Chapitre 1

Généralités sur l'estimation NP

Chapitre 2

Estimateur NP de la Densité

Soit X une variable aléatoire de densité de probabilité inconnue f. Supposons que nous avons n observations $X_1, X_2, ..., X_n$ provenant de X. Le problème consiste à trouver un estimateur pour la fonction f à partir de cet échantillon issu de X. Pour cela, l'approche non paramétrique est plus adéquate lorsqu'on ne possède aucune information précise sur la forme et la classe de la vraie densité. Dans cette approche, ce sont les observations qui vont nous permettre de déterminer un estimateur pour la densité f.

On observe $X_1, \ldots, X_n$ échantillon issu d'une v.a. réelle X, de fonction de répartition (fdr) F :

$$F(x) = P(X \leq x)$$

et de densité de probabilité f :

$$P\left(X \in [a, b]\right) = \int_a^b f\left(x\right) dx, \qquad \forall a, b \in \mathbb{R}.$$

Nous supposons de plus, que f est deux fois continûment différentiable.

2.1 Estimation de la densité par histogramme

En statistique, l'histogramee est une représentation graphique de la répartition d'une variable aléatoire X (Pearson, 1895). Supposons que f est à support compact inclus dans $[0,1]$. Soit $C_1, ..., C_m$ une partition uniforme de $[0,1[$:

$$C_k = [\frac{k-1}{m}, \frac{k}{m}[, \quad k = 1, ..., m.$$

Il est clair que f est bien approchée par des fonctions en escalier, constantes par morceaux sur les intervalles C_j. Posons $h = \frac{1}{m}$ et on approche f par la fonction

$$f_h(x) = \sum_{k=1}^{m} \frac{p_k}{h} I_{C_k(x)}.$$

avec $p_k = \int_{C_k} f(x)\, dx = E\left(I_{C_k(X)}\right)$. Alors, il est naturel d'estimer p par

$$\hat{p}_k = \frac{1}{n} \sum_{j=1}^{n} I_{C_k(X_j)}, \qquad k = 1, ..., m.$$

Observons que chaque $\hat{p}_k$ représente la proportion des observations X_j se trouvant dans l'intervalle C_k. Par substitution, nous définissons l'estimateur de f par histogramme à m classes comme suit :

$$\hat{f}_h(x) = \sum_{k=1}^{m} \frac{\hat{p}_k}{h} I_{C_k(x)}. \tag{2.1}$$

Remarque 2.1.1 *On dit que chaque C_k est une classe de longueur (ou fenêtre) h. La hauteur des rectangles représente les fréquences absolues (nombre d'observations dans chaque classe) ou bien il s'agit des fréquences relatives comme dans la figure suivante.*

2.1.1 Propriétés asymptotique de l'histogramme

Il est clair que la qualité d'ajustement par histogramme dépend fortement de la fenêtre h. Il est naturel donc, d'étudier le risque quadratique de $\hat{f}_h$ au point $x \in [0,1]$ comme étant l'erreur

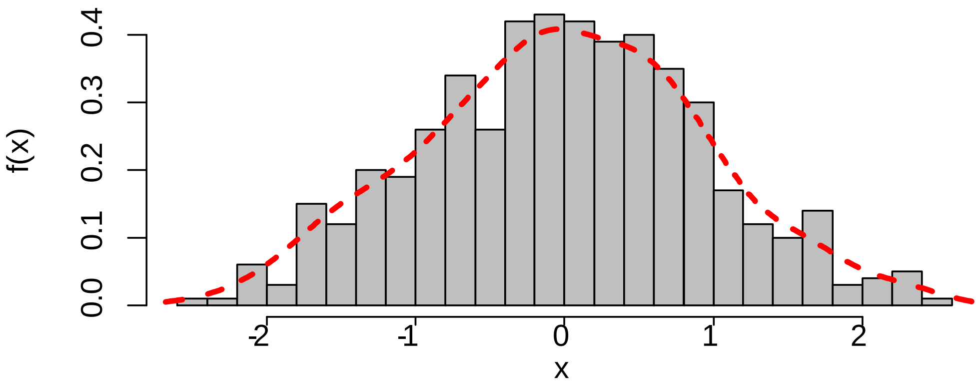

FIG. 2.1 – Estimation par histogramme basée sur un échantillon de taille $n = 500$, $X \sim N(0,1)$, $m = 30$, $h \simeq 0.2$

quadratique moyenne (MSE) définit par :

$$MSE\left(\hat{f}_h\right) = E\left(\left(\hat{f}_h(x) - f(x)\right)^2\right) = \left(E\hat{f}_h(x) - f(x)\right)^2 + Var\left(\hat{f}_h(x)\right)$$
$$= Biais^2\left(\hat{f}_h(x)\right) + Var\left(\hat{f}_h(x)\right)$$

Pour tout $x \in C_k$, on a

$$\hat{f}_h(x) = \frac{\hat{p}_k}{h} = \frac{1}{nh}\sum\nolimits_{j=1}^{n} I_{C_k(X_j)} = \frac{W_k}{nh},$$

avec W_k est la somme de n variables indépendantes de loi de Bernoulli de paramètre p_k :

$$P\left(I_{C_k(X_j)} = 1\right) = P\left(X_j \in C_k\right) = \int_{C_k} f(x)\,dx = p_k.$$

Donc, $\forall x \in C_k$:

$$E\left(\hat{f}_h(x)\right) = \frac{p_k}{h} \text{ et } Var\left(\hat{f}_h(x)\right) = \frac{np_k(1-p_k)}{n^2h^2} = \frac{p_k(1-p_k)}{nh^2}.$$

En déduit de ce dernier résultat que

$$MSE\left(\hat{f}_h\right) = \left(\frac{p_k}{h} - f(x)\right)^2 + \frac{p_k(1-p_k)}{nh^2}. \tag{2.2}$$

Afin d'avoir une évaluation globale valable pour tout point $x \in [0,1]$, on considère le risque quadratique intégré moyen $(MISE)$:

$$\begin{aligned} MISE\left(\hat{f}_h\right) &= \int MSE\left(\hat{f}_h(x)\right)dx = E\left(\int \left(\hat{f}_h(x) - f(x)\right)^2\right) \\ &= \int Biais^2\left(\hat{f}_h(x)\right)dx + \int Var\left(\hat{f}_h(x)\right)dx. \end{aligned}$$

Premièrement, pour le biais, on a

$$\begin{aligned} \int Biais^2\left(\hat{f}_h(x)\right)dx &= \sum\nolimits_{k=1}^m \int_{C_k} \left(\frac{p_k}{h} - f(x)\right)^2 dx \\ &= \sum\nolimits_{k=1}^m \frac{p_k^2}{h^2}\int_{C_k} dx - 2\sum\nolimits_{k=1}^m \frac{p_k}{h}\int_{C_k} f(x)\,dx + \sum\nolimits_{k=1}^m \int_{C_k} f^2(x)\,dx \\ &= \sum\nolimits_{k=1}^m \frac{p_k^2}{h} - 2\sum\nolimits_{k=1}^m \frac{p_k^2}{h} + \int_0^1 f^2(x)\,dx, \end{aligned}$$

puisque, on a

$$\int_{C_k=\left[\frac{k-1}{m},\frac{k}{m}\right[} dx = \frac{1}{m} = h, \quad \int_{C_k} f(x)\,dx = p_k \text{ et } \sum\nolimits_{k=1}^m \int_{C_k} f^2(x)\,dx = \int_0^1 f^2(x)\,dx.$$

Nous obtenons donc,

$$\int Biais^2\left(\hat{f}_h(x)\right)dx = \int \left(E\hat{f}_h(x) - f(x)\right)^2 dx = \int f^2(x)\,dx - \frac{1}{h}\sum\nolimits_{k=1}^m p_k^2. \tag{2.3}$$

Pour le terme de la variance, on a

$$\begin{aligned} \int Var\left(\hat{f}_h(x)\right)dx &= \sum\nolimits_{k=1}^m \int_{C_k} Var\left(\hat{f}_h(x)\right)dx \\ &= \sum\nolimits_{k=1}^m \frac{p_k(1-p_k)}{nh^2}\int_{C_k} dx = \frac{1}{nh}\sum\nolimits_{k=1}^m p_k(1-p_k) \\ &= \frac{1}{nh}\sum\nolimits_{k=1}^m p_k - \frac{1}{nh}\sum\nolimits_{k=1}^m p_k^2 \\ &= \frac{1}{nh}\sum\nolimits_{k=1}^m \int_{C_k} f(x)\,dx - \frac{1}{nh}\sum\nolimits_{k=1}^m p_k^2 \\ &= \frac{1}{nh}\int f(x)\,dx - \frac{1}{nh}\sum\nolimits_{k=1}^m p_k^2 = \frac{1}{nh} - \frac{1}{nh}\sum\nolimits_{k=1}^m p_k^2. \end{aligned} \tag{2.4}$$

Théorème 2.1.1 *Soit $X_1, \dots, X_n$ un échantillon issu d'une v.a. réelle X, de densité f et $\hat{f}_h$ est l'estimateur de f par histogramme, où $m = 1/h$ classes, alors :*

$$MISE\left(\hat{f}_h\right) = \int f^2(x)\,dx + \frac{1}{nh} - \frac{1}{h}\left(1 + \frac{1}{n}\right)\sum\nolimits_{kj=1}^{m} p_k^2.$$

Preuve. La preuve du résultat est détaillée dans (2.3) et (2.4). ■

Théorème 2.1.2 *Supposons que la densité f de X est deux fois continûment différentiable et s'annule en dehors de l'intervalle $[0,1]$. Sous la condition que*

$$h := h_n, \text{ et } h \to 0 \text{ quand } n \to \infty$$

Alors, lorsque $n \to \infty$,

$$MISE\left(\hat{f}_h\right) = \frac{h^2}{12}\int f'(x)^2\,dx + \frac{1}{nh} + O\left(h^3\right) + O\left(\frac{1}{n}\right). \tag{2.5}$$

Preuve. Du Théorème 1, on a

$$\begin{aligned} MISE\left(\hat{f}_h\right) &= \int f^2(x)\,dx + \frac{1}{nh} - \frac{1}{h}\left(1 + \frac{1}{n}\right)\sum\nolimits_{kj=1}^{m} p_k^2 \\ &= \sum\nolimits_{kj=1}^{m}\left\{\int_{C_k} f^2(x)\,dx - \frac{p_k^2}{h}\right\} + \frac{1}{nh}(1 - \sum\nolimits_{kj=1}^{m} p_k^2) \end{aligned}$$

Il est clair que,

$$\begin{aligned} \int_{C_k} f^2(x)\,dx - \frac{p_k^2}{h} &= \int_{C_k} f^2(x)\,dx - \frac{1}{h}\left(\int_{C_k} f(x)\,dx\right)^2 \\ &= \int_{C_k} f^2(x)\,dx - \frac{2}{h}\left(\int_{C_k} f(x)\,dx\right)^2 - \frac{1}{h}\left(\int_{C_k} f(x)\,dx\right)^2 \\ &= \int_{C_k} f^2(x)\,dx - \frac{2}{h}\int_{C_k} f(x)\,dx\int_{C_k} f(t)\,dt + \frac{1}{h^2}\int_{C_k}\left(\int_{C_k} f(x)\,dx\right)^2 dx \end{aligned}$$

donc

$$\begin{aligned}\int_{C_k} f^2(x)\,dx - \frac{p_k^2}{h} &= \int_{C_k} \left\{ f^2(x) - \frac{2}{h} f(x) \int_{C_k} f(t)\,dt + \frac{1}{h^2} \left(\int_{C_k} f(x)\,dx \right)^2 \right\} dx \\ &= \int_{C_k} \left(f(x) - \frac{1}{h} \int_{C_k} f(t)\,dt \right)^2 dx = \int_{C_k} \left(\frac{1}{h} \int_{C_k} f(x)\,dt - \frac{1}{h} \int_{C_k} f(t)\,dt \right)^2 dx \\ &= \frac{1}{h^2} \int_{C_k} \left(\int_{C_k} \{ f(x) - f(t) \}\,dt \right)^2 dx,\end{aligned}$$

Puisque f est deux fois continûment différentiable, alors pour tout $x, t \in C_k$:

$$f(x) - f(t) = (x - t) f'(\theta_k) + O\left(h^2\right),$$

avec θ_k désigne l'extrémité gauche de l'intervalle C_k. Donc,

$$\begin{aligned}\int_{C_k} f^2(x)\,dx - \frac{p_k^2}{h} &= \frac{1}{h^2} \int_{C_k} \left(\int_{C_k} \left\{ (x - t) f'(\theta_k) + O\left(h^2\right) \right\} dt \right)^2 dx \\ &= \frac{1}{h^2} f'(\theta_k)^2 \int_{C_k} \left(\int_{C_k} (x - t)\,dt \right)^2 dx + O\left(h^4\right).\end{aligned}$$

En utilisant le changement de variable $(x, t) = (\theta_k + yh, \theta_k + zh)$, on obtient

$$x, t \in [\theta_k, \theta_k + h[, \quad dx = hdy, \quad dt = hdz \quad et \quad y, z \in [0, 1[,$$

donc

$$\begin{aligned}\int_{C_k} \left(\int_{C_k} (x - t)\,dt \right)^2 dx &= \int_{C_k} \left(\int_{C_k} xdt - \int_{C_k} tdt \right)^2 dx \\ &= h^5 \int \left(\int (y - z)\,dz \right)^2 dy = \frac{h^5}{12}.\end{aligned}$$

Par conséquence,

$$\int_{C_k} f^2(x)\,dx - \frac{p_k^2}{h} = \frac{h^3}{12} f'(\theta_k)^2 + O\left(h^4\right) = \frac{h^2}{12} \int_{C_k} f'(x)^2\,dx + O\left(h^4\right).$$

Donc,

$$
\begin{aligned}
MISE\left(\hat{f}_h\right) &= \sum\nolimits_{kj=1}^{m}\left(\int_{C_k} f^2\left(x\right)dx - \frac{p_k^2}{h}\right) + \frac{1}{nh} - \frac{1}{nh}\sum\nolimits_{kj=1}^{m} p_k^2. \\
&= \frac{h^2}{12}\int f'\left(x\right)^2 dx + \frac{1}{nh} + O\left(h^3\right) + O\left(\frac{1}{n}\right).
\end{aligned}
$$

■

Corollaire 2.1.1 *Ce résultat nous permet de calculer la fenêtre h optimale notée h_{opt}, en minimisant la $MISE$ asymptotique :*

$$
\begin{aligned}
h_{opt} &= \arg\min_h AMISE\left(\hat{f}_h\right) = \arg\min_h \frac{h^2}{12}\int f'\left(x\right)^2 dx + \frac{1}{nh} \\
&= \left(\frac{n}{6}\int f'\left(x\right)^2 dx\right)^{-1/3} \simeq Cn^{-1/3}.
\end{aligned}
$$

Remarque 2.1.2 *Cette fenêtre optimale est en général incalculable (donc inutilisable, du point de vue pratique), car la densité f (ainsi que sa dérivée f') est inconnue. Cependant, cette fenêtre optimale est de l'ordre de $n^{-1/3}$ pour n assez grand.*

Remarque 2.1.3 *Dans le cas général où f est à support $[a,b]$, on peut poser*

$$\forall m \in \mathbb{N}^*, \quad h = \frac{b-a}{m}.$$

On définit alors les classes C_k comme suit :

$$C_k = [a+(k-1)h, a+kh[\quad pour \quad j = 1, ..., m-1 \quad et \quad C_m = [b-h, b].$$

2.2 Estimation de la densité par noyau

Par construction, les histogrammes ne sont pas des fonctions continues, ni lisse. Si on cherche donc, à estimer une fonction densité (supposée deux fois continûment différentiable), il sera plus naturel que l'estimateur possède les mêmes propriétés de continuité et de différentiabilité et lisse en plus, donc plus proche de la vraie densité que l'estimateur par histogramme.

Soit $x \in R$ et $h > 0$. Si l'on suppose que x est le centre d'une classe de l'histogramme et que h est la longueur des classes, l'estimateur de la densité $f(x)$ par histogramme peut s'écrire :

$$\hat{f}_h(x) = \frac{1}{nh}\sum\nolimits_{j=1}^{n} I_{\left(|x-X_j|\leq\frac{h}{2}\right)} = \frac{1}{nh}\sum\nolimits_{j=1}^{n} I_{\left(\left|\frac{x-X_j}{h}\right|\leq\frac{1}{2}\right)}.$$

Une façon de généraliser les histogrammes consiste donc à utiliser la formule ci-dessus pour tout $x \in R$ et pas seulement pour les centres des classes. Par conséquent, en remplaçant $I_{\left(|z|\leq\frac{1}{2}\right)}$ par une fonction quelconque K, on obtient l'estimateur

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

A partir de la définition d'une densité de probabilité et en utilisant la distribution empirique comme estimateur NP de la distribution F, on aura pour h assez petite $(h \to 0,\ \text{quand } n \to \infty)$:

$$f\,(x) = \frac{F(x+h) - F(x-h)}{2h},$$

ce qui implique

$$\begin{aligned}
f_n(x) &= \frac{F_n(x+h) - F_n(x-h)}{2h} \\
&= \frac{1}{2h}\left(\frac{1}{n}\sum_{j=1}^{n} I_{(X_j\leq x+h)} - \frac{1}{n}\sum_{j=1}^{n} I_{(X_j\leq x-h)}\right) \\
&= \frac{1}{2nh}\sum_{j=1}^{n}\left(I_{\left(\frac{X_j-x}{h}\leq 1\right)} - I_{\left(\frac{X_j-x}{h}\leq -1\right)}\right) = \frac{1}{2nh}\sum_{j=1}^{n} I_{\left(-1\leq\frac{X_j-x}{h}\leq 1\right)} \\
&= \frac{1}{nh}\sum_{j=1}^{n}\frac{1}{2} I_{\left(-\left|\frac{X_j-x}{h}\right|\leq 1\right)}.
\end{aligned}$$

Cette dernière peut être réécrite, en ses points de continuité, sous la forme suivante :

$$f_n(x) = \frac{1}{nh}\sum_{j=1}^{n} K\left(\frac{x-X_j}{h}\right) \tag{2.6}$$

avec $h := h_n$ est le paramètre de lissage (fenêtre ou bandwidth en anglais) choisi en fonction de n telle que

$$\lim_{n\to\infty} h = 0,$$

et K est la fonction des poids (noyau, kernel en anglais) où :

$$K(t) := \frac{1}{2} I_{(|t|<1)}.$$

Ce dernier est l'estimateur à noyau uniforme dit de Rosenblatt, proposé pour la premiere fois en 1956 par Rosenblatt. Six ans après, cet estimateur a été généralisé par Parzen (1962). A partir de cette date, cet estimateur a pris le nom de l'estimateur de Parzen-Rosenblatt. L'idée de l'estimateur par la méthode du noyau consiste donc, à évaluer la densité $f(x)$ au point x en comptant le nombre d'observations tombées dans un certain voisinage de x sur $\mathbb{R}$.

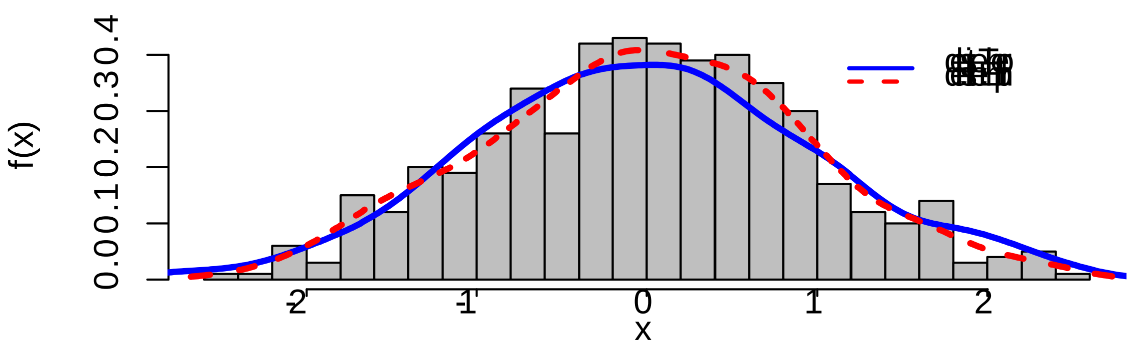

Fig. 2.2 – Estimation de la densité normale par noyau et histogramme

Définition 2.2.1 *Soit $K : R \to R$ une fonction (un noyau), $h > 0$ (une fenêtre), l'estimateur à noyau de la densité f est*

$$f_n(x) = \frac{1}{nh} \sum_{j=1}^{n} K\left(\frac{x - X_j}{h}\right).$$

Les noyaux les plus utilisés dans l'estimation de la densité de probabilité sont donnés dans le tableau suivant :

Proposition 2.2.1 *Si K est un noyau positif et $\int K(t)\,dt = 1$. Alors, f_n est une densité de probabilité. De plus, f_n a les mêmes propriétés de continuité et de différentiabilité que K.*

TAB. 2.1 – Noyaux usuels

Noyau	Fonction $K(t)$
Rectangulaire	$\frac{1}{2}1_{(\|t\|<1)}$
Triangulaire	$(1-\|t\|)1_{(\|t\|<1)}$
Gaussien	$\frac{1}{\sqrt{2\pi}}e^{-\frac{t^2}{2}},\ t\in R$
Quartique	$\frac{15}{16}(1-t^2)^2 1_{(\|t\|<1)}$
Epanechinkov	$\frac{3}{4}(1-t^2)1_{(\|t\|<1)}$

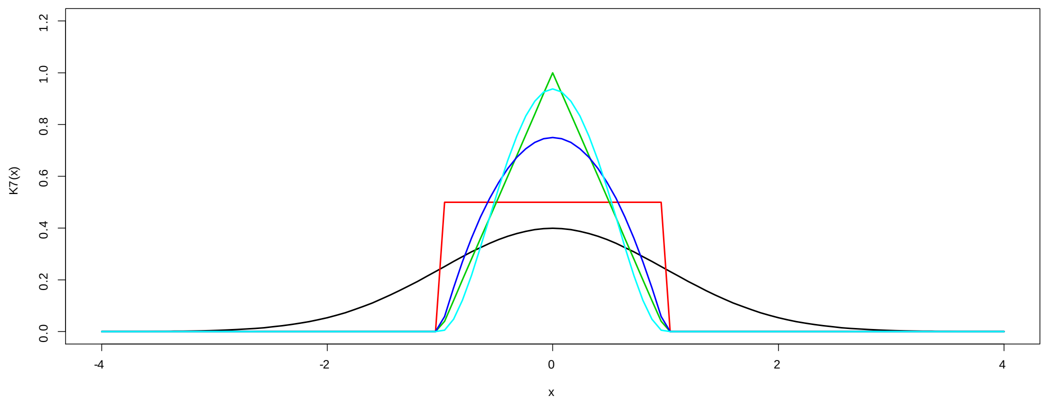

FIG. 2.3 – Courbes des noyaux usuels

Preuve. On a, $f_n(x) > 0$ et

$$\int f_n(x)\,dx = \frac{1}{nh}\int\sum_{j=1}^{n}K\left(\frac{x-X_j}{h}\right)dx = \frac{1}{nh}\sum_{j=1}^{n}\int hK(t)\,dt$$
$$= \int K(t)\,dt = 1.$$

avec le ch.v $(x = X_j + th)$. Donc f_n est bien une densité de probabilité. ■

2.2.1 Propriétés Asymptotique

Supposons que le noyau K vérifier les conditions suivantes :

(K1) $\int K(t)dt = 1$.

(K2) K symétrique autour de zéro, c.à.d $\int tK(t)dt = 0$.

(K3) K possède un moment d'ordre 2 fini, c.à.d $\int t^2K(t)dt < \infty$.

(K4) $\int t^2 \left|K(t)\right| dt < \infty.$

Proposition 2.2.2 *Si la densité f est bornée et que f'' existe et bornée aussi. Sous (K1-K3) :*

$$\left|Biais\left(f_n\right)\right| \leq C_1 h^2, \qquad avec\ C_1 = \frac{1}{2}\sup_z \left|f''\left(z\right)\right| \int t^2 \left|K(t)\right| dt.$$

Si, de plus, la condition (K4) est satisfaite, alors

$$Var\left(f_n\right) \leq \frac{C_2}{nh}, \quad avec\ C_2 = \sup_z f\left(z\right) \int\int K(t)^2 dt.$$

Preuve. Pour le biais, on a

$$\begin{aligned}
E\left(f_n\right) &= E\left(\frac{1}{nh}\sum_{j=1}^{n} K\left(\frac{x - X_j}{h}\right)\right) = \frac{1}{nh}\sum_{j=1}^{n} E\left(K\left(\frac{x - X_j}{h}\right)\right) \\
&= \frac{1}{nh}\sum \int K\left(\frac{y - x}{h}\right) f\left(y\right) dy \\
&= \int K\left(t\right) f\left(x + th\right) dt \qquad \left(ch.v : y = x + th, \quad dy = hdt\right)
\end{aligned}$$

En effectuons un développement limité d'ordre 2 à f, il vient

$$\begin{aligned}
E\left(f_n\right) &= \int K\left(t\right)\left\{f\left(x\right) + thf'\left(x\right) + \frac{t^2h^2}{2} f''\left(\theta\right)\right\} dt \qquad \theta \in \left[x, x + th\right] \\
&= f\left(x\right)\int K\left(t\right) dt + hf'\left(x\right)\int tK\left(t\right) dt + \frac{h^2}{2}\int t^2 f''\left(\theta\right) K\left(t\right) dt \\
&= f\left(x\right) + \frac{h^2}{2}\int t^2 f''\left(\theta\right) K\left(t\right) dt.
\end{aligned}$$

Alors,

$$\begin{aligned}
\left|Biais\left(f_n\right)\right| = \left|E\left(f_n\right) - f\right| &= \frac{h^2}{2}\left|\int t^2 f''\left(\theta\right) K\left(t\right) dt.\right| \\
&\leq \frac{h^2}{2}\int \left|t^2 f''\left(\theta\right) K\left(t\right)\right| dt. \\
&\leq \frac{h^2}{2}\sup_z \left|f''\left(z\right)\right| \int t^2 \left|K(t)\right| dt =: C_1 h^2.
\end{aligned}$$

Deuxièment, les variables aléatoires $Y_i = K\left(\frac{x-X_j}{h}\right)$ sont i.i.d. donc

$$\begin{aligned} Var\left(f_n\right) &= \frac{1}{n^2h^2}\sum_{j=1}^{n} Var\left(K\left(\frac{x-X_j}{h}\right)\right) \\ &\le \frac{1}{nh^2}E\left(K\left(\frac{x-X_j}{h}\right)^2\right) \end{aligned}$$

$(ch.v : y = x + th, \quad dy = hdt)$,

$$\begin{aligned} Var\left(f_n\right) &\le \frac{1}{nh^2}\int K^2\left(\frac{y-x}{h}\right) f\left(y\right) dy \\ &= \frac{1}{nh}\int K^2\left(t\right) f\left(x+th\right) dt \\ &\le \frac{1}{nh}\sup_z f\left(z\right) \int\int K(t)^2 dt =: \frac{C_2}{nh}. \end{aligned}$$

■

Remarque 2.2.1 *On déduit de cette Proposition que le risque MSE de f_n admet la majoration suivante :*

$$MSE\left(f_n\right) = Biais^2\left(f_n\right) + Var\left(f_n\right) \le C_1^2h^4 + \frac{C_2}{nh}.$$

La valeur de la fenêtre h qui minimise ce majorant du MSE est

$$h_{opt} = \left(C_2/4C_1^2\right)^{1/5} n^{-1/5}.$$

En injectant cette valeur dans l'expression du MSE on obtient : $MSE\left(f_n\right) \le C.n^{-4/5}$. Cela montre que la vitesse de convergence de l'estimateur à noyau est de $n^{-4/5}$. Elle est donc meilleure que la vitesse $n^{-2/3}$ obtenue pour les histogrammes.

Remarque 2.2.2 *Lorsque la fenêtre h est très petit $(h \to 0)$, le biais de l'estimateur à noyau est très petit face à sa variance et c'est cette dernière qui détermine la vitesse de convergence du risque quadratique. Dans ce type de situation, l'estimateur est très volatile et on parle de sous-lissage (under-smoothing, en anglais). En revanche, lorsque h grandit, la variance devient petite et* *c'est le biais qui devient dominant. L'estimateur est alors très peu variable et est de moins à moins influencé par les données. On parle alors d'un effet de sur-lissage (over-smoothing en anglais). En pratique, il est primordial de trouver la bonne dose de lissage qui permet d'éviter le sous-lissage et le sur-lissage.*

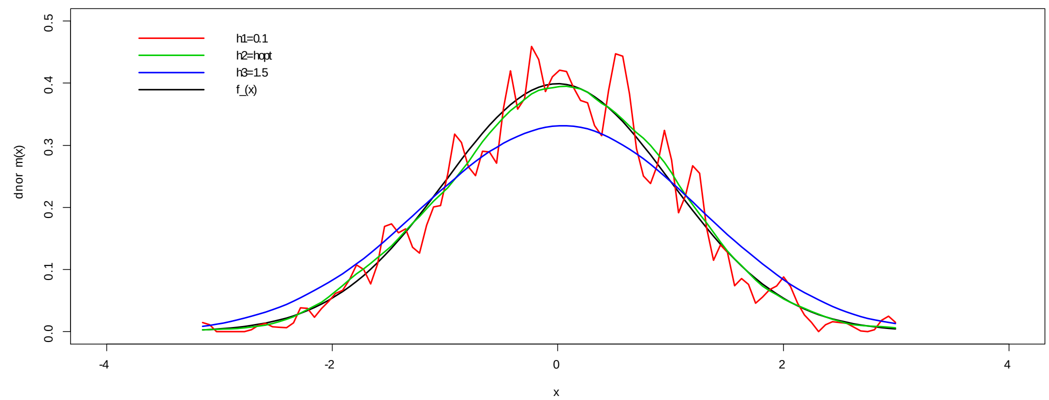

FIG. 2.4 – Biais, Var et choix de h : sur et sous lissage.

Proposition 2.2.3 *Supposons que $f \in C^2$ de carré intégrable, $h = h_n$, telle que*

$$h \to 0 \quad et \;\; nh \to \infty, \quad quand \; n \to \infty.$$

De plus, la fonction noyau K est supposée (densité, bornée, symétrique, de moment d'ordre 4 fini). Alors,

$$i) \;\; Biais(f_n(x)) = E(f_n(x)) - f(x) = \frac{h^2}{2}\mu_2(K)f''(x) + o(h^2),$$
$$ii) \;\; Var(f_n(x)) = \frac{1}{nh}f(x)R(K) + o(\frac{1}{nh}),$$

avec $\mu_2(K) := \int t^2 K(t)dt \;\; et \;\; R(g) := \int g^2(t)\,dt.$

Preuve. Pour i), on a

$$\begin{aligned}
E\left(\left(f_n\right)\right) &= \int K\left(t\right) f\left(x+th\right) dt \\
&= \int K\left(t\right) \left\{ f\left(x\right) + thf'\left(x\right) + \frac{t^2h^2}{2} f''\left(x\right) + o\left(h^2\right) \right\} dt \qquad \text{DT au voisinage de } x \\
&= f\left(x\right) + \frac{h^2}{2} f''\left(x\right) \int t^2 K\left(t\right) dt + o\left(h^2\right).
\end{aligned}$$

De même, pour ii)

$$\begin{aligned}
Var(f_n(x)) &= \frac{1}{n^2h^2} \sum_{j=1}^{n} Var\left(K\left(\frac{x-X_j}{h}\right)\right) \\
&= \frac{1}{nh^2} E\left(K\left(\frac{x-X_j}{h}\right)^2\right) - \frac{1}{nh^2} E\left(K\left(\frac{x-X_j}{h}\right)\right)^2 \\
&= \frac{1}{nh^2} \int K^2\left(\frac{y-x}{h}\right) f\left(y\right) dy - \frac{1}{n}\left(\frac{1}{h}\int K\left(\frac{y-x}{h}\right) f\left(y\right) dy\right)^2
\end{aligned}$$

En utilisant le changement de variable ($y = x + th, \quad dy = hdt$) et un développement de Taylor du f au voisinage de x d'ordre 1 :

$$\begin{aligned}
Var(f_n(x)) &= \frac{1}{nh} \int K^2\left(t\right) f\left(x+th\right) dt - \frac{1}{n}\left(\int K\left(t\right) f\left(x+th\right) dt\right)^2 \\
&= \frac{1}{nh} \int K^2\left(t\right) \left(f\left(x\right) + o\left(1\right)\right) dt - \frac{1}{n}\left(\int K\left(t\right) \left(f\left(x\right) + o\left(1\right)\right) dt\right)^2 \\
&= \frac{1}{nh} f\left(x\right) \int K^2\left(t\right) dt + o\left(1/nh\right) - o\left(1/n\right) \\
&= \frac{1}{nh} f\left(x\right) R\left(K\right) + o\left(1/nh\right),
\end{aligned}$$

car $1/n = o\left(1/nh\right)$. Le résultat est donc démontré. ■

Notation 2.2.1 *1)* $A_n = o\left(B_n\right) \Leftrightarrow \frac{A_n}{B_n} \to 0$

2) $A_n = O\left(B_n\right) \Leftrightarrow \lim \frac{A_n}{B_n} < \infty$.

Remarque 2.2.3 *L'estimateur $f_n(x)$ est asymptotiquement sans biais, i.e,*

$$\lim_{n\to\infty} E(\hat{f}_n(x)) = f(x).$$

2.2.2 Erreur quadratique moyenne (MSE) et intégré (MISE)

Sous les conditions sur K et en supposant que la densité de probabilité f avait toutes les dérivées (continues) nécessaires. On peut obtenir facilement les approximations suivantes pour la MSE et la $MISE$:

$$\begin{aligned} MSE(f_n(x)) &= E[(f(x) - f_n(x))^2] = Biais^2(f_n(x)) + Var(f_n(x)) \\ &= \frac{h^4}{4}\mu_2^2(K) f''(x)^2 + \frac{1}{nh} f\left(x\right) R\left(K\right) + o(\frac{1}{nh}) + o(h^2). \end{aligned}$$

D'ou la MSE asymptotique :

$$AMSE(f_n(x)) = \frac{h^4}{4}\mu_2^2(K) f''(x)^2 + \frac{1}{nh} f\left(x\right) R\left(K\right).$$

Par minimisation en h de la $AMSE$, on trouve la fenêtre optimale locale, notée h_{opt} :

$$\begin{aligned} h_{opt} &= Arg \min_h AMSE(f_n(x)) \\ &= \left(\frac{f\left(x\right) R\left(K\right)}{\mu_2^2(K) f''(x)^2}\right)^{1/5} n^{-1/5} \simeq C n^{-1/5}. \end{aligned} \tag{2.7}$$

Par intégration, la AMSE intégré notée AMISE est donnée par :

$$\begin{aligned} AMISE(f_n(x)) &= \int AMSE(f_n(x)) dx \\ &= \frac{h^4}{4}\mu_2^2(K) \int f''(x)^2 dx + \frac{1}{nh} R\left(K\right) \\ &= \frac{h^4}{4}\mu_2^2(K) R\left(f''\right) + \frac{1}{nh} R\left(K\right). \end{aligned}$$

Il est clair que cette quantité ne dépend pas de x, d'où le h optimal calculer par minimisation de la $AMISE$ est globale ;

$$\begin{aligned} h^*_{opt} &= Arg\min_h AMISE(f_n(x)) \\ &= \left(\frac{R(K)}{\mu_2^2(K)R(f'')}\right)^{1/5} n^{-1/5} \simeq Cn^{-1/5}. \end{aligned} \tag{2.8}$$

Exercice 2.2.1 *Calculer $R(K)$, $\mu_2^2(K)$ et $R(f'')$ dans les cas suivants :*

1) K noyau gaussien et f densité normale, $N(\mu, \sigma^2)$.

2) K noyau d'Epanechinkov et f densité exponentielle, $Exp(\lambda)$.

En déduire le h_{opt} et h^*_{opt} en fonction de n dans les deux cas.

2.2.3 Choix du noyau

Lorsqu'on définit un estimateur à noyau, on a non-seulement le choix de la fenêtre h, mais aussi celui du noyau K. Pour désigner un noyau optimal dans l'estimation NP de la densité, il suffit d'insérer la valeur du h^*_{opt} dans la $AMISE$:

$$\begin{aligned} AMISE(f_n(x)) &= \frac{\left(h^*_{opt}\right)^4}{4}\mu_2^2(K)R(f'') + \frac{1}{nh^*_{opt}}R(K) \\ &= \frac{1}{4}\left(\frac{R(K)}{\mu_2^2(K)R(f'')}\right)^{4/5} n^{-4/5}\mu_2^2(K)R(f'') + \left(\frac{R(K)}{\mu_2^2(K)R(f'')}\right)^{-1/5} n^{-4/5}R(K) \\ &= \frac{5}{4}\left\{\mu_2^2(K)R(f'')R^4(K)\right\}^{1/5} n^{-4/5}. \end{aligned} \tag{2.9}$$

Cette dernière expression ne dépend plus de h ou de x, elle dépend seulement du noyau K. La minimisation de (2.9) par rapport à K donne comme solution le noyau dite d'Epanechinkov (1969) :

$$K_{opt}(t) = \frac{3}{4}\left(1 - t^2\right) I_{(|t|<1)}.$$

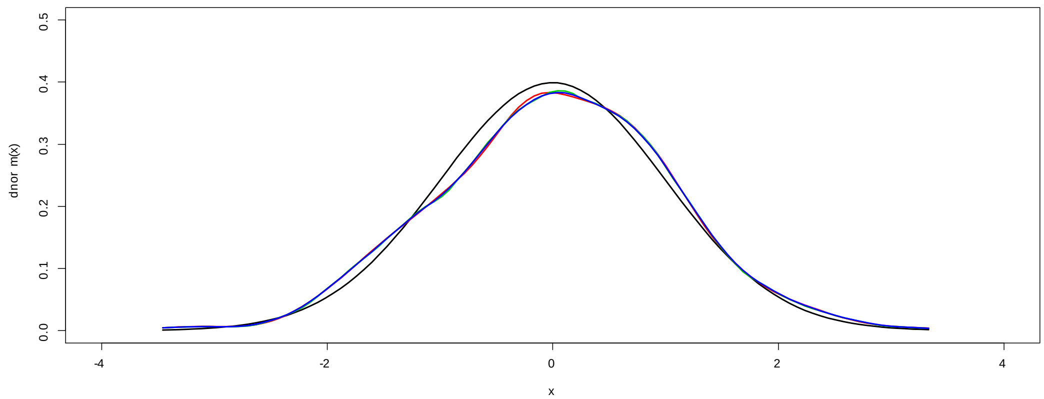

FIG. 2.5 – Effet du noyau : estimation d'une densité normale par déffirentes noyaux.

Définition 2.2.2 *L'efficacité relative d'un noyau K par rapport à K_{opt} est donnée par :*

$$eff\,(K) = \frac{AMISE\,(K_{opt})}{AMISE\,(K)} = \left(\frac{\mu_2^2(K_{opt})R^4\,(K_{opt})}{\mu_2^2(K)R^4\,(K)}\right)^{1/5} \leq 1. \tag{2.10}$$

Exercice 2.2.2 *Vérifier les résultats du tableau suivant :*

Noyau	$eff\,(K)$
Epanechinkov	1.000
Gaussien	0.951
Uniform	0.930
Triangulair	0.986
Quartique	0.994

Remarque 2.2.4 *Il est clair que le choix du noyau n'influe pas trop dans le cas des noyaux symétriques.*

2.2.4 Choix du paramètre de lissage

Contrairement au noyau K, la fenêtre h à un rôle important dans l'estimation à noyau. Elle détermine la qualité d'estimation et contrôle le lissage (sur lissage, sous lissage) d'une façon très sensible. En effet, le choix de h tour au tour de la condition :

$$h \to 0 \quad \text{et} \quad nh \to \infty.$$

Des faibles valeurs de h implique un sous-lissage ($Biais \to 0$) mais la variance augmente, et lorsque h grandit, la variance devient petite et c'est le biais qui augmente, donc sur-lissage de l'estimateur.

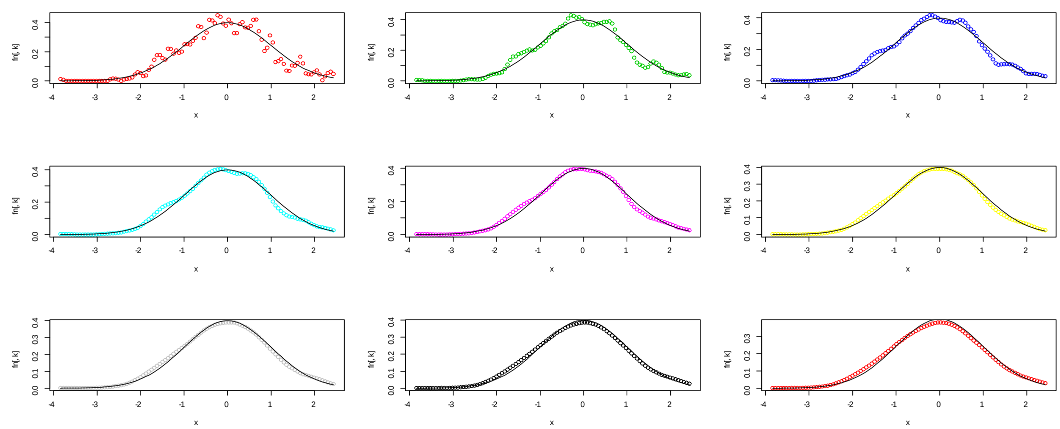

FIG. 2.6 – Effet de la fenêtre h : estimation d'une densité normale par déffirentes valeurs de h.

De plus, puisque h_{opt} et h^*_{opt} dépend des quantités inconnues (f, f''), donc pratiquement ne sont plus calculable. Plusieurs méthodes ont été développées pour résoudre se problème : Validation croisée, Plug-in et la règle de référence. C'est cette dernière qu'on va donner en détail dans la suite.

Règle de référence à une loi normale :

Silverman (1986) à proposer de se référer à une loi normale pour le calcul de h^*_{opt}. Soit $X_1, X_2, ..., X_n$ une suite de variables aléatoires de densité de probabilité f, supposons que f appartient à une famille de distributions normales, $N(\mu; \sigma^2)$. Alors $f(x) = \frac{1}{\sigma}\varphi\left(\frac{x-\mu}{\sigma}\right)$, avec

$$\varphi(x) = \frac{1}{\sqrt{2\pi}}e^{-\frac{x^2}{2}}, \;\; f''(x) = \frac{1}{\sigma^3}\varphi''\left(\frac{x-\mu}{\sigma}\right) \quad et \quad \varphi''(x) = \frac{1}{\sqrt{2\pi}}(x^2-1)e^{-\frac{x^2}{2}}$$

La quantité inconnue $R(f'')$ s'écrit alors

$$\begin{aligned} R(f'') &= \int f^{''}(x)^2 dx \\ &= \frac{1}{\sigma^6} \int \left\{ \varphi'' \left(\frac{x-\mu}{\sigma} \right) \right\}^2 dx = \frac{1}{\sigma^5} \int \left\{ \varphi^{''}(\upsilon) \right\}^2 d\upsilon \end{aligned}$$

Nous avons :

$$\begin{aligned} \varphi(\upsilon) = \frac{1}{\sqrt{2\pi}} e^{-\frac{\upsilon^2}{2}} &\Rightarrow \varphi'(\upsilon) = -\frac{\upsilon}{\sqrt{2\pi}} e^{-\frac{\upsilon^2}{2}} \\ &\Rightarrow \varphi''(\upsilon) = \frac{1}{\sqrt{2\pi}} (\upsilon^2 - 1) e^{-\frac{\upsilon^2}{2}}. \end{aligned}$$

Alors,

$$\begin{aligned} R(f'') &= \frac{1}{\sigma^5} \int \left\{ \frac{1}{\sqrt{2\pi}} (\upsilon^2 - 1) e^{-\frac{\upsilon^2}{2}} \right\}^2 d\upsilon \\ &= \frac{1}{\sigma^5} \frac{1}{\sqrt{2\pi}} \left\{ \int \upsilon^4 e^{-\upsilon^2} d\upsilon - 2 \int \upsilon^2 e^{-\upsilon^2} d\upsilon + \int e^{-\upsilon^2} d\upsilon \right\} \\ &= \frac{1}{\sigma^5} \frac{1}{\sqrt{2\pi}} \left\{ -\frac{1}{2} \int \upsilon^2 e^{-\upsilon^2} d\upsilon + \int e^{-\upsilon^2} d\upsilon \right\} \\ &= \frac{1}{\sigma^5} \frac{1}{\sqrt{2\pi}} \left\{ -\frac{1}{2} \int \frac{\mu^2}{2} e^{-\frac{\mu^2}{2}} \frac{1}{\sqrt{2}} d\mu + \int \frac{1}{\sqrt{2}} e^{-\frac{\mu^2}{2}} d\mu \right\} \qquad \text{avec } \mu = \sqrt{2}\upsilon \\ &= \frac{1}{\sigma^5} \frac{1}{2\pi} \left\{ -\frac{1}{4} \sqrt{\pi} + \sqrt{\pi} \right\} = \frac{1}{\sigma^5} \frac{3}{8\sqrt{\pi}}. \end{aligned}$$

Donc, l'expression du paramètre de lissage optimal devient

$$h^*_{opt} = \left(\frac{8\sqrt{\pi} R(K)}{3\mu_2^2(K)} \right)^{\frac{1}{5}} \hat{\sigma} n^{-\frac{1}{5}}$$

où $\hat{\sigma}^2$ est la variance empirique (estimateur sans biais) de la variance σ^2 de X :

$$\hat{\sigma}^2 = \frac{1}{n-1} \sum_{i=1}^{n} (X_i - X)^2, \qquad \bar{X} = \frac{1}{n} \sum_{i=1}^{n} X_i.$$

Remarque 2.2.5 (cas particulier) *i) Si K est un noyau d'Epanechnikov, alors*

$$h^*_{opt} = 2.34\hat{\sigma}n^{-1/5}.$$

ii) Si K est un noyau gaussien, alors

$$h^*_{opt} = 1.06\hat{\sigma}n^{-1/5}.$$

iii) Si K est un noyau Quartique, i.e., $K(t) = \frac{15}{16}\left(1-t^2\right)^2 I_{(|t|<1)}$, donc

$$h^*_{opt} = 2.78\hat{\sigma}n^{-1/5}.$$

Exercice 2.2.3 *Montrer les trois résultats précédents.*

2.2.5 Comportement Asymptotique

Dans cette section, nous allons étudier le comportement asymptotique de f_n l'estimateur à noyau de la densité f (Normalité asymptotique, convergence presque sure est uniforme).

Théorème 2.2.1 *Supposons que $h := h_n$, est tel que*

$$h \to 0, \quad nh^3 \to 0 \text{ et } nh \to \infty, \text{ quand } n \to \infty.$$

f est dérivable sur un voisinage de $x \in \mathbb{R}$ (où $f(x) > 0$) de dérivées bornées. Alors,

$$\sqrt{2nh}\left(\frac{f_n(x) - f(x)}{\sqrt{f_n(x)}}\right) \to N(0,1) \text{ en loi, quand } n \to \infty.$$

Corollaire 2.2.1 *Sous les mêmes conditions du théorème, l'intervalle de confiance au niveau*

$(1-\alpha)\,\%$ *de* f *est*

$$\left[f_n(x) - z_{1-\alpha/2}\sqrt{\frac{f_n(x)}{2nh}}; f_n(x) - z_{1-\alpha/2}\sqrt{\frac{f_n(x)}{2nh}}\right],$$

où $z_{1-\alpha/2}$ *est le quantile d'ordre* $\left(1-\frac{\alpha}{2}\right)$ *de la loi* $N(0,1)$.

Théorème 2.2.2 *Si* f *est continue sur le voisinage de* x *et si*

$$h \to 0, \quad \frac{\log n}{nh} \to 0 \;\; et \;\; nh \to \infty, \;\; quand \; n \to \infty.$$

Supposons que le noyau K *est d'ordre* 2*, borné, intégrable et à support compact. Alors,*

$$f_n(x) \to f(x), \quad Ps, \;\; quand \; n \to \infty.$$

Si de plus, f *est deux fois continuement dérivable. Alors,*

$$f_n(x) - f(x) = O\left(h^2\right) + O\left(\frac{\log n}{nh}\right).$$

Théorème 2.2.3 *Sous les mêmes conditions du théorème précidente. Supposons de plus que* Ω *est un compact de* $\mathbb{R}$ *sur le quelle* f *est deux fois continuement dérivable et que le noyau* K *est Lipschitzien sur* Ω :

$$\exists c < \infty, \; \forall x, y \in \Omega : |K(x) - K(y)| \leq c\,|x-y|.$$

Alors,

$$\sup_{x\in\Omega} |f_n(x) - f(x)| = O\left(h^2\right) + O\left(\frac{\log n}{nh}\right), \quad Ps.$$