

ESTIMATION ADAPTATIVE PAR NOYAUX DÉFORMÉS.

Gaëlle Chagny

Laboratoire Map5, UMR CNRS 8145, Université Paris Descartes

**Colloque "Jeunes Probabilistes et Statisticiens", CIRM,
Lundi 16 avril 2012.**

Introduction

Cadre général

Estimateurs à noyaux

Estimateur avec fenêtre
fixée

Déformation

Estimateur

Risque

Adaptativité

Sélection de la fenêtre

Résultat principal

Cas F_X inconnue

Illustrations

Régression

Censure par intervalle

Conclusion

Références

Plan

Introduction

Cadre général

Estimateurs à noyaux

Estimateur avec fenêtre fixée

Déformation

Estimateur

Risque

Adaptativité

Sélection de la fenêtre

Résultat principal

Cas F_X inconnue

Illustrations

Régression

Censure par intervalle

Conclusion

Références

Introduction

Cadre général

Estimateurs à noyaux

Estimateur avec fenêtre
fixée

Déformation

Estimateur

Risque

Adaptativité

Sélection de la fenêtre

Résultat principal

Cas F_X inconnue

Illustrations

Régression

Censure par intervalle

Conclusion

Références

Cadre statistique

Estimation non paramétrique

- ▶ **Modèle** : (X, Y) couple de variables aléatoires réelles, de support $A \times B \subset \mathbb{R}^2$.
 - ▶ X variable de support $A \subset \mathbb{R}$
 - densité : f_X
 - fonction de répartition : F_X .
- ▶ **Objectif** : Reconstruire une fonction s , liée aux variables (X, Y) avec un estimateur à noyau.
- ▶ **Observations** : $(X_i, Y_i)_{i \in \{1, \dots, n\}}$ ($n \in \mathbb{N} \setminus \{0\}$), i.i.d de même loi que (X, Y) .

Introduction

Cadre général

Estimateurs à noyaux

Estimateur avec fenêtre
fixée

Déformation

Estimateur

Risque

Adaptativité

Sélection de la fenêtre

Résultat principal

Cas F_X inconnue

Illustrations

Régression

Censure par intervalle

Conclusion

Références

Exemples étudiés

Exemples	(X, Y)	s	Hypothèses
Ex1 Régression additive	$Y = s(X) + \varepsilon$	s	$\varepsilon \in L^2(\mathbb{P}), \mathbb{E}[\varepsilon] = 0$ $\varepsilon \perp\!\!\!\perp X$
Ex2 Régression multiplicative	$Y = \sigma(X)\varepsilon$	σ^2	$\varepsilon \in L^4(\mathbb{P}), \mathbb{E}[\varepsilon] = 0$ $\mathbb{E}[\varepsilon^2] = 1, \varepsilon \perp\!\!\!\perp X$
Ex3 Censure par intervalle	$(X, Y = \mathbf{1}_{Z \leq X})$	F_Z	$Z, X \geq 0$ $Z \perp\!\!\!\perp X$
Ex4 Censure droite	$(X = Z \wedge C, Y = \mathbf{1}_{Z \leq C})$	$\frac{f_Z}{1-F_Z}$	$Z, C \geq 0$ $F_Z < 1, Z \perp\!\!\!\perp C$

Méthode : Estimateurs à noyaux

▶ **Noyau** : $K : \mathbb{R} \rightarrow \mathbb{R}$, fonction intégrable, tq $\int_{\mathbb{R}} K(x)dx = 1$.

▶ **Approximation de l'unité** :

▶ $\mathcal{H} \subset \mathbb{R}_+^*$ (fenêtres), et

$$\forall h \in \mathcal{H}, K_h : x \mapsto \frac{1}{h} K\left(\frac{x}{h}\right)$$

▶ **Propriété** : si s est suffisamment régulière,

$$K_h * s = \int_{\mathbb{R}} K_h(\cdot - x)s(x)dx \xrightarrow{h \rightarrow 0} s \quad \text{en un certain sens.}$$

▶ **Principe** de l'estimation d'une fonction s avec des noyaux :

- ▶ Trouver ψ telle que, pour tout $h \in \mathcal{H}$, $\mathbb{E}[\psi((X, Y), K_h)] = K_h * s$.
- ▶ Proposer un estimateur de type "moment" pour s :

$$\hat{s}_h = \frac{1}{n} \sum_{i=1}^n \psi((X_i, Y_i), K_h)$$

- ▶ Choisir une "bonne" fenêtre $\hat{h} \in \mathcal{H}$ sur la base des observations

Introduction

Cadre général

Estimateurs à noyaux

Estimateur avec fenêtre
fixée

Déformation

Estimateur

Risque

Adaptativité

Sélection de la fenêtre

Résultat principal

Cas F_x inconnue

Illustrations

Régression

Censure par intervalle

Conclusion

Références

Méthode : Estimateurs à noyaux

Un exemple classique

- ▶ **Ex1** : Régression additive, $Y = s(X) + \varepsilon$
- ▶ **Estimateur de Nadaraya-Watson** (1964)
 - ▶ $\mathbb{E}[YK_h(x - X)] = K_h * (sf_X)(x)$ et $\mathbb{E}[K_h(x - X)] = K_h * f_X(x)$.

- ▶ *Idée* : $s = \frac{sf_X}{f_X} \approx \frac{K_h * (sf_X)}{K_h * f_X}$.

- ▶ *Estimateur* :

$$\hat{s}^{NW}(x) = \frac{\frac{1}{n} \sum_{i=1}^n Y_i K_h(x - X_i)}{\frac{1}{n} \sum_{i=1}^n K_h(x - X_i)}$$

... problème : quotient.

- ▶ ... de même dans les autres exemples étudiés : on ne peut pas trouver ψ telle que $\mathbb{E}[\psi((X, Y), K_h)] = K_h * s$.

Introduction

Cadre général

Estimateurs à noyaux

Estimateur avec fenêtre
fixée

Déformation

Estimateur

Risque

Adaptativité

Sélection de la fenêtre

Résultat principal

Cas F_X inconnue

Illustrations

Régression

Censure par intervalle

Conclusion

Références

Estimateur pour une fenêtre fixée

Méthode de déformation

- ▶ **Point clé** : Il existe une fonction $\phi_X : A \rightarrow \phi_X(A)$ bijective, telle que

$$\mathbb{E}[\theta(Y)K_h(u - \phi_X(X))] = K_h * (s \circ \phi_X^{-1})(u),$$

$$\text{avec } \theta(Y) = \begin{cases} Y & (\text{Ex1,3,4}) \\ Y^2 & (\text{Ex2}) \end{cases}$$

- ▶ **Déformation** ϕ_X :

- ▶ $\phi_X = F_X$, fonction de répartition de X pour Ex1, Ex2 (régression) et Ex3 (censure par intervalle) bijective si $f_X > 0$ sur A .
- ▶ $\phi_X = \phi : x \mapsto \int_0^x (1 - F_X(t)) dt$, dans l'Ex4 (censure à droite) bijective si $F_X(x) < 1$ pour tout $x \in \mathbb{R}_+$.

- ▶ **Méthode** :

1. Estimer $g = s \circ \phi_X^{-1}$ par un estimateur à noyau \hat{g} .
2. Poser

$$\hat{s} = \begin{cases} \hat{g} \circ \phi_X & \text{si } \phi_X \text{ est connue,} \\ \hat{g} \circ \hat{\phi}_X & \text{sinon.} \end{cases}$$

- ▶ **Références** : Yang (1981), Stute (1984,1986), Kerkycharian et Picard (2004)

Introduction

Cadre général

Estimateurs à noyaux

Estimateur avec fenêtre fixée

Déformation

Estimateur

Risque

Adaptativité

Sélection de la fenêtre

Résultat principal

Cas F_X inconnue

Illustrations

Régression

Censure par intervalle

Conclusion

Références

Estimateur pour une fenêtre fixée

Soit $h \in \mathcal{H}$ fenêtre fixée.

1. **Estimateur pour $g = s \circ \phi_X^{-1} : \phi_X(A) \rightarrow \mathbb{R}$**

$$\forall u \in \phi_X(A), \hat{g}_h(u) = \frac{1}{n} \sum_{i=1}^n \theta(Y_i) K_h(u - \phi_X(X_i)).$$

2. **Estimateur pour $s : A \rightarrow \mathbb{R}$**

$$\forall x \in A, \hat{s}_h(x) = \hat{g}_h \circ \phi_X(x) = \frac{1}{n} \sum_{i=1}^n \theta(Y_i) K_h(\phi_X(x) - \phi_X(X_i)).$$

→ estimateur simple (moyenne empirique), sans quotient.

Introduction

Cadre général

Estimateurs à noyaux

Estimateur avec fenêtre
fixée

Déformation

Estimateur

Risque

Adaptativité

Sélection de la fenêtre

Résultat principal

Cas F_X inconnue

Illustrations

Régression

Censure par intervalle

Conclusion

Références

Estimateur pour une fenêtre fixée

Risque

- ▶ **Risque quadratique intégré** pondéré par ϕ'_X :

$$\mathbb{E} \left[\|\hat{s}_h - s\|_{\phi'_X}^2 \right] = \int_A (\hat{s}_h(x) - s(x))^2 \phi'_X(x) dx = \mathbb{E} \left[\|\hat{g}_h - g\|^2 \right].$$

- ▶ **Décomposition biais-variance**

$$\mathbb{E} \left[\|\hat{s}_h - s\|_{\phi'_X}^2 \right] = \underbrace{\mathbb{E} \left[\|\hat{g}_h - K_h * g\|^2 \right]}_{\text{Variance}} + \underbrace{\|K_h * g - g\|^2}_{\text{Biais}}.$$

- ▶ **Majoration du risque** : si g est régulière d'indice α (espaces de Hölder ou Nikol'skii), et si K est d'ordre suffisant,

$$\mathbb{E} \left[\|\hat{s}_h - s\|_{\phi'_X}^2 \right] \leq c(Y_1) \|K\|_{L^2(\mathbb{R})}^2 \frac{1}{nh} + ch^{2\alpha}.$$

avec $c(Y_1) = \mathbb{E}[Y_1^2]$, (Ex1), $\mathbb{E}[Y_1^4]$, (Ex2), 1 (Ex3 et 4).

→ pour h bien choisi, vitesse de convergence du risque
 $n^{-\frac{2\alpha}{2\alpha+1}}$.

Introduction

Cadre général

Estimateurs à noyaux

Estimateur avec fenêtre fixée

Déformation

Estimateur

Risque

Adaptativité

Sélection de la fenêtre

Résultat principal

Cas F_X inconnue

Illustrations

Régression

Censure par intervalle

Conclusion

Références

Sélection automatique de la fenêtre

Méthode de Goldenshluger-Lepski (2011)

1. Approximation du terme de variance

$$\forall h \in \mathcal{H}, V(h) = \delta(1 + \|K\|_{L^1(\mathbb{R})}^2) \frac{\|K\|_{L^2(\mathbb{R})}^2}{nh}.$$

2. Approximation du terme de biais

$$\forall h \in \mathcal{H}, A(h, \phi_X) = \max_{h' \in \mathcal{H}} \left\{ \|\hat{g}_{h'} - \hat{g}_{h,h'}\|^2 - V(h') \right\}_+.$$

avec $\hat{g}_{h,h'} = K_{h'} * \hat{g}_h$.

- ▶ **Règle de sélection** : $\hat{h} \in \operatorname{argmin}_{h \in \mathcal{H}} \{A(h, \phi_X) + V(h)\}$
- ▶ **Estimateur** :
 - ▶ pour g : $\hat{g}_{\hat{h}}$
 - ▶ pour s :

$$\hat{s}_{\hat{h}} = \hat{g}_{\hat{h}} \circ \phi_X$$

Introduction

Cadre général

Estimateurs à noyaux

Estimateur avec fenêtre
fixée

Déformation

Estimateur

Risque

Adaptativité

Sélection de la fenêtre

Résultat principal

Cas F_X inconnue

Illustrations

Régression

Censure par intervalle

Conclusion

Références

Résultat principal

Borne non asymptotique pour le risque : résultat adaptatif

Hypothèses

- ▶ $s \in L^\infty(A)$.
- ▶ Collection \mathcal{H} pas trop "grosse"
 - ▶ $\mathcal{H} = \{k^{-1}, k = 1, \dots, [\sqrt{n}]\}$,
 - ▶ $\mathcal{H} = \{2^{-k}, k = 1, \dots, [\log_2(n)]\}$.
- ▶ hypothèse d'existence de moments pour le bruit ε dans les Ex1 et Ex2 (régression)

• Théorème

Il existe $c_i, i = 1, 2$ et C des constantes, telles que

$$\mathbb{E} \left[\|\hat{s}_{\hat{h}} - s\|_{\phi'_X}^2 \right] \leq \min_{h \in \mathcal{H}} \left\{ c_1 \|K\|_{L^2(\mathbb{R})}^2 \frac{1}{nh} + c_2 \|K_h * g - g\|^2 \right\} + \frac{C}{n} \bullet$$

→ pour g d'indice de régularité α , et K d'ordre suffisant, vitesse de convergence du risque $n^{-\frac{2\alpha}{2\alpha+1}}$.

Introduction

Cadre général

Estimateurs à noyaux

Estimateur avec fenêtre
fixée

Déformation

Estimateur

Risque

Adaptativité

Sélection de la fenêtre

Résultat principal

Cas F_X inconnue

Illustrations

Régression

Censure par intervalle

Conclusion

Références

Cas général F_X inconnue

Méthode de plug-in

- ▶ Remplacer ϕ_X par un estimateur, partout où elle intervient.
- ▶ **Version empirique de ϕ_X**
 - ▶ *Observations supplémentaires* : $(X_{-i})_{i \in \{1, \dots, n\}}$, indépendantes des $(X_i)_i$ et de même loi.
 - ▶ *Estimateur $\hat{\phi}_X$ de F_X ou de ϕ* :

$$\hat{F}_n = \frac{1}{n} \sum_{i=1}^n \mathbf{1}_{]-\infty; X_{-i}]}, \quad \hat{\phi}_n(x) = \int_0^x (1 - \hat{F}_n(t)) dt = \frac{1}{n} \sum_{i=1}^n X_{-i} \wedge x.$$

- ▶ **Estimateur pour g et pour s** :

$$\hat{g}_{\hat{h}}^{\hat{\phi}}(u) = \frac{1}{n} \sum_{i=1}^n \theta(Y_i) K_{\hat{h}}(u - \hat{\phi}_n(X_i)) \quad \hat{s}_{\hat{h}}(x) = \hat{g}_{\hat{h}}^{\hat{\phi}} \circ \hat{\phi}_n(u).$$

→ Mêmes résultats théoriques, avec un peu plus d'hypothèses (contrainte sur \mathcal{H} , et hypothèse $s \in C^1(\mathbf{A})$).

Introduction

Cadre général

Estimateurs à noyaux

Estimateur avec fenêtre
fixée

Déformation

Estimateur

Risque

Adaptativité

Sélection de la fenêtre

Résultat principal

Cas F_X inconnue

Illustrations

Régression

Censure par intervalle

Conclusion

Références

1. Objectifs

- ▶ Illustrer la méthode,
- ▶ Comparer des estimateurs de type 'noyaux déformés' aux estimateurs de type 'moindres carrés' (en base trigonométrique, ou fondée sur des polynômes par morceaux)
→ toolbox Matlab FY3P de [Y. Rozenholc](#)

2. Fenêtres et noyaux

- ▶ **Fenêtres** : $\mathcal{H}_n = \{k^{-1}, k = 1, \dots, [\sqrt{n}]\}$,
- ▶ **Noyau** : gaussien
→ noyau d'ordre 1 \implies Polynômes par morceaux de degré au plus 1.

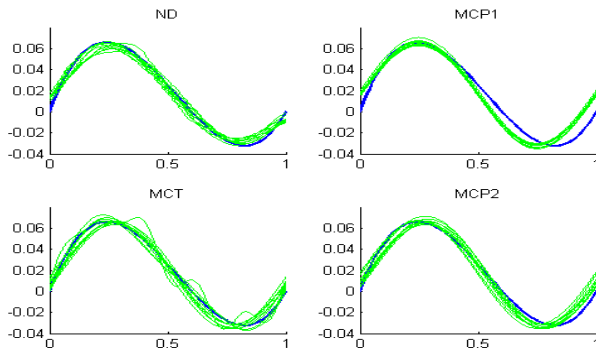
3. Exemples :

- ▶ en régression additive ([Baraud \(2002\)](#))
- ▶ en censure par intervalle ([Brunel et Comte \(2009\)](#)).

Simulations : Ex1 Régression additive

(X, Y) tels que $Y = s(X) + \varepsilon$, $s(x) = x(x-1)(x-0.6)$.

Observations : $X \sim \mathcal{U}_{[0;1]}$, $\varepsilon \sim \mathcal{N}(0, 0.006)$, $n = 1000$.



Comparaison des risques ($\times 10^3$)

n=60	200	500	2000	Méthode
0.3747	0.1279	0.0604	0.0324	ND
0.5222	0.447	0.5846	0.6469	MCT
0.3772	0.1283	0.0802	0.0666	MCP1
0.3892	0.1293	0.0681	0.0446	MCP2

Introduction

Cadre général

Estimateurs à noyaux

Estimateur avec fenêtre
fixée

Déformation

Estimateur

Risque

Adaptativité

Sélection de la fenêtre

Résultat principal

Cas F_x inconnue

Illustrations

Régression

Censure par intervalle

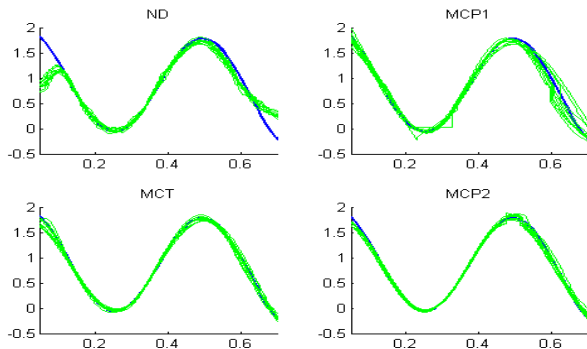
Conclusion

Références

Simulations : Ex1 Régression additive

(X, Y) tels que $Y = s(X) + \varepsilon$, $s(x) = \cos(4\pi x) + \exp(-x^2)$.

Observations : $X \sim \gamma(4, 0.8)$, $\varepsilon \sim \mathcal{N}(0, 0.194)$, $n = 1000$.



Comparaison des risques ($\times 10^3$)

n=60	200	500	2000	Méthode
67.535	15.735	5.22	2.311	ND
14.177	13.374	13.579	13.149	MCT
41.261	13.34	4.808	3.727	MCP1
23.213	5.549	2.059	0.86	MCP2

Introduction

Cadre général

Estimateurs à noyaux

Estimateur avec fenêtre
fixée

Déformation

Estimateur

Risque

Adaptativité

Sélection de la fenêtre

Résultat principal

Cas F_X inconnue

Illustrations

Régression

Censure par intervalle

Conclusion

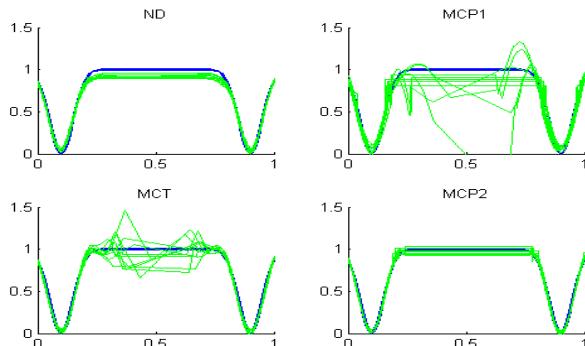
Références

Simulations : Ex1 Régression additive

(X, Y) tels que $Y = s(X) + \varepsilon$,

$$s(x) = -\exp(-200(x - 0.1)^2) - \exp(-200(x - 0.9)^2) + 1.$$

Observations : $X \sim 0.5N(0.05, 0.01) + 0.5N(0.05, 0.95)$, $\varepsilon \sim N(0, 0.18)$, $n = 1000$.



Comparaison des risques ($\times 10^3$)

n=60	200	500	2000	Méthode
120.367	36.804	9.737	3.22	ND
9.701	12.174	31.112	78.242	MCT
61.715	26.986	15.08	8.284	MCP1
52.668	11.009	5.817	1.215	MCP2

Introduction

Cadre général

Estimateurs à noyaux

Estimateur avec fenêtre
fixée

Déformation

Estimateur

Risque

Adaptativité

Sélection de la fenêtre

Résultat principal

Cas F_X inconnue

Illustrations

Régression

Censure par intervalle

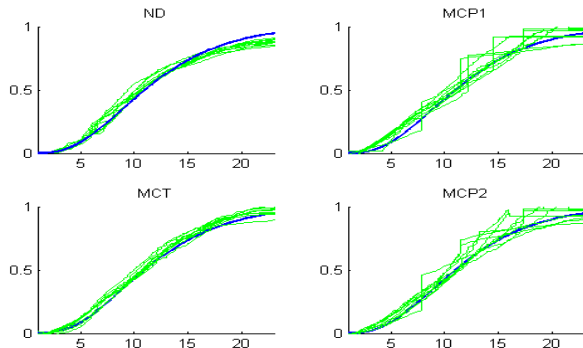
Conclusion

Références

Simulations : Ex4 Censure par intervalle, Cas I

$(X, Y = \mathbf{1}_{Z \leq X})$, estimation de F_Z

Observations : $X \sim \mathcal{E}(0.1)$, $Z \sim \gamma(4, 3)$, $n = 1000$.



Comparaison des risques ($\times 10^2$)

n=60	200	500	2000	Méthode
15.125	4.432	2.428	1.544	ND
25.383	21.553	2.536	1.733	MCT
28.452	14.315	8.815	7.158	MCP1
19.825	11.797	9.738	5.898	MCP2

Introduction

Cadre général

Estimateurs à noyaux

Estimateur avec fenêtre
fixée

Déformation

Estimateur

Risque

Adaptativité

Sélection de la fenêtre

Résultat principal

Cas F_X inconnue

Illustrations

Régression

Censure par intervalle

Conclusion

Références

Conclusion

→ Noyaux déformés, sélection de la fenêtre par méthode de Goldenshluger-Lepski :

▶ **Avantages de la méthode**

- ▶ résolution de problèmes variés d'estimation non paramétrique (régression, estimation pour des données censurées),
- ▶ extension à l'estimation de fonctions de 2 variables : densité conditionnelle d'un couple (X, Y) ,
- ▶ estimateurs ayant une expression simple, sans quotient donc facilement implémentables, avec une seule fenêtre à sélectionner,
- ▶ estimateurs à noyaux adaptatifs, satisfaisants des bornes de risque non asymptotique.

▶ **Inconvénients**

- ▶ hypothèse de régularité portant sur la fonction auxiliaire g et non sur la fonction cible s pour la majoration du biais.
- ▶ substitution de $\hat{\phi}_n$ à ϕ_X naturelle, mais nécessitant des calculs techniques.

Introduction

Cadre général

Estimateurs à noyaux

Estimateur avec fenêtre
fixée

Déformation

Estimateur

Risque

Adaptativité

Sélection de la fenêtre

Résultat principal

Cas F_X inconnue

Illustrations

Régression

Censure par intervalle

Conclusion

Références

Références

- ▶ **Baraud, Y.** Model selection for regression on a random design. *ESAIM Probab. Statist.* 6 (2002), 127–146.
- ▶ **Brunel, E. ; Comte, F.** Cumulative distribution function estimation under interval censoring case 1. *Electron. J. Stat.* 3 (2009), 1–24.
- ▶ **Goldenshluger, A. ; Lepski, O.** Bandwidth selection in kernel density estimation : oracle inequalities and adaptive minimax optimality. *Ann. Statist.*, 39 (2011), no. 3, 1608-1632.
- ▶ **Kerkycharian, G. ; Picard, D.** Regression in random design and warped wavelets. *Bernoulli* 10 (2004), no. 6, 1053–1105.
- ▶ **Stute, W.** Asymptotic normality of nearest neighbor regression function estimates. *Ann. Statist.*, 12 (1984), no. 3, 917–926.
- ▶ **Rozenholc, Y.**, Toolbox Matlab FY3P : Penalized Piecewise Polynomials for Regression (v002).

MERCI !

Introduction

Cadre général

Estimateurs à noyaux

Estimateur avec fenêtre
fixée

Déformation

Estimateur

Risque

Adaptativité

Sélection de la fenêtre

Résultat principal

Cas F_x inconnue

Illustrations

Régression

Censure par intervalle

Conclusion

Références