Université Paris Dauphine
Année 2016-2017
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
LE 30 MAI 2017 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 excellente 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 inclura 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 2015-2016 Apprentissage statistique et grande dimension

---

Projet 1 : Estimation d'une fonction de régression

Le problème 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 connait 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 problème 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 $\vartheta_j = \langle f, \varphi_j \rangle = \int_{[0,1]^d} f\varphi_j$ v ?rifiant $\sum_{j=1}^{\infty} \vartheta_j^2 < \infty$. 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 $j^{?\text{me}}$ 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 $\boldsymbol{A}_N = \boldsymbol{\Phi}_N(\boldsymbol{\Phi}_N'\boldsymbol{\Phi}_N)^{-1}\boldsymbol{\Phi}_N'$ est un projecteur orthogonal sur le sous-espace vectoriel de $\mathbb{R}^n$ engendré par les colonnes de la matrice $\boldsymbol{\Phi}_N$.

c) Montrer que lorsque $n \to \infty$, la matrice $\frac{1}{n}\boldsymbol{\Phi}_N'\boldsymbol{\Phi}_N$ converge vers la matrice identité. Vérifier qu'en remplaçant $\boldsymbol{\Phi}_N'\boldsymbol{\Phi}_N$ par l'approximation $nI_{N\times N}$ 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 $\sum_{j=N+1}^{\infty} \vartheta_j^2 + N(\|f\|_\infty^2 + \sigma^2)/n$, o ?

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

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

g) Supposons maintenant que f est bornée par M et l'on connait un entier $k > 0$ et un réel $L > 0$ tels que $\sum_{j=1}^{\infty} j^{2k}\vartheta_j^2 \le L$. Prouver que $\sum_{j>N} \vartheta_j^2 \le LN^{-2k}$ 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 $\{(\varphi_j(X_1), \ldots, \varphi_j(X_n))';\ 1 \le j \le N_0\}$. Montrer que $\hat{\sigma}_{N_0}^2 = \frac{1}{n-N_0}\|(I_{n\times n} - A_{N_0})\boldsymbol{Y}\|^2$ est un estimateur sans biais de σ^2.

2 Simulations

On considère 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 r ?ésultats 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 gaussien $\mathcal{N}(0, I_{n\times n})$.

- ▷ 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).

Université Paris Dauphine Année 2015-2016 Apprentissage statistique et grande dimension

---

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 d ?montrer ce r ?sultat n ?gatif), 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 $\vartheta_1, \dots, \vartheta_n$ sont les coordonnées de ϑ et $\xi_1, \dots, \xi_n$ 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 : $\hat{\vartheta}^B = \mathbb{E}(\vartheta|X) = (\mathbb{E}(\vartheta_1|X), \dots, \mathbb{E}(\vartheta_n|X))$, 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 $\sigma_i^2 > 0$, $i = 1, \dots, n$. 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

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

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 $R(\hat{\vartheta},\vartheta) \geq R(\hat{\vartheta}',\vartheta)$, $\forall \vartheta \in \Theta(Q)$. 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 :

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

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 $\vartheta \in \mathbb{R}^n$,

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

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

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

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 2015-2016 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 $\hat{\beta}_{OLS} = \text{argmin}_\beta \|\mathrm{Y} - \mathrm{X}\beta\|^2$. 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 :

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

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

Remarque : attention la normalisation est différente de celle du cours (cf paragraphe 3.7.1 du cours)

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) ? Indication pour la réciproque : consulter un livre d'optimisation et en particulier la notion de sous gradient.

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.

Indication pour cette partie : il s'agit de l'algorithme glmnet (cf articles sur internet).

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 $k = 1, \dots, K$ on note $I_k = \{1+(k-1)n/K, \dots, kn/K\}$ et $I_{-k} = \{1, \dots, n\} \setminus I_k$. 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 $\lambda > 0, \mu \geqslant 0$

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

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)

Bibliographie :

Regularization and variable selection via the elastic net, Hui Zou and Trevor Hastie (2005)

Regularization Paths for Generalized Linear Models via Coordinate Descent, Jerome Friedman, Trevor Hastie, Rob Tibshirani (2009)

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

$$\|P_j f - f\|_{L^2} \leqslant L2^{-j\alpha}.$$

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

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

o ? $c_{j,k}(f) = \int_{[0,1]} f(t)\varphi_{jk}(t)dt$. 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 ?crire

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

o ? 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.

Projet 5 : La méthode EM "Expectation-Maximization"

responsable : Guillaume Lecué guillaume.lecue@cmap.polytechnique.fr

L'objectif de ce projet est de présenter une méthode d'approximation de l'estimateur du maximum de vraisemblance dans le cas de données à valeurs manquantes et de mélanges.

Les estimateurs du maximum de vraisemblance (EMV) font vraisemblablement partie des méthodes statistiques les plus populaires pour l'analyse de données. Pour mettre en oeuvre une telle méthode, une condition importante est de travailler avec une log-vraisemblance facile à minimiser sur l'espace des paramètres. Cela implique de trés bonnes propriétés de la fonction de log- vraisemblance (comme de la convexité). Cependant, dans de nombreux problèmes d'analyse de données, l'application du principe de maximum de vraisemblance conduit à des problèmes numériques d'une grande complexité. De plus, la fonction de vraisemblance étudiée possède souvent plusieurs extrema. Par conséquent, dans de nombreux cas, les EMV sont calculés grâce à des techniques d'optimisation numérique, statique, dynamique ou mêlant les deux approches.

Un cas remarquable de calcul récursif des EMV est la méthode appelée **Expectation-Maximisation** (EM) (cf. [1] and [2]). Cette approche est privilégiée lorsque la difficulté pour obtenir des EMV provient de la présence de valeurs manquantes (encore appelées variables cachées ou latentes). Si les variables manquantes avaient été observées, l'estimation MV en aurait été largement simplifiée. Dans ce contexte, la méthode EM opère de manière récursive. Chaque récursion consiste en une étape E dans laquelle on calcule l'espérance conditionnelle par rapport aux données inconnues, étant données les variables observées, et une étape M dans laquelle on maximise par rapport aux paramètres. La construction de l'algorithme est telle que l'on peut garantir qu'a chaque itération la valeur de la fonction de vraisemblance augmente. En raison de sa simplicité et de sa robustesse, la méthode EM est très largement employée, et bien qu'elle converge de manière relativement lente, de nouvelles améliorations sont publiées régulièrement dans la littérature spécialisée.

1 Construction de l'algorithme

Le principe de l'algoritme EM repose sur une inégalité portant sur l'espérance conditionnelle de la log-vraisemblance de variables manquantes. Nous présentons maintenant quelques éléments nécessaires à la compréhension de l'algorithme EM.

On se donne un modèle statistique d'estimation de densité à partir de données partiellement obervées. Soit $\{f(\cdot,p) : p \in \mathcal{P}\}$ un ensemble de densité (par rapport à la mesure de Lebesgue) sur $\mathbb{R}^d$ où $\mathcal{P}$ est un ensemble de paramètres. On suppose que $d = d_1 + d_2$. On observe $X \in \mathbb{R}^{d_1}$ et on note par $X^m \in \mathbb{R}^{d_1}$ une valeur *manquante* (non-observée) telle que $X^c = (X, X^m)$ est la valeur *complète*. C'est-à-dire X^c est distribuée selon une certaine densité f_{p^*} pour un certain $p^* \in \mathcal{P}$. On cherche à estimer le paramètre p^* à partir de la donnée obervée X.

Sous p (càd quand X^c est distribuée selon $f(\cdot,p)$), on note par $f_X(\cdot,p)$ une densité marginale de X donnée pour Lebesgue presque tout $x \in \mathbb{R}^{d_1}$ par

$$f_X(x,p) = \int_{\mathbb{R}^{d_2}} f((x,x^m),p)dx^m.$$

De même, on note $f_{X^m}(\cdot,p)$ une densité marginale de X^m sous p. On rappelle aussi que des densités conditionelles de X^m sachant X et X sachant X^m sont données respectivement pour Lebesgue presque tout $x \in \mathbb{R}^{d_1}$ et $x^m \in \mathbb{R}^{d_2}$ par

$$f(x^m|x,p) = \frac{f((x,x^m),p)}{f_X(x,p)} \text{ et } f(x|x^m,p) = \frac{f((x,x^m),p)}{f_{X^m}(x^m,p)}.$$

Question 1.1 *Soit $X^c = (X, X^m)$ une loi normale sur $\mathbb{R}^d$ (où $d = d_1 + d_2$) de moyenne μ et de matrice de covariance Σ données par la décomposition en bloques selon $d = d_1 + d_2$,*

$$\mu = \begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix} \text{ et } \Sigma = \begin{pmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{12}^\top & \Sigma_{22} \end{pmatrix}.$$

On suppose que Σ_{11} est inversible. Montrer que pour le paramètre $p = (\mu, \Sigma)$, la densité conditionalle de X^m sachant $X = x$ (sous p), $x^m \mapsto f(x^m|x,p)$ est la densité d'une Gaussienne sur $\mathbb{R}^{d_2}$ de moyenne $\mu_2 + \Sigma_{12}^\top \Sigma_{11}^{-1}(x - \mu_1)$ et de matrice de covariance $\Sigma_{22} - \Sigma_{12}^\top \Sigma_{11}^{-1}\Sigma_{12}$.

La fonction de log-vraisemblance pour une observation x est donnée par : $p \in \mathcal{P} \mapsto \log f_X(x,p)$. On se fixe une estimation a priori $p^{old} \in \mathcal{P}$ du paramètre p à estimer. On va mettre à jour récursivement le paramètre p^{old}. On rappelle que x est connu et fixé. On peux alors donner une estimation de la densité conditionelle de X^m sachant $X = x$ par $x^m \mapsto f(x^m|x,p^{old})$.

Question 1.2 *On introduit les fonctions (qui dépendent de x)*

$$Q(p,p^{old}) = \mathrm{E}\left[\log f(X^c,p)|X = x, p^{old}\right] = \int_{\mathbb{R}^{d_2}} \log f((x,x^m),p) f(x^m|x,p^{old})dx^m$$

et

$$H(p,p^{old}) = \mathrm{E}\left[\log f(X^m|x,p)|X = x, p^{old}\right] = \int_{\mathbb{R}^{d_2}} \log f(x^m|x,p) f(x^m|x,p^{old})dx^m.$$

Montrer que

$$\log f_X(x,p) = Q(p,p^{old}) - H(p,p^{old})$$

et, à l'aide des propriétés fondamentales de la divergence de Kullback-Leibler :

$$H(p^{old},p^{old}) - H(p,p^{old}) = -\int_{\mathbb{R}^{d_2}} f(x^m|x,p^{old}) \log\Big(\frac{f(x^m|x,p)}{f(x^m|x,p^{old})}\Big)dx^m \geqslant 0.$$

Par conséquent, si on est capable de trouver un paramètre p^{new} tel que $Q(p^{new},p^{old}) > Q(p^{old},p^{old})$ alors

$$\log f_X(x,p^{new}) > \log f_X(x,p^{old})$$

et donc on a réussi à augmenter la vraisemblance. Typiquement, p^{new} sera choisi en maximisant $p \mapsto Q(p,p^{old})$.

En résumant, la construction de l'algorithme EM est une succession d'itérations des deux étapes Expectation et Maximization

Etape E : Calculer $Q(p, p^{old})$.

Etape M : Calculer $p^{new} \in \arg\max Q(p, p^{old})$.

Une réitère en prenant $p^{new} \hookrightarrow p^{old}$. On augmente ainsi à chaque étape la log-vraisemblance du problème. Dans la plupart des cas, cette approche itérative converge vers un maximum global unique et permet ainsi d'obtenir un EMV. Toutefois, les itérations EM peuvent également se terminer sur un maximum local (d'où l'importance de "bien" choisir la valeur d'initialisation p^{old}), voire ne pas converger du tout.

2 Exemple 1 : distribution exponentielle avec données censurées

Les données censurées se rencontrent fréquemment en épidémiologie, et en particulier dans les études de survie (cf. [3]). On les retrouve également pour l'analyse de données issues d'instruments de mesure pour lesquels la gamme de mesure observables est trop limitée.

Ici, on considère un variable aléatoire T de loi exponentielle et de paramètre $a > 0$. On rappelle que T admet une densité (par rapport à la mesure de Lebesgue) donnée par

$$f(t, a) = \begin{cases} 0 & \text{si } t \leqslant 0 \\ a \exp(-at) & \text{si } t > 0. \end{cases}$$

Question 2.1 *Monrer que T est une variable aléatoire "avec perte de mémoire", c'est-à-dire : pour tout $s, t > 0$, $\mathbb{P}[T > s + t | T > t] = \mathbb{P}[T > s]$.*

L'objectif est d'estimer le paramètre a à partir d'un ensemble de N observations, en tenant compte du fait qu'il existe un mécanisme de censure de seuil constant C : si une mesure T est plus grande que C alors on ne connaît pas sa valeur, mais on sait seulement que le seuil a été dépassé. Supposons que les obervations $t_1, \ldots, t_k$ n'ont pas excédé le seuil et que $t_{k+1}, \ldots, t_N$ sont au-dessus du seuil. Les informations disponibles sont donc $t_1, \ldots, t_k$ et $[t_{k+1}, \ldots, t_N \geqslant C]$. L'information complète est constituée par le vecteur $t^c = (t_1, \ldots, t_k, t_{k+1}, \ldots, t_N)^\top$.

Question 2.2 *On se fixe une valeur d'initialisation a^{old} de l'algorithme EM. Montrer que*

$$\begin{aligned} Q(a, a^{old}) &= \mathrm{E}\left[\log f(T^c, a) | t_1, \ldots, t_k, t_{k+1} \geqslant C, \ldots, t_N \geqslant C, a^{old} \right] \\ &= N \log a - a \Big[\sum_{i=1}^{k} t_i + (N-k) \Big(C + \frac{1}{a^{old}} \Big) \Big]. \end{aligned}$$

En déduire que si a^{new} maximise $a \mapsto Q(a, a^{old})$ alors

$$a^{new} = \frac{N}{\sum_{i=1}^{k} t_i + (N-k) \Big(C + \frac{1}{a^{old}} \Big)}.$$

En déduire, que l'algo EM converge vers

$$\hat{a} = \frac{k}{\sum_{i=1}^{k} t_i + (N-k)C}.$$

Montrer que $\hat{a}$ est l'EMV d'une expérience où les données observées sont N variables aléatoires i.i.d. $X_1, \ldots, X_N$ distribuées comme $X = TI(T \leqslant C) + CI(T > C)$ et $k = |\{i : X_i \leqslant C\}|$.

Question 2.3 *Montrer expérimentalement la convergence de l'algo EM pour ces données censurées. Estimer la vitesse de convergence en fonction du nombre d'itérations.*

3 Exemple 2 : Modèle de mélange Gaussien

Etant donné un ensemble de densités (appelées composantes du mélange) sur $\mathbb{R}^d$ noté $\{f_k(\cdot, p_k) : k \in \{1, \ldots, K\}, p_k \in \mathcal{P}_k\}$ où pour tout $k = 1, \ldots, K$, $\mathcal{P}_k$ est un ensemble de paramètres indexant la k-ième composante $f_k(\cdot, p_k)$ du mélange. Un mélange étant une densité de la forme

$$f^{mix}(x, (\alpha_1, \ldots, \alpha_K, p_1, \ldots, p_K)) = \sum_{k=1}^{K} \alpha_k f_k(x, p_k) \tag{1}$$

où $(\alpha_1, \ldots, \alpha_K, p_1, \ldots, p_K)$ est le paramètre de la loi composée. Les poids $\alpha_1, \ldots, \alpha_K$ sont non négatifs et somment à 1, i.e. $\sum_{k=1}^{K} \alpha_k = 1$. La plupart du temps, les densités $f_x(\cdot, p_k)$ sont du même type (par exemple Gaussiennes - on parle alors de mélanges Gaussiens, etc.)

Question 3.1 *Mettre en oeuvre une méthode de simulation d'un mélange Gaussien. Représenter graphiquement un mélange de 10 gaussiennes dans* $\mathbb{R}^2$.

Supposons qu'un échantillon de taille N soit tiré selon la distribution d'un mélange. Le calcul des paramètres $\alpha_1, \ldots, \alpha_K, p_1, \ldots, p_K$ pose typiquement des problèmes numériques. L'algorithme EM propose une approche naturelle à ce problème. En effet, on suppose que l'information complète est donnée par $x^c = k_1, \ldots, k_N, x_1, \ldots, x_N$; en d'autres termes, on fait l'hypothèse que l'on connaît l'indice k_n de la composante $f_{k_n}(\cdot, p_{k_n})$ qui a permis de générer l'observation x_n. Avec cette information complète, la log-vraisemblance s'écrit (pour l'information complète) :

$$\log f(x^c, p) = \sum_{n=1}^{N} \log \alpha_{k_n} + \sum_{n=1}^{N} \log f_{k_n}(x_n, p_{k_n}).$$

On a donc ainsi transformer un problème de calcul de l'EMV dans un modèle de mélange en un problème de données manquantes pour lequel on sait appliquer l'algorithme EM. On initialise le paramètre $p^{old} = (\alpha_1^{old}, \ldots, \alpha_K^{old}, p_1^{old}, \ldots, p_K^{old})^\top$.

Etape E : On note par $x = (x_1, \ldots, x_N)$ l'information observée et par $x^m = (k_1, \ldots, k_N)$ l'information manquante.

Question 3.2 *Montrer que*

$$Q(p, p^{old}) = \sum_{n=1}^{N} \sum_{k=1}^{K} f(k|x_n, p^{old}) \log \alpha_k + \sum_{n=1}^{N} \sum_{k=1}^{K} f(k|x_n, p^{old}) \log f_k(x_n, p_k)$$

où $f(\cdot|x_n, p)$ est la densité conditionnelle de k (l'indice de la densité, vu comme une variable aléatoire à valeurs dans $\{1, \ldots, K\}$) sachant qu'on a obervé x_n sous p.

Etape M : On optimise l'expression $Q(p, p^{old})$ directement par rapport aux poids $\alpha_1, \ldots, \alpha_K$.

Question 3.3 *Montrer que pour tout $k = 1, \ldots, K$,*

$$\alpha_k^{new} = \frac{1}{N} \sum_{n=1}^{N} f(k|x_n, p^{old}).$$

Les itérations ci-dessus pour les poids restent valides quelle que soit la forme des composantes du mélange. Pour l'estimation des paramètres $p_1^{new}, \ldots, p_K^{new}$, on étudie le cas des mélanges Gaussiens sur $\mathbb{R}$. La k-ième composante Gaussienne $f_k(\cdot, (\mu_k, \sigma_k))$ est paramétrée par sa moyenne μ_k et sa variance σ_k^2.

Question 3.4 *Montrer que la loi conditionelle de la donnée manquante $k = k_n$ par rapport à la donnée observée x_n sous p^{old} est donnée par*

$$f(k|x_n, p^{old}) = \frac{(\alpha_k^{old}/\sigma_k^{old}) \exp\big[-(x_n - \mu_k^{old})^2/(2(\sigma_k^{old})^2)\big]}{\sum_{\kappa=1}^K (\alpha_\kappa^{old}/\sigma_\kappa^{old}) \exp\big[-(x_n - \mu_\kappa^{old})^2/(2(\sigma_\kappa^{old})^2)\big]}$$

où $p^{old} = (\alpha_1^{old}, \ldots, \alpha_K^{old}, \mu_1^{old}, \ldots, \mu_K^{old}, \sigma_1^{old}, \ldots, \sigma_K^{old})$. En déduire, que la maximisation par rapport aux μ_k, σ_k de $p \mapsto Q(p, p^{old})$ donne pour tout $k = 1, \ldots, K$,

$$\mu_k^{new} = \frac{\sum_{n=1}^N x_n f(k|x_n, p^{old})}{\sum_{n=1}^N f(k|x_n, p^{old})}$$

et

$$(\sigma_k^{new})^2 = \frac{\sum_{n=1}^N (x_n - \mu_k^{new})^2 f(k|x_n, p^{old})}{\sum_{n=1}^N f(k|x_n, p^{old})}.$$

Implémenter l'algorithme EM pour les mélanges Gaussiens sur $\mathbb{R}$. Représenter graphiquement l'évolution des itérations. Estimer l'erreur d'approximation L_2 entre le vrai mélange et le mélange estimé par l'algo EM

Question 3.5 *Proposer un algorithme de "Clustering" à l'aide d'un modèle de mélange Gaussien et de la méthode EM. Etendre cette méthode sur $\mathbb{R}^2$ et simuler cette méthode de clustering.*

Références

[1] Dempster, A. P., Laird, N. M. and Rubin, D. B. Maximum likelihood from incomplete data via the em algorithm (with discussion). 1977. J. R. Statist. Soc., Ser. B, 39 :1-38. `http://www.aliquote.org/pub/EM.pdf`.

[2] McLachan, G. J. and Krishnan, T. The EM Algorithm and Extensions. 1997. Wiley.

[3] Cox, D. R. and Oakes, D. Analysis of survival data. 1984. Monographs on Statistics and Applied Probability, Chapman & Hall, London.

Université Paris Dauphine Année 2015-2016 Apprentissage statistique et grande dimension

Projet 6 : Estimation dans le modèle linéaire fonctionnel

L'objectif de la statistique pour données fonctionnelles est de traiter des données sous forme de courbes. Contrairement à la statistique classique, les observations ne sont pas des vecteurs de $\mathbb{R}^d$ mais des fonctions. La dimension des observations est donc supposée infinie.

Nous nous intéressons plus particulièrement dans ce projet au modèle linéaire fonctionnel. Nous avons des observations supposées i.i.d. (X_i, Y_i), $i = 1, ..., n$ avec X_i une fonction et Y_i une quantité d'intérêt réelle tels que

$$Y_i = \int_0^1 X_i(t)\beta(t)dt + \xi_i \;\; i = 1, ..., n, \tag{1}$$

avec β une fonction inconnue et ξ_i le bruit, supposé centré, indépendant de X_i, et de variance σ^2.

Remarque : Nous supposerons que X_1 est une variable aléatoire à valeurs dans $\mathbb{L}^2([0,1])$ et que le paramètre $\beta \in \mathbb{L}^2([0,1])$. Notons $\langle \cdot, \cdot \rangle$ le produit scalaire usuel de $\mathbb{L}^2([0,1])$, c'est-à-dire

$$\langle f, g \rangle = \int_0^1 f(t)g(t)dt \text{ pour tout } f, g \in \mathbb{L}^2([0,1]),$$

et $\|f\|^2 = \langle f, f \rangle$ sa norme associée.

Nous pouvons réecrire (1) de la façon suivante :

$$Y_i = \langle f, X_i \rangle + \xi_i.$$

Le modèle ci-dessus est donc l'exact équivalent du modèle linéaire classique pour des observations dans $\mathbb{R}^d$ en remplaçant le produit scalaire usuel de $\mathbb{L}^2([0,1])$ par le produit scalaire usuel de $\mathbb{R}^d$.

Nous considèrerons dans ce projet une application à l'analyse de la concentration en nitrates des eaux usées. Dans ce contexte, nous avons :

— X_i : courbe spectromètrique du i-ème échantillon d'eau ;

— Y_i : concentration en nitrate de l'échantillon.

L'objectif final est de définir plusieurs procédures d'estimation du paramètre β, d'étudier leurs propriétés numériques permettant de prédire correctement la concentration en nitrate à partir de la courbe spectrométrique de cet échantillon.

1 Étude du modèle

Pour simplifier nous supposerons sans perte de généralité que X_1 est centré (sinon il suffit de remplacer partout X_1 par $X_1 - \mathbb{E}[X_1]$ ce qui ne change rien et alourdit les notations).

a) Réécrire le modèle sous une forme vectorielle

$$\mathbf{Y} = \boldsymbol{\beta} + \boldsymbol{\xi},$$

avec $\mathbf{Y}, \boldsymbol{\beta}, \boldsymbol{\xi} \in \mathbb{R}^n$ et $\boldsymbol{\beta}$ et $\boldsymbol{\xi}$ indépendants.

b) Montrer que le modèle (1) implique

$$\mathbb{E}[Y_1 X_1] = \Gamma\beta,$$

avec $\Gamma f \in \mathbb{L}^2([0,1]) \mapsto \mathbb{E}[\langle f, X_1\rangle X_1]$. Γ est un opérateur, c'est-à-dire dans notre cas une application linéaire de $\mathbb{L}^2([0,1])$ dans $\mathbb{L}^2([0,1])$. Quel est l'équivalent de Γ pour un modèle de régression linéaire multivarié ?

c) Nous définissons la semi-norme empirique associée à ce modèle

$$\|f\|_n^2 = \frac{1}{n}\sum_{i=1}^{n}\langle f, X_i\rangle^2,$$

et la semi-norme

$$\|f\|_\Gamma^2 = \langle \Gamma f, f\rangle.$$

Remarquer que $\mathbb{E}[\|f\|_n^2] = \|f\|_\Gamma^2$. Quelle hypothèse doit-on faire pour que $\|\cdot\|_\Gamma$ soit une norme sur $\mathbb{L}^2([0,1])$?

d) (Difficile) On admettra que l'opérateur Γ est compact. Vérifier que Γ est auto-adjoint. Que peut-on dire sur ses valeurs propres et ses fonctions propres ?

e) (Difficile) Nous supposons que X_1 est périodique ($X_1(0) = X_1(1)$ p.s.), continue p.s. et stationnaire au second ordre (i.e. il existe une fonction $C : [-1,1] \to \mathbb{R}$ telle que $\mathbb{E}[X_1(t_1)X_2(t_2)] = C(t_1 - t_2)$). Montrer que la base trigonométrique définie par $\varphi_1 \equiv 1$, et, pour tout $j \geqslant 1$, $\varphi_{2j}(x) = \sqrt{2}\cos(2\pi jx)$ et $\varphi_{2j+1}(x) = \sqrt{2}\sin(2\pi jx)$ est une base de fonctions propres de Γ.

2 Estimation par projection

Soit $(\varphi_j)_{j\geqslant 1}$ la base trigonométrique $\mathbb{L}^2([0,1])$. Nous cherchons des estimateurs de β sous la forme

$$\widehat{\beta}_M(t) = \sum_{j=1}^{M}\widehat{\vartheta}_j\varphi_j(t).$$

a) Définir un estimateur des moindres carrés (EMC) dans ce modèle.

b) Nous allons étudier, de manière théorique, le risque

$$R(\widehat{\beta}_M^{(EMC)}, \beta) := \mathbb{E}[\|\widehat{\beta}_M^{(EMC)} - \beta\|_n^2]$$

de l'estimateur des moindres carrés.

(a) Soient, pour tout $f \in \mathbb{L}^2([0,1])$,

$$\gamma_n(f) = \frac{1}{n}\sum_{i=1}^{n}(Y_i - \langle f, X_i\rangle)^2 \text{ et } \nu_n(f) = \sum_{i=1}^{n}\xi_i\langle f, X_i\rangle.$$

Montrer que, pour

$$\beta_M = \sum_{i=1}^{M}\vartheta_j\varphi_j, \text{ avec } \vartheta_j = \langle\beta, \varphi_j\rangle$$

nous avons

$$\gamma_n(\beta_M) - \gamma_n(\widehat{\beta}_M^{(EMC)}) = \|\beta_M - \beta\|_n^2 - \|\widehat{\beta}_M^{(EMC)} - \beta\|_n^2 + 2\nu_n(\beta_M - \widehat{\beta}_M^{(EMC)}).$$

(b) En utilisant le fait que, pour tous réels x, y, pour tout $\delta > 0$, $2xy \leqslant \delta x^2 + \delta^{-1} y^2$, montrer que, pour tout $\delta > 0$,

$$\nu_n(\beta_M - \widehat{\beta}_M^{(EMC)}) \leqslant \delta \|\beta_M - \widehat{\beta}_M^{(EMC)}\|_n^2 + \delta^{-1} \sup_{t \in B_M} \nu_n^2(t),$$

avec $B_M^{(n)} = \{t \in \mathrm{Vect}\{\varphi_1, ..., \varphi_M\}, \|t\|_n^2 \leqslant 1\}$.

(c) Montrer que

$$\mathbb{E}\left[\sup_{t \in B_{M,1/2}} \nu_n^2(t)\right] \leqslant \frac{\sigma^2}{2n} \sum_{j=1}^{M} \langle \Gamma\varphi_j, \varphi_j \rangle$$

avec

$$B_{M,1/2} := \{t \in \mathrm{Vect}\{\varphi_1, ..., \varphi_M\}, \|t\|^2 \leqslant 1/2\}.$$

(d) En remarquant que $\gamma_n(\widehat{\beta}_M^{(EMC)}) \leqslant \gamma_n(\beta_M)$, déduire que, pour tout $\delta \in]0, 1/2[$,

$$(1 - 2\delta)\mathbb{E}\left[\|\widehat{\beta}_M^{(EMC)} - \beta\|_n^2 \mathbf{1}_{\left\{B_M^{(n)} \subset B_{M,1/2}\right\}}\right] \leqslant (1 + 2\delta)\mathbb{E}[\|\beta_M - \beta\|_n^2] + \delta^{-1} \frac{\sigma^2}{2n} \sum_{j=1}^{M} \langle \Gamma\varphi_j, \varphi_j \rangle.$$

(e) On admettra que $\mathbb{P}\left(B_M^{(n)} \subset B_{M,1/2}\right) \geqslant 1 - M^2 \exp(-n/M^2)$. Montrer qu'il existe une constante $C > 0$ vérifiant, pour tout $M \leqslant \sqrt{n/\log^3(n)}$,

$$R(\widetilde{\beta}_M^{(EMC)}, \beta) \leqslant C\left(R(\beta_M, \beta) + \frac{\sigma^2}{n} \sum_{j=1}^{M} \langle \Gamma\varphi_j, \varphi_j \rangle + \frac{1 + \|\beta\|_\Gamma^2}{n}\right),$$

avec

$$\widetilde{\beta}_M^{(EMC)} = \widehat{\beta}_M^{(EMC)} \mathbf{1}_{\{\|\widehat{\beta}_M^{(EMC)}\|_n^2 \leqslant \exp(n \log^2(n))\}}.$$

c) On suppose que la base trigonométrique $(\varphi_j)_{j \geqslant 1}$ est une base de fonctions propres pour l'opérateur Γ, c'est-à-dire que, pour tout $j \geqslant 1$, il existe λ_j tel que $\Gamma\varphi_j = \lambda_j \varphi_j$ et qu'il existe une constante $a > 0$ et deux constantes $C, c > 0$ telles que les valeurs propres $(\lambda_j)_{j \geqslant 1}$ de l'opérateur Γ vérifient

$$cj^{-a} \leqslant \lambda_j \leqslant Cj^{-a}, \text{ pour tout } j \geqslant 1.$$

On suppose de plus que β est dans une boule de Sobolev de régularité α.
Montrer qu'il existe un choix de M pour lequel

$$R(\widetilde{\beta}_M^{(EMC)}, \beta) \leqslant C' n^{-(2a+2\alpha)/(2a+2\alpha+1)}.$$

3 ACP fonctionnelle

En règle générale, les fonctions propres de l'opérateur Γ sont inconnues. Nous les estimons donc en prenant la base des fonctions propres de l'opérateur Γ en prenant $(\widehat{\varphi}_j)_{j\geqslant 1}$ de l'opérateur Γ_n définit par

$$\Gamma_n f(t) = \frac{1}{n}\sum_{i=1}^{n}\langle f, X_i\rangle X_i(t).$$

Nous étudions des estimateurs du type

$$\widehat{\beta}^{(PCA)}(t) = \sum_{j=1}^{M}\widehat{\vartheta}^{(PCA)}\widehat{\varphi}(t).$$

a) Étudier brièvement l'opérateur Γ_n et justifier l'existence d'une base $(\widehat{\varphi}_j)_{j\geqslant 1}$ de fonctions propres de Γ_n.

b) Montrer que pour l'estimateur des moindres carrés

$$\widehat{\vartheta}^{(PCA,MC)} = \langle \mathbb{E}[YX], \widehat{\varphi}_j\rangle/\widehat{\lambda}_j$$

où $\widehat{\lambda}_j$ est la valeur propre de Γ_n associée à $\widehat{\varphi}_j$.

4 Régression ridge

On peut définir également un estimateur de type ridge en minimisant le critère suivant :

$$\widehat{\beta}^{(R)} \in \arg\min_{\beta\in\mathbb{L}^2([0,1])}\left\{\frac{1}{n}\sum_{i=1}^{n}(Y_i - \langle\beta, X_i\rangle)^2 + \rho\|\beta\|^2\right\}.$$

a) Donner une expression de $\widehat{\beta}^{(R)}$ en fonction de $(\Gamma - \rho I)^{-1}$.

b) En pratique, on cherche souvent $\widehat{\beta}^{(R)}$ dans un espace de dimension finie comme dans la Section 2 mais en choisissant M grand. Réecrire le problème de minimisation dans ce cas-là et donner une forme explicite pour l'estimateur.

5 Étude numérique

L'objectif de cette section est de comparer plusieurs estimateurs de la fonction β sur des données simulées. Des suggestions sont données pour simuler l'échantillon et proposer une procédure d'estimation.

Pour la simulation de l'échantillon, vous pouvez prendre par exemple :

— X un mouvement brownien sur $[0, 1]$,

— X un pont brownien $X(t) = W(t) - tW(1)$ avec W un mouvement brownien.

— $X(t) = a_i + b_i t + c_i e^t + \sin(d_i t)$ avec $a_i \sim \mathcal{U}([0, 100])$, $b_i \sim \mathcal{U}([-30, 30])$, $c_i \sim \mathcal{U}([0, 10])$ et $d_i \sim \mathcal{U}([1, 3])$.

— ...

Vous pouvez considérer différentes lois pour le bruit, à condition d'avoir toujours un bruit centré.

Concernant la procédure d'estimation, différents choix de base sont possibles :

— bases d'histogrammes,
— base trigonométrique,
— base des fonctions propres de Γ (lorsqu'elle est connue en théorie) ou de Γ_n (pour son estimation on pourra utiliser la fonction *pca.fd* du package *fda* [2]).

6 Application

L'objectif est de proposer une méthode satisfaisante pour mesurer la concentration en nitrite d'un échantillon d'eau. Pour cela, nous disposons de 49 mesures sur des échantillons dont la concentration en nitrite est donnée ou déterminée par une autre méthode plus coûteuse. L'objectif est d'apprendre la relation entre la courbe spectrométrique X d'un échantillon d'eau et la concentration en nitrites Y de façon à pouvoir déterminer automatiquement la concentration en nitrite d'un nouvel échantillon d'eau à partir de sa courbe spectrométrique.

Vous séparerez dans un premier temps l'échantillon en deux parties (échantillon d'apprentissage et échantillon de test) et vous comparerez différents estimateurs de la fonction β. Finalement, vous proposerez en la justifiant une seule méthode pour répondre au problème proposé.

Références

[1] H.N. Pham et al. Estimation simultanée et en ligne de nitrates et nitrites par identification spectrale UV en traitement des eaux usées, *L'eau, l'industrie, les nuisances*, n. 335, pp. 61-69, 2010.

[2] J. O. Ramsay, H. Wickham, S. Graves, and G. Hooker. *fda : Functional Data Analysis*, 2014. URL `http://CRAN.R-project.org/package=fda`. R package version 2.4.4.

(auteur du texte : A. Roche)