Un estimateur robuste et adaptatif pour la régression

Lucien Birgé

Séminaire Point de Vue – LPMA – 5 Mai 2014

Le problème de l'estimation d'une fonction de régression est un des problèmes importants de la Statistique. Il s'agit de l'estimation de la fonction inconnue f dans un modèle de la forme

$$Y_i = f(X_i) + \varepsilon_i, \quad 1 \leq i \leq n,$$

où les ε_i sont des erreurs i.i.d. centrées (en des sens variés) en zéro et les X_i sont soit déterministes, auquel cas on souhaite estimer le vecteur de $\mathbb{R}^n$ de coordonnées $f_i = f(X_i)$, soit aléatoires, i.i.d. et indépendants des ε_i.

Même lorsque l'on considère un cas particulier très simple tel que le modèle de translation sur $\mathbb{R}$ qui correspond à une fonction f constante ($= \theta$), le choix d'un estimateur adapté dépend de la loi (typiquement inconnue) des erreurs. Dans cette situation, les Y_i sont i.i.d. de densité $p_\theta(y) = p(y - \theta)$ où p est la densité des ε_i par rapport à la mesure de Lebesgue. L'estimateur classique des moindres carrés qui est excellent si les erreurs sont gaussiennes devient calamiteux si elles sont Cauchy. On lui préfèrera alors la médiane empirique, mais si la loi des ε_i est uniforme sur $[-1, 1]$ elle ne donnera pas la bonne vitesse d'estimation laquelle sera atteinte par l'estimateur du maximum de vraisemblance. Par contre, si la densité p n'est pas bornée, ce dernier n'existera même pas ! Tout cela pour dire qu'aucun de ces estimateurs classiques ne convient indépendamment de la loi des erreurs.

Je voudrais présenter ici le résultat d'un travail en commun avec **Yannick Baraud** et **Mathieu Sart** qui permet de fournir une solution à ce problème ainsi qu'à de nombreuses autres situations de régression. Les estimateurs en question sont fondés sur une estimation des différences de distances de Hellinger entre la vraie loi jointe des Y_i et une famille (éventuellement très large) de lois possibles. L'évaluation de la différence des distances entre deux lois et la vraie revient en fait à tester entre ces deux lois de manière robuste. Cette procédure conduit à une famille d'estimateurs bien adaptés à notre problème de régression. Ils permettent, par sélection de modèle, une adaptation simultanée à la loi des erreurs et à des hypothèses variées concernant la fonction de régression f. Ils sont en outre robustes et conduisent à des résultats raisonnables même si la modélisation initiale est approximative.