# Estimation par noyaux associés mixtes d'un modèle de mélange

Francial G. Libengué, Sobom M. Somé & Célestin C. Kokonendji

*Université de Franche-Comté*
*Laboratoire de Mathématiques de Besançon - UMR 6623 CNRS-UFC*
*16 route de Gray, 25030 Besançon cedex, France.*
*E-mail : prenom.nom@univ-fcomte.fr*

**Résumé.** Dans cette communication, nous montrons une méthode non-paramétrique d'estimation d'une densité qui est mélange de fonctions discrètes et continues, par utilisation des outils unifiant les analyses discrètes et continues. Nous présentons d'abord le modèle étudié, ensuite nous définissons le noyau associé mixte correspondant et, enfin, nous étudions les principales propriétés des estimateurs qui en découlent, en particulier aux points frontières lors de passage du continu au discret et inversement.

**Mots-clés.** Densité mixte, échelle de temps, noyau associé.

**Abstract.** In this communication, we introduce a nonparametric method for estimating a density which is a mixture of discrete and continuous functions, from the tools unifying discrete and continuous analysis. We first present the studied model, then we define the mixed corresponding associated kernel and, finally, we study the main properties of the arising estimators, in particular at frontier points during the passage of continuous to the discrete and conversely.

**Keywords.** Associated kernel, mixed density, time-scales.

# 1 Introduction

Nous nous intéressons à l'estimation non-paramétrique de densités gouvernant les variables aléatoires réelles à support connu $\mathbb{T} \subseteq \mathbb{R}$, constitué à la fois des intervalles et des ensembles discrets deux à deux disjoints. Par exemple, on étudie un processus (fonction) de suivis d'un malade à $p$ phases et au cours desquelles le patient subit alternativement des soins intensifs avec des prélèvements quasi-instantanés (i.e. continus) correspondant à une période d'hospitalisation, et des soins externes où les prélèvements sont périodiques ou séquentiels dans le temps (i.e. discrets). Ce type de fonctions est généralement un modèle de mélange fini et peut s'écrire de la forme suivante :

$$f(x) = \sum_{j=1}^{p} \beta_j f_j(x) \mathbf{1}_{\mathbb{T}_j}(x), \tag{1}$$

où les $f_j$ sont des fonctions de densité ou masse de probabilité (f.d.m.p.) et les constantes $\beta_j$ sont les proportions du mélange (supposées connues dans ce travail pour simplifier) sur chaque composante $\mathbb{T}_j$ partition de $\mathbb{T}$. Sans perte de généralité, nous noterons $\mathbb{T}_I$ et $\mathbb{T}_N$ les parties constituées respectivement d'intervalles et d'ensembles discrets de $\mathbb{T} \subseteq \mathbb{R}$. On prendra alors

$$\mathbb{T} = T_I \cup \mathbb{T}_N = [t_0, t_1] \cup \{t_2, t_3, ...\}. \tag{2}$$

Cette communication propose une méthode d'estimation non-paramétrique de fonctions de type (1), avec $p = 2$ pour simplifier, sur le support $\mathbb{T}$ défini en (2) tout en respectant la structure topologique de ce dernier. La fonction définie en (1) est une f.d.m.p. dite *fonction mixte univariée*. La mixité est due ici au fait que la densité $f$ à estimer est partiellement continue et discrète. Une méthode appropriée pour ce type d'estimation est celle des noyaux associés puisque ces derniers sont construits dans l'esprit du strict respect de la nature topologique du support de $f$ ; cependant, nous devrions rester attentifs aux différents changements de structures de supports. Il faut noter que la force des noyaux associés pour ce type d'estimation réside dans leur capacité à dépendre intrinsèquement du point d'estimation $x$ et de la fenêtre de lissage $h$ interprétée comme paramètre de dispersion et qui joue le même rôle tant dans le cas discret que continu ; voir, p. ex., Jørgensen (1997) et Jørgensen et Kokonendji (2011, 2013). Enfin, une dernière des raisons est leur flexibilité dans l'utilisation de l'analyse unifiant le discret et le continu que nous détaillons plus tard. Rappelons que lorsque $\mathbb{T}$ est homogène (i.e. restreint à $T_I$ ou à $\mathbb{T}_N$), le noyau associé (continu ou discret) est une f.d.m.p. $K_{x,h}$ paramétrée par le point d'estimation $x$ et le paramètre de lissage $h$, sur le support $\mathbb{S}_{x,h}$ et qui vérifie :

$$x \in \mathbb{S}_{x,h}, \quad \mathbb{E}(\mathcal{Z}_{x,h}) = x + A(x,h), \text{ et } Var(\mathcal{Z}_{x,h}) = B(x,h), \tag{3}$$

où $A(x,h)$ et $B(x,h)$ tendent vers 0 lorsque $h$ tend 0 et $\mathcal{Z}_{x,h}$ est une variable aléatoire de loi $K_{x,h}$.

Pour une suite $X_1, X_2, \cdots, X_n$ de variables aléatoires indépendantes et identiquement distribuées (i.i.d.) de densité inconnue $g$ sur le support $\mathbb{T} = \mathbb{T}_I$ ou $\mathbb{T} = \mathbb{T}_N$, l'estimateur $\widehat{g_n}$ de $g$ à noyau associé $K_{x,h}$ est de la forme :

$$\widehat{g_n}(x) = \frac{1}{n} \sum_{i=1}^{n} K_{x,h}(X_i), \quad x \in \mathbb{T} \subseteq \mathbb{R}. \tag{4}$$

On peut se référer à Chen (1999, 2000), Kokonendji et Zocchi (2010), Kokonendji et Senga Kiessé (2011) puis Kokonendji et Libengué (2013) pour une présentation détaillée avec des multiples exemples.

Dans ce qui suit, nous présentons les estimateurs de f.d.m.p. par noyaux associés mixtes puis nous étudions leurs propriétés tout en mettant un accent particulier sur les points limites (points frontières situés au passage de $T_I$ à $\mathbb{T}_N$).

## 2 Estimateurs par noyaux associés mixtes

Considérons $X_1, X_2, \cdots, X_n$ une suite de variables aléatoires *i.i.d.* de densité mixte inconnue $f$ définie en (1) et de support $\mathbb{T} = \bigcup_{j=1}^{p} \mathbb{T}_j$ où les $\mathbb{T}_j$ sont les composantes homogènes continues ou discrètes deux à deux disjoints de $\mathbb{T}$. L'estimateur $\widehat{f}$ de $f$ est de la forme

$$\widehat{f(x)} = \sum_{j=1}^{p} \beta_j \widehat{f_j}(x) \mathbf{1}_{\mathbb{T}_j}(x), \tag{5}$$

où les $\widehat{f_j}$ sont les estimateurs à noyaux associés définis en (4) pondérés par les poids $\beta_j$ (connus) sur des composantes $\mathbb{T}_j$. En exprimant $\widehat{f_j}$ comme dans (4) puis en le remplaçant dans (5) par son expression obtenue, on déduit **l'estimateur à noyau associé mixte** comme suit :

$$\widehat{f(x)} = \sum_{j=1}^{p} \frac{\beta_j}{n_j} \sum_{i=1}^{n_j} K_{x,h}^{[j]}(X_i) \mathbf{1}_{\mathbb{T}_j}(x) \mathbf{1}_{\mathbb{T}_j}(X_i), \tag{6}$$

avec $n_j$ le nombre d'obeservations tombant dans $\mathbb{T}_j$ et $K_{x,h}^{[j]}(\cdot)\mathbf{1}_{\mathbb{T}_j}(x)\mathbf{1}_{\mathbb{T}_j}(\cdot)$ le noyau associé sur $\mathbb{T}_j$. De l'expression (6), on déduit le **noyau associé mixte** par

$$K_{\theta(x,h)}(\cdot) = \sum_{j=1}^{p} K_{x,h}^{[j]}(\cdot) \mathbf{1}_{\mathbb{T}_j}(x) \mathbf{1}_{\mathbb{T}_j}(\cdot). \tag{7}$$

Il est de support $\mathbb{S}_{\theta(x,h)} = \bigcup_{j=1}^{p} \mathbb{S}_{x,h}^{[j]}$, où les $\mathbb{S}_{x,h}^{[j]}$ sont les supports de $K_{x,h}^{[j]}(\cdot)\mathbf{1}_{\mathbb{T}_j}(x)\mathbf{1}_{\mathbb{T}_j}(\cdot)$.

On vérifie facilement que $K_{\theta(x,h)}(\cdot)$ en (7) satisfait les conditions des noyaux associés données en (3) que l'on peut récrire sous la forme

$$x \in \mathbb{S}_{\theta(x,h)}, \quad \mathbb{E}(\mathcal{Z}_{\theta(x,h)}) = x + A_\theta(x,h), \text{ et } Var(\mathcal{Z}_{\theta(x,h)}) = B_\theta(x,h), \tag{8}$$

où $A_\theta(x,h) = \sum_{j=1}^{p} A_j(x,h)\mathbf{1}_{\mathbb{T}_j}(x)$ et $B_\theta(x,h) = \sum_{j=1}^{p} B_j(x,h)\mathbf{1}_{\mathbb{T}_j}(x)$ tendent vers 0 lorsque $h$ tend 0 et $\mathcal{Z}_{\theta(x,h)}$ est une variable aléatoire de loi $K_{\theta(x,h)}$.

**Proposition 2.1** *L'estimateur $\widehat{f}$ donné en (6), vérifie pour tout x dans $\mathbb{T}$ :*

$$\mathbb{E}\left\{\widehat{f(x)}\right\} = \mathbb{E}\left\{f(\mathcal{Z}_{\theta(x,h)})\right\}.$$

La démonstration se déduit de celle des cas discret et continu.

**Proposition 2.2** *Soit $f \in \mathscr{C}^2(\mathbb{T})$, la densité mixte à estimer et $\widehat{f}$ son estimateur à noyau associé mixte en (6). Pour tout x dans $\mathbb{T}$ et $h = h_n > 0$, alors*

$$Bias\left\{\widehat{f(x)}\right\} = A_\theta(x,h) f^{(1)}(x) + \frac{1}{2}\left\{A_\theta^2(x,h) + B_\theta(x,h)\right\} f^{(2)}(x) + o(h^2). \tag{9}$$

*Si de plus, $f$ est bornée sur $\mathbb{T}$ alors*

$$Var\left\{\widehat{f(x)}\right\} = \frac{1}{n} f(x) \left\|K_{\theta(x,h)}\right\|_2^2 + o\left(\frac{1}{nh^{r_2}}\right), \tag{10}$$

*où* $r_2 = r_2\left(K_{\theta(x,h)}\right) > 0$ *est le plus grand réel tel que* $\left\|K_{\theta(x,h)}\right\|_2^2 = \int_{\mathbb{S}_{\theta(x,h)} \cap \mathbb{T}} K^2_{\theta(x,h)}(u)du \leq c_2(x) h_n^{-r_2}$ *et* $0 < c_2(x) < +\infty$.

Les quantités $f^{(1)}$ et $f^{(2)}$ dans (9) désignent respectivement les dérivées ou différences finies de premier et second ordre de $f$ sur $\mathbb{T}$ et la quantité $\left\|K_{\theta(x,h)}\right\|_2^2$ dans (10) est l'intégrale de carré du noyau associé mixte $K_{\theta(x,h)}(\cdot)$. Toutefois, il est fondamental de connaître les valeurs explicites de $f^{(1)}$, $f^{(2)}$ et $\left\|K_{\theta(x,h)}\right\|_2^2$ plus précisément aux frontières lors de passage de $\mathbb{T}_I$ à $\mathbb{T}_N$, surtout si l'on veut avoir des informations aussi loin que possible dans un voisinage fixé de ces types de points. Ainsi, nous proposons l'utilisation des outils efficaces pour unifier les analyses discrète et continue, et dévéloppés dans Hilger (1990), Agarwal et Bohner (1999), Bohner et Peterson (2001), et Sanyal (2008). Ces auteurs ont considéré $\mathbb{T}$ comme une *échelle de temps* ("Time-scale" en anglais). C'est un sous ensemble fermé de $\mathbb{R}$ (muni de sa topologie naturelle) qui est une réunion des intervalles avec des ensembles discrets comme défini en (2). Deux opérateurs y sont créés pour unifier l'analyse discrète et continue. On a :

(i) l'**opérateur de saut en avant** (forward jump operator)

$$\sigma : \mathbb{T} \to \mathbb{T}, \quad x \mapsto \sigma(x) := \inf\{s \in \mathbb{T}\,;\, s > x\}, \quad \forall x \in \mathbb{T}\,;$$

(ii) l'**opérateur de saut en arrière** (backward jump operator)

$$\rho : \mathbb{T} \to \mathbb{T}, \quad x \mapsto \rho(x) := \sup\{s \in \mathbb{T}\,;\, s < x\}, \quad \forall x \in \mathbb{T}.$$

Pour mieux distinguer les points de bords, nous désignons par **point discret d'ordre** $k$, tout point situé à $k$-pas symétriques des bords gauche et droit de $\mathbb{T}_N$ et **point discret d'ordres** $k_1$ **à droite et** $k_2$ **à gauche** tout point situé à $k_1$-pas et $k_2$-pas respectivement de bords droit et gauche de $\mathbb{T}_N$. La continuité sur $\mathbb{T}_j$ (i.e. $\mathbb{T}_I$ ou $\mathbb{T}_N$) se traduit par :

$$\forall \varepsilon > 0,\ \exists \mathcal{V}_{x_j}\ :\ s \in \mathcal{V}_{x_j} \cap \mathbb{T}_j \Rightarrow |f(s) - f(x_j)| < \varepsilon,$$

où $\mathcal{V}_{x_j}$ est un voisinage de $x_j$ selon la topologie induite sur $\mathbb{T}_j$. Les notions usuelles de dérivées à droite et à gauche pour $x \in \mathbb{T} = \mathbb{T}_I$ et celles des différences finies décentrées à droite et à gauche lorsque $x \in \mathbb{T} = \mathbb{T}_N$ sont unifiées dans celles de **Δ-dérivée** et **∇-dérivée** de $f$, notées $f^{\Delta}$ et $f^{\nabla}$, puis définies respectivement par

$$f^{\Delta}(x) = \frac{f\{\sigma(x)\} - f(s)}{\sigma(x) - s} = \lim_{\substack{x \to s \\ x > s}} \frac{f(x) - f(s)}{x - s} \quad \text{et} \quad f^{\nabla}(x) = \frac{f\{\rho(x)\} - f(s)}{\rho(x) - s} = \lim_{\substack{x \to s \\ x < s}} \frac{f(x) - f(s)}{x - s}.$$

Par conséquent, le nombre dérivé en $x \in \mathbb{T} = \mathbb{T}_I$ et les différences finies du premier odre en $x \in \mathbb{T} = \mathbb{T}_N$ sont unifiées et définies à travers :

$$f^{(1)}(x) = \frac{1}{\sigma(x) - \rho(x)} \left[ \{\sigma(x) - x\} f^{\Delta}(x) + \{x - \rho(x)\} f^{\nabla}(x) \right], \quad \forall x \in \mathbb{T}.$$

Ceci conduit à la formule de Taylor donnée par :

$$f(x) = \sum_{j}^{k} \frac{(x-s)^j}{\{\sigma(x) - \rho(x)\} j!} \left[ \{\sigma(x) - x\} f^{\Delta^j}(x) + \{x - \rho(x)\} f^{\nabla^j}(x) \right] + o\{(x-s)^k\}. \tag{11}$$

Pour les points discrets d'ordres $k_1$ à droite et $k_2$ à gauche, l'expression (11) s'écrit :

$$f^{(k_1,k_2)}(x) = \frac{1}{\sigma(x) - \rho(x)} \left[ \{\sigma(x) - x\} f^{\Delta^{k_1}}(x) + \{x - \rho(x)\} f^{\nabla^{k_2}}(x) \right], \quad \forall (k_1, k_2) \in \mathbb{N}^* \times \mathbb{N}^*. \tag{12}$$

Notons qu'en général $f^{(j)}(\cdot) = f^{(j,j)}(\cdot)$ et en particulier pour $x \in \mathbb{T}_N = \mathbb{N}$, on a :

$$f^{(2)}(x) = \begin{cases} \{f(x+2) - 2f(x+1) + 2f(x) - 2f(x-1) + f(x-2)\}/2 & \text{si } x \in \mathbb{N} \setminus \{0,1\} \\ \{f(3) - 2f(2) + 2f(1) - f(0)\}/2 & \text{si } x = 1 \\ \{f(2) - 2f(1) + f(0)\}/2 & \text{si } x = 0. \end{cases} \tag{13}$$

Il est important de remarquer que la différence finie d'ordre $k$ d'une fonction $f$ en $x \in \mathbb{T}_N$ existe si et seulement si $x$ est au minimum discret d'ordre $k$. De même, une fonction $f$ est dite de classe $\mathscr{C}^k$ si et seulement $f^{(k)}$ existe et est continue. Enfin, en utilisant les notions de **$\Delta$-intégrabilité** et **$\nabla$-intégrabilité**, induites respectivement par **$\Delta$-mesure** et **$\nabla$-mesure** (lesquelles sont similaires à la mesure de Lebesgue sur $\mathbb{T}_I$ et la mesure de dénombrement sur $\mathbb{T}_N$), on a la **$\Delta$-intégrale** de $f$ comme suit

$$\int_a^b f^{\Delta}(t)\Delta t = f(b) - f(a), \text{ si } \mathbb{T} = \mathbb{T}_I \quad \text{et} \quad \int_a^b f(t)\Delta t = \sum_{t=a}^{b} f(t), \text{ si } \mathbb{T} = \mathbb{T}_N,$$

avec $\Delta t$ la **$\Delta$-mesure** sur $\mathbb{T}$. On obtient la **$\nabla$-intégrale** de $f$ gâce aux relations

$$f^{\Delta}(x) = f^{\nabla}\{\sigma(x)\} \quad \text{et} \quad f^{\nabla}(x) = f^{\Delta}\{\rho(x)\}. \tag{14}$$

Ainsi, en réunissant ces ingrédients d'unification des analyses discrète et continue de $\mathbb{T}$, on peut explicitement exprimer les quantités $f^{(1)}$ et $f^{(2)}$ dans (9) ainsi que $\left\| K_{\theta(x,h)} \right\|_2^2$ dans (10). Les preuves des propositions précédentes en découlent.

On obtient à partir de (9) et (10), le risque quadratique moyen intégré et asymptotique ("Asymptotic Mean Integrated Squared Error" (AMISE) en anglais) de l'estimateur (5) de la forme :

$$AMISE(\widehat{f_{n,h,K,f}}) = \int_{\mathbb{T}} \left( \left[ A_\theta(x,h) f^{(1)}(x) + \frac{1}{2}\{A_\theta^2(x,h) + B_\theta(x,h)\} f^{(2)}(x) \right]^2 \right) dx + \frac{1}{n} \left\| K_{\theta(x,h)} \right\|_2^2 f(x) dx. \tag{15}$$

Les études numériques (par simulation et sur des données réelles) démontrent une efficacité de la procédure de (5) et seront présentées lors de la communication orale. Pour des supports $\mathbb{T}$ du genre (2), nous illustrons des noyaux associés du type bêta étendu (Chen, 1999 ; Kokonendji et Libengué, 2013) et du type triangulaire discret (Kokonendji et Zocchi, 2010). Par la suite, nous discutons du cas général où les $\beta_j$ sont des paramètres inconnus dans (1). Le paramètre de dispersion $h$ est global et commun à toutes les composantes $K_{x,h}^{[j]}(\cdot)\mathbf{1}_{\mathbb{T}_j}(x)\mathbf{1}_{\mathbb{T}_j}(\cdot)$ de $K_{\theta(x,h)}(\cdot)$. Son choix se fait directement à partir de $K_{\theta(x,h)}(\cdot)$ et indépendament de ceux des fenêtres locales de lissage $h_j$ par validation croisé ou par l'approche bayésienne.

# Bibliographie


[1] Agarwal, R.P. et Bohner. M. (1999), Basic calculus on time scales and some of its applications, *Results Mathematics* **35**, 3-22.
[2] Bohner, M. et Peterson, A. (2001), *Dynamic Equations on Time Scales*, Birkhäusera Boston Inc., Boston.
[3] Chen, S.X. (1999), Beta kernel estimators for density functions, *Computational Statistics and Data Analysis* **31**, 131-145.
[4] Chen, S.X. (2000), Gamma kernel estimators for density functions, *Annals of Institute for Statistics and Mathematics* **52**, 471-480.
[5] Hilger, S. (1990), Analysis on measure chains – a unified approach to continuous and discrete calculus, *Results Mathematics* **18**, 18-56.
[6] Jørgensen, B. (1997), *The Theory of Dispersion Models*, Chapman & Hall, London.
[7] Jørgensen, B. et Kokonendji, C.C. (2011), Dispersion models for geometric sums, *Brazilian Journal of Probability and Statistics* **25**, 263-293.
[8] Jørgensen, B. et Kokonendji, C.C. (2013), Discrete dispersion models and their Tweedie asymptotics, *In Preparation*.
[9] Kokonendji, C.C. et Libengué, F.G. (2013), Non-classical associated kernels for non-standard density estimators, *Soumis pour publication*.
[10] Kokonendji, C.C. et Senga Kiessé, T. (2011), Discrete associated kernels method and extensions, *Statistical Methodology* **8**, 497-516.
[11] Kokonendji, C.C. et Zocchi, S.S. (2010), Extensions of discrete triangular distribution and boundary bias in kernel estimation for discrete functions, *Statistics and Probability Letters* **80**, 1655-1662.
[12] Sanyal, S. (2008), Stochastic dynamic equations, PhD thesis, Missouri University of Sciences and Technology.