambda,\mu} = (1+\mu)\times \operatorname*{argmin}_{\beta} F_{\lambda,\mu}(\beta) \quad \text{où} \quad F_{\lambda,\mu}(\beta) = \frac{1}{2}\|Y - X\beta\|^2 + \frac{1}{2}\mu\|\beta\|^2 + \lambda \sum_{j=1}^{p} |\beta_j|. \tag{2}$$

T9. Que dire de la fonction $F_{\lambda,\mu}$? En vous inspirant de la Partie 2, proposez un algorithme pour réaliser la minimisation $\operatorname{argmin}_{\beta} F_{\lambda,\mu}(\beta)$.

N7. Pour chaque $\mu \in \{0, 0.01, 0.02, 0.05, 0.1, 1\}$, tracez les valeurs de $\hat{\beta}_j^{\lambda,\mu}$ en fonction de $\lambda \in \Lambda$.

N8. Calculez l'estimateur cross-validé $\hat{\beta}^{\hat{\lambda},\hat{\mu}}$ en faisant varier μ dans

$$\{0, 0.01, 0.02, 0.05, 0.1, 1\}$$

et λ dans Λ. Quelles sont les variables sélectionnées ? Que vaut le risque $\mathcal{R}(\hat{\lambda}, \hat{\mu})$? a-t-on un gain comparativement au Lasso ?

(auteur du texte : C. Giraud)

Projet 4 : Débruitage d'un signal par seuillage d'ondelettes

Résumé : Nous abordons certaines questions relatives à la reconstruction d'un signal 1-dimensionnel à partir d'observations bruitées.
Mots-clés : vecteur gaussien, simulation de variables aléatoires, projection orthogonale.

1 Approximation par moyennes locales

Soit f une fonction de $L^2([0,1])$. On définit son approximation à l'échelle $j \geqslant 0$ en posant

$$f_j(x) = 2^j \int_{I_{j,k}} f(t)dt \quad \text{si} \quad x \in I_{j,k},\ k = 0, \ldots, 2^j - 1,$$

où $I_{j,k}$ désigne l'intervalle $[k2^{-j}, (k+1)2^{-j}[$. Autrement dit, f est approchée par sa moyenne sur chaque intervalle $I_{j,k}$. L'approximation f_j peut aussi s'interpréter comme une projection : si P_j désigne la projection orthogonale sur le sous-espace vectoriel V_j de $L^2([0,1])$ défini par

$$V_j = \{f \in L^2([0,1])\ ;\ f \text{ est constante sur } I_{j,k},\ k = 0, \ldots, 2^j - 1\},$$

on a le résultat suivant :

$$f_j = P_j f. \qquad (\star)$$

T1 Prouver $(\star)$.
(Indication : on pourra introduire les fonctions

$$\big(\varphi_{j,k}(x) = 2^{j/2}\varphi(2^j x - k),\ k = 0, \ldots, 2^j - 1\big),$$

avec $\varphi(x) = 1$ si $x \in [0,1]$ et 0 sinon.)

T2 Prouver que l'approximation $P_{j+1}f$ contient plus d'information sur f que $P_j f$ dans le sens suivant :

$$P_j f_{|I_{j,k}} = \frac{1}{2}\left[P_{j+1}f_{|I_{j+1,2k}} + P_{j+1}f_{|I_{j+1,2k+1}}\right]. \qquad (1)$$

Notons de plus que l'on dispose d'un contrôle de l'erreur d'approximation de f par $P_j f$ dès lors que f possède suffisamment de régularité :

Définition 1 *Soit $0 < \alpha \leqslant 1$ et $L > 0$. Une fonction $f : [0,1] \to \mathbb{R}$ vérifie la condition de régularité $H(\alpha, L)$ si pour tout $x, y \in [0,1]$:*

$$|f(y) - f(x)| \leqslant L|y - x|^{\alpha}.$$

T3 Prouver que si f vérifie la condition $H(\alpha, L)$, alors

$$\|P_j f - f\|_{L^2} \leqslant L 2^{-j\alpha}.$$

2 Lissage par projection

2.1 Un modèle de "signal plus bruit"

On suppose que l'on observe la réalisation de $(Y_{J,k}, k = 0, \ldots, 2^J - 1)$, avec

$$Y_{J,k} = 2^J \int_{I_{J,k}} f(t)dt + \xi_{J,k} \tag{2}$$

où J est un niveau de résolution maximal, et $\xi_{J,k}$ représente une erreur expérimentale systématique. On suppose que les $\xi_{J,k}$ sont des variables aléatoires gaussiennes centrées réduites, indépendantes. Lorsque J est grand, le modèle postulé par (2) correspond à l'échantillonnage bruité d'un signal. En posant $\sigma_J = 2^{-J/2}$ et $Z_{J,k} = \sigma_J Y_{J,k}$, on se ramène donc à l'observation de

$$Z_{J,k} = c_{J,k}(f) + \sigma_J \xi_{J,k}, \quad k = 0, \ldots, 2^J - 1,$$

où $c_{j,k}(f) = \int_{[0,1]} f(t)\varphi_{jk}(t)dt$. Ceci donne lieu à la reconstruction *bruitée* de f à l'échelle J :

$$\hat{f}_J := \sum_{k=0}^{2^J-1} Z_{J,k}\varphi_{J,k}.$$

Bien que l'on ait $\mathbb{E}\{Z_{J,k}\} = c_{J,k}(f)$ ($\mathbb{E}$ désigne l'espérance mathématique sur un espace de probabilité adéquat) et donc $\mathbb{E}\{\hat{f}_J\} = f_J$, l'estimateur $\hat{f}_J$ de f n'est pas bon : on peut écrire $\hat{f}_J = f_J + h_J$, avec

$$h_J = \sigma_J \sum_{k=0}^{2^J-1} \xi_{J,k}\varphi_{J,k},$$

et, pour chaque $x \in [0,1]$, $h_J(x)$ est une variable gaussienne, centrée, de variance 1 qui n'est donc pas "petite", même lorsque J est grand.

T4 Formaliser cette dernière remarque.

2.2 L'estimateur par projection $f_j^\star$

Dans ce contexte, l'idée de projection consiste à *lisser* les observations $Z_{J,k}$, en projetant $\hat{f}_J$ sur un espace d'approximation V_j plus *grossier* que V_J, c'est-à-dire tel que j soit petit devant J. On définit alors

$$f_j^\star := P_j \hat{f}_J,$$

et il convient de choisir judicieusement le niveau de projection, ou de lissage j. Pour cela, étudions l'erreur moyenne quadratique $e_{J,j} = \mathbb{E}\{\|f - f_j^\star\|_{L^2}^2\}$ sous l'hypothèse $H(\alpha, L)$.

T5 Montrer

$$e_{J,j} = \|f - P_j f\|_{L^2}^2 + \mathbb{E}\{\|P_j h_J\|_{L^2}^2\}.$$

T6 Montrer que l'on peut écrire

$$P_j h_J = \sum_{k=0}^{2^j-1} \eta_{j,k}^{(J)} \varphi_{jk},$$

où les $\eta_{jk}^{(J,k)}$ sont des variables aléatoires gaussiennes centrées, dont la variance ne dépend pas de j et vaut $2^{-J} = \sigma_J^2$.

T7 En déduire que l'erreur moyenne quadratique $e_{J,j}$, sous l'hypothèse $H(\alpha, L)$, est majorée par

$$L^2 2^{-2j\alpha} + 2^{j-J},$$

ce qui fournit une erreur minimale de l'ordre de

$$c(\alpha, L) 2^{-2J\alpha/(2\alpha+1)}.$$

Choisir j trop grand revient a *sous-lisser* le signal bruité, et choisir j trop petit revient à le *sur-lisser*. L'inconvénient de cette méthode est que le choix optimal de j dépend explicitement de la connaissance *a priori* de la régularité α du signal inconnu f, ce qui, sauf exception notoire, est peu réaliste. On va circonvenir à cet inconvénient en raffinant l'analyse de l'approximation par moyennes locales.

3 Représentation multi-échelle d'un signal

Avec les notations du paragraphe 1, écrivons

$$P_{j+1} f = P_j f + Q_j f,$$

où $Q_j f = (P_{j+1} - P_j) f$ désigne la projection orthogonale sur le complémentaire W_j de V_j dans V_{j+1}. La propriété (1) montre que $Q_j f$ *oscille*, dans le sens où :

$$Q_j f_{|I_{j+1,2k}} = -Q_j f_{|I_{j+1,2k+1}}. \tag{3}$$

La propriété d'oscillation (3) nous permet d'écrire

$$Q_j f = \sum_{k=0}^{2^j-1} d_{jk}(f)\psi_{j,k},$$

où $\psi_{j,k}(x) = 2^{j/2}\psi(2^j x - k)$, avec $\psi(x) = 2^{j/2}\psi(2^j x - k)$, où $\psi(x) = 1$ si $x \in [0, \frac{1}{2}[$, -1 si $x \in [\frac{1}{2}, 1[$ et 0 sinon. La famille $\big(\psi_{j,k}, k = 0, \ldots, 2^j - 1\big)$ constitue une base orthonormée de W_j. On a donc nécessairement $d_{j,k}(f) = \int_{[0,1]} f(t)\psi_{j,k}(t)dt$. En itérant cette décomposition, on obtient, pour $0 \leqslant j_0 < j_1$:

$$P_{j_1} f = P_{j_0} f + \sum_{j=j_0}^{j_1-1} Q_j f,$$

ou encore

$$\sum_{k=0}^{2^{j_1}-1} c_{j_1,k}(f)\varphi_{j_1 k} = \sum_{k=0}^{2^{j_0}-1} c_{j_0,k}(f)\varphi_{j_0,k} + \sum_{j=j_0}^{j_1-1}\sum_{k=0}^{2^j-1} d_{j,k}(f)\psi_{jk}, \tag{4}$$

ce qui exprime la décomposition de f à l'échelle d'approximation *fine* j_1 comme somme d'une décomposition *grossière* à l'échelle j_0 à laquelle on adjoint une somme de détails ou de fluctuations à des échelles intermédiaires.

La formule (4) doit être comprise comme un changement de base orthonormal de V_{j_1} : les $\big(\varphi_{j_1 k}, k = 0, \ldots, 2^{j_1} - 1\big)$ d'une part, et les $\big(\varphi_{j_0 k}, k = 0, \ldots, 2^{j_0} - 1, \psi_{jk}, j = j_0, \ldots, j_1 - 1, k = 0, \ldots, 2^j - 1\big)$ sont des bases orthonormées de V_{j_1} ; toute fonction de V_{j_1} admet une décomposition unique dans chacune de ces bases :

$$c_{j,k} = \frac{1}{\sqrt{2}}[c_{j+1,2k} + c_{j+1,2k+1}], \text{ et } d_{j,k} = \frac{1}{\sqrt{2}}[c_{j+1,2k} - c_{j+1,2k+1}], \tag{5}$$

et la transformation inverse est donnée par

$$c_{j+1,2k} = \frac{1}{\sqrt{2}}[c_{j,k} + d_{j,k}], \text{ et } c_{j+1,2k+1} = \frac{1}{\sqrt{2}}[c_{j,k} - d_{j,k}]. \tag{6}$$

On peut alors récapituler cette décomposition par les deux algorithmes suivants :

Décomposition (échelle fine j_1 vers échelle grossière j_0 plus les détails)
- Se donner des coefficients $c_{j_1 k}$.
- Calculer les $c_{j_1-1,k}$ et les $d_{j_1-1,k}$ en utilisant (5).
- Garder les détails $d_{j_1-1,k}$ et itérer la décomposition sur les c_{j_1-1} et ainsi de suite.
- Stopper à l'échelle j_0.

Reconstruction (échelle grossière j_0 plus les détails vers échelle fine j_1)
- Partir des coefficients $c_{j_0 k}$ et $d_{j_0 k}$.
- Calculer les $c_{j_0+1,k}$ en utilisant (6).
- Itérer la reconstruction en utilisant les $d_{j_0+1,k}$ et ainsi de suite.
- Stopper à l'échelle j_1 lorsque les $c_{j_1,k}$ sont calculés.

8 Démontrer les propriétés (3), (4), (5) et (6). Implémenter cet algorithme et le tester numériquement sur plusieurs exemples de signaux. En particulier, réfléchir à une représentation graphique de la décomposition multiéchelle. Quelle est la complexité de l'algorithme ?

4 Application à l'estimation d'un signal bruité : le seuillage

On part de l'observation (2). En appliquant l'algorithme de décomposition entre les échelles $j_1 = J$ et $j_0 = 0$, on observe aussi

$$\begin{cases} W_{jk} &= d_{jk}(f) + \sigma_J \widetilde{\xi_{jk}},\ k = 0, \dots 2^j - 1, j = 1, \dots, J \\ W_0 &= c_{00}(f) + \sigma_J \widetilde{\xi_{00}}, \end{cases}$$

où les $\widetilde{\xi_{jk}}$ sont des variables gaussiennes centrées réduites.

T9 Montrer que sous l'hypothèse $H(\alpha, L)$, la propriété d'oscillation

$$\int_{[0,1]} \psi(t)dt = 0$$

entraîne l'estimation

$$|d_{j,k}(f)| \leqslant L 2^{-j(\alpha+1/2)}.$$

Les $d_{j,k}(f)$ sont d'autant plus petits que f est régulière (c'est-à-dire α grand) ou que j est grand. Par ailleurs, le terme de bruit $\sigma_J \widetilde{\xi_{jk}}$ est grossièrement de l'ordre de $\sigma_J = 2^{-J/2}$, au sens où $\mathbb{E}\{(\widetilde{\xi_{jk}})^2\} = 1$. En conclusion, lorsque

l'observation $W_{j,k}$ n'est pas significativement plus grande que σ_J, elle n'apporte pas d'information sur f, au sens où le coefficient $d_{j,k}$ est dominé par le niveau de bruit σ_J. Ce principe donne lieu à l'algorithme de seuillage :

$$\hat{f}_J^{seuillage} = W_0 + \sum_{j=0}^{J} \sum_{k=0}^{2^j-1} T_{\sigma_J}(W_{jk})\psi_{j,k},$$

où $T_{\sigma_J}(x) = x$ si $|x| \geqslant \sigma_J\sqrt{2|\log \sigma_J|}$ et 0 sinon. Le choix de T_{σ_J} est motivé par la propriété suivante :

10 Montrer que

$$\mathbb{P}\{|d_{j,k} - W_{j,k}| \geqslant \sigma_J\sqrt{2|\log \sigma_J|}\} \text{ est "petit"}$$

et quantifier cette affirmation précisément. Montrer en particulier que cette probabilité est petite devant la vitesse optimale (renormalisée) de l'estimateur par projection.

On obtient ainsi $\hat{f}_J^{seuillage}$ en décomposant $\hat{f}_J$ dans la base

$$\{\varphi_{00}\} \cup \{\psi_{j,k}, k = 0, \ldots, 2^j - 1, j = 0, \ldots, J - 1\},$$

en ne conservant toutefois que les coefficients $W_{j,k}$ significatifs, ce qui permet de réduire la variance de l'estimation.

T10 (Difficile) Montrer que l'estimateur par seuillage $\hat{f}_J^{seuillage}$ atteint la vitesse optimale de l'estimateur par projection (à un facteur logarithmique près), sans avoir besoin de connaître α et L.

11 Implémenter l'estimateur par projection $\hat{f}_j$ pour différents niveaux de lissage j et différentes fonctions test. En particulier, on pourra remarquer que l'algorithme de décomposition de la section 3 fournit un procédé de calcul rapide de la projection $P_j g$ à partir de $P_J g$.
Pour simplifier la mise en oeuvre des algorithmes, on pourra faire (et justifier) l'approximation

$$2^J \int_{I_{J,k}} f(t)dt \text{ proche de } f(k2^{-J})$$

à l'échelle la plus fine J.

T12 On pourra justifier le calcul de la variance des $\eta_{j,k}^{(J)}$ ainsi que le calcul de l'erreur moyenne quadratique optimale de l'estimateur par projection.

Pour le choix de signaux sur lesquels tester la méthode, on pourra, par exemple, choisir les signaux

$$g_1(x) = \sin(2\pi x), \quad g_2(x) = 1_{[0,\frac{1}{3}[}(x) + \tfrac{1}{2}1_{[\frac{1}{3},\,1]}(x), \quad g_3(x) = \exp(x)$$

en discutant à chaque fois les méthodes de reconstruction selon les propriétés de régularité de g_i, $i = 1, 2, 3$. On pourra, en particulier, reconsidérer la méthode dans le cas (très particulier) où le signal f est constant.

Université Paris Dauphine Année 2014-2015 Apprentissage statistique et grande dimension

Projet 5 : Minimisation du risque empirique pour le problème d'agrégation convexe

1 Introduction au problème d'agrégation convexe

L'objectif de ce Projet Individuel est de proposer une introduction à l'apprentissage statistique, aux méthodes de processus empiriques et à la méthode de Maurey par le biais du probléme d'agrégation convexe.

On se propose d'étudier des données de type entrée/sortie. L'objectif étant d'inférer ou prédire une sortie associée à une nouvelle entrèe en fonction des données précédemment observées. On dispose de n données $(X_i, Y_i)_{i=1}^n$ où X_i est une donnée d'entrée à valeurs dans un espace mesurable quelconque $\mathcal{X}$ et Y_i est un "label" ou sortie associée à l'entrée X_i à valeurs dans un intervalle borné $[-b, b]$ pour un certain $b > 0$. On reÁoit une nouvelle entrée X et on souhaite prédire la sortie Y la plus naturellement associée à X en restant en accord avec ce qui a été observé avant. Ce problème a de multiple applications concrétes.

On modèlise ce problème de la manière suivante : les données (X_i, Y_i) pour $i = 1, \ldots, n$ et le "nouveau" couple entrée/sortie (X, Y) sont supposés indépendants et identiquement distribués. On souhaite construire des fonctions qui dépendent des données $\mathcal{D} := \{(X_1, Y_1), \ldots, (X_n, Y_n)\}$ et de la nouvelle entrée X pour prédire au mieux la sortie Y. De telles procédures sont appelées procèdures d'apprentissage, estimateurs ou statistiques. On va donc s'intérésser aux fonctions mesurables $\hat{f}_n : (\mathcal{X} \times [-b, b])^n \times \mathcal{X} \longmapsto \mathbb{R}$ telles que la distance moyenne entre $\hat{f}_n(\mathcal{D}, X)$ (la prédiction qui est faite à partir des données $\mathcal{D}$ pour l'entrée X) et Y (en quelque sorte la "vraie" sortie) soient la plus petite possible. On considére la distance L_2 (mais d'autres distances sont aussi envisageables). On définit alors le risque quadratique de $\hat{f}_n$ par

$$\begin{aligned} R(\hat{f}_n) &= \mathrm{E}_{(X,Y)}(\hat{f}_n(\mathcal{D}, X) - Y)^2 = \mathrm{E}\left[(\hat{f}_n(\mathcal{D}, X) - Y)^2 | \mathcal{D}\right] \\ &= \int_{\mathcal{X}\times\mathbb{R}} (\hat{f}_n(\mathcal{D}, x) - y)^2 d\mathbb{P}_{(X,Y)}(x, y) \end{aligned} \tag{1}$$

où $\mathrm{E}_{(X,Y)}$ est l'espérance par rapport à (X, Y) et $\mathbb{P}_{(X,Y)}$ est la mesure de probabilité de (X, Y). Pour simplifier l'écriture, on ne présisera plus que

les statistiques $\hat{f}_n$ dépendent des données $\mathcal{D}$. Suivant cette convention, le risque quadratique d'un estimateur $\hat{f}_n$ s'écrit $R(\hat{f}_n) = \mathrm{E}\left[(\hat{f}_n(X) - Y)^2 | \mathcal{D}\right]$. On cherche donc à constuire des estimateurs $\hat{f}_n$ ayant le plus petit risque quadratique $R(\hat{f}_n)$.

Dans ce projet individuel, on s'intéréssera à un certain type d'estimateur : ceux qui peuvent s'écrire comme combinaison convexe d'éléments d'un ensemble fini de fonctions de $\mathcal{X}$ dans $[-b, b]$. Un tel ensemble s'appelle un *dictionnaire*. Un dictionnaire peut se construire à partir d'éléments d'une base qu'on pense particulièrement bien adaptée au probléme traité, ou d'un grand nombre de fonctions simples comme des indicatrices de demi-espace ou encore, si on dispose d'autres données, on peut aussi construire une multitude d'estimateurs (possiblement non-adaptatifs) et en faire un dictionnaire, etc.. Qu'importe la manière dont a été construit ce dictionnaire, pour notre problème d'agrégation, on notera ses éléments par $f_1, \ldots, f_M$. Les f_j sont donc des fonctions de $\mathcal{X}$ à valeurs dans $[-b, b]$. On s'intéressera alors à des estimateurs de la forme

$$\hat{f}_n = \sum_{j=1}^{M} w_j f_j \tag{2}$$

où les poids w_j sont positifs et de somme égale à 1 (de telle sorte que $\hat{f}_n$ est bien une combinaison convexe d'éléments du dictionnaire). Un tel estimateur est appelé *méthode d'agrégation.* On souhaite choisir les poids w_j de telle sorte que (2) fasse aussi bien que la meilleure combinaison convexe d'élément dans $F = \{f_1, \ldots, f_M\}$, le dictionnaire. Les poids w_j devront donc être choisis à l'aide des données $\mathcal{D}$.

D'un point de vue mathématique, ce problème d'optimalité (i.e. "faire mieux que la meilleure combinaison convexe dans F) se traduit par une *inégalité oracle* : construire $\hat{f}_n$ telle que "avec grande probabilité (vis-à-vis des données)", on a

$$R(\hat{f}_n) \leqslant \inf_{f \in \mathrm{conv}(F)} R(f) + r(n, M) \tag{3}$$

où $\mathrm{conv}(F)$ est l'enveloppe convexe de F définie par

$$\mathrm{conv}(F) = \Big\{ \sum_{j=1}^{M} \lambda_j f_j : \lambda_j \geqslant 0, \sum_j \lambda_j = 1 \Big\}$$

et $r(n, M)$ est le terme résiduel qu'on souhaite aussi petit que possible. On sera aussi intéressé par des résultats en espérance, càd des inégalités oracle

du type :

$$\mathrm{E}\, R(\hat{f}_n) \leqslant \inf_{f \in \mathrm{conv}(F)} R(f) + r(n, M) \tag{4}$$

où l'espérance E est prise par rapport aux données $\mathcal{D}$. La construction d'estimateur tels que (3) et/ou (4) sont satisfaites avec un terme résidual $r(n, M)$ aussi petit que possible s'appele le probléme d'agrégation convexe. Il existe d'autres problèmes d'agrégation : faire mieux que le meilleur élément dans F, faire mieux que le meilleur élément dans l'espace linéaire engendré par F etc.. Pour ce Projet Individuel, on s'interessera d'abord au problème d'agrégation convexe.

2 Vitesse optimale d'agrégation et minimiseur du risque empirique

Un exemple de méthode d'agrégation est le *minimiseur du risque empirique* défini par :

$$\hat{f}_n^{ERM} \in \operatorname*{argmin}_{f \in \mathrm{conv}(F)} R_n(f) \tag{5}$$

où $R_n(f)$ est le risque empirique de f défini par

$$R_n(f) = \frac{1}{n} \sum_{i=1}^{n} (f(X_i) - Y_i)^2. \tag{6}$$

L'objectif de ce PI est de montrer que cette méthode est optimale pour le problème d'agrégation convexe. On doit d'abord définir ce qui est entendu par optimal. On introduit ici une définition d'optimalité pour ce problème.

Définition. *Soit n (nombre d'observations) et M (nombre d'éléments dans le dictionnaire) deux entiers. On dit que $\hat{f}_n$ est une* ***procédure optimale d'agrégation convexe*** *et que $r(n, M)$ est une* ***vitesse optimale d'agrégation*** *quand il existe deux constantes absolues $c_0 > 0$ et $c_1 > 0$ telles que les deux points suivants sont vérifiés :*

- *Pour tout dictionnaire $F = \{f_1, \ldots, f_M\}$ de cardinal M et tout couple (X, Y) de variables aléatoires tels que $|Y| \leqslant b$ et $|f_j(X)| \leqslant b, \forall j = 1, \ldots, M$ p.s., on a*

$$\mathrm{E}\, R(\hat{f}_n) \leqslant \min_{f \in \mathrm{conv}(F)} R(f) + c_0 r(n, M).$$

- *Pour toute statistique $\tilde{f}_n$, il existe un dictionnaire $F = \{f_1, \ldots, f_M\}$ et un couple (X, Y) de variables aléatoires tels que $|Y| \leqslant b$ et $|f_j(X)| \leqslant$*

$b, \forall j = 1, \dots, M$ *p.s. et*

$$\mathrm{E}\, R(\tilde{f}_n) \geqslant \min_{f \in \mathrm{conv}(F)} R(f) + c_1 r(n, M).$$

On remarque que la vitesse optimale d'agrégation convexe est définie ici à une constante absolue prés. La théorie minimax nous apprend que la vitesse minimale d'agrégation convexe est donnée par

$$\psi_{n,M}^{(C)} = \begin{cases} \frac{M}{n} & \text{quand } M \leqslant \sqrt{n} \\ \sqrt{\frac{1}{n} \log\left(\frac{eM}{\sqrt{n}}\right)} & \text{sinon.} \end{cases} \tag{7}$$

L'objectif de ce PI est de démontrer que le minimiseur du risque empirique défini en (5) atteint cette vitesse. C'est-à-dire que $\hat{f}_n^{ERM}$ vérfie une inégalité oracle comme (4) où le terme résiduel est proportionnel à $\psi_{n,M}^{(C)}$.

Théoréme 1 *Il existe une constante absolue $c_0 > 0$ telle que pour tout $n \geqslant 1$ et $M \geqslant 1$, ce qui suit est vérifié. Pour tout dictionnaire F de cardinal M et tout couple (X, Y) de variables aléatoires tels que $|Y| \leqslant b$ p.s. et $|f(X)| \leqslant b, \forall f \in F$ p.s., on a pour $\hat{f}_n^{ERM}$ le minimiseur du risque empirique sur* $\mathrm{conv}(F)$*,*

$$\mathrm{E}\, R(\hat{f}_n^{ERM}) \leqslant \min_{f \in F} R(f) + c_0 b^2 \psi_{n,M}^{(C)}.$$

On ne prouvera ce résultat que dans le cas (le plus intéressant) $M \geqslant \sqrt{n}$. Pour cela, on aura recours à un résultat sur les processus empiriques qu'on pourra admettre dans une première lecture et à la méthode de Maurey. C'est cette méthode qu'on introduit en premier lieu.

3 La méthode empirique de Maurey

La méthode empirique de Maurey a été introduite pour le calcul de l'entropie de la boule unité B_1^d par rapport à la métrique euclidienne de $\mathbb{R}^d$. On rappelle ici ce calcul.

On commence par quelques notations. Les boules unités pour les normes ℓ_1^d et ℓ_2^d sont

$$B_1^d = \Big\{x \in \mathbb{R}^d : \sum_{j=1}^M |x_j| \leqslant 1\Big\} \text{ et } B_2^d = \Big\{x \in \mathbb{R}^d : \sum_{j=1}^M x_j^2 \leqslant 1\Big\}.$$

Pour tout ensemble $T \subset \mathbb{R}^d$, on note par $N(T, \varepsilon B_2^d)$ le plus petit nombre de translatés de εB_2^d nécessaires pour recouvrir entièrement T. L'entropie de T par rapport à ℓ_2^d est la fonction $\varepsilon \longmapsto \log N(T, \varepsilon B_2^d) := \mathcal{N}(T, \varepsilon, \ell_2^d)$. On va démontrer la proposition suivante (qui est optimale à des constantes absolues prés) par la méthode empirique de Maurey.

Proposition 3.1 *Il existe une constante absolue $c_0 > 0$ telle que pour tout $\varepsilon > 0$,*

$$\log N(B_1^d, \varepsilon B_2^d) \leqslant c_0 \begin{cases} 0 & si\ \varepsilon \geqslant 1, \\ \frac{1}{\varepsilon^2} \log(d\varepsilon^2) & si\ d^{-1/2} \leqslant \varepsilon \leqslant 1, \\ d \log\left(\frac{e}{d\varepsilon^2}\right) & si\ \varepsilon \leqslant d^{-1/2}. \end{cases}$$

Q1.1 Montrer le cas $\varepsilon \geqslant 1$.

Q1.2 Soit $x \in B_1^d$ et $d^{-1/2} \leqslant \varepsilon \leqslant 1$. On veut montrer que x est proche (au sens ℓ_2^d) d'un sous-ensemble Λ de B_1^d dont le logarithme du cardinal est plus petit que $c_0 \varepsilon^{-2} \log(d\varepsilon^2)$. Pour cela, on utilise la méthode empirique de Maurey. On écrit $x = \sum_{j=1}^d x_j e_j$ où $(e_1, \dots, e_M)$ est la base canonique de $\mathbb{R}^d$ et $\sum_j |x_j| \leqslant 1$. On considère une variable aléatoire Z à valeurs dans $\{0, \pm e_1, \dots, \pm e_d\}$ telle que $\mathbb{P}[Z = 0] = 1 - \|x\|_1$ et $\mathbb{P}[Z = \operatorname{sign}(x_i) e_i] = |x_i|$. Montrer que $\mathrm{E}\, Z = x$.

Q1.3 Soit $Z_1, \dots, Z_p$ des variables aléatoires i.i.d. distribuées comme Z. Montrer que

$$\mathrm{E}\, \|x - \frac{1}{m} \sum_{i=1}^m Z_i\|_2^2 = \frac{\mathrm{E}\, \|Z - \mathrm{E}\, Z\|_2^2}{m} \leqslant \frac{4}{m}.$$

Q1.4 En déduire que pour m_ε le plus petit entier m tel que $4/m \leqslant \varepsilon^2$, l'ensemble

$$\Lambda := \Big\{ \frac{1}{m_\varepsilon} \sum_{i=1}^{m_\varepsilon} z_i : z_1, \dots, z_{m_\varepsilon} \in \{0, \pm e_1, \dots, \pm e_d\} \Big\} \tag{8}$$

est un ε-réseau de B_1^d pour ℓ_2^d (c'est-à-dire que pour tout $x \in B_1^d$ il existe $y \in \Lambda$ tel que $\|x - y\|_2 \leqslant \varepsilon$).

Q1.5 Montrer que le cardinal de Λ est tel que

$$\log |\Lambda| \leqslant \frac{c_0}{\varepsilon^2} \log(d\varepsilon^2).$$

En déduire le cas $d^{-1/2} \leqslant \varepsilon \leqslant 1$ de Proposition 3.1.

Q1.6 Montrer que pour tout $\varepsilon, \eta > 0$, on a

$$\log N(B_1^d, \varepsilon B_2^d) \leqslant \log N(B_1^d, \eta B_2^d) + \log N(\eta B_2^d, \varepsilon B_2^d).$$

Q1.7 Par un argument volumique, montrer que

$$N(\eta B_2^d, \varepsilon B_2^d) \leqslant \Big(1 + \frac{2\eta}{\varepsilon}\Big)^d. \tag{9}$$

Q1.8 Déduire le troisième cas de Proposition 3.1 de Q1.7, Q1.6 et du deuxième cas.

4 Un résultat sur les processus empirique

On introduit quelques notations classique en apprentissage statistique. La fonction de perte quadratique d'une fonction $f : \mathcal{X} \longmapsto \mathbb{R}$ est donnée par,

$$\ell_f(x, y) = (y - f(x))^2, \quad \forall x \in \mathcal{X}, y \in \mathbb{R}.$$

Le risque quadratique d'une fonction f s'écrit alors $R(f) = \mathrm{E}\, \ell_f(X, Y)$.

Soit f, g deux fonctions. On note par $[f, g]$ le ségment de f à g.

Q2.1 Montrer que $R(\cdot)$ atteint son minimum sur $[f, g]$.

Soit $f^* \in \operatorname{argmin}_{h \in [f,g]} R(h)$. Pour tout $h \in [f, g]$, on note par

$$\mathcal{L}_h(x, y) = \ell_h(x, y) - \ell_{f^*}(x, y), \quad \forall x \in \mathcal{X}, y \in \mathbb{R}$$

la *fonction de perte en excés* de h. Par ailleurs, on note

$$P\mathcal{L}_h = \mathrm{E}\, \mathcal{L}_h(X, Y) \text{ et } P_n\mathcal{L}_h = \frac{1}{n}\sum_{i=1}^{n} \mathcal{L}_h(X_i, Y_i). \tag{10}$$

On admettra le résultat suivant.

Proposition 4.1 *Il existe une constante absolue $c_0 > 0$ telle que ce quit suit a lieu. Pour tout $x > 0$, avec probabilité plus grande que $1 - 4\exp(-x)$, pour tout $h \in [f, g]$,*

$$\big|P\mathcal{L}_h - P_n\mathcal{L}_h\big| \leqslant \frac{1}{2}\max\big(P\mathcal{L}_h, \frac{c_0 x b^2}{n}\big).$$

5 Preuve du Théorème 1 pour le cas $M \geqslant \sqrt{n}$

On considère l'entier

$$m = \left\lceil \sqrt{\frac{n}{\log\left(eM/\sqrt{n}\right)}} \right\rceil$$

et le sous-ensemble $\mathcal{C}' \subset \mathcal{C} := \mathrm{conv}(F)$ défini par

$$\mathcal{C}' = \Big\{ \frac{1}{m} \sum_{j=1}^{m} h_j : h_1, \ldots, h_m \in F \Big\}$$

où, on rappelle que $F = \{f_1, \ldots, f_M\}$ est le dictionnaire.

Q3.1 Montrer en utilisant la méthode de Maurey que

$$\min_{f \in \mathcal{C}'} R(f) \leqslant \min_{f \in \mathcal{C}} R(f) + \frac{4b^2}{m}.$$

Pour cela, on pourra introduire $f_{\mathcal{C}}^* \in \mathrm{argmin}_{f \in \mathcal{C}} R(f)$.

Q3.2 En utilisant la Proposition 4.1 et une "union bound", prouver que pour $N = |\mathcal{C}'|$ et $\mathcal{C}' = \{g_1, \ldots, g_N\}$ tel que $R(g_1) = \min_{g \in \mathcal{C}'} R(g)$, on a pour tout $x > 0$, avec probabilité au moins $1 - 4\exp(-x)$, pour tout ségment $[g_1, g_j], j = 1, \ldots, N$,

$$\left| P\mathcal{L}_g^{1j} - P_n \mathcal{L}_g^{1j} \right| \leqslant (1/2) \max\left(P\mathcal{L}_g^{1j}, \gamma(x) \right), \quad \forall g \in [g_1, g_j] \tag{11}$$

où $\gamma(x) = c_0 b^2 (x + \log N)/n$ et $\mathcal{L}_g^{1j}$ est la fonction d'excés de risque de g par rapport au ségment $[g_1, g_j]$ (càd si $g_{1j}^* \in \mathrm{argmin}_{g \in [g_1, g_j]} R(h)$ alors $\mathcal{L}_g^{1j} = \ell_g - \ell_{g_{1j}^*}$). On note par $\Omega(x)$ l'événement sur lequel (11) a lieu (pour tout j).

On fixe $X_1, \ldots, X_n$. On écrit $\hat{f}_n^{ERM} = \sum_{j=1}^{M} \beta_j f_j$ et on considère $\Theta : \Omega' \to F$ defini sur un autre espace de probabilité $(\Omega', \mathcal{A}', \mathbb{P}')$ tel que $\mathbb{P}'[\Theta = f_j] = \beta_j, \forall j = 1, \ldots, M$ et on prend m copies i.i.d. $\Theta_1, \ldots, \Theta_m$ de Θ. On note par E'_Θ l'espérance par rapport à $\Theta_1, \ldots, \Theta_m$ et par $\tilde{V}_\Theta$ la variance par rapport à Θ.

Q3.3 Montrer que $\mathrm{E}'_\Theta \, \Theta_j = \hat{f}_n^{ERM}$ pour tout $j = 1, \ldots, m$ et en utilisant la méthode de Maurey que

$$\mathrm{E}'_\Theta \, R\Big(\frac{1}{m} \sum_{j=1}^{m} \Theta_j \Big) = R(\hat{f}_n^{ERM}) + \frac{\mathrm{E}\, \tilde{V}'_\Theta (Y - \Theta(X))}{m}. \tag{12}$$

Montrer que la méthode de Maurey fournit une preuve que, pour le risque empirique, on a aussi

$$\mathrm{E}'_\Theta\, R_n\Big(\frac{1}{m}\sum_{j=1}^m \Theta_j\Big) = R_n(\hat{f}_n^{ERM}) + \frac{1}{m}\Big(\frac{1}{n}\sum_{i=1}^n \tilde{V}'_\Theta(Y_i - \Theta(X_i))\Big). \quad (13)$$

On introduit la notation suivante :

$$g_\Theta = \frac{1}{m}\sum_{j=1}^m \Theta_j \text{ et } i_\Theta \in \{1,\dots,N\} \text{ tel que } g_{i_\Theta} = g_\Theta.$$

On remarque que g_Θ est un point aléatoire prenant ses valeurs dans $\mathcal{C}'$ (en tant que fonction mesurable de Ω' dans $\mathcal{C}'$) et que sur l'événement $\Omega(x)$, on a la propriété d'isomorphie suivante sur le segment $[g_1, g_\Theta]$:

$$\big|P_n\mathcal{L}_g^{1i_\Theta} - P\mathcal{L}_g^{1i_\Theta}\big| \leqslant (1/2)\max\big(P\mathcal{L}_g^{1i_\Theta}, \gamma(x)\big), \quad \forall g \in [g_1, g_{i_\Theta}]. \quad (14)$$

Q3.4 On fixe $\Theta_1,\dots,\Theta_m$. En introduisant le risque de

$$g^*_{1i_\Theta} \in \operatorname*{argmin}_{g\in[g_{i_\Theta},g_1]} R(g),$$

montrer que

$$R(\hat{f}_n^{ERM}) \leqslant \min_{f\in\mathcal{C}} R(f) + \frac{4b^2}{m} + P\mathcal{L}_{g_\Theta}^{1i_\Theta} + R(\hat{f}_n^{ERM}) - R(g_\Theta). \quad (15)$$

Q3.5 Montrer que sur l'événement $\Omega(x)$, on a pour tout $\Theta_1,\dots,\Theta_m$

$$R(\hat{f}_n^{ERM}) \leqslant \min_{f\in\mathcal{C}} R(f) + \frac{4b^2}{m} + \gamma(x) + 2\big(R_n(g_\Theta) - R_n(\hat{f}_n^{ERM})\big) + R(\hat{f}_n^{ERM}) - R(g_\Theta).$$

Q3.6 En prenant l'espérance par rapport à $\Theta_1,\dots,\Theta_m$ (defini sur Ω'), montrer que sur $\Omega(x)$,

$$R(\hat{f}_n^{ERM}) \leqslant \min_{f\in\mathcal{C}} R(f) + \frac{4b^2}{m} + \gamma(x) + 2\,\mathrm{E}'_\Theta\big(R_n(g_\Theta) - R_n(\hat{f}_n^{ERM})\big) + \mathrm{E}'_\Theta\Big(R(\hat{f}_n^{ERM}) - R(g_\Theta)\Big).$$

Q3.7 Démontrer le Théorème 1 dans le cas $M \geqslant \sqrt{n}$. On montrera d'abord un résultat en déviation : pour tout $x > 0$, avec probabilité plus grande que $1 - 4\exp(-x)$,

$$R(\hat{f}_n^{ERM}) \leqslant \min_{f \in \mathrm{conv}(F)} R(f) + c_0 b^2 \max\Big(\sqrt{\frac{1}{n}\log\Big(\frac{eM}{\sqrt{n}}\Big)}, \frac{x}{n}\Big).$$

On concluera par intégration de ce résultat pour obtenir un résultat en espérance.

Références

[1] Guillaume Lecué. Empirical risk minimization is optimal for the convex aggregation problem. Bernoulli journal, 2011.

[2] Alexandre Tsybakov. Optimal rate of aggregation. In *Computational Learning Theory and Kernel Machines (COLT-2003)*, volume 2777 of *Lecture Notes in Artificial Intelligence*, pages 303–313. Springer, Heidelberg, 2003.

(auteur du texte : G. Lecué)