



**HAL**  
open science

# Estimation localement suroptimale de la densité par projection tronquée

Jean-Baptiste Aubin

► **To cite this version:**

Jean-Baptiste Aubin. Estimation localement suroptimale de la densité par projection tronquée. Annales de l'ISUP, 2005, XLIX (2-3), pp.3-17. hal-03640184

**HAL Id: hal-03640184**

**<https://hal.science/hal-03640184v1>**

Submitted on 13 Apr 2022

**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.

Pub. Inst. Stat. Univ. Paris  
 XLIX, fasc. 2-3, 2005, 3 à 17

## ESTIMATION LOCALEMENT SUROPTIMALE DE LA DENSITÉ PAR PROJECTION TRONQUÉE

Jean-Baptiste Aubin

L.S.T.A, Université Paris VI, 175, rue du Chevaleret, 75013 Paris, FRANCE,

e-mail: jbaubin@ccr.jussieu.fr

### RÉSUMÉ

On compare les performances de l'estimateur localement suroptimal introduit par Bosq [1] à celles de deux autres estimateurs par projection à l'aide de simulations. En premier lieu, on détermine empiriquement l'influence d'un paramètre de l'estimateur étudié sur le choix de l'indice de troncature et plus généralement la qualité de l'estimation. Ensuite, on effectue les comparaisons entre estimateurs. Celles-ci sont notamment limitées au cas où la base de projection est trigonométrique. Elles se basent sur les erreurs  $L_1$ ,  $L_2$  et  $L_\infty$ . Les simulations ont montré des disparités locales entre les estimateurs en cohérence avec la théorie et apporté un éclairage nouveau sur le comportement à distance finie de l'estimateur étudié.

Mots-clés: *Estimation de la densité par projection, suroptimalité locale, choix de l'indice de troncature.*

### ABSTRACT

By Monte-Carlo experiments, we study a truncated version of the density projection estimator. The choice of the truncation index is important for this kind of estimator. Otherwise, this choice is particularly function of the data and  $c$ , a constant chosen by the statistician. The aim of the study is twice: first we precise the influence of the constant  $c$  on the truncation index, then, two others projection estimators respectively defined in [2] and [3] are compared to the one introduced in [1]. Criterion of comparison are uniform error,  $L_1$  error, and mean integrated square error. Moreover, functions to estimate are taken in and out of the space of superoptimality of our estimator.

Key words: *density projection estimator, local superoptimality, choice of truncation index.*

## 1 Introduction

Bosq [1] a introduit un estimateur de la densité par projection localement suroptimal. Plus précisément, cet estimateur atteint asymptotiquement un taux de convergence pour l'erreur quadratique intégrée de l'ordre de  $\frac{1}{n}$  sur un ensemble  $\mathcal{F}_0$  dense dans l'ensemble  $\mathcal{F}$  des densités considérées.

Les aspects abordés ici concernent le comportement de cet estimateur à distance finie; nous nous intéressons à l'influence d'un de ses paramètres sur le choix de la dimension du sous-espace de projection ainsi que la comparaison à d'autres estimateurs par projection plus classiques.

Nous rappelons d'abord la méthode utilisée pour construire l'estimateur en question, ensuite, nous décrivons le plan de simulations adopté. Enfin, nous examinons les résultats obtenus avant d'en tirer quelques conclusions.

## 2 Rappels Théoriques

Une méthode alternative à celle de l'estimation de la densité par la méthode du noyau est celle de l'estimation par projection. Elle consiste à projeter un paramètre fonctionnel de dimension a priori infinie sur un sous espace de dimension finie et d'estimer cette projection par la méthode des moments.

Concrètement, on dispose de  $(X_1, \dots, X_n)$  une suite de v.a. i.i.d. définies sur l'espace probabilisé  $(\Omega, \mathcal{A}, P)$  et à valeurs dans un espace probabilisé  $(E, \mathcal{B}, \mu)$ . On suppose que  $X_1$  admet une densité  $f$  par rapport à  $\mu$ , mesure de probabilité laissée au choix du statisticien, telle que  $f \in \mathcal{F}$  avec  $\mathcal{F} \subseteq L^2(\mu)$ . L'espace de Hilbert  $(L^2(\mu), \|\cdot\|)$  est supposé séparable, de dimension infinie, et muni d'une base orthonormale, appelée par la suite base de projection,  $(e_j)_{j \in \mathbb{N}^*}$  où  $e_1 = 1$ .

On se limite au cas où  $(e_j)_{j \in \mathbb{N}^*}$  est uniformément bornée. L'estimateur par projection basique de  $f = \sum_{j=1}^{\infty} a_j e_j$  s'écrit  $f_n = \sum_{j=1}^{k_n} \hat{a}_{j_n} e_j$  avec  $\hat{a}_{j_n} = \frac{1}{n} \sum_{i=1}^n e_j(X_i)$  (c.f. [3]) alors que l'estimateur par projection tronquée est défini par

$$\hat{f}_n = \sum_{j=1}^{\hat{k}_n} \hat{a}_{j_n} e_j$$

où  $\hat{k}_n = \max\{j : 1 \leq j \leq k_n, |\hat{a}_{j_n}| \geq \gamma_n\}$ . Comme dans [1], nous prenons

$\gamma_n = \min \left( c \sqrt{\frac{\log n}{n}}, 1 \right)$ ,  $c > 0$  et la suite d'entiers  $(k_n)_{n \in \mathbb{N}}$  telle que  $k_n \rightarrow \infty$ , et  $\frac{k_n}{n} \rightarrow 0$  lorsque  $n \rightarrow \infty$ .

Dans la suite, on distingue les densités qui admettent un développement fini par rapport à la base de projection des autres:

$$\mathcal{F}_0 = \bigcup_{K=1}^{\infty} \mathcal{F}_0(K) \quad \text{où} \quad \mathcal{F}_0(K) = \left\{ f \in L^2(\mu) : f = \sum_{j=1}^K a_j e_j, \quad a_K \neq 0 \right\}$$

et  $\mathcal{F}_1 = L^2(\mu) \setminus \mathcal{F}_0$

### 3 Les Simulations

Un des deux objectifs principaux de cette étude est de préciser l'influence de  $c$  sur l'indice de troncature  $\hat{k}_n$ . Asymptotiquement, Bosq a démontré que, lorsque  $f \in \mathcal{F}_0(K)$  (c'est-à-dire admet un développement fini d'ordre  $K$  par rapport à la base de projection)  $\hat{k}_n$  converge vers  $K$  presque sûrement. Dans le cadre de notre étude, pour une densité  $f$  appartenant à  $\mathcal{F}_0(K)$  préalablement établie, il s'agit de donner un ordre de grandeur pour  $c$  permettant à l'indice de troncature de se fixer le plus vite possible sur la "vraie" valeur  $K$ .

L'autre objectif est la comparaison en termes d'erreurs  $L_1$ ,  $L_2$  et  $L_\infty$  de notre estimateur à deux autres estimateurs par projection définis notamment dans [2]:

- un estimateur basique (pour lequel l'indice de troncature est choisi égal à  $k_n$ ).
- un estimateur classique (pour lequel l'indice de troncature choisi est inférieur à  $k_n$  et minimise une estimation de l'erreur  $L_2$  commise).

Les simulations effectuées ici l'ont été à l'aide du logiciel R version 1.5.1..

Dans un soucis de simplicité, les exemples considérés ont été restreints aux cas suivants:

### 3.1 Facteurs Contrôlés

Les facteurs sont les suivants:

1. L'espace  $\mathcal{F}$  des densités considérées: nous prenons  $L^2(\mu)$  où  $\mu = \mathcal{U}_{[-\pi, \pi]}$
2. La base de projection:

La base de projection est prise trigonométrique sur  $[-\pi, \pi]$ :

$$\begin{cases} e_1(t) = 1_{[-\pi, \pi]}(x), & \text{et } \forall k \in \mathbb{N}^* \\ e_{2k}(t) = \sqrt{2} \cos(kt) 1_{[-\pi, \pi]}(x) \\ e_{2k+1}(t) = \sqrt{2} \sin(kt) 1_{[-\pi, \pi]}(x) \end{cases}$$

3. La dimension maximale de projection  $k_n$ , égale à la partie entière de  $2 \log(n)$ .

*Remarque: Une vitesse de ce type -lente- a des inconvénients sur  $\mathcal{F}_0$  en augmentant le risque que  $K$  soit supérieur à  $k_n$  et ainsi de ne pas chercher la bonne dimension où elle se trouve. De même sur  $\mathcal{F}_1$ , une vitesse plus importante permettrait de projeter sur plus de composantes et ainsi d'améliorer la qualité de l'estimation. Cependant, un tel choix permet d'obtenir un estimateur plus simple et un gain important en terme de temps de calcul.*

4. La taille des échantillons

Elle est fixée successivement à 100 puis à 200 afin d'étudier de quelle façon  $\hat{k}_n$  tend vers  $K$ , toujours selon  $c$ .

5. les densités à estimer:

- Sur  $\mathcal{F}_0$ , les densités  $f_j$  à estimer sont prises dans  $\mathcal{F}_0(j)$ :

$$f_1(x) = e_1(x)$$

$$f_j(x) = e_1(x) + \frac{1}{\sqrt{2}} e_j(x), \text{ pour } j \in [2; 10] \cap \mathbb{N}.$$

- Sur  $\mathcal{F}_1$ , les densités à estimer sont tirées de [4] et sont pour trois d'entre elles des normales ou mélanges de normales et pour la quatrième une densité monotone:

- La normale:  $\mathcal{N}(0; 0, 3\pi)$ .

- La bimodale:  $\mathcal{N}(-0, 2\pi; 0, 24\pi) + \mathcal{N}(0, 4\pi; 0, 16\pi)$ .
- La delta:  $\mathcal{N}(0; 0, 04\pi)$
- La monotone: admet la densité  $f(x) = (\frac{x}{2\pi^2} + \frac{1}{2\pi})1_{[-\pi; \pi]}(x)$

Remarque: A support sur  $\mathbb{R}$ , les trois premières densités, de part leur structure, peuvent être assimilées en première approximation à des densités à support sur  $[-\pi, \pi]$ . De plus, en toute rigueur, seules la bimodale et la monotone sont des éléments de  $\mathcal{F}_1$  pour la base considérée. Cependant, l'ordre du développement pour la normale et la delta étant plus élevés que les  $k_n$  pour les  $n$  considérés, nous assimilerons celles-ci dans notre étude à des éléments de  $\mathcal{F}_1$ .

6.  $c$  a été étudié entre 0 et 5.

L'intervalle d'étude  $[0; 5]$  est légitimé par les considérations suivantes:

D'après la définition de  $\hat{k}_n$ , il vient que si l'on veut conserver, à  $n$  donné, les composantes de la base de projection dont les coefficients  $a_j$  sont tels que  $|a_j| > c'$ , alors il convient de prendre  $c \leq c' \sqrt{\frac{n}{\log n}}$ .

Dans le cas d'une densité de la forme  $f_{K_0}$ , il apparaît que si  $1 \geq c' \geq \frac{1}{\sqrt{2}}$ , alors on négligera souvent le  $K_0$ ième coefficient, conduisant à une estimation de la forme  $\hat{f}_n = \overline{X}_n$ .

A l'inverse, un  $c$  trop petit fera naître une sensibilité de  $\hat{k}_n$  aux variations des  $|\hat{a}_{j,n}|$ ,  $j > K_0$  et aura pour conséquence l'obtention d'un  $\hat{k}_n$  strictement supérieur à  $K_0$ .

Il semble donc raisonnable de chercher  $c$  juste inférieur à  $\sqrt{\frac{n}{2 \log n}}$ , soit respectivement pour  $n = 100$  et  $n = 200$ , à 3,3 et 4,3. La volonté d'observer également la décroissance de  $\hat{k}_n$  a conduit à choisir l'intervalle  $[0; 5]$ .

### 3.2 Le plan de Simulation

Afin d'étudier les conséquences du choix de  $c$  et de comparer les estimateurs, on fait varier  $c$ ,  $n$  et les densités soit  $16 \times 2 \times 14 = 448$  combinaisons différentes. Pour chacune d'elles, on a calculé pour les trois estimateurs obtenus:

- l'indice de troncature
- l'erreur  $L_1$
- l'erreur  $L_2$
- l'erreur  $L_\infty$

Dans le but de réduire la variabilité résultant de l'échantillonnage, ces calculs ont été répétés 10 fois créant ainsi un vecteur-indice à 10 composantes et une matrice-erreur  $10 \times 3$ .

N.B: pour les deux estimateurs classique et basique, on a en fait seulement 28 combinaisons différentes car leur comportement est indépendant de  $c$ .

## 4 Optimisation du choix de $c$

Les résultats se présentent sous la forme de lignes de niveau. Pour une densité à estimer dans  $\mathcal{F}_0(K)$ , on regarde, pour des valeurs de  $c$  et de  $n$  fixées, la moyenne et la variance des composantes du vecteur-indice correspondant.

Remarque: *L'étude analogue sur  $\mathcal{F}_1$  est inutile car le meilleur indice de troncature est le plus grand et est par définition obtenu pour  $c = 0$ .*

Les figures 1 et 2 donnent ainsi les moyennes des indices pour les  $c$  de 0 à 5 en abscisse et les  $K$  de 1 à 10 en ordonnée respectivement pour  $n = 100$  et  $n = 200$ , alors que les figures 3 et 4 représentent les variances correspondantes:

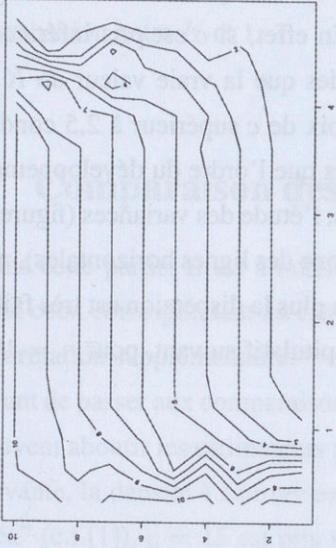


Fig.2 : Moyennes des  $\hat{k}_n$  pour  $n = 200$

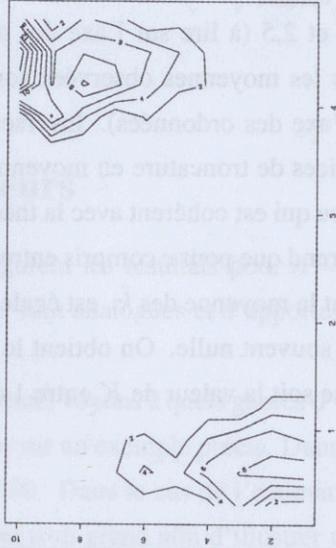


Fig.4 : Variance des  $\hat{k}_n$  pour  $n = 200$

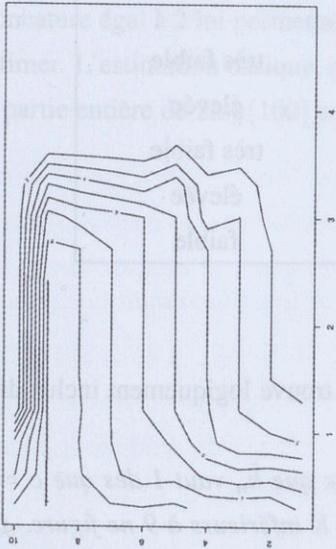


Fig.1 : Moyennes des  $\hat{k}_n$  pour  $n = 100$

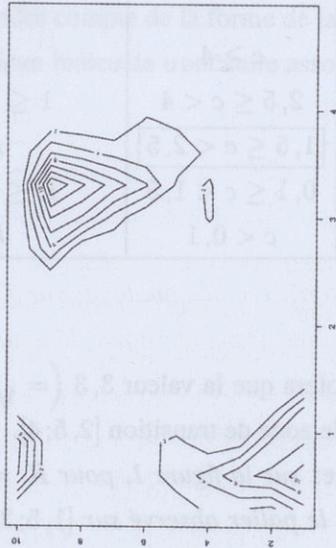


Fig.3 : Variance des  $\hat{k}_n$  pour  $n = 100$

Remarque: l'aspect de ces tableaux ne doit pas faire oublier qu'en ordonnée,  $K$  prend des valeurs entières.

Au vu de la figure 1, il apparaît qu'un choix pertinent de  $c$  pour  $n = 100$  se situe entre 1,5 et 2,5 (à lire sur l'axe des abscisses). En effet, si  $c$  est pris inférieur à 1,5, alors les moyennes observées sont plus grandes que la vraie valeur de  $K$  (à lire sur l'axe des ordonnées). Inversement, un choix de  $c$  supérieur à 2,5 conduit à des indices de troncature en moyenne plus faibles que l'ordre du développement attendu, ce qui est cohérent avec la théorie. De plus, l'étude des variances (figure 3) nous apprend que pour  $c$  compris entre 1,5 et 2,5 (zone des lignes horizontales), non seulement la moyenne des  $\hat{k}_n$  est égale à  $K$ , mais en plus la dispersion est très faible et même souvent nulle. On obtient le tableau récapitulatif suivant, pour  $n = 100$ , quelle que soit la valeur de  $K$  entre 1 et 9:

	Moyenne des $\hat{k}_n$	Variance des $\hat{k}_n$
$c \geq 4$	1	très faible
$2,5 \leq c < 4$	$1 \leq \leq K$	élevée
$1,5 \leq c < 2,5$	$K$	très faible
$0,1 \leq c < 1,5$	$K \leq \leq k_n$	élevée
$c < 0,1$	$k_n$	faible

On notera que la valeur  $3,3 \left( = \sqrt{\frac{100}{2 \log 100}} \right)$  se trouve logiquement inclus dans la seconde zone de transition  $[2,5; 4]$ .

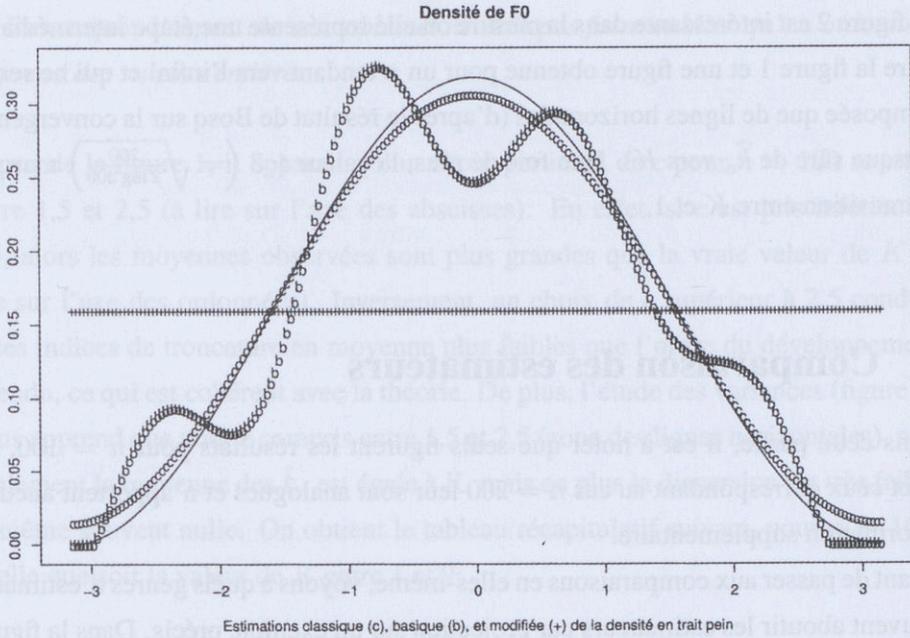
Remarque: sur la figure 1, pour  $K = 10$ , on note que  $\hat{k}_n$  vaut 1 dès que  $c = 1$  sans que le palier observé sur  $[1,5; 2,5]$  pour les  $K$  inférieurs à 9 ne figure. Cet "effondrement" s'explique par le fait que  $k_{100} = 9$  et que par construction,  $\hat{k}_n \leq 9$ .  $\hat{k}_n$  se fixe sur 1 rapidement puisque pour  $f_{10}$ , les  $a_j$  sont nuls pour tout  $j$  entre 2 et 9. Ce phénomène ne se retrouve pas sur la figure 2 car alors  $k_{200} > 10$ .

La figure 2 est intéressante dans la mesure où elle représente une étape intermédiaire entre la figure 1 et une figure obtenue pour un  $n$  tendant vers l'infini et qui ne serait composée que de lignes horizontales (d'après le résultat de Bosq sur la convergence presque sûre de  $\hat{k}_n$  vers  $K$ ). Une fois de plus, la valeur  $4,3 \left( = \sqrt{\frac{200}{2 \log 200}} \right)$  marque la transition entre  $K$  et 1.

## 5 Comparaison des estimateurs

Dans cette partie, il est à noter que seuls figurent les résultats pour  $n = 100$ , en effet ceux correspondant au cas  $n = 200$  leur sont analogues et n'apportent aucune information supplémentaire.

Avant de passer aux comparaisons en elles-mêmes, voyons à quels genres d'estimation peuvent aboutir les estimateurs par projection sur un exemple précis. Dans la figure suivante, la densité à estimer est  $f_2$  et  $n = 100$ . Dans le cas de l'estimation "modifiée" (c.f.[1]),  $c = 15$  est pris volontairement trop grand afin d'illustrer un cas où l'indice de troncature est trop petit. L'estimation classique (c.f.[2]) a ici un indice de troncature égal à 2 lui permettant de bien rendre compte de la forme de la densité à estimer. L'estimation basique, quant à elle, a un indice de troncature associé valant la partie entière de  $2 \log(100)$  soit 9:



On peut voir dans ce graphe l'illustration que l'idée répandue selon laquelle la qualité de l'estimation croît avec la dimension  $\hat{k}_n$  du sous espace de projection est fautive sur  $\mathcal{F}_0(K)$  dès que  $\hat{k}_n$  est strictement supérieur à  $K$ . Ceci s'explique par l'accumulation des erreurs d'estimation des coefficients de Fourier.

## 5.1 Sur $\mathcal{F}_0$

Les résultats se présentent sous la forme d'un tableau. Pour chaque  $c$ , type d'erreur et densité de  $\mathcal{F}_0$ , on observe si la moyenne de l'erreur commise par notre estimateur est plus faible que celle commise par l'estimateur comparé. Si cela est vérifié, on dira qu'on remporte une "victoire" sur l'autre estimateur.

Plutôt que d'exposer le tableau avec les 25 valeurs de  $c$  entre 0 et 5, nous nous sommes servi des résultats précédents afin d'exhiber 3 classes de valeurs de  $c$  dans lesquelles les résultats sont plus homogènes. Voici le tableau de synthèse en tenant compte des 10 densités de  $\mathcal{F}_0$  considérées pour chaque estimateur de comparaison: La table 1 rend compte de la supériorité de l'estimateur étudié sur l'estimateur basique et ce quels que soient le type d'erreur et  $c$ .

A l'inverse de la Table 1, la Table 2 laisse apparaître des disparités importantes de

Choix de $c$	Pour l'erreur $L_1$	Pour l'erreur $L_2$	Pour l'erreur $L_\infty$
$c > 3$	7,3	8,9	8,8
$3 \geq c \geq 2,2$	8,3	9,3	9,1
$2,2 > c$	7,7	8,9	9

Table 1: Nombre moyen de "victoires" contre l'estimateur basique (sur 10)

Choix de $c$	Pour l'erreur $L_1$	Pour l'erreur $L_2$	Pour l'erreur $L_\infty$
$c > 3$	2,9	3	2,5
$3 \geq c \geq 2,2$	6	6,7	7
$2,2 > c$	3,1	3,2	2,7

Table 2: Nombre moyen de "victoires" contre l'estimateur classique (sur 10)

comportement selon la valeur de  $c$ . Ainsi, seul un  $c$  choisi entre 2,2 et 3 permet à l'estimateur par troncature d'être en moyenne pour les erreurs considérées meilleur que l'estimateur classique.

## 5.2 Sur $\mathcal{F}_1$

Les résultats, s'il sont analogues à ceux de la section précédente (un vecteur indice et une matrice erreur pour chaque combinaison), sont exposés différemment. Au lieu de faire la moyenne des erreurs et des indices sur les densité, nous analysons individuellement les résultats.

Notons par ailleurs que l'estimateur basique devrait donner de meilleurs résultats dans cette section car, par construction, son indice de troncature est plus élevé que ceux des deux autres estimateurs. Cependant, les résultats de nos simulations n'ont pas permis de noter une différence significative entre les estimateurs basique et classique. De même, lorsque  $c$  est petit (nous verrons par la suite ce que nous entendons par "petit" suivant chaque densité), l'estimateur étudié donne des résultats très voisins de ceux obtenus pour les deux autres estimateurs. Il est à noter que pour  $c = 0$ , celui-ci est identique à l'estimateur basique.

Les Tables suivantes donne pour chaque densité, le comportement simplifié de l'estimateur étudié en fonction de  $c$  pour les erreurs  $L_2$  et  $L_\infty$ . Le comportement des deux autres estimateurs peut être déduit de ces tables en considérant qu'ils ont les mêmes erreurs que celles obtenues pour notre estimateur à  $c = 0$ .

$c$	0	1,5	4
Normale	0,001	0,001	↗ 0,015
Bimodale	0,002	↗ 0,01	0,01
Delta	0,001		0,001

Table 3: Erreur  $L_2$  pour l'estimateur étudié en fonction de  $c$

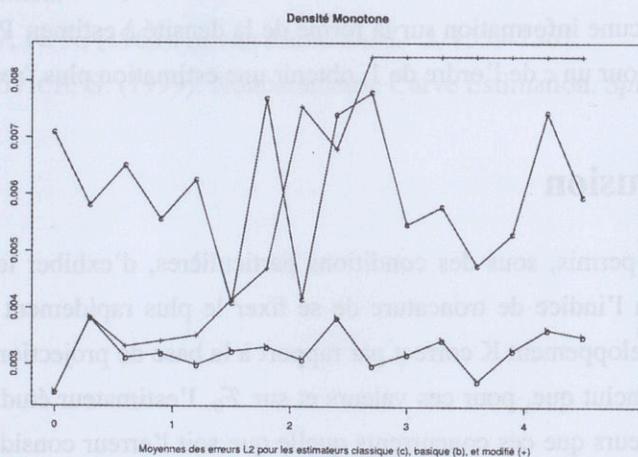
$c$	0	1,5	4
Normale	0,01	0,01	↗ 0,2
Bimodale	0,1	↗ 0,3	0,3
Delta	0,01		0,01

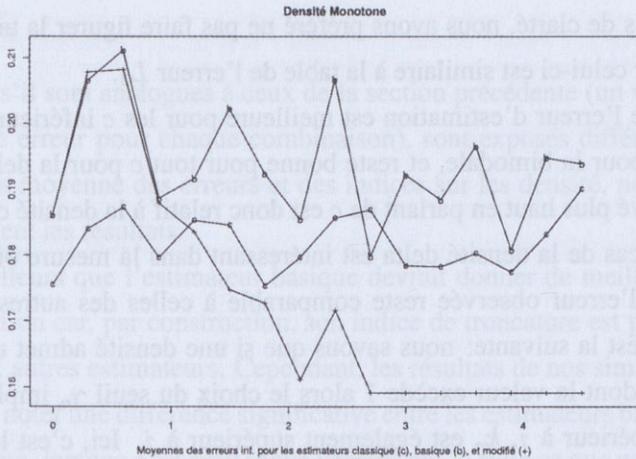
Table 4: Erreur  $L_\infty$  pour l'estimateur étudié en fonction de  $c$

Dans un soucis de clarté, nous avons préféré ne pas faire figurer la table relative à l'erreur  $L_1$  car celui-ci est similaire à la table de l'erreur  $L_2$ .

Il apparaît que l'erreur d'estimation est meilleure pour les  $c$  inférieurs à 4 pour la normale, 1,5 pour la bimodale, et reste bonne pour tout  $c$  pour la delta. Le terme "petit" employé plus haut en parlant de  $c$  est donc relatif à la densité considérée. A ce propos, le cas de la densité delta est intéressant dans la mesure où même pour des  $c$  élevés, l'erreur observée reste comparable à celles des autres estimateurs. L'explication est la suivante: nous savons que si une densité admet un coefficient de Fourier  $a_j$  dont la valeur excède 1 alors le choix du seuil  $\gamma_n$  implique que dès que  $k_n$  est supérieur à  $j$ ,  $\hat{k}_n$  est également supérieur à  $j$ . Ici, c'est le cas pour le huitième coefficient de Fourier de la densité delta, ce qui permet à l'estimation de rester correcte même à  $c$  élevé.

La dernière densité non présente dans les tables précédentes est la densité monotone. Voici le comportement des erreurs  $L_2$  et  $L_\infty$  commises pour les trois estimateurs basique, classique et modifié (ou étudié) en fonction de  $c$ :





Ce cas illustre l'intérêt de considérer plus d'un type d'erreur. En effet, les comportements de ces deux erreurs sont inverses; l'erreur  $L_2$  croît avec  $c$ , l'erreur  $L_\infty$  décroît. De plus, dès que  $c$  est supérieur à 3, celles-ci se stabilisent (en fait, l'indice de troncature s'est fixé sur 1). Il convient donc d'être prudent quant au choix d'un critère de sélection de  $c$ , car certains donnent des conclusions différentes. Ici, un  $c$  trop grand est à proscrire car, même si l'erreur  $L_\infty$  est la plus faible, l'estimation ne donnera aucune information sur la forme de la densité à estimer. Par contre, on peut espérer, pour un  $c$  de l'ordre de 1, obtenir une estimation plus "ressemblante".

## 6 Conclusion

Cette étude a permis, sous des conditions particulières, d'exhiber les valeurs de  $c$  permettant à l'indice de troncature de se fixer le plus rapidement possible sur l'ordre de développement  $K$  correct par rapport à la base de projection choisie. De plus, elle a conclu que, pour ces valeurs et sur  $\mathcal{F}_0$ , l'estimateur étudié donne des résultats meilleurs que ces concurrents quelle que soit l'erreur considérée même à distance finie. Sur  $\mathcal{F}_1$ , l'estimateur étudié fait aussi bien que les autres pour des  $c$  assez faibles, mais nous avons montré que cette notion était relative aux densités à estimer.

Cependant, il ne faut pas oublier que cette étude porte sur un nombre restreint de densités et a été menée pour une base de projection, une suite  $(k_n)_{n \in \mathbb{N}}$  et des tailles

d'échantillon spécifiques. L'aspect parcellaire des résultats est donc important et appelle d'autres simulations. Ainsi, tous les paramètres cités plus haut mais également la mesure de probabilité  $\mu$  peuvent être modifiés pour donner une vue plus large du comportement à distance finie de l'estimateur en question. Outre les travaux pour continuer d'étudier le comportement de l'indice de troncature pour d'autres choix de bases, de  $k_n$ , ou de densités à estimer, d'autres simulations peuvent être envisagées afin de confronter les résultats de cet estimateur à d'autres estimateurs par projection.

Par ailleurs, un travail d'affaiblissement des conditions (notamment pour pouvoir considérer des bases non uniformément bornées) est en cours, ainsi que l'application de cette technique à des problèmes d'estimation de la régression.

## Bibliographie

- [1] BOSQ, D. (2002). Estimation localement suroptimale et adaptative de la densité. *C.R. Acad. Sci. Paris, Ser.I* **334** 591-595.
- [2] BOSQ, D. et Lecoutre, J.P. (1987). Théorie de l'Estimation Fonctionnelle. *Editions Economica*.
- [3] CENCOV, N. N. (1962). *Soviet Math. Dokl.* **3**, 1559-1562.
- [4] EFROMOVICH, S. (1999). Nonparametric Curve Estimation. *Springer, N.Y.*

*Key words and phrases:* Optimal and asymptotically optimal stepwise regression, optimal linear estimator, simple linear regression, nonparametric estimation.