Estimation adaptative de la régression multivariée par noyaux déformés

Thomas Laloë[1] & Gaëlle Chagny[2] & Rémi Servien[3]

[1] *LJAD, Parc Valrose, 06108 Nice Cedex 02, thomas.laloe@unice.fr*
[2] *LMRS, UMR CNRS 6085, Université Rouen Normandie, gaelle.chagny@univ-rouen.fr*
[3] *Toxalim, Université de Toulouse, INRA, INP-ENVT, Toulouse, remi.servien@inra.fr*

Résumé. Nous considérons le problème de l'estimation non-paramétrique d'une fonction de régression multivariée sans hypothèse sur la compacité du support du design aléatoire, via une méthode dite de "déformation" (warping). Un estimateur à noyau déformé adaptatif, dont on prouve qu'il est optimal au sens oracle, est tout d'abord défini dans le cas où la loi du design est connue. Dans un second temps, nous proposons d'estimer également celle-ci : les marginales sont estimées via les fonctions de répartition empiriques. Quant à la structure de dépendance, elle est reconstruite via l'estimation, à noyau toujours, de la densité de copule. Le plug-in de ces estimateurs dans celui de la fonction de régression permet ensuite d'obtenir un estimateur dans le cas général. Des simulations illustrent la méthode.

Mots-clés. estimation adaptative, fonction de régression, méthode à noyaux, déformation, densité de copule, design non compact ...

Abstract. We deal with the problem of nonparametric estimation of a multivariate regression function without any assumption on the compacity of the support of the random design, thanks to a "warping" device. An adaptive warped kernel estimator, which is proved to be optimal in the oracle sense, is first defined in the case of known design distribution. Then, we propose to also estimate it : marginal distributions of the design are estimated by the empirical cumulative distribution functions, and the dependance structure is built thanks the kernel estimation of the copula density. The plug-in of these estimates in the regression function estimator permits to obtain an estimator in the general case. A numerical study is also carried out.

Keywords. adaptive estimation, regression, kernel methods, warping device, copula density, non-compact design ...

1 Introduction

Nous considérons un couple $(\mathbf{X}, Y)$ de variables aléatoires à valeurs dans $\mathbb{R}^d \times \mathbb{R}$ liées par la relation fonctionnelle suivante, à un bruit près :

$$Y = r(\mathbf{X}) + \varepsilon \tag{1}$$

où ε est une variable centrée, admettant un moment d'ordre 2, et indépendante de la variable explicative $\mathbf{X}$. L'estimation de la fonction de régression $r : A \subset \mathbb{R}^d \to \mathbb{R}$ à partir d'un échantillon $(\mathbf{X}_i, Y_i)_{i\in\{1,\dots,n\}}$ indépendant et de même distribution que $(\mathbf{X}, Y)$ constitue un problème extrêmement classique en statistique, et de nombreux travaux sont dévolus au sujet. Les principales méthodes non-paramétriques ont été largement étudiées dans ce cadre : estimateurs à noyaux (Nadaraya 1964, Watson 1964), de type polynômes locaux, estimateurs par projection dans des bases orthonormées ou de splines, estimateurs de type plus proche voisins...

Ce travail est consacré à la méthode dite de "déformation" ("warping"). Cette méthode est basée, dans le cas où $d = 1$, sur l'utilisation des données transformées ou déformées $(F_{\mathbf{X}}(\mathbf{X}_i), Y_i)_i$, où $F_{\mathbf{X}}$ est la fonction de répartition de $\mathbf{X}$. Le but est de bâtir un estimateur ayant une expression simple permettant de gérer l'irrégularité éventuelle du design $\mathbf{X}$ et le caractère non-compact de son support. Le principe consiste, toujours dans le cas $d = 1$, à proposer d'abord un estimateur $\widehat{g}$ de la fonction auxiliaire $g = r \circ F_{\mathbf{X}}^{-1}$, puis à estimer la fonction cible r par $\widehat{r} = \widehat{g} \circ \widehat{F}_{\mathbf{X}}$, où $\widehat{F}_{\mathbf{X}}$ estime $F_{\mathbf{X}}$.

Initiée par les travaux de Yang (1981) dans le cas de méthodes à noyaux, l'étude d'estimateurs par déformation pour la régression est ensuite reprise par Kerkyacharian et Picard (2004), qui proposent des estimateurs par projection dans des bases déformées. Étendue également à d'autres cadres statistiques (Chesneau et Willer 2015 et Chagny 2015 par exemple) les estimateurs par déformation se révèlent performants aussi bien en théorie qu'en pratique, bien que le comportement de leur biais reste une question complexe. Cependant, les travaux précédents se concentrent tous sur des problèmes d'estimation de fonctions d'une seule variable, et l'extension à la dimension supérieure, c'est à dire le cas $d > 1$ dans le modèle (1) n'est pas automatique. Nous nous proposons donc d'étudier comment généraliser la méthode dans ce cadre (Section 2), d'étudier l'estimateur adaptatif obtenu dans un cas simple (Section 3), avant de discuter le cas général (Section 4).

2 Méthode d'estimation

On note dans la suite F_l (resp. f_l) la fonction de répartition (resp. une densité) de chaque marginale X_l, $l \in \{1, \dots, d\}$ de $\mathbf{X}$, et $F_{\mathbf{X}}$ (resp. $f_{\mathbf{X}}$) la fonction de répartition de $\mathbf{X}$ (resp. sa densité). Soit également $\widetilde{F}_{\mathbf{X}} : \mathbf{x} = (x_l)_{l=1,\dots,d} \in \mathbb{R}^d \mapsto (F_1(x_1), \dots, F_d(x_d))$. Nous supposons l'existence de son "inverse" $\widetilde{F}_{\mathbf{X}}^{-1} : \mathbf{u} \in [0,1]^d \mapsto (F_1^{-1}(u_1), \dots, F_d^{-1}(u_d))$, et nous introduisons la fonction auxiliaire $g = r \circ \widetilde{F}_{\mathbf{X}}^{-1}$ de telle sorte que $r = g \circ \widetilde{F}_{\mathbf{X}}$.

L'application généralisant naturellement au cas $d > 1$ l'estimateur proposé par Yang (1981) et repris par Chagny (2015) est

$$\mathbf{u} \in [0,1]^d \mapsto \frac{1}{n} \sum_{i=1}^{n} Y_i K_{\mathbf{h}}(\mathbf{u} - \widetilde{F}_{\mathbf{X}}(\mathbf{X}_i)), \tag{2}$$

où $K_{\mathbf{h}}(\mathbf{x}) = K_{1,h_1}(x_1)K_{2,h_2}(x_2)\dots K_{d,h_d}(x_d)$, pour $K_{l,h_l}(x) = K_l(x/h_l)/h_l$, $h_l > 0$, et

$K_l : \mathbb{R} \to \mathbb{R}$ un noyau, c'est-à-dire $\int_{\mathbb{R}} K_l(x)dx = 1$, $l \in \{1, \ldots, d\}$, le paramètre de lissage $\mathbf{h} = (h_1, \ldots, h_d)$ désignant classiquement la fenêtre.

En supposant un instant les marginales de $\mathbf{X}$ indépendantes, la densité de $\mathbf{X}$ s'écrit $f_{\mathbf{X}}(\mathbf{x}) = \prod_{l=1}^{d} f_l(x_l)$, et un rapide calcul entraîne

$$\begin{aligned}
\mathbb{E}\left[\frac{1}{n}\sum_{i=1}^{n} Y_i K_{\mathbf{h}}(\mathbf{u} - \widetilde{F}_{\mathbf{X}}(\mathbf{X}_i))\right] &= \mathbb{E}\left[r(\mathbf{X})K_{\mathbf{h}}(\mathbf{u} - \widetilde{F}_{\mathbf{X}}(\mathbf{X}))\right], \\
&= \int_{\mathbb{R}^d} r(\mathbf{x})K_{\mathbf{h}}(\mathbf{u} - \widetilde{F}_{\mathbf{X}}(\mathbf{x})) \prod_{l=1}^{d} f_l(x_l)d\mathbf{x}, \\
&= \int_{[0,1]^d} g(\mathbf{u}')K_{\mathbf{h}}(\mathbf{u} - \mathbf{u}')d\mathbf{u}' = K_{\mathbf{h}} \star (g\mathbf{1}_{[0,1]^d})(\mathbf{u}),
\end{aligned}$$

en effectuant le changement de variables $\mathbf{u}' = \widetilde{F}_{\mathbf{X}}(\mathbf{x})$ et où $\star$ désigne la convolution sur $\mathbb{R}^d$. Ainsi (2) estime bien l'auxiliaire g. Dans le cas général, une dépendance apparait entre les coordonnées X_l de $\mathbf{X}_i$, et l'on peut la modéliser par une copule. Grâce au théorème de Sklar (1959), il existe en effet C telle que la répartition $F_{\mathbf{X}}$ de $\mathbf{X}$ se ré-écrit $F_{\mathbf{X}}(\mathbf{x}) = C(F_1(x_1), \ldots, F_d(x_d)) = C(\widetilde{F}_{\mathbf{X}}(\mathbf{x}))$. La densité de copule est alors $c(\mathbf{u}) = \partial^d C/(\partial u_1 \ldots \partial u_d)(\mathbf{u})$, $\mathbf{u} \in [0;1]^d$, et la densité $f_{\mathbf{X}}$ s'exprime par $f_{\mathbf{X}}(\mathbf{x}) = c(\widetilde{F}_{\mathbf{X}}(\mathbf{x})) \prod_{l=1}^{d} f_l(x_l)$, $\mathbf{x} = (x_l)_{l=1,\ldots,d} \in \mathbb{R}^d$. Le calcul précédent entraîne alors cette fois $\mathbb{E}[n^{-1}\sum_{i=1}^{n} Y_i K_{\mathbf{h}}(\mathbf{u} - \widetilde{F}_{\mathbf{X}}(\mathbf{X}_i))] = K_{\mathbf{h}} \star (cg\mathbf{1}_{[0,1]^d})(\mathbf{u})$. Ceci justifie l'introduction de

$$\widehat{g}_{\mathbf{h}}(\mathbf{u}) = \frac{1}{n\widehat{c}(\mathbf{u})}\sum_{i=1}^{n} Y_i K_{\mathbf{h}}(\mathbf{u} - \widehat{\widetilde{F}}_{\mathbf{X}}(\mathbf{X}_i)), \quad \mathbf{u} \in \widetilde{F}_{\mathbf{X}}^{-1}(A),$$

où $\widehat{c}$ et $\widehat{\widetilde{F}}_{\mathbf{X}}$ sont respectivement des estimateurs de la densité de copule c et de la fonction $\widetilde{F}_{\mathbf{X}}$, puis

$$\widehat{r}_{\mathbf{h}}(\mathbf{x}) = \widehat{g}_{\mathbf{h}} \circ \widehat{\widetilde{F}_{\mathbf{X}}}(\mathbf{x}) = \frac{1}{n\widehat{c}(\widehat{\widetilde{F}}_{\mathbf{X}}(\mathbf{x}))}\sum_{i=1}^{n} Y_i K_{\mathbf{h}}(\widehat{\widetilde{F}}_{\mathbf{X}}(\mathbf{x}) - \widehat{\widetilde{F}}_{\mathbf{X}}(\mathbf{X}_i)), \quad \mathbf{x} \in A \tag{3}$$

comme estimateur de la fonction cible r.

3 Résultats théoriques dans un cas simple

Nous considérons d'abord le cas jouet où à la fois $\widetilde{F}_{\mathbf{X}}$ et la densité de copule c sont connues, ce qui revient globalement à connaître la loi du design.

3.1 Risque d'un estimateur à fenêtre fixée

Soit $\|.\|_{f_{\mathbf{X}}}$ la norme L^2 sur l'espace des fonctions intégrables sur A pondérée par la densité jointe $f_{\mathbf{X}}$. Pour toute fonction t sur cet espace, $\|t\|^2_{f_{\mathbf{X}}} = \int_A t^2(\mathbf{x}) f_{\mathbf{X}}(\mathbf{x}) d\mathbf{x} = \int_{\widetilde{F}_{\mathbf{X}}(A)} t^2 \circ \widetilde{F}^{-1}_{\mathbf{X}}(\mathbf{u}) c(\mathbf{u}) d\mathbf{u}$. Le risque quadratique intégré de l'estimateur $\widehat{r}_{\mathbf{h}}$ est $\mathcal{R}(\widehat{r}_{\mathbf{h}}, r) = \mathbf{E}[\|\widehat{r}_{\mathbf{h}} - r\|^2_{f_{\mathbf{X}}}]$, et peut donc s'écrire $\mathcal{R}(\widehat{r}_{\mathbf{h}}, r) = \mathbf{E}[\int_{\widetilde{F}_{\mathbf{X}}(A)} (\widehat{g}_{\mathbf{h}}(\mathbf{u}) - g(\mathbf{u}))^2 c(\mathbf{u}) d\mathbf{u}]$.

Pour contrôler ce risque, nous introduisons les hypothèses suivantes :

$(H_{c,low})$ La densité de copule est minorée par $\widetilde{F}_{\mathbf{X}}(A)$: $\exists m_C > 0,\ \forall \mathbf{u} \in [0,1]^d,\ c(\mathbf{u}) \geq m_C$.

$(H_{cg,\beta})$ La fonction cg appartient à une boule $\mathcal{N}_2(\beta, L)$ d'un espace de Nikol'skiĭ pour $L > 0$ et $\beta = (\beta_1, \ldots, \beta_d) \subset (\mathbb{R}^*_+)^d$ fixés (Nikol'skiĭ 1975).

$(H_{K,l})$ Le noyau K est d'ordre l (voir par exemple Goldenshluger et Lepski 2010).

Proposition 3.1 *Supposons que les hypothèses $(H_{c,low})$, $(H_{cg,\beta})$ et $(H_{K,l})$ pour un indice $l \in \mathbb{R}^d_+$ tel que $l_j \geq \lfloor \beta \rfloor$ sont vérifiées. Alors,*

$$\mathcal{R}(\widehat{r}_{\mathbf{h}}, r) \leq \frac{1}{m_c} \left(L \sum_{l=1}^{d} h_l^{2\beta_l} + \|K\|^2 \mathbb{E}[Y_1^2] \frac{1}{n h_1 \ldots h_d} \right).$$

Le résultat de la Proposition 3.1 est une décomposition biais-variance classique du risque quadratique intégré de l'estimateur : le premier terme du majorant est un terme de biais, d'autant plus petit que la fenêtre $\mathbf{h}$ l'est, alors que le second augmente quand $\mathbf{h}$ décroit. Un compromis est donc nécessaire pour choisir une fenêtre optimale dans une collection finie donnée $\mathcal{H}_n \subset (\mathbb{R}^*_+)^d$ de fenêtres possible. L'optimalité est entendue ici au sens où le risque quadratique de l'estimateur doit être le plus petit possible dans la collection : l'estimateur sélectionné doit satisfaire une inégalité de type oracle. Dans le cas où la régularité β de la fonction cg est connue, l'argument minimum $\mathbf{h}(\beta)$ du majorant de l'inégalité de la Proposition 3.1 peut-être calculé, et le pseudo-estimateur résultant atteint la vitesse $n^{-2\bar{\beta}/(2\bar{\beta}+d)}$, où $\bar{\beta}$ est la moyenne harmonique des β_l, $l = 1, \ldots, d$. Il s'agit de la vitesse optimale en estimation non-paramétrique en dimension d.

3.2 Sélection de fenêtre

L'objectif est maintenant de choisir l'estimateur (ce qui revient à choisir la fenêtre) optimale au sens de l'oracle dans la collection $(\widehat{r}_{\mathbf{h}})_{\mathbf{h} \in \mathcal{H}_n}$, de manière automatique, uniquement sur la base des observations, sans supposer connue la régularité des fonctions estimées. Nous appliquons une méthode inspirée de Goldenshluger et Lepski (2011) : définissons la fenêtre sélectionnée $\widehat{\mathbf{h}} = \arg\min_{h \in \mathcal{H}_n} \{\widehat{B}(\mathbf{h}) + V(\mathbf{h})\}$ avec

$$\widehat{B}(\mathbf{h}) = \max_{\mathbf{h}' \in \mathcal{H}_n} \left\{ \left\| \frac{K_{\mathbf{h}} \star (c\widehat{g}_{\mathbf{h}'})}{c} \circ \widetilde{F}_{\mathbf{X}} - \widehat{r}_{\mathbf{h}'} \right\|^2_{f_{\mathbf{X}}} - V(\mathbf{h}') \right\}_+$$

et $V(\mathbf{h}) = \kappa \mathbb{E}[Y_1^2]/(nm_c h_1 \dots h_d),\ \kappa > 0$. Ce dernier terme a le même ordre de grandeur que le majorant du terme de variance du risque de l'estimateur $\widehat{r}_{\mathbf{h}}$ et désigne une pénalité.On ajoute les hypothèses suivantes, pour prouver une borne de risque.

(H_ε) Le bruit ε admet un moment d'ordre p, pour un $p > 2$: $\mathbb{E}[|\varepsilon|^p] < \infty$.

$(H_{c,high})$ La densité de copule est majorée sur $[0,1]^d$: $\exists M_C > 0,\ \forall \mathbf{u} \in [0,1]^d,\ c(\mathbf{u}) \leq M_C$.

Théorème 3.1 *Supposons que les hypothèses (H_ε), $(H_{c,low})$ et $(H_{c,high})$ sont vérifiées. Alors, sous des hypothèses sur la collection de fenêtres $\mathcal{H}_n$, il existe deux constantes c_1 et c_2 telles que*

$$\begin{aligned}\mathcal{R}(\widehat{r}_{\widehat{\mathbf{h}}}, r) \quad \leq \quad & c_1 \min_{\mathbf{h}\in\mathcal{H}_n} \left\{ \frac{1 + \|K\|^2_{L^1([0,1]^d)}}{m_c} \left\| K_{\mathbf{h}} \star (cg)\mathbf{1}_{\widetilde{F}_{\mathbf{X}}(A)} - (cg)\mathbf{1}_{\widetilde{F}_{\mathbf{X}}(A)} \right\|^2_{L^2([0,1]^d)} \right. \\ & \left. + \|K\|^2 \mathbb{E}[Y_1^2] \frac{1}{nm_c h_1 \dots h_d} \right\} + \frac{c_2}{n}.\end{aligned}$$

L'estimateur sélectionné est ainsi aussi "bon", en terme de risque quadratique pondéré, à constante multiplicative et terme de reste près que le meilleur des estimateurs dans la collection, puisqu'il réalise le meilleur compromis biais-variance possible. Il est adaptatif, car défini sans la connaissance de la régularité de la fonction estimée. Les quantités inconnues $\mathbb{E}[Y^2]$ et m_c présente dans le terme V de la méthode de Lepski, peuvent, au prix de calculs supplémentaires, être remplacés par des estimateurs, et l'on obtient alors un résultat similaire. Les hypothèses sur la collection de fenêtres ne sont pas détaillées ici, mais portent principalement sur sa taille.

4 Cas général et perspectives

Le cas général de l'étude de l'estimateur (3) requiert le "plug-in" d'estimateurs de la fonction $\widetilde{F}_{\mathbf{X}}$ ainsi que de la copule c. Pour la première fonction, un estimateur naturel est constitué des répartitions empiriques $\widehat{F}_l$ des marginales X_l de $\mathbf{X}$: on introduit donc $\widehat{\widetilde{F}}_{\mathbf{X}}(\mathbf{x}) = (\widehat{F}_1(x_1), \dots, \widehat{F}_d(x_d))$, $\mathbf{x} \in \mathbb{R}^d$. Les résultats de Kerkyacharian et Picard (2004) et ceux de Chagny (2015) s'étendent sans difficultés, mais au prix de calculs longs et fastidieux, à la dimension supérieure, et une inégalité oracle peut-également être obtenue pour l'estimateur $\widehat{r}_{\mathbf{h}}$ défini à l'aide de $\widehat{\widetilde{F}}$. Par ailleurs, l'on peut choisir raisonnablement de supposer que $\widetilde{F}_{\mathbf{X}}$ est connue et que seule la copule doit être estimée. Cela revient à dire que l'on connait la distribution de chaque covariable mais pas la structure de dépendance.

Pour estimer la densité de copule, on reprend l'estimateur à noyau de Fermanian (2005), pour en fournir une étude non-asymptotique similaire à celle menée ci-dessus : soit $\mathbf{b} = (b_1, \dots, b_d) \in (\mathbb{R}_+^*)^d$ une fenêtre multivariée, et $W_{\mathbf{b}}(\mathbf{u}) = W_{1,b_1}(u_1) \dots W_{d,b_d}(u_d)$,

pour $W_{l,b_l}(u) = W_l(u/b_l)/b_l$ for $b_l > 0$, et $W_l : \mathbb{R} \to \mathbb{R}$ un noyau univarié. L'estimateur de c est

$$\widehat{c}_{\mathbf{b}}(\mathbf{u}) = \frac{1}{n}\sum_{i=1}^{n} W_{\mathbf{b}}\left(\mathbf{u} - \widehat{\widetilde{F}}_{\mathbf{X}}(\mathbf{X}_i)\right), \quad \mathbf{u} \in [0,1]. \tag{4}$$

On remarquera qu'il fait lui-même à nouveau intervenir une déformation des données par $\widetilde{F}_{\mathbf{X}}$ quand celle-ci est connue, fonction que l'on remplace par $\widehat{\widetilde{F}}_{\mathbf{X}}$ dans le cas général. En notant $\|.\|_{L^2([0,1]^d)}$ la norme classique sur $L^2([0,1]^d)$, on obtient la borne de risque suivante, dans le cas $\widetilde{F}_{\mathbf{X}}$ connue,

$$\mathbb{E}\left[\|\widehat{c}_{\mathbf{b}} - c\|^2_{L^2([0,1])}\right] \leq \left\|W_{\mathbf{b}} \star (c\mathbf{1}_{[0,1]^d}) - c\right\|^2_{L^2([0,1]^d)} + \frac{\|W\|^2_{L^2([0,1]^d)}}{nb_1 \cdots b_d}.$$

Une méthode de Goldenshluger et Lepski, dans l'esprit de celle détaillée pour la régression à la Section 3 permet à nouveau de sélectionner la meilleure fenêtre $\widehat{\mathbf{b}}$ dans une collection donnée, et sous l'hypothèse $(H_{c,high})$, on prouve une inégalité oracle pour l'estimateur $\widehat{c}_{\widehat{\mathbf{b}}}$.

L'étude de l'estimateur $\widehat{r}_{\mathbf{h}}$ dans lequel on a remplacé c par son estimateur, ainsi que la sélection de $\mathbf{h}$ dans ce cadre, est un travail en cours.

Des simulations seront présentées afin d'illustrer ces résultats théoriques. L'ensemble, ainsi que les preuves, pourront être retrouvés dans Chagny *et al.* (2017).

Bibliographie

[1] Chagny, G. (2015), Adaptive warped kernel estimators. *Scand.J.Stat.* 42 (2), 336–360.
[2] Chagny, G., Laloë, T. et Servien, R. (2017), Adaptive multivariate regression estimation using warped kernels. *En préparation.*
[3] Chesneau, C. et Willer, T. (2015), Estimation of a cumulative distribution function under interval censoring 'case 1" via warped wavelets. *Comm. Statist. Theory Methods*, 44 (17), 3680–3702.
[4] Goldenshluger, A. et Lepski, O. (2011), Bandwidth selection in kernel density estimation : oracle inequalities and adaptive minimax optimality. *Ann. Statist.*, 39(3), 1608–1632.
[5] Fermanian, J-D. (2005), Goodness-of-fit tests for copulas. *J. Multivariate Anal.* 95(1) 119–152.
[6] Kerkyacharian, G. et Picard, D. (2004), Regression in random design and warped wavelets. *Bernoulli*, 10(6), 1053–1105.
[7] Nikol′skĭi, S.M. (1975), *Approximation of functions of several variables and imbedding theorems.* Springer-Verlag, New York-Heidelberg.
[8] Yang, S. (1981), Linear functions of concomitants of order statistics with application to nonparametric estimation of a regression function. *J. Amer. Statist. Assoc.* 76(375), 658–662.