

Cours STAT 2150

"Statistique non paramétrique:

Méthodes de lissage"

Année académique 2008 - 2009

Séance 1

Table de matière du cours

1. Introduction (Fonction de répartition, histogramme, propriétés d'un estimateur)
2. Estimation d'une densité (méthode à noyau)
3. Régression non paramétrique (méthode à noyau)
4. Aspects supplémentaires (traitement des bornes, estimation des dérivées, aspects multivariés)
5. Lissage à méthodes différentes des noyaux
6. Un peu de théorie asymptotique et de l'inférence

Support du cours et Evaluation

- Copies des transparents du cours - en ligne
- Chapitre 3 du syllabus "Statistique non paramétrique" (I. Gijbels) - en ligne
- Extraits des chapitres 2 et 3 du syllabus "Advanced nonparametric statistics" (I. Gijbels) - sur demande
- Littérature anglaise :
 - J. Simonoff (1996): Smoothing Methods in Statistics
 - W. Härdle (1990): Applied Nonparametric Regression
 - W. Härdle (1990): Smoothing Techniques, With Implementations in S
 - B. Silverman (1986): Density estimation for Statistics and Data Analysis
 - M. Wand, R. Jones (1995): Kernel smoothing
 - J. Fan, I. Gijbels (1996): Local Polynomial Modelling and Its Applications
 - R. Eubank (1988): Spline Smoothing and Nonparametric Regression
- 2 séances Travaux sur ordinateur (Assistante: Olga Reznikova)
- Manuel de références pour le logiciel R
- Evaluation: Préparation d'un projet sur ordinateur et interrogation orale

Chapitre 1 : Introduction

Motivations :

1ère motivation : régression (non) paramétrique

Données des observations (Y_i, X_i) , $i = 1, \dots, n$

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$$

Estimation "très précise" des paramètres $\beta_0, \beta_1, \sigma_\varepsilon^2$

mais modèle pas très robuste car étroit ("quel modèle choisir?")

Alternative non paramétrique

$$Y_i = m(X_i) + \varepsilon_i$$

lien "fonctionnelle" entre X et Y , avec certaines conditions de régularité (continuité, monotonie, ...) sur la fonction m

Modèle "large" mais estimation de m moins précise car plus "difficile"

Ici: développement des conditions sur m et des méthodes comment construire des estimateurs de m (par "lissage")

Motivations (suite) :

2ème motivation : densité (non) paramétrique

Données des n observations (i.i.d.) $X_i, \quad i = 1, \dots, n,$

on cherche de modéliser leur densité :

Approche paramétrique : la densité appartient à une classe paramétrique, par exemple normale, gamma, ...

$$f_{\mu, \sigma^2}(x) = \frac{1}{\sqrt{2\pi}} \exp -\frac{(x - \mu)^2}{2\sigma^2}$$

Estimation "très précise" des paramètres μ et σ^2

mais modèle pas très robuste car étroit ("Hypothèse du modèle correcte? Densité paramétrique compatible avec données?")

Alternative d'une densité non paramétrique : $f_X(x)$ est une courbe non négative avec

$$\int f_X(x) dx = 1$$

et certaines conditions de régularité (continuité, symétrie, unimodalité, ...) sur la fonction f_X

Modèle "large" mais estimation de f_X moins précise car plus "difficile"

Ici: développement des conditions sur f_X et des méthodes comment construire des estimateurs de f_X (par "lissage")

Histogramme de densité :

Soient données X_1, \dots, X_n (i.i.d.).

Choisir point d'origine t_0 et longueur de classe $h > 0$.

Définir la k -ème classe par $B_k = [t_k; t_{k+1}[$; $k \in \mathbb{N}$ avec $t_{k+1} = t_k + h$.

Un estimateur de f est donné par

$$\hat{f}_H(x) = \frac{1}{nh} \#\{i : X_i \text{ est dans la classe qui contient } x\}.$$

Soit ν_k le nombre d'observations dans une classe B_k , alors l'estimateur du type histogramme de densité s'écrit

$$\hat{f}_H(x) = \frac{\nu_k}{nh} = \frac{1}{nh} \sum_{i=1}^n I_{[t_k; t_{k+1}[}(X_i) \quad \text{pour } x \in B_k .$$

Remarques : 1) Ceci est un estimateur d'un "lissage" très élémentaire (mais utile). Il s'agit d'une fréquence relative locale selon une fenêtre rectangulaire.

2) L'estimateur dépend de deux paramètres : le point d'origine t_0 et la largeur de classe ("bin width") h .

Etudier leur influence sur l'histogramme (à travers des exemples...)

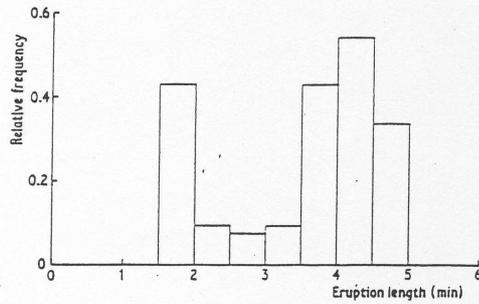
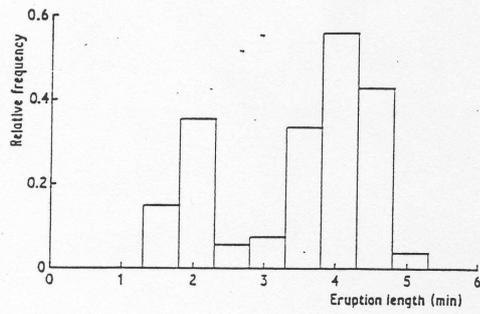


Figure 2.2: Histograms of eruptions lengths of the Old Faithful geyser (from Silverman (1986)).

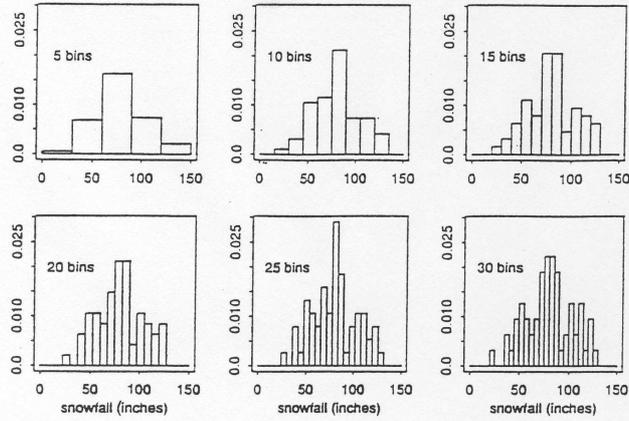


Figure 2.4: Histograms of the Buffalo snowfall data with bin origin $t_0 = 0$, and bin widths of 30, 15, 10, 7.5, 6 and 5 inches (from Scott (1992)).

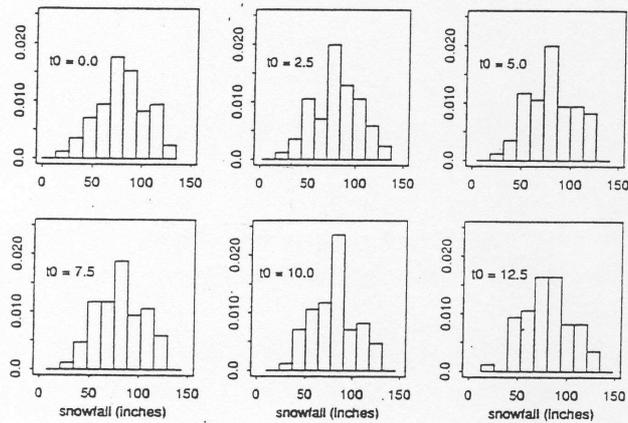


Figure 2.5: Histograms of the Buffalo snowfall data. All have a bin width of 13.5 inches, but different origins t_0 (from Scott (1992)).

L'histogramme motive un estimateur simple "à noyau" :

"Astuce" : Prendre des classes de longueur $2h$ autour de chaque point x : $\hat{f}(x) =$

$$\frac{1}{2h} \frac{\#\{i : x - h < X_i \leq x + h\}}{n} = \frac{1}{2hn} \sum_{i=1}^n I\left\{-1 \leq \frac{x - X_i}{h} < 1\right\}.$$

Clairement, cas particulier d'une construction générale

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

avec une fenêtre ("un noyau") rectangulaire

$$w(y) = 1/2 I\{-1 \leq y < 1\}$$

Avantage : Cet estimateur ne dépend plus du choix d'origine.

Desavantage : Comme histogramme, cet estimateur est une fonction discontinue (en points $X_i \pm h$) et constante par morceaux.

Motivation pour la suite : Remplacer le fenêtre rectangulaire par une fenêtre plus lisse, le noyau.

Avant : Dérivation des propriétés de cet estimateur.

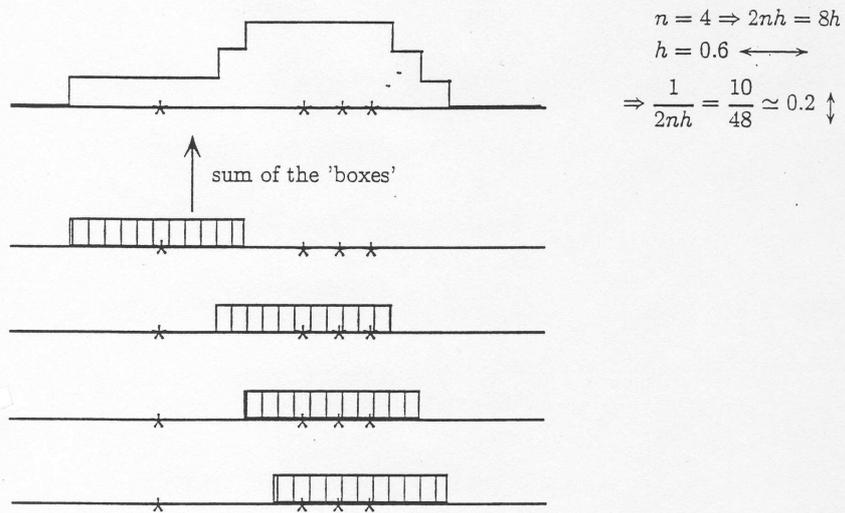


Figure 2.8: Construction of the simple estimator ('sum of boxes').

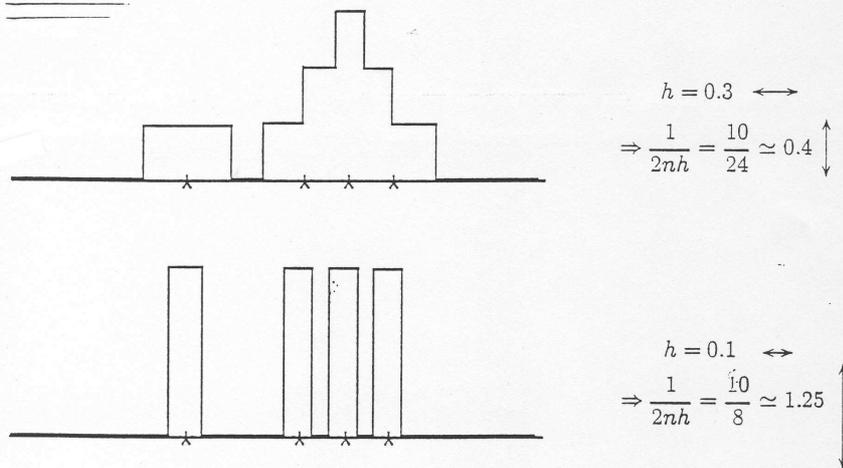


Figure 2.9: Illustration of the influence of the parameter h on the simple estimator.

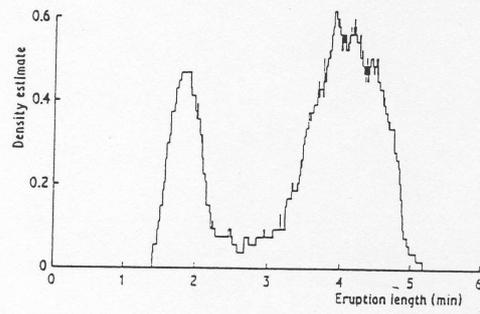


Figure 2.10: The simple estimator for the Old Faithful geyser data set with $h = 0.25$ (from Silverman (1986)).

Les propriétés théoriques d'un estimateur

Critères de comparaison des estimateur: biais, variance, Mean Squared Error

1) Le "biais" d'un estimateur $\hat{f}(x)$ d'une courbe $f(x)$:

$$\text{Biais}[\hat{f}(x)] := E[\hat{f}(x)] - f(x)$$

Un estimateur est dit (asymptotiquement) sans biais si son biais est zéro (tend vers zero pour $n \rightarrow \infty$).

2) La variance d'un estimateur $\hat{f}(x)$

$$\text{var}[\hat{f}(x)] = E\left[\left(\hat{f}(x) - E[\hat{f}(x)]\right)^2\right]$$

3) Erreur quadratique en moyenne (MSE) (pour un point x fixé) :

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

4) Un estimateur est dit "consistant" (en moyenne quadratique) si son MSE tend vers zero pour $n \rightarrow \infty$.

5) D'autres critères : IMSE (ou MISE) = MSE intégrée par rapport à tout x

MSSE = MSE sommée par rapport à une grille de points x_i (par exemple, $x_i = i/n, i = 1, \dots, n$)

et finalement, ISE / SSE = integrated (sum of) squared errors (réalisation par réalisation, donc une v.a.)

Estimation non paramétrique d'une fonction de répartition

Soit $X \sim F$, avec $F(x) = P(X \leq x)$ fonction de répartition de X .

Soient X_1, \dots, X_n échantillon i.i.d. de F et soient

$X_{(1)} \leq \dots \leq X_{(n)}$ l'échantillon ordonné.

Comment estimer F sans hypothèse (paramétrique)?

La fonction de répartition empirique :

$$F_n(x) = \frac{\#\{i : X_i \leq x\}}{n} = \frac{1}{n} \sum_{i=1}^n I\{X_i \leq x\} = \frac{1}{n} \sum_{i=1}^n I\{X_{(i)} \leq x\}.$$

Note que

$$F_n(x) = \begin{cases} 0 & x \leq X_{(1)} \\ k/n & X_{(k)} \leq x < X_{(k+1)} \quad \text{si } k = 1, \dots, n-1 \\ 1 & x \geq X_{(n)}. \end{cases}$$

Exemple : les durées de l'éruption du "Geyser Old Faithful"

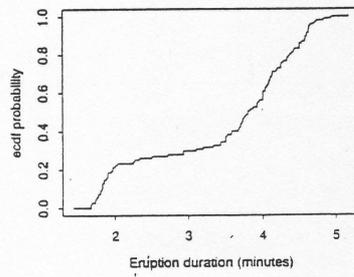


Figure 1.1: Empirical cumulative distribution function of the Old Faithful geyser data (from Scott (1992)).

Les propriétés de la fonction de répartition empirique :

1) $F_n(x)$ est un estimateur sans biais de $F(x)$

2) La variance de $F_n(x)$:

$$n F_n(x) = \sum_{i=1}^n I\{X_i \leq x\} \sim \text{Bin}(n, F(x)) .$$

Donc, $\text{var}[n F_n(x)] = n F(x) (1 - F(x))$.

Par conséquent (loi des grands nombres):

$F_n(x) \rightarrow F(x)$ en probabilité pour tout $x \in \mathbb{R}$.

3) Théorème central limite :

$$\sqrt{n} (F_n(x) - F(x)) \sim \mathcal{N}(0, F(x) (1 - F(x))) \quad \text{si } n \rightarrow \infty .$$

(NB : 4) Distance de Kolmogorov-Smirnov :

$$\sup_x |F_n(x) - F(x)|$$

utile pour comparer deux fonctions de répartition)

Les propriétés de l'estimateur simple à noyau

Motivation : $\hat{f}(x) \rightarrow f(x)$ car

$$f(x) = \lim_{h \rightarrow 0} \frac{F(x+h) - F(x-h)}{2h},$$

$$\hat{f}(x) = \frac{F_n(x+h) - F_n(x-h)}{2h}$$

et

$$\lim_{n \rightarrow \infty} F_n(x) = F(x).$$

Calcul rigoureux :

Biais, variance via $n F_n(x) \sim \text{Bin}(n, F(x))$

1) $E[\hat{f}(x)] = \frac{F(x+h) - F(x-h)}{2h} \rightarrow f(x)$ si $h = h_n \rightarrow 0$.

2) $\text{var}[\hat{f}(x)] \sim \frac{1}{2nh} f(x) \rightarrow 0$ si $n \rightarrow \infty$ et $h = h_n \rightarrow 0$.

Donc, l'estimateur $\hat{f}(x)$ est consistant pour $f(x)$, plus précisément, consistant en moyenne quadratique :

Son MSE tend vers zero lors $n \rightarrow \infty$ et $h = h_n \rightarrow 0$.

$$E\left[\left(\hat{f}(x) - f(x)\right)^2\right] = \text{var}[\hat{f}(x)] + \left(\text{Biais}[\hat{f}(x)]\right)^2 \rightarrow 0.$$

Estimateurs à noyaux lisses

Idée : Remplacer la fenêtre rectangulaire $w(x)$ par une fonction plus lisse $K(x)$ avec la ou les propriété(s) suivante(s):

- $K(x) \geq 0$ pour tout x (positivité)
- $\int K(x) dx = 1$ ("densité")
- $\int x K(x) dx = 0$ ("symétrie")

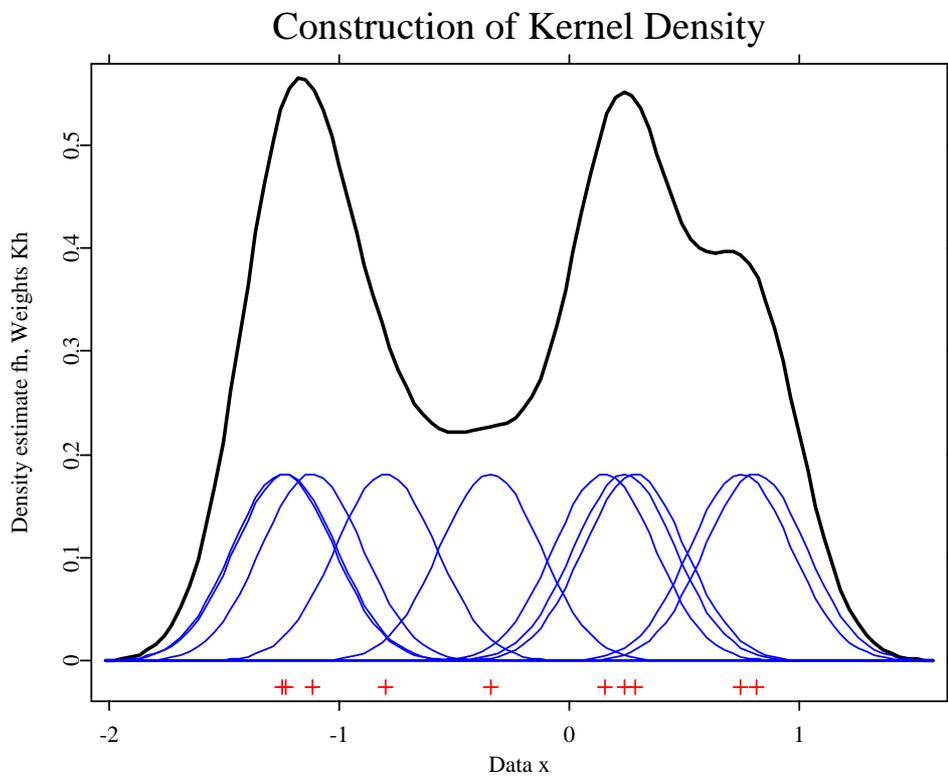
L'estimateur à noyau (de densité) est

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

avec $K =$ la fonction des poids ("noyau, kernel")
et $h =$ le paramètre de lissage ("bandwidth")

Voir les figures sur la construction de cet estimateur et sur l'influence de K et de h sur la forme de cet estimateur.

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



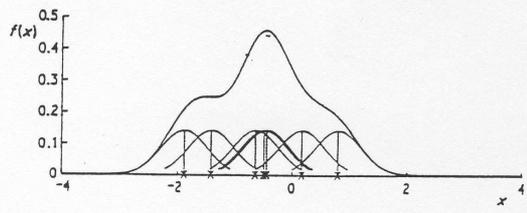


Figure 2.11: Kernel estimate showing individual kernels. Window width 0.4 (from Silverman (1986)).

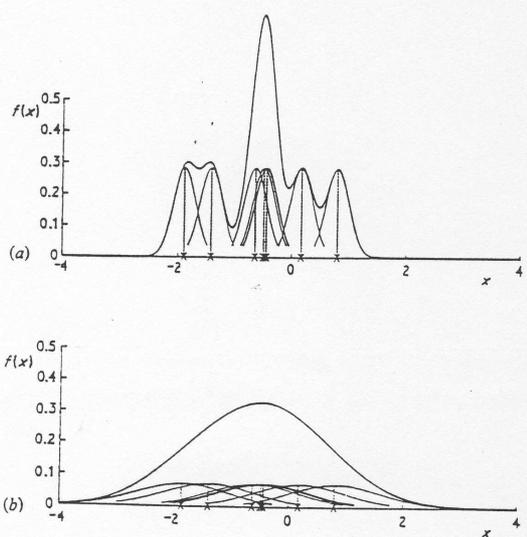
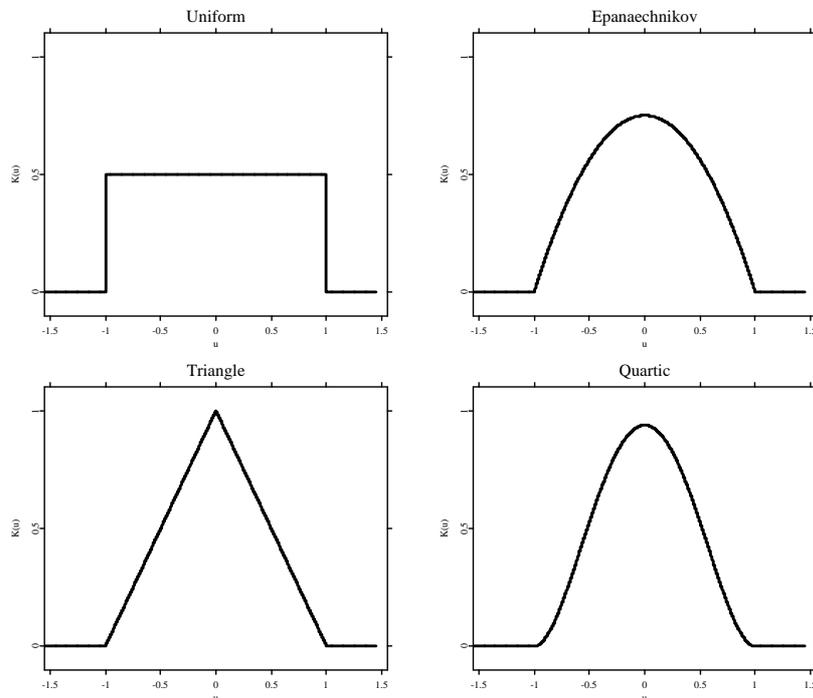


Figure 2.12: Kernel estimates showing individual kernels. Window widths: (a) 0.2; (b) 0.8 (from Silverman (1986)).

Quelques formes des noyaux

Kernel	$K(u)$
Uniform	$\frac{1}{2}I(u \leq 1)$
Triangle	$(1 - u)I(u \leq 1)$
Epanechnikov	$\frac{3}{4}(1 - u^2)I(u \leq 1)$
Quartic	$\frac{15}{16}(1 - u^2)^2I(u \leq 1)$
Triweight	$\frac{35}{32}(1 - u^2)^3I(u \leq 1)$
Gaussian	$\frac{1}{\sqrt{2\pi}} \exp(-\frac{1}{2}u^2)$
Cosinus	$\frac{\pi}{4} \cos(\frac{\pi}{2}u)I(u \leq 1)$



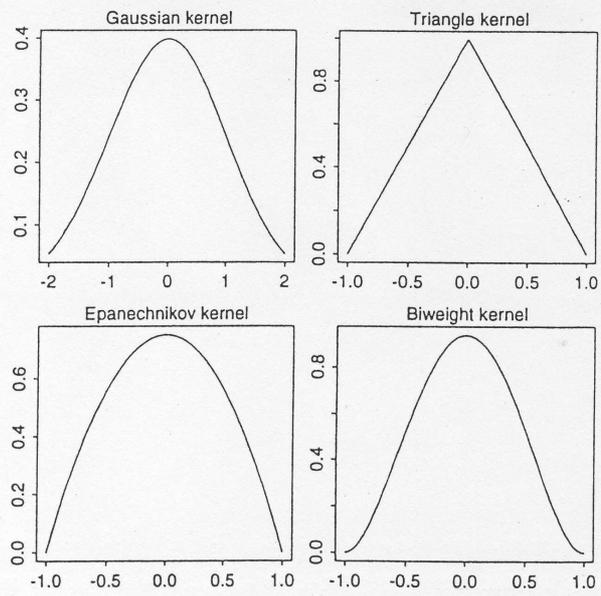


Figure 2.2: Four kernel functions

De l'histogramme aux noyaux - données "Geyser"

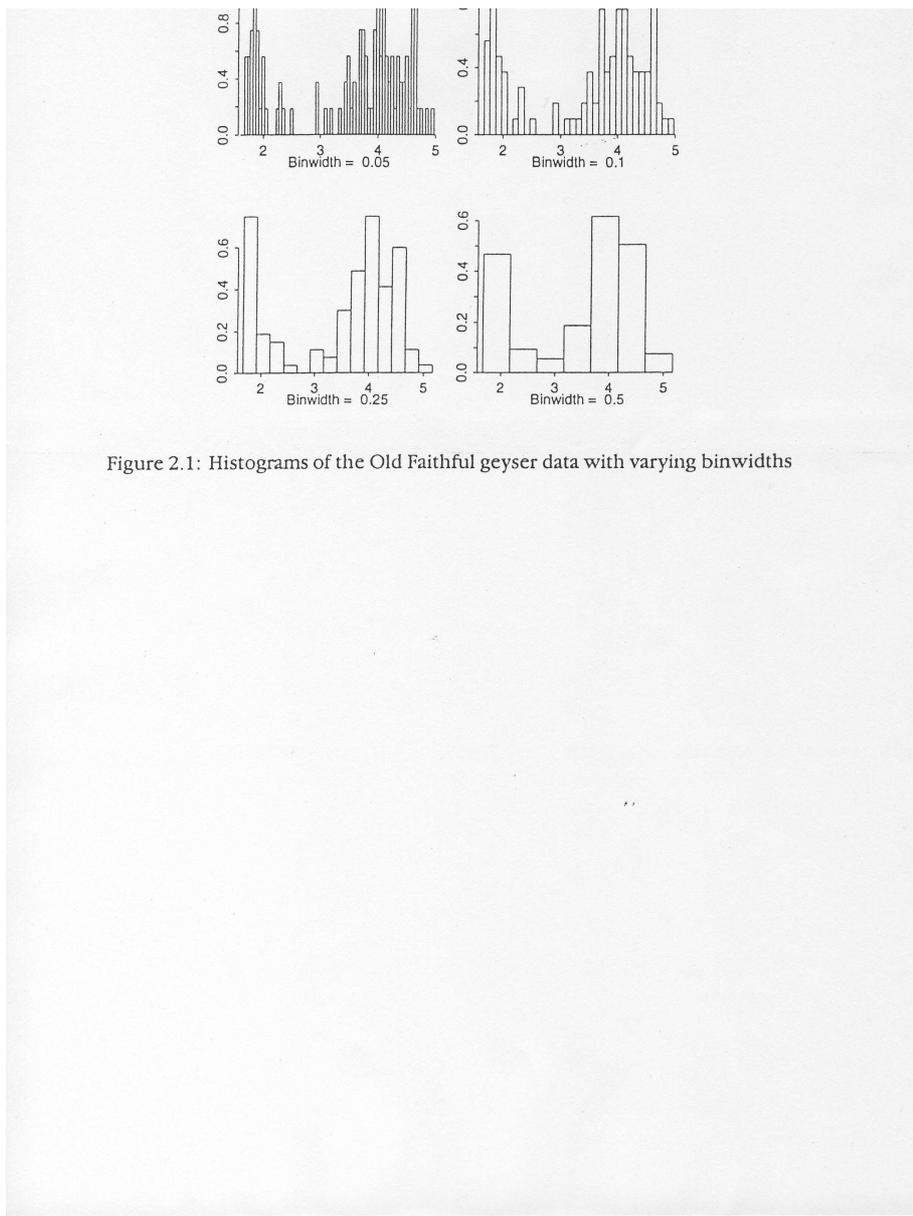


Figure 2.1: Histograms of the Old Faithful geyser data with varying binwidths

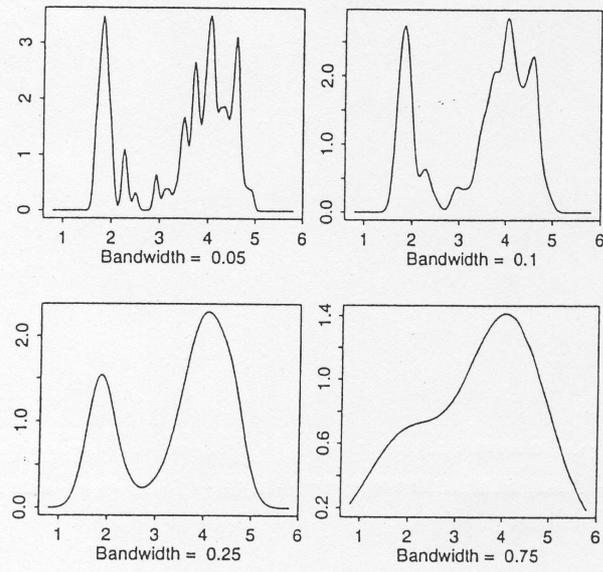


Figure 2.3: Kernel density estimates for the Old Faithful geyser data, using the Gaussian kernel and varying bandwidths

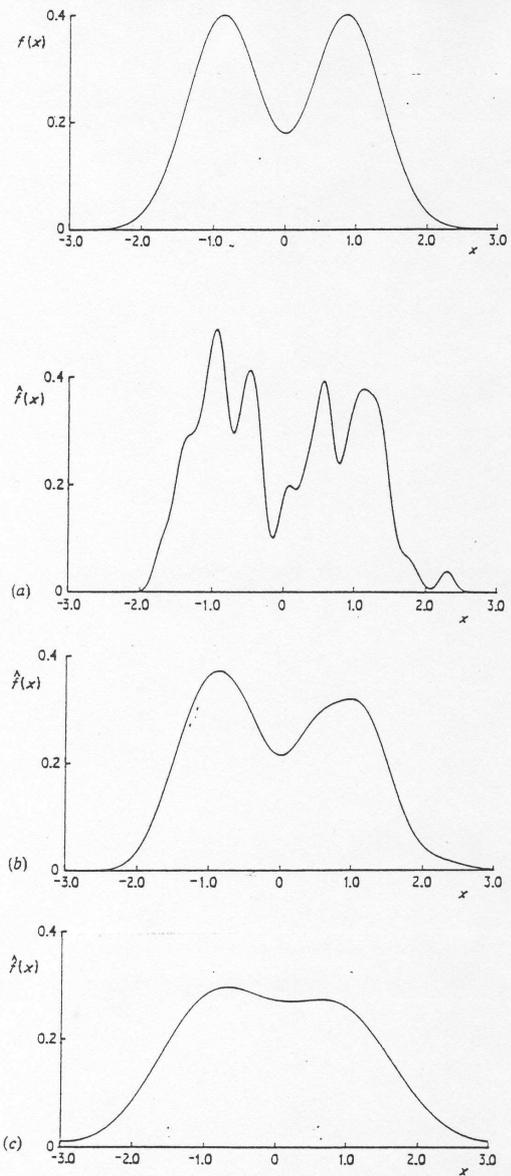
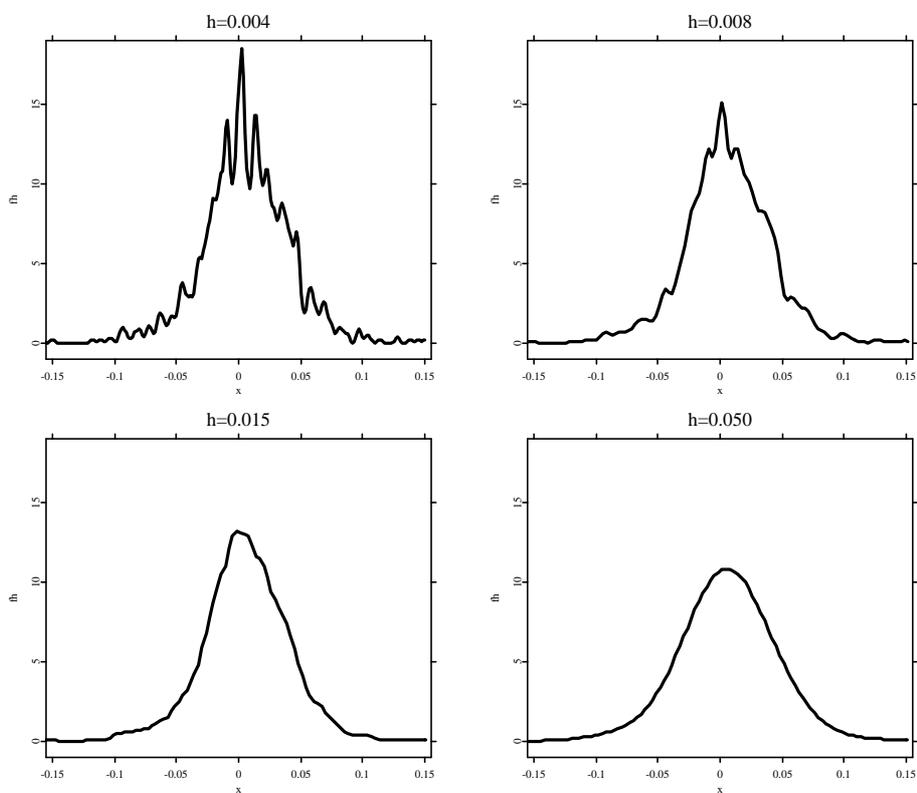
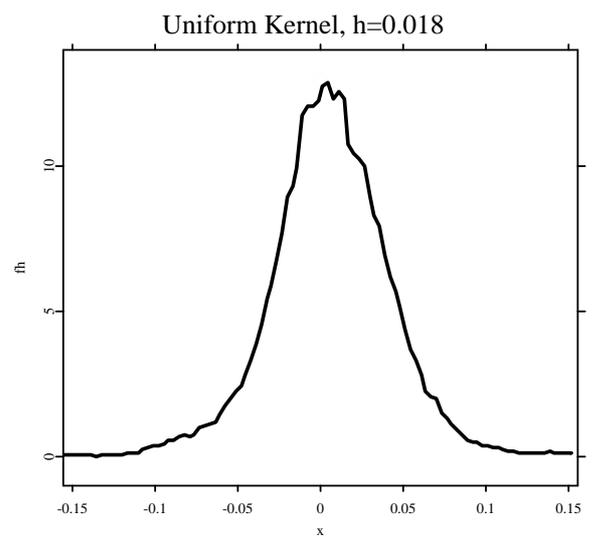
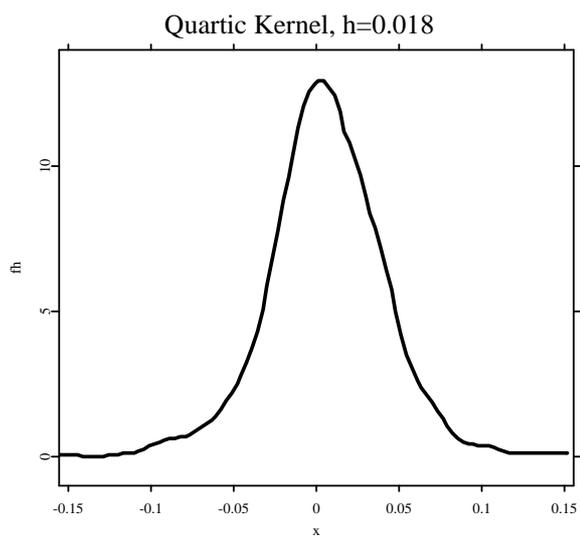


Figure 2.13: Kernel estimates of a bimodal density (top panel) for 200 simulated data points drawn from the density. Window widths: (a) 0.1; (b) 0.3; (c) 0.6. (from Silverman (1986)).

Estimateur à noyau de la densité de "stock returns" avec
différentes h



Estimateur à noyau de la densité de "stock returns" avec des noyaux différents



Quelques propriétés de ces estimateurs à noyau

- K non négative implique \hat{f} non négative
- K densité implique \hat{f} densité
- K continu implique \hat{f} continu
- K différentiable implique \hat{f} différentiable

Les expressions du biais et de la variance de \hat{f} :

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

où $K_h(\cdot) := \frac{1}{h} K\left(\frac{\cdot}{h}\right)$.

(1) Pour calculer le biais notons :

$$\mathbb{E}[\hat{f}(x)] = \mathbb{E}[K_h(x - X)] = \int K_h(x - y) f(y) dy .$$

Définissons la convolution entre deux fonctions f et g

$$(f * g)(x) := \int f(x - y) g(y) dy ,$$

on constate (traitement de l'approximation plus tard) que

$$\begin{aligned} \mathbb{E}[\hat{f}(x)] - f(x) &= (K_h * f)(x) - f(x) \\ &\approx \frac{h^2}{2} f''(x) \int u^2 K(u) du =: \frac{h^2}{2} f''(x) \mu_2 , \end{aligned}$$

sous certaines conditions sur le noyau K .

(2) Pour la variance on calcule :

$$\text{var}[\widehat{f}(x)] = \mathbb{E}[\widehat{f}^2(x)] - \mathbb{E}[\widehat{f}(x)]^2 = \frac{1}{n} ((K_h^2 * f)(x) - (K_h * f)^2(x))$$

Plus loin, on montrera l'approximation (asymptotique)

$$\text{var}[\widehat{f}(x)] \approx \frac{1}{nh} f(x) \int K^2(u) du =: \frac{1}{nh} f(x) R(K) .$$

(3) L'erreur quadratique moyenne ("MSE") :

$$\begin{aligned} MSE[\widehat{f}(x)] &= \mathbb{E}[(\widehat{f}(x) - f(x))^2] = \text{var}[\widehat{f}(x)] + \text{biais}^2[\widehat{f}(x)] = \\ &= \frac{1}{n} (K_h^2 * f)(x) + (1 - \frac{1}{n}) (K_h * f)^2(x) - 2(K_h * f)(x) f(x) + f^2(x) \end{aligned}$$

et son comportement asymptotique,

$$MSE[\widehat{f}(x)] \approx \frac{1}{nh} f(x) R(K) + \frac{h^4}{4} (f''(x))^2 \mu_2^2 .$$

(4) L'erreur quadratique moyenne intégrée ("MISE") :

$$\begin{aligned} MISE[\widehat{f}(\cdot)] &= \int MSE[\widehat{f}(x)] dx \\ &\approx \frac{1}{nh} R(K) + \frac{h^4}{4} \int (f''(x))^2 dx \mu_2^2 . \end{aligned}$$

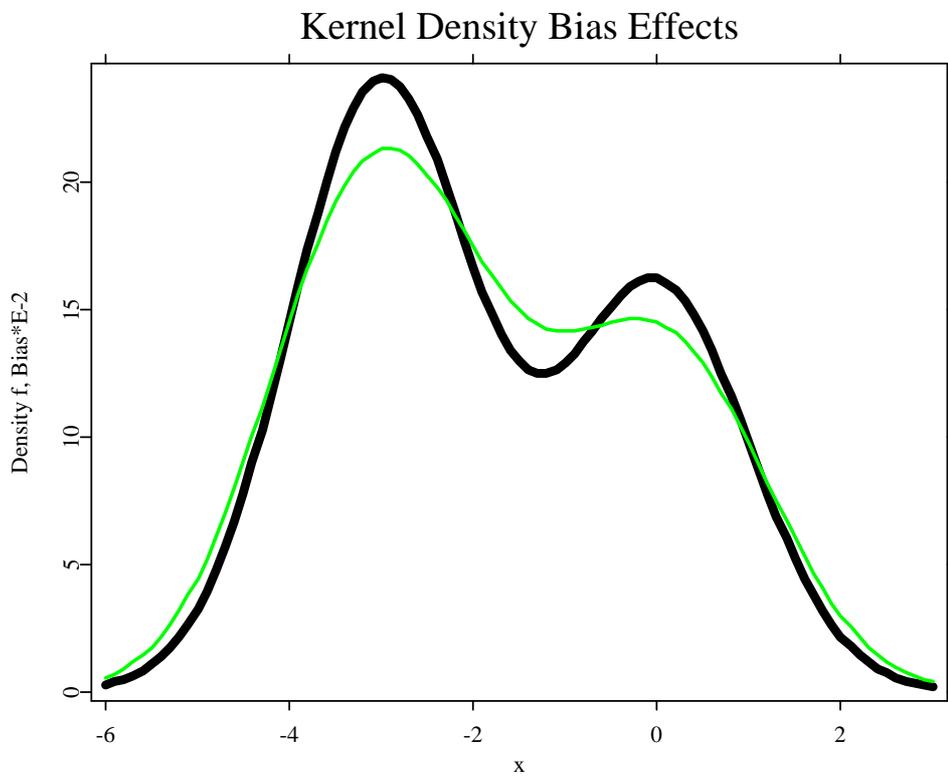
Discussion du comportement de biais et variance

$$E[\hat{f}_h(x) - f(x)] \approx \frac{h^2}{2} f''(x) \int u^2 K(u) du$$

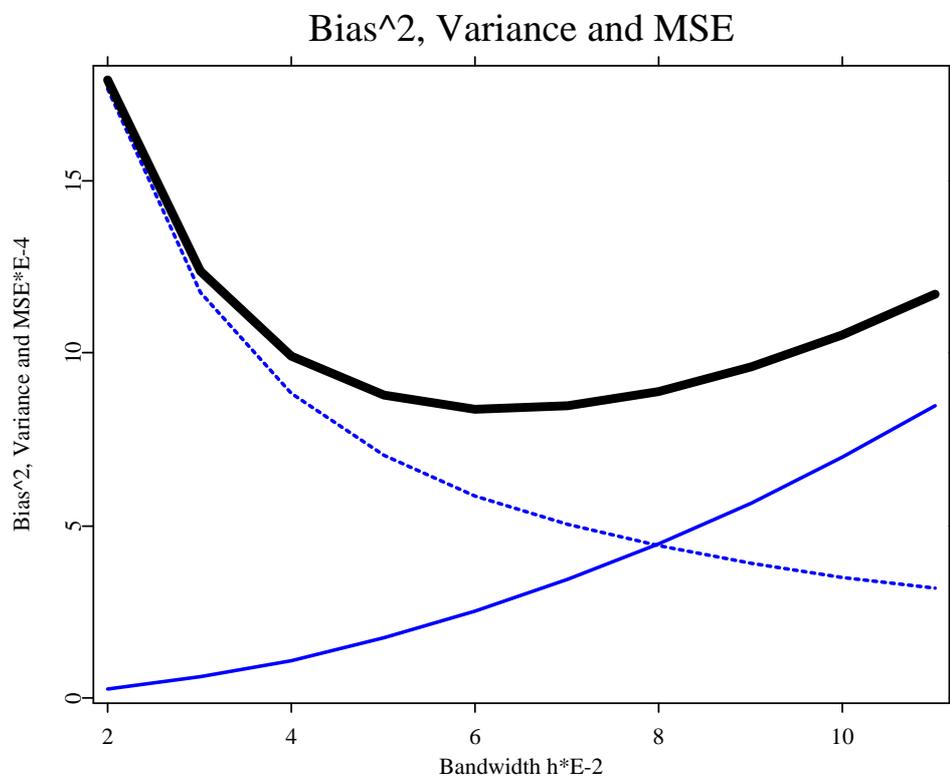
$$\text{Var}(\hat{f}_h(x)) \approx \frac{1}{nh} f(x) \int K^2(u) du$$

- Le biais décroît si h diminue mais la variance augmente.
- La variance diminue si h augmente mais le biais augmente.
- Pour que la variance tende vers zéro, il faut que $nh \rightarrow \infty$.
- Plus la courbure de la densité est haute en x , plus le biais est grand.
- La variance est plus grande pour des valeurs plus grande de la densité.

Le biais depend de la courbure de la densité ($f''(x)$)



Le "trade-off" biais-variance en fonction de h



Expressions asymptotiques du biais et de la variance

Par développement de Taylor de la densité f , supposée deux fois dérivable, on trouve

$$\begin{aligned} \mathbb{E}[\widehat{f}(x)] &= f(x) \int K(u) \, du - f'(x) h \int u K(u) \, du + \\ &\quad + \frac{h^2}{2} f''(x) \int u^2 K(u) \, du + o(h^2), \end{aligned}$$

d'où vient les conditions suivantes au noyau pour simplifier de suite :

- $\int K(u) \, du = 1$
- $\int u K(u) \, du = 0$ (par exemple, si K est symétrique)
- $0 < \mu_2 := \int u^2 K(u) \, du < \infty$.

Donc, pour le biais,

$$\mathbb{E}[\widehat{f}(x)] - f(x) = \frac{h^2}{2} f''(x) \mu_2 + o(h^2).$$

Et idem pour la variance :

$$\text{var}[\widehat{f}(x)] = \frac{1}{nh} f(x) R(K) + o\left(\frac{1}{nh}\right).$$

Expressions asymptotiques du MSE/MISE

$$MSE[\hat{f}(x)] = \frac{1}{nh} f(x) R(K) + \frac{h^4}{4} (f''(x))^2 \mu_2^2 + o\left(h^4 + \frac{1}{nh}\right)$$

et

$$MISE[\hat{f}(x)] = \frac{1}{nh} R(K) + \frac{h^4}{4} \int (f''(x))^2 dx \mu_2^2 + o\left(h^4 + \frac{1}{nh}\right)$$

Ceci implique une possibilité pour choisir le paramètre de lissage h par minimisation du AMSE/AMISE :

Pour l'AMSE, $h = h(x)$, paramètre de lissage variable (**local**), si $f''(x) \neq 0$:

$$h_{AMSE}(x) = \left(\frac{f(x)R(K)}{(f''(x))^2 \mu_2^2} \right)^{1/5} n^{-1/5}$$

Pour l'AMISE, h , paramètre de lissage constant (**global**) :

$$h_{AMISE} = \left(\frac{R(K)}{R(f'') \mu_2^2} \right)^{1/5} n^{-1/5}$$

avec

$$\lim_{n \rightarrow \infty} \frac{h_{MISE}}{h_{AMISE}} = 1 .$$

Problème : dépendance de f'' (et de f) inconnue - possibilité de les remplacer par un estimateur (méthode "plug-in") ce qui est parfois "lourd"

Taux de convergence des M(I)SE

Insérer la bandwidth $h_{opt} = cn^{-1/5}$ dans le M(I)SE donne un taux de convergence de l'ordre

$$M(I)SE(\hat{f}) \sim n^{-4/5}$$

pour un noyau avec $\int u K(u) du = 0$.

Comparaison avec l'histogramme

$$M(I)SE(\hat{f}) \sim n^{-2/3}$$

et le taux de convergence "paramétrique"

$$E[(\hat{\theta} - \theta)^2] \sim n^{-1}$$

Choix optimal du noyau

Insérer la bandwidth $h_{opt} = cn^{-1/5}$ dans le MISE

$$MISE(\hat{f}) = 5/4 (\mu_2^2 R^4(K))^{1/5} (R(f''))^{1/5} n^{-4/5}$$

ce qui dépend du noyau k à travers

$$C(K) = (\mu_2^2 R^4(K))^{1/5}$$

Minimisation (sous contrainte) par rapport à K mène à la solution (Epanechnikov, 1969)

$$K_{opt}(x) = \frac{3}{4}(1 - x^2)I(|x| \leq 1)$$

Quel est l'effet de ne pas utiliser le noyau optimal?

Concept de l'efficacité de K relative à K_{opt} :

$$\text{eff}(K) = \left(\frac{C(K_0)}{C(K)} \right)^{5/4} \leq 1 .$$

Interprétation : rapport de taille d'échantillon pour K_{opt} nécessaire pour atteindre même AMISE qu'avec K

Tableau des efficacités relatives de plusieurs noyaux

	Kernel K	eff(K)
Epanechnikov	$\frac{3}{4}(1 - u^2)I(u \leq 1)$	1.000
Quartic (biweight)	$\frac{15}{16}(1 - u^2)^2I(u \leq 1)$	0.994
Triweight	$\frac{35}{32}(1 - u^2)^3I(u \leq 1)$	0.987
Gaussian	$\frac{1}{\sqrt{2\pi}} \exp(-\frac{1}{2}u^2)$	0.951
Uniform	$\frac{1}{2}I(u \leq 1)$	0.930
Triangle	$(1 - u)I(u \leq 1)$	0.986

Choix pratiques du paramètre du lissage

Trois méthodes à voir:

- la règle simple de référence à une distribution normale
- la méthode de "plug-in"
- la méthode de la validation croisée

1) La règle simple de référence à une distribution normale ("Rule-of-thumb" bandwidth selector) :

Rappelons

$$h_{AMISE} = \left(\frac{R(K)}{R(f'')\mu_2^2} \right)^{1/5} n^{-1/5}$$

et calculons $R(f'')$ pour $f =$ densité de la $\mathcal{N}(\mu, \sigma^2)$:

$$R(f'') = \int ((f''(x))^2) dx = \frac{3}{8\sqrt{\pi} \sigma^5} .$$

Le bandwidth du type "normal reference" est défini par

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

où $\hat{\sigma}$ = estimateur de l'écart-type de X , par exemple

$$\hat{\sigma} = \min\left\{S, \frac{R}{1.349}\right\} ,$$

où S = l'écart-type et R = l'écart interquartile empirique.

Pour un noyau Gaussien/Epanechnikov/biweight

$$\left(\frac{8\sqrt{\pi}R(K)}{3\mu_2^2} \right)^{1/5} = 1.06/2.34/2.78 , \text{ respectivement.}$$

2) La méthode du "plug-in" :

Rappelons

$$h_{AMISE} = \left(\frac{R(K)}{R(f'')\mu_2^2} \right)^{1/5} n^{-1/5}$$

Comment estimer $R(f'')$?

La méthode de Sheather et Jones (1991): par $R(\widehat{f}'')$,
où \widehat{f} est un autre estimateur de f qui se base sur un choix de bandwidth sous-optimal.

Une autre idée est d'estimer $R(f'') = \dots = E[f^4(X)]$ par un estimateur de la dérivée f^4 nécessitant de nouveau le choix d'un autre bandwidth.

Eventuellement passer par une itération (voir le chapitre "Régression non paramétrique").

2) La méthode de la validation croisée :

Idée: Estimer $MISE[\hat{f}]$ et minimiser cet estimateur en fonction de h .

$$MISE[\hat{f}] = E \int \hat{f}^2(x) dx - 2 E \int \hat{f}(x) f(x) dx + \int f^2(x) dx.$$

Minimiser ceci en h , nécessite d'estimer $E \int \hat{f}(x) f(x) dx$.

Un estimateur sans biais est donné par "leave-one-out"

$$\frac{1}{n} \sum_{i=1}^n \hat{f}_{-i}(X_i) = \frac{1}{n} \sum_{i=1}^n \frac{1}{n-1} \sum_{j=1, j \neq i}^n K_h(X_i - X_j)$$

Ici, $\hat{f}_{-i}(X_i)$ = l'estimateur à noyau basé sur l'échantillon réduit $X_1, \dots, X_{i-1}, X_{i+1}, \dots, X_n$.

Ainsi (la méthode "least-squares cross-validation"),

$$\hat{h}_{LSCV} = \arg \min_h LSCV(h) = \arg \min_h \int \hat{f}^2(x) dx - \frac{2}{n} \sum_{i=1}^n \hat{f}_{-i}(X_i).$$

Critique: fort variable et souvent sous-estimant le h_{opt}
(donc susceptible de produire des fausses modes)