Estimation de densité conditionnelle en grandes dimensions par algorithme glouton

Minh-Lien Jeanne Nguyen[1]

[1] *LM-Orsay, Université Paris-Sud,*
91405 Orsay Cedex
minh-lien.nguyen@math.u-psud.fr

Résumé. Dans cet exposé, une nouvelle méthode par algorithme glouton est présentée pour l'estimation de densité conditionnelle. Plusieurs objectifs sont recherchés : contourner le fléau de la grande dimension sous une hypothèse de parcimonie, converger à vitesse adaptative optimale au sens minimax, assurer une exécution rapide.

Mots-clés. Modèles non paramétriques, Grande dimension, Statistique mathématique, Parcimonie, Densité conditionnelle, Estimateurs à noyau, Algorithme glouton.

Abstract. In this talk, a new method using greedy algorithm is presented for conditional density estimation. Several goals are pursued: avoiding the curse of high dimensionality under a sparsity condition, achieving an adaptive optimal minimax rate, being computationally expedient.

Keywords. Nonparametric models, High-dimensional problems, Mathematical statistics, Sparsity, Conditional density, Kernel density estimator, Greedy algorithm.

1 Introduction

Dans de nombreuses situations, l'étude d'une donnée d'intérêt Y peut s'expliquer en fonction de données auxiliaires X. Dans ce type de contexte, la densité conditionnelle de Y sachant X s'avère bien plus riche que la fonction de régression, définie comme l'espérance conditionnelle de Y sachant X. Ansi la littérature propose de nombreuses méthodes et applications de l'estimation de la densité conditionnelle : en économie [3], en médecine [8], en météorologie [4] entre autres.

Il s'agit aussi d'offrir une alternative fréquentiste aux méthodes bayésiennes de type ABC [1, 7] pour sélectionner la loi a posteriori.

Un point essentiel de notre approche est de traiter le problème lorsque les observations évoluent dans un espace de grande dimension. En effet, pour des dimensions supérieures à 3, la plupart des méthodes pré-citées sont coûteuses algorithmiquement.

De manière plus fondamentale, dans le cadre de la théorie minimax, le fléau de la grande dimension restreint la vitesse de convergence de tout estimateur, d'autant plus

en cas de la grande dimension et de fonction à estimer peu régulière. Plus précisément, sans hypothèse supplémentaire, en notant d la dimension du couple (X, Y), si la densité conditionnelle f appartient à une boule de Hölder de régularité s, le risque quadratique de tout estimateur construit à partir d'un échantillon de taille n converge au mieux à vitesse $n^{-\frac{2s}{2s+d}}$ (à une constante près).

En adaptant l'algorithme RODEO, introduit par Lafferty et Wasserman pour les problèmes de régression [5] et d'estimation de densité [6], on propose une méthode qui combine rapidité algorithmique, convergence minimax quasi-optimale, adaptativité en la régularité de la densité conditionnelle. De plus, si la fonction à estimer est parcimonieuse, au sens où elle ne dépend que d'un nombre inconnu r de ses composantes, la spécificité de RODEO est qu'il détecte les r composantes pertinentes et contourne ainsi le fléau de la grande dimension en convergeant à la vitesse $n^{-\frac{2s}{2s+r}}$ (à un facteur logarithmique près).

2 Méthode

Commençons par préciser le modèle et quelques notations.
On dispose d'un n-échantillon $\{W_i\}_{i=1}^n$ du couple (X, Y) : pour $i = 1, \dots, n$, $W_i = (X_i, Y_i)$ avec X_i à valeurs dans $\mathbb{R}^{d_1}$ et Y_i à valeurs dans $\mathbb{R}^{d_2}$, et on note $d := d_1 + d_2$.
On suppose que la loi conditionnelle de Y sachant X est absolument continue par rapport à la mesure de Lebesgue et pour tout $x \in \mathbb{R}^{d_1}$, on note $f(x, \cdot)$ la densité de Y sachant $X = x$.
On suppose de même pour la loi marginale de X et on note f_X sa densité. De plus, on suppose que l'on dispose d'un estimateur $\tilde{f}_X$ de f_X construit à partir d'un échantillon de X indépendant de $\{(X_i, Y_i)\}_{i=1}^n$.

On s'intéresse à une estimation ponctuelle de f. Fixons le point d'intérêt $w := (x, y) \in \mathbb{R}^d$. Comme dans [2], on considère la famille d'estimateurs à noyau adaptée à la densité conditionelle : pour $K : \mathbb{R} \to \mathbb{R}$ un noyau, $h \in (\mathbb{R}_+^*)^d$ une fenêtre,

$$\hat{f}_h(w) := \frac{1}{n} \sum_{i=1}^{n} \frac{1}{\tilde{f}_X(X_i)} \prod_{j=1}^{d} \frac{K(\frac{w_j - W_{ij}}{h_j})}{h_j}$$

Alors que le choix du noyau s'avère peu influente sur la vitesse de de convergence, choisir la fenêtre se ramène à un problème de compromis biais-variance.
Pour sélectionner cette fenêtre, on adapte l'algorithme glouton RODEO, acronyme de Regularization Of Derivative Expectation Operator. Le principe général de cet algorithme est de faire décroître itération après itération les composantes de la fenêtre jusqu'à ce que les dérivées partielles associées de l'estimateur deviennent « petites ». Pour $j = 1, \dots, d$, notons $Z_{h,j} := \frac{\partial}{\partial h_j} \hat{f}_h(w)$.

L'heuristique repose sur le fait que plus f varie dans une direction j, plus la composante de la fenêtre h_j doit être petite autour de w, et cette variation est quantifiée par $Z_{h,j}$ que l'on compare alors à un seuil $\lambda_{h,j}$ choisi pour satisfaire le compromis biais-variance.
Plus précisément, on suit la procédure décrite à l'Algorithme 1, pour l'initialisation de la fenêtre $h_0 := \frac{1}{\log n}$ et le seuil $\lambda_{h,j} := \sqrt{\frac{4C_K \log(n)^a}{nh_j^2 \prod_{k=1}^{d} h_k}}$ pour $a > 1$ et C_K une constante dépendant uniquement du noyau.

Algorithm 1 RODEO

1. *Entrées :* w le point d'intérêt, $\{W_i\}_{i=1}^n$ le jeu de données, $\tilde{f}_X$ l'estimateur de f_X, $\beta > 0$ le facteur de décroissance de la fenêtre, h_0 la valeur d'initialisation de la fenêtre.
2. *Initialisation :*
   (a) Initialisation de la fenêtre : *pour* $j = 1 : d$, $\hat{h}_j^{(0)} \leftarrow h_0$.
   (b) Ensemble des composantes actives (*ie*: les composantes de la fenêtre que l'on continue à faire décroître) : $\mathcal{A}^{(0)} \leftarrow \{1, \dots, d\}$.
   (c) On pose $t := 0$ le compteur d'itérations.
3. *Tant qu'* il reste des composantes actives, *ie* : $\mathcal{A}^{(t)} \neq \emptyset$, et *tant que* la fenêtre courante n'est pas trop petite, *ie* : $\prod_{k=1}^{d} \hat{h}_k^{(t)} > \frac{1}{n}$:
   (a) Ensemble possible des composantes actives à l'itération suivante : $\mathcal{A}^{(t+1)} \leftarrow \mathcal{A}^{(t)}$
   (b) *Pour toute* composante active $j \in \mathcal{A}^{(t)}$:
      i. Mise à jour de $Z_{\hat{h}^{(t)},j}$ et de $\lambda_{\hat{h}^{(t)},j}$ pour la fenêtre courante $\hat{h}^{(t)}$.
      ii. *Si* $|Z_{\hat{h}^{(t)},j}| > 2\lambda_{\hat{h}^{(t)},j}$, mise à jour de la fenêtre : $\hat{h}_j^{(t+1)} \leftarrow \beta \hat{h}_j^{(t)}$.
      Sinon, désactivation de la composante j, ie : on retire j de $\mathcal{A}^{(t+1)}$.
   (c) On passe à l'itération suivante : $t \leftarrow t + 1$.
4. *Sortie :* $\hat{h}^{(t)}$ (et $\hat{f}_{\hat{h}^{(t)}}(w)$).

3 Résultat principal

On présente dans cette section les résultats théoriques associés à notre procédure.
Commençons par exposer les hypothèses.

HK *Hypothèses sur le noyau :* On choisit $K \in \mathcal{C}^1$, à support compact et d'ordre p, c'est-à-dire : $\int_{\mathbb{R}} t^p K(t)dt \neq 0$ et pour $l \in \{1, \dots, p-1\}$, $\int_{\mathbb{R}} t^l K(t)dt = 0$.

HR *Régularité de f :* On suppose que f appartient à une boule de Hölder de régularité $s \in (1, p)$ au voisinage de w. Plus précisément, sur le voisinage $\mathcal{V}_n(w) := \{z \in \mathbb{R}^d : \forall j \in \{1, \dots, d\}, \log n(w_j - z_j) \in \mathrm{supp}(K)\}$, f est q-fois différentiable avec $q := \lceil s - 1 \rceil$, le plus grand entier strictement plus petit que s, et il existe une constante $C > 0$ telle que : $\forall z, z' \in \mathcal{V}_n(w)$, $\left| \frac{\partial^q}{\partial z_j^q} f(z) - \frac{\partial^q}{\partial z_j^q} f(z') \right| < C \|z - z'\|^{s-q}$.

HX *Hypothèses sur f_X :* On suppose f_X bornée supérieurement et loin de 0 au voisinage de x. En notant le voisinage $\mathcal{U}_n(x) := \{u \in \mathbb{R}^{d_1} : \forall j \in \{1, \dots, d_1\}, \log n(x_j - u_j) \in \mathrm{supp}(K)\}$, $\delta := \inf_{u \in \mathcal{U}_n(x)} f_X(u) > 0$ et $\|f_X\|_{\infty, \mathcal{U}} := \sup_{u \in \mathcal{U}_n(x)} f_X(u) < \infty$.

HP *Hypothèse de parcimonie :* On suppose que f ne dépend que de r de ses d composantes, Plus précisément, il existe un sous-ensemble $\mathcal{R} \subset \{1, \dots, d\}$ de cardinal r et $f_{\mathcal{R}} : \mathbb{R}^r \to \mathbb{R}_+$ telle que pour tout $z \in \mathbb{R}^d$, en notant $z_{\mathcal{R}}$ la restriction de z à ses composantes indicées dans $\mathcal{R}$, $f(z) = f_{\mathcal{R}}(z_{\mathcal{R}})$.

HM *Hypothèse de monotonie :* Pour $j = 1, \dots, d$, les fonctions $h_j \mapsto |\mathbb{E}[Z_{h,j}]|$ sont croissantes sur $]0, \frac{1}{\log n}]$.

De plus, on se munit d'un estimateur $\tilde{f}_X$ tel qu'il existe des constantes $0 < M_X \leq 1$ et $C_X > 0$ telles que : $\mathbb{P}\left(\sup_{u \in \mathcal{U}_n(x)} \left| \frac{\tilde{f}_X(u) - f_X(u)}{\tilde{f}_X(u)} \right| > M_X \right) \leq C_X e^{-(\log n)^{3/2}}$. L'existence d'un tel estimateur est prouvée dans [2]. Sans perte de restriction, on peut aussi assurer pour n assez grand que : $\inf_{u \in \mathcal{U}_n(x)} \tilde{f}_X(u) > \frac{1}{\log n}$.

Théorème 1 *Sous les hypothèses **HK**, **Hf**, **HX**, **HP** et **HM**, la fenêtre $\hat{h}$ sélectionnée par* RODEO *vérifie pour n assez grand :*

$$\mathbb{E}\left[\left(\hat{f}_{\hat{h}}(w) - f(w) \right)^2 \right] \leq C \, (\log n)^{\frac{2s(d-r+a)}{2s+r}} n^{-\frac{2s}{2s+r}} + o(n^{-1}),$$

où $C > 0$ est une constante dépendant de f, f_X, K, r, β et a.

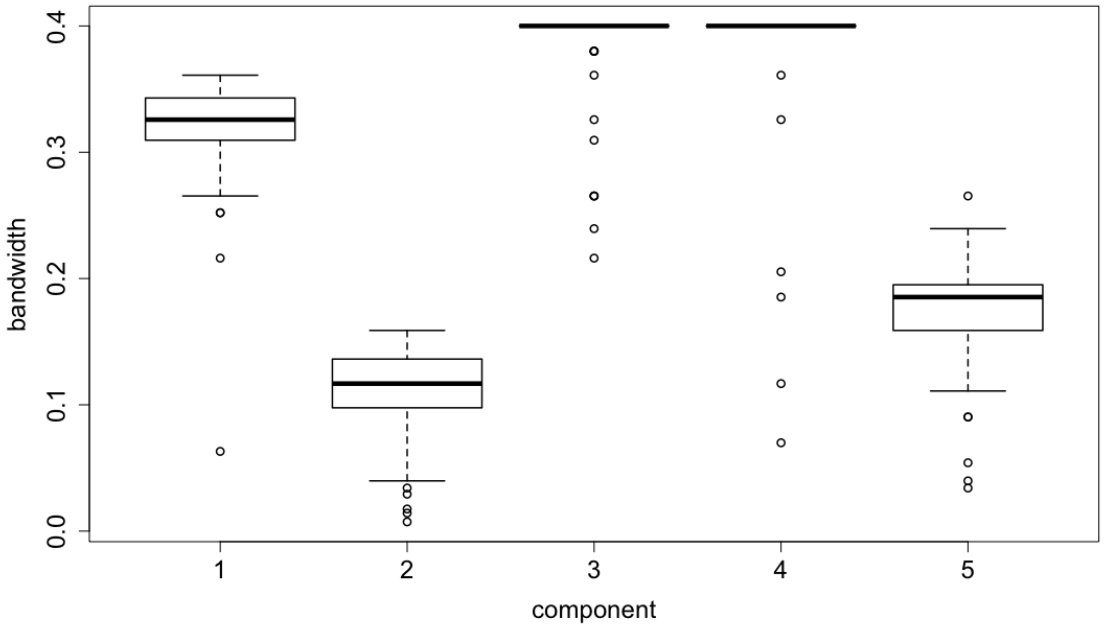
Ce théorème montre que quand $d = r$ c'est-à-dire sans l'hypothèse de parcimonie **HP**, notre estimateur atteint la vitesse minimax optimale $n^{-\frac{2s}{2s+d}}$ (au facteur logarithmique près). De plus, quand **HP** est vérifiée, la vitesse de convergence $n^{-\frac{2s}{2s+r}}$ est meilleure que $n^{-\frac{2s}{2s+d}}$. Ce résultat est donc doublement adaptatif puisqu'aussi bien la régularité de f que les composantes pertinentes sont inconnues. À noter que les résultats antérieurs sur RODEO[5, 6] ne traite que le cas $s = 2$.
Les hypothèses, exceptées **HM**, sont classiques : en particulier, la minoration de $f_X(x)$

loin de 0 est naturelle puisque que l'on estime la densité conditionnellement à $X = x$. Un travail est en cours pour relaxer **HM**.

Par ailleurs, le nombre d'opérations de RODEO est de l'ordre de $n \log n$, assurant une exécution rapide.

4 Simulations

Figure 1: Boxplot pour 100 simulations de la fenêtre sélectionnée par RODEO au point $x = (0, 1, 0, 0)$, $y = 1$

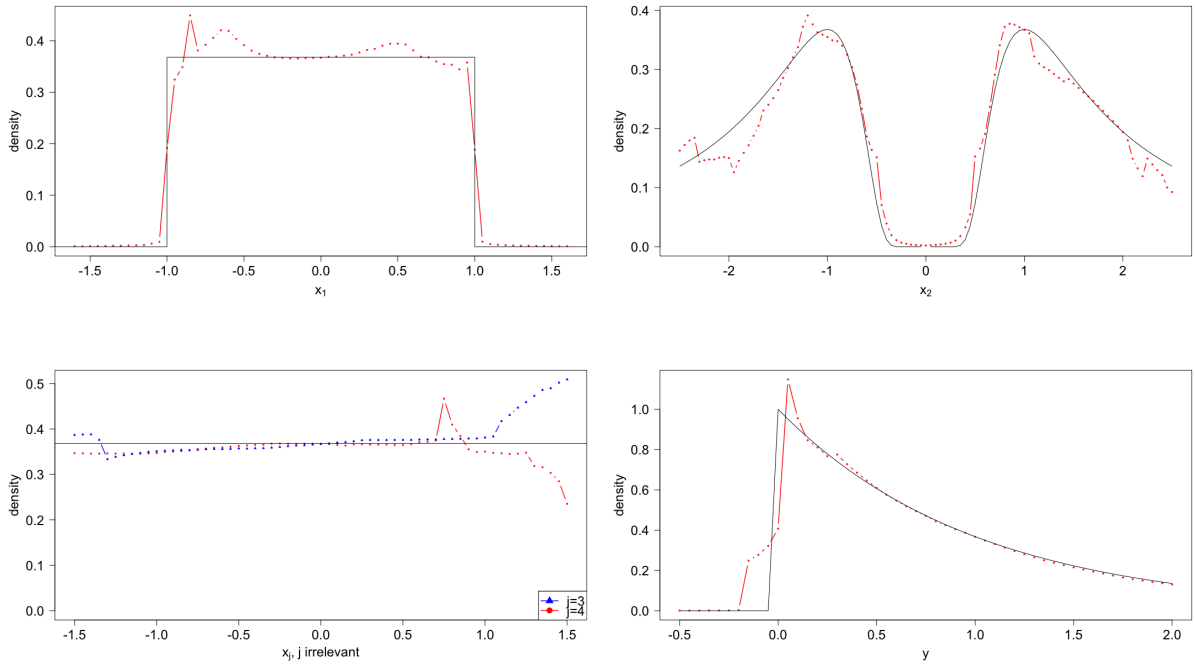


Commençons par préciser notre exemple de simulation : $d = 5$ avec $d_1 = 4$, $d_2 = 1$. L'échantillon est de taille $n = 200000$, et sa loi est décrite ainsi : pour $i = 1, \dots, n$, $X_{i1} \sim \mathcal{U}_{[0,1]}$; pour $j = 2, \dots, 4$, $X_{ij} \overset{\text{iid}}{\sim} \mathcal{N}(0, 1)$ indépendamment de X_{i1} ; et $Y_i | X_{i2} \sim \mathcal{E}(X_{i2}^{-2})$ indépendamment de (X_{i1}, X_{i3}, X_{i4}).
La densité conditionnelle de Y sachant X vaut alors : $f(x, y) = \mathbb{1}_{[-1,1]}(x_1) \times \mathbb{1}_{[0,+\infty[}(y) \frac{e^{-y/x_2^2}}{x_2^2}$.
On a pris K le noyau gaussien, $\tilde{f}_X \equiv f_X$, $\beta = 0.95$, $h_0 = 0.4$, $a = 1.1$.

La figure 1 illustre la détection des composantes non pertinentes par RODEO. En effet, presque systématiquement, les composantes x_3 et x_4 sont immédiatement désactivées et gardent leurs valeurs initiales $h_0 = 0.4$. De plus, la composante x_1 est peu pertinente autour de 0 puisque f est constante dans cette direction, ce que RODEO quantifie en sélectionnant la valeur de $\hat{h}_1$ grande par rapport à $\hat{h}_2$ et $\hat{h}_5$.

Quelques reconstructions de l'estimateur final sont tracées en figure 2.

Figure 2: Pour chaque graphique, la vraie densité conditionnelle (ligne pleine noire) est représentée en fonction de chacune des composantes, les autres étant fixées suivant $x = (0, 1, 0, 0)$, $y = 1$, et on la compare avec notre estimateur (en rouge ou bleu) évalué pour une grille de points.



Remerciements

L'auteure tient à remercier ses directeurs de thèse Claire Lacour et Vincent Rivoirard pour leurs précieux conseils et suggestions.

Bibliographie

[1] Beaumont M.A., Cornuet J.-M., Marin J.-M., Robert C.P. (2009) *Adaptive approximate Bayesian computation.* Biometrika 96: 983-990.

[2] Bertin K., Lacour C., Rivoirard V. (2016) *Adaptive pointwise estimation of conditional density function.* Ann. Inst. H. Poincaré Probab. Statist., Vol. 52, No. 2, 939-980.

[3] Hall, P., Racine, J., Li, Q. (2004) *Cross-validation and the estimation of conditional probability densities.* Journal of the American Statistical Association, Vol. 99, No. 468, 1015-1026.

[4] Jeon, J., Taylor, J. W. (2012). *Using conditional kernel density estimation for wind power density forecasting.* Journal of the American Statistical Association, Vol. 107, No 497, 66-79.

[5] Lafferty J.D., Wasserman L.A. (2008) *Rodeo: Sparse, greedy nonparametric regression.* Annals of Statistics, Vol. 36, No. 1, 28-63.

[6] Liu H., Lafferty J.D., Wasserman L.A. (2007) *Sparse Nonparametric Density Estimation in High Dimensions Using the Rodeo.* AISTATS, 283-290.

[7] Marin J.-M., Pudlo P., Robert C.P., Ryder R.J. (2012) *Approximate Bayesian computational methods.* Statistics and Computing, Vol. 22, No. 6, 1167-1180.

[8] Takeuchi, I., Nomura, K., Kanamori, T. (2009) *Nonparametric conditional density estimation using piecewise-linear solution path of kernel quantile regression.* Neural Computation, Vol. 21, No. 2, 533-559.