Sélection optimale de modèles à base localisée en régression hétéroscédastique

Fabien Navarro [1] & Adrien Saumard [2]

[1] *CREST, ENSAI, Campus de Ker-Lann, Rue Blaise Pascal, BP 37203, 35172 Bruz cedex, France, fabien.navarro@ensai.fr*

[2] *CREST, ENSAI, Campus de Ker-Lann, Rue Blaise Pascal, BP 37203, 35172 Bruz cedex, France, adrien.saumard@ensai.fr*

Résumé. La notion de base localisée est un concept fécond en théorie de l'approximation linéaire, qui unifie entre autres les histogrammes, les polynômes par morceaux et les ondelettes à support compact. Nous considérons donc le problème de sélection du nombre de coefficients non nuls dans un développement en base localisée, pour l'estimation non-paramétrique d'une fonction de régression, avec design aléatoire et bruit hetéroscédastique.

Nous démontrons alors par l'établissement d'inégalités oracles l'optimalité asymptotique de l'heuristique de pente et d'une stratégie de pénalisation V-fold. Nous montrons aussi que la procédure classique de validation croisée V-fold est asymptotiquement sous-optimale au sens où elle retrouve à l'infini l'oracle construit à partir d'une fraction des données initiales égale à $(V-1)/V$.

Nous concluons l'exposé par une étude simulatoire sur des modèles d'ondelettes, où nous notons une différence sensible entre les résultats asymptotiques du cadre théorique et la pratique à distance finie. En effet, la validation croisée V-fold et sa variante par pénalisation donnent dans nos expérimentations des résultats comparables alors que la supériorité asymptotique de la pénalisation est avérée.

Mots-clés. Régression non-paramétrique, bruit hétéroscédastique, sélection de modèles, inégalités oracles, heuristique de pente, validation croisée, base localisée, ondelettes.

Abstract. The concept of localised basis is a useful tool in linear approximation theory, as it unifies histograms, piecewise polynomials and wavelets. We thus consider the selection of the order of a linear expansion into a localised basis, for the nonparametric estimation of a regression function, with random design and heteroscedastic noise.

We formulate oracle inequalities that prove the asymptotic optimality of the so-called slope heuristics and a V-fold penalisation procedure. Furthermore, we show that the classical V-fold cross-validation procedure is asymptotically suboptimal as it produces an estimate that converges to the oracle corresponding to a fraction of the initial sample equal to $(V-1)/V$.

We conclude the presentation with a simulation study on wavelets. Particularly, we exhibit a gap between the asymptotic theoretic results and the pratice at a finite horizon.

Indeed, V-fold cross-validation and penalisation give similar results on our experiments, although the asymptotic superiority of the penalisation scheme is proved.

Keywords. Nonparametric regression, heteroscedastic noise, model selection, oracle inequality, slope heuristics, cross-validation, localised basis, wavelets.

1 Heuristique de pente

L'heuristique de pente, initalement formulée par Birgé et Massart (2007), est une méthode générique de calibration empirique des pénalités en sélection de modèles. Appliquée à la sélection de modèles linéaires en régression, elle permet ainsi par exemple de calibrer la pénalité linéaire classique de Mallows dans le cas d'un niveau de bruit inconnu.

La validité de l'heuristique de pente se démontre en établissant notamment une inégalité oracle, comparant l'excès de risque de l'estimateur sélectionné avec le meilleur de la collection considérée et impliquant l'optimalité asymptotique de la procédure. Pareil résultat a été obtenu en régression hétéroscédastique avec design aléatoire par Arlot et Massart (2009) pour le cas des histogrammes et par Saumard (2012) pour le cas des polynômes par morceaux.

Nous généralisons ici les précédents résultats théoriques en établissant la validité de l'heuristique de pente pour une variété beaucoup plus grande de modèles, dits à base localisée et en particulier pour des modèles d'ondelettes.

Nous montrons sur des simulations pour des modèles d'ondelettes que l'heuristique de pente fournit en effet un estimateur quasi-optimal lorsque le niveau de bruit est inconnu ou qu'il varie peu selon la valeur du design. Par contre, si le niveau de bruit est très hétéroscédastique, la performance d'une pénalité linéaire, même calibrée par heuristique de pente, se dégrade inévitablement et il est alors judicieux de faire appel à une méthode de rééchantillonnage.

2 Validation croisée et pénalisation V-fold

Nous démontrons, également à travers l'obtention d'inégalités oracles, l'optimalité asymptotique dans le cadre hétéroscédastique d'une stratégie de pénalisation V-fold, issue des travaux de Arlot (2008). Nous montrons aussi que la procédure classique de validation croisée V-fold est asymptotiquement sous-optimale au sens où elle retrouve à l'infini l'oracle construit à partir d'une fraction des données initiales égale à $(V-1)/V$.

Nous montrons par des simulations sur des modèles d'ondelettes que la validation croisée et la pénalisation V-fold permettent de s'adapter correctement à bruit hétéroscédastique et de dépasser ainsi largement les performances d'une pénalité linéaire. Les deux méthodes fournissent à distance finie des performances comparables. Nous retrouvons ainsi, dans un contexte plus général, des conclusions similaires à celles établies par Arlot (2008) sur des modèles par histogrammes.

Bibliographie

[1] Arlot, S. (2008) V-fold cross-validation improved: V-fold penalization, preprint, arXiv:0802.0566v2.

Auteurs (année), Titre, revue, localisation.

[2] Arlot, S. et Massart, P. (2009) Data-driven calibration of penalties for least-squares regression. *J. Mach. Learn. Res.*, 10:245–279.

[3] Birgé, L. et Massart, P. (2007) Minimal penalties for Gaussian model selection. *Probab. Theory Related Fields*, 138(1-2):33–73.

[4] Saumard, A. (2012) Optimal upper and lower bounds for the true and empirical excess risks in heteroscedastic least-squares regression. *Electron. J. Statist.*, 6(1-2):579–655.