Estimation non-paramétrique robuste pour données fonctionnelles

Christophe Crambes, Laurent Delsol, Ali Laksaci

Estimation nonparamétrique robuste pour données fonctionnelles

Christophe Crambes[1], Laurent Delsol[2] & Ali Laksaci[3]

[1] *Université Montpellier 2, place Eugène Bataillon, 34095 Montpellier cedex.*
[2] *Université Catholique de Louvain, 20 voie du Roman Pays, 1348 Louvain-la-Neuve.*
[3] *Université Djillali Liabès, BP 89, 22000 Sidi Bel Abbès, Algeria.*

Résumé. L'estimation robuste présente une approche alternative aux méthodes de régression classiques, par exemple lorsque les observations sont affectées par la présence de données aberrantes. Récemment, ces estimateurs robustes ont été considérés pour des modèles avec données fonctionnelles. Dans cet exposé, nous considérons un modèle de régression robuste avec une variable d'intérêt réelle et une variable explicative fonctionnelle. Nous définissons un estimateur non-paramétrique de la fonction de régression, et nous nous intéressons à ses propriétés asymptotiques, relativement à des erreurs $\mathbb{L}^q$. Notre procédure d'estimation est ensuite évaluée sur un jeu de données réelles.

Abstract. It is well known that robust estimation provides an alternative approach to classical methods affected for instance by the presence of outliers. Recently, these robust estimators have been considered for models with functional data. In this talk, we focus on asymptotic properties of a conditional nonparametric estimation of a real valued variable with a functional covariate. We present results dealing with convergence in probability, asymptotic normality and $\mathbb{L}^q$ errors. Our estimation procedure is also evaluated on a real dataset.

Mots clés. Données fonctionnelles, statistique robuste, estimation non-paramétrique, erreurs $\mathbb{L}^q$.

1. Introduction

Un problème courant en statistiques est d'essayer d'expliquer comment une variable d'intérêt Y est reliée à une variable explicative X. Cet exposé se rapporte à ce cadre, dans lequel on suppose de plus que la variable à expliquer Y est à valeurs réelles et la variable explicative X est à valeurs dans un espace de fonctions $\mathcal{F}$ muni d'une semi-métrique d. Ce type de variables, connu sous le nom de variables fonctionnelles dans la littérature, permet de considérer des variables aléatoires en tant que fonctions (du temps par exemple), ce qui semble être le plus adapté dans le cas où les observations sont par nature fonctionnelles: on renvoie notamment à Ramsay et Silverman (2002, 2005) pour une vue d'ensemble sur les données fonctionnelles. Dans ce contexte, le modèle le plus général est le modèle de régression lorsque la variable explicative est une courbe, qui s'écrit

$$Y = r(X) + \epsilon,$$

où r est un opérateur de $\mathcal{F}$ dans $\mathbb{R}$ et ϵ est une variable aléatoire d'erreur. Ce modèle a déjà été étudié d'un point de vue non-paramétrique (c'est-à-dire lorsque l'on fait uniquement des hypothèses de régularité sur r). La monographie de Ferraty et Vieu (2006) recense les principaux résultats obtenus pour l'estimateur non-paramétrique à noyau de r. Cependant, cette estimation de r vu comme la moyenne conditionnelle de Y sachant X peut être inadaptée dans certaines situations. Par exemple, la présence de données aberrantes peut amener à des résultats non pertinents. La régression robuste a été introduite pour résoudre ce genre de problèmes. Depuis les premiers résultats obtenus dans les années soixante, notamment par Huber (1964), de nombreux auteurs ont développé le domaine: Robinson (1984), Collomb et Härdle (1986), Boente et Fraiman (1990), et Laïb ! et Ould-Saïd (2000), ... Concernant les données en dimension infinie, la littérature est beaucoup plus restreinte: Cadre (2001), ainsi que Cardot *et al.* (2005) sont les principales références. Récemment, Azzedine *et al.* (2008) ont étudié la convergence presque complète d'estimateurs robustes à noyau. Dans le même cadre, Attouch *et al.* (2007) ont étudié la normalité asymptotique de ces estimateurs.

Dans ce travail, on se propose de poursuivre l'étude de ces estimateurs à noyau. Après avoir rappelé le résultat de convergence en probabilité ainsi que la normalité asymptotique de Attouch *et al.* (2007), on donne les expressions asymptotiques de la vitesse de convergence vis-à-vis des normes $\mathbb{L}^q$, étendant ainsi les résultats de Delsol (2007) obtenus dans un cadre non robuste. En guise d'illustration, on applique nos résultats en prévision de séries temporelles sur un jeu de données réelles de consommation d'énergie.

2. Modèle

Considérons un couple (X, Y) de variables aléatoires à valeurs dans $\mathcal{F} \times \mathbb{R}$. Pour $x \in \mathcal{F}$, on considère une fonction mesurable ψ_x. Le paramètre fonctionnel étudié dans ce travail, noté θ_x, est la solution (supposée unique) de l'équation en t définie par

$$\Psi(x, t) := \mathbb{E}\left[\psi_x\left(Y, t\right) | X = x\right] = 0. \tag{1}$$

En général, la fonction ψ_x est fixée par le statisticien en fonction de la situation à laquelle il est confronté. Des exemples classiques de ψ_x conduisent à l'estimation de la moyenne conditionnelle (si $\psi_x(Y, t) = Y - t$) ou de quantiles conditionnels (si $\psi_x(Y, t) = \mathbb{1}_{\{Y \geq t\}} - \mathbb{1}_{\{Y < t\}}$): voir Ferraty et Vieu (2006) et Attouch *et al.* (2007). Étant donné un échantillon $(X_i, Y_i)_{i=1,\ldots,n}$ de même loi que (X, Y), un estimateur à noyau de $\Psi(x, t)$ est donné par

$$\widehat{\Psi}(x, t) = \frac{\sum_{i=1}^{n} K\left(h^{-1} d(x, X_i)\right) \psi_x\left(Y_i, t\right)}{\sum_{i=1}^{n} K\left(h^{-1} d(x, X_i)\right)}, \tag{2}$$

où K est un noyau et $h = h_n$ est la largeur de fenêtre. Alors, un estimateur à noyau naturel de θ_x est $\widehat{\theta}_n = \widehat{\theta}_n(x)$ donné par

$$\widehat{\Psi}(x, \widehat{\theta}_n) = 0. \tag{3}$$

On remarque que, sous la condition $\sum_{i=1}^{n} K\left(h^{-1} d(x, X_i)\right) \neq 0$, la définition de l'estimateur par (3) est équivalente à

$$\widehat{\rho}_n(x, \widehat{\theta}_n) := \sum_{i=1}^{n} K\left(h^{-1} d(x, X_i)\right) \psi_x\left(Y_i, \widehat{\theta}_n\right) = 0. \tag{4}$$

3. Résultats asymptotiques

3.1. Convergence en probabilité et normalité asymptotique

On donne ici certains résultats obtenus par Attouch *et al.* (2007) pour des données $(X_i, Y_i)_{i=1,\ldots,n}$ indépendantes et identiquement distribuées. On pose $F(h) = \mathbb{P}\left(d(X, x) \leq h\right)$, connu sous le nom de probabilités de petites boules dans la littérature. Sous des conditions techniques (non citées ici) mais relativement classiques dans ce contexte non-paramétrique fonctionnel, ces auteurs montrent

$$\widehat{\theta}_n - \theta_x \xrightarrow[n \to +\infty]{\mathbb{P}} 0.$$

et

$$\left(\frac{nF(h_n)}{V_n(x)}\right)^{1/2} \left(\widehat{\theta}_n - \theta_x - B_n(x)\right) \xrightarrow[n \to +\infty]{\mathcal{L}} \mathcal{N}(0, 1),$$

avec des expressions explicites pour $V_n(x)$ et $B_n(x)$.

3.2. Un résultat d'uniforme intégrabilité

On donne dans ce paragraphe un résultat utile dans la suite pour obtenir la convergence des moments de $\widehat{\theta}_n$. Soit $t \in \mathbb{R}$ fixé. On donne le résultat pour des données $(X_i, Y_i)_{i=1,\ldots,n}$ indépendantes et identiquement distribuées. On peut également montrer, sous des hypothèses plus fortes, le même type de résultat pour des données arithmétiquement α-mélangeantes. On considère les hypothèses suivantes.

($H.1$) Il existe $p > 2$ et $C > 0$, tels que, pour X dans un voisinage ouvert de x, on a presque sûrement

$$\mathbb{E}\left[|\psi_x(Y, t)|^p \,|X\right] \leq C.$$

($H.2$) On suppose que $\lim_{n \to +\infty} nF(h_n) = +\infty$.

($H.3$) K est à support $[0,1]$, est borné, et $K(1) > 0$.

Sous les hypothèses $(H.1) - (H.3)$, pour $0 \leq q < p$, la quantité

$$\left| \sqrt{nF(h_n)} \left(\Psi_n(x,t) - \mathbb{E}\left[\Psi_n(x,t) \right] \right) \right|^q,$$

est uniformément intégrable, où $\Psi_n(x,t) = \dfrac{1}{nF(h_n)} \widehat{\rho}_n(x,t)$.

3.3. Convergence des moments

On donne le résultat pour des données $(X_i, Y_i)_{i=1,\ldots,n}$ indépendantes et identiquement distribuées. On peut également montrer, sous des hypothèses plus fortes, le même type de résultat pour des données arithmétiquement α-mélangeantes. On suppose que ψ_x est de classe $\mathcal{C}^1$ vis-à-vis de son second argument sur un voisinage de θ_x. On note ζ_n la variable aléatoire (entre θ_x and $\hat{\theta}_n$) telle que $\hat{\theta}_n - \theta_x = -\frac{\Psi_n(x,\theta_x)}{\frac{\partial \Psi_n}{\partial t}(x,\zeta_n)}$ et on définit $B_n := -\frac{\mathbb{E}[\Psi_n(x,\theta_x)]}{\mathbb{E}\left[\frac{\partial \Psi_n}{\partial t}(x,\zeta_n)\right]}$. On suppose que

$$Z_n := \sqrt{\frac{nF\left(h_n\right)}{V_n(x)}} \left(\widehat{\theta}_n - \theta_x - B_n(x) \right) \xrightarrow[n \to +\infty]{\mathcal{L}} W, \tag{5}$$

où W est une variable aléatoire gaussienne centrée réduite, et avec des expressions explicites de $B_n(x)$ et $V_n(x)$. On suppose que les hypothèses $(H.1) - (H.3)$ sont vérifiées (avec $t = \theta_x$), ainsi que certaines conditions techniques données ci-dessous.

($H.4$) $t \mapsto \sup_y \left(\dfrac{\partial \psi_x}{\partial t}(y,t) - \dfrac{\partial \psi_x}{\partial t}(y,\theta_x) \right)$ est continue dans un voisinage de θ_x.

($H.5$) Il existe une constante N telle que, presque sûrement dans un voisinage de x

$$\mathbb{E}\left[\left(\frac{\partial \psi_x}{\partial t}(Y_i, \zeta_n) - \frac{\partial \psi_x}{\partial t}(Y_i, \theta_x) \right)^2 \mid X_i \right] \leq N.$$

($H.6$) Il existe des constantes γ et δ telles que

$$\mathbb{E}\left[\frac{\partial \psi_x}{\partial t}(Y, \theta_x) \mid X \right] \mathbb{1}_{\{d(X,x) \leq \delta\}} \geq \gamma \mathbb{1}_{\{d(X,x) \leq \delta\}}.$$

($H.7$) $B_n(x)$ vérifie $\sqrt{nF\left(h_n\right)} B_n = O\left(1\right)$.

($H.8$) Il existe $p' > 2$ et une constante $0 < C' < +\infty$ telle que, pour X dans un voisinage ouvert de x, on a presque sûrement

$$\mathbb{E}\left[\left| \frac{\partial \psi_x}{\partial t}(Y, \zeta_n) \right|^{p'} | X \right] \leq C'.$$

($H.9$) Il existe r et une constante $0 < M_0 < +\infty$ tels que $\mathbb{E}\left[\left|\widehat{\theta}_n - \theta_x\right|^r\right] \leq M_0$.

Alors, pour $q < q'$ (avec une expression explicite de q', qui n'est pas donnée ici), on a

$$\mathbb{E}\left[\left|\widehat{\theta}_n - \theta_x\right|^q\right] = \mathbb{E}\left[\left|B_n(x) + \sqrt{\frac{V_n(x)}{nF(h_n)}}W\right|^q\right] + o\left(\frac{1}{\sqrt{nF(h_n)}^q}\right).$$

Des expressions asymptotiques plus explicites des erreurs $\mathbb{L}^q$ peuvent être déduites à partir des expressions de $B_n(x)$ et $V_n(x)$ données par Attouch *et al.* (2007) en utilisant la même approche que Delsol (2007). Ces expressions peuvent être utilisées par exemple pour choisir la fenêtre optimale h. Ce sont les premiers résultats de convergence dans $\mathbb{L}^q$ pour des estimateurs robustes dans un modèle de régression non-paramétrique fonctionnel.

4. Application

Dans cet exemple, on s'intéresse à l'application de notre modèle comme outil de prévision. On utilise les données de consommation d'énergie[1] étudiées dans la monographie de Ferraty et Vieu (2006). L'objectif est de prévoir la consommation une année connaissant la courbe l'année précédente. Pour éviter l'hétéroscédasticité de ces données, Ferraty et Vieu (2006) ont transformé ces données à l'aide de différences logarithmiques. On souhaite utiliser ici les données initiales, notre estimateur robuste ne devant pas être affecté par l'hétéroscédasticité. On considère la fonction objectif $\psi_x(\cdot) = \psi\left(\frac{\cdot}{S(x)}\right)$ où $\psi(\cdot)$ est la fonction de Huber définie par

$$\psi(u) = \begin{cases} u & \text{si } u \in [-1.34, 1.34], \\ 1.34 sgn(u) & \text{sinon,} \end{cases}$$

et la quantité $S(x)$ est la médiane de $|Y - med_x|$, avec med_x la médiane conditionnelle de Y sachant $X = x$. Le paramètre de lissage est choisi en utilisant la validation croisée de type $\mathbb{L}^1$. Le noyau utilisé est le noyau quadratique. Un autre paramètre à fixer et dont l'importance est cruciale est la semi-métrique d. On utilise une famille de semi-métriques induites par l'analyse en composantes principales fonctionnelle avec plusieurs dimensions: voir Besse *et al.* (1997) pour plus de détails. Les résultats que nous avons obtenus dans cet exemple sont relativement satisfaisants en termes de prédiction.

Bibliographie

[1] Attouch, M., Laksaci, A. and Ould-Saïd, E. (2007). Asymptotic distribution of robust estimator for functional nonparametric models. Prépublication, LMPA No 314.

[2] Azzeddine, N., Laksaci, A. and Ould-Saïd, E. (2008) On the robust nonparametric regression estimation for functional regressor. *Statistic and Probabability Letters*, **78**, 3216-3221.

[1] données disponibles à l'adresse www.economagic.com

[3] Besse, P., Cardot, H. and Ferraty, F. (1997). Simultaneous nonparametric regressions of unbalanced longitudinal data. *Computational Statistics and Data Analysis*, **24**,255-270.

[4] Boente, G. and Fraiman, R. (1990). Asymptotic distribution of robust estimators for nonparametric models from mixing processes. *Annals of Statistics*, **18**, 891-906.

[5] Cadre, B. (2001). Convergent estimators for the L_1-median of a Banach valued random variable. *Statistics*, **35**, 509-521.

[6] Cardot, H., Crambes, C. and Sarda, P. (2005). Quantiles regression when the covariates are functions. *Journal of Nonparametric Statistics*, **17**, 841-856.

[7] Collomb, G. and Härdle, W. (1986). Strong uniform convergence rates in robust nonparametric time series analysis and prediction: kernel regression estimation from dependent observations. *Stochastic Processes and Applications*, **23**, 77-89.

[8] Delsol, L. (2007). Régression non-paramétrique fonctionnelle: expressions asymptotiques des moments. *Annales de l'ISUP*, **LI**, 43-67.

[9] Ferraty, F. and Vieu, P. (2006). *Nonparametric functional data analysis: theory and practice.* Springer-Verlag, New York.

[10] Huber, P.J. (1964). Robust estimation of a location parameter. *Annals of the Mathematical Statistics*, **35**, 73-101.

[11] Laïb, N. and Ould-Sïad, E. (2000). A robust nonparametric estimation of the autoregression function under an ergodic hypothesis. *Canadian Journal of Statistics*, **28**, 817-828.

[12] Robinson, R. (1984). *Robust Nonparametric Autoregression.* Lecture Notes in Statistics, Springer-Verlag, New York, **26**, 247-255.

[13] Ramsay, J.O. and Silverman, B.W. (2002). *Applied functional data analysis.* Springer-Verlag, New York.

[14] Ramsay, J.O. and Silverman, B.W. (2005). *Functional data analysis* (Sec. Ed.). Springer-Verlag, New York.