Estimateurs polynomiaux locaux généralisés des fonctionnelles d'une fonction de répartition à support positif.

LAÏB Naâmane [1]

Laboratoire AGM, CY Cergy Paris Université, Cergy

JS2O,Perpignon, April 2-4, 2025

[1] Joint work with Y. Chaubey (Concordia Univ.) and K. Ghoudi (United Arab Emirates University).
published in the Journal of Nonparametric Statistics (2024)

Table of contents

Introduction: Biais aux bords

▷ Biais aux bords dans l'estimation de densité à noyau symétrique

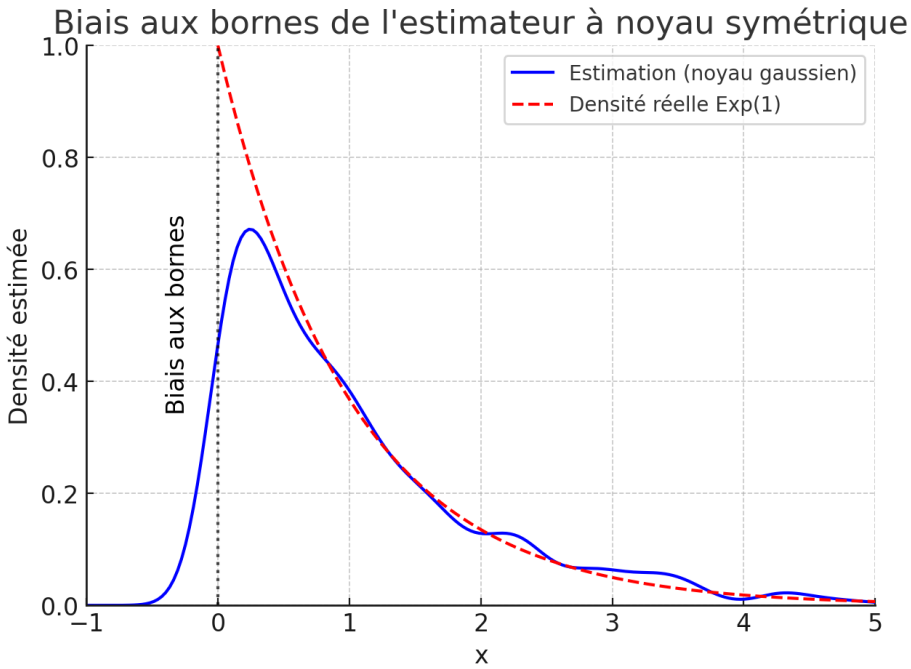
- Les estimateurs à **noyau symétrique** souffrent d'un biais aux bords pour des densités sur $[0,1]$ ou $[0,\infty)$

Par example:

- Le noyau gaussien (symétrique) accorde une probabilité non nulle à des valeurs négatives, bien qu'elles ne soient pas présentes dans les données.
- Il en résulte un biais asymptotique aux bords, donnée par:

$$\widehat{f}_n(0) \to \frac{f(0)}{2}, \quad n \to \infty.$$

Introduction: Biais aux bords (Illustration)



- La densité estimée (bleue) s'écarte de la densité vraie (rouge pointillée) en $x = 0$.
- Une **méthode de correction**, comme l'ajustement des noyaux aux bords, est nécessaire.

Introduction : Méthodes pour corriger le biais aux frontières

- Silverman (1986) a suggéré une transformation logarithmique de la variable aléatoire.
- Wand et al. (1991) ont proposé une méthode de transformation générale, mais coûteuse en calcul.
- Une autre approche repose sur les **noyaux asymétriques** (qui respecte le support des données):
  - Noyau exponentiel : Bagai and PrakasaRao (1995)
  - Noyau Gamma : Chen (2000); Chaubey et al. (2012)
  - Noyau Gamma inverse : Balakrishna and Koul (2017); Kakizawa and Igarashi (2017)
  - Noyaux Gaussien inverse et Gaussien inverse réciproque : Scaillet (2004)

Introduction (suite)

- Chaubey et al. (2012) ont proposé un estimateur lisse de la densité basé sur la fonction de répartition empirique.

Notre objectif est de construire

- des estimateurs polynomiaux locaux (EPL) pour des fonctionnelles lisses d'une fonction de répartition à support positif.
- Nous nous intéressons également à estimer les dérivées de ces fonctionnelles.

Contexte et Méthodologie

- Nous considérons un fonctionnel $\Phi(x, F)$ de la fonction de répartition F, avec un **support** $0 \leq x < \infty$.
- **Exemples** de fonctionnels :

$$\Phi(x, F) = F(x) \quad \text{ou} \quad \Phi(x, F) = \int \phi(x, s)\, dF(s).$$

- L'**objectif** est d'**estimer** $\Phi(x, F)$ et ses **dérivées** par rapport à x,
  - en utilisant des noyaux asymétriques pour réduire le biais aux bords.
  - Cette approche fournit un **cadre général** incluant des estimateurs, de la fonction de risque (hazard function) et du rapport de densité.

Contexte et Méthodologie (suite)

- Les noyaux **asymétriques** utilisés sont des densités, notées $\mathbf{q}_h$, à support positif. Le paramètre de lissage h contrôle leur variance.
  - Un **exemple** d'un tel noyau est le noyau gamma(α, β) :

$$\mathbf{q}_h(z) = \frac{1}{\beta^\alpha \Gamma(\alpha)} e^{-z/\beta} z^{\alpha-1},$$

où $\alpha = 1/h^2$, $\beta = h^2$, et $h = h_n$ est un paramètre de lissage qui tend vers zéro lorsque n tend vers l'infini.

Définition des Estimateurs

Idée principale :

- On approxime localement $\Phi(y, F)$, si elle est r-fois dérivable, par un polynôme autour de x :

$$\Phi(y, F) \approx \sum_{k=0}^{r} \frac{a_k(x)}{k!}(y-x)^k.$$

- L'estimation consiste à trouver les coefficients $(\hat{a}_0(x), \hat{a}_1(x), \ldots, \hat{a}_r(x))$ en minimisant l'erreur quadratique pondérée :

$$J(a_0, \ldots, a_r, x) = \int_0^{\infty} \frac{1}{x}\mathbf{q}_h\left(\frac{z}{x}\right)\left\{\Phi(z, F_n) - \sum_{k=0}^{r} \frac{a_k(x)}{k!}(z-x)^k\right\}^2 dz.$$

Définition des Estimateurs (suite)

- Ce problème se résout dans l'espace des polynômes $L^2(\mathbf{q}_h)$, où la solution optimale est la projection de $\Phi(z, F_n)$ **observée** sur $L^2(\mathbf{q}_h)$.
- Le vecteur $(\hat{a}_0(x), \ldots, \hat{a}_r(x))$ donne le meilleur ajustement polynomial local.
- Il permet d'estimer de manière consistante :

$$(\Phi(x, F), \Phi^{(1)}(x, F), \ldots, \Phi^{(r)}(x, F)).$$

Produit scalaire fonctionnel et norme

Pour préciser notre cadre de travail, nous introduisons un produit scalaire et une norme induite dans un espace fonctionnel.

- **Contexte :** Nous considérons une famille de fonctionnels $\Phi(x, F)$, avec $x \in [0, \infty)$ fixé, où F est une fonction de distribution continue à support positif.
- **Produit scalaire :** entre deux fonctions f et g sur $[0, \infty)$ est défini par l'intégrale suivante :

$$\langle f, g \rangle := \int_0^\infty f(u) g(u) \mathbf{q}_h(u)\, du,$$

où $\mathbf{q}_h(u)$ est une fonction de pondération.

- **Norme induite :** associée à ce produit scalaire est donnée par :

$$\|f\| := \langle f, f \rangle^{1/2}.$$

Espace Polynomial et Noyau de Reproduction

Pour caractériser ces estimateurs, nous introduisons un nouvel espace polynomial ainsi qu'un noyau de reproduction

- **Espace polynomial :** $\mathcal{P}_r(\mathbf{q}_h)$.
  - l'espace vectoriel des polynômes de degré au plus r,
    - équipé du produit scalaire précédent.
    - D'une base orthonormée : $P_0, P_1, \ldots, P_r$
- **Noyau de reproduction :** associé est donné par la somme des produits des polynômes $P_k(u)$ et $P_k(v)$, soit :

$$K_r(u,v) = \sum_{k=0}^{r} P_k(u)P_k(v).$$

[2] [2]Ce noyau est utilisé dans l'estimation locale pour projeter les données sur l'espace des polynômes

Hypothèses

A.1 Pour tout $x \in [0, \infty)$, $\Phi(x, F)$ est **bien définie** pour toute fonction de répartition F.

A.2 La fonction Φ est r fois différentiable par rapport à x.

A.3 Le noyau $\mathbf{q}_h$ est une fonction de densité bornée de moyenne 1 et de variance h^2 :

$$\int_0^\infty u\mathbf{q}_h(u)du = 1, \quad \int_0^\infty (u-1)^2\mathbf{q}_h(u)du = h^2.$$

A.4 Les conditions sur les moments:

$$\int_0^\infty u^{2r+1}\mathbf{q}_h(u)du < \infty, \text{ et } \frac{1}{h^k}\int_0^\infty |u-1|^k\mathbf{q}_h(u)du \leq C_k < B < \infty,$$

$k = 1, \ldots, 2r+1$, C_k est une constante

Exemples de noyaux $\mathbf{q}_h$ satisfaisant A.3 et A.4

1. **Noyau de densité triangulaire** symétrique autour de 1

$$\mathbf{q}_h(u) = \frac{h\sqrt{6} - |u-1|}{6h^2}, \quad \text{pour } 1 - h\sqrt{6} < u < 1 + h\sqrt{6}$$

2. Le **noyau Beta décalé**, donné par

$$\mathbf{q}_h(u) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} \left(t - \frac{1}{2}\right)^{\alpha-1} \left(\frac{3}{2} - t\right)^{\beta-1}, \quad \text{pour } \frac{1}{2} < t < \frac{3}{2}.$$

Choisissant $\alpha = \beta = 8h^2 - \frac{1}{2}$.

3. Le **noyau Gamma asymétrique**, défini par :

$$\mathbf{q}_h(u) = \frac{t^{\alpha-1} e^{-t/\beta}}{\beta^\alpha \Gamma(\alpha)}, \quad \text{pour } t > 0, \quad \alpha = \frac{1}{h^2} \text{ et } \beta = h^2.$$

Caractéristiques des Estimateurs Polynômiaux Locaux

Théorème. Supposons que **A.1–A.4** soient vérifiées et que $P_0(z), \ldots, P_r(z)$ forment une base orthonormée de l'espace $\mathcal{P}_r(\mathbf{q}_h)$. Définissons le noyau

$$K^{[m,r]}(v) = \sum_{k=0}^{r} P_k(v) \left. \frac{d^m}{du^m} P_k(u) \right|_{u=1} \mathbf{q}_h(v).$$

3

[3] $K^{[m,r]}(v)$: est obtenu en multipliant le noyau $\mathbf{q}_h(v)$ par la dérivée d'ordre m du noyau de reproduction par rapport à u, évaluée en $u = 1$.

Caractéristiques des EPL (suite)

▷ Les **estimateurs polynomiaux locaux**(EPL) de

$$\Phi(x,F),\Phi^{(1)}(x,F),\ldots,\Phi^{(r)}(x,F)$$

sont données (pour $x>0$) par :

$$\hat{\Phi}_n^{[m,r]}(x):=\hat{a}_m(x)=\frac{1}{x^{m+1}}\int_0^{\infty}K^{[m,r]}\left(\frac{z}{x}\right)\Phi(z,F_n)dz,\ m=0,1,\ldots,r.$$

- $\hat{\Phi}_n^{[m,r]}(x)$ est un (**EPL**) de la m-ème dérivée de Φ
- r : l'ordre du polynôme de lissage local

Propriétés de $K^{[m,r]}$

Définition. Un noyau K est dit **d'ordre** (m, p) avec $(m, p) \in \mathbb{N}^2$ et $m \leq p-1$, s'il satisfait la propriété suivante :

$$\int_0^\infty (u-1)^k K(u)\, du = \begin{cases} 0 & \text{pour } k = 0, \ldots, p-1 \text{ et } k \neq m \\ m! & \text{pour } k = m \\ C_p \neq 0 & \text{pour } k = p \end{cases}$$

Pour tout noyau $\mathbf{q}_h$ vérifiant **A.3 et A.4** :

- $K^{[m,r]}$ est un noyau reproduisant d'ordre $(m, r+1)$.
- De plus, il vérifie une condition de décroissance en h

$$\int_0^\infty |(u-1)|^\ell K^{[m,r]}|\, du = O(h^{\ell-m}) \quad \text{pour tout } 0 \leq \ell \leq r+1.$$

Biais et Consistance de l'Estimateur pour un Fonctionnel Général Φ

Propriétés : Biais

Proposition

Supposons que :

- $\mathbb{E}\{\Phi(x, F_n)\} = \Phi(x, F)$ *for all* $x \geq 0$.
- $\Phi(x, F)$ *admet*$(r+2)$ *dérivées bornées..*
- *Le noyau* $\mathbf{q}_h$ *vérifie les conditions (***A3-A4)**

Alors, le biais de l'estimateur (PL) est donné par :

$$\mathbb{E}\{\hat{\Phi}_n^{[m,r]}(x)\} - \Phi^{(m)}(x, F) = \frac{x^{r+1-m}\,\Phi^{(r+1)}(x, F)}{(r+1)!} \times$$

$$\int_0^{\infty} (u-1)^{r+1} K^{[m,r]}(u) du + x^{r+2-m} O(h^{r+2-m}).$$

$\triangleright$ *Si* $h = h_n \to 0$ *lorsque* $n \to \infty$, *alors*

- $\hat{\Phi}_n^{[m,r]}(x)$ *est un estimateur asymptotiquement sans biais de* $\Phi^{(m)}(x, F)$ *pour* $0 \leq m \leq r$.

Propriétés : Convergence

Theorem

Supposons que les conditions suivantes sont remplies :

- $\mathbb{E}[\Phi(t, F_n)] = \Phi(t, F)$ *pour tout* $t \geq 0$.
- $\Phi(t, F)$ *admet* $(r+1)$ *dérivées bornées par rapport à* t.
- *Le noyau* $\mathbf{q}_h$ *vérifie* **(A3-A4)**.
- $h = h_n \to 0$ *lorsque* $n \to \infty$.

De plus, supposons que $\forall\ x > 0$, *il existe un voisinage ouvert* $\mathcal{N}_x$ *de* x *tel que :*

$$\lim_{n\to\infty} \sup_{y\in\mathcal{N}_x} h^{-r} \left|\Phi(y, F_n) - \Phi(y, F)\right| = 0 \quad \textit{(p.s.)}.$$

Alors, l'estimateur $\hat{\Phi}_n^{[m,r]}(x)$ *converge p.s. vers* $\Phi^{(m)}(x, F)$.

Fonctionnels Linéaires

- Dans de nombreuses applications pratiques,
- le fonctionnel Φ est linéaire

Fonctionnels Linéaires

- On dit que Φ est un **fonctionnel linéaire (FL)** s'il peut être exprimé sous la forme :

$$\Phi(x, F) = \int_0^\infty \phi(x, s)\, dF(s).$$

- Si Φ est un (FL), alors l'estimateur $\hat{\Phi}_n^{[m,r]}(x)$ est donné par :

$$\hat{\Phi}_n^{[m,r]}(x) = \frac{1}{x^{m+1}} \frac{1}{n} \sum_{i=1}^{n} \int_0^\infty \phi(z, X_i) K^{[m,r]}\left(\frac{z}{x}\right) dz.$$

Notation et hypothèses supplémentaires

- Soit $G(x,y) = \mathbb{E}[\phi(x,X_1)\phi(y,X_1)]$ et supposons qu'elle possède des dérivées partielles continues et bornées d'ordre 1 et 2 pour $0 < y \leq x < \infty$.
- Soit $G_{(1,0)}$ et $G_{(0,1)}$ ses dérivées partielles par rapport à x et y, respectivement.

$$b_h^{[m,r]} = h^{2m-1} \int_0^\infty \bar{K}^{[m,r]}(u)^2 \, du, \text{ où } \bar{K}^{[m,r]}(v) = \int_v^\infty K^{[m,r]}(u) \, du,$$

et définissons

$$C_{h,r+1} = h^{m-r-1} \int_0^\infty (v-1)^{r+1} K^{[m,r]}(v) \, dv.$$

Remarque. Sous **A3-A4**: $b_h^{[m,r]} = O(1)$ et $C_{h,r+1} = O(1)$.

Propriétés : Fonctionnels linéaires

Proposition

Soit $x > 0$ un nombre réel fixe et m un entier tel que $0 \leq m \leq r$. Sous des hypothèses de régularité appropriées, nous avons :

i) *pour $0 \leq m \leq r$,*

$$\mathbb{E}\left(\hat{\Phi}_n^{[m,r]}(x)\right) - \Phi^{(m)}(x, F)$$
$$= \frac{h^{r+1-m} x^{r+1-m} \Phi^{(r+1)}(x, F)}{(r+1)!} \mathcal{C}_{h,r+1} + o(h^{r+1-m}),$$

ii) *pour $m = 0$,*

$$Var\left(\hat{\Phi}_n^{[0,r]}(x)\right) = \frac{Var\{\phi(x, X_1)\}}{n} + O\left(\frac{h}{n}\right),$$

Propriétés : Fonctionnels linéaires

Proposition (Proposition Continue)

iii) *pour* $1 \leq m \leq r$,

$$Var\left(\hat{\Phi}_n^{[m,r]}(x)\right) = \frac{G^{(0,1)}(x,x) - G^{(1,0)}(x,x)}{nh^{2m-1}x^{2m-1}} b_h^{[m,r]} + o\left(\frac{1}{nh^{2m-1}}\right).$$

iv) *Si, de plus,* $h_n \longrightarrow 0$ *et* $nh_n^{2m-1} \longrightarrow \infty$ *lorsque* $n \longrightarrow +\infty$, *alors*

$$\mathbb{E}\left(\hat{\Phi}_n^{[m,r]}(x) - \Phi^{(m)}(x,F)\right)^2 \longrightarrow 0.$$

Normalité asymptotique $m > 0$

1. Si $1 \leq m \leq r$,

   $h_n \to 0$, $nh_n^{2m-1} \to \infty$ lorsque $n \to \infty$, et $\lim_{h_n \to 0} b_{h_n}^{[m,r]} = b^{[m,r]}$, alors

$$\sqrt{nh_n^{2m-1}}\Big(\hat{\Phi}_n^{[m,r]}(x) - \mathbb{E}\{\hat{\Phi}_n^{[m,r]}(x)\}\Big) \xrightarrow{\mathcal{L}} \mathcal{N}\Big(0, \frac{b^{[m,r]}\{G^{(0,1)}(x,x) - G^{(1,0)}(x,x)\}}{x^{2m-1}}\Big),$$

2. Si, en plus des conditions 1), $nh_n^{2r+1} \longrightarrow 0$, $n \to \infty$, alors

$$\sqrt{nh_n^{2m-1}}\Big(\hat{\Phi}_n^{[m,r]}(x) - \Phi^{(m)}(x,F)\Big) \xrightarrow{\mathcal{L}} \mathcal{N}\Big(0, \frac{b^{[m,r]}\{G^{(0,1)}(x,x) - G^{(1,0)}(x,x)\}}{x^{2m-1}}\Big),$$

Normalité asymptotique $m = 0$

1. Si $h_n \longrightarrow 0$ lorsque $n \longrightarrow +\infty$, alors

$$\sqrt{n}\left(\hat{\Phi}_n^{[0,r]}(x) - \mathbb{E}\{\hat{\Phi}_n^{[0,r]}(x)\}\right) \underrightarrow{\mathcal{L}} \mathcal{N}\left(0,\ \mathrm{Var}\{\phi(x,X)\}\right),$$

2. Si $h_n \longrightarrow 0$ et $nh_n^{2r+2} \longrightarrow 0$ lorsque $n \longrightarrow +\infty$, alors

$$\sqrt{n}\left(\hat{\Phi}_n^{[0,r]}(x) - \Phi(x,F)\right) \underrightarrow{\mathcal{L}} \mathcal{N}\left(0,\ \mathrm{Var}\{\phi(x,X)\}\right).$$

Paramètre de lissage optimale

Soit

$$c_1^\star(m) = b^{[m,r]} \int_0^\infty \frac{G^{(0,1)}(x,x) - G^{(1,0)}(x,x)}{x^{2m-1}} dx$$

et

$$c_2^\star(m,r) = \left(\frac{\bar{C}_{r+1}}{(r+1)!}\right)^2 \int_0^\infty (\Phi^{(r+1)}(x,F))^2 x^{2(r+1-m)} dx,$$

alors la **paramètre de lissage optimale**

$$h_n^{o\star} = \left(\frac{(2m-1)c_1^\star(m)}{2(r+1-m)c_2^\star(m,r)}\right)^{\frac{1}{2r+1}} n^{-\frac{1}{2r+1}}$$

et le **AMISE optimale**

$$\mathrm{AMISE}^{o\star}(\hat{\Phi}_n^{[m,r]}(x)) = C^\star(m,r) \frac{1}{n^{1-\frac{2m-1}{2r+1}}}.$$

Applications

Applications

Nous explorons diverses applications de la procédure d'estimation proposée :

- Estimation de la fonction de répartition et de ses dérivées.
- Estimation de densité avec échantillonnage biaisé [4]
- Estimation non paramétrique de rapport des densités.
- Estimation du taux de risque (statistiques de survie).

[4] Par exemple, si une enquête sur les revenus ne concerne que les personnes employées, l'estimation de la distribution des revenus sera biaisée car elle exclut les chômeurs

Estimation de la fonction de distribution et de ses dérivées

- Nous considérons l'estimation de la **fonction de distribution cumulative** (CDF) F et de **ses dérivées** en $x \in \mathbb{R}^+$ en utilisant une approche fonctionnelle linéaire.
- Dans ce cadre :

$$\phi(x,s) = \mathbf{1}_{(0,x]}(s) \quad \text{et} \quad \Phi(x,F) = F(x).$$

- Notre objectif est d'étudier les propriétés des estimateurs $\Phi_n^{[m,r]}(x)$ pour $0 \leq m \leq r$.

Remarque

- Pour $m = 0$, $\hat{\Phi}_n^{[0,r]}(x) := \hat{F}_n^{[0,r]}(x)$ est un estimateur de la fonction de répartition F.
- Pour $m = 1$, $\hat{\Phi}_n^{[1,r]}(x) := \hat{f}_n^{[0,r]}(x)$ est un estimateur de la fonction de densité f.
- Pour $m > 1$, nous avons :

$$\hat{\Phi}_n^{[m,r]}(x) = \hat{F}_n^{[m,r]}(x) = \hat{f}_n^{[m-1,r]}(x) = \frac{1}{nx^m} \sum_{i=1}^{n} \int_{X_i/x}^{\infty} K^{[m,r]}(u)\, du$$

$$= \frac{1}{nx^m} \sum_{i=1}^{n} \bar{K}^{[m,r]}\left(X_i/x\right)$$

qui est un **estimateur de la dérivée d'ordre** $(m-1)$ de la fonction de densité.

Exemple : Estimation de la CDF et de ses dérivées

Si $\Phi(x, F) = F(x)$, alors la **variance** des estimateurs satisfait :

- Pour l'**estimateur de la CDF** :

$$\mathrm{Var}\left(\hat{F}_n^{[0,r]}(x)\right) = \frac{F(x)(1-F(x))}{n} + O\left(\frac{h}{n}\right).$$

- Pour l'**estimateur de la densité** et ses dérivées :

$$\mathrm{Var}\left(\hat{f}_n^{[m-1,r]}(x)\right) = \frac{f(x)b_h^{[m,r]}}{nh^{2m-1}x^{2m-1}} + o\left(\frac{1}{nh^{2m-1}}\right), \quad \text{pour } 1 \leq m \leq$$

Sous les conditions : $h_n \to 0$, $nh_n^{2m-1} \to \infty$ lorsque $n \to \infty$,

• l'**erreur quadratique moyenne** de l'estimateur tend vers zéro asymptotiquement :

$$\mathbb{E}\left(\hat{f}_n^{[m-1,r]}(x) - f^{(m-1)}(x)\right)^2 \to 0.$$

Estimation de la CDF et de ses dérivées (suite)

- Normalité asymptotique de l'estimateur de la densité :
  Si $1 \leq m \leq r$

  $$\text{et} \quad nh_n^{2r+1} \to 0, \quad nh_n^{2m-1} \to \infty \quad \text{lorsque} \quad n \to \infty,$$

  alors l'estimateur de la $(m-1)$-ième dérivée de la fonction de densité satisfait :

  $$\sqrt{nh^{2m-1}}\left(\hat{f}_n^{[m-1,r]}(x) - f^{(m-1)}(x)\right) \xrightarrow[\mathcal{L}]{} \mathcal{N}\left(0, \frac{b^{[m,r]}f(x)}{x^{2m-1}}\right).$$

- Normalité asymptotique de l'estimateur de la CDF :

  $$\textbf{Si} \qquad nh_n^{2r+2} \to 0 \quad \text{lorsque} \quad n \to \infty,$$

  alors l'estimateur empirique de la CDF satisfait :

  $$\sqrt{n}\left(\hat{F}_n^{[0,r]}(x) - F(x)\right) \xrightarrow[\mathcal{L}]{} \mathcal{N}\Big(0, F(x)(1-F(x))\Big).$$

Estimation de la densité dans les modèles avec biais de sélection

Les données biaisées par sélection se rencontrent dans des contextes tels que: les données manquantes, l'échantillonnage, les observations endommagées et l'économie.

- La variable aléatoire cible Y avec densité f n'est pas observée directement.
- Nous observons une variable X avec fonction de distribution G et densité g. La relation entre ces densités est :

$$g(x) = \frac{w(x)f(x)}{\mu_w}, \quad x > 0,$$

où $w(x)$ est une fonction de poids positive connue, et

$$\mu_w = \int_0^\infty w(x)f(x)dx < \infty.$$

Estimation de la densité dans les modèles avec biais de sélection (suite)

▷ Pour **estimer** f à partir d'un échantillon observé $X_1, \ldots, X_n$ tiré de G, nous utilisons la fonctionnelle

$$\Phi(x, F) = \int_0^\infty \phi(x, u) dG(u), \quad \text{où} \quad \phi(x, u) = \frac{\mu_w \mathbf{1}_{\{u \leq x\}}}{w(u)}.$$

- Cela conduit à l'estimateur

$$\hat{F}_n^{[m,r]}(x) = \hat{f}_n^{(m-1)}(x) = \frac{\hat{\mu}_w}{n x^m} \sum_{i=1}^n \bar{K}^{[m,r]}\left(\frac{X_i}{x}\right) w(X_i),$$

où $\hat{\mu}_w = \frac{1}{n}\sum_{i=1}^n w(X_i)$.

- L'estimateur de la densité est obtenue en posant $m = 1$.

Estimation Non Paramétrique du Ratio de Densité

Soit G une distribution connue avec une densité $g(x) > 0$.

• **Objectif**. Estimer la densité de $R = f/g$ à partir d'un échantillon $Y_1, \ldots, Y_n$ tiré de F.

En utilisant la représentation fonctionnel linéaire de F:

$$G(x) = \Phi(x, F) = \int \frac{\mathbf{1}\{z \leq x\}}{g(z)} dF(z).$$

- Un estimateur de G et de ses dérivées, pour $x > 0$, est donné :

$$\hat{G}_n^{[m,r]}(x) = \frac{1}{nx^m} \sum_{i=1}^{n} \bar{K}^{[m,r]}\left(\frac{Y_i}{x}\right) \frac{1}{g(Y_i)}.$$

- Pour le ratio de densité R, en choisissant $m = 1$, on obtient

$$\hat{R}_n(x) = \frac{1}{nx} \sum_{i=1}^{n} \bar{K}^{[1,r]}\left(\frac{Y_i}{x}\right) \frac{1}{g(Y_i)}.$$

Estimation de la Fonction de Risque

Les fonctions de risque jouent un rôle important en analyse de survie. Étant donné un temps de survie $X \sim F_0$ et un temps de censure $Y \sim H$ indépendants, on **observe** $\delta = \mathbf{1}(X \leq Y)$ et $Z = \min(X, Y)$.

- La **fonction de survie** de Z est :

$$S_Z(x) = (1 - F_0(x))(1 - H(x)).$$

- La **fonction de risque** $\alpha(x)$ est définie par :

$$\alpha(x) = \frac{(1 - H(x))f_0(x)}{S_Z(x)}.$$

Estimation de la Fonction de Risque

- Le **fonctionnel linéaire** associé est :

$$\Phi(x, F_0, H) = \int_0^x \alpha(u)\, du.$$

- Un estimateur de ce fonctionnel est :

$$\hat{\Phi}_n^{[m,r]}(x) = \frac{1}{nx^m} \sum_{i=1}^{n} \mathbf{1}_{\{\delta_i=1\}} \bar{K}^{[m,r]}\left(\frac{Z_i}{x}\right) \frac{1}{S_n(Z_i)}.$$

- L'**estimateur du taux de risque**, correspondant à $m = 1$, est:

$$\hat{\alpha}_n(x) = \frac{1}{nx} \sum_{i=1}^{n} \mathbf{1}_{\{\delta_i=1\}} \bar{K}^{[1,r]}\left(\frac{Z_i}{x}\right) \frac{1}{S_n(Z_i)}.$$

Simulations

- Dans notre comparaison numérique, nous considérons les estimateurs de densité suivants :
  - Estimateurs de densité à **noyau symétrique** classique.
  - Estimateurs de densité à **noyau asymétrique.**

Estimateurs de densité concurrents

Nous considérons les estimateurs concurrents suivants issus de la littérature :
Chen (2000) ont proposé deux estimateurs ($j = 1, 2$) :

- $\hat{f}_{ch_j}(x) = n^{-1} \sum_{i=1}^{n} K_j(X_i)$, où K_j est une densité Gamma avec paramètres $\alpha = \rho_{b,j}(x)$, $\beta = b = h_n^2$.
- $\rho_{b,1}(x) = (x/b) + 1$,
  $\rho_{b,2}(t) = (t/b)\mathbb{1}\{t \geq 2b\} + \left[(1/4)(t/b)^2 + 1\right] \mathbb{1}\{0 \leq t < 2b\}$.

Scaillet (2004) a introduit deux estimateurs :

- $\hat{f}_{IG}(x) = n^{-1} \sum_{i=1}^{n} K_{IG(x,1/b)}(X_i)$ (Gaussienne inverse).
- $\hat{f}_{RIG}(x) = n^{-1} \sum_{i=1}^{n} K_{IG(1/(x-b),1/b)}(X_i)$ (Gaussienne inverse réciproque).

Estimateurs de densité concurrents (suite)

▷ **L'estimateur à noyau tronqué et normalisé** est défini par :

$$\hat{f}_{cn}(x) = \frac{1}{nh}\sum_{i=1}^{n}\left[K_{cn}\left(\frac{x-X_i}{h}\right)\mathbb{1}\{[0,h)\}(x) + K\left(\frac{x-X_i}{h}\right)\mathbb{1}\{[h,\infty)\}(x)\right], \quad \text{où}$$

$$K(t) = \frac{3}{4}(1-t^2)\mathbb{1}\{[-1,1]\}(t),$$

$$K_{cn}(t) = \frac{(1-t^2)}{\int_{-1}^{c}(1-t^2)dt}\mathbb{1}\{[-1,c]\}, \quad c \geq 0.$$

▷ **L'estimateur à noyau au bord** est donné par :

$$\hat{f}_B(x) = \begin{cases} \frac{1}{nh}\sum_{i=1}^{n} K_c\left(\frac{x-X_i}{h}\right), & x \in [0,h), \quad c = x/h, \\ \frac{1}{nh}\sum_{i=1}^{n} K\left(\frac{x-X_i}{h}\right), & x \in [h,\infty), \quad \text{où} \end{cases}$$

$$K_c(t) = \frac{12(1+t)}{(1+c)^4}\left[t(1-2c) + \frac{3c^2-2c+1}{2}\right]\mathbb{1}\{[-1,c]\}(t), \quad c = x/h.$$

Étude de Simulation

Les données sont générées selon l'une des cinq densités suivantes :

1. Densité demi-normale : $f(x) = \sqrt{\frac{2}{\pi}} e^{-x^2/2}$ pour $x \geq 0$.
2. Densité exponentielle : $f(x) = e^{-x}$ pour $x \geq 0$.
3. Densité de Weibull : $f(x) = 2xe^{-x^2}$ pour $x \geq 0$.
4. Chi-deux avec 6 degrés de liberté : $f(x) = \frac{1}{16} x^2 e^{-x/2}$ pour $x \geq 0$.
5. Log-normale : $f(x) = \frac{1}{x\sqrt{2\pi}} \exp\left(-\frac{(\ln x)^2}{2}\right)$ pour $x > 0$.

Étude de Simulation

- Le paramètre de lissage a été sélectionnée pour minimiser l'Erreur Quadratique Intégrée (ISE), estimée à l'aide de l'Erreur Quadratique Intégrée Moyenne (AISE) sur 1000 réplicas de simulation.
- L'ISE a été calculée numériquement comme suit :

$$\int (\widehat{f}(x) - f(x))^2 dx, \quad \text{où } \widehat{f} \text{ représente l'un des estimateurs.}$$

- Les estimateurs considérés sont basés sur différents noyaux :

$\widetilde{f}_{\gamma}^{[0,r]}$ (Gamma), $\widetilde{f}_{\beta}^{[0,r]}$ (Bêta décalé), $\widetilde{f}_{T}^{[0,r]}$ (Triangulaire),

$\widehat{f}_{RIG}$ (Inverse Gaussien réciproque), $\widehat{f}_{ch2}$ (Chen 2),

$\widehat{f}_{Ep}$ (Epanechnikov), $\widehat{f}_{B}$ (Au bord), $\widehat{f}_{ch}$ (Chen 1).

results for n=100

Table: Optimal bandwidth and optimal AISE for $n = 100$

Dist.		$\tilde{f}_{\gamma}^{[0,1]}$	$\tilde{f}_{\gamma}^{[0,2]}$	$\tilde{f}_{\beta}^{[0,1]}$	$\tilde{f}_{T}^{[0,1]}$	$\hat{f}_{RIG}$	$\hat{f}_{ch2}$	$\hat{f}_{Ep}$	$\hat{f}_{B}$	$\hat{f}_{cn}$
1	h_n^*	0.365	0.337	0.390	0.3136	0.042	0.175	0.350	1.530	0.748
	ϵ_n^*	0.258	0.322	0.354	0.3511					
	AISE	0.011	0.014	0.015	0.0138	0.028	0.009	0.025	0.021	0.006
2	h_n^*	0.501	0.442	0.466	0.3741	0.034	0.171	0.283	1.525	0.592
	ϵ_n^*	0.124	0.181	0.194	0.1733					
	AISE	0.012	0.016	0.021	0.0197	0.035	0.011	0.035	0.021	0.011
3	h_n^*	0.263	0.251	0.294	0.2549	0.055	0.055	0.428	0.662	0.346
	ϵ_n^*	0.027	0.029	0.020	0.0159					
	AISE	0.018	0.021	0.021	0.0206	0.013	0.013	0.012	0.040	0.018
4	h^*	0.2838	0.2718	0.4234	0.3438	0.3598	0.3565	1.9994	1.9999	1.9880
	ϵ^*	0.0200	0.0200	0.0200	0.0300					
	AISE	0.0024	0.0027	0.0019	0.0019	0.0019	0.0019	0.0025	0.0112	0.0031
5	h^*	0.4405	0.4040	0.4416	0.3515	0.0704	0.0650	0.4129	0.8066	0.3599
	ϵ^*	0.0100	0.0100	0.0100	0.0100					
	AISE	0.0103	0.0138	0.0145	0.0136	0.0121	0.0128	0.0181	0.0436	0.0245

Results for n=200

Table: Optimal bandwidth and optimal AISE for $n = 200$

Dist.		$\tilde{f}_{\gamma}^{[0,1]}$	$\tilde{f}_{\gamma}^{[0,2]}$	$\tilde{f}_{\beta}^{[0,1]}$	$\tilde{f}_{T}^{[0,1]}$	$\hat{f}_{RIG}$	$\hat{f}_{ch2}$	$\hat{f}_{Ep}$	$\hat{f}_{B}$	$\hat{f}_{cn}$
1	h_n^*	0.3109	0.2783	0.3326	0.2776	0.0289	0.1324	0.2621	1.4488	0.6590
	ϵ_n^*	0.2583	0.3099	0.3171	0.3150					
	AISE	0.0077	0.0091	0.0094	0.0095	0.0186	0.0061	0.0190	0.0193	0.0037
2	h_n^*	0.4303	0.3662	0.4147	0.3384	0.0213	0.1300	0.2003	1.4421	0.4713
	ϵ_n^*	0.1061	0.1427	0.1417	0.1397					
	AISE	0.0079	0.0103	0.0128	0.0123	0.0225	0.0063	0.0250	0.0181	0.0071
3	h_n^*	0.2254	0.2151	0.2392	0.2208	0.0425	0.0426	0.3717	0.6573	0.2943
	ϵ_n^*	0.0100	0.0100	0.0100	0.0100					
	AISE	0.0113	0.0125	0.0125	0.0126	0.0076	0.0077	0.0075	0.0362	0.0110
4	h^*	0.2405	0.2273	0.3564	0.2938	0.2689	0.2746	1.9990	1.9999	1.6847
	ϵ^*	0.0201	0.0200	0.0300	0.0300					
	AISE	0.0014	0.0016	0.0011	0.0011	0.0011	0.0011	0.0014	0.0095	0.0018
5	h^*	0.3764	0.3370	0.4026	0.3124	0.0534	0.0493	0.3266	0.6121	0.2534
	ϵ^*	0.0100	0.0100	0.0100	0.0100					
	AISE	0.0060	0.0079	0.0077	0.0076	0.0069	0.0072	0.0115	0.0379	0.0158

Résumé des Résultats

- Les résultats des Tableaux 1 et 2 indiquent que l'estimateur $\widetilde{f}_{\gamma}^{[0,1]}$ est plus précis que les autres estimateurs par polynômes locaux :

$$\left(\widetilde{f}_{\gamma}^{[0,2]}, \quad \widetilde{f}_{\beta}^{[0,1]}, \quad \widehat{f}_{T}^{[0,1]}\right).$$

- En comparant les estimateurs par polynômes locaux avec d'autres méthodes :
  - Les performances de $\widetilde{f}_{\gamma}^{[0,1]}$ sont, en général, très proches de celles des principaux concurrents $\widehat{f}_{ch2}$ et $\widehat{f}_{ch1}$.
  - Il présente un avantage clair pour certaines distributions, comme la log-normale.

Merci

Merci !

Questions ?

Bagai, I. and PrakasaRao, B. (1995). Kernel type density estimates for positive valued random variables. *Sankhyā B*, 57(1):56–67.

Balakrishna, N. and Koul, H. (2017). Varying kernel marginal density estimator for a positive time series. *Journal of Nonparametric Statistics*, 29(3):531– 552.

Chaubey, Y., Li, J., Sen, A., and Sen, P. (2012). A new smooth density estimator for non-negative density etimator. *Journal of Indian Statistical Association*, 50:83–104.

Chen, S. X. (2000). Probability density function estimation using gamma kernels. *Annals of the Institute of Statistical Mathematics*, 52(3):471–480.

Kakizawa, Y. and Igarashi, G. (2017). Inverse gamma kernel density estimation for nonnegative data. *Journal of the Korean Statistical Society*, 46(2):194–207.

Scaillet, O. (2004). Density estimation using inverse and reciprocal

inverse gaussian kernels. *Journal of Nonparametric Statistics*, 16(1-2):217–226.

Silverman, B. (1986). *Density Estimation*. Chapman and Hall.

Wand, M. P., Marron, J. S., and Ruppert, D. (1991). Transformations in density estimation. *Journal of the American Statistical Association*, 86(414):343–353.