Estimation plug-in d'ensembles de niveau de la fonction de régression

Dau Hai Dang [1] & Thomas Laloë [1] & Rémi Servien [2]

[1] *Université de Nice Sophia-Antipolis, Laboratoire J-A Dieudonné, Parc Valrose, 06108 Nice Cedex 02, France.*

[2] *INTHERES, Université de Toulouse, INRA, ENVT, Toulouse, France.*

Résumé. Dans cette communication, nous étudions le comportement asymptotique d'un estimateur plug-in à noyau des ensembles de niveau de la régression. Une vitesse asymptotique exacte est obtenue pour un niveau fixé pour la différence symétrique. Ensuite nous obtenons un résultat similaire quand le niveau correspond à une probabilité fixée et est par conséquent inconnu.

Mots-clés. Régression, Ensembles de niveau, Estimateur à noyau, Statistique non-paramétrique ...

Abstract. In this communication, the asymptotic behavior of a plug-in kernel estimator of the regression level sets is studied. The exact asymtotic rate is derived for a given level for the symmetric difference. Then, we study the case when the level corresponds to a fixed probability and is by consequence unknown and we provide the exact asymptotic rate.

Keywords. Regression, Level sets, Kernel estimator, Non-parametric statistic ...

1 Introduction

L'estimation d'ensembles de niveau d'une fonction g consiste à estimer l'ensemble $\{x : g(x) \geq t\}$, à partir d'un échantillon. Selon le contexte, la fonction g peut-être une densité (Hartigan, 1987; Tsybakov, 1997; Cadre, 2006; Mason and Polonik, 2009; Chen et al., 2017), une fonction de répartition (Di Bernardino et al., 2013, 2015) ou encore une fonction de régression, ce qui sera l'objet de cette communication. Plus formellement, considérons (X, Y) une paire de variables aléatoires prenant ses valeurs dans $\mathbb{R}^d \times \mathbb{R}$. Notre but est d'estimer les ensembles $\mathcal{L}(t) = \{x : r(x) \geq t\}$, où r est la fonction de régression définie par $r(x) = \mathbb{E}[Y|X = x]$. Les applications potentielles sont multiples et nous renvoyons le lecteur intéressé à Scott and Davenport (2007). Par exemple, la gravité du cancer peut-être caractérisée par une variable Y qui influence directement le choix d'une chimiothérapie (standard ou agressive). Malheureusement Y est souvent mesurée à l'aide d'une biopsie invasive. Dans ce cas il est intéressant de pouvoir

étudier les ensembles de niveau de Y au travers d'une régression sur des caractéristiques plus facilement mesurables.

L'estimation des ensembles de niveau de la fonction de régression n'a été que légerement étudiée. Un estimateur fondé sur des partitions diadiques récursives a été étudié dans Willett and Nowak (2007). Un estimateur différent fondé sur la maximisation d'une "masse en excès" est étudié dans Cavalier (1997) et Polonik and Wang (2005). Dans cette présentation, nous proposerons une méthode plug-in (voir par exemple Laloë and Servien (2013)). Plus précisément nous proposons un estimateur $\mathcal{L}_n(t)$ défini par

$$\mathcal{L}_n(t) = \{x \in \mathbb{R}^d : r_n(x) > t\}$$

où $r_n(x)$ est un estimateur à noyau de r. L'avantage principal de cet estimateur est sa simplicité de calcul, héritée de l'approche plug-in. De plus cet estimateur permet une grande souplesse sur la forme des ensembles de niveau à estimer. Nous avons retenu comme critère d'erreur le volume de la différence symétrique entre les ensembles réels et estimés :

$$d_\lambda(\mathcal{L}_n(t), \mathcal{L}(t)) = \lambda\left(\mathcal{L}_n(t) \Delta \mathcal{L}(t)\right),$$

où $\mathcal{L}_n(t)\Delta\mathcal{L}(t) = (\mathcal{L}_n(t) \cap \mathcal{L}^C(t)) \cup (\mathcal{L}_n^C(t) \cap \mathcal{L}(t))$ et λ est une mesure sur $\mathbb{R}^d$ (par exemple la mesure de Lebesgue). Ce critère d'erreur offre l'avantage d'une lisibilité (et d'une interprétation) très aisée. Dans le cas où λ est la mesure de Lebesgue il mesure juste le volume de la partie de l'espace mal attribuée par l'estimateur. Une autre possibilité serait d'utiliser la mesure de probabilité des données pour évaluer la probabilité d'être dans une zone mal attribuée. Nous prouverons la consistance de notre estimateur et expliciterons la limite et la vitesse de convergente, dans le cas ou le niveau t est donné et dans celui ou il est inconnu et qu'il faut l'estimer. Finalement nous présenterons quelques résultats sur simulations.

2 Définitions, notations et hypothèses

Soient (X, Y) une paire de variables aléatoires prenant leurs valeurs dans $\mathbb{R}^d \times J$ où $d \geq 2$ et J est un ensemble borné de $\mathbb{R}$. Supposons que X admet une densité f et soit K une densité de probabilité de $\mathbb{R}^d$ avec $\tilde{K} = \int K^2$ et rappelons que

$$r_n(x) = \begin{cases} \frac{\phi_n(x)}{f_n(x)} & \text{pour } f_n(x) > 0; \\ 0 & \text{sinon,} \end{cases} \tag{1}$$

avec $\phi_n(x) = \frac{1}{nh^d} \sum_{i=1}^n Y_i K\left(\frac{x - X_i}{h}\right)$ et $f_n(x) = \frac{1}{nh^d} \sum_{i=1}^n K\left(\frac{x - X_i}{h}\right)$ où $h = h(n) \to 0$. Nous utilisons λ pour noter la mesure de Lebesgue et λ_m la mesure induite par une fonction bornée

m dans le sens où $\lambda_m(S) = \int m \mathbb{1}_{x \in S} dx$ pour des ensembles mesurables S. Φ représente la fonction de répartion d'une loi Normale $\mathcal{N}(0,1)$. Nous pouvons maintenant introduire nos hypothèses :

- Hypothèses sur (X, Y) :
  - **A0** Il existe $t^- \geq \inf r$ tel que $\{x \mid r(x) \geq t^-\}$ soit compact;
  - **A1** Les fonctions r et f appartiennent à la classe $C^2(\mathbb{R}^d)$ et $\inf_M f > 0$ (voir la définition de M ci-dessous). Soit Θ un intervalle fermé contenu dans $]t^-, \sup_M r[$ et soit t tel que $t \in \Theta \setminus \partial\Theta$;
  - **A2** Nous avons $\inf_{r^{-1}(\Theta)} \|\nabla r\| > 0$;
  - **A3** La fonction $q(x) = \mathrm{Var}\,(Y|X = x)$ appartient à la classe C^2 et vérifie $\inf_{r^{-1}(\Theta)} q > 0$.
- Hypothèses sur le noyau :
  - **B1** Le noyau K appartient à la classe C^2, a un support compact qui est supposé inclus dans $B(0,1)$ et vérifie $K(x) = K(-x)$;
  - **B2** La fenêtre h vérifie $\frac{nh^d}{\log^8 n} \to \infty$ et $nh^{\min(d+4,2d)} \log^8 n \to 0$.

En pratique, le choix de M n'est pas très difficile. Pour estimer $L(t)$, la seule exigence est que $M \supset L(t^-)$ pour un $t^- < t$ (*i.e.* M doit contenir l'ensemble $L(t)$). Si ceci est vérifié, alors, comme nous pourrons le voir par la suite, la vitesse de convergence et la limite asymptotique sont indépendants de M.

L'hypothèse **A0** est standard en théorie de l'estimation. En effet, il est extrêmement difficile de mesurer des erreurs si l'ensemble à estimer n'est pas borné. L'hypothèse **A2** implique quet r n'ait pas de plateau pour des niveaux autour de t. En effet, si ce n'est pas le cas, alors même un excellent estimateur de r (dans le sens où $|r_n - r|$ est petit) pourrait donner un très mauvais estimateur de $L(t)$.

3 Résultats théoriques

3.1 t connu

Nous considérons tout d'abord le cas où le niveau t est connu et nous obtenons le théroème suivant.

Théorème 1. *Sous les hypothèses **A0** à **B2**, nous avons*

$$\sqrt{nh^d}\lambda(L_n(t)\Delta L(t)) \xrightarrow{\mathbb{P}} \sqrt{\frac{2\tilde{K}}{\pi}} \int_{r^{-1}\{t\}} \frac{\sqrt{q/f}}{\|\nabla r\|}\, \mathrm{d}\mathcal{H}, \tag{2}$$

et

$$\sqrt{nh^d}\mathbb{E}\left[\lambda\left(L_n(t)\Delta L(t)\right)\right] \to \sqrt{\frac{2\tilde{K}}{\pi}} \int_{r^{-1}\{t\}} \frac{\sqrt{q/f}}{\|\nabla r\|}\, \mathrm{d}\mathcal{H}. \tag{3}$$

Ce théorème fournit la limite asymptotique exacte pour la mesure de Lebesgue de la différence symétrique entre le vrai ensemble de niveau $L(t)$ et son estimation $L_n(t)$. Remarquons que (2) est une extension naturelle du Théorème 2.1 de Cadre (2006) au cas de la régression et que (3) est une amélioration du Théorème 2.1 de Laloë and Servien (2013) où seule une vitesse de convergence de $O(\sqrt{nh^d})$ est obtenue.

3.2 t inconnu

Nous nous intéressons maintenant à l'estimation de $L(t)$ quand t est inconnu car défini via une probabilité p telle que $\mathbb{P}(r(X) \geq t) = p$. Nous définissons alors un estimateur t_n de t et l'estimateur de $L(t)$ devient alors $L_n(t_n)$. Nous avons alors besoin d'une hypothèse un peu plus restrictive sur la fenêtre de l'estimateur à noyau (impliquant l'hypothèse précédente **B2**):

B2' $d \geq 3$ et la fenêtre vérifie $\frac{nh^{d+2}}{\log n} \to \infty$ et $nh^{\min(d+4,2d)} \log^8 n \to 0$.

Théorème 2. *Soit la fonction P définie par $P(u) = \mathbb{P}(r(X) \geq u)$ et soit $\Theta_s = [s_1, s_2]$ un sous-intervalle strict de Θ. Soit p tel que $p \in P(\Theta_s)$ et t tel que$P(t) = p$. Alors, sous les hypothèses **A0-B2'**:*

1. *Presque sûrement pour $n > n_0 = n_0(\omega)$, il existe un unique t_n tel que $\int_M f_n \mathbb{1}_{r_n \geq t_n} = p$. En particulier, $\mathbb{P}\left(\exists!\ t_n \text{ s.t. } \int_M f_n \mathbb{1}_{r_n \geq t_n} = p\right) \to 1$;*
2. *Nous avons $t_n \xrightarrow{a.s.} t$;*
3. *Nous avons $\sqrt{nh^d}\lambda\left[L_n(t_n)\Delta L(t)\right] \xrightarrow{\mathbb{P}} \sqrt{\frac{2\tilde{K}}{\pi}} \int_{r^{-1}\{t\}} \frac{\sqrt{q/f}}{\|\nabla r\|}\, \mathrm{d}\mathcal{H}$.*

Remarquons que par l'hypothèse **A0**, pour $u \geq t^-$, nous pouvons définir de manière équivalente $P(u)$ comme $\int_M f(x)\mathbb{1}_{r(x)\geq u}$. Cette seconde définition nous permet d'avoir une certaine harmonie dans les notations avec la définition de t_n. De plus, il est intéressant de remarquer que même si le problème d'estimer $L(t)$ avec t inconnu semble plus difficile, la vitesse de convergence reste identique.

3.3 Discussion

Tout d'abord il faut noter que la vitesse de convergence et la forme de la limite exacte sont similaires dans les deux théorèmes. De plus, si nous considérons les restrictions sur la fenêtre nous remarquons que la meilleure vitesse de convergence que nous puissions obtenir est $O\left(n^{\frac{2}{d+4}}\right)$ à un facteur $\log n$ près. Nous sommes dans une situation similaire à celle envisagée par Cadre (2006) avec un fléau de la dimension assez clair.

Discutons maintenant la forme de la limite exacte :

- Le terme $\sqrt{q/f}$ est naturel: estimer $L(t)$ est plus facile quand la variabilité est faible et que la densité est élevée;
- Remarquons que dans le cas trivial $q = 0$ partout, Y serait une fonction déterministe de X. Estimer $L(t)$ serait alors bien plus facile et nous mènerait vers une vitesse de convergence plus rapide qu'en $\sqrt{nh^d}$.
- La limite exacte dépend de l'intégrale sur $r^{-1}\{t\}$. Ceci n'est pas surprenant car $r^{-1}\{t\}$ correpond aux frontières de $L(t)$ qui est l'endroit où l'estimation est la plus difficile.

Les résultats présentés dans les Théorèmes 1 et 2 nous fournissent une extension naturelle et élégante des résultats précédemment démontrés dans Cadre (2006); Laloë and Servien (2013). Une perspective intéressante à ce travail serait d'obtenir un résultat de normalité asymptotique à la manière de celui obtenu par Polonik and Wang (2005) dans le cas de la fonction de densité. D'un point de vue plus pratique, les problèmes du choix de la fenêtre la plus appropriée et de l'estimation de l'intégrale $\int_{r^{-1}\{t\}} \frac{\sqrt{q/f}}{\|\nabla r\|} \mathrm{d}\mathcal{H}$ restent ouverts.

Bibliography

Cadre, B. (2006). Kernel estimation of density level sets. *J. Multivariate Anal.*, 97(4):999–1023.

Cavalier, L. (1997). Nonparametric estimation of regression level sets. *Statistics*, 29(2):131–160.

Chen, Y.-C., Genovese, C. R., and Wasserman, L. (2017). Density level sets: Asymptotics, inference, and visualization. *Journal of the American Statistical Association*, 112(520):1684–1696.

Di Bernardino, E., Laloë, T., Maume-Deschamps, V., and Prieur, C. (2013). Plug-in estimation of level sets in a non-compact setting with applications in multivariate risk theory. *ESAIM Probab. Stat.*, 17:236–256.

Di Bernardino, E., Laloë, T., and Servien, R. (2015). Estimating covariate functions associated to multivariate risks: a level set approach. *Metrika*, 78(5):497–526.

Hartigan, J. A. (1987). Estimation of a convex density contour in two dimensions. *J. Amer. Statist. Assoc.*, 82(397):267–270.

Laloë, T. and Servien, R. (2013). Nonparametric estimation of regression level sets using kernel plug-in estimator. *J. Korean Statist. Soc.*, 42(3):301–311.

Mason, D. M. and Polonik, W. (2009). Asymptotic normality of plug-in level set estimates. *Ann. Appl. Probab.*, 19(3):1108–1142.

Polonik, W. and Wang, Z. (2005). Estimation of regression contour clusters: an application of the excess mass approach to regression. *Journal of multivariate analysis*, 94:227–249.

Scott, C. and Davenport, M. (2007). Regression level set estimation via cost-sensitive classification. *IEEE Trans. Signal Process.*, 55(6, part 1):2752–2757.

Tsybakov, A. B. (1997). On nonparametric estimation of density level sets. *Ann. Statist.*, 25(3):948–969.

Willett, R. M. and Nowak, R. D. (2007). Minimax optimal level-set estimation. *IEEE Trans. Image Process.*, 16(12):2965–2979.