Université Paris Dauphine
Année 2015-2016
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 10 JUIN 2016 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 ?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).