Estimation non paramétrique des ensembles de niveaux de la régression

Thomas Laloë

Estimation non paramétrique des ensembles de niveaux de la régression

Thomas Laloë

ISFA 50 avenue Tony Garnier 69007 Lyon

1 Résumé en Français

Soit (X, Y) un couple aléatoire à valeurs dans $\Lambda \times J$, où $\Lambda \subset \mathbb{R}^d$ et $J \subset \mathbb{R}$ sont supposés bornés. Nous allons chercher à estimer les ensembles de niveau de la fonction de régression r de Y sur X, définie pour tout $x \in \Lambda$ par

$$r(x) = \mathbb{E}[Y|X = x].$$

Pour $t > 0$, un ensemble de niveau pour r est défini par

$$\mathcal{L}(t) = \{x \in \Lambda : r(x) > t\}.$$

On suppose disposer d'un échantillon indépendant et identiquement distribué (i.i.d.) $\Big((X_1, Y_1), \ldots, (X_n, Y_n)\Big)$ de même loi que (X, Y). On considère alors un estimateur de type plug-in de $\mathcal{L}(t)$. Plus précisément, à partir d'un estimateur consistant $\hat{r}_n$ de r, on estime $\mathcal{L}(t)$ par

$$\mathcal{L}_n(t) = \{x \in \Lambda : \hat{r}_n(x) > t\}.$$

On utilise ainsi la méthode présentée par Cadre (2006) dans le cas de la fonction de densité.

L'essentiel des travaux de recherche sur le thème de l'estimation d'ensembles de niveau concerne la fonction de densité. On peut citer par exemple les travaux de Cadre (2006), Cuevas et Fraiman (1997), Hartigan (1995), Polonik (1995), Tsybakov (1997), Walther (1997). Ce foisonnement de travaux sur le sujet est motivé par le grand nombre d'applications possibles. L'estimation de ces ensembles de niveau est notamment utile en estimation du mode (Müller et Stawitzki (1991), Polonik (1995)), ou encore en clustering (Biau, Cadre et Pelletier (2007), Cuevas, Febrero et Fraiman (2000,2001)). En particulier, Biau, Cadre et Pelletier (2007) utilisent un estimateur des ensembles de niveau de la fonction de densité pour apporter des éléments de réponse au problème de la détermination du nombre de clusters.

Les mêmes applications sont envisageables dans le cas de la fonction de régression. Par ailleurs, il est par exemple possible d'utiliser un estimateur des ensembles de niveau

de la fonction de régression pour déterminer le trajet de l'écoulement de l'eau à partir de représentations numériques de la topographie d'une zone géographique. De la même manière, en imagerie médicale, il peut être utile d'estimer les zones où certaines fonctions de l'image dépassent un seuil fixé, par exemple pour déterminer automatiquement le lieu ou la nature d'une tumeur. On remarque que, dans ces deux exemples, l'utilisation de domaines compacts Λ et J se justifie pleinement. C'est en fait le cas dans la plupart des situations pratiques, et plus particulièrement en analyse d'images.

En dépit des nombreuses applications possibles, l'estimation des ensembles de niveau de la fonction de régression reste relativement peu étudiée dans la littérature. Müller (1993) en parle brièvement dans son *survey*. On peut également citer les travaux plus récents de Cavalier (1997), Scott et Davenport (2007), et Willett et Nowak (2007). L'estimateur proposé par Cavalier est fondé sur la maximisation de la masse en excès, et adapte celui proposé par Tsybakov (1997) dans le cas de la fonction de densité. Sous certaines hypothèses, notamment sur la régularité de la fonction de régression et sur la forme de ses ensembles de niveau, il obtient une vitesse minimax. Scott et Davenport adoptent quant-à eux une approche de type *cost sensitive*. Pour résumer, ces auteurs mesurent la qualité de l'approximation par l'espérance du coût de mauvaise classification. La méthode se révèle efficace mais souffre cependant d'une trop grande dépendance au choix de la fonction de coût.

Les différents résultats de convergence seront donnés au sens de la différence symétrique, définie par

$$\mathcal{L}_n \Delta \mathcal{L} = (\mathcal{L}_n \cap \mathcal{L}^C) \cup (\mathcal{L}_n^C \cap \mathcal{L}),$$

où, comme à chaque fois qu'il n'y aura pas de confusion possible, $\mathcal{L}_n = \mathcal{L}_n(t)$ et $\mathcal{L} = \mathcal{L}(t)$. Notre approche consiste à établir des résultats sous des hypothèses sur r et $\hat{r}_n$ aussi raisonnables que possible. En utilisant les résultats de Cuevas, González-Manteiga et Rodríguez-Casal (2006), nous commençons par présenter un résultat de convergence valable pour tout estimateur consistant $\hat{r}_n$ de r. Ensuite, nous particularisons notre approche en considérant le cas de l'estimateur à noyau de la régression. Pour cet estimateur, nous obtenons une vitesse de convergence du même ordre que celle obtenu par Cadre (2006) dans le cas de la densité.

Mots clefs : Fonction de régression, Ensembles de niveau, Estimateur plug-in, Estimateur à noyaux.

2 Summary in English

Let (X, Y) be a random pair taking values in $\Lambda \times J$, where $\Lambda \subset \mathbb{R}^d$ and $J \subset \mathbb{R}$ are supposed to be bounded. The goal is to build an estimator of the level sets of the regression function

r of Y on X, defined for all $x \in \Lambda$ by

$$r(x) = \mathbb{E}\left[Y|X = x\right].$$

For $t > 0$, a level set for r is defined by

$$\mathcal{L}(t) = \{x \in \Lambda : r(x) > t\}.$$

Assume that we have an independent and identically distributed sample (i.i.d.) $\Big((X_1, Y_1), \ldots, (X_n, Y_n)\Big)$ with same distribution than (X, Y). We then consider a plug-in estimator of $\mathcal{L}(t)$. More precisely, we use a consistent estimator $\hat{r}_n$ of r to estimate $\mathcal{L}(t)$ by

$$\mathcal{L}_n(t) = \{x \in \Lambda : \hat{r}_n(x) > t\}.$$

We adapt here the method proposed by Cadre (2006) for the density function.

Most of the research works concerning the estimation of level sets concern the density function. One can cite the works of Cadre (2006), Cuevas and Fraiman (1997), Hartigan (1995), Polonik (1995), Tsybakov (1997), Walther (1997). This proliferation of works on this subject is motivated by the high number of possible applications. Estimating this level sets can be useful in mode estimation (Müller and Stawitzki (1991), Polonik (1995)), or in clustering (Biau, Cadre and Pelletier (2007), Cuevas, Febrero and Fraiman (2000,2001)). In particular, Biau, Cadre and Pelletier (2007) use an estimator of the level sets of the density function to determine the number of clusters.

The same applications are possible with the regression function. Moreover, it is for example possible to use an estimator of the level sets of the regression function to determinate the paths of water flow from a digital representation of an area. In the same vein, in medical imaging, people want to estimate the areas where some function of the image exceeds a fixed threshold. It may be useful, for instance in order to automatically determine the location or nature of a tumor. Note that, in this both examples, the use of compact sets Λ and J is fully justified. This is generally the case in most practical situations, particularly in image analysis.

Despite the many potential applications, the estimation of the level sets of the regression function has not been extensively studied yet. Müller (1993) mentioned it briefly in his survey. One can also cite the recent work of Cavalier (1997), Scott and Davenport (2007), and Willett and Nowak (2007). The estimator proposed by Cavalier is based on the maximization of the excess mass, and adapt the estimator proposed by Tsybakov (1997) for the density function. Scott and Davenport use a cost sensitive approach.

All our consistency results are in the sense of the symmetrical difference, defined by

$$\mathcal{L}_n \Delta \mathcal{L} = (\mathcal{L}_n \cap \mathcal{L}^C) \cup (\mathcal{L}_n^C \cap \mathcal{L}),$$

where $\mathcal{L}_n = \mathcal{L}_n(t)$ and $\mathcal{L} = \mathcal{L}(t)$.

Our goal is to establish some consistency results under reasonable assumptions on r and $\hat{r}_n$. We first use the results by Cuevas, González-Manteiga and Rodríguez-Casal (2006) to state a consistency result for any consistent estimator $\hat{r}_n$ of r. Then we particularize our approach by considering a kernel estimator of the regression function. For this estimator, we get a rate of convergence equivalent to the one obtained by Cadre (2006) for the density function.

keywords : Regression function, Level sets, Plug-in estimator, Kernel estimator.

Références

[1] G. Biau, B. Cadre, and B. Pelletier. A graph-based estimator of the number of clusters. *ESAIM. Probability and Statistics*, 11 :272–280, 2007.

[2] B. Cadre. Kernel estimation of density level sets. *Journal of Multivariate Analysis*, 97(4) :999–1023, 2006.

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

[4] A. Cuevas, M. Febrero, and R. Fraiman. Estimating the number of clusters. *The Canadian Journal of Statistics*, 28(1) :367–382, 2000.

[5] A. Cuevas, M. Febrero, and R. Fraiman. Cluster analysis : a further approach based on density estimation. *Computational Statistics and Data Analysis*, 36(4) :441–459, 2001.

[6] A. Cuevas and R. Fraiman. A plug-in approach to support estimation. *The Annals of Statistics*, 25(6) :2300–2312, 1997.

[7] A. Cuevas, W. González-Manteiga, and A. Rodríguez-Casal. Plug-in estimation of general level sets. *Australian and New Zealand Journal of Statistics*, 48(1) :7–19, 2006.

[8] J. A. Hartigan. Estimation of a convex density contour in two dimensions. *Journal of the American Statistical Association*, 82(397) :267–270, 1987.

[9] D. Müller. The excess mass approach in statistics. *Beiträge zur Statistik, Universität Heidelberg*, 3, 1993.

[10] D. W. Müller and G. Sawitzki. Excess mass estimates and tests for multimodality. *Journal of the American Statistical Association*, 86(415) :738–746, 1991.

[11] R. D. Nowak and R. M. Willett. Minimax optimal level-set estimation. *IEEE Transactions on Image Processing*, 16(12) :2965–2979, 2007.

[12] W. Polonik. Measuring mass concentrations and estimating density contour clusters–an excess mass approach. *The Annals of Statistics*, 23(3) :855–881, 1995.

[13] C. Scott and M. Davenport. Regression level set estimation via costsensitive classification. *IEEE Transaction on Signal Processing*, 55 :2752–2757, 2007.

[14] A. B. Tsybakov. On nonparametric estimation of density level sets. *The Annals of Statistics*, 25(3) :948–969, 1997.

[15] G. Walther. Granulometric smoothing. *The Annals of Statistics*, 25(6) :2273–2299, 1997.