
STAT2—Introduction à la statistique non paramétrique

Mathieu Ribatet

École Centrale de Nantes

Quelques références bibliographiques

- [1] B. W. Silverman. *Density Estimation for Statistics and Data Analysis*. Chapman & Hall, 1996.
- [2] A. B. Tsybakov. *Introduction to Nonparametric Estimation*. Springer Series in Statistics. Springer, 2009.

1. Introduction

2. Estimation de la densité

2. Régression non paramétrique
(en bien trop rapide)

1. Introduction

Motivations

- Le choix d'une loi de probabilité spécifique pourra toujours être sujet à débat

All models are wrong, but some are useful

[George Cox]

Motivations

- Le choix d'une loi de probabilité spécifique pourra toujours être sujet à débat

All models are wrong, but some are useful [George Cox]

- Pouvoir se détacher d'une loi paraît alors séduisant :
 - plus de modèle donc plus d'erreur¹
 - test d'hypothèse plus générique
 - permettra de vérifier la qualité d'un modèle paramétrique

¹personne n'y croit non ?

Exemples : paramétrique \rightarrow non paramétrique

Exemple 1. On dispose d'un n -échantillon iid X_1, \dots, X_n . On souhaite estimer la densité de probabilité associée au modèle Gaussien

$$\left\{ f(x; \theta) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x - \mu)^2}{2\sigma^2}\right) : x \in \mathbb{R}, (\mu, \sigma) \in \mathbb{R} \times (0, \infty) \right\},$$

de sorte que l'estimation se résume à estimer μ et σ .

En non paramétrique on pourra s'intéresser au problème de l'estimation d'une fonction f inconnue vivant dans l'espace

$$\{f : \text{densité de probabilité continue et Lipschitz}\}.$$

Exemples : paramétrique \rightarrow non paramétrique

Exemple 1. On dispose d'un n -échantillon iid $(X_1, Y_1), \dots, (X_n, Y_n)$. On souhaite régresser Y en fonction de X par le modèle linéaire

$$Y = \beta_0 + X^\top \boldsymbol{\beta} + \varepsilon, \quad \varepsilon \sim N(0, \sigma^2),$$

de sorte que l'estimation se résume à estimer $\beta_0, \boldsymbol{\beta}$ et σ^2 .
En non paramétrique on pourra s'intéresser au problème

$$Y = f(X) + \varepsilon, \quad \mathbb{E}(\varepsilon) = 0,$$

où f est une fonction inconnue vivant dans l'espace

$$\{f : \text{fonction de classe } \mathcal{C}^1\}.$$

1. Introduction

2. Estimation de la densité

2. Régression non paramétrique
(en bien trop rapide)

2. Estimation de la densité

Objectif

- On dispose d'un n -échantillon iid X_1, \dots, X_n de fonction de répartition F inconnue mais admettant une densité de probabilité, i.e.,

$$F(x) = \int_{-\infty}^x f(u)du.$$

Objectif

- On dispose d'un n -échantillon iid X_1, \dots, X_n de fonction de répartition F inconnue mais admettant une densité de probabilité, i.e.,

$$F(x) = \int_{-\infty}^x f(u) du.$$

- On désire estimer la densité f . Construisons ensemble un estimateur...

Objectif

- On dispose d'un n -échantillon iid X_1, \dots, X_n de fonction de répartition F inconnue mais admettant une densité de probabilité, i.e.,

$$F(x) = \int_{-\infty}^x f(u) du.$$

- On désire estimer la densité f . Construisons ensemble un estimateur...
- Par la loi des grands nombres, pour tout $x \in \mathbb{R}$,

$$\hat{F}_n(x) = \frac{1}{n} \sum_{i=1}^n 1_{\{X_i \leq x\}} \xrightarrow{\text{ps}} F(x), \quad n \rightarrow \infty.$$

- Les différences finies (symétrisées) suggèrent d'estimer $f(x) = F'(x)$ par

$$\tilde{f}_{n,h}(x) = \frac{\hat{F}_n(x+h) - \hat{F}_n(x-h)}{2h} = \frac{1}{2nh} \sum_{i=1}^n 1_{\{x-h \leq X_i \leq x+h\}}.$$

L'estimateur de Parzen–Rosenblatt de la densité

Définition 1. L'estimateur de [Parzen–Rosenblatt](#) de la densité de probabilité f est donné par

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

où la fonction $K(\cdot)$ est un [noyau](#)² (cf. plus loin) et le paramètre $h > 0$ la [fenêtre](#).

L'estimateur de Parzen–Rosenblatt de la densité

Définition 1. L'estimateur de **Parzen–Rosenblatt** de la densité de probabilité f est donné par

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

où la fonction $K(\cdot)$ est un **noyau**² (cf. plus loin) et le paramètre $h > 0$ la **fenêtre**.

 L'estimateur que nous avons construit ensemble est un cas particulier de Parzen–Rosenblatt avec

$$K(u) = \frac{1}{2} 1_{\{-1 \leq u \leq 1\}}.$$

²Je vous mens. . . cf. plus loin

Définition 2. Une fonction $K: \mathbb{R} \rightarrow \mathbb{R}_+$ est appelée **noyau** lorsque

- $\int_{\mathbb{R}} K(u)du = 1$;
- $K(u) = K(-u)$ pour tout $u \in \mathbb{R}$.

Noyau

Définition 2. Une fonction $K: \mathbb{R} \rightarrow \mathbb{R}_+$ est appelée **noyau** lorsque

- $\int_{\mathbb{R}} K(u) du = 1$;
- $K(u) = K(-u)$ pour tout $u \in \mathbb{R}$.

Table 1: Quelques exemples de famille de noyaux couramment utilisés.

Nom	Expression
Quadratique	$K(u) = \frac{15}{16}(1 - u^2)^2 1_{\{-1 \leq u \leq 1\}}$
Rectangulaire	$K(u) = \frac{1}{2} 1_{\{-1 \leq u \leq 1\}}$
Triangulaire	$K(u) = (1 - u) 1_{\{-1 \leq u \leq 1\}}$
Epanechnikov	$K(u) = \frac{3}{4}(1 - u^2) 1_{\{-1 \leq u \leq 1\}}$
Gaussien	$K(u) = \frac{1}{\sqrt{2\pi}} \exp(-u^2/2)$

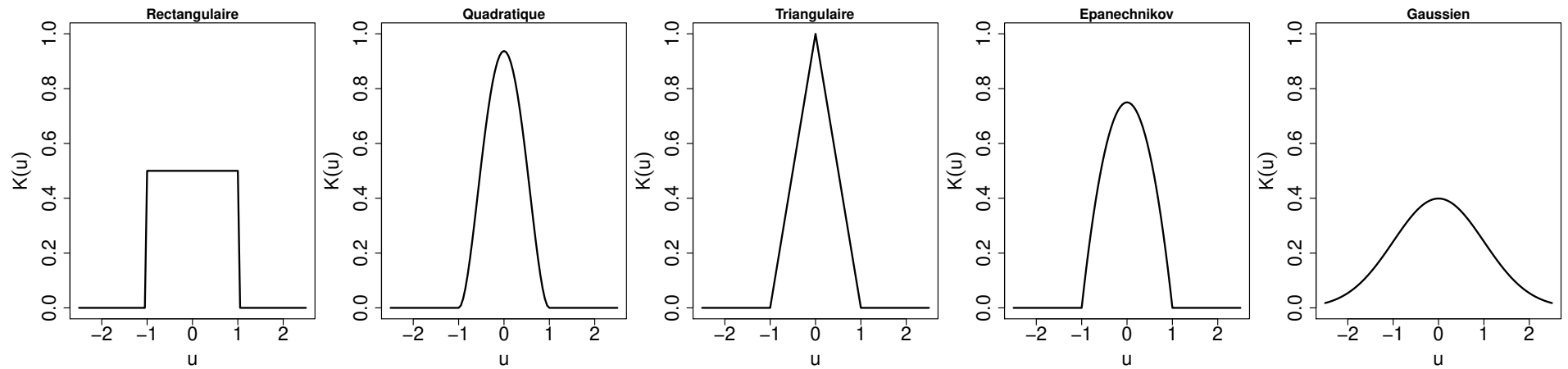


Figure 1: *Graphe de quelques noyaux fréquemment utilisés.*

Une somme de contributions

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

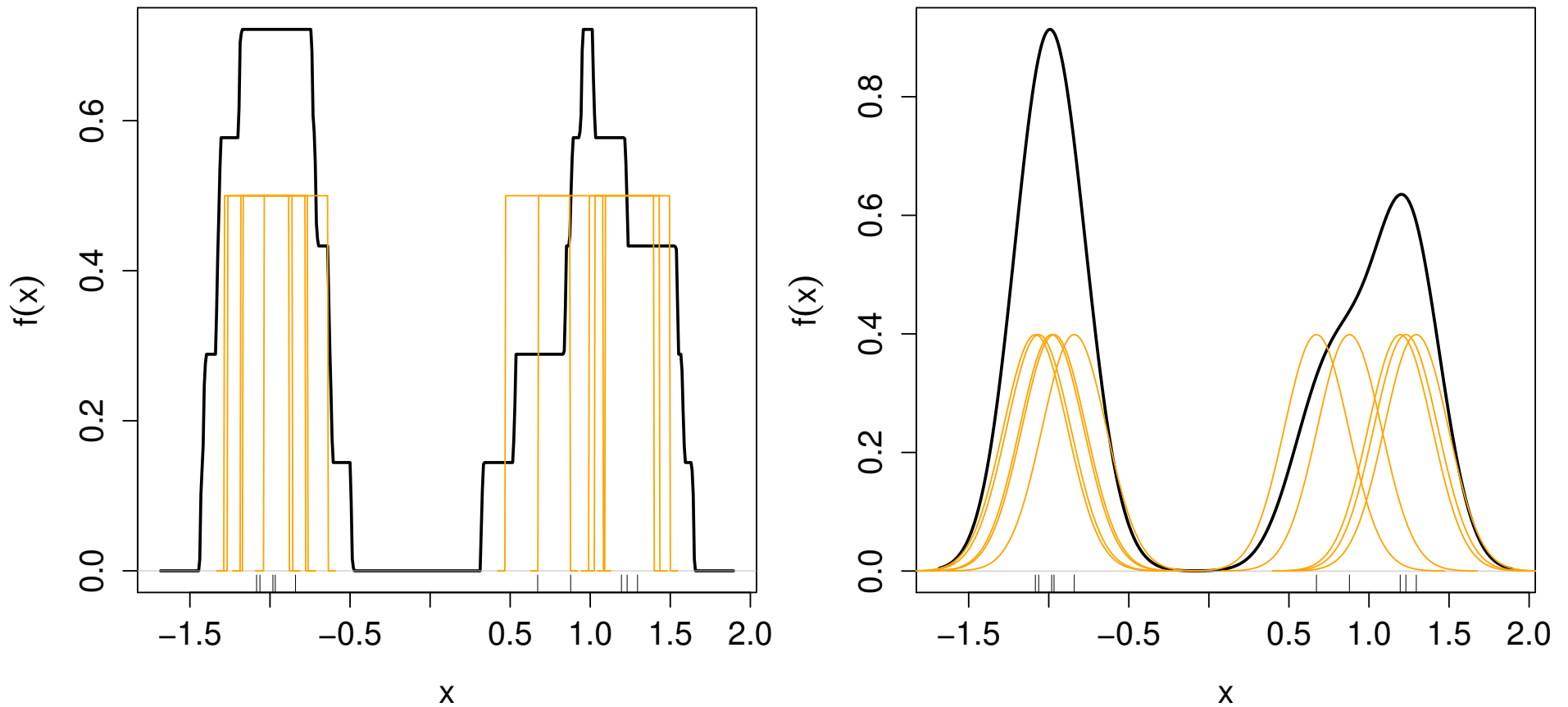


Figure 2: *Illustration graphique de l'estimateur de Parzen-Rosenblatt.*

Impact du noyau

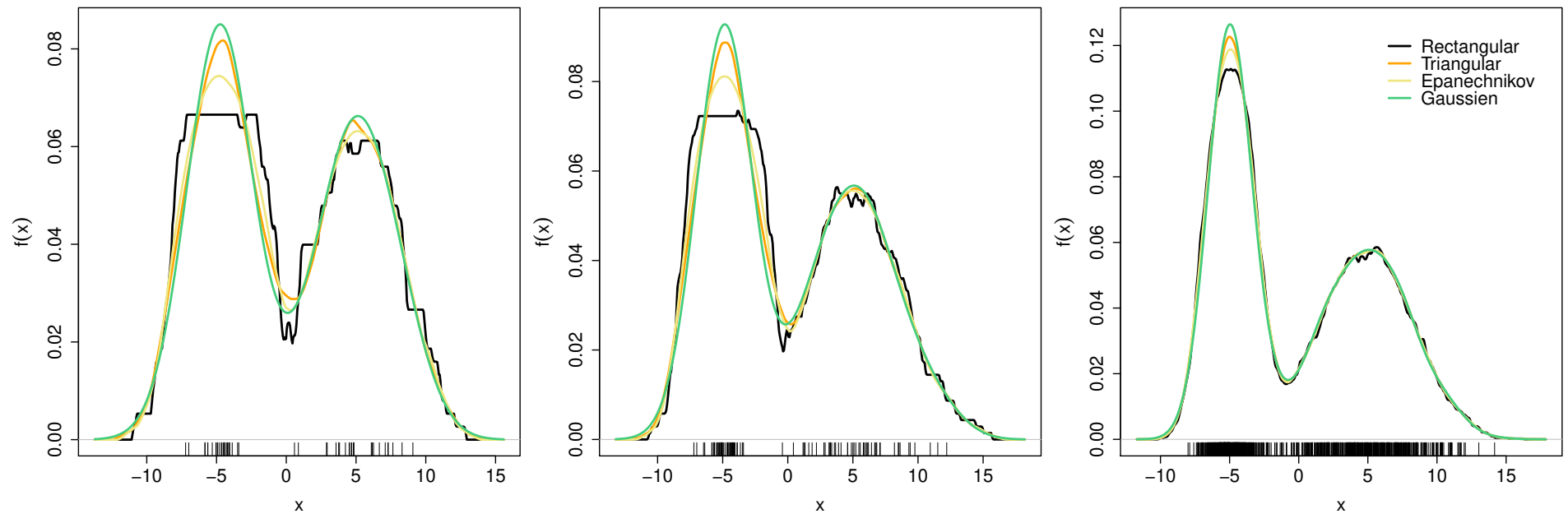


Figure 3: *Évolution de l'impact du choix du noyau sur l'estimation de la densité. De gauche à droite : $n = 50, 100, 1000$.*

Impact du noyau

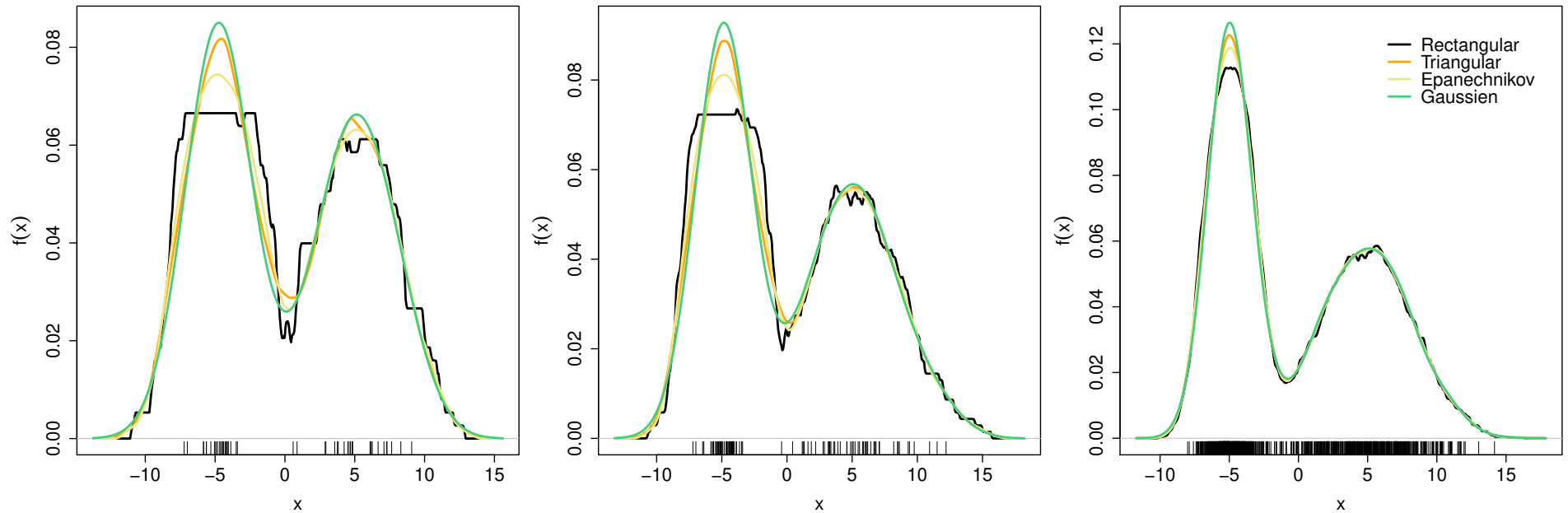


Figure 3: *Évolution de l'impact du choix du noyau sur l'estimation de la densité. De gauche à droite : $n = 50, 100, 1000$.*

👉 Il semblerait que le choix du noyau ait de moins en moins de conséquence lorsque $n \rightarrow \infty$.

Impact de la fenêtre

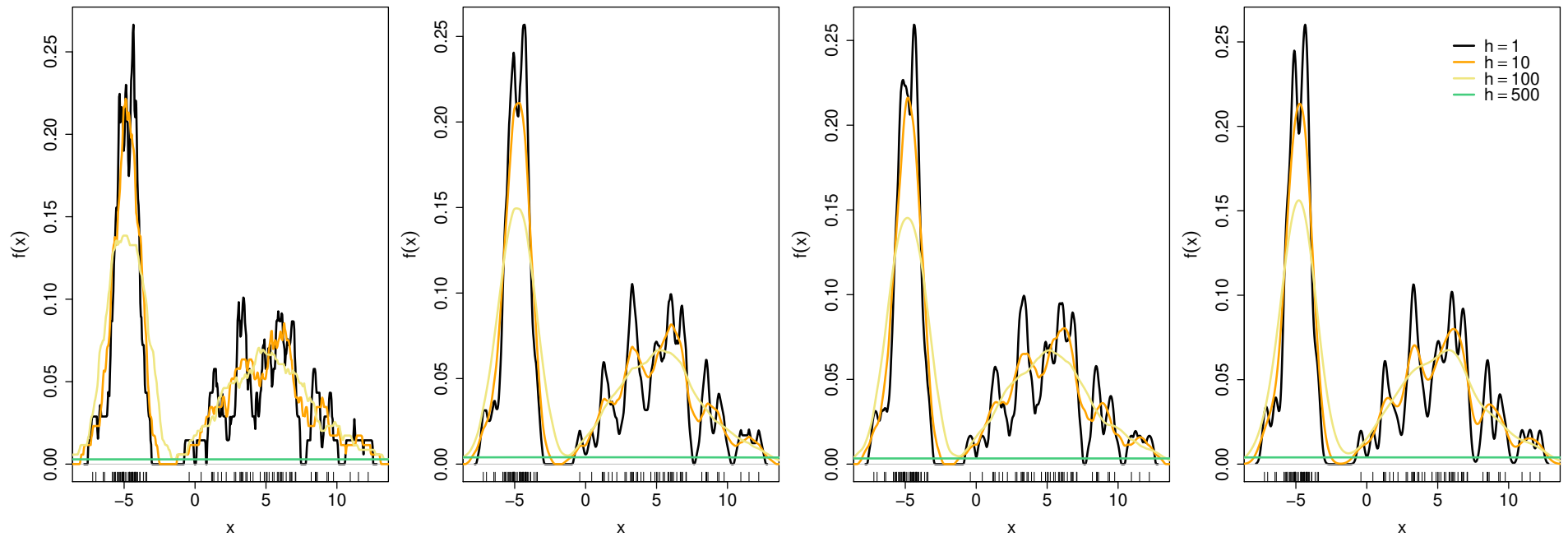


Figure 4: Evolution de l'impact du choix de la fenêtre sur l'estimation de la densité— $n = 100$. De gauche à droite : noyau rectangulaire, triangulaire, Epanechnikov et Gaussien.

Impact de la fenêtre

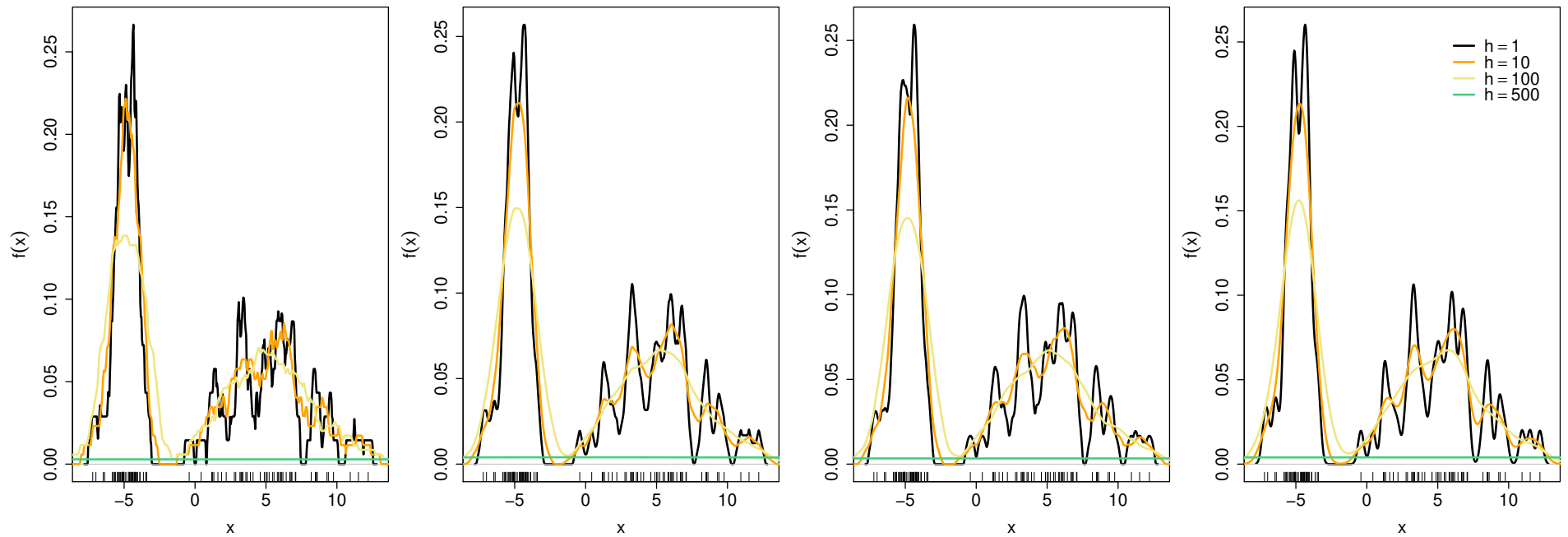


Figure 4: Evolution de l'impact du choix de la fenêtre sur l'estimation de la densité— $n = 100$. De gauche à droite : noyau rectangulaire, triangulaire, Epanechnikov et Gaussien.

👉 Il semblerait que le choix de la fenêtre soit bien plus important que le choix du noyau. Cela dit \hat{f} hérite de la régularité de K .

Biais de $\hat{f}_h(x)$

Proposition 1. *Si la “vraie” densité f est au minimum \mathcal{C}^2 , on a pour tout $x \in \mathbb{R}$*

$$\text{Biais}\{\hat{f}_h(x)\} = \frac{h^2}{2} f''(x) \mu_2(K) + o(h^2), \quad h \rightarrow 0,$$

où $\mu_2(K) = \int u^2 K(u) du$.

Biais de $\hat{f}_h(x)$

Proposition 1. *Si la “vraie” densité f est au minimum \mathcal{C}^2 , on a pour tout $x \in \mathbb{R}$*

$$\text{Biais}\{\hat{f}_h(x)\} = \frac{h^2}{2} f''(x) \mu_2(K) + o(h^2), \quad h \rightarrow 0,$$

où $\mu_2(K) = \int u^2 K(u) du$.

- ☞ Le biais décroît en $h^2 \Rightarrow$ suggère de prendre h petit ;
- ☞ Le biais dépend de la courbure de f , i.e., $f''(x) \Rightarrow$ on “lissera les pics et les vallées”

Lissage des pics et des vallées

$$\text{Biais}\{\hat{f}_h(x)\} = \frac{h^2}{2} f''(x) \mu_2(K) + o(h^2), \quad h \rightarrow 0,$$

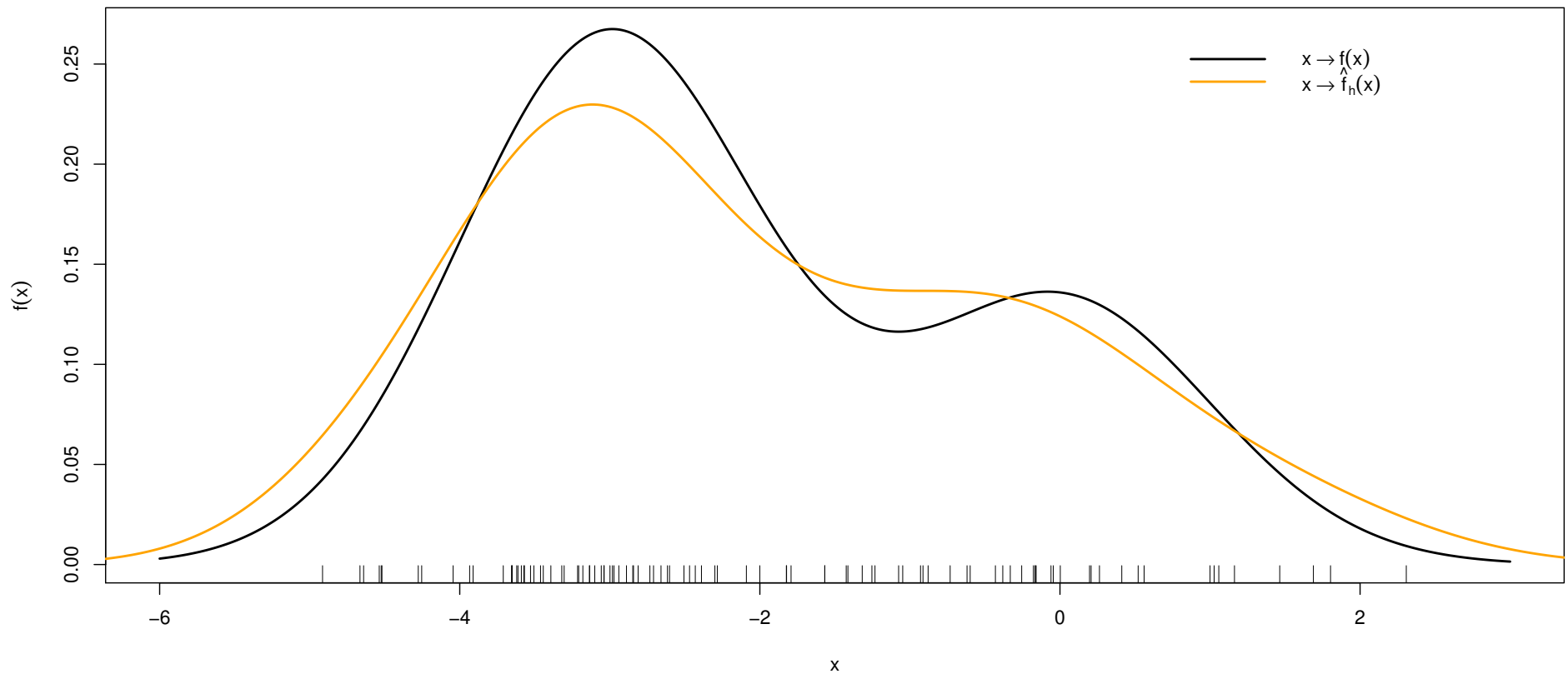


Figure 5: *Illustration du lissage des pics et des vallées.*

Variance de $\hat{f}_h(x)$

Proposition 2. *Si la “vraie” densité f est au minimum \mathcal{C}^1 et que le noyau $K \in L^2$, on a pour tout $x \in \mathbb{R}$*

$$\text{Var}\{\hat{f}_h(x)\} = \frac{1}{nh} f(x) \int K(u)^2 du + o\left(\frac{1}{nh}\right), \quad nh \rightarrow \infty.$$

Variance de $\hat{f}_h(x)$

Proposition 2. *Si la “vraie” densité f est au minimum \mathcal{C}^1 et que le noyau $K \in L^2$, on a pour tout $x \in \mathbb{R}$*

$$\text{Var}\{\hat{f}_h(x)\} = \frac{1}{nh} f(x) \int K(u)^2 du + o\left(\frac{1}{nh}\right), \quad nh \rightarrow \infty.$$

- La variance décroît en $nh \Rightarrow$ suggère h grand
- La variance croît avec $\int K(u)^2 du \Rightarrow$ suggère de prendre des noyaux plutôt réguliers

Compromis biais // variance

- Le biais croît en $h^2 \Rightarrow$ suggère de prendre h petit
- La variance décroît en $nh \Rightarrow$ suggère h grand

Compromis biais // variance

- Le biais croît en $h^2 \Rightarrow$ suggère de prendre h petit
- La variance décroît en $nh \Rightarrow$ suggère h grand

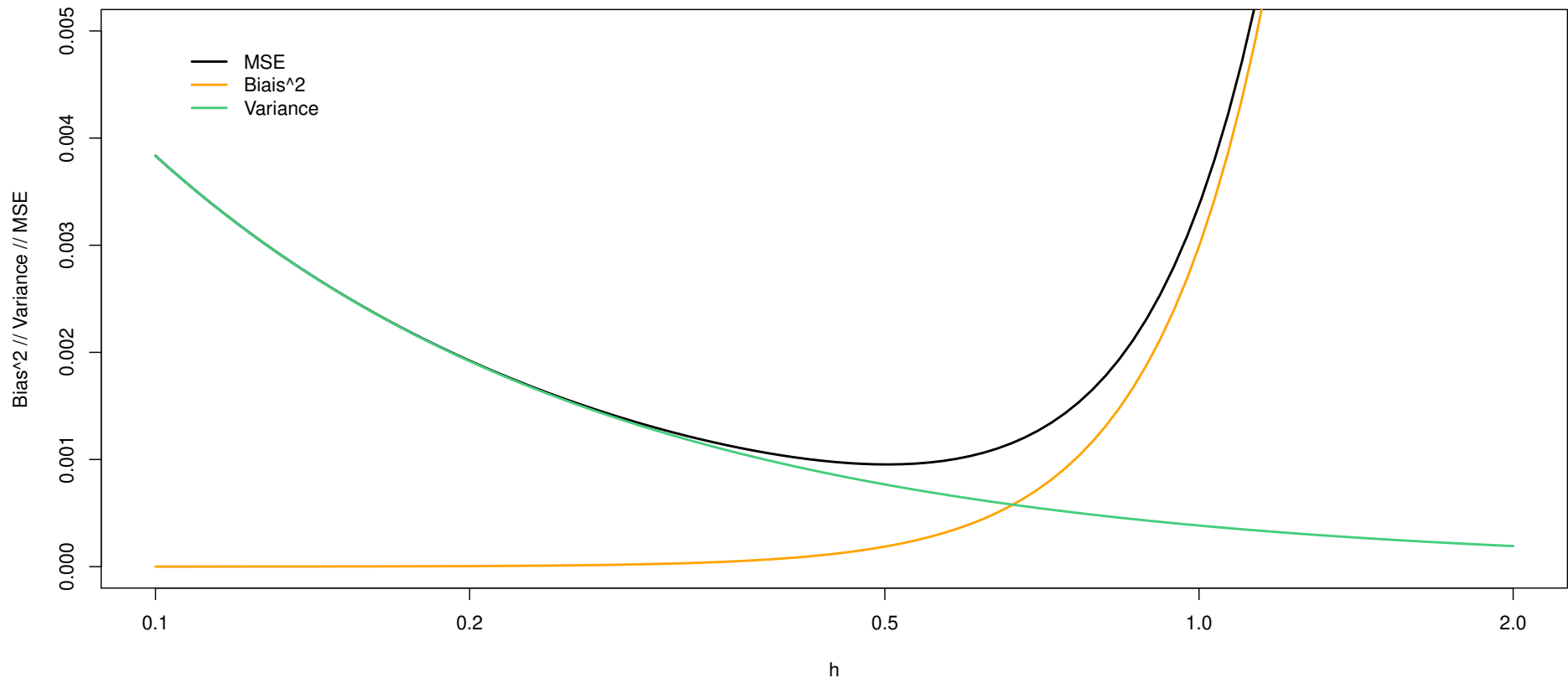


Figure 6: Illustration du compromis biais // variance.

Choisir un h optimal ?

- Pour évaluer la qualité d'un estimateur on s'intéressera souvent à l'erreur quadratique moyenne

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

Choisir un h optimal ?

- Pour évaluer la qualité d'un estimateur on s'intéressera souvent à l'erreur quadratique moyenne

$$MSE\{\hat{f}_h(\boldsymbol{x})\} = \mathbb{E} \left[\{\hat{f}_h(\boldsymbol{x}) - f(\boldsymbol{x})\}^2 \right] = \text{Biais}\{\hat{f}_h(\boldsymbol{x})\}^2 + \text{Var}\{\hat{f}_h(\boldsymbol{x})\}.$$

- Cependant c'est une mesure locale

Choisir un h optimal ?

- Pour évaluer la qualité d'un estimateur on s'intéressera souvent à l'erreur quadratique moyenne

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

- Cependant c'est une mesure locale et on préférera alors l'erreur quadratique moyenne intégrée

$$MISE(\hat{f}_h) = \int MSE\{\hat{f}_h(x)\} dx$$

Choisir un h optimal ?

- Pour évaluer la qualité d'un estimateur on s'intéressera souvent à l'erreur quadratique moyenne

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

- Cependant c'est une mesure locale et on préférera alors l'erreur quadratique moyenne intégrée

$$MISE(\hat{f}_h) = \int MSE\{\hat{f}_h(x)\} dx$$

 Il paraît séduisant de choisir h minimisant $h \mapsto MISE(\hat{f}_h)$!

Une formule (presque) inutile

Proposition 3. *La fenêtre optimale minimisant le critère MISE (approché) est donnée par*

$$h_* = \mu_2(K)^{-2/5} n^{-1/5} \left(\int K(u)^2 du \right)^{1/5} \left(\int f''(x)^2 dx \right)^{-1/5},$$

pour laquelle $MISE(\hat{f}_h)$ vaut (approximativement)

$$\frac{5}{4} C(K) \left(\int f''(x)^2 dx \right)^{1/5} n^{-4/5}, \quad C(K) = \mu_2(K)^{2/5} \left(\int K(u)^2 du \right)^{4/5}.$$

Une formule (presque) inutile

Proposition 3. *La fenêtre optimale minimisant le critère MISE (approché) est donnée par*

$$h_* = \mu_2(K)^{-2/5} n^{-1/5} \left(\int K(u)^2 du \right)^{1/5} \left(\int f''(x)^2 dx \right)^{-1/5},$$

pour laquelle $MISE(\hat{f}_h)$ vaut (approximativement)

$$\frac{5}{4} C(K) \left(\int f''(x)^2 dx \right)^{1/5} n^{-4/5}, \quad C(K) = \mu_2(K)^{2/5} \left(\int K(u)^2 du \right)^{4/5}.$$

□ C'est décevant car h_* dépend de f'' inconnue !

Une formule (presque) inutile

Proposition 3. *La fenêtre optimale minimisant le critère MISE (approché) est donnée par*

$$h_* = \mu_2(K)^{-2/5} n^{-1/5} \left(\int K(u)^2 du \right)^{1/5} \left(\int f''(x)^2 dx \right)^{-1/5},$$

pour laquelle $MISE(\hat{f}_h)$ vaut (approximativement)

$$\frac{5}{4} C(K) \left(\int f''(x)^2 dx \right)^{1/5} n^{-4/5}, \quad C(K) = \mu_2(K)^{2/5} \left(\int K(u)^2 du \right)^{4/5}.$$

- C'est décevant car h_* dépend de f'' inconnue !
- Néanmoins...

Une formule (presque) inutile

Proposition 3. *La fenêtre optimale minimisant le critère MISE (approché) est donnée par*

$$h_* = \mu_2(K)^{-2/5} n^{-1/5} \left(\int K(u)^2 du \right)^{1/5} \left(\int f''(x)^2 dx \right)^{-1/5},$$

pour laquelle $MISE(\hat{f}_h)$ vaut (approximativement)

$$\frac{5}{4} C(K) \left(\int f''(x)^2 dx \right)^{1/5} n^{-4/5}, \quad C(K) = \mu_2(K)^{2/5} \left(\int K(u)^2 du \right)^{4/5}.$$

- C'est décevant car h_* dépend de f'' inconnue !
- Néanmoins... elle nous apprend que :
 - $h_* \downarrow 0$ as $n \rightarrow \infty$;
 - f irrégulière, i.e., $\int f''$ grand $\Rightarrow h$ devrait être petit.

Existe-t-il un noyau optimal ?

$$MISE \approx \frac{5}{4}C(K) \left(\int f''(x)^2 dx \right)^{1/5} n^{-4/5}.$$

- Trouver le noyau optimal revient donc à minimiser $C(K)$ sous la contrainte $\int K(u)du = \int u^2 K(u)du = 1$ et $K \geq 0$.
- Hodges and Lehman [1956] ont montré que le noyau optimal est alors de la forme

$$K_e(u) = \frac{3}{4}(1 - u^2)1_{\{-1 \leq u \leq 1\}}, \quad [\text{Noyau d'Epanechnikov}].$$

Existe-t-il un noyau optimal ?

$$MISE \approx \frac{5}{4} C(K) \left(\int f''(x)^2 dx \right)^{1/5} n^{-4/5}.$$

- Trouver le noyau optimal revient donc à minimiser $C(K)$ sous la contrainte $\int K(u) du = \int u^2 K(u) du = 1$ et $K \geq 0$.
- Hodges and Lehman [1956] ont montré que le noyau optimal est alors de la forme

$$K_e(u) = \frac{3}{4} (1 - u^2) 1_{\{-1 \leq u \leq 1\}}, \quad [\text{Noyau d'Epanechnikov}].$$

- On peut “s’amuser” à comparer l’optimalité d’un noyau par rapport au noyau de référence Epanechnikov, i.e., en calculant

$$\text{Efficacité} = \left(\frac{C(K)}{C(K_e)} \right)^{5/4} = \frac{\int K(u)^2 du}{\int K_e(u)^2 du}.$$

Oui mais... pas tant que cela

Table 2: *Efficacité (au sens $C(K)/C(K_e)$) de quelques familles de noyaux couramment utilisés.*

Nom	Expression	Efficacité
Epanechnikov	$K(u) = \frac{3}{4}(1 - u^2)1_{\{-1 \leq u \leq 1\}}$	1
Quadratique	$K(u) = \frac{15}{16}(1 - u^2)^2 1_{\{-1 \leq u \leq 1\}}$	≈ 0.9939
Triangulaire	$K(u) = (1 - u)1_{\{-1 \leq u \leq 1\}}$	≈ 0.9859
Gaussien	$K(u) = \frac{1}{\sqrt{2\pi}} \exp(-u^2/2)$	≈ 0.9512
Rectangulaire	$K(u) = \frac{1}{2}1_{\{-1 \leq u \leq 1\}}$	≈ 0.9295

Oui mais... pas tant que cela

Table 2: *Efficacité (au sens $C(K)/C(K_e)$) de quelques familles de noyaux couramment utilisés.*

Nom	Expression	Efficacité
Epanechnikov	$K(u) = \frac{3}{4}(1 - u^2)1_{\{-1 \leq u \leq 1\}}$	1
Quadratique	$K(u) = \frac{15}{16}(1 - u^2)^2 1_{\{-1 \leq u \leq 1\}}$	≈ 0.9939
Triangulaire	$K(u) = (1 - u)1_{\{-1 \leq u \leq 1\}}$	≈ 0.9859
Gaussien	$K(u) = \frac{1}{\sqrt{2\pi}} \exp(-u^2/2)$	≈ 0.9512
Rectangulaire	$K(u) = \frac{1}{2}1_{\{-1 \leq u \leq 1\}}$	≈ 0.9295

👉 On comprend mieux pourquoi le choix de la famille du noyau n'est pas si critique. En pratique on choisira soit l'optimal soit un noyau avec la régularité désirée, i.e., $\mathcal{C}^1, \mathcal{C}^2, \dots$

Fenêtre optimale : La règle empirique de Silverman

$$h_* = \mu_2(K)^{-2/5} n^{-1/5} \left(\int K(u)^2 du \right)^{1/5} \left(\int f''(x)^2 dx \right)^{-1/5},$$

- Principe : Comme f inconnue on fera les calculs pour $f \sim N(\mu, \sigma^2)$ et K noyau Gaussien
- Cela nous donnera une “**indication**” de la fenêtre optimale.
- Cette dernière est donnée par

$$h_{\text{Silverman}} = \left(\frac{4}{3n} \right)^{1/5} \hat{\sigma} \approx 1.06 \hat{\sigma} n^{-1/5}, \quad \hat{\sigma} = \sqrt{\frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X})^2}.$$

Fenêtre optimale : La règle empirique de Silverman

$$h_* = \mu_2(K)^{-2/5} n^{-1/5} \left(\int K(u)^2 du \right)^{1/5} \left(\int f''(x)^2 dx \right)^{-1/5},$$

- Principe : Comme f inconnue on fera les calculs pour $f \sim N(\mu, \sigma^2)$ et K noyau Gaussien
- Cela nous donnera une “indication” de la fenêtre optimale.
- Cette dernière est donnée par

$$h_{\text{Silverman}} = \left(\frac{4}{3n} \right)^{1/5} \hat{\sigma} \approx 1.06 \hat{\sigma} n^{-1/5}, \quad \hat{\sigma} = \sqrt{\frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X})^2}.$$

👉 Cette fenêtre sera adaptée lorsque la vraie densité f est assez proche d'une loi normale, i.e., unimodale, à peu près symétrique, ayant des queues de distributions légères.

Fenêtre optimale : Validation croisée

$$\arg \min_{h>0} MISE(\hat{f}_h) = \arg \min_{h>0} \mathbb{E} \left[\int \left\{ \hat{f}_h(x) - f(x) \right\}^2 dx \right]$$

- Nous avons vu que la solution dépend malheureusement de f ...
- Pourquoi ne pas minimiser un problème indépendant de f qui serait un **estimateur** du premier ?
- Travaillons un peu sur cette expression à minimiser...

$$\begin{aligned}
J(h) &= \mathbb{E} \left[\int \left\{ \hat{f}_h(x) - f(x) \right\}^2 dx \right] \\
&= \mathbb{E} \left[\int \hat{f}_h(x)^2 dx - 2 \int \hat{f}_h(x) f(x) dx \right] + \underbrace{\int f(x)^2 dx}_{\text{indépendant de } h, \text{ on ignore}} \\
&= \mathbb{E} \left[\int \hat{f}_h(x)^2 dx \right] - 2 \mathbb{E} \left[\mathbb{E}_X \left\{ \hat{f}_h(X) \right\} \right]
\end{aligned}$$

$$\begin{aligned}
J(h) &= \mathbb{E} \left[\int \left\{ \hat{f}_h(x) - f(x) \right\}^2 dx \right] \\
&= \mathbb{E} \left[\int \hat{f}_h(x)^2 dx - 2 \int \hat{f}_h(x) f(x) dx \right] + \underbrace{\int f(x)^2 dx}_{\text{indépendant de } h, \text{ on ignore}} \\
&= \mathbb{E} \left[\int \hat{f}_h(x)^2 dx \right] - 2 \mathbb{E} \left[\mathbb{E}_X \left\{ \hat{f}_h(X) \right\} \right]
\end{aligned}$$

👉 On estimera $J(h)$ (sans le terme indépendant de h) par

$$\tilde{J}(x) = \int \hat{f}_h(x)^2 dx - \frac{2}{n} \sum_{i=1}^n \hat{f}_{h,-i}(X_i), \quad \hat{f}_{h,-i}(x) = \frac{1}{(n-1)h} \sum_{\substack{j=1 \\ j \neq i}}^n K \left(\frac{X_j - x}{h} \right).$$

Validation croisée : Leave one out

- On vient de définir le critère de validation croisée “leave one out”

$$CV(h) = \int \hat{f}_h(x)^2 dx - \frac{2}{n(n-1)h} \sum_{i=1}^n \sum_{\substack{j=1 \\ j \neq i}}^n K\left(\frac{X_j - X_i}{h}\right).$$

- On choisira donc pour fenêtre optimale (selon ce critère)

$$h_{CV} = \arg \min_{h>0} CV(h).$$

Oui mais... pas vraiment en fait ;-) (Tsybakov, Section 1.2.4)

- Nous avons vu que le noyau d'Epanechnikov était optimal...

Oui mais... pas vraiment en fait ;-) (Tsybakov, Section 1.2.4)

- Nous avons vu que le noyau d'Epanechnikov était optimal...
- C'est vrai sous la contrainte de **positivité** du noyau K .
- Cela dit on peut tout à fait utiliser des noyaux admettant des valeurs négatives et poser l'estimateur

$$\hat{f}_h(x) = \frac{1}{nh} \max \left\{ \sum_{i=1}^n K \left(\frac{X_i - x}{h} \right), 0 \right\}.$$


- Alors le noyau d'Epanechnikov n'est alors plus le noyau optimal.

Oui mais... pas vraiment en fait ;-) (Tsybakov, Section 1.2.4)

- Nous avons vu que le noyau d'Epanechnikov était optimal...
- C'est vrai sous la contrainte de **positivité** du noyau K .
- Cela dit on peut tout à fait utiliser des noyaux admettant des valeurs négatives et poser l'estimateur

$$\hat{f}_h(x) = \frac{1}{nh} \max \left\{ \sum_{i=1}^n K \left(\frac{X_i - x}{h} \right), 0 \right\}.$$

- Alors le noyau d'Epanechnikov n'est alors plus le noyau optimal.

 On choisira alors le noyau selon ses “convictions” quant à la régularité de la densité f .

1. Introduction

2. Estimation de la densité

2. Régression non paramétrique
(en bien trop rapide)

2. Régression non paramétrique (en bien trop rapide)

Cadre de travail

- Soit (X, Y) un couple de variables aléatoires de densité $f(x, y)$ et tel que $\mathbb{E}(|Y|) < \infty$.
- Nous cherchons à estimer la fonction

$$m: \mathbb{R}^p \longrightarrow \mathbb{R}$$
$$x \longmapsto m(x) = \mathbb{E}(Y \mid X = x).$$

- Contrairement aux modèles linéaires, nous ne supposerons pas que $m(x)$ est de la forme $x^\top \beta$.
- Au contraire nous supposerons que m est de forme inconnue.

Avant de commencer

- Il existe plusieurs possibilités pour cette problématique :
 - polynômes locaux ;
 - spline ;
 - plus proches voisins
- Mais ici, pour être cohérent, nous allons uniquement aborder la **régression non paramétrique par noyaux**.

Construction de l'estimateur

$$m(x) = \mathbb{E}(Y \mid X = x) = \frac{\int y f(x, y) dy}{f_X(x)}, \quad f_X(x) = \int f(x, y) dy$$

- Il nous faut donc estimer deux quantités :
 - $f_X(\cdot)$ mais ça on sait déjà le faire ;-)
 - $f(x, y)$ qui est une densité bivariée.
- Pour $f(x, y)$ on va utiliser l'estimateur suivant

$$\hat{f}_{h_1, h_2}(x, y) = \frac{1}{nh_1 h_2} \sum_{i=1}^n K_{h_1} \left(\frac{X_i - x}{h_1} \right) K_{h_2} \left(\frac{Y_i - y}{h_2} \right).$$

Construction de l'estimateur

$$m(x) = \mathbb{E}(Y \mid X = x) = \frac{\int y f(x, y) dy}{f_X(x)}, \quad f_X(x) = \int f(x, y) dy$$

- Il nous faut donc estimer deux quantités :
 - $f_X(\cdot)$ mais ça on sait déjà le faire ;-)
 - $f(x, y)$ qui est une densité bivariée.
- Pour $f(x, y)$ on va utiliser l'estimateur suivant

$$\hat{f}_{h_1, h_2}(x, y) = \frac{1}{nh_1 h_2} \sum_{i=1}^n K_{h_1} \left(\frac{X_i - x}{h_1} \right) K_{h_2} \left(\frac{Y_i - y}{h_2} \right).$$

 On utilise ici une forme spécifique de noyaux sur \mathbb{R}^2 , le noyau produit.

L'estimateur de Nadaraya–Watson

Définition 3. L'estimateur de [Nadaraya–Watson](#) pour l'espérance conditionnelle $m(x) = \mathbb{E}(Y \mid X = x)$ est donné par

$$\hat{m}_h(x) = \begin{cases} \frac{n^{-1} \sum_{i=1}^n Y_i K\left(\frac{X_i - x}{h}\right)}{n^{-1} \sum_{i=1}^n K\left(\frac{X_i - x}{h}\right)}, & \sum_{i=1}^n K\left(\frac{X_i - x}{h}\right) \neq 0 \\ 0, & \text{sinon.} \end{cases}$$

L'estimateur de Nadaraya–Watson

Définition 3. L'estimateur de [Nadaraya–Watson](#) pour l'espérance conditionnelle $m(x) = \mathbb{E}(Y \mid X = x)$ est donné par

$$\hat{m}_h(x) = \begin{cases} \frac{n^{-1} \sum_{i=1}^n Y_i K\left(\frac{X_i - x}{h}\right)}{n^{-1} \sum_{i=1}^n K\left(\frac{X_i - x}{h}\right)}, & \sum_{i=1}^n K\left(\frac{X_i - x}{h}\right) \neq 0 \\ 0, & \text{sinon.} \end{cases}$$

□ Notons que l'on peut également l'écrire sous la forme

$$\hat{m}_h(x) = \sum_{i=1}^n W_{hi}(x) Y_i, \quad W_{hi}(x) = \frac{K\{(X_i - x)/h\}}{\sum_{j=1}^n K\{(X_j - x)/h\}}$$

L'estimateur de Nadaraya–Watson

Définition 3. L'estimateur de **Nadaraya–Watson** pour l'espérance conditionnelle $m(x) = \mathbb{E}(Y \mid X = x)$ est donné par

$$\hat{m}_h(x) = \begin{cases} \frac{n^{-1} \sum_{i=1}^n Y_i K\left(\frac{X_i - x}{h}\right)}{n^{-1} \sum_{i=1}^n K\left(\frac{X_i - x}{h}\right)}, & \sum_{i=1}^n K\left(\frac{X_i - x}{h}\right) \neq 0 \\ 0, & \text{sinon.} \end{cases}$$

□ Notons que l'on peut également l'écrire sous la forme

$$\hat{m}_h(x) = \sum_{i=1}^n W_{hi}(x) Y_i, \quad W_{hi}(x) = \frac{K\{(X_i - x)/h\}}{\sum_{j=1}^n K\{(X_j - x)/h\}}$$

 L'estimateur de Nadaraya–Watson est donc une moyenne pondérée des Y_i .

Design fixe

- Si l'on se place dans le cas où la densité f_X de X est connue, i.e., design connu, alors on utilisera l'estimateur

$$\tilde{m}_h(x) = \frac{1}{nhf_X(x)} \sum_{i=1}^n Y_i K\left(\frac{X_i - x}{h}\right).$$

- En particulier si $f_X \sim U\{0, \dots, T\}$, alors

$$\tilde{m}_h(x) = \frac{T+1}{nh} \sum_{i=1}^n Y_i K\left(\frac{X_i - x}{h}\right).$$


Design fixe

- Si l'on se place dans le cas où la densité f_X de X est connue, i.e., design connu, alors on utilisera l'estimateur

$$\tilde{m}_h(x) = \frac{1}{nhf_X(x)} \sum_{i=1}^n Y_i K\left(\frac{X_i - x}{h}\right).$$

- En particulier si $f_X \sim U\{0, \dots, T\}$, alors

$$\tilde{m}_h(x) = \frac{T+1}{nh} \sum_{i=1}^n Y_i K\left(\frac{X_i - x}{h}\right).$$

 Ce dernier résultat est également utilisé pour estimer la tendance d'une série temporelle, cf. intro. séries temporelles, même si le design n'est alors plus aléatoire.

Parce que nous n'avons pas assez de temps...

- La régression non paramétrique est un **vaste** domaine des statistiques
- Nous n'avons vu (trop rapidement) qu'une petite partie
- D'autres approches très utilisées existent, parmi lesquelles
 - les splines ;
 - les polynômes locaux ;
 - k -plus proches voisins ;
 - ...