Détection par Boosting de Données Aberrantes en Régression

Nathalie Cheze[*,**], Jean-Michel Poggi[*,***]

[*]Université Paris-Sud, Lab. Mathématique, Bât. 425, 91405 Orsay, France
jean-michel.poggi@math.u-psud.fr
[**]Université Paris 10-Nanterre, Modal'X, France
cheze@u-paris10.fr
[***]Université Paris 5 Descartes, France

Résumé. Nous proposons une méthode basée sur le boosting, pour la détection des données aberrantes en régression. Le boosting privilégie naturellement les observations difficiles à prévoir, en les surpondérant de nombreuses fois au cours des itérations. La procédure utilise la réitération du boosting pour sélectionner parmi elles les données effectivement aberrantes. L'idée de base consiste à sélectionner l'observation la plus fréquemment rééchantillonnée lors des itérations du boosting puis de recommencer après l'avoir retirée. Le critère de sélection est basé sur l'inégalité de Tchebychev appliquée au maximum du nombre moyen d'apparitions dans les échantillons bootstrap. Ainsi, la procédure ne fait pas d'hypothèses sur la loi du bruit. Des exemples tests bien connus sont considérés et une étude comparative avec deux méthodes classiques illustrent le comportement de la méthode.

1 Introduction

Rousseeuw et Leroy (1987) proposent un panorama très complet des problèmes de détection de données aberrantes en régression. Le modèle sous-jacent, la méthode d'estimation et le nombre de données aberrantes par rapport à la taille de l'échantillon conduisent à définir différents types de données aberrantes. Par exemple, plusieurs voies de contamination sont distinguées : dans l'espace de la variable réponse, dans celui des covariables ou dans les deux. De nombreuses méthodes ont été développées pour traiter ces situations.

Une idée majeure est néanmoins facile à dégager : la robustesse et la référence à un modèle paramétrique sous-jacent, le plus souvent linéaire. Par exemple, on peut citer outre Rousseeuw et Leroy (1987), Pena et Yohai (1999) et pour un rapide panorama, saisi au travers de la présentation d'un logiciel, on pourra consulter Verboven et Hubert (2005)). Parmi les méthodes classiques évoquées on peut en dégager deux dont les principes marquent fortement ce type de méthodes.

Le premier contexte est celui des méthodes factorielles robustes (voir Jolliffe (2002)) qui sont basées sur des estimateurs robustes de la matrice de covariance comme l'estimateur MCD (pour Minimum Covariance Determinant, voir Rousseeuw et Van Driessen (1999)). Les données (non aberrantes) sont assimilées à un vecteur gaussien dont les caractéristiques sont estimées sur un échantillon, éventuellement contaminé par des observations aberrantes, par des