Séminaire Mathématique de Béjaïa (LaMOS)
Volume 4, 2006, pp. 77 - 82

12

Equivalence des choix du paramètre de lissage dans l'estimation des fonctions densité et d'intensité

Aïcha BARECHE[1]

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

Résumé Le choix de la fenêtre dans l'estimation des fonctions de densité et d'intensité a été largement étudié dans la littérature. L'objectif de ce travail est d'expliciter l'équivalence existant entre les deux choix de la fenêtre en fournissant les justificatifs de chaque méthode, et de voir les profits des résultats de chacune des deux méthodes dans le contexte de l'autre.

Mots clés : Estimation non paramétrique, Méthode du noyau, Paramètre de lissage, Densité, Intensité.
.

12.1 Introduction

L'estimation non-paramétrique par la méthode du noyau est une attractive méthode de lissage pour l'estimation des fonctions densité de probabilité et d'intensité d'un processus de poisson non stationnaire [4, 5]. Dans les deux cas, le choix du paramètre de lissage (fenêtre) est crucial pour la performance des estimateurs.
Le choix de la fenêtre par la méthode "cross-validation" a été largement étudié pour l'estimation densité [2, 8]. Lorsque la densité est définie sur un support borné, les effets de bord sont présents. Pour remédier au problème, on fait appel à la méthode "image miroir" [10] ou aux estimateurs basés sur les noyaux asymétriques ou les histogrammes lissés [1].
Dans le cas de l'estimation de l'intensité, le choix de la fenêtre qui a été proposé minimise l'estimateur de l'erreur quadratique moyenne sous la supposition que les données sont générées par un processus de Cox stationnaire. Malgré que les deux méthodes sont motivées de différentes manières, il y a une équivalence entre les deux choix de la fenêtre.
En plus de fournir les justifications de chaque méthode, cette équivalence des choix de la fenêtre permet de voir les profits des résultats de chaque méthode dans le contexte de l'autre [5].

12.2 Les estimateurs

Les données brutes pour les estimateurs des deux fonctions densité et d'intensité consistent à un ensemble de points $X_1, ..., X_n \in \mathbb{R}$. Pour l'estimation de densité, ils sont pris comme la réalisation de n variables aléatoires indépendantes ayant toutes la fonction densité de probabilité $f(x)$. Pour l'estimation de la fonction d'intensité, les points sont pris comme les réalisations d'un processus de poisson non-stationnaire sur un intervalle $[0, T]$ de fonction d'intensité $\lambda(x)$ donnée par :

$$\forall t, \forall x > 0 : \lambda(x) = \frac{\text{prob}\{\text{événement}(t+x, t+x+\Delta x)/\text{événement en t}\}}{\Delta x}. \tag{12.1}$$

Une raisonnable estimation des fonctions f ou λ doit être une fonction qui prend de larges valeurs dans des régions où les données sont denses et des valeurs proches de 0 quand les données sont dispersées.
Pour la construction d'une telle fonction, la méthode du noyau prend :

$$\hat{f}_t(x) = n^{-1} \sum_{i=1}^{n} \delta_t (x - X_i), \tag{12.2}$$

pour l'estimation de densité, ou :

$$\hat{\lambda}_t(x) = \sum_{i=1}^{n} \delta_t (x - X_i), \tag{12.3}$$

pour l'estimation de l'intensité. Où :

$$\delta_t(.) = (1/t)\delta(./t) \text{ pour } t > 0. \tag{12.4}$$

et $\delta(.)$ est une densité de probabilité symétrique. Le paramètre t contrôle la qualité du lissage fait, et est appelé fenêtre ou paramètre de lissage. Le facteur de normalisation (n^{-1}) fait de $\hat{f}(x)$ une densité de probabilité.

Généralement, le choix de la fenêtre t est plus important que le choix du noyau $\delta(.)$ pour la performance effective de l'estimateur à noyau. Une discussion théorique est donnée dans [11].
Pour l'estimation de densité, Rudemo et Bowman [8, 2] proposent de choisir t par la méthode "least squares cross-validation". Ils notent $\hat{t}_{CV}$ le minimum de la quantité (CV) "cross-validation" :

$$CV(t) = \int [\hat{f}_t(x)]^2 dx - 2n^{-1} \sum_{i=1}^{n} f_{t,j}(X_j). \tag{12.5}$$

où :

$$f_{t,j}(X_j) = n^{-1} \sum_{\substack{j=1 \\ j \neq i}}^{n} \delta_t(x - X_i). \tag{12.6}$$

Pour l'estimation de l'intensité, Diggle [4] propose de choisir t par la méthode du processus de Cox. Il suppose que la fonction d'intensité $\lambda(x)$ est une réalisation d'un processus aléatoire stationnaire à valeurs non-négatives $\{\Lambda(x) : x \in \mathbb{R}\}$, avec $E[\Lambda(x)] = \mu$ et $E[\Lambda(x)\Lambda(y)] = \nu(|x-y|)$, et $X_1, ..., X_n$ forment une réalisation partielle d'un processus de Cox.

Diggle montre que pour x plus grand que t unités dans $[0, T]$,

$$\begin{aligned} MSE(t) &= E([\hat{\lambda}_t(x) - \Lambda(x)]^2) \\ &= \nu(0) + \mu[1 - 2\mu K(t)]/2t + (\mu/2t)^2 \int_0^{2t} K(y)dy \end{aligned} \tag{12.7}$$

où

$$K(t) = 2\mu^{-2} \int_0^t \nu(x)dx. \tag{12.8}$$

et le cas spécial du noyau uniforme :

$$\delta(.) = \frac{1}{2} I_{[-1,1]}(.). \tag{12.9}$$

est utilisé dans l'estimateur $\hat{\lambda}_t$.

Dans le cas de processus ponctuel unidimentionnel, la fonction $K(t)$ peut être estimée par :

$$\hat{K}(t) = Tn^{-2} \sum \sum_{i \neq j} I_{[-t,t]}(X_i - X_j). \tag{12.10}$$

Alors, la fenêtre qui minimise l'erreur quadratique moyenne (MSE) donnée dans (12.7), doit être approximée par la fenêtre $\hat{t}_M$ qui minimise :

$$\hat{M}(t) = (2\hat{\mu}t)^{-1} - t^{-1}\hat{K}(t) + (2t)^{-2} \int_0^{2t} \hat{K}(y)dy. \tag{12.11}$$

où μ est estimé par :

$$\hat{\mu} = n/T. \tag{12.12}$$

12.3 L'équivalence

L'équivalence entre les deux fenêtres $\hat{t}_{CV}$ et $\hat{t}_M$ est donnée par Diggle et Marron [5] dans le théorème suivant :

Théorème 12.1 *Dans le cas du noyau uniforme (12.9),* $\hat{t}_{CV} = \hat{t}_M$*, dans le sens où chaque minimum de* $CV(t)$ *est un minimum de* $\hat{M}(t)$ *et vice versa.*

La preuve de ce théorème découle immédiatement du lemme suivant :

Lemme 12.1. Pour le noyau uniforme, $\hat{M}(t) = T.CV(t)$.

12.4 Les profits

Cette équivalence permet :

1- d'utiliser les résultats de chaque méthode dans le contexte de l'autre.

2- d'appliquer la méthode du processus de Cox avec des noyaux non-uniformes pour l'estimation de l'intensité.

3- d'appliquer la théorie asymptotique bien développée dans l'estimation de densité pour générer de nouvelles idées dans l'estimation d'intensité.

4- de motiver l'application de la méthode du processus de Cox dans l'estimation densité.

12.5 Les noyaux asymétriques et les histogrammes lissés

Le plus populaire estimateur non-paramétrique pour une fonction densité de probabilité inconnue f est l'estimateur à noyau. Quand la densité est définie sur $[0, +\infty]$ et qu'elle est bornée autour de 0 et nulle à son extérieur, on est confronté au problème des effets de bord. Une simple méthode pour corriger ce problème est "l'image miroir", ajustement considéré par Schuster [10]. Cette correction prend les fonctions noyau qui s'étendent au-delà du bord et les ramènent dans le bord, alors toute leur masse est dans l'intervalle. Le biais au bord de l'estimateur à noyau est dû à l'allocation de poids par le noyau symétrique fixé en dehors du support quand l'estimation de la densité est faite près du bord. Pour éviter ce problème, une simple idée est l'utilisation d'un noyau flexible, qui n'assigne jamais un poids en dehors du support de la fonction densité. La première catégorie de ces noyaux flexibles est les noyaux asymétriques donnés sous la forme :

$$\hat{f}_b(x) = \frac{1}{n}\sum_{i=1}^{n} K(x, b)(X_i), \tag{12.13}$$

où b est la fenêtre et le noyau asymétrique K peut être :

1. Une densité Gamma K_G avec les paramètres $(x/b + 1, b)$ donnée par :

$$K_G(\frac{x}{b} + 1, b)(t) = \frac{t^{x/b}e^{-t/b}}{b^{x/b+1}\Gamma(x/b + 1)}, \tag{12.14}$$

2. Un inverse d'une densité gaussienne K_{IG} avec les paramètres $(x, 1/b)$ donné par :

$$K_{IG}(x, \frac{1}{b})(t) = \frac{1}{\sqrt{2\pi bt^3}} \exp(-\frac{1}{2bx}(\frac{t}{x} - 2 + \frac{x}{t})), \tag{12.15}$$

3. Une réciproque de l'inverse d'une densité gaussienne K_{RIG} avec les paramètres $(1/(x - b), 1/b)$ donné par :

$$K_{RIG}(\frac{1}{x-b}, \frac{1}{b})(t) = \frac{1}{\sqrt{2\pi bt}} \exp(-\frac{x-b}{2b}(\frac{t}{x-b} - 2 + \frac{x-b}{t})). \tag{12.16}$$

L'estimateur $\hat{f}_b$ basé sur le noyau Gamma K_G a été proposé par Chen [3] et les estimateurs basés sur les noyaux K_{IG} et K_{RIG} ont été proposés par Scaillet [9].

La deuxième catégorie des noyaux flexibles est les histogrammes lissés donnés par Gawronski et stadtmüller [6, 7] sous la forme :

$$\hat{f}_k(x) = k \sum_{i=0}^{+\infty} \omega_{i,k} p_{ki}(x), \tag{12.17}$$

où les poids $\omega_{i,k}$ sont aléatoires et donnés par :

$$\omega_{i,k} = F_n(\frac{i+1}{k}) - F_n(\frac{i}{k}), \tag{12.18}$$

où F_n est la distribution empirique, l'entier k est le paramètre de lissage et $P_{ki}(.)$ peut être :

$$P_{ki}(x) = e^{-kx} \frac{(kx)^i}{i!}, \quad i = 0, 1, ..., \tag{12.19}$$

ou bien :

$$P_{ki}(x) = \int_{i/k}^{(i+1)/k} K(x, 1/k)(t)dt. \tag{12.20}$$

où $K(x, 1/k)$ est soit le noyau K_{IG} ou le noyau K_{RIG} avec une fenêtre égale à $1/k$.

La consistence uniforme faible et la convergence faible en L_1 de ces estimateurs ont été établies par Bouezmarni et Scaillet [1]

12.6 conclusion

Les résultats vus précédemment représentent une petite partie de ce qui peut être fait en termes de la recherche des analogues des résultats connus pour l'estimation densité dans

le contexte de l'estimation d'intensité. On pourrait s'intéresser entre autres à la recherche des analogues des résultats d'optimalité asymptotique, les bruits dans les idées du choix de la fenêtre et les idées du choix du noyau.

Lorsque les effets de bord sont présents dans l'estimation densité, plusieurs récentes études notent que l'utilisation des noyaux asymétriques et des histogrammes lissés sont plus appropriés et réduisent le biais. Il serait intéressant, d'un côté, de comparer les résultats de ces récentes méthodes avec les anciennes et d'un autre côté de voir l'analogue de ceci dans l'estimation d'intensité.

Références

1. T. Bouezmarni, and O. Scaillet. Consistency of Asymetric Kernel Density Estimators and Smoothed Histograms with Application to Income Data. http ://www.stat.ucl.ac.be, 2003
2. A.W. Bowman. An alternative method of cross-validation for the smoothing of density estimates. *Biometrika.*, 71 :353–360, 1984.
3. S.X. Chen. Probability Density Function Estimation Using Gamma Kernels. *Ann. Inst. Statis. Math.*, 52 :471–480, 2000.
4. P. Diggle. A kernel Method for Smoothing Point Process Data. *Appl. Statist.*, 34 :138–147, 1985.
5. P. Diggle, J.S. Marron. Equivalence of Smoothing Parameter Selectors in Density and Intensity Estimation. *J. Amer. Statist. Assoc.*, 83 :793–800, 1988.
6. N. Gawronski, U. Stadtmüller. On Density Estimation by Means of Poisson's Distribution. *Scand. J. Statist.*, 7 :90–94, 1980.
7. N. Gawronski, U. Stadtmüller. Smoothing hostograms by means of lattice and continuous distributions. *Metrika.*, 28 :155–164, 1981.
8. M. Rudemo. Empirical choice of histograms and kernel density estimators. *Scand. J. Statist.*, 9 :65–78, 1982.
9. O. Scaillet. Density Estimation Using Inverse and Reciprocal Inverse Gaussian Kernels. *IRES DP.*, 17, 2001.
10. E.F. Schuster. Incorporating support constraints into nonparametric estimation of densities. *Commun. Statist. Theory Meth.*, 14 :1123–1136, 1985.
11. B.W. Silverman. Density Estimation for Statistics and Data Analysis. estimation. Chapman and Hall, London, 1986.