Université de Montréal

Estimation non-paramétrique de la fonction de répartition et de la densité

par

Mohammed HADDOU

Département de mathématiques et de statistique
Faculté des arts et des sciences

Thèse présentée à la Faculté des études supérieures
en vue de l'obtention du grade de
Philosophiæ Doctor (Ph.D.)
en statistique

mars 2007

Université de Montréal

Faculté des études supérieures

Cette thèse intitulée

Estimation non-paramétrique de la fonction de répartition et de la densité

présentée par

Mohammed HADDOU

a été évaluée par un jury composé des personnes suivantes :

Martin Bilodeau
(président-rapporteur)

François Perron
(directeur de recherche)

Bruno Rémillard
(membre du jury)

Luc Devroye
(examinateur externe)

Pierre Poulin
(représentant du doyen)

Thèse acceptée le:
05 mars 2007

SOMMAIRE

Cette thèse porte sur l'estimation non-paramétrique de la fonction de répartition (f.r.) et de la densité. Dans le premier essai, nous proposons une nouvelle méthode d'estimation adaptative de la f.r. Nous sommes capable de contrôler la distance, dans le sens de la norme du supremum, de l'estimateur à la fonction de répartition échantillonnale. Ceci nous permet d'obtenir tous les résultats asymptotiques de cette dernière sous les mêmes conditions *minimes* de régularité. L'estimateur proposé est plus lisse, dépend de trois méta-paramètres dont une fonction (instrumentale). Nous pensons que cette dernière est propre à notre méthode. Elle permet d'inclure de l'information *a priori* sur la fonction cible et ce faisant contribue à l'amélioration de l'estimation.

Le second essai traite de l'estimation de la fonction densité. L'estimateur proposé s'obtient en dérivant l'estimateur pour la f.r. proposé dans le premier essai. Cet estimateur consiste en une combinaison convexe (finie) de densités dont les supports sont déterminés de manière aléatoire par les espacements des statistiques d'ordre. Dans une certaine mesure, notre estimateur est une généralisation de l'estimateur (histogramme) de la partition aléatoire et de l'estimateur connu sous le nom de "histo-spline".

Mots Clés

Estimation non-paramétrique, fonction de répartition, densité, lissage adaptatif, splines, convergence uniforme.

SUMMARY

This thesis is about non-parametric estimation of the cumulative distribution function (cdf) and the density function. In the first work, we propose a new adaptive method for estimating a cdf. In the supnorm, we are able to control the distance of the estimator to the empirical distribution function (edf). This allows us to achieve the same asymptotic results like those obtained using the edf and this is done under the same *minimum* regularity conditions. The proposed estimator is, however, smoother and depends on three parameters of which an instrumental function H. This function allows us to include prior information about the target density function and therefore helps improve the estimation.

The second work deals with non-parametric density estimation. The proposed estimator is obtained by differentiating the estimator for the cdf proposed in the first work. This estimator consists of a finite convex combination of densities with supports that are randomly determined by the spacings of the order statistics. In a certain way, the proposed estimator is a generalization of the histogram with random partition and the histo-spline.

Key Words

Non-parametric estimation, cumulative distribution function, density function, adaptive smoothing, splines, uniform convergence.

DÉDICACE

A mes parents.

REMERCIEMENTS

Je voudrais tout d'abord exprimer ma profonde gratitude à Monsieur François Perron, mon directeur de thèse. Je le remercie pour la qualité de sa direction, sa rigueur scientifique et sa disponibilité.

Je suis très sensible à l'honneur que me font Messieurs Luc Devroye et Bruno Rémillard en acceptant d'être rapporteurs de cette thèse.

Je suis très reconnaissant à Monsieur Martin Bilodeau d'avoir accepté de présider le jury et à Monsieur Pierre Poulin, représentant du doyen.

Je voudrais remercier mon directeur de recherche François Perron, le département de mathématiques et de statistique, la Faculté des Études Supérieures de l'Université de Montréal (FES), le fond CRSNG, l'ISM et le CRM pour leur support financier.

Ma reconnaissance va également aux membres du personnel du département de mathématiques et de statistique pour leur amabilité et serviabilité.

Je remercie l'ensemble des collègues qui m'ont apporté aide, soutien, sympathie et amitié ainsi que tous ceux pour qui le mot science rime encore avec conscience. En particulier, je tiens à remercier mon ami Sévérien Nkurunziza pour son amitié et les bons moments passés ensemble.

Un grand merci au professeur Martin Goldstein pour ses cours, les bonnes discussions qu'on a eues et pour son aide.

Je suis très reconnaissant envers mes parents, ma grande famille et tous mes amis pour leur soutien et leur patience.

Mes remerciements vont évidemment à Eva (ma femme) et à Meriem (ma petite fille) pour leur patience envers Papi !, pour me remonter le moral et me supporter. Tout simplement "*grazie di esistere !*"

Table des matières

Sommaire iii

Summary iv

Dédicace v

Remerciements vi

Introduction 1

0.0.1. Plan 3

0.1. Estimation de la fonction de répartition 4

0.1.1. La fonction de répartition échantillonnale 4

0.1.1.1. Propriétés 4

0.1.2. L'estimateur à noyau 8

0.2. Estimation de la densité 9

0.2.1. L'histogramme 9

0.2.2. Estimateur dit simple ou naïf ("naive estimator") 10

0.2.2.1. Propriétés et remarques 10

0.2.3. L'estimateur à noyau 11

0.2.3.1. Exemples de fonction noyau 11

0.2.3.2. Propriétés immédiates de l'estimateur à noyau 13

0.2.3.3. Choix du méta-paramètre h 14

0.2.3.4. Méthode du noyau adaptatif ou variable 15

Bibliographie 17

Chapitre 1. Estimating the Cdf by a Perturbation of the Edf ... 20

1.1. Introduction 22

1.2. Approximation of functions and properties 26
1.2.1. Approximation and loss 26
1.2.2. The mixture 27
1.2.3. The basis 28
1.2.4. The choice of the nodes and the parameters H and k 30

1.3. Estimation and statistical results 35
1.3.1. Distance of $\hat{F}$ to F_n 36
1.3.2. Uniform convergence of $\hat{F}$ to F (L_∞ convergence) 37
1.3.3. L_p convergence of $\hat{F}$ to F 38
1.3.4. A Cramér-Von Mises like statistic 39
1.3.5. Asymptotic behavior of $\hat{F}$ 40
1.3.6. Some uniform results concerning the bias, variance, MSE, and other quantities 42

1.4. Guidelines on the choice of the parameters k, H, m and Simulations 44
1.4.1. Numerical examples 45

Bibliographie 50

Chapitre 2. Adaptive estimation of a density function using finite mixtures 53

2.1. Introduction 55
2.1.1. Approximation of the density function 56
2.1.2. The mixture 56
2.1.3. The basis 57
2.1.4. The choice of the nodes and the parameters H and k 57

2.2. The approximation step 58

2.3. The estimation step 60

2.4. Simulation study 62

Bibliographie 68

INTRODUCTION

«Dans le petit nombre de choses qu'il sait et qu'il sait bien,
la plus importante est qu'il y en a beaucoup qu'il ignore.»
J.-J. Rousseau, L'Emile, III.

"He uses statistics as a drunken man uses lamp posts -
for support rather than for illumination."
(*«Il utilise les statistiques comme l'ivrogne les lampadaires,*
pour s'appuyer plutôt que pour s'éclairer.»)
Andrew Lang.

Dans cette thèse, on propose une nouvelle méthode d'estimation de la fonction de répartition F et de ses dérivées basée sur un échantillon $X_1, \ldots, X_n$ issu de F. Cette nouvelle méthode se veut une alternative intéressante aux méthodes classiques comme celle du noyau. On adopte pour cela une approche non-paramétrique. La procédure utilisée s'appuie uniquement sur des hypothèses qualitatives sur la fonction à estimer telles que la continuité, le fait d'être Lipschitz, la différentiabilité, etc. Lorsque le contexte ne suggère aucune structure *a priori* sur le modèle, l'approche non-paramétrique apparaît comme la méthode la plus appropriée. Elle nous permet, dans ce cas, d'aller au-delà du cadre paramétrique qui est souvent difficile à justifier en pratique. Cette approche prend d'ailleurs de plus en plus de place en statistique étant donné sa flexibilité et la facilité (relative) à utiliser le calcul intensif sur ordinateur. Cependant, la souplesse de ces méthodes ne s'avère pas sans coût. En effet, les méthodes d'estimation non-paramétriques dépendent, en général, d'un vecteur de paramètres (dits "méta-paramètres") qui contrôle le degré de lissage de l'estimateur non-paramétrique. Le choix de ces méta-paramètres s'avère être crucial quant au résultat (final) de l'estimation surtout dans le cas d'échantillons de taille relativement petite. On peut penser, par exemple, au paramètre de lissage appelé *fenêtre* dans l'estimation par la méthode du noyau. L'analyste cherchera alors, étant donné un certain critère, le (méta-)paramètre *optimal*. Souvent, ce dernier, va dépendre de quantités inconnues comme la fonction qu'on désire estimer (méthode du noyau.) D'autres fois, il s'obtient par des formules qui tirent leurs justifications de résultats asymptotiques ou par validation croisée et donc sujet à l'erreur statistique. Dans la plupart des méthodes non-paramétriques de lissage, on va supposer que la fonction à estimer est lisse et souvent deux fois différentiable. Notre méthode se veut une alternative nettement moins restrictive. Elle permet, par exemple, de bien estimer des fonctions de répartition lisses par morceaux avec un minimum d'hypothèses. En effet, on obtient, par exemple, la convergence L_∞ de l'estimateur en ne requérant de F que le fait d'être une fonction de répartition.

0.0.1. Plan

Deux chapitres constituent cette thèse. **Le chapitre 1** est consacré à l'estimation de la fonction de répartition (f.r.) et est divisé en trois parties. Dans la *première partie*, on propose une méthode d'*approximation* d'une f.r. Certains résultats de cette partie sont indépendants et représentent un intérêt en soi. Dans la *deuxième partie*, on utilise les résultats obtenus pour l'approximation et on les applique à l'*estimation* d'une f.r. Plusieurs résultats à distance finie et asymptotiques sont obtenus. Remarquons qu'on obtient aussi les mêmes résultats (asymptotiques) que ceux obtenus en utilisant la fonction de répartition échantillonnale et ce avec les mêmes hypothèses minimales. Enfin, la *troisième partie* est consacrée aux évaluations numériques où plusieurs simulations sont faites pour visualiser les performances de l'estimateur.

Le chapitre 2 est dédié à l'estimation de la fonction densité. On utilise pour cela l'estimateur de la f.r. proposé au chapitre 1. Comme pour le chapitre 1, ce chapitre se divise en tois parties ; la partie approximation, la partie estimation et une partie simulation. Dans la partie simulation, on compare les performances de l'estimateur à celles obtenues en utilisant la méthode du noyau et ce pour des données fictives et des données réelles.

0.1. Estimation de la fonction de répartition

Dans cette section, on présente deux estimateurs de la f.r. qui ont été, sans doute, les plus étudiés et utilisés en pratique.

0.1.1. La fonction de répartition échantillonnale

Soit X une variable aléatoire (v.a.) réelle de f.r. F, où $F(x) = \Pr(X \leq x) = \mathrm{E}\,[\mathrm{I}(X \leq x)]$. Supposons qu'on dispose d'un échantillon $X_1, \ldots, X_n$ de F. La fonction de répartition échantillonnale (empirique) notée f.r.e. est alors définie, pour tout $x \in \mathbb{R}$, par

$$\begin{aligned} F_n(x) &= \frac{\#\{i\colon X_i \leq x\}}{n} = \frac{1}{n}\sum_{i=1}^{n} I(X_i \leq x) = \frac{1}{n}\sum_{i=1}^{n} I(X_{(i)} \leq x) \\ &= \begin{cases} 0 & \text{si} \quad x < X_{(1)}, \\ \dfrac{k}{n} & \text{si} \quad X_{(k)} \leq x < X_{(k+1)} \qquad k = 1, \ldots, n-1, \\ 1 & \text{si} \quad x \geq X_{(n)}, \end{cases} \end{aligned}$$

La figure 0.1 montre un exemple de f.r.e. pour un échantillon de taille 10. La f.r.e. F_n est une fonction en escalier qui met des sauts de hauteur $1/n$ en chaque observation X_i. Elle indique (caractérise) la position des observations et permet donc de recouvrir l'ensemble des observations (en ignorant l'ordre d'apparition, cependant) d'où son rôle important en statistique.

0.1.1.1. *Propriétés*

Notons d'abord que la v.a. $nF_n(x)$ admet comme distribution la loi binomiale $B(n,\, F(x))$. De sorte que, $F_n(x)$ est un estimateur sans biais de $F(x)$ de variance égale à $F(x)(1-F(x))/n = \mathrm{MSE}\,(F)$. La f.r.e. possède une longue liste de bonnes propriétés statistiques comme le fait d'être efficace au sens minimax (*first order efficient in the minimax sense*) et que $F_n(x)$ est l'unique estimateur sans biais à variance minimale pour $F(x)$ (voir Dvoretzky, Kiefer et Wolfowitz (1956) et

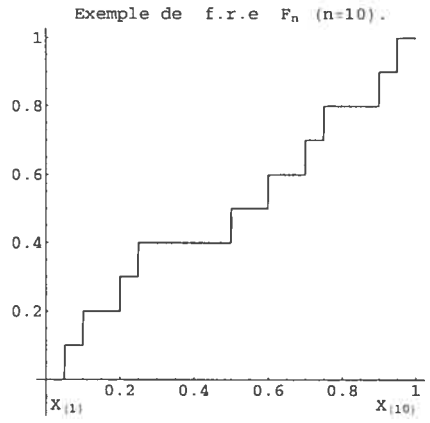


FIG. 0.1. Exemple de f.r.e..

Lehmann et Cassella (1998), chapitre 2). De plus, elle est l'estimateur du maximum de vraisemblance non-paramétrique de F et joue un rôle central dans les méthodes de simulation non-paramétriques et de ré-échantillonnage (voir Efron et Tibshirani (1993) p. 310). Pour une revue de certaines des propriétés de la f.r.e. voir par exemple, Csáki (1984), Stute (1982), Serfling (1980) et Devroye (2001). On trouvera, dans ce qui suit, une liste de quelques propriétés de F_n pertinentes à notre travail. Notons ici que l'estimateur que l'on propose admet ces mêmes propriétés sans aucune condition supplémentaire sur F que celles citées ici.

(1) **Loi des grands nombres.**

$$F_n(x) \xrightarrow[n\to\infty]{p.s.} F(x) \quad \text{pour tout } x \in \mathbb{R}, \quad \text{(p.s.=presque sûrement)}.$$

(2) **Convergence en loi (ponctuelle).**

$$\sqrt{n}\,(F_n(x) - F(x)) \xrightarrow[n\to\infty]{\mathcal{D}} N\,(\,0, F(x)(1-F(x)\,) \quad \text{pour tout } \ x.$$

(3) **Convergence p.s. uniforme (lemme de Glivenko-Cantelli (1933)).**

$$\|F - F_n\|_\infty = \sup_{x\in\mathbb{R}} |F_n(x) - F(x)| \xrightarrow[n\to\infty]{p.s.} 0.$$

(4) **Loi du logarithme itéré (LLI) (vitesse de convergence).**

$$\overline{\lim_n} \|F_n - F\|_\infty \sqrt{\frac{n}{2 \log\log(n)}} = c(F), \quad \text{avec probabilité un,}$$

où $c(F) = \sup_{x \in \mathbb{R}} [F(x)(1 - F(x))]^{1/2}$. Il s'en suit que si F est continue, on obtient $c(F) = 1/2$.

(5) **La propriété dite de Chung-Smirnov (la même que (4)).**

$$\overline{\lim_n} (2n/\log\log n)^{1/2} \|F_n - F\|_\infty \leq 1, \quad \text{avec probabilité un.}$$

(6) **Inégalité en probabilité de type exponentiel.**

Cette inégalité est due à Dvoretzky, Kiefer et Wolfowitz (1956).

$$\Pr\{\|F - F_n\|_\infty > t\} \leq C\, e^{-2nt^2}, \quad \text{pour tout } t > 0.$$

La meilleure constante C est obtenue par Massart (1990) et vaut 2. Voir aussi Devroye (2001).

(7) **La distance de Kolmogorov-Smirnov et convergence.**

La quantité $\|F - F_n\|_\infty$ est appelée distance de Kolmogorov-Smirnov et est notée D_n.

Théorème (Kolmogorov, 1933).

Si F est continue, alors on obtient

$$\lim_{n \to \infty} \Pr(\sqrt{n}\, D_n \leq d) = \sum_{j=-\infty}^{\infty} (-1)^j\, e^{-2j^2 d^2}, \quad d > 0.$$

(8) **La statistique de Cramér-Von Mises.**

La statistique de Cramér-Von Mises est définie par (voir, e.g., Serfling (1980))

$$C_n = n \int_{-\infty}^{\infty} [F_n(x) - F(x)]^2\, dF(x).$$

On a le résultat suivant,

Lemme (Finkelstein (1971)).

Avec probabilité un,

$$\overline{\lim_n} \frac{C_n}{2 \log\log n} = \frac{1}{\pi^2}.$$

Convergence en loi de C_n.

On a le résultat suivant (voir, par exemple, Serfling (1980))

$$\lim_n \Pr(C_n \leq c) = \Pr(Y \leq c), \quad c > 0,$$

où Y est une v.a. qui peut être représentée sous la forme

$$Y = \sum_{j=1}^{\infty} \frac{X_j^2}{j^2 \pi^2}.$$

où les v.a. X_j^2 sont indépendantes de loi khi-deux de degré un.

(9) **Distances de Kolmogorov-Smirnov unilatérales.**

Théorème (Smirnov, 1941.)

Introduisons les deux quantités suivantes :

$$D_n^+ = \sup_x [F_n(x) - F(x)] \quad \text{et} \quad D_n^- = \sup_x [F(x) - F_n(x)].$$

Si F est continue, on obtient

$$\lim_n \Pr(\sqrt{n}\, D_n^+ > d) = \lim_n \Pr(\sqrt{n}\, D_n^- > d) = e^{-2\,d^2}, \quad d > 0.$$

0.1.2. L'estimateur à noyau

Le fait que F_n soit une fonction en escalier même lorsque la distribution sous-jacente est continue, interpelle la nécessité (dans certains domaines d'application) à considérer des estimateurs (plus) lisses pour F. Beaucoup d'estimateurs *lisses* ont été proposés dans la littérature. La plupart de ces estimateurs sont basés sur le lissage de F_n. L'estimateur à noyau, que l'on notera F_h, est peut-être celui qui a été le plus étudié. Il est donné sous la forme suivante

$$F_h(x) = \int_{-\infty}^{x} f_h(t)\, dt = \frac{1}{n} \sum_{i=1}^{n} \bar{K}_h\left(x, X_i\right),$$

où

$$f_h(x) = \frac{1}{nh} \sum_{i=1}^{n} K\left(\frac{x - X_i}{h}\right),\ \bar{K}_h(x, y) = \int_{-\infty}^{x} K_h(t - y)\, dt,\ K_h(x) = K(x/h)/h.$$

La fonction K (fonction de poids) est dite "noyau". Elle est souvent choisie comme une densité de probabilité symétrique. Le paramètre positif h (paramètre de lissage) est appelé "fenêtre". Le paramètre de lissage h a une grande influence sur la performance de l'estimateur. Un h trop petit produit une courbe qui fluctue beaucoup donnant trop de détails. Un choix de h trop grand aura pour résultat une courbe trop lisse qui montre peu de détails. Le choix de h est plus important que celui de K. Notons que F_h admet la réécriture suivante

$$F_h(x) = \int_{-\infty}^{\infty} \bar{K}_h(x - u)\, F_n(du) = \int_{-\infty}^{\infty} K_h(x - u)\, F_n(u)\, du = (K_h * F_n)(x).$$

Cet estimateur est initialement étudié par Nadaraya (1964) qui établit, sous certaines conditions de régularité, sa convergence presque sûre uniforme. Watson et Leadbetter (1964) montrent la normalité asymptotique de F_h, Winter (1979) montre que F_h admet la propriété de Smirnov-Chung. Azzalini (1981) donne une expression asymptotique de l'erreur quadratique moyenne de F_h. Voir aussi Reiss

(1981), Falk (1983), Mammitzsch (1984), Swanepoel (1988), Jones (1990), Shirahata et Chu (1992.) D'un autre côté, Sarda (1993), Altman et Léger (1995) et Chu (1995) traitent du problème du choix du paramètre de lissage h. Pour d'autres résultats, voir aussi Shao et Xiang (1997), Bowman, Hall et Prvan (1998), et Alvarez, Manteiga et Suárez (2000).

0.2. Estimation de la densité

Dans ce qui suit, on présente quelques estimateurs de la densité qui ont été, sans doute, les plus étudiés et utilisés en pratique.

0.2.1. L'histogramme

Pour construire l'histogramme, on fixe une origine t_0 et une largeur de classe $h\,(h > 0.)$ Le support de la densité est alors subdivisé en classes de types

$$C_k = [t_k, t_{k+1}) = [t_0 + kh, t_0 + (k+1)h)\,, \qquad k \in \mathbb{Z}.$$

L'estimateur est alors défini par

$$\begin{aligned} \hat{f}_H(x) &= \frac{1}{n} \frac{\#\{i :\ X_i \text{ est dans la même classe que } x\}}{\text{largeur de la classe contenant } x} \\ &= \frac{1}{nh} \sum_{i=1}^{n} I_{C_k}(X_i), \qquad \text{si} \quad x \in C_k. \end{aligned}$$

Remarques :

- L'histogramme est une fonction étagée, donc discontinue.
- L'utilisation de l'histogramme n'est pas appropriée dans des applications requérant l'emploi de la dérivée de l'estimateur.
- L'histogramme dépend de deux paramètres : le point d'origine t_0 et la largeur des classes h. Le paramètre h contrôle la qualité du lissage.
- Des versions modifiées de $\hat{f}_H$ sont possibles en permettant à h de varier.

0.2.2. Estimateur dit simple ou naïf ("naive estimator")

Partant du fait suivant

$$f(x) = \lim_{h\to 0} \frac{F(x+h) - F(x-h)}{2h} = \lim_{h\to 0} \frac{\Pr(x-h < X \leq x+h)}{2h},$$

un estimateur naturel pour f serait

$$\begin{aligned} \hat{f}_h(x) &= \frac{1}{2h} \frac{\#\{i:\ x-h < X_i \leq x+h\}}{n} = \frac{1}{2nh} \sum_{i=1}^{n} I\left\{-1 \leq \frac{x-X_i}{h} < 1\right\} \\ &= \frac{1}{n} \sum_{i=1}^{n} \frac{1}{h} \omega\left(\frac{x-X_i}{h}\right) = \frac{F_n(x+h) - F_n(x-h)}{2h}, \end{aligned}$$

où la fonction poids ω est donnée par

$$\omega(x) = \frac{1}{2}\, I_{[-1,1)}(x).$$

Cet estimateur est construit en plaçant un plateau de largeur $2h$ et de hauteur $(2nh)^{-1}$ en chaque observation et en sommant. Le paramètre de lissage $h = h_n$ dépend de la taille n de l'échantillon et est tel que $\lim_{n\to\infty} h_n = 0$.

0.2.2.1. *Propriétés et remarques*

- Distribution :

  $2nh\, \hat{f}_h(x) \sim Bin(n, \Delta_h(x)), \quad$ où $\quad \Delta_h(x) = F(x+h) - F(x-h).$

- Biais :

  $$\mathrm{E}\,(\hat{f}_h(x)) = \frac{F(x+h) - F(x-h)}{2h}.$$

  L'estimateur est donc asymptotiquement sans biais.

- Variance :

  $$\begin{aligned} \text{Variance} &= \frac{[F(x+h) - F(x-h)]\,[1 - F(x+h) + F(x-h)]}{4nh^2} \\ &\sim \frac{f(x)}{2}\,\frac{1}{nh}. \end{aligned}$$

- Erreur quadratique moyenne :

$$\mathrm{MSE}\left(\hat{f}_h(x)\right) \xrightarrow[n\to\infty]{} 0 \quad \text{si} \quad h \xrightarrow[n\to\infty]{} 0 \quad \text{et} \quad nh \xrightarrow[n\to\infty]{} \infty.$$

- $\hat{f}_h$ est une fonction en escalier, discontinue aux points $X_i \pm h$.
- Avec $\hat{f}_h$, on n'a plus le problème du point d'origine comme pour l'histogramme.

0.2.3. L'estimateur à noyau

Une généralisation de l'estimateur "naïf" est possible en remplaçant la fonction de poids ω (uniforme sur $[-1, 1)$) par une fonction plus générale pouvant être une densité de probabilité. On obtient alors la réécriture suivante

$$\hat{f}_h(x) = \frac{1}{nh}\sum_{i=1}^{n} K\left(\frac{x - X_i}{h}\right) = \frac{1}{n}\sum_{i=1}^{n} K_h\left(x - X_i\right).$$

La fonction de poids K est dite "noyau". Elle est souvent choisie comme une densité de probabilité symétrique. Le paramètre positif h (paramètre de lissage) est appelé "fenêtre". Le paramètre de lissage h a une grande influence sur la performance de l'estimateur. Un choix de h trop petit résulte en une courbe oscillante donnant trop de détails. Un h trop grand résulte en une courbe trop lisse qui montre peu de détails. Le choix de h est plus important que celui de K. L'estimateur à noyau est donc obtenu en mettant des "bosses" sur chaque observation et ensuite de sommer ces bosses.

0.2.3.1. *Exemples de fonction noyau*

Un des noyaux les plus utilisés est le noyau gaussien donné par

$$K(t) = \frac{1}{\sqrt{2\pi}} e^{-t^2/2}, \; t \in \mathbb{R}.$$

Dans les noyaux à supports compacts on peut citer le noyau "cosinus" donné par

$$K(t) = \frac{\pi}{4}\cos\left(\frac{\pi}{2}t\right) \mathrm{I}(|t| \leq 1).$$

Beaucoup de noyaux (à supports compacts) qui reviennent souvent dans la littérature appartiennent à la famille suivante

$$K(t) = c_{rs}\,(1-|t|^r)^s\,\mathrm{I}(|t|\le 1), \quad c_{rs} = \frac{r}{2\mathrm{Bêta}(s+1,1/r)},\ r>0, s\ge 0.$$

En voici quelques-uns

(K1)	*Uniforme ou rectangulaire* :	$s=0,$	$c_{r0}=1/2.$
(K2)	*Triangulaire* :	$r=s=1,$	$c_{11}=1.$
(K3)	*Epanechnikov* :	$r=2,\ s=1,$	$c_{21}=3/4.$
(K4)	*"Biweight"* ou *"Quartic"* :	$r=s=2,$	$c_{22}=15/16.$
(K5)	*"Triweight"* :	$r=2,\ s=3,$	$c_{23}=35/32.$

Ces fonctions noyaux sont représentées dans la figure 0.2.

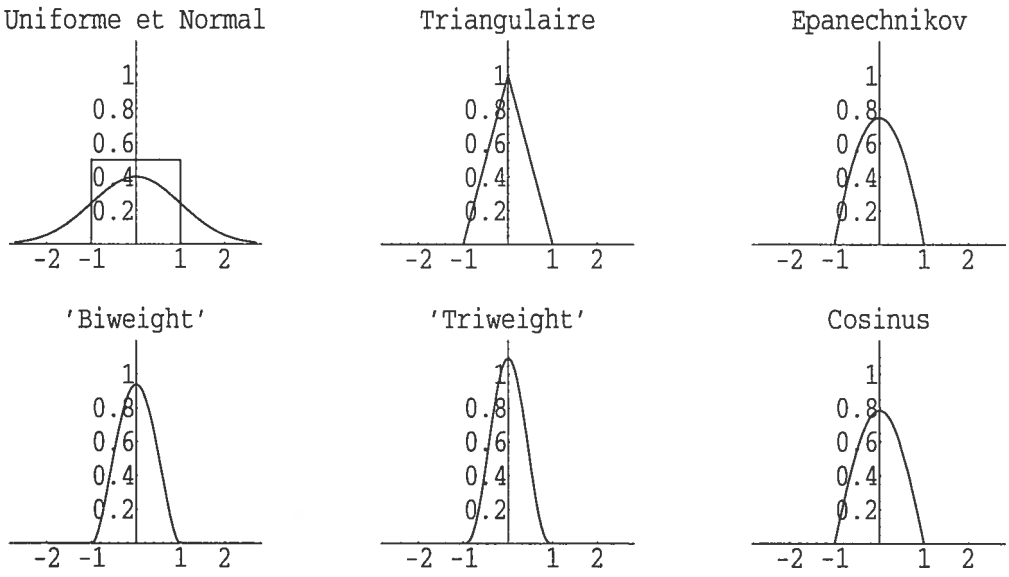


FIG. 0.2. Graphiques de quelques fonctions noyaux.

Example 0.2.1. *On considère le mélange de deux normales suivant*

$$f:\ 0.5\,\mathrm{N}(0,0.09)\ +\ 0.5\,\mathrm{N}(1.5,0.025)$$

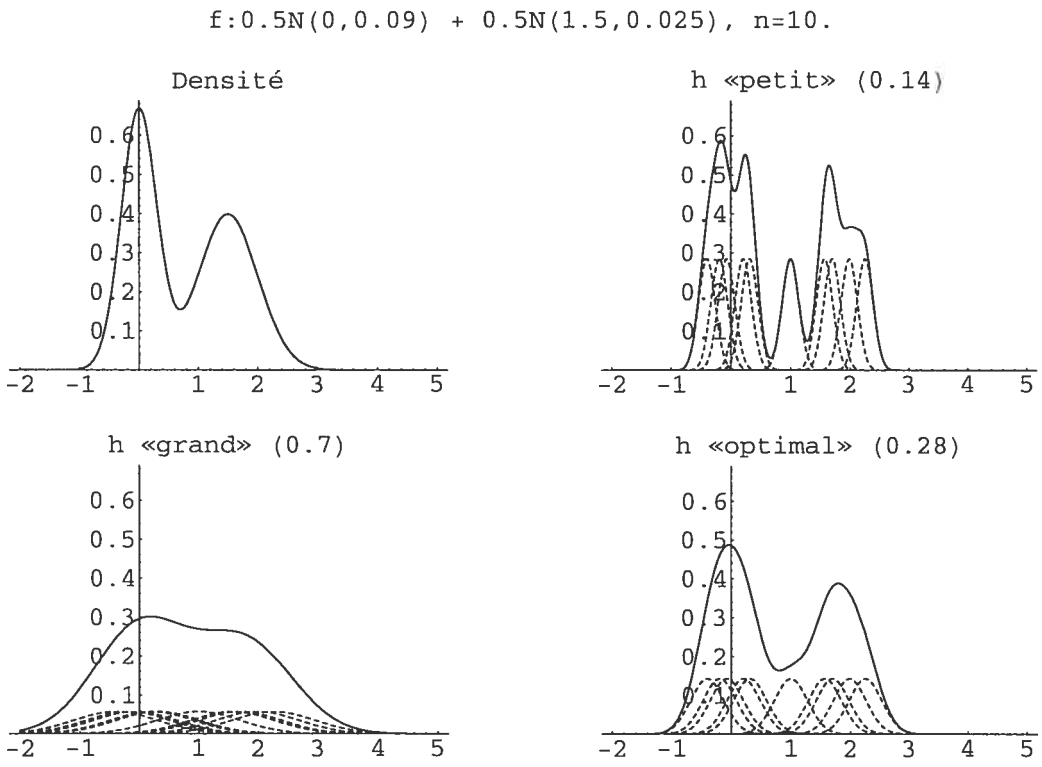


FIG. 0.3. Différents résultats pour différentes valeurs de h.

et on dispose de $n = 10$ *observations. À la figure 0.2., on trouve le graphique de la vraie densité* f *ainsi que trois estimations correspondant à trois choix du paramètre* h.

0.2.3.2. *Propriétés immédiates de l'estimateur à noyau*

- L'estimateur $\hat{f}_h$ hérite, en général, des propriétés de la fonction K. Il sera, par exemple, continu, différentiable et une densité si K l'est. Aussi, il pourra prendre des valeurs négatives dans le cas où K en fait autant.
- Biais :

$$\mathrm{E}\left(\hat{f}_h(x)\right) = \mathrm{E}\left(K_h(x - X)\right) = (K_h * f)(x).$$

- Variance :

$$\mathrm{Var}\left(\hat{f}_h(x)\right) = \frac{1}{n}\left\{(K_h^2 * f)(x) - (K_h * f)^2(x)\right\}.$$

En pratique, c'est plutôt avec des expressions asymptotiques du biais et de la variance qu'on travaille. Pour cela, on ajoute des hypothèses sur f et K comme exiger de f d'être deux fois différentiable et les conditions suivantes sur K

$$K \geq 0, \quad \int K(u)\,du = 1, \quad \int uK(u)\,du = 0, \quad \int u^2K(u)\,du < \infty.$$

Sous ces hypothèses de régularité on obtient

$$\text{Biais}\{\hat{f}_h(x)\} = A(f,K)\,h^2 + o(h^2) \tag{0.2.1}$$

$$\text{Var}\{\hat{f}_h(x)\} = B(f,K)\,\frac{1}{nh} + o\left(\frac{1}{nh}\right), \tag{0.2.2}$$

où

$$A(f,K) = \frac{1}{2}f''(x)\int u^2K(u)du \quad \text{et} \quad B(f,K) = f(x)\int K^2(u)du.$$

Il en résulte que

- Biais$\{\hat{f}_h(x)\} \to 0$, lorsque $h \to 0$.
- Var$\{\hat{f}_h(x)\} \to 0$, lorsque $h \to 0$ et $nh \to \infty$.
- D'où MSE$\{\hat{f}_h(x)\} \to 0$, lorsque $h \to 0$ et $nh \to \infty$.
  Aussi, la partie principale du développement du biais est croissante en h alors que celle de la variance est décroissante. Dans le choix du paramètre h, on devra faire un compromis entre le carré du biais et la variance.

0.2.3.3. *Choix du méta-paramètre h*

Les critères les plus utilisés pour le choix de h sont la MSE et la MISE, où

$$\text{MISE}\{\hat{f}_h\} = \int \text{MSE}\left(\{\hat{f}_h(x)\}\,dx.\right.$$

Il s'agira alors de trouver le paramètre h qui minimisera l'une ou l'autre de ces quantités selon le choix. On distinguera ici les deux cas suivants :

(a) Le cas où h est constant et on parlera alors de paramètre *global*. Pour l'obtenir, on minimisera la MISE. En fait, en pratique c'est plutôt une approximation de la MISE qu'on essayera de minimiser pour obtenir la quantité suivante

$$h_{\text{MISE}} = C_1(f,K)\frac{1}{n^{1/5}}, \quad \text{avec} \quad C_1(f,K)^5 = \frac{\int K^2(u)\,du}{\int u^2K(u)\,du \int f''^2(u)\,du}.$$

La valeur de h obtenue dépend donc de quantités inconnues.

(b) Le cas où $h = h(x)$ est variable. On parlera dans ce cas d'un paramètre *local*. On le déduit en minimisant une approximation de la MSE. On obtient l'expression suivante

$$h_{\text{MSE}} = C_2(f,K)\frac{1}{n^{1/5}}, \quad \text{avec} \quad C_2(f,K)^5 = \frac{f(x)\int K^2(u)\,du}{f''^2(x)\left(\int u^2K(u)\,du\right)^2}.$$

En pratique, on remplacera f dans les expressions précédentes par la densité de la loi normale par exemple. D'autres méthodes sont aussi utilisées pour déterminer le paramètre *optimal* h. La plus utilisée en pratique est la méthode dite de validation croisée. Avec l'histogramme, l'estimateur à noyau est le plus répandu et le plus étudié.

0.2.3.4. *Méthode du noyau adaptatif ou variable*

Une modification de l'estimateur à noyau classique est faite en faisant varier le paramètre h. Deux telles versions ont reçu une attention particulière :

- **Fenêtre locale** ("local bandwidth") ; le cas $h = h(x)$.
  Plusieurs techniques ("observation dépendantes") ont été proposées pour choisir la fenêtre (locale). On peut citer ici les travaux de Fan et *al.* (1996), Hazelton (1996, 1999) et Farmen et Marron (1999.) Cependant, à distance finie, les études de simulations dans ces études n'ont pas montré de résultats prometteurs même lorsque comparés à des estimateurs à fenêtre fixe comme celui de Sheather et Jones (1991) (voir Hazelton, 2003).

- **Fenêtre adaptative** ("sample-point adaptive bandwidth") ; $h = \alpha(X_i)$.
  La fonction α est choisie de sorte que AMISE($\hat{f}$) est petite. Abramson (1982) propose de prendre $h_i \propto f(X_i)^{-1/2}$. Une seconde façon de faire est celle de restreindre α à une certaine classe de fonctions, ensuite optimiser relativement à un certain critère (Sain et Scott, 1996.) Pour plus de détails, voir Hazelton (2003).

BIBLIOGRAPHIE

[1] Altman, N., Léger, C., (1995). *"Bandwidth selection for kernel distribution function estimation"*. J. Statist. Plann. Inference **46**, 195-214.

[2] Azzalini, A., 1981. *"A note on the estimation of a distribution function and quantiles by a kernel method."* Biometrika **68** (1), 326-328.

[3] Bowman, A., Hall, P., Prvan, T., 1998. *"Bandwidth selection for the smoothing of distribution functions."* Biometrika **85** (4), 799-808.

[4] Chu, I.S., (1995). *"Bootstrap smoothing parameter selection for distribution function estimation."* Math. Japon. **41** (1), 189-197.

[5] Csáki E. (1984). *"Empirical Distribution Function."*, Handbook of Statistics (P. R. Krishnaiah and P. K. Sen, eds), **vol. 4,** 405-430.

[6] Devroye, L. and Lugosi, G. (2001). *"Combinatorial Methods in Density Estimation"*, Springer-Verlag, New York, Inc.

[7] Dvoretzky, A. Kiefer, J. and Wolfowitz, A. J. (1956) *"Asymptotic minimax character of the sample distribution function and of the classical multinomial estimator,"* Ann. Math. Statist., 27, No. 3, 642669.

[8] Efron, B., Tibshirani, R.J., (1993). *"An Introduction to the Bootstrap."* Chapman & Hall, London.

[9] Falk, M., (1983). *"Relative eficiency and deficiency of kernel type estimators of smooth distribution functions"*. Statist. Neerlandica **37**, 73-83.

[10] Fan, J., Hall, P., Martin, M. & Patil, P. (1996). "On local smoothing of nonparametric curve estimators." J. Amer. Statist. assoc. **91**, 258-266.

[11] Farmen, M. et Marron, J. (1999) "An assesment of finite sample performance of adaptive methods in density estimation." Comput. Statist. Data Anal. **30**, 143-168.

[12] Finkelstein, H. (1971). *"The Law of the Iterated Logarithm for empirical distributions,"* Ann. Math. Statist., **42,**, 607-615.

[13] Hazelton, M. (1996). "Bandwidth selection for local density estimators." Scand. J. statist. **23**, 221-232.

[14] Hazelton M. (1999). "An optimal local bandwidth selector for kernel density estimation." J. Statist. Plann. Inference **77**, 37-50.

[15] Hazelton M. (2003). "Variable kernel density estimation." Aust. N. Z. J. stat. **45**(3), 271-284.

[16] Jones, M.C., 1990. *"The performance of kernel density functions in kernel distribution function estimation."* Statist. Probab. Lett. **9**, 129-132.

[17] Lehmann, E.L., Casella, G. (1998.) *" Theory of Point Estimation."* 2nd ed., Springer, New York.

[18] Mammitzsch, V., (1984). *"On the asymptotically optimal solution within a certain class of kernel type estimators."* Statist. Decisions **2**, 247-255.

[19] Massart, P. (1990). *"The Tight Constant in the Dvoretsky-Kiefer-Wolfowitz Inequality"*. Annals of probability, **18**, 1269-1283.

[20] Nadaraya, E.A., (1964). *"Some new estimates for distribution functions"*. Theory Probab. Appl. **15** , 497-500.

[21] Reiss, R.D., (1981). *"Nonparametric estimation of smooth distribution functions"*. Scand. J. Statist. **8**, 116-119.

[22] Sain, S. et Scott, D. (1996).*"On locally adaptive density estimation"*, J.A.S.A., **91**, 1525-1534.

[23] Sarda, P., (1993). *"Smoothing parameter selection for smooth distribution functions"*. J. Statist. Plann. Inference **35**, 65-75.

[24] Serfling, J. Robert (1980). *"Approximation Theorems of Mathematical Statistics"*, John Wiley & Sons, Inc.

[25] Shao, Y., Xiang, X., (1997). *"Some extensions of the asymptotics of a kernel estimator of a distribution function."* Statist. Probab. Lett. **34**, 301-308.

[26] Sheather, S., et Jones, M. (1991). *"A reliable data-based bandwidth selection method for kernel density estimation"*, J.R.S.S., Series B **53**, 683-690.

[27] Shirahata, S., Chu, I.S., (1992). *"Integrated squared error of kernel-type estimator of distribution function".* Ann. Inst. Statist. Math. **44** (3), 579-591.

[28] Stute, W. (1982). *"The oscilation behavior of empirical processes."* Ann. Probability, **10**, 86-107.

[29] Swanepoel, J.W.H., (1988). *"Mean integrated squared error properties and optimal kernels when estimating a distribution function".* Comm. Statist. Theory Methods **17** (11), 3785-3799.

[30] de Uña-Álvarez, J., González-Manteiga, W., Cadarso-Suárez, C., (2000) *"Kernel distribution function estimation under the Koziol-Green model."* Journal of Statistical Planning and Inference **87**, pp. 199-219.

[31] Watson, G. S., Leadbetter, M. R. (1964) *"On the estimation of the probability density."* I. Ann. Math. Statist. **34** 480–491.

[32] Winter, B. B. (1979) *"Convergence rate of perturbed empirical distribution functions."* J. Appl. Probab. **16**, no. 1, 163–173.

Chapitre 1

ESTIMATING THE CDF BY A PERTURBATION OF THE EDF

"Distributions are the numbers of the future."

Berthold Schweizer.

Abstract.

In this paper, we propose a new nonparametric approach for estimating a cumulative distribution function F using finite mixtures. The properties of the proposed estimator are studied. We are able to obtain the same asymptotic statistical properties as those obtained using the empirical distribution function, *under the same minimum regularity conditions*. Uniform results for fixed sample sizes are also obtained. Simulations and examples illustrate the approach.

Key Words : Non-parametric cdf estimation, Adaptive smoothing, Splines, Uniform consistency, Convergence in law.

1.1. INTRODUCTION

We are interested in estimating a cumulative distribution function (cdf) F with support on an interval I of $\mathbb{R}$, bounded or not, based on a sample $X_1, \ldots, X_n$ from F. Several works have been devoted to the estimation of a cdf. Most of these works require regularity conditions such as the existence of the density function, the density being Lipschitz or F being twice differentiable, for example. Our aim is to obtain a *smooth* estimation when necessary with strong asymptotic results such as those obtained using the empirical distribution function under the same regularity conditions. Furthermore, we wish to obtain interesting results even for small sample sizes. We hope, on the basis of the theoretical and simulation results we obtain, that the reader will be convinced we achieved these goals.

The most frequently used estimator for a cdf F is the empirical (sample) distribution function (edf) F_n, where $F_n(x) = \sum_{i=1}^{n} I(X_i \leq x)/n$ (I being the indicator function). Here $nF_n(x)$ has a binomial distribution $B(n, F(x))$ and F_n indicates the location of the observations. Also, the edf has a long list of good statistical properties such as it is first order efficient in the minimax sense and $F_n(x)$ is the unique minimum variance and unbiased estimator of $F(x)$ (see Dvoretzky & al. (1956) and Lehmann & Cassella (1998), chapter 2). Furthermore, the edf is the nonparametric maximum likelihood estimator of F and plays a central role in nonparametric simulation and bootstrap (see Efron and Tibshirani (1993), p. 310.) For a review of some properties of the edf see, e.g., Csáki E. (1984), Stute (1982) and Serfling (1980). The fact that F_n is a step function even when the underlying cdf F is continuous, has called for the need (in certain areas of application like estimating the density) for smooth(er) estimators of F. Many *smooth* estimators have been proposed in the literature. Most of these estimators are based on smoothing the edf. One that has been extensively studied is the kernel estimator, say F_h. Nadaraya (1964) established, under appropriate regularity conditions, the almost sure uniform convergence of F_h. Watson and Leadbetter (1964) proved the asymptotic normality of F_h, Winter (1979) showed that F_h has

the Smirnov-Chung property. Azzalini (1981) derived an asymptotic expression for the mean squared error of F_h. Falk (1983), Mammitzsch (1984), Swanepoel (1988) and Jones (1990) analyzed mean integrated squared error properties of F_h, proving that the smoothed estimate is asymptotically more efficient than the empirical one. Shirahata and Chu (1992) showed that the superiority of kernel estimators is not necessarily true in the sense of the integrated squared error. Sarda (1993), Altman and Leger (1995) and Chu (1995) are devoted to the problem of bandwidth selection for F_h. See as well Shao & Xiang (1997), Bowman and *al.* (1998) and Alvarez and *al.* (2000). Alternative methods have been proposed as well.

Several estimators using splines have been investigated. Wahba (1976) proposed the method called histospline (a spline-smoothed histogram) which uses a cubic spline to smooth the edf. Restle (1999) propose another estimator based on smoothing the edf using cubic splines. He and Shi (1998) use quadratic splines to estimate the cdf. We may mention as well Ramsay (1998, 1988) who uses monotone (regression) splines to estimate a monotone function.

Nonparametric Bayesian estimators have been proposed by, e.g., Perron & Mengersen (2001) who use mixtures of triangular distributions (quadratic splines). Hansen & Lauritzen (2002) use Dirichlet processes to model the prior to estimate a *concave* cdf.

Chaubey and Sen (1996, 2002 (multivariate)) propose the estimation of a smooth cdf F, based on a Poisson operator (using Hille's theorem) to smooth the edf. Babu and *al.* (2002) propose an estimator based on Bernstein (operator) polynomials to smooth the edf. For other approaches and references see, e.g., Efromovitch (2001).

In order to estimate a cdf F, we start from the fact that the best estimation based on the observations cannot do better than the best approximation based on the fact that F is known. The present work has two goals. The first one is to develop a method for *approximating* a cdf F. The aim is to approximate the

space of all cdf on the interval I by a finite dimensional space, of dimension m, say. The second one is to apply the approximation to the edf F_n and therefore to estimate F. We would like to mention here that our estimator is not necessarily smooth. In fact, our approach concerns the estimation of a cdf without further conditions on the function F.

When F is known (section 2), a basic approximation G of F is a step function, with (discontinuity) jumps of the same amplitude. It is then enough to choose m jumps in the interval I in order to obtain an approximation G such that $\|F - G\|_\infty \leq 1/2(m+1)$ (uniform upper bound). The problem then reduces to determining the location of the jumps (i.e., the nodes), this is an m-dimensional problem. Still, one might prefer other options than working on step functions. Therefore, we seek for smoother alternatives to the step function G. In section 2, we develop an approach, where by using a *smoothing parameter* k, we are able to construct an approximation G_k, in the form of a finite mixture of cdfs (we call *basis functions*), such that, using m nodes, we obtain $\|F - G_k\|_\infty \leq k/2(m+1)$. Furthermore, in our construction, we allow the possibility of using an instrumental function H which, when chosen close to F, helps in obtaining a better approximation (in fact, when H is taken as F, then $G = F$, see lemma 2.3). By analogy to the Bayesian approach, the function H is seen here as a "prior" distribution (a pseudo prior.) The basis functions, that enter in the definition of G, possess a hierarchical structure, with supports that depend on a vector of nodes. Each cdf (descendent), element of the basis, is a mixture of two cdfs (generators) at a lower level, and the mixture of the two is performed using a third (fixed) cdf, say H. A choice of the nodes that ensures a uniform bound to $\|F - G\|_\infty$ is given in Lemma 2.2. A construction of the basis is given in section 2.3. The approximation G is smooth in general but it allows for discontinuity jumps when necessary/suitable.

We think that the approximation results we obtain are new and could be used in approximation theory. In particular, we give a probabilistic interpretation to our construction which allows, among other things, to show in a simple

way how the construction of monotone splines gives the property of monotonicity. In particular, we construct the monotone spline basis in a probabilistic manner (which presents an interest in itself). Furthermore, we think that, the (parameter) function H and the role it plays in the approximation process, is unique to our method.

When F is unknown (section 3), we apply the approximation to the edf F_n. The edf is then used to choose the nodes (among the order statistics) that define the supports of the basis functions and thus define the estimator $\hat{F}$ of F (Lemma 3.1). With the right choice of the nodes, we obtain $\|F_n - \hat{F}\|_\infty \leq k/2(m+1)$, which allows us to prove almost sure uniform convergence of $\hat{F}$ to F (Lemma 3.2). We then give, under certain conditions on m and n, the rates of the L_∞ convergence and the L_p convergence for $p \geq 1$. We show as well that the estimator $\hat{F}$ has the Chung-Smirnov property. In section 3.5, we establish the asymptotic behavior of the estimator in terms of the convergence in law of $\sqrt{n}\|F - \hat{F}\|_\infty$. In sections 3.4 and 3.5, equivalent statistics to the Kolmogorov-Smirnov and the Cramér-Von Mises statistics are introduced (where F_n is replaced by $\hat{F}$) . We then obtain similar asymptotic results to those obtained using F_n. In section 3.6, we give some uniform upper bounds to the variance, bias, MSE, and other quantities (lemma 3.9). Note that, no other conditions on F than the fact of being a cdf is supposed here compared to other methods where, usually, F is set to be differentiable and sometimes twice differentiable. Furthermore, our approach works on bounded and unbounded supports with no need for transformations and adjustments that may cause a lost of certain (asymptotic) properties of the estimator.

In section 4, we present numerical simulations to illustrate the performance of the proposed estimator. We think that a series of examples will help in understanding the choices to make concerning different parameters involved in the construction of the estimator. General guidelines are given relatively to the choice of these parameters. Comparison to recent works are done throughout the paper.

Finally, we would like to mention that "direct" applications are considered in future works. Among these, the estimation of a density function (in progress), the estimation of the survival function and related functions, and some bootstrap applications (smoothed bootstrap) are considered.

1.2. Approximation of functions and properties

In this section we discuss the problem of approximating a cumulative distribution function (cdf) F defined on an interval $I \subset \mathbb{R}$, i.e., a non-decreasing, continuous from the right function with values on the interval $[0,1]$ such that $F(x)(1-F(x)) = 0$ for all $x \notin I$.

Let us denote by $\mathcal{F}(I)$ the space of all cdfs on I, and let us define for $F_1,\ F_2 \in \mathcal{F}(I)$ the (usual supremum) metric, $d_{sup}(F_1, F_2) = \sup_{x\in\mathbb{R}} |F_1(x) - F_2(x)|$, which we denote by $\|F_1 - F_2\|_\infty$. The space $(\mathcal{F}(I),\ d_{sup})$ is a complete metric space.

In a statistical setup, we may have $X_1, X_2, \ldots, X_n$, a sample from a common distribution P_θ, where the parameter θ depends on different quantities including a parameter F, $F \in \mathcal{F}$. Under a frequentist approach, we may consider the maximum likelihood estimator of θ, but this is well known to lead to overfitting problems with the space $\mathcal{F}$ being too large. In a Bayesian context, we may instead establish a prior on Θ, the space of θ, but this may be complicated since the space is infinite dimensional. The technical difficulties of working on $\mathcal{F}$ itself thus impel us to consider alternative finite dimensional approximate spaces.

1.2.1. Approximation and loss

Assume that $\mathcal{F}_m = \{H_\omega \colon \omega \in \Omega \subset \mathbb{R}^m\}$ is a space of dimension m approximating $\mathcal{F}$.

Two questions arise : What is the loss associated with the use of $\mathcal{F}_m$ instead of $\mathcal{F}$, and can we find an approximate space $\mathcal{F}_m$ for which this loss is small ?

A measure of loss associated with the approximation of $\mathcal{F}$ by $\mathcal{F}_m$ can be given by

$$\lambda(\mathcal{F}, \mathcal{F}_m) = \sup_{F \in \mathcal{F}} \inf_{H \in \mathcal{F}_m} d_{sup}(F, H).$$

Hence, $\lambda(\mathcal{F}, \mathcal{F}_m)$ is always bounded from above by one.
The second question may now be phrased as follows. Given $\epsilon > 0$, can we find an approximate space $\mathcal{F}_m$ such that $\lambda(\mathcal{F}, \mathcal{F}_m) < \epsilon$? And, if so, is there a simple upper bound on m? In practice we would prefer to have both m and ϵ small!.
In the next section, we introduce an approximate space that will achieve the goals set in this section. This approximate space consists of finite mixtures of cdfs.

1.2.2. The mixture

Let $a = \inf\{x\colon x \in I\}$ and $b = \sup\{x\colon x \in I\}$. The approximate space $\mathcal{F}_m$, we choose to work with is based on m points $a \le y_1 \le \cdots \le y_m \le b$ called nodes (knots) and a parameter k called the smoothing parameter. For convenience, we set $y_j = a$ for $j = -k+2, \ldots, 0$ and $y_j = b$ for $j = m+1, \ldots, m+k-1$ (multiple nodes of order at least $(k-1)$ at the end points a and b). Any $G_k \in \mathcal{F}_m$ is a finite mixture and satisfies the following representation

$$G_k = \sum_{j=1}^{m+k-1} \omega_{kj}\, G_{k,j} = \frac{1}{(k-1)(m+1)} \sum_{j=1}^{m+1} \sum_{l=j}^{j+k-2} G_{k,l},$$

with weights

$$\omega_{kj} = \frac{\min(j, k-1, m+k-j)}{(k-1)(m+1)} \quad \text{for} \quad j = 1, 2, \ldots, m+k-1.$$

The weights ω_{kj} are therefore positive and add up to 1.
Each function $G_{k,j}$ is a cdf on the interval $I_{k,j} = [y_{j-k+1}, y_j]$ for $j = 1, 2, \ldots, m+k-1$. The elements $G_{k,j}$ are called basis functions. A construction of $G_{k,j}$ is given in the next section. We further suppose that $m \ge k-1 > 0$. The function G_k thus defined is therefore an element of $\mathcal{F}(I)$ (hence a cdf by construction).

1.2.3. The basis

The basis elements are such that $G_{k,j} \geq G_{k,j+1}$ for $j = 1, 2, \ldots, m+k-2$. For convenience, we set $G_{k,0} = 1$ and $G_{k,m+k} = 0$.

There is a hierarchical structure in the construction of the basis based on a fixed cdf H on I. Let us define on $I_{k,j}$, $j = 1, 2, \ldots, m+k-1$, the following functions

$$H_{k,j}(x) = \begin{cases} \frac{H(x)-H(y_{j-k+1})}{H(y_j)-H(y_{j-k+1})} & \text{if} \quad H(y_{j-k+1}) < H(y_j), \\ I(x \geq y_j) & \text{if} \quad H(y_{j-k+1}) = H(y_j). \end{cases}$$

Then the hierarchical structure is the following

Step 1 (initial step) : $G_{1,j}(x) = I(x \geq y_j)$ for all $x \in I$ and $j = 1, \ldots, m$ (Dirac cdf).

Step $l+1$ **:** $G_{l+1,j} = H_{l+1,j}\, G_{l,j-1} + (1 - H_{l+1,j})\, G_{l,j}$.

Clearly $G_{k,j} \geq G_{k,j+1}$ for all k and j, $j = 1, 2, \ldots, m+k-2$.

The fact that $G_{k,j}$ is a cdf on $I_{k,j} = [y_{j-k+1}, y_j]$ comes from the next lemma.

Lemma 1.2.1. *We have the following results,*

(a) Let F_1, F_2 be two cdfs on I, an interval of $\mathbb{R}$, such that $F_1 \geq F_2$, and let F_3 be a cdf on $\mathbb{R}$.

The function $F_{(2)}$ defined by

$$F_{(2)} = F_3\, F_1 + (1 - F_3)\, F_2$$

is then a cdf on I.

(b) The function $F_{2\wedge 3} = F_3 + (1 - F_3)F_2 = F_2 + (1 - F_2)F_3$ is a cdf on I (the extreme case $F_1 \equiv 1$.)

In particular, if $F_1 = F_2 = 1$, then $F_{(2)} = 1$.

(c) The function $F_{1\vee 3} = F_3 F_1$ is a cdf on I (the extreme case $F_2 \equiv 0$.)

In particular, if $F_1 = F_2 = 0$, then $F_{(2)} = 0$.

Proof.

(a) Let X_1 be a random variable (r.v.) with cdf F_1.

Let U be a r.v. independent of X_1 and uniformly distributed on $[0,1]$.

Set $X_2 = F_2^{-1}\left[(1-U)\sup\{F_1(x)\colon x < X_1\} + U\,F_1(X_1)\right]$.

The cdf of X_2 is then F_2 and we have $P[X_2 \geq X_1] = 1$.

Let X_3 be a r.v. independent of (X_1, X_2) and with cdf F_3.

If we denote by $X_{(2)}$ the second order statistic based on X_1, X_2, and X_3, then

$$\begin{aligned} P[X_{(2)} \leq x] &= P[X_{(2)} \leq x | X_3 \leq x]P[X_3 \leq x] + P[X_{(2)} \leq x | X_3 > x]P[X_3 > x] \\ &= F_1(x)F_3(x) + F_2(x)(1 - F_3(x)) \\ &= F_{(2)}(x) \quad \text{for all } x \in I, \end{aligned}$$

recall that $X_1 \leq X_2$ with probability one.

Note that $F_2 \leq F_{(2)} \leq F_1$ and $X_1 \leq X_{(2)} \leq X_2$ with probability one.

(b) The function $F_{2\wedge 3}$ is the cdf of the r.v. $\min(X_2, X_3)$.

(c) The function $F_{1\vee 3}$ is the cdf of the r.v. $\max(X_1, X_3)$. □

Remarks :

(1) Suppose we have multiple nodes of multiplicity l at y_j, e.g., $y_j = y_{j+1} = \cdots = y_{j+l-1}$, then we obtain $G_{1,j} = G_{1,j+1} = \cdots = G_{1,j+l-1}$ and therefore, $G_{l,\,j+l-1}(x) = G_{l-1,j+l-2}(x) = \cdots = G_{1,j}(x)$ if $y_j = y_{j+l-1}$.

(2) If a node y_j is of multiplicity l with $l \geq k$, then the function G will have a jump at y_j.

If H is continuous, then the amplitude of the jump is given by $(l - k + 1)_+/(m+1)$, where $a_+ = \max(0, a)$.

(3) If H is chosen as the cdf of a uniform distribution on I (bounded), then the basis functions $G_{k,j}$ become piecewise polynomials of degree $k-1$ and the function G_k becomes a *monotone* spline. So, for $k = 4$, e.g., the approximation is a cubic (*monotone*) spline.

(4) The function H may depend on the nodes.

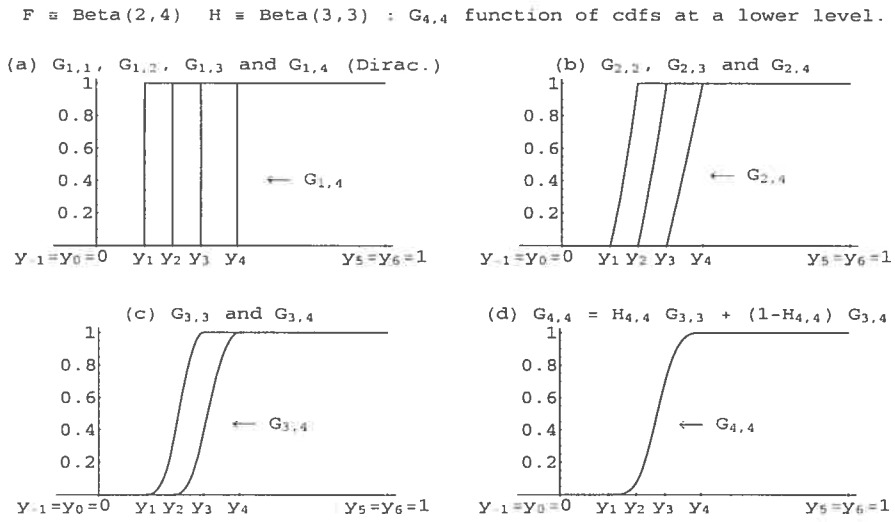


FIG. 1.1. The building of the basis function $G_{4,4}$.

(5) If H is flat between two nodes, then G_k is flat between these two nodes.

(6) If H has a discontinuity jump at a point x_0, then G_k has a jump at x_0.

(7) In figure 1.1, we consider the case where F is the cdf of a $Beta(2,4)$, H is taken as the cdf of a $Beta(3,3)$ and we show the building of an element of the basis for $k=4$, say $G_{4,4}$ (cdf on $[y_1, y_4]$) throw the different stages involved. The function in (d) $(G_{4,4})$ is the mixture of the two functions in (c) ($G_{3,3}$ and $G_{3,4}$). Each function in (c) is a mixture of two functions in (b), etc.

1.2.4. The choice of the nodes and the parameters H and k

When $k=1$, G_k is a step function. In order to obtain a good approximation for F, it is natural to set

$$y_j = F^{-1}\left(\frac{j}{m+1}\right)$$

where F^{-1} is the generalized inverse of F,

$$F^{-1}(x) = \inf\{t\colon F(t) \geq x\}.$$

For $k > 1$, we shall keep the same nodes. In fact, we have the following lemma on the quality of the approximation.

Lemma 1.2.2. (Choice of the nodes)

If we take $y_i = F^{-1}\left(\frac{i}{m+1}\right)$ *for* $i = 1, \ldots, m$, *then we obtain*

$$\|F - G_k\|_\infty \leq \frac{k}{2(m+1)}.$$

Proof.

Let $x \in I\backslash\{a, y_1, \ldots, y_m, b\}$, then there exists $i \in \{0, 1, \cdots, m\}$ for which $y_i < x < y_{i+1}$. We have, on the one hand

$$\frac{i}{m+1} \leq F(x) < \frac{i+1}{m+1}$$

and on the other hand

$$\frac{i+1}{m+1} - \frac{k}{2(m+1)} \leq \sum_{j=1}^{i} \omega_{kj} \leq G_k(x) \leq \sum_{j=1}^{i+k-1} \omega_{kj} \leq \frac{i}{m+1} + \frac{k}{2(m+1)}.$$

Hence for $x \in I\backslash\{a, y_1, \ldots, y_m, b\}$

$$-\frac{k}{2(m+1)} \leq F(x) - G_k(x) < \frac{k}{2(m+1)},$$

that is,

$$|F(x) - G_k(x)| \leq \frac{k}{2(m+1)}, \quad \text{for all } x \in I\backslash\{a, y_1, \ldots, y_m, b\}.$$

Finally, since F and G are continuous from the right, we obtain

$$\|F - G_k\|_\infty \leq \frac{k}{2(m+1)}. \quad \square$$

It follows that for $\epsilon > 0$ given, we have $\lambda(\mathcal{F}, \mathcal{F}_m) \leq \epsilon$ whenever $m > \frac{k}{2\epsilon} - 1$, where $\mathcal{F}_m$ is the approximate space of dimension m. We need to emphasize here that the choice of the nodes is really critical. The function H is like a guess for F. When the nodes are adequately selected, a poor choice for H cannot be dramatic.

However, we hope that a perfect guess will imply that $G_k = F$ for all $k > 1$. In fact, this is true in the case where F is continuous (see next lemma).

Lemma 1.2.3. $(H \equiv F \implies G_k \equiv F)$

If F is a continuous cdf on an interval $I \subset \mathbb{R}$, $y_j = F^{-1}(j/(m+1))$ for $j = 1, 2, \ldots, m$ and $H = F$ then $G_k = F$ for all k, $k \leq m+1$.

Proof.

The proof is done by induction on k, m being fixed. The result holds for $k = 2$. Suppose it is also true for $k = 2, \ldots, l$, where $l \leq m+1$.

For all $x \in \mathbb{R}$ we have,

$$\begin{aligned}
G_{l+1}(x) &= \sum_{j=1}^{m+l} \omega_{l+1,j}\, G_{l+1,j}(x) \\
&= \sum_{j=1}^{m+l} \omega_{l+1,j} \left\{ H_{l+1,j}(x)\, G_{l,j-1}(x) + (1 - H_{l+1,j}(x))\, G_{l,j}(x) \right\} \\
&= \omega_{l+1,1} H_{l+1,1}(x)\, G_{l,0}(x) \\
&+ \sum_{j=1}^{m+l-1} \left\{ \omega_{l+1,j+1} H_{l+1,j+1}(x) + \omega_{l+1,j}(1 - H_{l+1,j}(x)) \right\} G_{l,j}(x) \\
&+ \omega_{l+1,m+l}\, (1 - H_{l+1,m+l}(x))\, G_{l,m+l}(x).
\end{aligned}$$

Note that $G_{l,0}(x) = 1$ and $G_{l,m+l}(x) = 0$, for all x. Therefore,

$$\begin{aligned}
G_{l+1}(x) &= \frac{1}{l} F(x)\, I_{(-\infty, y_1]}(x) + \frac{1}{l(m+1)} I_{(y_1, \infty)}(x) \\
&+ \sum_{j=1}^{m+l-1} \left\{ \frac{l-1}{l} \omega_{l,j}\, G_{l,j}(x) + \frac{1}{l} F(x)\, I_{(y_j, y_{j+1}]}(x) - \frac{j}{l(m+1)} I_{(y_j, y_{j+1}]}(x) \right. \\
&\qquad \left. + \frac{1}{l(m+1)} I_{(y_{j+1}, \infty)}(x) \right\}.
\end{aligned}$$

By induction, we have $\sum_{j=1}^{m+l-1} \omega_{l,j}\, G_{l,j}(x) = F(x)$ for all x. Furthermore, we have

(1) $I_{(-\infty, y_1]}(x) + \sum_{j=1}^{m+l-1} I_{(y_j, y_{j+1}]}(x) = 1.$

(2) $I_{(y_1, \infty)}(x) - \sum_{j=1}^{m+l-1} j\, I_{(y_j, y_{j+1}]}(x) = -\sum_{j=1}^{m} (j-1)\, I_{(y_j, y_{j+1}]}(x)$

(3)

$$\begin{aligned}\sum_{j=1}^{m+l-1} I_{(y_{j+1},\infty)}(x) &= \sum_{j=1}^{m} I_{(y_{j+1},\infty)}(x) \\ &= \sum_{j=1}^{m}\sum_{i=j+1}^{m} I_{(y_i,y_{i+1}]}(x) = \sum_{i=2}^{m}\sum_{j=1}^{i-1} I_{(y_i,y_{i+1}]}(x) \\ &= \sum_{i=2}^{m} (i-1)\, I_{(y_i,y_{i+1}]}(x) = \sum_{j=1}^{m} (j-1)\, I_{(y_j,y_{j+1}]}(x)\end{aligned}$$

Thus, $G_{l+1}(x) = F(x)$ for all x. □

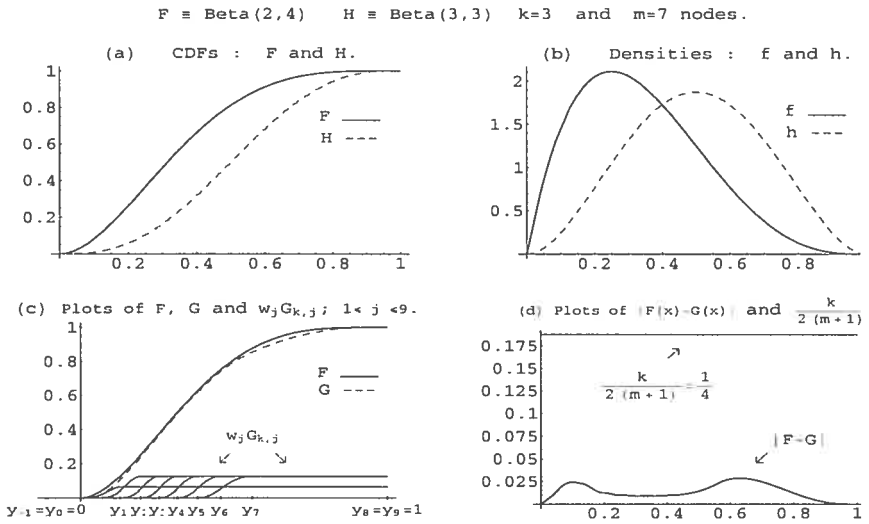
Next, we give two examples. The first one concerns a smooth cdf, i.e., a $Beta(2,4)$, and the second one concerns a cdf with a discontinuity jump.

Example 1.2.1. *Let F be the cdf of a $Beta(2,4)$ distribution. For illustration, we take $k = 3$, $m = 7$ nodes only and for H we take a symmetric distribution on $[0,1]$, a $Beta(3,3)$ distribution. The function H is seen as a prior distribution. When no knowledge about F is available and I is bounded, we could choose H as the cdf of a uniform distribution for example. In figure 1.2., we have plotted the cdfs F and H ("the prior") and the corresponding densities f and h. Figure 2.c shows the plots of the approximation G together with F and the 9 functions $w_{kj}G_{k,j}$. The weights at the end points are equal to $1/2(m+1)$. In figure 2.d we have plotted the (absolute) error function $|F(x) - G(x)|$ and we can see that $\|F - G\|_\infty$ is relatively far from the uniform upper bound $k/2(m+1) = 1/4$.*

Example 1.2.2. *In this example, we consider a cdf F on $[0,1]$ with a discontinuity jump at $x = 1/2$. The function F is given by,*

$$F(x) = \begin{cases} x^2 & \text{if } x < \dfrac{1}{2} \\ \dfrac{1+(2x-1)^2}{2}, & \text{otherwise.} \end{cases}$$

There are multiple nodes at $x = 1/2$ so that G jumps at $x = 1/2$. For this example we choose to take $k = 3$, $m = 10, 30, 50,$ and 100. The function H is chosen as the cdf of a uniform distribution on $[0,1]$. In this case, the elements



y_j	0	0	y_1	y_2	y_3	y_4	y_5	y_6	y_7	1	1
y_j	0	0	0.1275	0.1937	0.2537	0.3138	0.3785	0.4541	0.5563	1	1
$F(y_j)$	0	0	0.125	0.25	0.375	0.5	0.625	0.75	0.875	1	1
$G_k(y_j)$	0	0	0.1023	0.2375	0.3652	0.4910	0.6158	0.7390	0.8530	1	1
$w_{k,j}$		1/16	1/8	1/8	1/8	1/8	1/8	1/8	1/8	1/16	

FIG. 1.2. Example of a $Beta(2,4)$.

$G_{k,j}$ of the basis are piecewise polynomials of degree $2 = k-1$, so that G becomes a quadratic (monotone) spline. It is clear that a prior knowledge about this feature of F (the jump) would have made us choose another H that reflects this feature and therefore obtain a better approximation. This ability of the approximation G, and thus of the estimator $\hat{F}$ (see section 3), to be smooth in the regions where we expect F to be so, and to follow the jumps of F whenever there are, makes our method more general than the competitors and helps us obtain better estimations for F. See figure 1.3. for graphics.

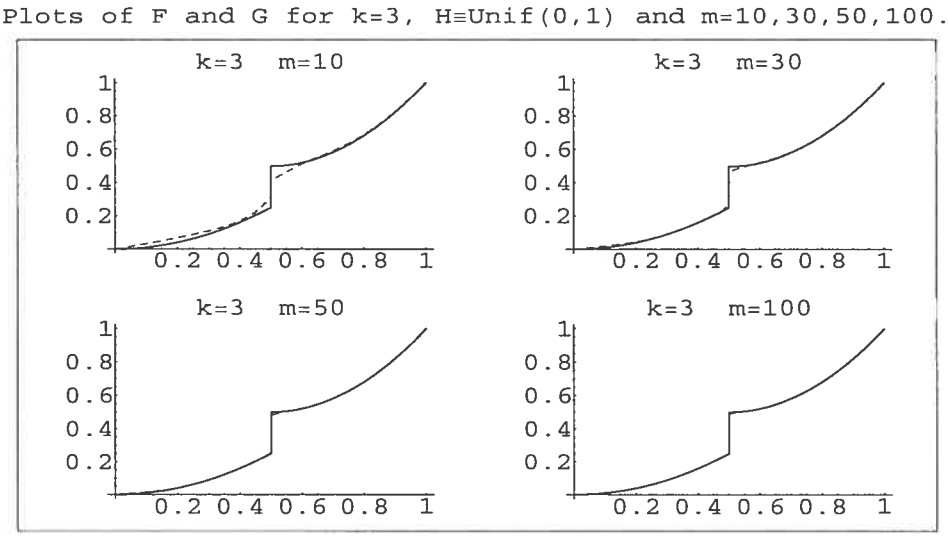


FIG. 1.3. Example of a cdf on $[0,1]$ with a discontinuity jump at $1/2$.

1.3. ESTIMATION AND STATISTICAL RESULTS

In this section we apply the results of section 1.2 to the estimation of an unknown distribution function $F \in \mathcal{F}(I)$ based on a sample $X_1, \ldots, X_n$ from F. Let $F_n(x) = \sum_{i=1}^{n} I(X_i \leq x)/n$ denote the empirical distribution function and let us denote G by $\hat{F}$ (where we drop the subscript k for simplicity). In the next subsections, we are going to look at the asymptotic behavior of the estimator $\hat{F}$ of F. In section 3.1, we give a uniform upper bound to $\|F_n - \hat{F}\|_\infty$. In section 3.2 we establish the (almost sure) uniform convergence of $\hat{F}$ to F (lemma 3.2), then by adding some conditions on m and n, we are able to give, in lemma 3.3, the rate of the uniform convergence. In section 3.3., we give a uniform upper bound to the L_p norm $\|\hat{F} - F\|_p$ $(p \geq 1)$ (and by doing that we prove the L_p convergence of $\hat{F}$ to F and we give the rate of this convergence). In section 3.4, we define a Cramér-Von Mises like statistic and give a result relative to this one (lemma 3.6). In section 3.5., we define a statistic similar to the Kolmogorov-Smirnov one. We establish the convergence in law of $\sqrt{n}\,\|\hat{F} - F\|_\infty$ (lemma 3.9) and we show that the estimator $\hat{F}$ has the Chung-Smirnov property. We end the section by giving some uniform results (uniform upper bounds) concerning the bias, variance, MSE,

and other quantities (lemma 3.11). We would like to note here that by adding some conditions on m and n, we are able to obtain the asymptotic results and properties that the Kolmogorov-Smirnov and the Cramér-Von Mises statistics have.

1.3.1. Distance of $\hat{F}$ to F_n

When F is unknown, we use the edf F_n to choose the nodes (among the order statistics). The selected nodes are given by $y_j = F_n^{-1}\left(\frac{j}{m+1}\right)$ for $j = 1, \ldots, m$. In other words, we approximate the edf. We have the next lemma

Lemma 1.3.1. (Distance of $\hat{F}$ to F_n)

The above choice of the nodes implies that $y_j = x_{(\lceil n\frac{j}{m+1} \rceil)}$, *and in this case we obtain*

$$\|F_n - \hat{F}\|_\infty \leq \frac{k}{2(m+1)},$$

where $\lceil a \rceil = \min\{n \in \mathbb{Z} \colon a \leq n\}$ *(the ceiling), and* $x_{(.)}$ *denotes the order statistics. In the case where* $m = n$, *we have* $y_j = x_{(j)}$ *for* $1 \leq j \leq n$ *and we obtain*

$$\|F_n - \hat{F}\|_\infty \leq \frac{k}{2(n+1)}.$$

We may compare this upper bound to the one obtained in the paper by Babu and al. (2002) (theorem 2.2). In (4), the authors propose a (degree m polynomial) estimator $\tilde{F}_{n,m}$ based on Bernstein polynomials and show that if F is differentiable with density f that is Lipschitz of order 1, then at best we have

$$\|\tilde{F}_{n,m} - F_n\|_\infty = O\left(\left(\frac{\log(n)}{n}\right)^{3/4}\right) \quad \text{almost surely} \quad (\text{when } m = n).$$

Compare this rate of convergence with the one in lemma 1.3.1. where we suppose F to be a cdf without further conditions.

1.3.2. Uniform convergence of $\hat{F}$ to F (L_∞ convergence)

Using the previous lemma, we show that one can obtain (almost sure) uniform convergence of $\hat{F}$ to F.

Lemma 1.3.2. *We have*

$$\lim_{m,n\to\infty} \|\hat{F} - F\|_\infty = 0 \quad a.s.\,(\textit{almost surely}).$$

Proof.

By taking $y_j = x_{(\lceil n\frac{j}{m+1}\rceil)}$, we obtain

$$\|\hat{F} - F\|_\infty \leq \|\hat{F} - F_n\|_\infty + \|F_n - F\|_\infty \leq \frac{k}{2(m+1)} + \|F_n - F\|_\infty.$$

According to the Glivenko-Cantelli theorem, $\|F_n - F\|_\infty$ converges almost surely to 0 as n tends to ∞. Therefore $\|\hat{F} - F\|_\infty$ converges almost surely to 0 as m and n tend to ∞ . □

By adding conditions on m and n, we are able to control the rate of the uniform converge.

Lemma 1.3.3. (Rate of the uniform convergence)

If $\sup_n[n/(m^2 \log\log(n))] < \infty$, *then* $\|\hat{F} - F\|_\infty$ *tends to* 0 *almost surely at the rate* $\sqrt{\frac{2\log\log(n)}{n}}$.

Proof. We have

$$\begin{aligned}\overline{\lim}\|\hat{F} - F\|_\infty \sqrt{\frac{n}{2\log\log(n)}} &\leq \overline{\lim}\frac{k}{2(m+1)}\sqrt{\frac{n}{2\log\log(n)}} \\ &+ \overline{\lim}\|F_n - F\|_\infty \sqrt{\frac{n}{2\log\log(n)}}.\end{aligned}$$

And we have the result by noting that, by the law of the iterated logarithm, we have

$$\overline{\lim}\|F_n - F\|_\infty \sqrt{\frac{n}{2\log\log(n)}} \leq \frac{1}{2}, \quad \text{with probability one.}$$

□

1.3.3. L_p convergence of $\hat{F}$ to F

Let us first recall the following inequalities :

$$x^s + y^s \leq (x+y)^s \leq 2^{s-1}(x^s + y^s), \quad \text{for every } s \geq 1 \text{ and } x, y > 0,$$

$$2^{s-1}(x^s + y^s) \leq (x+y)^s \leq x^s + y^s, \quad \text{for every } 0 \leq s \leq 1 \text{ and } x, y > 0,$$

and

$$\mathrm{P}\{\|F - F_n\|_\infty > t\} \leq 2\,e^{-2nt^2}, \quad \text{for all } t > 0.$$

For the last inequality, see Massart (1990) and Devroye (2001).

We have the next lemma,

Lemma 1.3.4. *Let $p \geq 1$ and let us denote by $\|\cdot\|_p$ the L_p-norm, i.e.,*

$$\|G\|_p = \|G\|_{p,K} = \left(\int_{-\infty}^{\infty} |G(x)|^p dK(x) \right)^{1/p}, \quad \textit{for all} \quad G, K \in \mathcal{F}(I).$$

We have

$$\begin{aligned} \mathrm{E}\|\hat{F} - F\|_{p,F} &\leq \left\{ 2^{p-1} \left[\frac{p\,\Gamma(p/2)}{(2n)^{p/2}} + \left[\frac{k}{2(m+1)} \right]^p \right] \right\}^{1/p} \\ &\leq \frac{1}{2^{1/p}} \left\{ \frac{\sqrt{2}\,[p\,\Gamma(p/2)]^{1/p}}{\sqrt{n}} + \frac{k}{m+1} \right\}. \end{aligned}$$

Proof. Since

$$|\hat{F}(x) - F(x)| \leq \|F - F_n\|_\infty + \frac{k}{2(m+1)}, \quad \text{for every } x \in I,$$

we obtain

$$\begin{aligned} |\hat{F}(x) - F(x)|^p &\leq \left\{ \|F - F_n\|_\infty + \frac{k}{2(m+1)} \right\}^p \\ &\leq 2^{p-1} \left\{ \|F - F_n\|_\infty^p + \left[\frac{k}{2(m+1)} \right]^p \right\}, \end{aligned}$$

it follows that,

$$\mathrm{E}\|\hat{F} - F\|_{p,F}^p \leq 2^{p-1} \left\{ \mathrm{E}\|F - F_n\|_\infty^p + \left[\frac{k}{2(m+1)} \right]^p \right\}$$

We obtain

$$\begin{aligned} \mathrm{E}\|F - F_n\|_\infty^p &= \int_0^\infty \mathrm{P}(\|F - F_n\|_\infty > t^{1/p})\, dt \\ &\leq 2 \int_0^\infty e^{-2nt^{2/p}}\, dt \\ &= p \int_0^\infty u^{p/2-1}\, e^{-2nu}\, du \\ &= \frac{p\,\Gamma(p/2)}{(2n)^{p/2}}. \end{aligned}$$

Hence,

$$\begin{aligned} \mathrm{E}\|\hat{F} - F\|_{p,F} &\leq \left\{ 2^{p-1} \left[\frac{p\,\Gamma(p/2)}{(2n)^{p/2}} + \left[\frac{k}{2(m+1)} \right]^p \right] \right\}^{1/p} \\ &\leq \frac{1}{2^{1/p}} \left\{ \frac{\sqrt{2}\,[p\,\Gamma(p/2)]^{1/p}}{\sqrt{n}} + \frac{k}{m+1} \right\}. \end{aligned}$$

□

1.3.4. A Cramér-Von Mises like statistic

The Cramér-Von Mises statistic is given by (see, e.g., Serfling (1980))

$$C_n = n \int_{-\infty}^{\infty} [F_n(x) - F(x)]^2\, dF(x).$$

We have the following well known result,

Lemma 1.3.5. (Finkelstein (1971))

With probability 1,

$$\overline{\lim} \frac{C_n}{2 \log\log n} = \frac{1}{\pi^2}.$$

Let us define a similar statistic based on $\hat{F}$ in the following way

$$C_n^* = n \int_{-\infty}^{\infty} [\hat{F}(x) - F(x)]^2 \, dF(x),$$

so that we obtain the next lemma

Lemma 1.3.6. *With probability* 1,

$$\overline{\lim} \frac{C_n^*}{2 \log\log n} \leq \frac{k^2}{4} \overline{\lim} \frac{n}{m^2 \log\log n} + \frac{1}{\pi^2}.$$

1.3.5. Asymptotic behavior of $\hat{F}$

Let us recall the following well known results (see Serfling (1980), for example).

Theorem 1.3.1. (Kolmogorov, 1933)

If F is continuous and if we set $D_n = \|F - F_n\|_\infty$, we obtain

$$\lim_{n\to\infty} \Pr(\sqrt{n}\, D_n \leq d) = \sum_{j=-\infty}^{\infty} (-1)^j \, e^{-2j^2 d^2}, \quad d > 0.$$

Theorem 1.3.2. (Smirnov, 1941)

Let us introduce the two quantities

$$D_n^+ = \sup_x [F_n(x) - F(x)] \quad \text{and} \quad D_n^- = \sup_x [F(x) - F_n(x)].$$

If F is continuous, then we have the following result,

$$\lim_n \mathrm{P}(\sqrt{n}\, D_n^+ > d) = \lim_n \mathrm{P}(\sqrt{n}\, D_n^- > d) = e^{-2d^2}, \quad d > 0.$$

Lemma 1.3.7. (Asymptotic behavior of $\hat{F}$)

If F is continuous and m is chosen such that $\frac{n}{m^2} \xrightarrow[n,m\to\infty]{} 0$, then

$$\lim_{m,n\to\infty} \Pr(\sqrt{n}\, \hat{D}_n \leq d) = \sum_{j=-\infty}^{\infty} (-1)^j \, e^{-2j^2 d^2}, \quad d > 0.$$

where $\hat{D}_n = \|\hat{F} - F\|_\infty$.

Proof. We have

$$|\,\|\hat{F}-F\|_\infty - \|F-F_n\|_\infty\,| \le \|\hat{F}-F_n\|_\infty \le \frac{k}{2(m+1)}.$$

Hence,

$$-\frac{k\sqrt{n}}{2(m+1)} + \sqrt{n}\,\hat{D}_n \le \sqrt{n}\,D_n \le \sqrt{n}\,\hat{D}_n + \frac{k\sqrt{n}}{2(m+1)}.$$

We obtain,

$$\Pr\left(\sqrt{n}\,D_n \le d - \frac{k\sqrt{n}}{2(m+1)}\right) \le \Pr\left(\sqrt{n}\,\hat{D}_n \le d\right) \le \Pr\left(\sqrt{n}\,D_n \le d + \frac{k\sqrt{n}}{2(m+1)}\right),$$

and since (Kolmogorov's theorem)

$$\lim_{n\to\infty} \Pr(\sqrt{n}\,D_n \le d) = \sum_{j=-\infty}^{\infty} (-1)^j\, e^{-2j^2d^2}, \quad d>0,$$

the result then follows. □

Lemma 1.3.8. *Let us put,*

$$\hat{D}_n^+ = \sup_x[\hat{F}(x)-F(x)] \quad \textit{and} \quad \hat{D}_n^- = \sup_x[F(x)-\hat{F}(x)].$$

If F is continuous and m is chosen such that $\frac{n}{m^2} \xrightarrow[n,m\to\infty]{} 0$, then

$$\lim_n \mathrm{P}(\sqrt{n}\,\hat{D}_n^+ > d) = \lim_n \mathrm{P}(\sqrt{n}\,\hat{D}_n^- > d) = e^{-2d^2}, \quad d>0.$$

Remarks

(1) **The Chung-Smirnov property**

Note that (cf. lemma 3.3.) if m is chosen such that $\overline{\lim}\; n/(m^2\,\log\log(n)) = 0$, then $\hat{F}$ has the Chung-Smirnov property, i.e.,

$$\overline{\lim}\;(2n/\log\log n)^{1/2}\|\hat{F}-F\|_\infty \le 1, \quad \text{with probability one.}$$

(2) **Pointwise convergence in law**

If m is chosen such that $\overline{\lim}\; n/m^2 = 0$, then

$$\sqrt{n}\left(\hat{F}(x)-F(x)\right) \xrightarrow[n,m\to\infty]{\mathcal{D}} N\left(0, F(x)(1-F(x)\right) \quad \text{for all } x.$$

(3) **A Dvoretzky-Kiefer-Wolfowitz like inequality**

Let us recall the two following inequalities

$$\|\hat{F} - F\|_\infty \leq \frac{k}{2(m+1)} + \|F_n - F\|_\infty,$$

and

$$\mathrm{P}\left\{\|F - F_n\|_\infty > t\right\} \leq 2\, e^{-2nt^2}, \quad \text{for all } t > 0.$$

Let $\{a_m\}$, $a_m \in (0,1)$, be a sequence of real numbers (close to 1), then we obtain

$$\mathrm{P}\left\{\|F - \hat{F}\|_\infty > t\right\} \leq 2\, e^{-2\,n\,a_m^2\,t^2},$$

whenever $(m+1)(1-a_m) \geq k/(2t)$.

1.3.6. Some uniform results concerning the bias, variance, MSE, and other quantities

In the next lemma, we give uniform upper bounds to the bias, variance, MSE, and other quantities.

Lemma 1.3.9. *We have the following results :*

(a) $|Bias(\hat{F})| = |\mathrm{E}\,(\hat{F} - F)| \leq \frac{k}{2(m+1)} = \delta_m$. *Therefore* $\hat{F}$ *is asymptotically unbiased.*

(b) $\mathrm{MSE}\,(\hat{F}) \leq (\delta_m + \frac{1}{2\sqrt{n}})^2 \leq 2\,(\delta_m^2 + \frac{1}{4n})$.

(c) $\mathrm{Var}\,(\hat{F}) \leq (\delta_m + \frac{1}{2\sqrt{n}})^2 \leq 2\,(\delta_m^2 + \frac{1}{4n})$.

(d) $\mathrm{E}\left(\|F - \hat{F}\|_\infty\right) \leq \delta_m + \frac{1}{\sqrt{n}}$.

(e) $\mathrm{Var}\left(\|F - \hat{F}\|_\infty\right) \leq 2\,(\delta_m^2 + \frac{1}{n})$.

Proof.

(a) **Bias**,

$$|Bias| = |\mathrm{E}\,(\hat{F} - F)| = |\mathrm{E}\,(\hat{F} - F_n)| \leq \mathrm{E}\,|\hat{F} - F_n| \leq \delta_m.$$

(b) **MSE**,

$$\begin{aligned}
\mathrm{MSE}\,(\hat{F}) &= \mathrm{E}\,(\hat{F}-F)^2 = \mathrm{E}\,(\hat{F}-F_n+F_n-F)^2 \\
&\leq \left\{\sqrt{\mathrm{E}\,(\hat{F}-F_n)^2}+\sqrt{\mathrm{E}\,(F_n-F)^2}\right\}^2 \\
&\leq (\delta_m+\frac{1}{2\sqrt{n}})^2 \leq 2\,(\delta_m^2+\frac{1}{4\,n}).
\end{aligned}$$

(c) **Variance**, $\mathrm{Var}\,(\hat{F}) \leq \mathrm{MSE}\,(\hat{F})$.

(d) We have,

$$\mathrm{E}\left(\|F-\hat{F}\|_\infty\right) \leq \mathrm{E}\,\left(\|F-F_n\|_\infty\right)+\mathrm{E}\,\left(\|F_n-\hat{F}\|_\infty\right) \leq \frac{1}{\sqrt{n}}+\delta_m.$$

(e) We have

$$\begin{aligned}
\mathrm{Var}\left(\|F-\hat{F}\|_\infty\right) &= \mathrm{E}\,(\|F-\hat{F}\|_\infty^2)-\mathrm{E}^{\,2}(\|F-\hat{F}\|_\infty) \leq \mathrm{E}\,(\|F-\hat{F}\|_\infty^2) \\
&\leq \mathrm{E}\left(\|F-F_n\|_\infty+\|F_n-\hat{F}\|_\infty\right)^2 \\
&\leq 2\,\mathrm{E}\left(\|F-F_n\|_\infty^2+\|F_n-\hat{F}\|_\infty^2\right) \\
&\leq 2\,\left(\mathrm{E}\,\|F-F_n\|_\infty^2+\delta_m^2\right) \leq 2\left(\frac{1}{n}+\delta_m^2\right). \quad \square
\end{aligned}$$

1.4. Guidelines on the choice of the parameters k, H, m and Simulations

We start this section by explaining the choice of the different parameters involved in the construction of the estimator, i.e., k, H, and m. When m and k are chosen, we work with the nodes as in lemma 3.1. Here is a summary guideline to the choice of the parameters.

1. **The choice of k ("the smoothing parameter ")**

   The parameter k is seen here as a smoothing parameter. The larger k, the smoother the estimator. When H is taken as the cdf of a uniform distribution (on a bounded support), the basis functions, $G_{k,j}$, become a basis for the monotone splines of order $k-1$ (with variable nodes). So, to have a cubic spline, for example, we need to take $k = 4$.

   However, a large value of k renders the jumps of $\hat{F}$ more difficult to obtain. In fact, to have a jump at y_j, a multiple node of order r (say $y_j = y_{j+1} = \cdots = y_{j+r-1}$), we need to have $r \geq k$.

   In our simulations, we take $k = 4$, unless we have some knowledge about certain features of F.

2. **Choice for H (the instrumental cdf)**

   By analogy to the Bayesian approach, H is seen as a prior distribution. In fact, if H is equal to F (not possible in practice), then we obtain excellent results as mentioned in section 2 (lemma 2.3). A choice of H that is close to F will help mostly in the regions where F is flat or almost flat (in particular in the tails), that is, in those regions where the density f of F (when it exists) is very small. A prior knowledge about certain features of the distribution, like unimodality, asymmetry, concavity, discontinuity jumps, etc., dictates the choice of H, and therefore helps obtain a better estimation. When the support I of F is bounded but no other knowledge about F is available, H might be taken as the cdf of a uniform distribution.

Note that when the support I of F is not bounded, another choice than the uniform distribution is necessary. This choice will have an influence on $\hat{F}$ in the tails. If H has a discontinuity jump at a given point, the estimator will present a jump at that same point. So a prior information of this kind help get a better estimation of F especially when the sample size is small.

3. **The choice of m (link between H and F_n)**
   The parameter $m = m(n)$ depends on n in general. When H is smooth, small values for m induce a very smooth $\hat{F}$. Whereas large values of m make $\hat{F}$ stick to F_n, the edf, but still remains smooth (unless we have enough multiples nodes to make $\hat{F}$ jump).
   By analogy to Kernel methods, we may compare $1/m$ to h, the bandwidth (window). Recall (lemma 3.7) that by choosing m such that $n/m^2 \to 0$, we are able to obtain the asymptotic distribution of $\sqrt{n}\|\hat{F} - F\|_\infty$. So, we might use this condition to help us choose m.
   The choice of m is also associated with H. If one strongly believes that H is close to F, then choosing m small is good. Furthermore, if m is small relatively to n then the spacings will be very stable. If m is small, $\hat{F}$ looks like H and if m is large, $\hat{F}$ looks like F_n. Note that the choice of m is more important than the choice of k.

1.4.1. Numerical examples

To illustrate the performance of the estimator, we have chosen six examples, four of which concern cdfs with bounded supports and two with unbounded ones. In figure 1.4. we have plotted the cdfs and corresponding densities of the selected distributions together with the chosen cdf H (dashed). For each of these cdfs, we have computed 10 estimates and plotted them together with the respective true distribution (cf figure 1.5) In what follows, we comment the results obtained for the corresponding examples (cf figure 1.5) :

(a) **The standard normal ; N(0, 1)**

In the case of the standard normal distribution, we choose to take H as a $N(-0.5, (1.5)^2)$. In figure 1.5 (a) we show the graphics of 10 estimates from 10 samples with $n = 100$, $m = 15$, and $k = 4$.

(b) **The exponential distribution ; Exp(1)**

For the exponential distribution, we choose to take H as the cdf of an $Expo(0.5)$ ("large variance.") In figure 1.5 (b). we show the graphics of 10 estimates from 10 samples with $n = 100$, $m = 15$ and $k = 4$.

(c) **Mixture of beta distributions**

We choose to take the following mixture :

$$\frac{1}{2}Beta(10, 20) + \frac{1}{2}Beta(30, 20).$$

The function H is taken symmetric to reflect "absence" of prior knowledge, we choose to use a $Beta(9, 9)$.

We have taken 10 samples with $n = 100$, $m = 15$, and $k = 4$. We can see in figure 1.5 (c). that the estimator does quite well. For our simulations, we take $k = 4$ in general, unless we know about a particular feature in F, like being very smooth. In this case, we may take k greater than 4.

(d) **Another mixture of beta distributions (with a flat region)**

We choose to take another mixture with a flat area, that's :

$$\frac{1}{2}Beta(10, 40) + \frac{1}{2}Beta(40, 20).$$

We choose for H the following mixture $2/3\ Beta(4, 13) + 1/3\ Beta(25, 20)$ (suppose we had a prior knowledge that F was bimodal and we had an idea where the modes were).

We have taken 10 samples with $n = 200$, $m = 30$, and $k = 4$. We can see in figure 1.5 (d). that the estimator does well. It is clear that a prior

knowledge about the flat region would have helped in obtaining a better estimation.

(e) **A cdf with an infinite derivative**

We take the following cdf on $[0,1]$,

$$F(x) = \begin{cases} \dfrac{1-(1-2x)^{\frac{1}{8}}}{2} & \text{if} \quad x < \dfrac{1}{2} \\ \dfrac{1+(2x-1)^{\frac{1}{8}}}{2}, & \text{otherwise.} \end{cases}$$

The cdf H is chosen as a $Beta(10,10)$ (symmetric). We have taken 10 samples with $n = 100$, $m = 20$, and $k = 4$. It is clear that no polynomial based estimator would have followed the infinite slope.

(f) **A distribution with a jump**

Like in example 1.2.2, we consider a cdf F on $[0,1]$ with a discontinuity jump at $x = 1/2$. The function F is given by,

$$F(x) = \begin{cases} x^2 & \text{if} \quad x < \dfrac{1}{2} \\ \dfrac{1+(2x-1)^2}{2}, & \text{otherwise.} \end{cases}$$

The cdf H is chosen as a $U(0,1)$ to reflect the fact that no prior knowledge is available. We have taken 10 samples with $n = 200$, $m = 30$, and $k = 4$. Figure 1.5 (f). shows that the estimator does well. An accumulation of nodes provokes the jump in $\hat{F}$. Clearly, another choice of H reflecting the discontinuity jump would have lead to a better estimation.

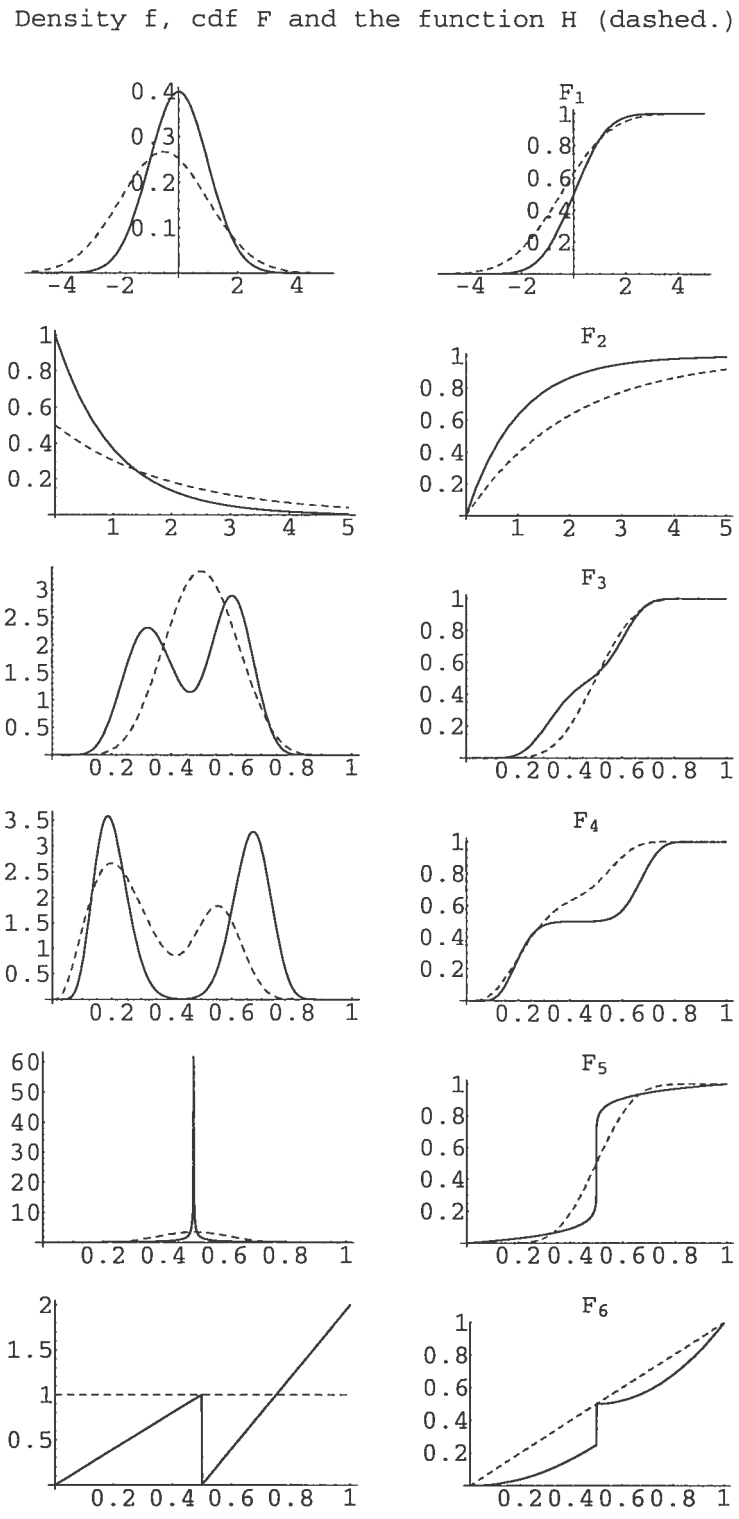


FIG. 1.4. Plots of the distributions used for numerical evaluations with the respective chosen H.

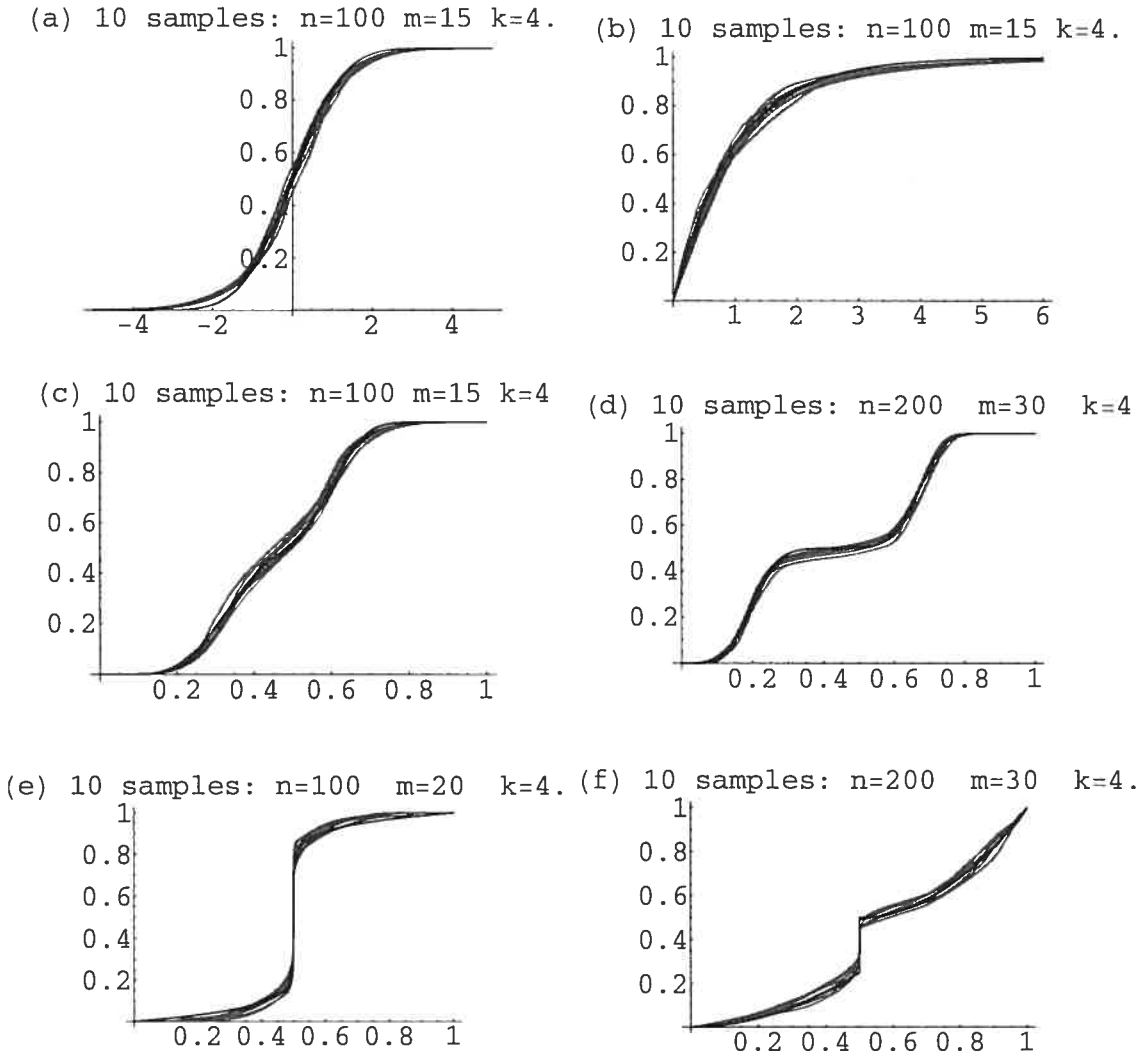


FIG. 1.5. Plots of the true distribution together with the estimates from 10 samples.

BIBLIOGRAPHIE

[1] Altman, N., Leger, C., (1995). *"Bandwidth selection for kernel distribution function estimation".* J. Statist. Plann. Inference **46**, 195-214.

[2] Azzalini, A., 1981. *"A note on the estimation of a distribution function and quantiles by a kernel method."* Biometrika **68** (1), 326-328.

[3] Bowman, A., Hall, P., Prvan, T., 1998. *"Bandwidth selection for the smoothing of distribution functions."* Biometrika **85** (4), 799-808.

[4] Babu J., G., Canty , A. J. and Chaubey, P.C. (2002). *"Application of Bernstein Polynomials for Smooth Estimation of a Distribution and Density Function."* J. Statist. plann. Inference **105,** No 2, 377-392.

[5] Chaubey, Y.P., Sen, P.K. (1996). *"On smooth estimation of survival and density functions."* Statist. Decisions **14,** 1-22.

[6] Chu, I.S., (1995). *"Bootstrap smoothing parameter selection for distribution function estimation."* Math. Japon. **41** (1), 189-197.

[7] Csáki E. (1984). *"Empirical Distribution Function."*, Handbook of Statistics (P. R. Krishnaiah and P. K. Sen, eds), **vol. 4,** 405-430.

[8] Devroye, L. and Lugosi, G. (2001). *"Combinatorial Methods in Density Estimation"*, Springer-Verlag, New York, Inc.

[9] Dvoretzky, A. Kiefer, J. and Wolfowitz, A. J. (1956) *"Asymptotic minimax character of the sample distribution function and of the classical multinomial estimator,"* Ann. Math. Statist., 27, No. 3, 642669.

[10] Efromovitch, S. (2001). *"Second Order Efficient Estimating a Smooth Distribution Function and its Applications."* Methodology and Computing in Applied Probability, **9**, 179-198.

[11] Efron, B., Tibshirani, R.J., (1993). *" An Introduction to the Bootstrap."* Chapman & Hall, London.

[12] Falk, M., (1983). *"Relative eficiency and deficiency of kernel type estimators of smooth distribution functions"*. Statist. Neerlandica **37**, 73-83.

[13] Finkelstein, H. (1971). *" The Law of the Iterated Logarithm for empirical distributions,"* Ann. Math. Statist., **42,**, 607-615.

[14] Hansen, B. M., Lauritzen S. L. (2002). *"Nonparametric Bayes inference for concave distribution functions."* Statistica Neerlandica, **56,** No 1, 110-127.

[15] He, X., Shi, P. (1998). *"Monotone B-Spline Smoothing,"* JASA, Theory and Methods, **93, No. 442**, 643-650.

[16] Jones, M.C., 1990. *"The performance of kernel density functions in kernel distribution function estimation."* Statist. Probab. Lett. **9**, 129-132.

[17] Lehmann, E.L., Casella, G. (1998.) *" Theory of Point Estimation."* 2nd ed., Springer, New York.

[18] Mammitzsch, V., (1984). *"On the asymptotically optimal solution within a certain class of kernel type estimators."* Statist. Decisions **2**, 247-255.

[19] Massart, P. (1990). *"The Tight Constant in the Dvoretsky-Kiefer-Wolfowitz Inequality"*. Annals of probability, **18**, 1269-1283.

[20] Nadaraya, E.A., (1964). *"Some new estimates for distribution functions"*. Theory Probab. Appl. **15** , 497-500.

[21] Perron, F. and Mengersen, K. (2001). *"Bayesian Nonparametric Modeling Using Mixtures of Triangular Distributions"* Biometrics **57,** 518-528.

[22] Restle E. Maria (1999). *" Estimating Distribution Functions with smoothing Splines"*, Technical Report Statap.1999.5 (DMA, EPF Lausanne, CH.)

[23] Sarda, P., (1993). *"Smoothing parameter selection for smooth distribution functions"*. J. Statist. Plann. Inference **35**, 65-75.

[24] Serfling, J. Robert (1980). *"Approximation Theorems of Mathematical Statistics"*, John Wiley & Sons, Inc..

[25] Shao, Y., Xiang, X., (1997). *"Some extensions of the asymptotics of a kernel estimator of a distribution function."* Statist. Probab. Lett. **34**, 301-308.

[26] Shirahata, S., Chu, I.S., (1992). *"Integrated squared error of kernel-type estimator of distribution function"*. Ann. Inst. Statist. Math. **44** (3), 579-591.

[27] Stute, W. (1982). *"The oscilation behavior of empirical processes."* Ann. Probability, **10**, 86-107.

[28] Swanepoel, J.W.H., (1988). *"Mean integrated squared error properties and optimal kernels when estimating a distribution function"*. Comm. Statist. Theory Methods **17** (11), 3785-3799.

[29] de Uña-Álvarez, J., González-Manteiga, W., Cadarso-Suárez, C., (2000) *"Kernel distribution function estimation under the Koziol-Green model."* Journal of Statistical Planning and Inference **87**, pp. 199-219.

[30] Wahba G. (1976). *"Histosplines with knots which are order statistics,"*, Journal of the Royal Statistical Society, **B, 38** 140-151.

Chapitre 2

ADAPTIVE ESTIMATION OF A DENSITY FUNCTION USING FINITE MIXTURES

«Le hockey sur glace est un savant mélange de glisse acrobatique et de Seconde Guerre mondiale.»
Alfred Hitchcock.

«L'intelligence de la vie... Ce mélange si particulier de respect des convenances et de largeur d'esprit, cette faculté de comprendre avant de savoir.»
Jean Dutourd.
(Extrait des Mémoires de Mary Watson.)

Abstract.

We propose an estimator based on a convex combination of densities (pdf) with supports that vary according to the order statistics. It therefore puts different amounts of smoothing at different locations. In a way, our estimator generalizes the histogram with random partitions (and the histo-spline) while being smooth(er) and a pdf by construction. The properties of the proposed estimator are studied. The rate for the uniform almost sure convergence is given. Simulations and examples illustrate the approach. The performances of the estimator are compared to that of the kernel method.

Key Words : Non-parametric density estimation, Adaptive smoothing, Splines, Uniform consistency.

2.1. INTRODUCTION

We are interested in estimating a probability density function (pdf) f with support on an interval I of $\mathbb{R}$, bounded or not, based on a sample $X_1, X_2, \ldots, X_n$ from f. We adopt a nonparametric approach, thus no functional form for the underlying density is assumed. Here, the estimator $\hat{f}$ for the density function is obtained by differentiating $\hat{F}$, the estimator for the cdf F obtained in Haddou and Perron (2006). Several works have been devoted to the estimation of a density function. Perhaps the most commonly used and studied estimators are the histogram and the kernel estimator. One of the advantages that made the kernel technique so popular is its ease of implementation. However, this method suffers from some well known drawbacks like its extreme sensitivity to the choice of the "window" width (bandwidth). It has as well the tendency to exhibit bumpy behavior in the tails and sometimes puts positive mass outside the support of the density (Silverman, 1983). Being designed for continuous pdf, the kernel method fails to estimate discontinuities at boundary (exponential pdf). The estimator we propose here is a pdf by construction and has some similar features with the variable (adaptive) kernel density estimator in that it puts different amounts of smoothing at different locations. The present work has two goals, approximating and estimating. In section 2, we develop a method for *approximating* a (known) pdf f. The approximation consists of a finite convex combination of densities (we call basis functions) with varying supports that depend on a vector of nodes. In the estimation step, the nodes are given by the quantiles and the support sizes are randomly determined by the spacings of the order statistics (section 3). The approximation depends on three meta-parameters, namely k a smoothing parameter, m the number of nodes and h an instrumental pdf (looked at as a pseudo-prior). For $k = 2$ and h uniform (I bounded) we obtain a version of the famous histogram estimator with random partitions (Van Ryzin 1973, Abou-Jaoude 1976, Devroye 1985, 1987). The choice $k > 2$ and h uniform or polynomial leads to a version of the histospline estimator (Wahba, 1976). To obtain a cubic

spline, we may take for example, $k = 5$ and h uniform. Uniform upper bounds in the sup norm are given in corollary 2.2.
In section 2.3, we use the approximation developed in section 2.2 to construct the estimator. Uniform almost sure convergence is established and the rate of this convergence is given in Lemma 2.3.1. Section 4 is dedicated to simulation studies. Guidelines for the choice of the parameters and for the implementation are given. The performances of the estimator are compared to that of the kernel method. The examples include densities with discontinuities at boundary, mass at endpoints of support, and a real data example (galaxies data).

The reader interested in the practical aspects of the method; implementation and choice of the different parameters may skip section 2. In this work, we consider densities with bounded support though the method works on unbounded ones as well.

2.1.1. Approximation of the density function

In this section we discuss the problem of *approximating* a (known) pdf f defined on an interval $I \subset \mathbb{R}$. The approximation we propose is obtained by differentiating G_k an approximation of F given in [1]. We present in what follows the approximation G_k together with some results.

2.1.2. The mixture

Let $a = \inf\{x\colon x \in I\}$ and $b = \sup\{x\colon x \in I\}$. The approximation G_k we choose to work with, is based on m points $a \leq y_1 \leq \cdots \leq y_m \leq b$ called nodes (knots) and a parameter k called the smoothing parameter. For convenience, we set $y_j = a$ for $j = -k+2, \ldots, 0$ and $y_j = b$ for $j = m+1, \ldots, m+k-1$ (multiples nodes of order at least $(k-1)$ at the end points a and b). The function G_k is a finite mixture and satisfies the following representation

$$G_k = \sum_{j=1}^{m+k-1} \omega_{kj}\, G_{k,j} = \frac{1}{(k-1)(m+1)} \sum_{j=1}^{m+1} \sum_{l=j}^{j+k-2} G_{k,l},$$

with weights

$$\omega_{kj} = \frac{\min(j, k-1, m+k-j)}{(k-1)(m+1)} \quad \text{for} \quad j = 1, 2, \ldots, m+k-1.$$

The weights ω_{kj} are therefore positive and add up to 1. Each function $G_{k,j}$ is a cdf on the interval $\mathrm{I}_{k,j} = [y_{j-k+1}, y_j)$ for $j = 1, 2, \ldots, m+k-1$. The elements $G_{k,j}$ are called basis functions. A construction of $G_{k,j}$ is given in the next section. We further suppose that $m \geq k-1 > 0$. Therefore, the function G_k is a cdf by construction.

2.1.3. The basis

The basis elements are such that $G_{k,j} \geq G_{k,j+1}$ for $j = 1, 2, \ldots, m+k-2$. For convenience, we set $G_{k,0} = 1$ and $G_{k,m+k} = 0$. There is a hierarchical structure in the construction of the basis based on a fixed cdf H on I. Let us define for $j = 1, 2, \ldots, m+k-1$, the following functions

$$H_{k,j}(x) = \frac{H(x) - H(y_{j-k+1})}{H(y_j) - H(y_{j-k+1})} \mathrm{I}_{kj}(x) + \mathrm{I}(x \geq y_j).$$

Then the hierarchical structure is the following

Step 1 (initial step) : $G_{1,j}(x) = I(x \geq y_j)$ for all $x \in I$ and $j = 1, \ldots, m$ (Dirac cdf).

Step $l+1$ **:** $G_{l+1,j} = H_{l+1,j}\, G_{l,j-1} + (1 - H_{l+1,j})\, G_{l,j}$.

Clearly $G_{k,j} \geq G_{k,j+1}$ for all k and j, $j = 1, 2, \ldots, m+k-2$. The fact that $G_{k,j}$ is a cdf on $I_{k,j} = [y_{j-k+1}, y_j)$ comes from lemma 2.1 in [1].

2.1.4. The choice of the nodes and the parameters H and k

Lemma 2.1.1. (Choice of the nodes)
If we take $y_i = F^{-1}(i/(m+1))$ *for* $i = 1, \ldots, m$, *where* F^{-1} *denotes the generalized inverse of* F, *then we obtain*

$$\|F - G_k\|_\infty \leq \frac{k}{2(m+1)}.$$

We need to emphasize here that the choice of the nodes is really critical. The function H is like a guess for F. When the nodes are adequately selected, a poor choice for H cannot be dramatic. However, we hope that a perfect guess will imply that $G_k = F$ for all $k > 1$. In fact, this is true in the case where F is continuous see next lemma.

Lemma 2.1.2. ($H \equiv F \implies G_k \equiv F$)
If F is a continuous cdf on an interval $I \subset \mathbb{R}$, $y_j = F^{-1}(j/(m+1))$ for $j = 1, 2, \ldots, m$ and $H = F$ then $G_k = F$ for all k, $1 < k \leq m+1$.

2.2. The approximation step

In this section, we suppose F continuously differentiable on $I = [a, b] \subset \mathbb{R}$. Let us denote the approximation G_k by $\hat{F}_k$ and the elements of the basis by F_{kj}. We have,

$$\hat{F}_k(x) = \sum_{j=1}^{m+k-1} \omega_{kj} F_{kj}(x) = \sum_{i=0}^{m} \left\{ \sum_{j=1}^{i} \omega_{kj} + \sum_{j=i+1}^{i+k-1} \omega_{kj} F_{kj}(x) \right\} \mathrm{I}_{[y_i, y_{i+1})}(x),$$

which gives by differentiation

$$\hat{f}_k(x) = \sum_{i=0}^{m} \left\{ \sum_{j=i+1}^{i+k-1} \omega_{kj} f_{kj}(x) \right\} \mathrm{I}_{[y_i, y_{i+1})}(x), \qquad \text{where} \quad f_{kj}(x) = F'_{kj}(x).$$

Suppose the nodes are chosen like in lemma 1.1 and let $\Delta_{kj} = y_j - y_{j-k+1}$, $\Delta F_{kj} = F(y_j) - F(y_{j-k+1})$ to have $\omega_{kj} = \Delta F_{kj}/(k-1)$. We obtain

$$(k-1)(\hat{f}_k - f) = \sum_{j=i+1}^{i+k-1} \left\{ \frac{\Delta F_{kj}}{\Delta_{kj}} - f \right\} \Delta_{kj} f_{kj} + f \left\{ S_{k,i} - (k-1) \right\}, \quad \text{for} \quad x \in [y_i; y_{i+1}), \tag{2.2.1}$$

where $S_{k,i} = \sum_{j=i+1}^{i+k-1} \Delta_{kj} f_{kj}$. We have the following lemma (for the proof see Appendix A).

Lemma 2.2.1. *With the above notations we obtain,*
(a) If H is the cdf of a uniform distribution on I, then $S_{k,i}(x) = k-1$ for all $i = 0, 1, \ldots, m$.

(b) If H is continuously differentiable, $h = H'$ Lipschitz with constant L_h, $I_h = \inf\{h(x),\, x \in I\} > 0$, then there exists constants $B_{k,h}$ and $C_{k,h}$ such that, $\|S_{k,i}\|_\infty \leq B_{k,h}$ and

$$\sup_{y_i \leq x < y_{i+1}} |S_{k,i}(x) - (k-1)| \leq C_{k,h} \max\{y_{i+k-1} - y_i, y_{i+1} - y_{i-k+1}\}. \qquad (2.2.2)$$

Corollary 2.2.1. *If F is differentiable with density f Lipschitz and $h = H'$ Lipschitz with $I_h = \inf_{x \in I}(h(x)) > 0$, then we obtain*

(a) There exists a constant $B_{f,h,k}$, depending on f, h and k such that

$$\sup_{y_i \leq x < y_{i+1}} |\hat{f}_k(x) - f(x)| \leq B_{f,h,k} \max\{y_{i+k-1} - y_i, y_{i+1} - y_{i-k+1}\},$$

Note that if H is the cdf of a uniform distribution on I, then $B_{f,h,k} = L_f$.

(b) If F is the cdf of a uniform distribution on I, then

$$\|\hat{f}_k - f\|_\infty \leq B_{h,k} \frac{k-1}{m+1}.$$

(c) If f is such that $I_f = \inf_{x \in I}(f(x)) > 0$, then

$$\|\hat{f}_k - f\|_\infty \leq \frac{B_{f,h,k}}{I_f} \frac{k-1}{m+1}.$$

Proof.

From equation (2.2.1) and for all $x \in [y_i, y_{i+1})$, there exists $x_{kj} \in I_{kj}, j = i+1, \ldots, i+k-1$ such that

$$\begin{aligned}
(k-1)|\hat{f}_k(x) - f(x)| &\leq \sum_{j=i+1}^{i+k-1} |f(x_{kj}) - f(x)| \Delta_{kj} f_{kj} + f(x)\, |S_{k,i}(x) - (k-1)| \\
&\leq L_f \sum_{j=i+1}^{i+k-1} |x_{kj} - x| \Delta_{kj} f_{kj} + f(x)\, |S_{k,i}(x) - (k-1)| \\
&\leq L_f \max\{y_{i+k-1} - y_i, y_{i+1} - y_{i-k+1}\}\, S_{k,i}(x) \\
&+ f(x)\, |S_{k,i}(x) - (k-1)|.
\end{aligned}$$

So that (a) and (b) follow from lemma 2.1. For (c), note that $y_j = F^{-1}(j/(m+1))$.

2.3. THE ESTIMATION STEP

In this section we suppose F continuously differentiable with density $f > 0$ Lipschitz and $h = H'$ Lipschitz with $I_h = \inf_{x \in I}(h(x)) > 0$. The estimator $\hat{f}$ for the density function is obtained by differentiating $\widehat{F}$ the estimator for the cdf F proposed in [4]. We obtain

$$\hat{f}_k(x) = \sum_{j=1}^{m+k-1} \omega_{kj}\, f_{kj}(x) = \sum_{i=0}^{m} \left\{ \sum_{j=i+1}^{i+k-1} \omega_{kj}\, f_{kj}(x) \right\} \mathrm{I}_{[y_i, y_{i+1})}(x),$$

where $f_{kj}(x) = F'_{kj}(x)$. This estimator is a generalization of some well known estimators. Indeed, the histogram with random partition and histosplines appear as special cases. This generalization is, however, a pdf by construction. When k, the smoothing parameter, is set to 2, the weights $\omega_{2,i}$ are all equal to $1/(m+1)$. Now if $h = H'$ is chosen as a uniform distribution on $I = [a, b]$, we obtain a version of the well-known histogram estimator with random partition ;

$$\hat{f}_2(x) = \sum_{j=1}^{m+1} \omega_{2j}\, f_{2j}(x) = \sum_{j=1}^{m+1} \omega_{2,j+1}\, h_{2j}(x) = \sum_{i=0}^{m} \frac{1}{(m+1)(y_{i+1} - y_i)} \mathrm{I}_{[y_i, y_{i+1})}(x).$$

If F_n, the empirical distribution function (edf), is used to choose the nodes then $y_i = F_n^{-1}(i/(m+1)) = x_{(\lceil n \frac{i}{m+1} \rceil)}$, $i = 1, \ldots, m$, $x_{(\cdot)}$ being the vector of the order statistics. This estimator has been extensively studied by many authors : Van Ryzin (1973), Abou-Jaoude (1976), Devroye (1985, 1987), and others. Assume that n is a multiple of $m+1$. In this case, we have $(k-1)\omega_{kj} = F_n(y_j) - F_n(y_{j-k+1}) \equiv \Delta F_{n,kj}$. We obtain,

$$\hat{f}_k(x) = \sum_{j=i+1}^{i+k-1} \frac{\Delta F_{n,kj}}{(k-1)} f_{kj}(x), \qquad \text{for} \quad x \in [y_i; y_{i+1}).$$

We may write,

$$(k-1)(\hat{f}_k - f) = \sum_{j=i+1}^{i+k-1} \left\{ \frac{\Delta F_{kj}}{\Delta_{kj}} - f \right\} \Delta_{kj} f_{kj} + f \{S_{k,i} - (k-1)\} + \sum_{j=i+1}^{i+k-1} \frac{\Delta F_{n,kj} - \Delta F_{kj}}{\Delta_{kj}} \Delta_{kj} f_{kj},$$

so that,

$$(k-1)|\hat{f}_k - f| \leq \sum_{j=i+1}^{i+k-1} \left| \frac{\Delta F_{kj}}{\Delta_{kj}} - f \right| \Delta_{kj} f_{kj} + f|S_{k,i} - (k-1)| + 2\, S_{k,i} \frac{\|F_n - F\|_\infty}{y_{i+1} - y_i}.$$

Using Lemma 2.1., we obtain the inequality

$$\sup_{y_i \leq x < y_{i+1}} |\hat{f}_k(x) - f(x)| \leq B_{f,h,k} \max\{y_{i+k-1} - y_i, y_{i+1} - y_{i-k+1}\} + C_{k,h} \frac{\|F_n - F\|_\infty}{y_{i+1} - y_i}. \tag{2.3.1}$$

Which gives, with probability one as $n \to \infty$

$$\|\hat{f}_k - f\|_\infty = O\left(\frac{1}{m}\right) + O\left(m\sqrt{\frac{\ln \ln n}{n}}\right), \tag{2.3.2}$$

since $\Delta y = O(1/m)$ and according to the law of the iterated logarithm we have $\|F_n - F\|_\infty = O\left(\sqrt{(\ln \ln n)/n}\right)$ *a.s. as* $n \to \infty$. At this stage we need to suppose that $m = m(n)$ satisfies $m^2(\ln \ln n)/n \to 0$ as $m, n \to \infty$. The rate of convergence of the second term in (2.3.2) might be improved using Bahadur's Lemma (Bahadur, 1966, Lemma 1). In fact, choosing m such that $m^{-1} = o(\ln(n)/\sqrt{n})$ will insure that $\Delta y < \ln(n)/\sqrt{n} = a_n$ (for n large enough) and therefore

$$|\Delta F_{n,kj} - \Delta F_{kj}| \leq \sup_{|x-t|<a_n} |F_n(x) - F_n(t) - F(x) + F(t)| = O(\ln(n)/n^{3/4}).$$

Which leads to next Lemma.

Lemma 2.3.1. *Suppose F differentiable with density f Lipschitz and $0 < m \leq f \leq M < \infty$. Suppose h has the same properties. Then, with probability one, we*

obtain as $n \longrightarrow \infty$

$$\|\hat{f}_k - f\|_\infty = O\left(\frac{1}{m}\right) + O\left(m\,\frac{\ln(n)}{n^{3/4}}\right). \tag{2.3.3}$$

2.4. SIMULATION STUDY

The parameter k is related with the smoothness at the nodes. In our simulations, we take $k = 5$ unless we have some knowledge about certain features of f. A choice of h that is close to f will help mostly in those regions where f is (very) small. A prior knowledge about certain features of the pdf, like bimodality, concavity, discontinuity jump, etc., dictates the choice of h, and therefore helps in obtaining a better estimation (see Lemma 1.2). When the support I of f is bounded but no other knowledge about f is available, h might be taken as the pdf of a uniform distribution. In this case, $\hat{f}$ becomes a spline of order less or equal to $k-2$. A cubic spline is obtained by choosing $k = 5$ (and h uniform). In general, splines are obtained by choosing h polynomial. Note that $\hat{f}$ is a pdf by construction unlike the histospline for example. If one strongly believes that h is close to f, then choosing m small is good. On the other hand, large values of m cause $\hat{f}$ to wiggle. The effect of m on the smoothness of f is similar to that of the bandwidth in kernel density estimation. We like to look at h as a pseudo-prior (by analogy to the Bayesian approach). Our estimator has some similar features with the variable/adaptive kernel density estimator in that it puts different amounts of smoothing at different locations. The estimator being a finite mixture of densities with supports that vary in length, smaller in the areas of accumulation of the observations.

To illustrate the performance of the estimator, we choose the three following pdfs : (a) a bimodal density ; mixture of two beta distributions (very smooth), (b) another mixture of beta distributions with a discontinuity at boundary, an almost flat area in the middle and a sharp slope at the end of the support ; and (c) the galaxies data. To obtain good results, we usually follow the following procedure

especially for small sample sizes : if a prior information is available on f, then we will incorporate it in h, otherwise we start by computing a pilot estimator ($k = 5$, h uniform or spread (large variance), m relatively large) to have an idea about the ("gross") shape of f. This pilot estimate will then be used to choose another h that mimics the shape of the pilot estimate (especially in those areas of accumulation of data) but preferably more spread than that one. This second choice will lead to an estimator that will confirm the first shape with probably other details. We may repeat this procedure one or two more times. After the second stage, the shape will not change that much. The last step would be to use the last h proposed but this time with a smaller m. In what follows, we comment on the results obtained for the chosen examples.

(a) **Mixture of two beta densities**

The mixture : $0.5\,\text{Beta}(10, 20) + 0.5\,\text{Beta}(30, 20)$. At top of figure 2.4, we find on the left three successive choices for h together with the pdf (solid), and on the right the corresponding estimates. Here $n = 200$ and k is set to 5. We proceed in the following way :

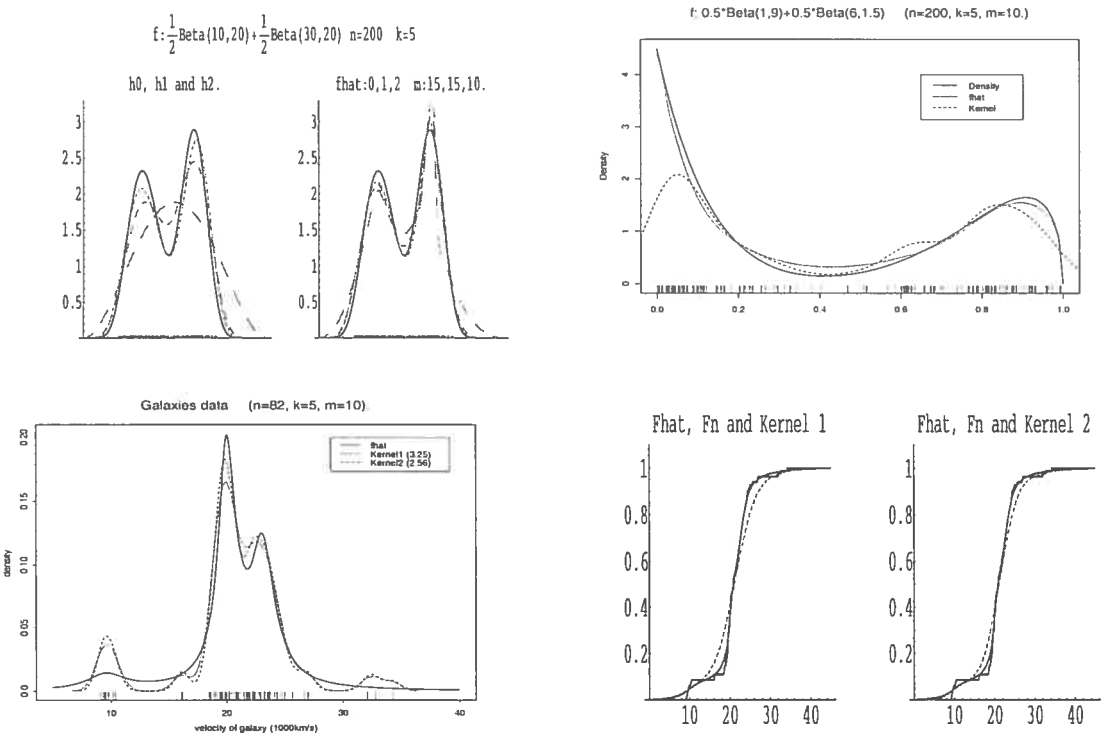
- `Step 1` : We start with $h_1 \equiv \text{Beta}(3, 3)$; a symmetric density, "large" variance to reflect "absence" of prior knowledge. The (pilot) estimate is then computed using $m = 15$.
- `Step 2` : A second choice for h that mimics the pilot estimate is then considered. The pilot showing two clear modes, we choose to take for h_2 a bimodal density on $[0, 1]$, that is a $0.5\,\text{Beta}(7, 13) + 0.5\,\text{Beta}(20, 13)$. This new h should be "flatter" than the estimate itself (in the middle). At this stage, since we are more confident in h (pseudo-prior), we may use a smaller m. We still decide to compute the new estimate using $m = 15$. This new estimate confirms the two modes and almost sticks to the target pdf.

- `Step 3` : The last choice for h confirms the two modes and the amplitude of the valley in the middle. The estimate is then computed using $m = 10$ and gives a very good result.

(b) **A mixture with a discontinuity, a flat region and a sharp slope.** The mixture : $0.5\,\text{Beta}(1, 9) + 0.5\,\text{Beta}(6, 1.5)$. This choice combines three interesting features; a discontinuity at $x = 0$, a quasi flat area in the middle and a sharp slope at the end support. The discontinuity is similar to that found in the exponential distribution. In Figure 2.1, we find the plots of the target pdf f, the estimate $\hat{f}$, and an estimate based on the kernel method using the default bandwidth in Splus. We see that our estimator handles well the discontinuity at $x = 0$ and the sharp slope at the end of the support unlike the kernel based estimator. Clearly one would expect the kernel to have that kind of noisy behavior near $x = 0$, being designed for continuous densities.

(c) **The galaxies data**

The data set was first described in Roeder (1990) and subsequently analyzed under different mixture models by several authors. It consists of the velocities of 82 distant galaxies, diverging from our own galaxy. Figure 2.1 bottom shows two (optimal) Gaussian kernel density estimates with bandwidths chosen by two variants of the Sheather-Jones method (bandwidth=2.56 and 3.25), see Venables & Ripley (2002.) The kernel shows the typical bumpy behavior in the tail. To smooth the bumps at the tails, one needs to take a bandwidth greater than that of Kernel 1 ($bandwidth = 3.25$), but this will smooth the valley in the middle resulting in an almost single mode! The estimator $\hat{f}$ however has heavy tails which is consequent with the sparse data we find in the tails but shows a deeper valley than the two kernel based estimators. Which one of these estimators best represents the data? To answer this (difficult) question, one may plot the corresponding cdf estimators together with F_n as shown



in Figure 2.1 (bottom). One area of interest is the middle part (accumulation of data). Unlike $\hat{F}$ (solid) and F_n, the two kernel based cdf estimates (dashed) do not (clearly) show the inflection point(s) corresponding to the valley. Furthermore, for a sample size of $n = 200$, one expects the estimates to "stick" to F_n which clearly the two kernel estimates do not do ($\mathrm{E}\|F_n - F\|_\infty \leq 1/\sqrt{n} \approx 0.07$).

Finally, the method has been tested on a variety of examples (not cited here) and gave very good results. The examples included densities with discontinuities at boundary (like the exponential pdf), multimodal pdf, pdf with mass at end of support (the Beta(0.5, 0.5)), and others.

APPENDIX

A. Proof of lemma 2.1 : The proof is done by induction on k.

(a) For $k = 2$ and $x \in [y_i, y_{i+1})$ we have $S_{2,i} = \Delta_{2i} h_{2i} = 1$. Then it is easy to show that $S_{k+1,i} = S_{k,i} + 1$.

(b) For $k = 2$ and $x \in [y_i, y_{i+1})$ we have

$$S_{2,i} = \Delta_{2i} h_{2i} = \frac{y_{i+1} - y_i}{H(y_{i+1}) - H(y_i)}\, h(x) = \frac{h(x)}{h(x^*_{2i})} \leq \frac{\|h\|_\infty}{I_h} = B_{k,h},$$

where $x^*_{2i} \in [y_i, y_{i+1})$. Furthermore,

$$|S_{2,i}(x) - 1| = \frac{h(x) - h(x^*_{2i})}{h(x^*_{2i})} \leq \frac{L_h}{I_h} |x - x^*_{2i}| \leq \frac{L_h}{I_h} (y_{i+1} - y_i).$$

Suppose equation (2.2.2) was true up to k. Write $S_{k+1} - k = \Sigma_1 - (k-1) + \Sigma_2 - 1$, which gives on the one hand,

$$\begin{aligned} |\Sigma_2 - 1| &\leq \sum_{i+1}^{i+k} |h_{k+1,j} - 1|(F_{k,j+1} - F_{k,j}) \leq \frac{L_h}{I_h} \sum_{i+1}^{i+k} |x - x^*_{k+1,j}|(F_{k,j+1} - F_{k,j}) \\ &\leq \frac{L_h}{I_h} \max(y_{i+k} - y_i; y_{i+1} - y_{i-k+1}) \leq \frac{L_h}{I_h} (y_{i+k} - y_{i-k+1}), \end{aligned}$$

and on the other hand,

$$\Sigma_1 - 1 = \sum_{j=i+1}^{i+k-1} (Q_{1,j} - 1)\Delta_{k,j} f_{kj} + [S_{k,i} - (k-1)],$$

where

$$\begin{aligned} Q_{1,j} &= \frac{\Delta_{k+1,j+1}}{\Delta H_{k+1,j+1}} \frac{H(x) - H(y_{j-k+1})}{\Delta_{k,j}} + \frac{\Delta_{k+1,j}}{\Delta H_{k+1,j}} \frac{H(y_j) - H(x)}{\Delta_{k,j}} \\ &= \frac{\Delta_{k+1,j+1}}{\Delta H_{k+1,j+1}} \frac{\Delta H_{k,j}}{\Delta_{k,j}} - \left(\frac{\Delta_{k+1,j+1}}{\Delta H_{k+1,j+1}} - \frac{\Delta_{k+1,j}}{\Delta H_{k+1,j}} \right) \frac{H(y_j) - H(x)}{\Delta_{k,j}}. \end{aligned}$$

Thus,

$$\begin{aligned} |Q_{1,j} - 1| &\leq \left| \frac{h(x_{kj})}{h(x^*_{k+1,j+1})} - 1 \right| + \left| \frac{1}{h(x^*_{k+1,j+1})} - \frac{1}{h(x^*_{k+1,j})} \right| \frac{\Delta H_{k,j}}{\Delta_{k,j}} \\ &\leq \frac{L_h}{I_h} |x^*_{k,j} - x^*_{k+1,j+1}| + \frac{L_h}{I_h^2} |x^*_{k+1,j} - x^*_{k+1,j+1}| B_h \\ &\leq \frac{L_h}{I_h} \left(1 + \frac{B_h}{I_h} \right) \max(y_{i+k} - y_i; y_{i+1} - y_{i-k+1}). \end{aligned}$$

So that,

$$|\Sigma_1 - (k-1)| \leq C_h \max(y_{i+k} - y_i; y_{i+1} - y_{i-k+1})\, S_{k,i} + |S_{k,i} - (k-1)|.$$

Finally,

$$\begin{aligned} |S_{k+1} - k| &\leq \left(C_h + \frac{L_h}{I_h}\right) \max(y_{i+k} - y_i; y_{i+1} - y_{i-k+1})\, S_{k,i} + |S_{k,i} - (k-1)| \\ &\leq B_{f,h,k}\,(y_{i+k} - y_{i-k+1}). \end{aligned}$$

BIBLIOGRAPHIE

[1] Abou-Jaoude S. (1976) *Sur la convergence L_1 et L_∞ de l'estimateur de la partition aléatoire pour une densité.* Ann. Inst. Henri Poincaré **Vol. XII, 4,** 299-317.

[2] Bahadur, R.R. (1966) *"A note on quantiles in large samples."* Ann. Math. Statist. **37,** 577-580.

[3] Devroye, L.,and Györfi, L. (1985). *Nonparametric Density Estimation : The L_1 View,* Wiley, N.Y.

[4] Haddou, M. and Perron, F. (2006) *Nonparametric estimation of a cdf with mixtures of cdf concentrated on small intervals.* Technical report CRM-3210, January 2006 (Université de Montréal).

[5] Kim B. K. and Van Ryzin J. (1975) *Uniform consistency of a histogram density estimator and modal estimation.* Communications in Statistics, **4(4),** 303-315.

[6] Kim B. K. and Van Ryzin J. (1980) *On the asymptotic distribution of a histogram density estimator.* Nonparametric statistical inference, Vol. I, II (Budapest, 1980), 483–499, Colloq. Math. Soc. János Bolyai, 32, North-Holland, Amsterdam, 1982.

[7] Roeder, K. (1990) *Density estimation with confidence sets exemplified by superclusters and voids in the galaxies.* J.A.S.A. **85** 617-624.

[8] Silverman B. (1993) *Density Estimation for Statistics and Data Analysis.* Chapman and Hall, London.

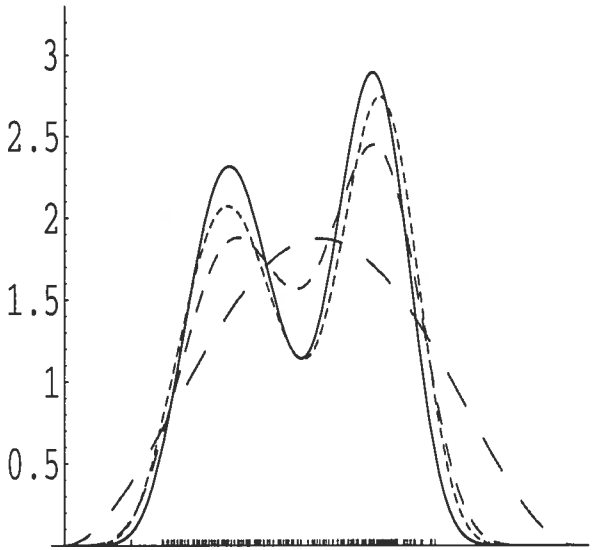
[9] Van Ryzin J. (1973) *A histogram method of density estimation.* Comm. in Stats.,**2(6),** 493-506.

[10] Venables, W.N., Ripley, B.D. (2002) *Modern applied statistics with S.* Springer, New York.

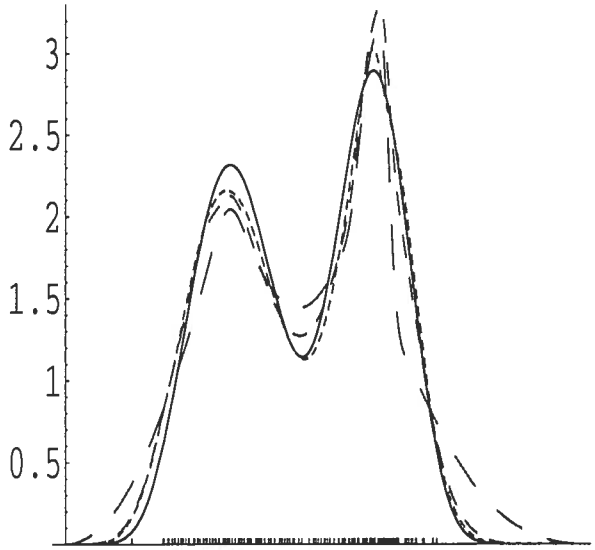
[11] Wahba G. (1976). *"Histosplines with knots which are order statistics"*, J.R.S.S. **B, 38** 140-151.

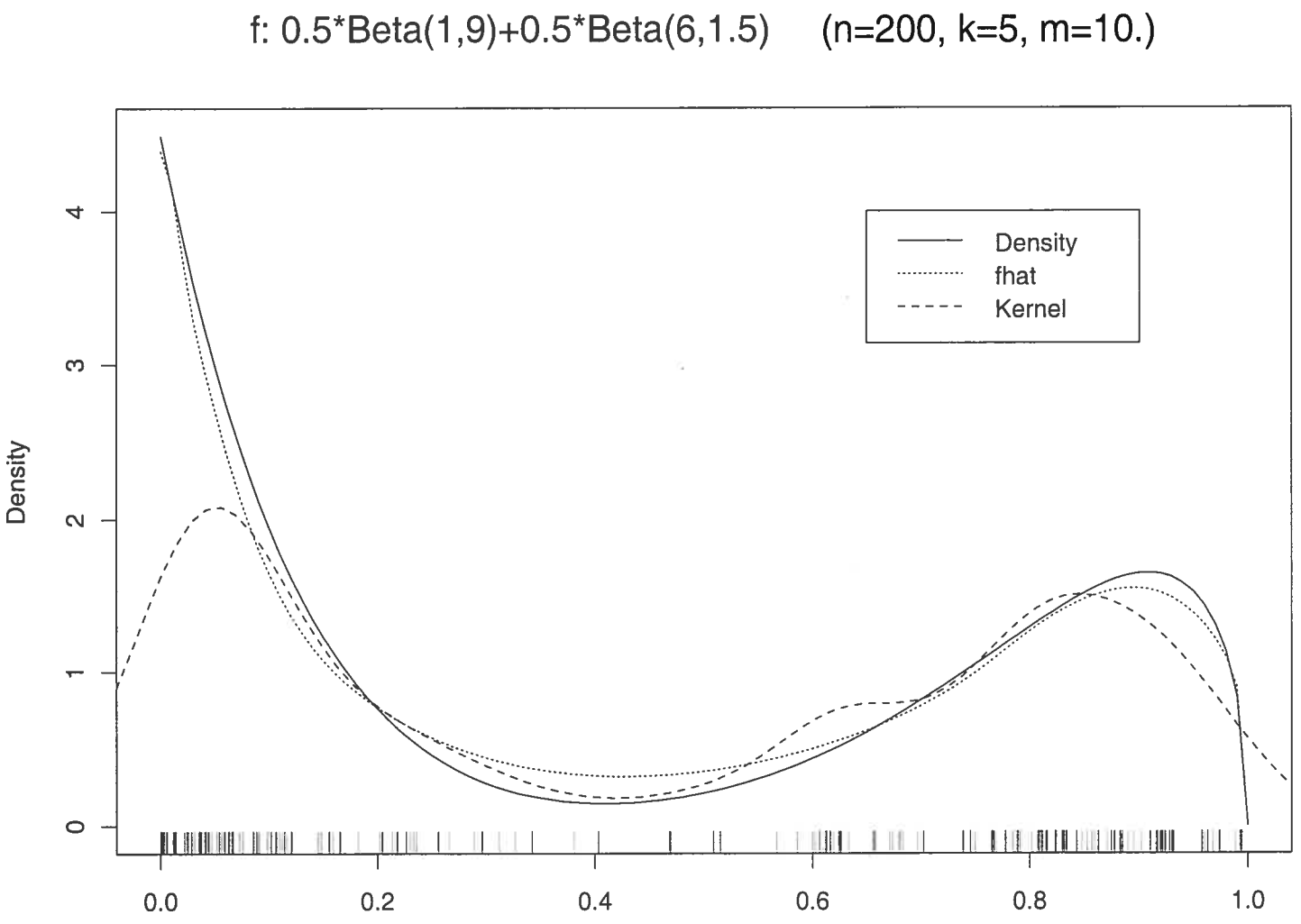
$$f: \frac{1}{2}\text{Beta}(10,20)+\frac{1}{2}\text{Beta}(30,20) \quad n=200 \quad k=5$$

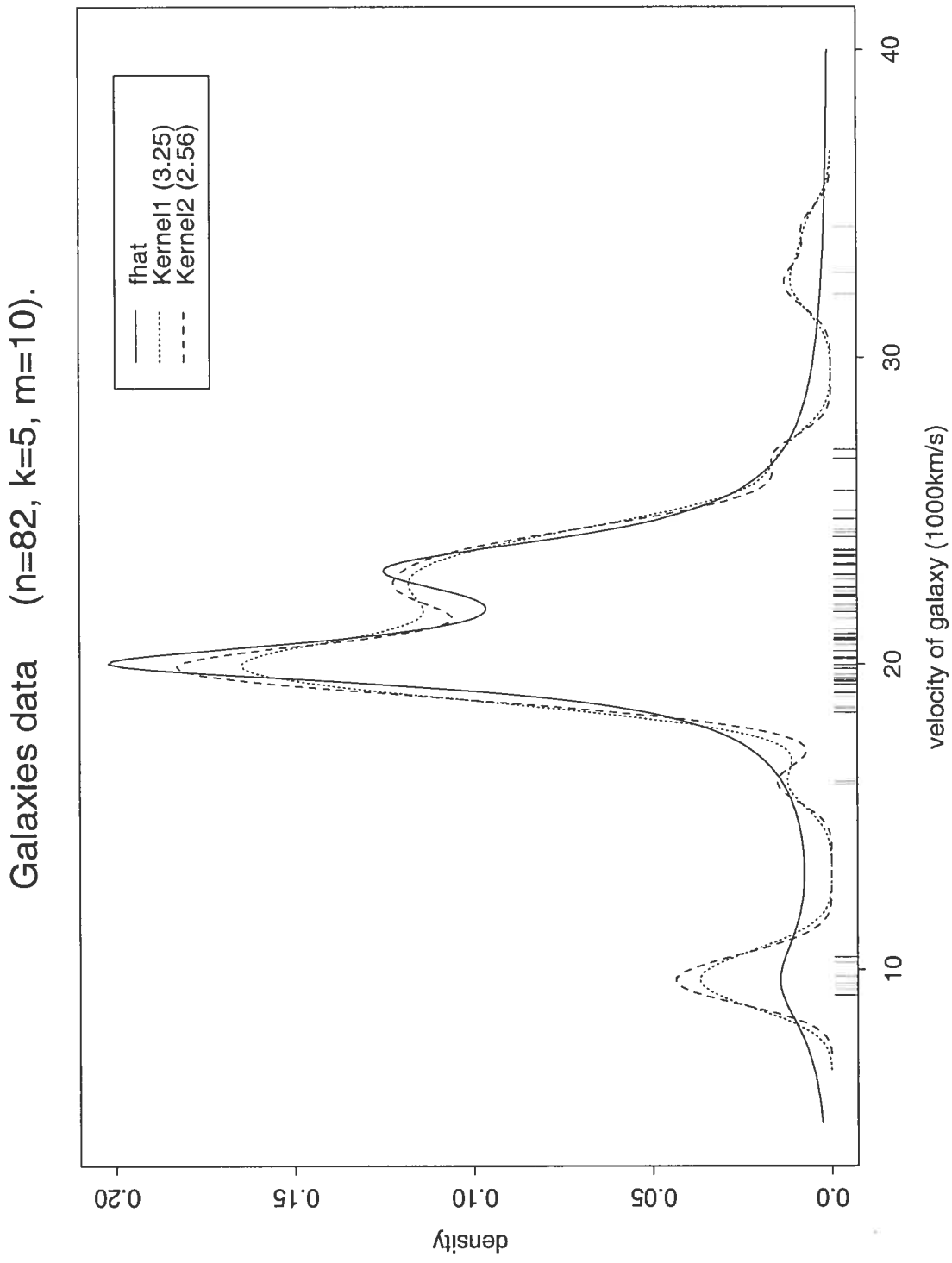
h0, h1 and h2.

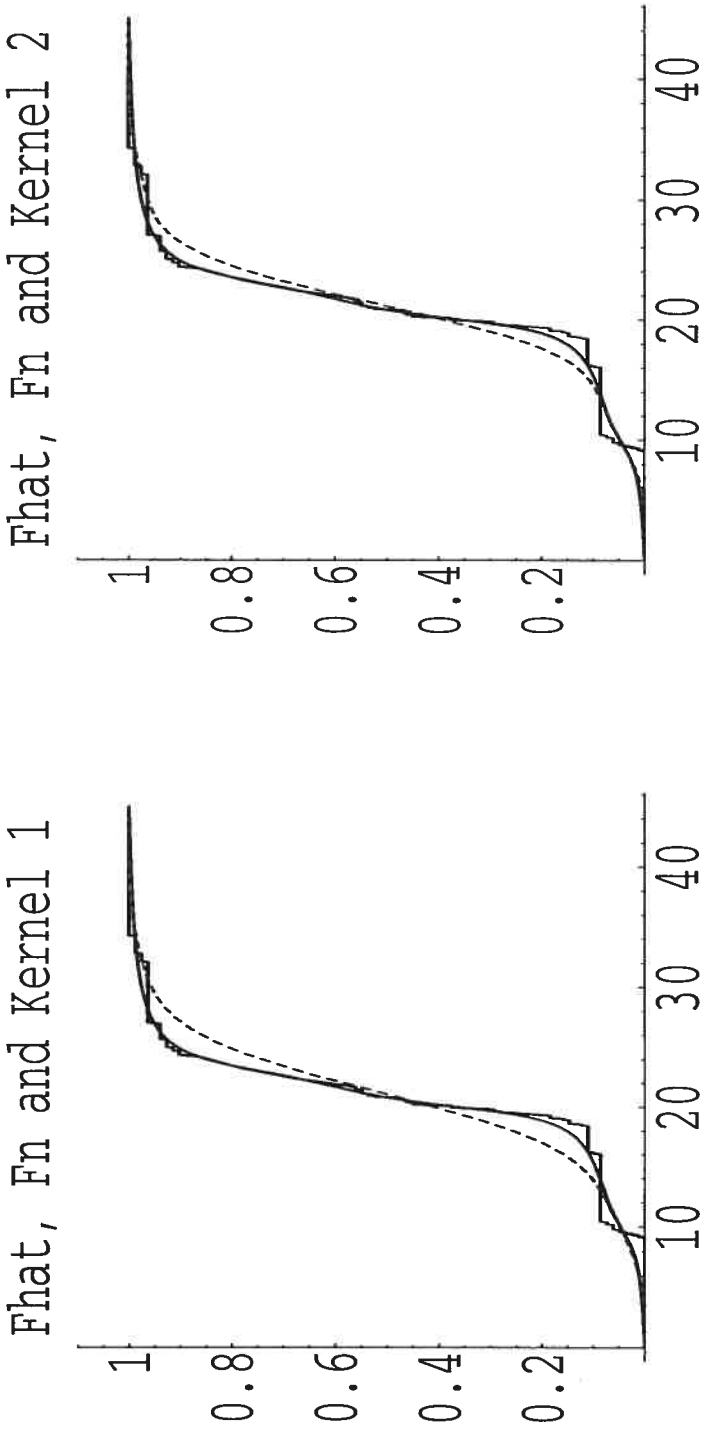


fhat:0,1,2 m:15,15,10.




f: 0.5*Beta(1,9)+0.5*Beta(6,1.5) (n=200, k=5, m=10.)
Density
fhat
Kernel
Density
0
1
2
3
4
0.0
0.2
0.4
0.6
0.8
1.0


Galaxies data (n=82, k=5, m=10).
fhat
Kernel1 (3.25)
Kernel2 (2.56)
density
0.0
0.05
0.10
0.15
0.20
10
20
30
40
velocity of galaxy (1000km/s)


Fhat, Fn and Kernel 2
1
0.8
0.6
0.4
0.2
10
20
30
40
Fhat, Fn and Kernel 1
1
0.8
0.6
0.4
0.2
10
20
30
40