

Estimation Non Paramétrique

Habilitation à diriger des recherches

E. Matzner-Løber

Laboratoire de Statistique, Univ. Rennes II

Thèmes de recherche

- Estimation et prévision non paramétriques
- Estimation non paramétrique lorsque les données sont mesurées avec erreur
- Analyse de données fonctionnelles
- Analyse des données.

PARTIE I

Estimation et prévision non paramétriques

- Rappels historiques
- Estimation du mode conditionnel
- Choix de la fenêtre
- Prévision
- Perspectives

Historique

Analyse des séries

Modèles linéaires

Modèles non linéaires

Estimation fonctionnelle

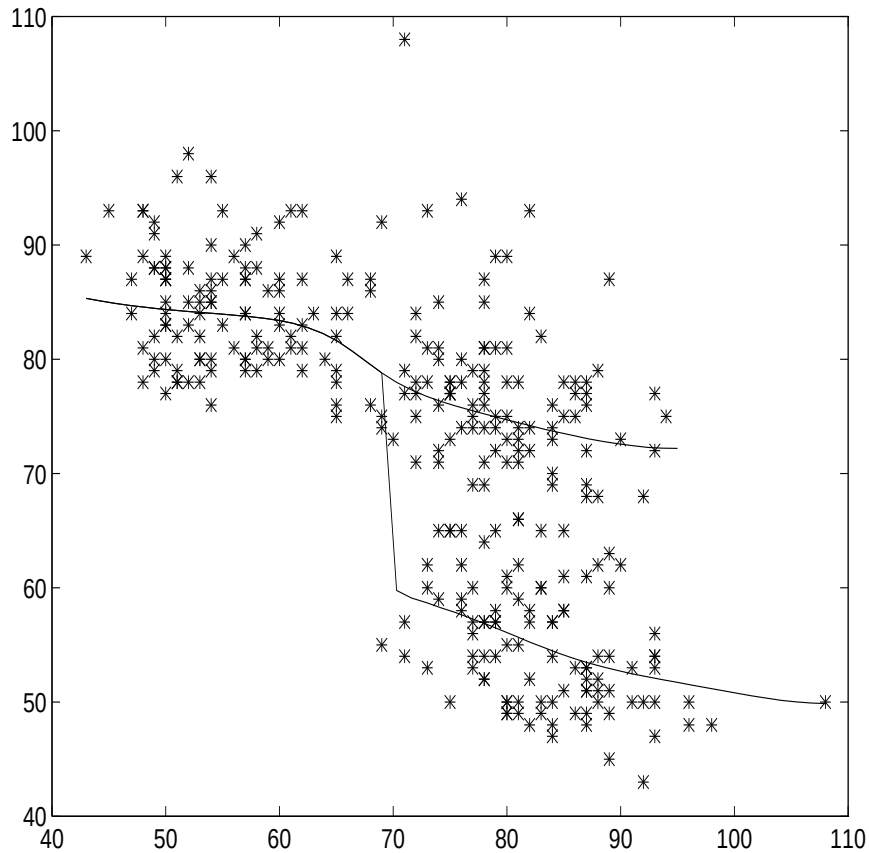
Densité, régression

Indép., mélange

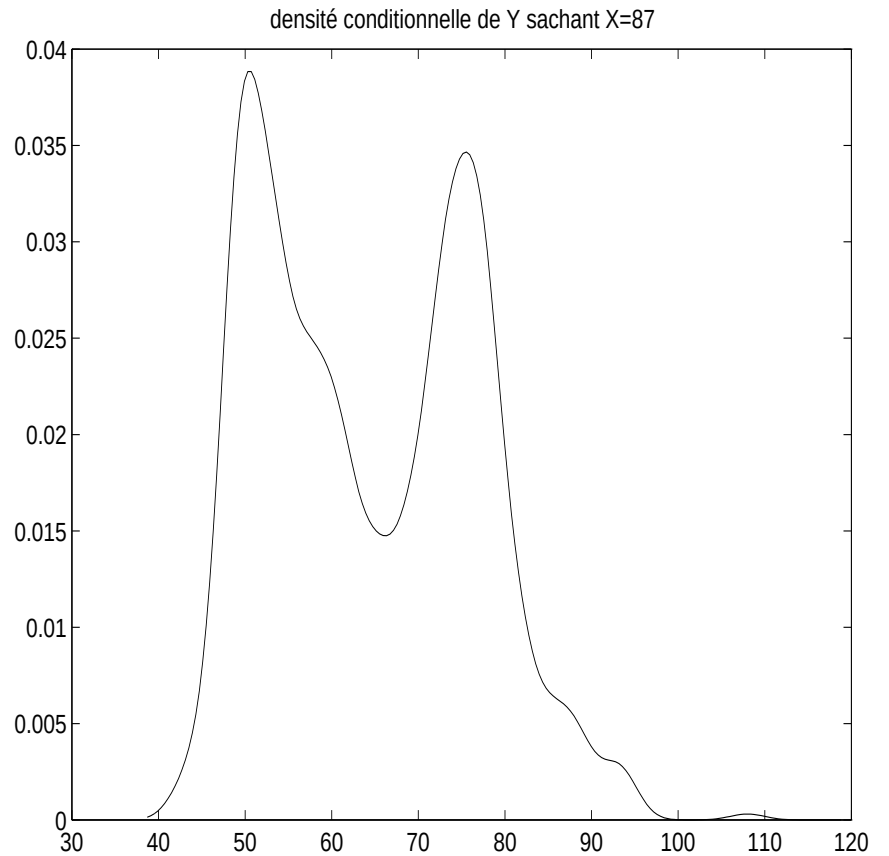
Prévision non paramétrique

Moyenne, mode, médiane conditionnels

Motivations



Régression modale



Prévision modale

Définitions

Le mode conditionnel

$$\Theta(x) = \operatorname{argmax}_{y \in \mathbb{R}} f(y|X = x) = \operatorname{argmax}_{y \in \mathbb{R}} \frac{f(x, y)}{f(x)}.$$

Estimateurs naturels

$$\Theta_{1,n}(x) = \operatorname{argmax}_{y \in \mathbb{R}} f_n(x, y)$$

$$\Theta_{2,n}(x) = \left\{ y : f_n^{(0,1)}(x, y) = 0 \right\}.$$

Résultat

Théorème (Berlinet et al. 1998)

si $\sqrt{nh_n^{d+7}} \rightarrow c$, nous avons

$$\sqrt{nh_n^{d+3}}(\Theta(x) - \Theta_{1,n}(x)) \xrightarrow{\mathcal{L}} \mathcal{N} \left(\frac{m(x, \Theta(x))}{f^{(0,2)}(x, \Theta(x))}, \frac{\sigma^2(x, \Theta(x))}{(f^{(0,2)}(x, \Theta(x)))^2} \right)$$

où

$$m(x, \Theta(x)) = \frac{c}{6} \left(f^{(2,1)}(x, \Theta(x)) \sigma_{K_x}^2 + 3f^{(0,3)}(x, \Theta(x)) \sigma_{K_y}^2 \right)$$

$$\sigma^2(x, \Theta(x)) = f(x, \Theta(x)) \int_{\mathbb{R}^{d+1}} \left(K_x(u) K_y^{(1)}(v) \right)^2 dudv.$$

Choix de la fenêtre

$$Y_i = g(X_i) + \varepsilon_i.$$

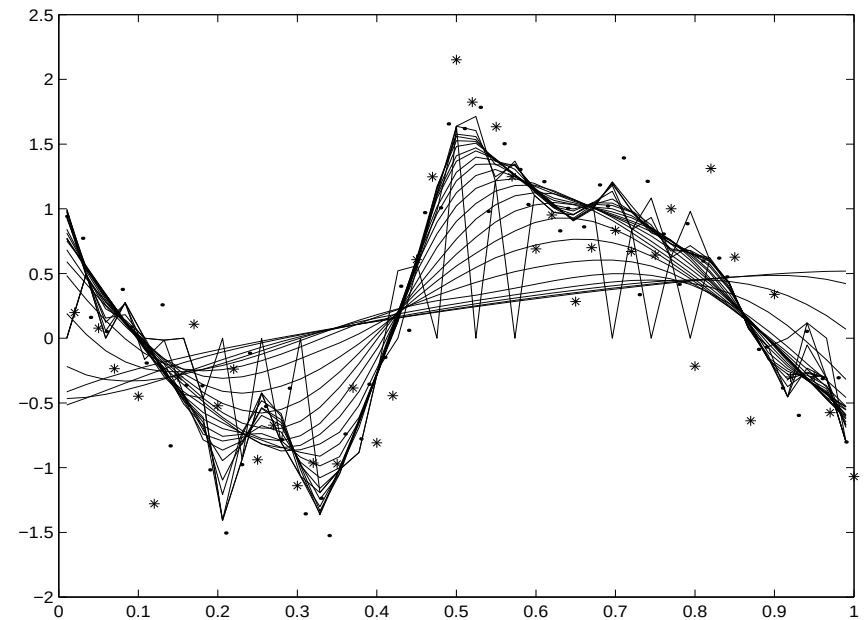
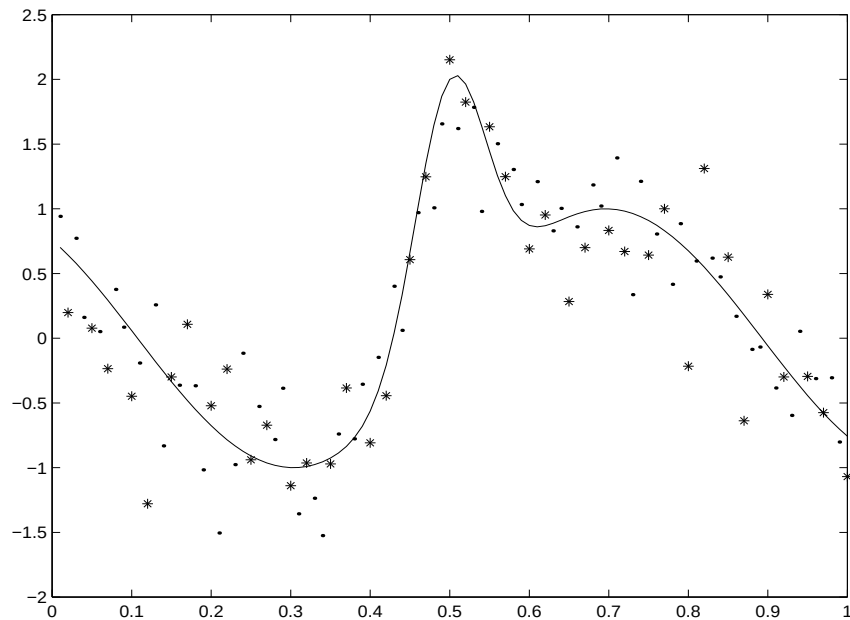
On estime g par les polynômes locaux

$$\sum_{i=1}^n K \left(\frac{x - X_i}{h} \right) (Y_i - \beta_0 - \beta_1(x - X_i) - \cdots - \beta_p(x - X_i)^p)^2.$$

On obtient $\hat{g}_{n,h} = \hat{\beta}_0(x)$.

- Validation croisée
- AIC
- Validation croisée généralisée
- ...

La méthode



Découpage des données
 n pour estimer, m pour valider

Différents estimateurs
 $g_{n,h}$ avec $h \in \mathcal{H}$

$$H = \operatorname{argmin}_{h \in \mathcal{H}} \frac{1}{m} \sum_{j=1}^m (Y_j - \hat{g}_{n,h}(X_i))^2.$$

Résultat

Théorème (Hengartner et al. 2002)

Pour tout $\alpha > 0$ et $\beta > 0$,

$$\begin{aligned} & \frac{2}{2 + \beta} \mathbb{E} \left\{ \int (\hat{g}_{n,H} - g)^2 dP \right\} \\ & \leq (1 + \alpha) \min_{h \in \mathcal{H}} \mathbb{E} \left\{ \int (\hat{g}_{n,h} - g)^2 dP \right\} \\ & \quad + \frac{8\sigma^2(1 + \alpha)}{m\alpha} \{2 + \log(2|\mathcal{H}|)\} + \frac{2(2B)^2}{m\beta} \log(4e|\mathcal{H}|). \end{aligned}$$

Résultat

Théorème (Hengartner et al. 2002)

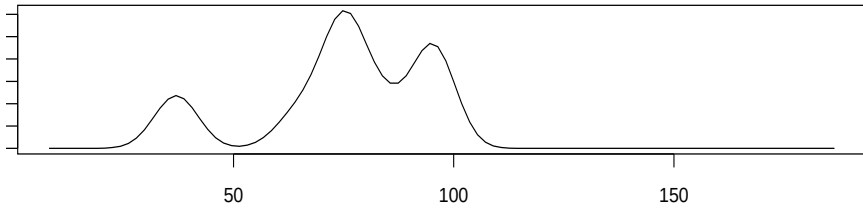
Pour tout $\alpha > 0$ et $\beta > 0$,

$$\begin{aligned} & \frac{2}{2 + \beta} \mathbb{E} \left\{ \int (\hat{g}_{n,H} - g)^2 dP \right\} \\ & \leq (1 + \alpha) \min_{h \in \mathcal{H}} \mathbb{E} \left\{ \int (\hat{g}_{n,h} - g)^2 dP \right\} \\ & \quad + \frac{8\sigma^2(1 + \alpha)}{m\alpha} \{2 + \log(2|\mathcal{H}|)\} + \frac{2(2B)^2}{m\beta} \log(4e|\mathcal{H}|). \end{aligned}$$

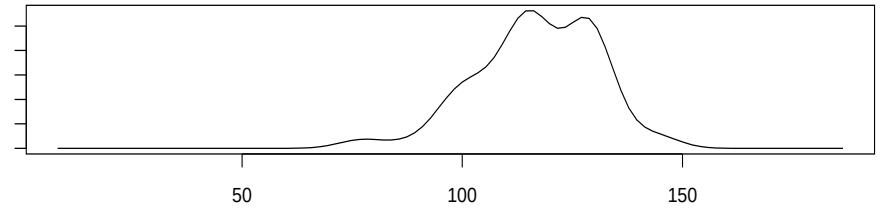
$$\mathcal{H} = \{a, a(1 + \delta), a(1 + \delta)^2, \dots, b\}.$$

Densité conditionnelle

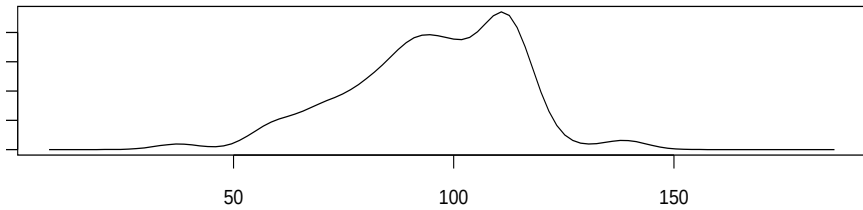
23 avril 99



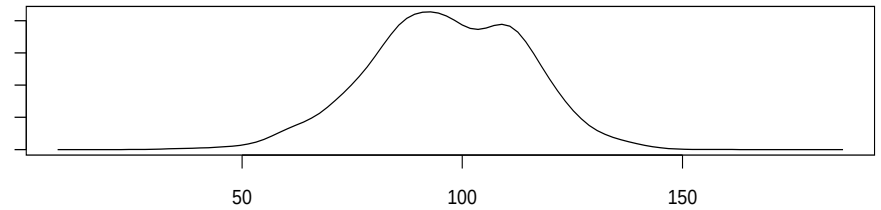
3 mai 99



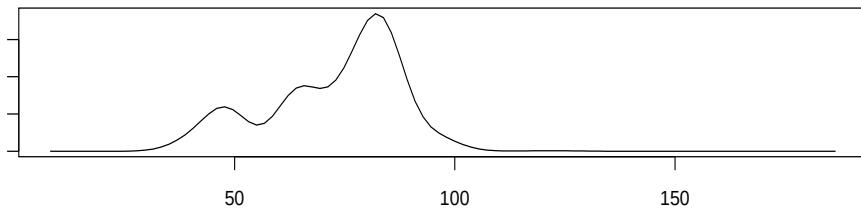
26 mai 99



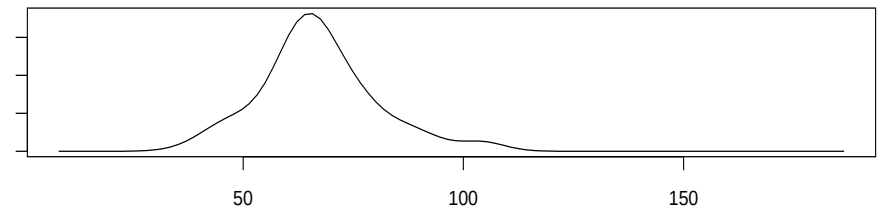
14 juin 99



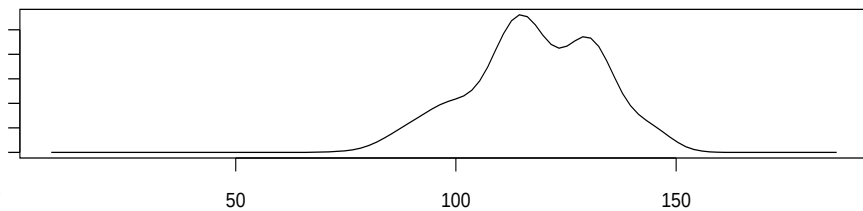
20 juin 99



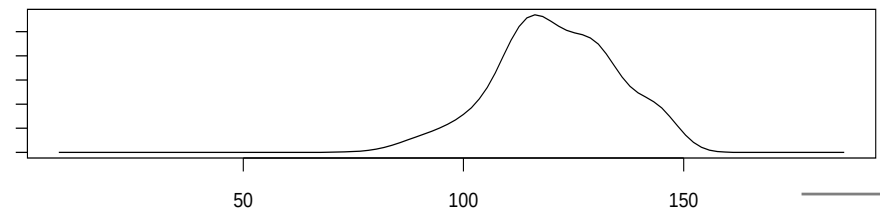
7 juil. 99



9 juil. 99



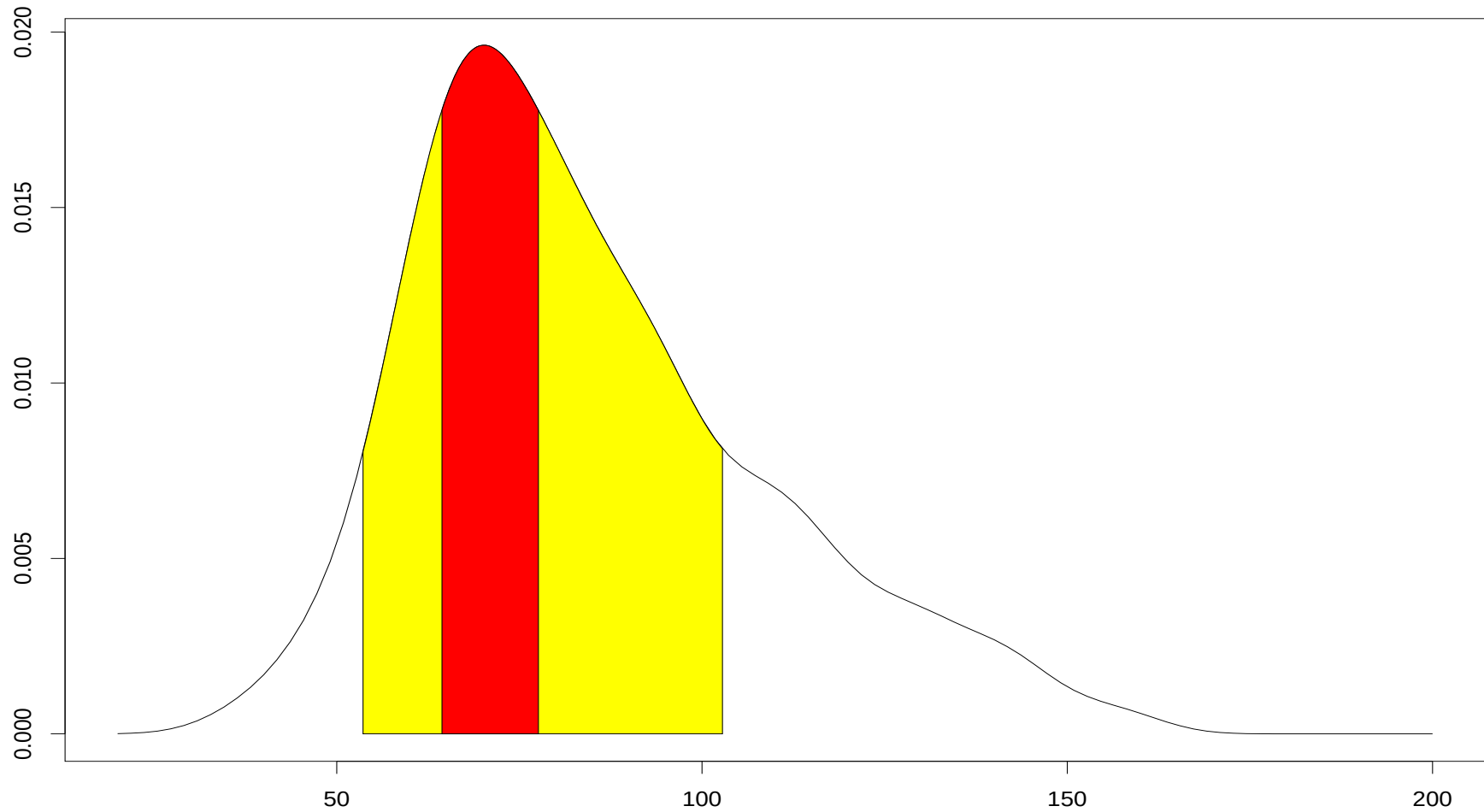
3 sept. 99



Prévision

- ponctuelle
 - moyenne
 - médiane
 - mode
- intervalle
 - via la loi asymptotique
 - via l'estimation des quantiles
 - le plus petit intervalle (ou union d'intervalles) de niveau α .

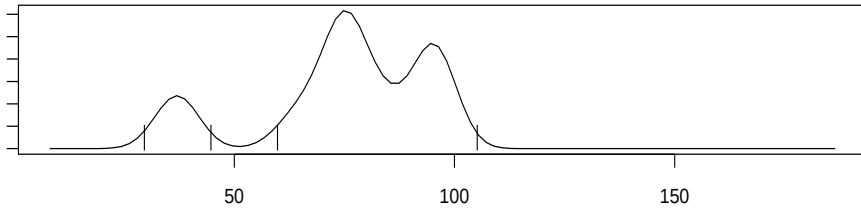
Le “shorth”



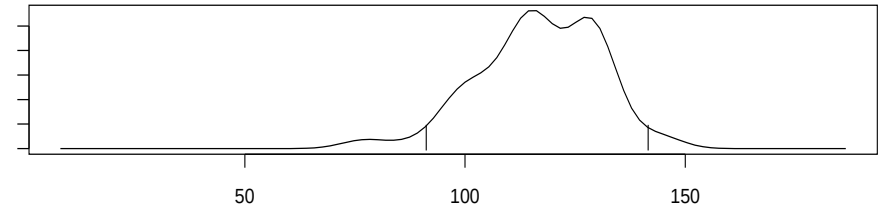
Shorth de niveau 0.25 et de niveau 0.75.

Le “shorth” en pratique

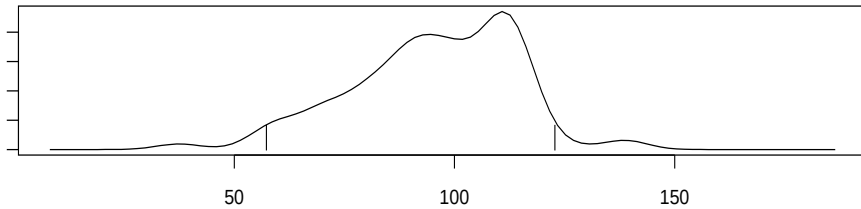
23 avril 99



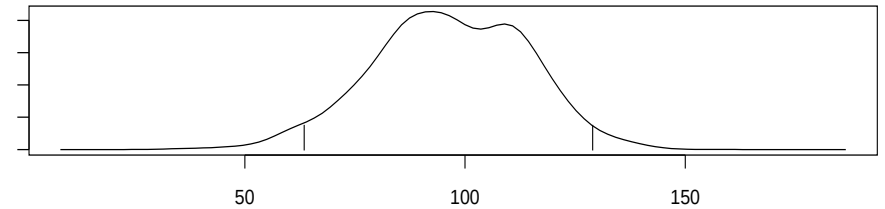
3 mai 99



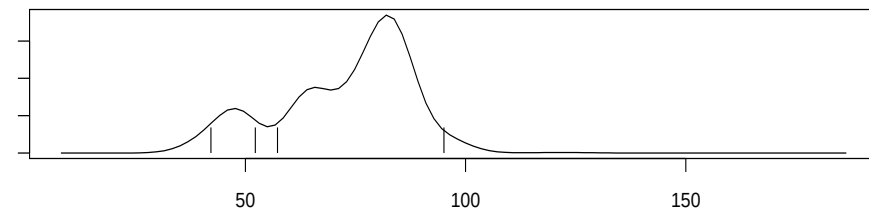
26 mai 99



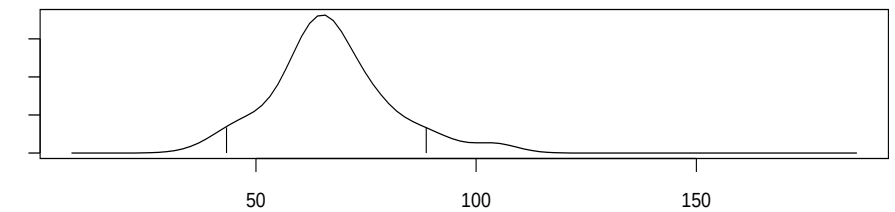
14 juin 99



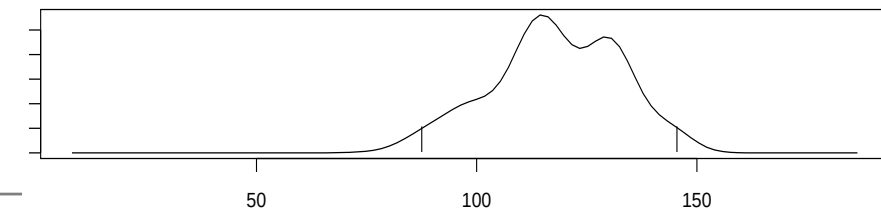
20 juin 99



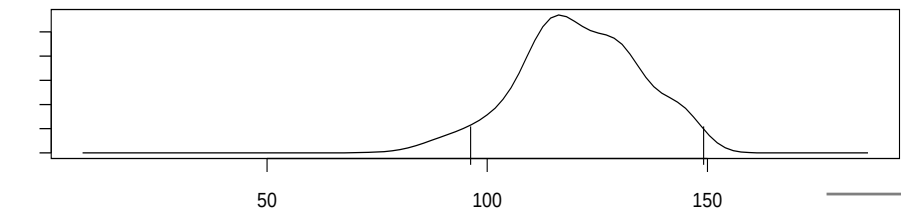
7 juil. 99



9 juil. 99



3 sept. 99



Perspective

Choix d'un modèle en ayant pour objectif l'utilisation du plus petit intervalle (ou union d'intervalles) de prévision.

PARTIE II

Données mesurées avec erreurs

- Rappels historiques
- Motivations
- Noyau de déconvolution
- Estimation des quantiles conditionnels
- Perspective

Historique

Méthodes paramétriques

Modèles linéaires

Modèles non linéaires

Méthodes NP

Densité,

Régression

SIMEX...

Mode et quantiles conditionnels

Motivations

Données mesurées avec erreur ou via un proxy

$$X^e = X + \eta \quad r = \frac{\text{Var}(X)}{\text{Var}(X^e)}.$$

- en économie : revenu (0.85)
- en sciences sociales : le chômage (0.77)
- en médecine : taux de caféine, taux de graisse dans le sang...

Noyau de déconvolution

$$X^e = X + \eta \quad \text{avec} \quad \Phi_\eta(t) \neq 0 \quad \forall t.$$

$$W_\eta(x) = \frac{1}{2\pi} \int_{-\infty}^{\infty} \exp(-itx) \frac{\Phi_{K_1}(t)}{\Phi_\eta\left(\frac{t}{h}\right)} dt$$

W_η dépend de h mais aussi de la loi de η .

- Ordinary smooth : $|\Phi_\eta(t)| \leq |t|^{-\beta}$
- Super smooth : $|\Phi_\eta(t)| \leq \exp(-a|t|^\beta)$

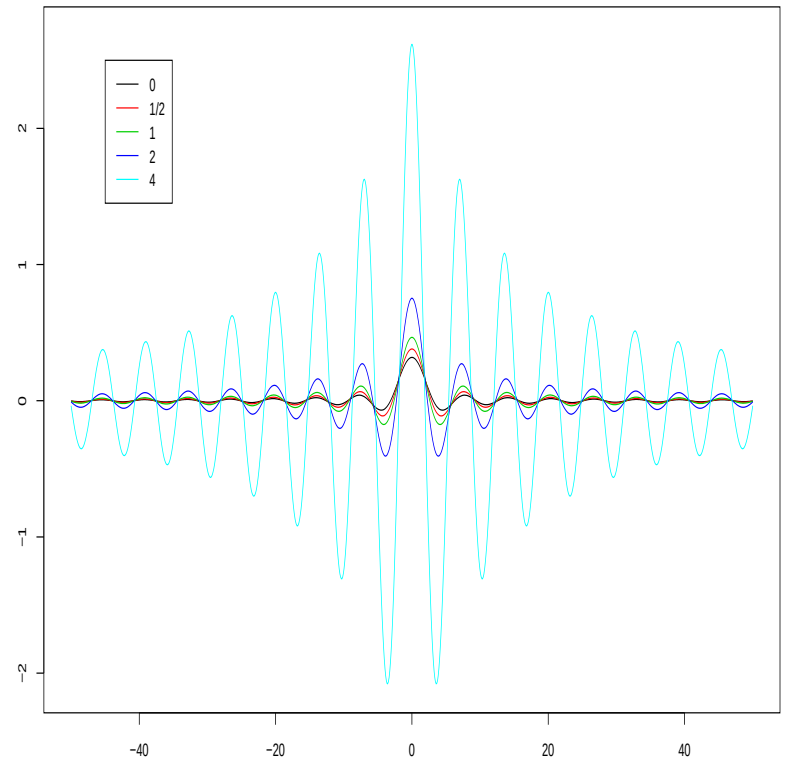
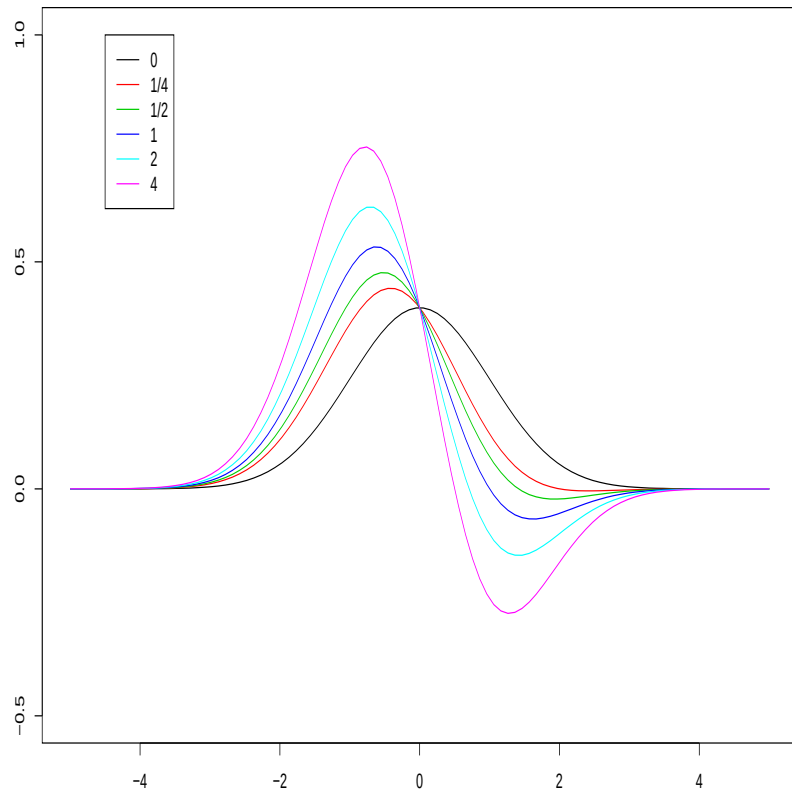
Représentation de 2 noyaux

Loi exponentielle

$$W_{\eta}(x) = \int e^{-itx} \Phi_K(t) \left[1 - i \frac{t}{h} \sigma_{\eta}\right]$$

Loi normale

$$W_{\eta}(x) = \int e^{-itx} \Phi_K(t) \exp \frac{\sigma_{\eta}^2 t^2}{2h^2}$$



Définitions

Un quantile conditionnel d'ordre p

$$F(q(x)|X = x) = p.$$

Estimateurs naturels

$$F_{1,n}(y|x) = \frac{1}{h} \frac{\sum_{i=1}^n W_{\eta} \left(\frac{x - X_i^e}{h_n} \right) \int_{-\infty}^y K_y \left(\frac{z - Y_i}{h_n} \right) dz}{\sum_{i=1}^n W_{\eta} \left(\frac{x - X_i^e}{h_n} \right)}$$

$$F_{2,n}(y|x) = \frac{1}{n} \frac{\sum_{i=1}^n W_{\eta} \left(\frac{x - X_i^e}{h_n} \right) \mathbb{1}_{\{Y_i < y\}}}{\sum_{i=1}^n W_{\eta} \left(\frac{x - X_i^e}{h_n} \right)}$$

Résultat

Théorème (Ioannides et Matzner-Løber, 2005)

si $\sqrt{nh_n^{5+2\beta_n}} \rightarrow 0$, nous avons

$$\sqrt{nh_n^{1+\beta_n}} (q_{1,n}(x) - q(x)) \xrightarrow{\mathcal{L}} \mathcal{N} \left(0, \frac{p(1-p)}{g(x)f^2(q(x)|x)} \int_{\mathbb{R}^d} K_0^2(u) du \right).$$

Perspective

Une firme fabrique un output Y avec un input X .

La frontière de production est donnée par

$$Y_i = g(X_i) + \varepsilon_i - \eta_i.$$

Estimation de g en utilisant des modèles avec erreurs de mesures.

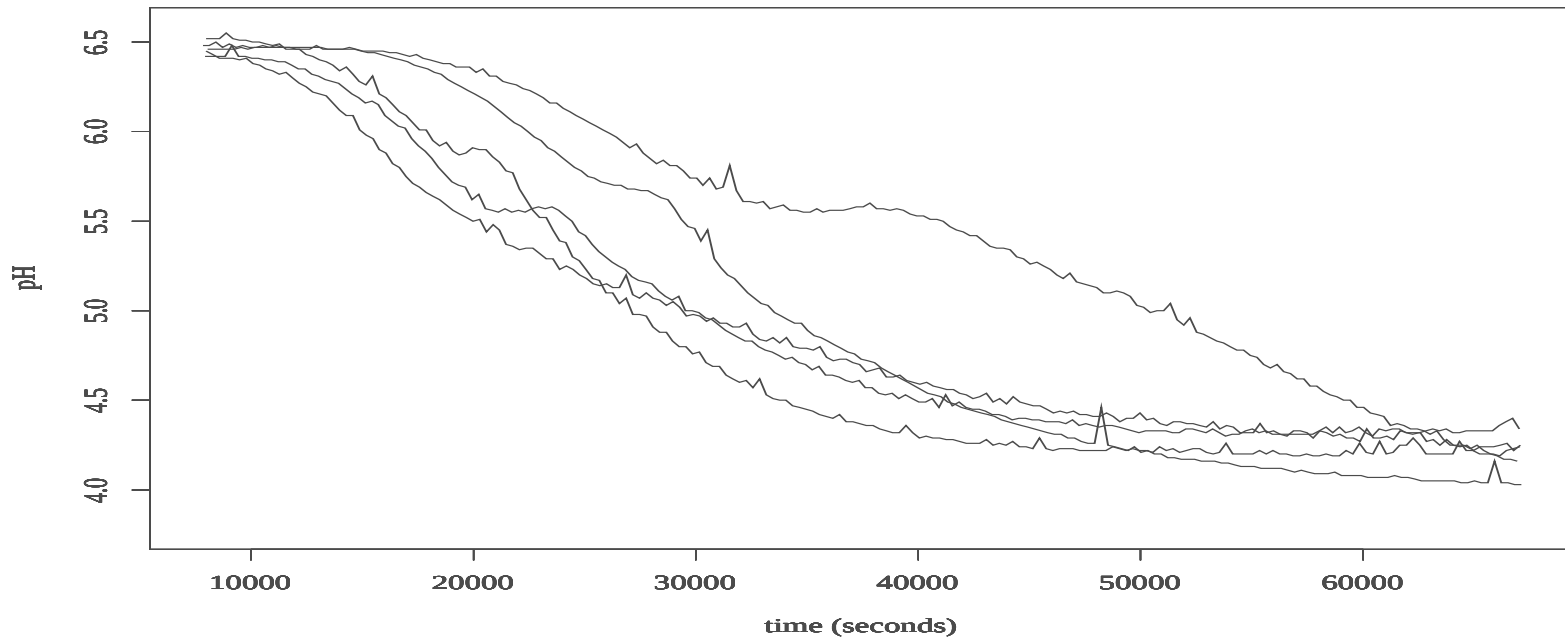
PARTIE III

Analyse de données fonctionnelles

- Motivations
- La méthode
- Résultat
- Perspective

Motivations

- évolution de la température au cours du temps
- évolution du pH au cours du temps
- évolution de l'absorbance en fonction de la longueur d'ondes (spectrométrie)



Classification

- Choix d'un espace d'approximation (S e.v. de dim finie)
- Estimation des coefficients
- Classification des coefficients par k -means
 1. $z^{(0)} = \{c^1, \dots, c^k\}$ avec $c^i \in \mathbb{R}^L$
 2. on classe les coefficients en tenant compte des centres initiaux il faut minimiser

$$\frac{1}{n} \sum_{i=1}^n \min_{c \in z} \|\text{coeff}_i - c\|^2$$

3. on recalcule les centres, on obtient $z^{(1)}$ et on itère
4. on suppose que cet algorithme atteint le meilleur ensemble de centres possible z^n .

Résultats : Abraham *et al.* (2001)

$$\begin{cases} G_1 \text{obs.}[a, b] \\ \vdots \\ G_n \text{obs.}[a, b] \end{cases}$$

↓

$$\begin{cases} \beta_1 \\ \vdots \\ \beta_n \end{cases}$$

$$\lim_n h(z^n, z^*) \rightarrow 0$$

Résultats : Abraham *et al.* (2001)

$$\left\{ \begin{array}{c} G_1 \text{obs.}[a, b] \\ \vdots \\ G_n \text{obs.}[a, b] \end{array} \right\} \quad \left\{ \begin{array}{c} y_1 = G^1(x^1) + \varepsilon^1 \\ \vdots \\ y_n = G^n(x^n) + \varepsilon^n \end{array} \right.$$

↓

$$\left\{ \begin{array}{c} \beta_1 \\ \vdots \\ \beta_n \end{array} \right.$$

↓

$$\left\{ \begin{array}{c} \hat{\beta}_1 \\ \vdots \\ \hat{\beta}_n \end{array} \right.$$

$$\lim_n h(z^n, z^*) \rightarrow 0$$

Résultats : Abraham *et al.* (2001)

$$\begin{cases} G_1 \text{obs.}[a, b] \\ \vdots \\ G_n \text{obs.}[a, b] \end{cases}$$

↓

$$\begin{cases} \beta_1 \\ \vdots \\ \beta_n \end{cases}$$

$$\lim_n h(z^n, z^*) \rightarrow 0$$

$$\begin{cases} y_1 = G^1(x^1) + \varepsilon^1 \\ \vdots \\ y_n = G^n(x^n) + \varepsilon^n \end{cases}$$

↓

$$\begin{cases} \hat{\beta}_1 \\ \vdots \\ \hat{\beta}_n \end{cases}$$

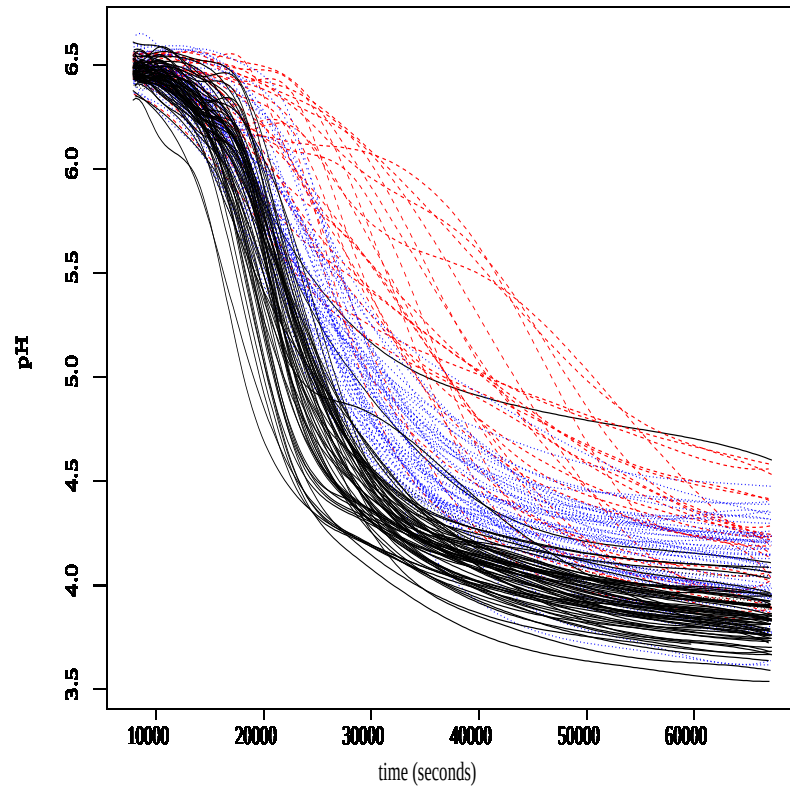
$$\lim_n \lim_m h(\hat{z}^n, z^*) \rightarrow 0$$

$$\begin{cases} y_1 \text{ long.} m_1 \\ \vdots \\ y_n \text{ long.} m_n \end{cases}$$

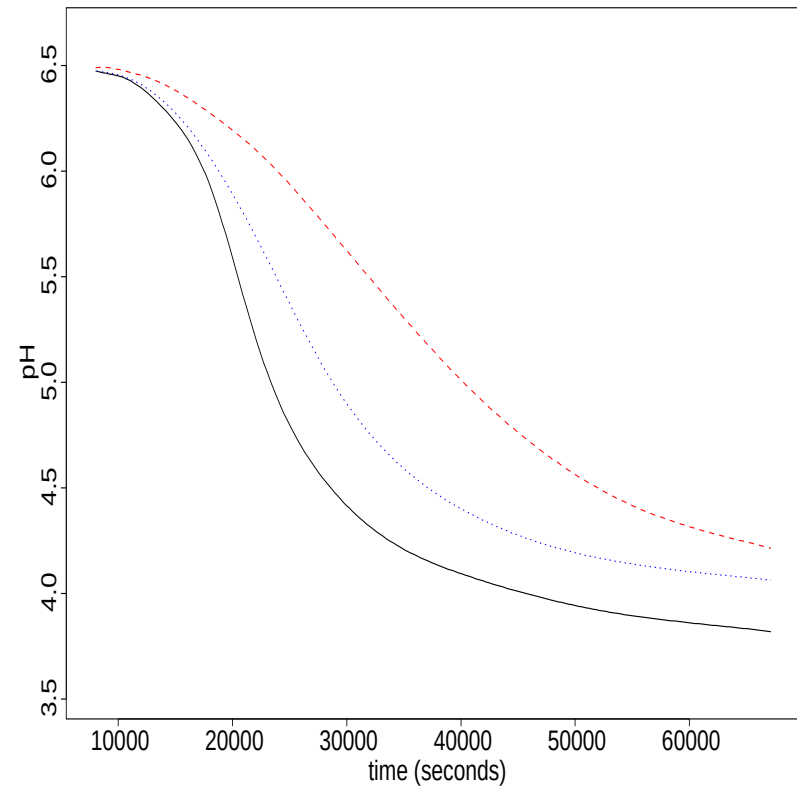
↓

$$\begin{cases} \hat{\beta}_1 \\ \vdots \\ \hat{\beta}_n \end{cases}$$

Exemple d'utilisation



Données lissées groupées



Centre des classes

Perspective

Discrimination de courbes, choix de la distance, agrégation de règles de décision.

PARTIE IV

Analyse des données

- Rappels historiques
- Présentation de la méthode
- Résultat

Historique

Analyse des données

ACP, ACPVI...

ACPVI spline...

Prévision

Prévision np

ACPVI spline

ACPVI

Nous avons 2 tableaux X et Y .

Objectif de l'ACPVI, trouver une métrique R telle que

$$\hat{R} = \arg \min_R ||YY' - XRX'||^2.$$

Solution

$$\hat{R} = [X'X]^+ X'YY'X[X'X]^+$$

ACPVI Spline

Trouver une métrique R et une transformation s tq

$$(\hat{R}, \hat{s}) = \arg \min_{(R, s)} \text{tr} \{ [YY' - X(s)RX'(s)]^2 \}$$

Pas de solution explicite, donc méthode itérative

1. $X(s_0) = X$
2. $\hat{R} = [X'(s_0)X(s_0)]^+ X'(s_0)YY'X(s_0)[X'(s_0)X(s_0)]^+$
3. Méthode du gradient pour trouver s
4. Calcul de R et ainsi de suite.

Résultat

Y univarié

● régression spline

$$\hat{Y}_{spl.reg} = P_B Y = B(B' B)^{-1} B' Y.$$

● ACPVI spline

$$\hat{Y}_{acpvi} = P_{X(\hat{s})} Y = B \hat{S} (\hat{S}' B' B \hat{S})^{-1} \hat{S}' B' Y.$$

Théorème (Cornillon et Matzner-Løber 2005)

$$\hat{Y}_{spl.reg} = \hat{Y}_{acpvi}.$$

Perspectives

- Shorth conditionnel, choix d'un modèle en ayant pour objectif l'utilisation du short
- Prévion sur des champs aléatoires avec pour application la prévion des pics d'ozone en Bretagne (14 sites de mesures)
- Faire le lien entre les modèles ayant des erreurs sur les variables et l'estimation de frontière en économétrie
- Discrimination de courbes, choix de la distance, agrégation de règles de décision.