C. R. Acad. Sci. Paris, Ser. I 348 (2010) 1115–1118

Contents lists available at ScienceDirect

C. R. Acad. Sci. Paris, Ser. I

www.sciencedirect.com

Statistique

Choix optimal du paramètre de lissage pour les U-statistiques conditionnelles

Optimal bandwidth selection for conditional U-statistics

Lynda Arezki, Djamal Louani

LSTA, université Pierre-et-Marie-Curie Paris 6, 175, rue du Chevaleret, 75013 Paris, France

I N F O A R T I C L E

Historique de l'article :
Reçu le 12 mai 2010
Accepté après révision le 31 août 2010
Disponible sur Internet le 14 octobre 2010

Présenté par Paul Deheuvels

R É S U M É

Dans cette Note, nous proposons une procédure basée sur la méthode de validation croisée pour choisir le paramètre de lissage pour les U-statistiques conditionnelles. Nous montrons, par ailleurs, que le paramètre sélectionné est asymptotiquement optimal pour divers critères quadratiques. Notons que pour un choix approprié du noyau (φ) de la U-statistique, nos résultats permettent de déduire le paramètre de lissage optimal pour différents estimateurs non paramétriques usuels.

A B S T R A C T

This Note proposes a procedure based on the cross-validation method to select the smoothing parameter of the conditional U-statistics. We state here that the obtained data-driven bandwidths are asymptotically optimal with respect to various criteria. Notice that by suitable choices of the U-statistic kernel, say φ, our results allow to deduce in a straightforward way the optimal smoothing parameter of various usual nonparametric estimates.

1. Introduction

Soit $(X_1, Y_1), \ldots, (X_n, Y_n)$ un n-échantillon i.i.d. d'un vecteur aléatoire (X, Y) à valeurs dans $\mathbb{R}^2$ de densité conjointe g. Soit f la densité marginale de X. Nous nous intéressons ici au problème d'estimation de la fonction de régression suivante :

$$m(\mathbf{x}) := \mathbb{E}\big(\varphi(Y_1, \ldots, Y_k) | (X_1, \ldots, X_k) = \mathbf{x}\big), \quad \mathbf{x} = (x_1, \ldots, x_k) \in \mathbb{R}^k,$$

où, pour $k \leqslant n$, $\varphi : \mathbb{R}^k \to \mathbb{R}$, est une fonction mesurable. Pour estimer la fonction $m(\mathbf{x})$, Stute [9] a introduit une classe générale de U-statistiques conditionnelles de la forme :

$$U_n(x_1, \ldots, x_k) = U_n^h(\mathbf{x}) = \frac{\sum_{I_n^k} \varphi(Y_{i_1}, \ldots, Y_{i_k}) \prod_{j=1}^k K((x_j - X_{i_j})/h)}{\sum_{I_n^k} \prod_{j=1}^k K((x_j - X_{i_j})/h)}, \tag{1}$$

où K est un noyau réel et h le paramètre de lissage tendant vers zéro avec une vitesse appropriée. Dans l'énoncé (1), les sommations sont considérées sur toutes les permutations $\mathbf{i} = (i_1, \ldots, i_k)$ de longueur k telles que $1 \leqslant i_j \leqslant n$, $j = 1, \ldots, k$.

Adresses e-mail : lynda.arezki@etu.upmc.fr (L. Arezki), djamal.louani@upmc.fr (D. Louani).

1631-073X/$ – see front matter
doi:10.1016/j.crma.2010.08.007

Comme pour les estimateurs non paramétriques usuels, l'utilisation de U_n nécessite le choix du paramètre de lissage h, ce dernier revêt une importance capitale pour sa performance. De nombreuses méthodes de sélection automatique du paramètre h ont été proposées dans la littérature pour l'estimation de la fonction de densité et de la fonction de régression, parmi lesquelles nous citons la méthode de validation croisée et la méthode d'injection (plug-in). Nous renvoyons aux travaux de Sarda et Vieu [6], Wand et Jones [10] et Hart [3] pour un aperçu de quelques résultats sur ce sujet. Toutefois, il reste à ce jour divers problèmes d'estimation où cette question de choix du paramètre de lissage n'a pas été étudiée et c'est le cas des U-statistiques conditionnelles. Pour cette classe d'estimateurs, nous proposons dans cette Note, une procédure empirique, basée sur la méthode de validation croisée globale, permettant ainsi de sélectionner une fenêtre de lissage qui soit asymptotiquement optimale au sens de l'erreur quadratique. Cette procédure peut être considérée comme une généralisation des résultats obtenus par Härdle et Marron [2] et par Marron et Härdle [4]. Rappelons que la consistance de la U-statistique conditionnelle a été établie par Sen [7] et que le livre de Serfling [8] comporte de nombreux éléments relatifs à la U-statistique.

Cette Note est organisée comme suit, la Section 2 est consacrée à la définition et à la justification du critère de sélection. L'optimalité asymptotique de la fenêtre sélectionnée est donnée dans la Section 3, qui contient également des résultats d'équivalences asymptotiques entre différentes mesures de risques quadratiques pour les U-statistiques conditionnelles.

2. Le critère de validation croisée

En s'inspirant des travaux de Härdle et Marron [2] et de Rachdi et Vieu [5], la règle de sélection proposée ici est l'une des plus populaires tant sur le plan pratique que du point de vue des études théoriques. Suivant cette méthodologie, soit h_0 le paramètre de lissage théorique minimisant l'erreur quadratique integrée $ISE(h)$. Du fait qu'il dépende des fonctions inconnues f et m, ce paramètre de lissage n'est pas directement calculable en pratique. Il est alors nécessaire de s'affranchir de ces quantités inconnues en utilisant la méthode de validation croisée. Pour cela, considérons d'abord la décomposition suivante :

$$ISE(h) = ISE(U_n, m) = \int_{\mathbb{R}^k} \big(U_n^h(\mathbf{x})\big)^2 \tilde{W}(\mathbf{x})\,\mathrm{d}\tilde{P}(\mathbf{x}) - 2\int_{\mathbb{R}^k} U_n^h(\mathbf{x}) m(\mathbf{x}) \tilde{W}(\mathbf{x})\,\mathrm{d}\tilde{P}(\mathbf{x}) + \int_{\mathbb{R}^k} m(\mathbf{x})^2 \tilde{W}(\mathbf{x})\,\mathrm{d}\tilde{P}(\mathbf{x}),$$

où $\tilde{W}(\mathbf{x}) = \prod_{j=1}^k w(x_j)$, avec w est une fonction de poids positive connue, et $\tilde{P}$ désigne la mesure de probabilité associée à $\mathbf{X} = (X_1, \ldots, X_k)$. Comme le dernier terme est indépendant de h, choisir h minimisant $ISE(h)$ revient alors à choisir h minimisant la quantité

$$\int_{\mathbb{R}^k} \big(U_n^h(\mathbf{x})\big)^2 \tilde{W}(\mathbf{x})\,\mathrm{d}\tilde{P}(\mathbf{x}) - 2\int_{\mathbb{R}^k} U_n^h(\mathbf{x}) m(\mathbf{x}) \tilde{W}(\mathbf{x})\,\mathrm{d}\tilde{P}(\mathbf{x}).$$

En utilisant l'idée de la validation croisée pour estimer cette quantité, la règle de sélection de la largeur de fenêtre est finalement obtenue en choisissant comme paramètre de lissage

$$h_{cv} = \arg\min_{h\in H_n} CV(h),$$

où

$$CV(h) = \big(A_n^k\big)^{-1} \sum_{I_n^k} \big[\varphi(\mathbf{Y_j}) - U_n^{h,\mathbf{j}}(\mathbf{X_j})\big]^2 \tilde{W}(\mathbf{X_j}).$$

Dans ces définitions, H_n désigne l'ensemble des valeurs possibles pour h, $A_n^k = card(I_n^k) = n!/(n-k)!$ et $U_n^{h,\mathbf{j}}(\mathbf{x})$ est l'estimateur validé croisé de $U_n^h(\mathbf{x})$ donné par :

$$U_n^{h,\mathbf{j}}(\mathbf{x}) = \frac{\sum_{\mathbf{i}\in I_n^k,\, \mathbf{i}\neq\mathbf{j}} \varphi(\mathbf{Y_i}) \tilde{K}_h(\mathbf{x}, \mathbf{X_i})}{\sum_{\mathbf{i}\in I_n^k,\, \mathbf{i}\neq\mathbf{j}} \tilde{K}_h(\mathbf{x}, \mathbf{X_i})}.$$

Pour la suite, nous définissons l'erreur quadratique moyenne et l'erreur quadratique moyenne integrée pour les U-statistiques conditionnelles respectivement par : $ASE(h) = (A_n^k)^{-1} \sum_{I_n^k} [U_n^h(\mathbf{X_i}) - m(\mathbf{X_i})]^2 \tilde{W}(\mathbf{X_i})$ et $MISE(h) = \int_{\mathbb{R}^k} \mathbb{E}[U_n^h(\mathbf{x}) - m(\mathbf{x})]^2 \tilde{W}(\mathbf{x})\,\mathrm{d}\tilde{P}(\mathbf{x})$.

Le critère de validation croisée proposé ici est un critère de choix global, à la différence du choix local, la fonction de poids w est une fonction indépendante de x et de n (voir, par exemple, Benhenni, Ferraty, Rachdi, et Vieu [1] pour le choix local établi dans le cas des variables fonctionnelles).

3. Optimalité asymptotique du paramètre de lissage sélectionné

3.1. Hypothèses

Considérons maintenant l'ensemble des hypothèses nécessaires pour établir notre résultat principal donné dans le paragraphe 3.2.

- **(H.1)** H_n est un ensemble fini de paramètres h, tels que, pour $\tau > 0$, $\gamma > 0$ et $C > 0$, on a $card(H_n) \leqslant C(A_n^k)^\tau$ et $C^{-1}(A_n^k)^{\gamma-1} \leqslant h^k \leqslant C(A_n^k)^{-\gamma}$.
- **(H.2)** La fonction w est une fonction positive, bornée et à support compact d'intérieur non vide.
- **(H.3)** Le noyau K est borné, à support compact et satisfaisant la condition, $|K(x) - K(t)| \leqslant C|x-t|^\tau$ pour tout $(x,t) \in \mathbb{R}^2$ et certaines constantes $C > 0$ et $\tau > 0$. En outre, $\int K(u)\,du = 1$.
- **(H.4)** La fonction de densité marginale f est positive, bornée sur le support de w, et continue au point x_j, pour $1 \leqslant j \leqslant k$.
- **(H.5)** La fonction φ est bornée. De plus, pour $p \in \mathbb{N}^*$, il existe une constante C_p telle que : pour tout $\mathbf{x} \in \mathbb{R}^k$, $\mathbb{E}(|\varphi(\mathbf{Y})|^p|\mathbf{X} = \mathbf{x}) \leqslant C_p$.
- **(H.6)** Les fonctions m et f sont de classe C^2, hölderiennes et dont les dérivées sont bornées.
- **(H.7)** La fonction $m_c(x_1,\dots,x_c,v_{c+1},\dots,v_k,\acute{v}_{c+1},\dots,\acute{v}_k) := m_c(\mathbf{x}_c,\mathbf{v}_{c+1},\acute{\mathbf{v}}_{c+1}) := \mathbb{E}(\varphi(Y_1,\dots,Y_c,B_{c+1},\dots,B_k)\varphi(Y_1,\dots,Y_c,\acute{Y}_{c+1},\dots,\acute{Y}_k)|X_i = x_i,\ V_j = v_j,\ \acute{V}_j = \acute{v}_j,\ 1 \leqslant i \leqslant c,\ c+1 \leqslant j \leqslant k)$ est positive, continue et différentiable au voisinage de $(x_1,\dots,x_c,v_{c+1},\dots,v_k,\acute{v}_{c+1},\dots,\acute{v}_k) \in \mathbb{R}^{2k-c}$. De plus ses dérivées sont bornées.

Remarque 1. Les hyphothèses (H.1)–(H.6) sont classiques et se rapprochent de celles qui sont disponibles dans la littérature relative à l'estimation non paramétrique à l'image des travaux de Härdle et Marron [2] et Rachdi et Vieu [5] par exemple. Il est clair que l'ensemble H_n des paramètres h considéré dans l'hyphothèse (H.1) est fini, mais que tous nos résultats restent valides en utilisant un intervalle des paramètres h et en imposant une condition de continuité hölderienne, i.e., pour tout $h \in I_h$, $h^* \in H_n$, $|h^k - (h^*)^k| \leqslant C(A_n^k)^{-\gamma}$, pour $\gamma > 0$. La condition (H.7) est de nature technique, elle est nécessaire à cause de la structure de la U-statistique.

3.2. Résultat principal

Nous pouvons maintenant énoncer le résultat principal de cette Note.

Théorème 3.1. *Sous les hypothèses* (H.1)–(H.7)*, le critère de sélection du paramètre de lissage qui consiste à choisir $h_{cv} \in H_n$ minimisant $CV(h)$ est asymptotiquement optimal pour les distances $d = ASE$ et $d = MISE$, dans le sens où l'on a*

$$\lim_{n\to\infty}\left[\frac{d(U_n^{h_{cv}},m)}{\inf_{h\in H_n} d(U_n^h,m)}\right] = 1, \quad p.s.$$

4. Eléments de la démonstration

Pour démontrer ce théorème, il suffit d'établir le résultat suivant :

$$\sup_{h,\tilde{h}}\left|\frac{d(U_n^h,m) - d(U_n^{\tilde{h}},m) - (CV(h) - CV(\tilde{h}))}{d(U_n^h,m) + d(U_n^{\tilde{h}},m)}\right| \to 0, \quad \text{p.s. quand } n \to \infty,$$

où $d \in \{ISE, ASE, MISE\}$ et d'utiliser les équivalences asymptotiques entre les distances ISE, ASE et $MISE$ données dans les trois lemmes suivants.

Lemme 4.1. *Sous les hypothèses* (H.1)–(H.7)*, nous avons*

$$\sup_{h\in H_n}\left|\frac{MISE(h) - ISE(h)}{MISE(h)}\right| \to 0, \quad p.s.\ quand\ n \to \infty.$$

Lemme 4.2. *Sous les hypothèses* (H.1)–(H.7)*, nous avons*

$$\sup_{h\in H_n}\left|\frac{ASE(h) - ISE(h)}{ISE(h)}\right| \to 0, \quad p.s.\ quand\ n \to \infty.$$

Lemme 4.3. *Sous les hypothèses* (H.1)–(H.7)*, nous avons*

$$\sup_{h\in H_n}\left|\frac{MISE - ASE(h)}{MISE(h)}\right| \to 0, \quad p.s.\ quand\ n \to \infty.$$

Références

[1] K. Benhenni, F. Ferraty, M. Rachdi, P. Vieu, Local smoothing regression with functional data, Comput. Statist. 22 (2007) 353–369.
[2] W. Härdle, J.S. Marron, Optimal bandwidth selection in nonparametric regression function estimation, Ann. Statist. 13 (1985) 1465–1481.
[3] J.D. Hart, Nonparametric Smoothing and Lack-of-Fit Tests, Springer Series in Statistics, Springer-Verlag, New York, 1997.
[4] J.S. Marron, W. Härdle, Random approximations to some measures of accuracy in nonparametric curve estimation, J. Multivariate Anal. 20 (1986) 91–113.
[5] M. Rachdi, P. Vieu, Nonparametric regression for functional data: automatic smoothing parameter selection, J. Statist. Plann. Inference 137 (2007) 2784–2801.
[6] P. Sarda, P. Vieu, Validation croisée pour l'estimation non-paramétrique de la densité conditionnelle, Publ. Inst. Univ. Paris 1 (1994) 57–80.
[7] A. Sen, Uniform strong consistency rates for conditional U-statistics, Sankhyā Ser. A 2 (1994) 179–194.
[8] R.J. Serfling, Approximation Theorems of Mathematical Statistics, Wiley Series in Probability and Mathematical Statistics, John Wiley & Sons, Inc., New York, 1980.
[9] W. Stute, Conditional U-statistics, Ann. Probab. 2 (1991) 812–825.
[10] M.P. Wand, M.C. Jones, Kernel Smoothing, Chapman & Hall, London, 1995.