

## Statistique

# Un choix de fenêtre optimal en estimation polynomiale locale de la fonction de répartition conditionnelle

Sandie Ferrigno <sup>a</sup>, Gilles R. Ducharme <sup>b</sup>

<sup>a</sup> *Equipe Probabilités et Statistiques, Institut Elie Cartan Nancy, Nancy-Université, CNRS, INRIA, Boulevard des Aiguillettes, B.P. 239, 54506 Vandœuvre-lès-Nancy cedex, France*

<sup>b</sup> *Equipe Probabilités et Statistique, Institut de Mathématiques et de Modélisation de Montpellier, cc051, Université Montpellier II, place E. Bataillon, 34095 Montpellier cedex 05, France*

Reçu le 22 avril 2005 ; accepté après révision le 22 novembre 2007

Disponible sur Internet le 26 décembre 2007

Présenté par Paul Deheuvels

---

### Résumé

Nous proposons un critère de choix optimal pour la fenêtre d'ajustement utilisée pour l'estimation polynomiale locale de la fonction de répartition conditionnelle. Ce critère est basé sur la minimisation de l'expression asymptotique de l'erreur quadratique moyenne intégrée pondérée. *Pour citer cet article : S. Ferrigno, G.R. Ducharme, C. R. Acad. Sci. Paris, Ser. I 346 (2008).*

© 2007 Académie des sciences. Publié par Elsevier Masson SAS. Tous droits réservés.

### Abstract

**An optimal choice for the bandwidth parameter in local polynomial estimation of the conditional distribution function.** We propose an optimal choice for the bandwidth parameter used in the local polynomial estimation of the conditional distribution function. This choice is approximated by the bandwidth which minimizes the asymptotic weighted mean integrated squared error (MISE). *To cite this article: S. Ferrigno, G.R. Ducharme, C. R. Acad. Sci. Paris, Ser. I 346 (2008).*

© 2007 Académie des sciences. Publié par Elsevier Masson SAS. Tous droits réservés.

---

## 1. Introduction

La fenêtre d'ajustement est un paramètre crucial en estimation non paramétrique puisqu'elle détermine le degré de lissage de l'estimateur et donc en particulier sa complexité. Une importante littérature a été consacrée au choix de cette fenêtre. Dans un contexte d'ajustement polynomial local de la fonction de régression, Doksum et al. [1], Hall et al. [5], Prewitt [7] mais aussi Fan et Gijbels [2] ont proposé des critères de choix pour cette fenêtre. Nous nous intéressons ici au choix de fenêtre dans un contexte d'estimation polynomiale locale de la fonction de répartition conditionnelle, cette dernière étant par exemple utilisée d'un point de vue pratique pour l'estimation des courbes de référence dans le domaine médical (voir Gannoun et al. [4]).

---

Adresses e-mail : Sandie.Ferrigno@iecn.u-nancy.fr (S. Ferrigno), ducharme@math.univ-montp2.fr (G.R. Ducharme).

## 2. Estimateur polynomial local de la fonction de répartition conditionnelle

Dénotons par  $F(y|x) = P(Y \leq y|X = x)$ , la fonction de répartition conditionnelle de  $Y$  sachant  $\{X = x\}$ . Considérons  $K(\cdot)$ , un noyau positif et  $h = h(n)$ , une largeur de fenêtre. Soit

$$(\hat{\beta}_0, \dots, \hat{\beta}_p) = \arg \min_{\beta_0, \dots, \beta_p} \sum_{i=1}^n \left[ I_{\{Y_i \leq y\}} - \sum_{v=0}^p \beta_v (X_i - x)^v \right]^2 K\left(\frac{X_i - x}{h}\right).$$

$\hat{\beta}_0$  est un estimateur de  $F(y|x)$  que l'on dénotera  $\hat{F}_n(y|x)$ . Cet estimateur n'est pas nécessairement une fonction de répartition et on peut parfois préférer une projection dans cet espace. Si on pose  $X_{n \times (p+1)} = ((X_i - x)^j)_{1 \leq i \leq n, 0 \leq j \leq p}$  et  $P_{n \times n} = \text{diag}(K(\frac{X_i - x}{h}))$ , on obtient :  $\hat{F}_n(y|x) = \sum_{i=1}^n W^n(\frac{X_i - x}{h}) I_{\{Y_i \leq y\}}$ , avec  $W^n(t) = e_0^T (X^T P X)^{-1} (1, ht, \dots, (ht)^p)^T K(t)$  où  $e_0^T = (1, 0, \dots, 0)$ . Dans la suite, on utilise  $p$  impair. Les autres hypothèses considérées sont les suivantes :

- (H.1) Le noyau  $K(\cdot)$  est une densité de probabilité à support compact  $[-1, 1]$ , symétrique autour de 0, bornée et de dérivée bornée.
- (H.2) La variable aléatoire  $X$  a le support compact  $[c, d]$ .
- (H.3) La densité marginale  $f(\cdot)$  de  $X$  satisfait  $0 < m < f(\cdot) < \infty$  pour tout  $x \in [c, d]$ . Elle est aussi de dérivée continue et bornée dans un voisinage de tout  $x \in ]c, d[$ .
- (H.4) Pour tout  $x \in [c, d]$ ,  $F(y|x)$  est continûment dérivable en  $y$ .
- (H.5) Pour tout  $y \in \mathbb{R}$ ,  $F(y|x)$  est  $p + 2$  fois continûment dérivable en  $x$  dans un voisinage  $V(x)$  de chacun des points de  $]c, d[$ . De plus, pour tout  $u \in V(x)$  avec  $x \in ]c, d[$ ,

$$\sup_{u \in V(x)} \sup_{y \in \mathbb{R}} |F^{(p+2)}(y|u)| \leq M.$$

- (H.6) On suppose que  $h \rightarrow 0$ ,  $nh \rightarrow \infty$  et  $\frac{\sqrt{\log n}}{\sqrt{nh^3}} \rightarrow 0$ .

Le noyau  $W^n(\cdot)$  est difficile à manipuler mathématiquement. On en donne une approximation plus stable. Soit  $K^*(t) = e_0^T S^{-1} (1, t, \dots, t^p)^T K(t)$  où  $S_{(p+1) \times (p+1)}$  est la matrice dont les éléments sont les  $(\int_{-1}^1 u^{i+j} \times K(u) du)_{0 \leq i, j \leq p}$ .

**Lemme 2.1.** Soit  $[a, b] \subset ]c, d[$ . Alors, sous les hypothèses (H.1)–(H.3) et (H.6), on a presque sûrement :

$$\sup_{t \in [-1, 1]} \sup_{x \in [a, b]} \left| nh W^n(t) - \frac{K^*(t)}{f(x)} \right| = O(h).$$

**Preuve.** Ce résultat est une amélioration du Lemme 2.1 de Huang et Fan [6]. On le démontre en étant soigneux dans la gestion des restes (voir Ferrigno et Ducharme [3]).  $\square$

Dans la suite, l'utilisation du Lemme 2.1 nous amène à travailler avec  $x \in [a, b] \subset ]c, d[$ .

## 3. Choix de fenêtre

La fenêtre d'ajustement  $h$  contrôle la largeur du voisinage considéré autour de  $x$  lors de l'ajustement polynomial. Une fenêtre trop petite reproduit presque intégralement les données et est donc très variable. Dans le cas de la fonction de répartition conditionnelle, avec  $p = 1$ , quand  $h \rightarrow 0$ , l'estimateur linéaire local de  $F(y|x)$  pour  $y$  fixe est proche d'une interpolation des données  $(I_{\{Y_i \leq y\}}, X_i)$  alors que pour  $x$  fixe,  $\hat{F}_n(y|x) = 0$  si  $X_i \neq x$  et vaut  $I_{\{Y_i \leq y\}}$  si  $X_i = x$ . A contrario, un choix du paramètre de lissage trop grand entraîne une augmentation du biais de l'estimateur. Le but est donc de choisir une fenêtre qui équilibre le biais et la variance de l'estimateur. A l'instar de Fan et Gijbels [2], on considère ici l'erreur quadratique moyenne intégrée (MISE) pondérée définie par :

$$\int_a^b \int_{\mathbb{R}} \left[ [\text{Biais}_F(\hat{F}_n(y|x))]^2 + \text{Var}_F(\hat{F}_n(y|x)) \right] \pi(x, y) F(dy|x) dx,$$

où  $\pi(x, y)$  est une fonction de poids supposée positive et bornée. Une fenêtre optimale est alors obtenue en minimisant l'expression asymptotique de cette fonction.

Calculons d'abord  $\text{Biais}_F(\hat{F}_n(y|x)|X_1, \dots, X_n)$ . Comme  $\sum_{i=1}^n (X_i - x)^k W^n(\frac{X_i - x}{h}) = 0$  pour tout  $0 < k \leq p$ ,

$$E_F(\hat{F}_n(y|x)|X_1, \dots, X_n) = \sum_{i=1}^n W^n\left(\frac{X_i - x}{h}\right) F(y|X_i).$$

De plus, par un développement de Taylor d'ordre  $p + 2$  de  $F(y|X_i)$ ,  $\exists x_i^* \in (X_i, x)$  tel que :

$$\text{Biais}_F(\hat{F}_n(y|x)|X_1, \dots, X_n) = Q_{xy} + R_{xy},$$

où

$$Q_{xy} = \frac{F^{(p+1)}(y|x)}{(p+1)!} \sum_{i=1}^n W^n\left(\frac{X_i - x}{h}\right) (X_i - x)^{p+1} \quad \text{et} \quad R_{xy} = \sum_{i=1}^n W^n\left(\frac{X_i - x}{h}\right) (X_i - x)^{p+2} \frac{F^{(p+2)}(y|x_i^*)}{(p+2)!}.$$

Par le Lemme 2.1 et en posant  $\gamma_{i,j} = \int_{-1}^1 u^{i+j} K^*(u) du$ ,

$$Q_{xy} = h^{p+1} \gamma_{p+1} \frac{F^{(p+1)}(y|x)}{(p+1)!} + o_p(h^{p+1}),$$

et  $R_{xy} = O_p(h^{p+2})$  de sorte que, uniformément pour  $y \in \mathbb{R}$  et  $x \in [a, b]$ ,

$$\text{Biais}_F(\hat{F}_n(y|x)|X_1, \dots, X_n) = \frac{F^{(p+1)}(y|x)}{(p+1)!} h^{p+1} \gamma_{p+1} + o_p(h^{p+1}).$$

Le terme dominant de ce biais conditionnel est indépendant des  $X_i$  ce qui permet de l'utiliser directement pour approximer l'expression asymptotique de la MISE. Maintenant, si on pose  $H = \text{diag}(1, \dots, h^p)$ , on a aussi

$$\text{Var}_F(\hat{F}_n(y|x)|X_1, \dots, X_n) = \frac{1}{nh} e_0^T S_n^{-1}(x) S_n^*(x, y) S_n^{-1}(x) e_0,$$

où  $S_n(x) = H^{-1}(\frac{1}{nh} X^T P X) H^{-1}$  et où  $S_n^*(x, y)$  est la matrice dont l'élément  $(i, j)_{1 \leq i, j \leq p+1}$  est  $S_n^*(x, y)_{i,j} = \frac{1}{nh} \sum_{i=1}^n K^2(\frac{X_i - x}{h})(\frac{X_i - x}{h})^{i+j-2} [F(y|X_i)(1 - F(y|X_i))]$ . De plus,

$$S_n^{-1}(x) = f^{-1}(x) S^{-1} + o_p(1),$$

et

$$S_n^*(x, y) = f(x) F(y|x) (1 - F(y|x)) S^* + o_p(1),$$

où  $S^*$  est la matrice dont les éléments sont les  $(\int_{-1}^1 u^{i+j} K^2(u) du)_{0 \leq i, j \leq p}$ . Ainsi,

$$\text{Var}_F(\hat{F}_n(y|x)|X_1, \dots, X_n) = \left[ \frac{F(y|x)(1 - F(y|x))}{nhf(x)} \right] K^{*(2)}(0) + o_p\left(\frac{1}{nh}\right),$$

puisque  $e_0^T S^{-1} S^* S^{-1} = K^{*(2)}(0)$ , la convolution du noyau  $K^*(\cdot)$  par lui-même. Comme pour le calcul du biais, le terme dominant de cette variance conditionnelle est indépendant des  $X_i$  ce qui permet de l'utiliser pour approximer l'expression asymptotique de la MISE. En injectant ces termes dominants dans l'expression de la MISE pondérée, on est amené à considérer la valeur de  $h$  qui minimise :

$$\int_a^b \int_{\mathbb{R}} \left[ \gamma_{p,1}^2 \left[ \frac{F^{(p+1)}(y|x)}{(p+1)!} \right]^2 h^{2(p+1)} + K^{*(2)}(0) \left[ \frac{F(y|x)(1 - F(y|x))}{nhf(x)} \right] \right] \pi(x, y) F(dy|x) dx.$$

Quelques manipulations algébriques montrent que :

$$h_{opt} = C_p(K) \left[ \frac{\int_a^b \int_{\mathbb{R}} \left[ \frac{F(y|x)(1 - F(y|x))}{f(x)} \right] \pi(x, y) F(dy|x) dx}{\int_a^b \int_{\mathbb{R}} [F^{(p+1)}(y|x)]^2 \pi(x, y) F(dy|x) dx} \right]^{\frac{1}{2p+3}} n^{-\frac{1}{2p+3}},$$

où  $C_p(K) = \left[ \frac{(p+1)!K^{*(2)}(0)}{2(p+1)\gamma_{p,1}^2} \right]^{\frac{1}{2p+3}}$ . Cette fenêtre optimale (Fan et Gijbels [2], Prewitt [7]) n'est pas directement utilisable en pratique en raison des paramètres inconnus dont elle dépend.  $f(\cdot)$  peut être estimée par noyau classique. Les autres termes peuvent l'être via un estimateur pilote issu d'une autre régression polynomiale locale d'ordre  $q \geq p+1$  puisque  $\hat{\beta}_0$  et  $\hat{\beta}_{p+1}$  estiment  $F(y|x)$  et  $F^{(p+1)}(y|x)$  (sous des conditions de régularité adaptées). Pour obtenir cet estimateur pilote, il est nécessaire de préciser une fenêtre  $h^*$ . Heureusement, pour bon nombre de cas importants, ce choix peut être fait sans user d'un soin extrême car le terme entre crochets s'avère assez stable. Par exemple si  $p=1$ ,  $a=0$ ,  $b=1$ ,  $\pi(x, y) = \pi(x)$  intégrant à 1, et dans le cas important de la régression linéaire où  $F(y|x) = F((y - \alpha_0 - \alpha_1 x)/\sigma)$ , le terme entre crochets à la puissance  $1/5$  varie de  $1.4(d\sigma^4/\alpha_1^4)^{\frac{1}{5}}$  pour la loi Normale à  $1.6(d\sigma^4/\alpha_1^4)^{\frac{1}{5}}$  pour la Cauchy, où  $d = \int_0^1 \pi(x) f^{-1}(x) dx$ . Des résultats similaires tiennent pour le cas de la régression polynomiale.

## Références

- [1] K. Doksum, D. Peterson, A. Samarov, On variable bandwidth selection in local polynomial regression, *J. R. Stat. Soc. Ser. B Stat. Methodol.* 62 (3) (2000) 431–448.
- [2] J. Fan, I. Gijbels, *Local Polynomial Modelling and its Applications*, Chapman & Hall, London, 1996.
- [3] S. Ferrigno, G.R. Ducharme, Un test d'adéquation global pour la fonction de répartition conditionnelle, *C. R. Acad. Sci. Paris, Ser. I* 341 (2005) 313–316.
- [4] A. Gannoun, S. Girard, C. Guinot, J. Saracco, Reference curves based on nonparametric quantile regression, *Statistics in Medicine* 21 (2002) 3119–3155.
- [5] P. Hall, B.U. Park, Bandwidth choice for local polynomial estimation of smooth boundaries, *J. Multivariate Anal.* 91 (2) (2004) 240–261.
- [6] L.-S. Huang, J. Fan, Nonparametric estimation of quadratic regression functionals, *Bernoulli* 5 (5) (1999) 927–949.
- [7] K.A. Prewitt, Efficient bandwidth selection in non-parametric regression, *Scandinavian J. Statist.* 30 (2003) 75–92.