HAL
open science

Estimation non-paramétrique de courbes (ou surfaces ou images): vers des régions de confiance pour l'estimateur (localement) L2-optimal

Didier A. Girard

To cite this version:

Didier A. Girard. Estimation non-paramétrique de courbes (ou surfaces ou images): vers des régions de confiance pour l'estimateur (localement) L2-optimal. 2010. hal-00450049

HAL Id: hal-00450049

https://hal.science/hal-00450049v1

Preprint submitted on 25 Jan 2010

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Estimation non-paramétrique de courbes (ou surfaces ou images): vers des régions de confiance pour l'estimateur (localement) L2-optimal

Didier Girard
CNRS et Université Joseph Fourier, Grenoble
lab. LJK, département Statistiques, équipe SMS

January 11, 2010

Résumé: Cette note et les 4 séquences qui l'accompagnent décrivent et illustrent succinctement une méthodologie qui a pour objectif de construire des régions de confiance pour l' "estimateur optimal parmi une famille, à un paramètre, d'estimateurs possibles" dans divers problèmes de "débruitage"; le cadre des illustrations est le lissage par noyau avec comme paramètre la largeur de fenêtre

Le cadre très simple des illustrations animées présentées ici est celui où l'on dispose de n (avec ici $n = 128$ ou 512) valeurs échantillonnées bruitées d'un signal régulier f. Plus précisément, en notant y_i pour un indice i dans l'ensemble $\{1, 2, \cdots, n\}$ les n données réelles, on admet qu'il existe une fonction f "lisse" du cercle unité telle que chaque y_i est une valeur "perturbée" de f évaluée en $x_i = i/n$ avec des perturbations i.i.d. gaussiennes (c'est-à-dire que les écarts $y_i - f(i/n)$, les n bruits de mesure par exemple, sont n réalisations indépendantes d'une variable aléatoire gaussienne centrée), où la variance de la perturbation, $sigma^2$, est supposée connue.

La procédure classique de production de courbes utilisée ici est connue sous le nom "lissage par noyau avec une largeur de fenêtre adaptée (dite "data-driven parameter")", l'adaptation se faisant par la minimisation d'un des critères du type validation croisée (par exemple le fameux critère CL de Mallows ou la validation croisée CV

ou sa version "généralisée" GCV qui, rappelons-le, choisissent très souvent 3 valeurs quasiment équivalentes pour la largeur, tout au moins asymptotiquement, cad pour n grand). L'originalité ici est la production, par une méthode simple et asymptotiquement valide, non-pas d'un seul choix pour la largeur de fenêtre mais de toute une plage (en fait un intervalle) de largeurs qui est approximativement un intervalle de confiance pour la largeur optimale, cad la largeur inconnue pour laquelle le lissage associé estime de façon la "meilleure possible" la fonction sous-jacente f. La "meilleure" est jugée par la version discrétisée en les x_i de la distance L_2 (notée Δ) à f.

Si l'on pouvait connaître la loi de probabilité de la différence entre le paramètre optimal inconnu et le paramètre "data-driven" produit par exemple avec le critère CL, ce serait d'une très grande utilité puisqu'il est facile de voir qu'un percentile, disons le 95ième, de cette loi donnerait, une fois translaté par le paramètre "data-driven" observé, une borne de confiance supérieure pour le paramètre optimal, dont la validité serait correcte avec une probabilité 0.95. De la même façon, le 5ème percentile toujours similairement translaté donnerait une borne inférieure de même niveau de confiance.

La méthodologie proposée ici est basée sur un des résultats exposés en [2] : la loi, conditionnée aux données, des largeurs de fenêtre que l'on obtient en minimisant un grand nombre (1000 ici) de répliques de la version "randomized-trace" [1] du critère utilisé, par ex. CL (en fait d'une version un peu modifiée, dite version avec "aléa augmenté", notée ARCL), puis que l'on recentre autour de la largeur de fenêtre obtenue par la version non-randomisée, est en fait une bonne estimation de la loi de la différence recherchée. Le résultat théorique établit une "légère" surestimation asymptotique (cad pour n grand) de l'écart type de cette loi. On peut parler de "légère" car le facteur de surestimation (> 1) est toujours borné par $\sqrt{3/2}$ pour des noyaux positifs, cette borne étant même très pessimiste dans des contextes classiques. Par exemple, une borne supérieure de 1.058 est obtenue dans le contexte des (très classiques) splines cubiques à pas réguliers.

On peut remarquer (simplement parce que le recentrage et la translation, évoqués ci-dessus, se compensent exactement) que la méthodologie consiste finalement à proposer le 95ième (resp. 5ème) percentile de la population simulée des choix ARCL comme borne supérieure (resp. inférieure) pour le choix optimal, avec un niveau de confiance approchant 95%.

Une propriété de cette méthodologie, très appréciable en pratique, est qu'elle est invariante par toute transformation monotone du paramètre de lissage (par exemple

si on avait considéré comme “paramètre” le log de la largeur au lieu de la largeur dans toute la construction).

Une remarque importante paraît appropriée maintenant : il est connu aujourd'hui que dans de nombreux contextes divers de “débruitage” adaptatif (estimation de surfaces, reconstruction d'images, problèmes inverses, ...), le calcul de, disons, “1000 randomized CL (or GCV) choices” est d'un coût très raisonnable (ou le sera bientôt ... peut-être grâce au développement des technologies de calcul parallèle); or il en est de même pour la version AR, et cette méthodologie apparaît clairement (tout au moins dans le cas d'un seul paramètre d'adaptation) applicable à tous ces contextes. Le potentiel d'applications (pour lesquelles performance et limitations de la méthodologie sont à évaluer théoriquement et expérimentalement) est donc vaste (voir Section 8.2 de [2] pour des détails sur quelques exemples).

Dans le cadre servant d'illustration ici, les intervalles de confiance auparavant proposés nécessitaient un choix assez sophistiqué d'au moins une largeur de fenêtre supplémentaire (choix d'une “pilot bandwidth”) si l'on souhaitait que la méthodologie soit asymptotiquement valide; une autre propriété appréciable pour la pratique de cette méthodologie est qu'elle n'a pas besoin d'un tel choix de pilot bandwidth. Signalons aussi (cf [2]) que les hypothèses de “condition de bord périodique”, d'uniformité de la répartition des x_i, d'un *sigma* connu, peuvent être relaxées, et qu'on aurait pu remplacer la distance L_2 par une version pondérée par une fonction positive ou nulle de x donnée (cela induit une modification simple des critères de choix et de leurs versions randomisées), de manière à s'intéresser à un lissage localement optimal.

Sur ces animations (fichiers au format mpeg4 attachés), sont montrés le lissage (courbe rouge) associé à un percentile donné p de la population des choix ARCL (on peut faire varier p de, disons, 5 % à 95%) et le lissage optimal (vert); les 2 noyaux respectifs sont aussi montrés autour de l'axe des x; et sont ajoutés le noyau correspondant au choix CL exact (violet) et à celui du percentile $100 - p$ (pointillé).

Les 6 valeurs de la graine (“seed”) ne sont pas complètement prises au hasard: elles correspondent à 6 jeux de données pour lesquels les largeurs des intervalles de confiances “equal tails” de niveau 0.90 (cette largeur mesure en gros le support de la population des choix ARCL) sont “assez” représentatives des valeurs que l'on peut rencontrer quand “seed” varie.

Ces premières expériences indiquent que la borne de confiance supérieure ainsi construite pourrait être très utile en pratique même avec seulement 128 observations; mais la borne inférieure semble, elle, soumise à plus de variabilité.

Remerciement: Ces animations ont été produites avec Mathematica 7 par Michel Girard, professeur de mathématiques au lycée Paul Cézanne

Références:

[1] Girard, D. (1991) Asymptotic optimality of the fast randomized versions of GCV and CL in ridge regression and regularization. Ann. Statist 19 , pp. 1950-1963 (`http://projecteuclid.org/euclid.aos/1176348380`)

[2] Girard, D. (2009) Estimating the accuracy of (local) cross-validation via randomised GCV choices in kernel or smoothing spline regression. Journal of Nonparametric Statistics (`http://dx.doi.org/10.1080/10485250903095820`)