Séminaire Mathématique de Béjaïa (LaMOS)
Volume 5, 2007, pp. 45 - 49

7

Le bootstrap pour le choix de la fenêtre dans la méthode du noyau

Aicha BARECHE

Laboratoire de Modélisation et d'Optimisation des Systèmes LAMOS
Université de Béjaïa 06000, Algerie.
email : aicha_barecheyahoo.fr

Résumé Le bootstrap est une méthode de ré-échantillonnage destinée à faciliter l'inférence dans les situations complexes où les méthodes analytiques ne suffisent pas. Son utilisation pour le choix de la fenêtre dans la méthode du noyau est très fréquente. Nous donnons dans ce travail un aperçu de cette application.

Mots clés : Méthode du noyau, Fenêtre, Ré-échantillonnage, Bootstrap.

7.1 Introduction

Le bootstrap est aujourd'hui une technique fréquemment utilisée en inférence statistique. Très souple à mettre en oeuvre, elle constitue une alternative intéressante aux méthodes d'estimation classiques surtout lorsque ces dernières ne peuvent être appliquées.

Les premiers éléments de la technique du bootstrap sont assez récents car cette dernière repose sur l'usage de calculateurs puissants. Ces éléments sont apparus dans la littérature statistique en 1979. Toutefois, il a fallu attendre 1993 pour voir paraître le livre d'Efron et Tibshirani [3] qui fait encore aujourd'hui référence sur le sujet.

7.2 Le bootstrap

Le bootstrap est, tout simplement, une méthode de ré-échantillonnage destinée à faciliter l'inférence dans les situations complexes où les méthodes analytiques ne suffisent pas. Son idée est d'utiliser l'échantillon des observations pour permettre une inférence statistique plus fine. Si l'échantillon initial observé est : $X = (X_1, X_2, ..., X_n)$, on réalise un certain nombre d'échantillons -qualifiés d'échantillons bootstrap- obtenus par tirage aléatoire de n observations parmi l'échantillon initial [2, 3] :

$$\begin{aligned} X^{*1} &= (X_1^*, X_2^*, ..., X_n^*) \\ X^{*2} &= (X_1^*, X_2^*, ..., X_n^*) \\ &\vdots \\ X^{*B} &= (X_1^*, X_2^*, ..., X_n^*) \end{aligned}$$

7.3 La méthode du noyau

L'estimation non-paramétrique par la méthode du noyau est une attractive méthode de lissage pour l'estimation des fonctions densité de probabilité. Etant donné un échantillon $X_1, ..., X_n$ de taille n d'une distribution de fonction densité f, l'estimateur à noyau de Rosenblatt [7] est donné par :

$$f_n(x; h) = \frac{1}{nh} \sum_{i=1}^{n} K\left(\frac{x - X_i}{h}\right). \tag{7.1}$$

Où K est une fonction densité symétrique appelée noyau avec $\int_{\mathbb{R}} K(x)dx = 1$ et h est appelé paramètre de lissage (fenêtre).

En pratique, lorsqu'on utilise la méthode du noyau pour estimer la densité de probabilité des observations indépendantes et identiquement distribuées, il est nécessaire de choisir la fonction noyau K et la largeur de la fenêtre h (ou h_n). Le choix optimal de (K, h_n) se fait généralement suivant le critère de minimisation de l'erreur quadratique moyenne (MSE) donnée par :

$$MSE(f_n(x; h)) = \mathbb{E}(f_n(x; h) - f(x))^2, \tag{7.2}$$

ou de l'erreur quadratique moyenne intégrée (MISE) donnée par :

$$MISE(f_n(x; h)) = \mathbb{E} \int_{-\infty}^{+\infty} (f_n(x; h) - f(x))^2 dx. \tag{7.3}$$

Beaucoup d'études ont été faites pour discuter du bon choix des deux paramètres de cette méthode (K, h_n). Plusieurs parmi elles, par exemple Epanechnikov [4], montre que le choix du noyau K n'est pas aussi important et qu'il est complètement satisfaisant de choisir la fonction noyau pour la convenance du calcul informatique comme par exemple le noyau gaussien.

Dans la pratique, la critique étape dans l'estimation densité est le choix de la fenêtre h, qui contrôle le lissage de l'estimateur densité. Ce dernier problème a été largement étudié

et plusieurs méthodes ont été proposées.

► Le choix asymptotique de h minimise l'erreur quadratique moyenne dans (7.2) et il est donné par :

$$h = \left[\frac{f(x) \int K^2(x)dx}{n\{f''(x) \int K(x)x^2dx\}^2} \right]^{1/5} \tag{7.4}$$

Une ancienne idée est de remplacer f'' par son estimée f''_n

► Bowman et Rudemo [1, 8] choisissent la fenêtre h pour minimiser le critère "least squares cross-validation" donné par :

$$LSCV(h) = \int f_n(x;h)^2 dx - \frac{2}{n} \sum_{i=1}^{n} f_{h,-i}(X_i), \tag{7.5}$$

où $f_{h,-i}(x_i)$ est donné comme suit :

$$f_{h,-i}(x_i) = \frac{1}{(n-1)h} \sum_{\substack{j=1 \\ j \neq i}}^{n} K\left(\frac{x_i - X_j}{h}\right). \tag{7.6}$$

7.4 Le bootstrap pour le choix de la fenêtre dans la méthode du noyau

L'idée de base est de ré-échantillonner à partir de la distribution empirique F_n de l'échantillon initial $X_1, ..., X_n$ obtenu d'une distribution F par remplacement pour avoir un nouvel échantillon $X_1^*, ..., X_n^*$ et par la suite construire les estimées bootstrap comme suit :

$$f^*_{nj}(x;h) = (nh)^{-1} \sum_{i=1}^{n} K\left(\frac{x - X_i^*}{h}\right), \tag{7.7}$$

pour $j = 1, ..., B$, où B est le nombre d'échantillons bootstrap pris.

Hall [6] a exploité cette idée pour estimer l'erreur quadratique moyenne (7.2) et la fenêtre h_1 en choisissant pour les échantillons bootstrap une taille n_1 plus petite que la taille n de l'échantillon initial comme suit :

$$\begin{aligned} MSE^*(x, n_1, h_1) &= E\left[\{f^*_{n_1}(x;h_1) - f_n(x,h)\}^2\right] \\ &= (nn_1)^{-1} \sum_{i=1}^{n} K^2_{h_1}(x - X_i^*) - n_1^{-1}\{f_n(x;h_1)\}^2 \\ &\quad + \{f_n(x;h_1) - f_n(x;h)\}^2 \end{aligned} \tag{7.8}$$

Taylor [10] a utilisé lui aussi le même principe pour estimer l'erreur quadratique moyenne intégrée (7.3) et la fenêtre h en prenant la même taille n pour les échantillons bootstrap et en choisissant le noyau gaussien. Le résultat est comparé à celui obtenu par la méthode "least squares cross-validation" toujours dans le cas du noyau gaussien.

Faraway et Jhun [5] utilisent eux aussi le bootstrap classique pour estimer l'erreur quadratique moyenne intégrée (7.3) qu'ils minimisent par la suite pour estimer la fenêtre h. Les auteurs montrent que l'utilisation du bootstrap classique échoue. Le MISE doit être décomposé en deux termes : variance et biais. Le bootstrap estime adéquatement la variance par :

$$B^{-1}\sum_{j=1}^{B}\int \left(f_{nj}^{*}(x;h)-\bar{f}_{nj}^{*}(x;h)\right)^{2}dx \tag{7.9}$$

où

$$\bar{f}_{nj}^{*}(x;h)=B^{-1}\sum_{j=1}^{B}f_{nj}^{*}(x;h) \tag{7.10}$$

Cependant, l'estimée bootstrap du biais donnée par :

$$f_n(x;h)-\bar{f}_{nj}^{*}(x;h) \tag{7.11}$$

disparait, car la composante du biais croît avec h et peut être considérable. D'où l'échec de la méthode.

Les auteurs proposent une autre méthode "le bootsrap lissé". Ils obtiennent d'abord une estimation initiale de la densité f par une fenêtre choisie par une autre procédure ensuite ils ré-échantillonnent à partir de cela. Cette méthode peut construire une estimée du MISE qui capte bien le terme biais, et tend à améliorer l'estimée initial de la densité.

Ils construisent une estimée initiale de la densité f notée $\hat{f}_n(x;h_o)$, ensuite ils ré-échantillonnent à partir de celle-ci en utilisant l'algorithme donné par Silverman et Young [9] qui consiste à rajouter une quantité aléatoire $h_o\varepsilon$ pour chaque valeur ré-échantillonnée X_j^*, où ε est distribuée avec densité $K(.)$. Alors $X_j^* \longrightarrow X_j^* + h_o\varepsilon$.
On doit alors construire $f_{nj}^{*}(x;h)$ comme précédemment, estimer le biais par :

$$\hat{f}_n(x;h_o)-\bar{f}_{nj}^{*}(x;h), \tag{7.12}$$

et estimer le MISE comme : variance + (biais)^2 par :

$$\text{MISE}^{*}(h,h_o)=B^{-1}\sum_{j=1}^{B}\int\left(f_{nj}^{*}(x;h)-\hat{f}_n(x;h_o)\right)^{2}dx \tag{7.13}$$

Ils obtiennent le choix bootstrap de la fenêtre $\hat{h}_b$ en minimisant $\text{MISE}^{*}(h,h_o)$ à travers h.

7.5 conclusion

Aujourd'hui, le bootstrap s'est imposé dans le domaine statistique comme une technique très pratique d'inférence statistique. Elle nécessite peu d'hypothèses et est relativement facile à programmer -ce ne sont, en effet, que des tirages aléatoires.

Le bootstrap permet d'associer une erreur standard et un biais à n'importe quelle statistique, malgré le fait qu'il n'existe pas de formule théorique connue pour estimer cette statistique.

La norme L_2 a été le critère le plus populaire pour le choix de la fenêtre, mais la norme L_1 a ses avantages. L'un des avantages du bootstrap est qu'il peut être facilement adapté au critère : erreur absolue moyenne intégrée par :

$$\text{MISE}^*(h, h_o) = B^{-1} \sum_{j=1}^{B} \int |f^*_{nj}(x; h) - \hat{f}_n(x; h_o)| dx \tag{7.14}$$

Lors de l'application de la stabilité forte, le bootstrap peut être appliqué pour déterminer la distance de variation w qui caractérise l'erreur de proximité entre deux système :

$$w = w(G, E_\lambda) = \int |G - E_\lambda|(dt) = \int |g_n - e_\lambda|(t) dt \tag{7.15}$$

Références

1. A.W. Bowman. An alternative method of cross-validation for the smoothing of density estimates. *Biometrika.*, 71 :353–360, 1984.
2. I. Buvat. Introduction à l'approche bootstrap. At : (guillemet.org/irene/equipe4/coursem/bootstrap)., 2000.
3. B. Efron, and R.J. Tibshirani. An Introduction to the Bootstrap. London : Champan & Hall. 1993.
4. V.A. Epanechnikov. Nonparametric estimation of a multidimensional probability density. theory Probab. Appl., 14 : 153–158, 1969.
5. J.J. Faraway, and M. Jhun. Bootstrap Choice of Bandwidth for Density Estimation. Journal of the American Statistical Association., 85 : 1119–1122, 1990.
6. P. Hall. Using the Bootstrap to Estimate Mean Squared Error and Select Smoothing Parameter in Nonparametric Problems. J. Multivariate Anal., 32 : 177–203, 1990.
7. M. Rosenblatt. Remarks on some nonparametric estimates of a density function. Ann. Math. Statist., 27 : 832–837, 1956.
8. M. Rudemo. Empirical choice of histograms and kernel density estimators. Scand. J. Statist., 9 : 65–78, 1982.
9. B.W. Silverman, and G. Young. The Bootstrap : To smooth or not to smooth ?. Biometrika., 74 : 469–479, 1987.
10. C.C. Taylor. Bootstrap choice of the smoothing parameter in kernel density estimation. Biometrika., 76 : 705–712, 1989.