Estimation non paramétrique des ensembles de niveau de la fonction de régression

Thomas Laloë

Université Lyon I

04 Mai 2010
Neuvième Colloque JPS, Le Mont-Dore

Problème général

- Fonction de densité ou de régression $r : \Lambda \to \mathbb{R}$;
- Estimer $\mathcal{L}(t) = \{x \in \Lambda : r(x) > t\}$;
- Méthode plug-in, masse en excès, “cost-sensitive” ;
- Applications en clustering, imagerie médicale, analyse spatiale, détection de flux, etc.

État de l'art

- De nombreux résultats pour la fonction de densité ;
- Quelques résultats pour la fonction de régression ;
- Hypothèses sur la fonction de regression souvent contraignantes.

Cadre général

- $r(x) = \mathbb{E}[Y|X = x]$;
- $(X, Y) \in \Lambda \times \mathbb{R}$;
- $\Lambda \subset \mathbb{R}^d$ compact.

Différents critères d'erreur

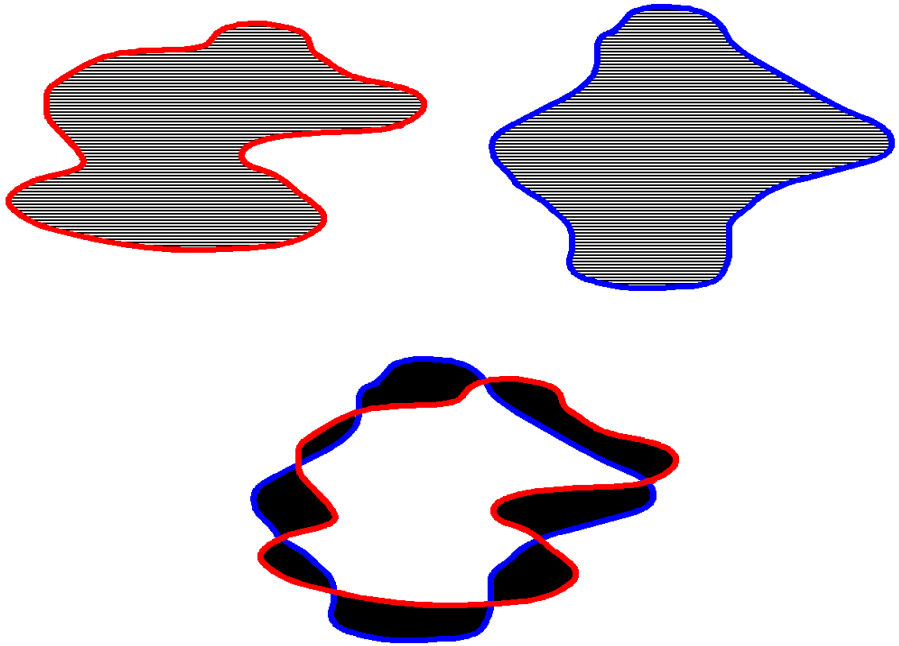
- Distance de Hausdorff :

$$d_H(A,B) = max\Big(\sup_{x\in A} d(x,B), \sup_{y\in B} d(y,A)\Big);$$

- Distance de Hausdorff entre les frontières ;

- Différence symétrique.

Différence symétrique



Principe général

- r_n estimateur convergent de r ;
- Méthode plug-in :

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

Convergence

- r_n estimateur quelconque de r ;
- On suppose que $\lambda(\{r = t\}) = 0$;

Théorème

Si

$$\mathbb{E}\|r_n - r\|_p \underset{n\to\infty}{\to} 0 \text{ } ou \text{ } \sup_{\Lambda} |r_n - r| \underset{n\to\infty}{\to} 0 \text{ } p.s,$$

alors

$$\mathbb{E}\ \lambda\Big(\mathcal{L}(t)\Delta\mathcal{L}_n(t)\Big) \underset{n\to\infty}{\to} 0.$$

Convergence

- Estimateur à Noyaux : $r_n(x) = \dfrac{\sum_{i=1}^{n} Y_i K_h(x - X_i)}{\sum_{i=1}^{n} K_h(x - X_i)}$;
- On suppose que $\lambda(\{r = t\}) = 0$;

Théorème

Si $nh^d / \log n \to \infty$, *alors*

$$\mathbb{E}\ \lambda\Big(\mathcal{L}(t)\Delta\mathcal{L}_n(t)\Big) \underset{n\to\infty}{\to} 0.$$

Vitesse de convergence

- Soit $\Theta \subset (0, \sup_\Lambda r)$ un intervalle ouvert ;

On suppose

H1 Les fonctions r et f sont deux fois continûment différentiables, et $\inf_\Lambda f > 0$;

H2 Pour tout $t \in \Theta$,

$$\inf_{\{r=t\}} \|\nabla r\| > 0.$$

Vitesse de convergence

Théorème

Si $h^d \to 0$, $nh^d/(\log n)^7 \to \infty$ *et* $nh^{d+4}\log(n) \to 0$, *alors pour presque tout* $t \in \Theta$, *il existe deux constantes strictement positives* C_1 *et* C_2 *telles que*

$$\liminf_{n\to\infty} \sqrt{nh^d}\mathbb{E}\ \lambda(\mathcal{L}_n(t)\Delta\mathcal{L}(t)) \geq C_1(t),$$

et

$$\limsup_{n\to\infty} \sqrt{nh^d}\mathbb{E}\ \lambda(\mathcal{L}_n(t)\Delta\mathcal{L}(t)) \leq C_2(t).$$

Difficultés liées a cet estimateur

- Choix du h optimal ;
- Fléau de la dimension.

Application

- Soit r : $[-6.5, 4.5] \times [-6.5, 4.5] \to [-2, 2]$ définie par

$$r : (u, v) \mapsto \sin(u) + \sin(v);$$

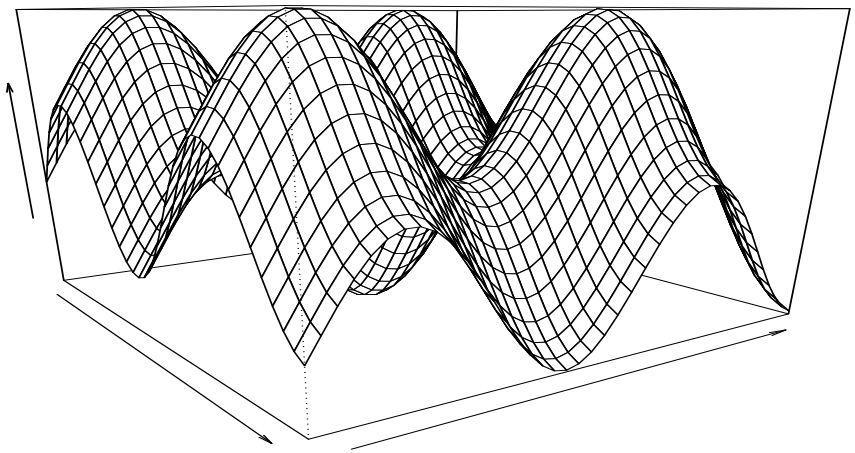


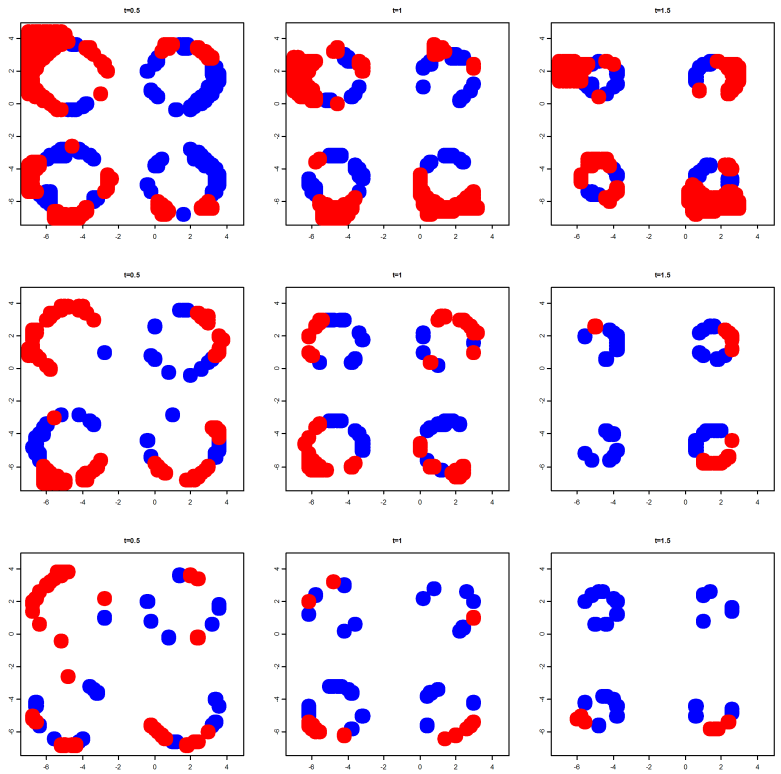
FIGURE: Représentation de $r(u, v) = \sin(u) + \sin(v)$ sur $[-6.5, 4.5] \times [-6.5, 4.5]$.

Illustration

- Soit X un couple aléatoire sur $[-6.5, 4.5] \times [-6.5, 4.5]$ de loi binormale centrée en $(-1, -1)$, et de matrice de covariance $6 * Id$;
- Soit $\Big((X_1, Y_1), \ldots, (X_n, Y_n)\Big)$ un échantillon i.i.d. avec $Y_i = r(X_i) + \varepsilon_i$, où $\varepsilon_i \sim \mathcal{N}(0, 0.01)$;
- On estime $\mathcal{L}(t)$ pour différentes valeurs de n et t.

Application

- Différences symétriques pour $n = 500, 2500, 7500$, $t = 0.5, 1, 1.5$:



Perspectives

- Vitesse exacte
- Application sur des données réelles
- Cadre fonctionnel ?