

ESTIMATION SEMI-PARAMÉTRIQUE
PAR MINIMISATION DE L'ENTROPIE DES RÉSIDUS

Éric Thierry & Luc Pronzato & Éric Wolsztynski

Laboratoire I3S

UMR Nice Sophia-Antipolis

FRANCE

Séminaire CEA 13/12/2005

Quel est le problème ?

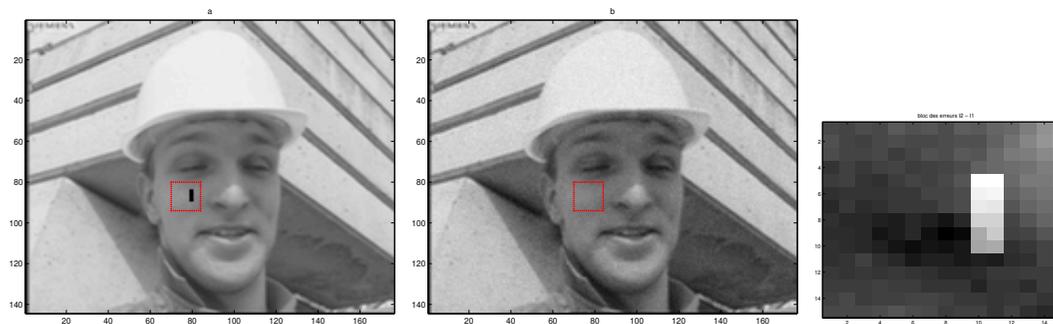
1 - Un problème d'estimation semi-paramétrique pour un modèle de régression, où l'on considère les observations

$$y_i = \eta(\bar{\theta}, x_i) + \varepsilon_i, \quad i = 1, \dots, n$$

avec $\eta(., .)$ connue et (ε_i) **i.i.d., de densité $f \in \mathcal{P}$**

- **Modèle paramétrique:** $\mathcal{P}_{(\theta, \nu)} = \{f(y - \eta(\theta, x), \nu), \theta \in \Theta \quad \nu \in \Omega\}$
 $dim(\Theta)$ et $dim(\Omega)$ finies. Efficacité asymptotique
- **Modèle semi-paramétrique:** $\mathcal{P}_{(\theta, f)} = \{f(y - \eta(\theta, x)); \theta \in \Theta \text{ et } f \in \mathfrak{F}\}$
 $dim(\Theta)$ finie mais $dim(\mathfrak{F})$ infinie. Perte d'efficacité mais pas toujours !

- 2 - Un problème de traitement d'image, où l'on considère 2 copies **a** et **b** bruitées et altérées d'une même image, à intensité lumineuse prêle**
- observations = le seul bloc rouge situé en $\bar{\theta}$ à l'intérieur de **a**
 - on parcourt **b** par blocs; on considère leur différence avec le bloc pris dans **a**



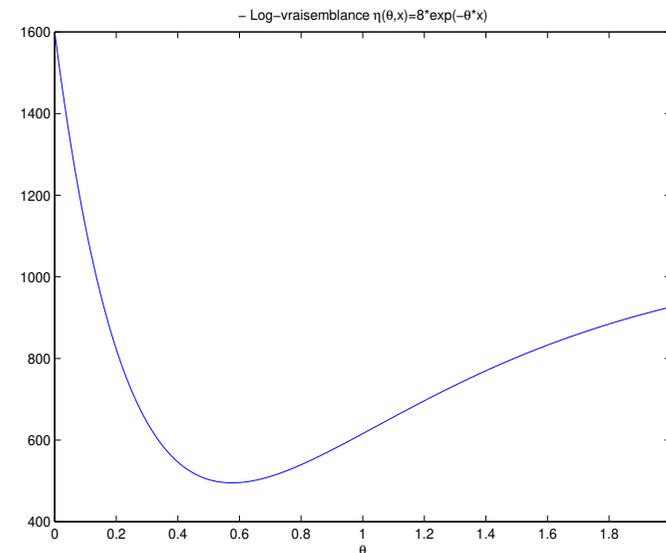
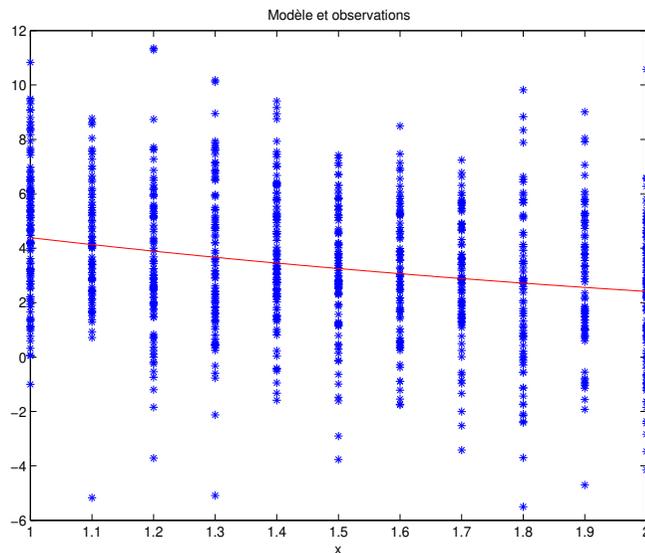
Difficultés: nature du bruit inconnue, images altérées, intensités \neq

Estimation Paramétrique

Modèle statistique: $\mathcal{P} = \{f(e(\theta)), \theta \in \Theta\}$ $e(\theta) = y - \eta(\theta, x)$ et observations i.i.d

Log-vraisemblance: $l(\theta; \mathbf{y} = (y_1, \dots, y_n)) = \sum_{i=1}^n \log f_{\epsilon}(e_i(\theta))$

Maximum de vraisemblance $\hat{\theta}_{MV} = \operatorname{argmax}_{\theta} l(\theta)$



$$\eta(\theta, x) = 8 \exp(-0.6x), \varepsilon_i \sim \mathcal{N}(0, 6)$$

$$-l(\theta, x) \quad n = 100$$

Performances asymptotiques

Consistance asymptotique: $\hat{\theta}_{MV} \xrightarrow[n \rightarrow \infty]{P} \bar{\theta}$

Normalité asymptotique: $\sqrt{n}(\hat{\theta}_{MV} - \bar{\theta}) \rightsquigarrow \mathcal{N}(0, I^{-1}(\bar{\theta}))$

Matrice d'information:

$$I(\vartheta) = E\left[\frac{\partial l(\vartheta, \mathbf{y})}{\partial \vartheta} \frac{\partial l(\vartheta, \mathbf{y})}{\partial \vartheta^t}\right]$$

Si $\vartheta = (\theta, \nu)$ avec θ paramètres d'intérêt et ν paramètres de nuisance

$$I(\vartheta) = \begin{pmatrix} E\left[\frac{\partial l(\theta)}{\partial \theta} \frac{\partial l(\theta)}{\partial \theta^t}\right] & E\left[\frac{\partial l(\theta)}{\partial \theta} \frac{\partial l(\nu)}{\partial \nu^t}\right] \\ E\left[\frac{\partial l(\nu)}{\partial \nu} \frac{\partial l(\theta)}{\partial \theta^t}\right] & E\left[\frac{\partial l(\nu)}{\partial \nu} \frac{\partial l(\nu)}{\partial \nu^t}\right] \end{pmatrix}$$

Adaptativité

$$I^{-1}(\vartheta) = \begin{pmatrix} I^{11} & -I^{11}I_{12}I_{22}^{-1} \\ -I^{22}I_{21}I_{11}^{-1} & I^{22} \end{pmatrix}$$

avec

$$(I^{11})^{-1} = I_{11} - I_{12}I_{22}^{-1}I_{21} \quad (I^{22})^{-1} = I_{22} - I_{21}I_{11}^{-1}I_{12}$$

Si les paramètres de nuisance sont connus alors

$$V[\hat{\theta}] = I_{11}^{-1}$$

Si les paramètres de nuisance sont inconnus alors

$$V[\hat{\theta}] = I^{11} = (I_{11} - I_{12}I_{22}^{-1}I_{21})^{-1}$$

Il y a adaptation si $I_{12} = 0$

Exemple Modèle Position/Echelle

Contexte paramétrique: $f(y; \theta) = \frac{1}{\sigma} f\left(\frac{y-\mu}{\sigma}\right)$

Ici $\theta = \mu, \nu = \sigma$ et $f(\cdot)$ est connue.

Log-vraisemblance: $l(\vartheta) = -\ln(\sigma) + \ln\left(f\left(\frac{y-\mu}{\sigma}\right)\right)$

Anti-diagonale de la matrice d'information:

$$\frac{\partial l(\vartheta)}{\partial \mu} = -\frac{1}{\sigma} \frac{f'\left(\frac{y-\mu}{\sigma}\right)}{f\left(\frac{y-\mu}{\sigma}\right)}$$

$$\frac{\partial l(\vartheta)}{\partial \sigma} = -\frac{1}{\sigma} - \frac{y-\mu}{\sigma^2} \frac{f'\left(\frac{y-\mu}{\sigma}\right)}{f\left(\frac{y-\mu}{\sigma}\right)}$$

$$I_{12}(\vartheta) = E\left[\frac{\partial l(\vartheta)}{\partial \mu} \frac{\partial l(\vartheta)}{\partial \sigma}\right] = \frac{1}{\sigma} \int u \left(\frac{f'(u)}{f(u)}\right)^2 f(u) du$$

Si $f(\cdot)$ est symétrique alors $I_{12}(\vartheta) = 0$, il y a adaptation !

Régression semi-paramétrique

On considère les observations

$$y_i = \eta(\bar{\theta}, x_i) + \varepsilon_i, \quad i = 1, \dots, n$$

- $\bar{\theta}$ vraie valeur inconnue de $\theta \in \Theta \subset \mathbb{R}^p$
- $\eta(\theta, X)$ fonction **connue** de θ et $X \in \mathbb{R}^q$
- (ε_i) **i.i.d., de densité f inconnue symétrique**
- X et ε_i indépendantes ($\text{Prob}[X \in dx] = \mu(dx)$)

Les résidus donnés par

$$e_i(\theta) = y_i - \eta(\theta, x_i)$$

sont i.i.d. de densité $f_e(u) = \int_{\mathcal{X}} f(u - \eta(\bar{\theta}, x) + \eta(\theta, x))\mu(dx)$.

Critère du minimum d'entropie

1- Motivations

- En $\theta = \bar{\theta}$, $e_i(\bar{\theta}) \equiv \varepsilon_i$ et $f_e \equiv f$. La vraisemblance coïncide avec l'entropie de f :

$$\bar{H}_n(\bar{\theta}) = -\frac{1}{n}l(\bar{\theta}) = -\frac{1}{n} \sum_i \log f(\varepsilon_i) \xrightarrow{LLN} H(f) = - \int f \log f$$

- $H(f_e)$ est une mesure de la dispersion de f_e

- **PB : L'entropie est invariante par translation**

\Rightarrow on considère les résidus symétrisés $e^s(\theta) = [e_i(\theta), -e_i(\theta)]_{i=1}^n$

Propriétés du Critère du minimum d'entropie

Critère: $E_X[H(f_{e,X}^s)]$

$$f_{e,X}^s = 0.5(f(u - \eta(\bar{\theta}, X) + \eta(\theta, X)) + f(u + \eta(\bar{\theta}, X) - \eta(\theta, X)))$$

Critère localement convexe, de dérivée nulle en $\theta = \bar{\theta}$

$$\nabla E_X[H(f_{e,X}^s)]|_{\theta=\bar{\theta}} = 0 \quad \nabla^2 E_X[H(f_{e,X}^s)]|_{\theta=\bar{\theta}} = I(\bar{\theta})$$

Critère $H(f_e^s)$

$$f_e^s = \int f_{e,X}^s(u) \mu(dx)$$

Propriétés locales identiques.

- $H(f_e^s) \geq E_X[H(f_{e,X}^s)] \geq H(f)$

- Hyp: modele identifiable: $H(f_e^s) = H(f)$ ssi $\theta = \bar{\theta}$

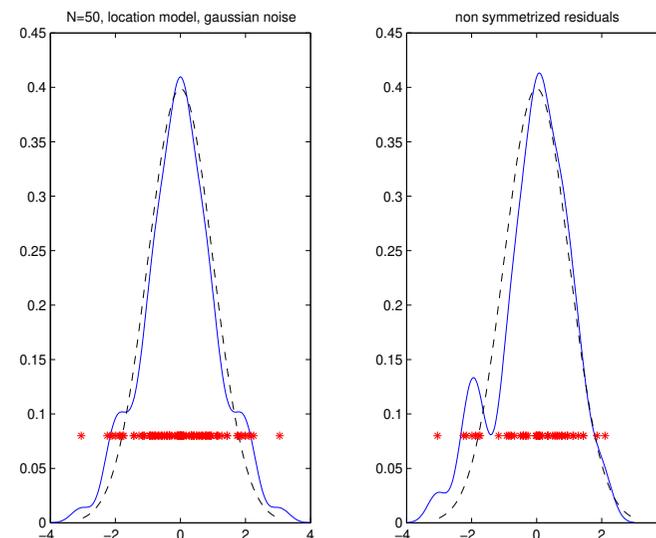
Construction d'estimateurs de l'entropie

Estimateurs par substitution:

estimateur à noyaux \hat{f}_n^θ de f_e à partir des résidus symétrisés:

$$\hat{f}_n^\theta(x) = \frac{1}{2nh_n} \sum_{i=1}^n \left[K\left(\frac{x - e_i(\theta)}{h_n}\right) + K\left(\frac{x + e_i(\theta)}{h_n}\right) \right]$$

avec $h_n \rightarrow 0$, $nh_n \rightarrow \infty$ quand $n \rightarrow \infty$.



- Intégrale tronquée (avec $(A_n) \rightarrow \infty$ suffisamment lentement)

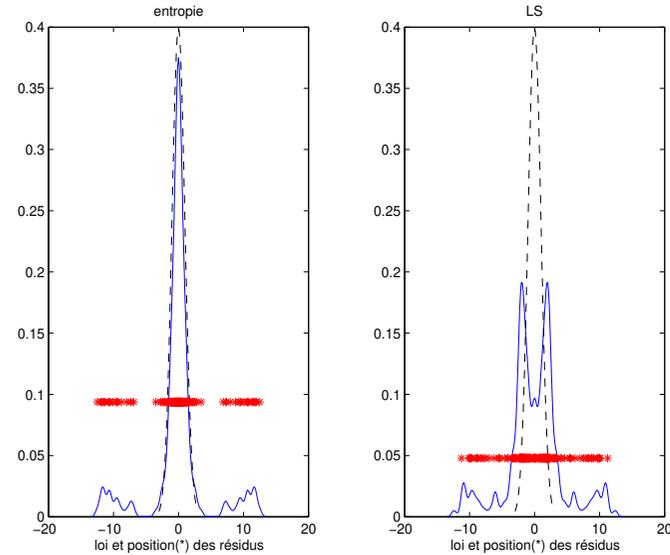
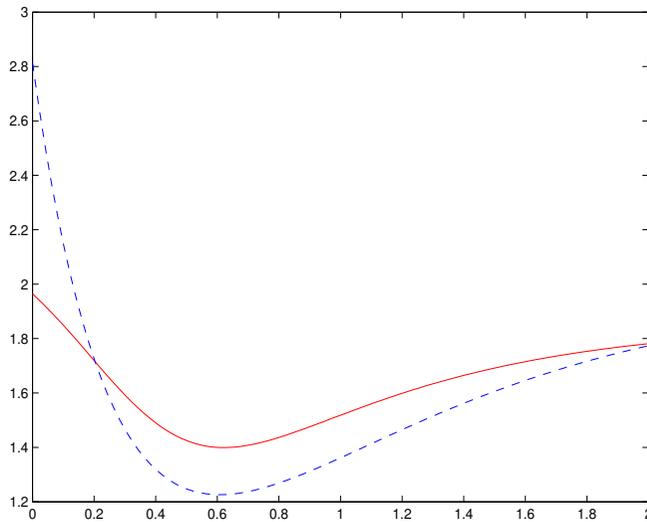
$$\hat{H}_n^1(\theta) = - \int_{-A_n}^{A_n} \hat{f}_n^\theta(x) \log \hat{f}_n^\theta(x) dx,$$

- Estimateur de Ahmad ($0 \leq U_n(x) \leq 1$, $U_n(x) \rightarrow 0$ suffisamment lentement)

$$\hat{H}_n^2(\theta) = - \frac{1}{n} \sum_{i=1}^n \log \hat{f}_n^\theta(e_i(\theta)) U_n(e_i(\theta))$$

Une approche efficace ET robuste ?

L'entropie suggère un critère d'estimation dont l'expression coïncide (localement) avec la vraisemblance, et qui semble être naturellement robuste aux données aberrantes, par sa propriété d'invariance par translation.



$$\hat{H}_n^1(\theta) \text{ et } \bar{H}_n(\theta), n = 100$$

$$\eta(\theta, x) = 8 \exp(-\theta x),$$

$$X_i \in \{1, 1.11\dots, 1.22\dots, \dots, 1.88\dots, 2\},$$

$$\varepsilon_i \sim \mathcal{N}(0, 6)$$

$$\hat{H}_n^1(\theta), f \text{ loi de Laplace}$$

$$\text{modèle de position } \eta(\theta, x) = \theta$$

$$20 \text{ aberrations ajoutées à}$$

$$n = 100 \text{ observations régulières}$$

CAS A)

Répétitions en des points fixés x^1, \dots, x^m

$\xi^j =$ masse en $x^j \Leftrightarrow n_j = n\xi^j$ observations en $x = x^j$ sur un total de n

- (i) Construire un estimateur à noyaux $\hat{f}^{j,\theta}$ en chaque x^j séparément, calculer les entropies correspondantes $H(\hat{f}^{j,\theta})$
- (ii) calculer pour $\hat{H}_n(\theta, x^j) = H(\hat{f}^{j,\theta})$, $j = 1, \dots, m$

$$\hat{\theta}^n = \arg \min_{\theta \in \Theta} \mathbf{E}_X \{ \hat{H}_n(\theta, X) \}$$

Pb: ne peut s'étendre à des expériences + générales

CAS B)

Général, pas de répétitions: ξ_n mesure empirique des X^j

- Mélanger tous les résidus (symétrisés), estimer l'entropie $\hat{H}_n(\theta)$

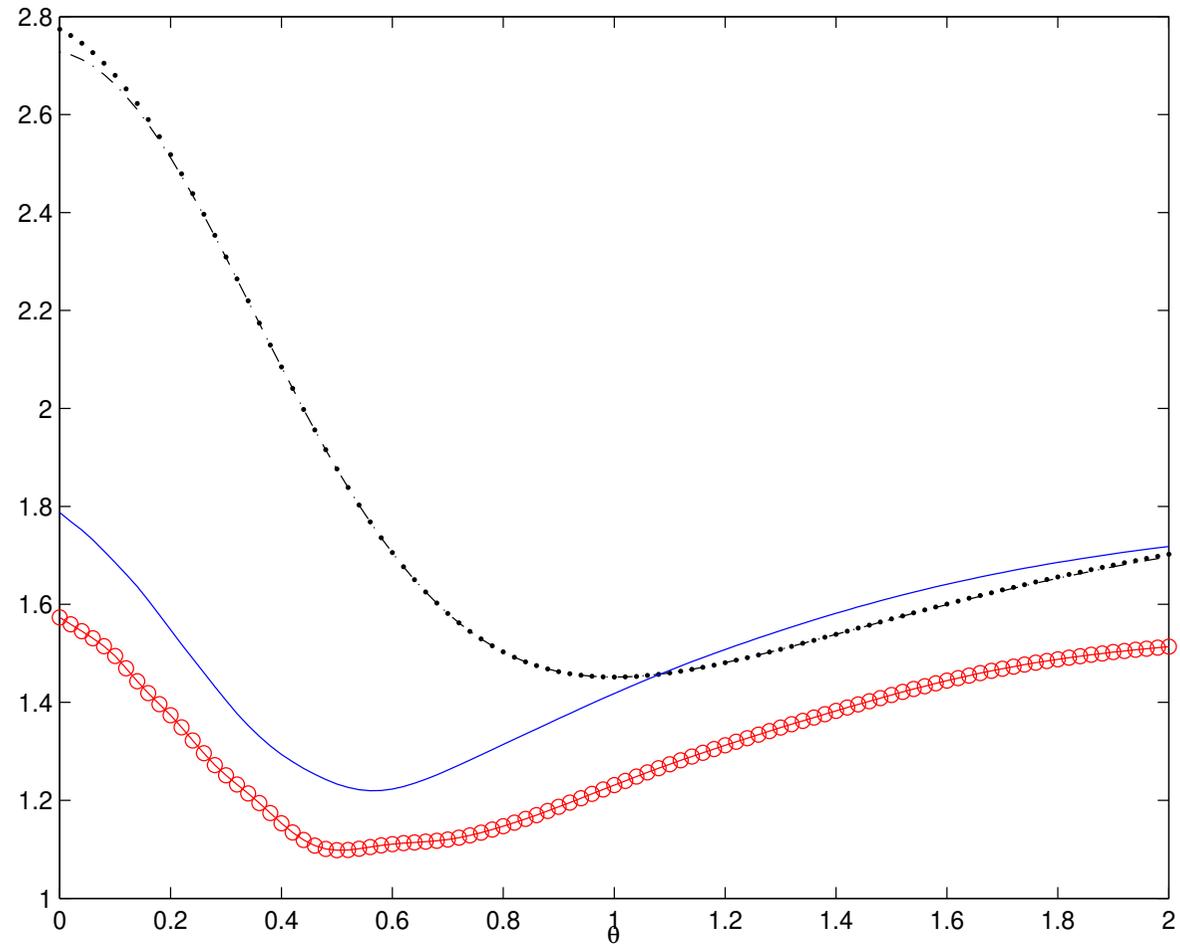
Soit $U =$ variable aléatoire, $f^{j,\theta}$ sa distribution conditionnelle à X

$$\begin{aligned} \mathbb{E}_{\xi_n} \{ \hat{H}_n(\theta, X) \} &= \mathcal{H}(U|X) \\ &= \text{entropie conditionnelle de } U \text{ à } X \text{ donné} \\ &\leq \mathcal{H}(U) = \hat{H}_n(\theta) \end{aligned}$$

Adaptativité ?

$$H(f_e^s), \mathbf{E}_X \{H(f_{e,X}^s)\}, \hat{H}_n(\theta) \text{ et } \hat{H}_n(\theta | X),$$

$$\eta(\theta, x) = \exp(-\theta x), n = 100$$



- $\eta(\theta, x) = \theta_1 \exp(-\theta_2 x)$, $\bar{\theta} = (100, 2)^\top$
 - 10 points (design) régulièrement répartis entre 1 et 2, masses $\xi^j = 1/10$
 - $A_n = \infty$, $U_n(x) \equiv 1$
 - $h_n \rightarrow$ méthode des double noyaux [Berlinet, Devroye, 1994] (à partir des résidus obtenus par un M -estimateur robuste)
- \rightarrow trace et det. de la matrice de cov. empirique \hat{C}_n de $\sqrt{n}(\hat{\theta} - \bar{\theta})$ (100 répétitions, $n = 100$ observations)

$$n = 100, (10^{-3} \text{trace}(\hat{C}_n), 10^{-2} \det(\hat{C}_n))$$

| f | $\mathcal{N}(0, 1)$ | exp | t_3 |
|---|---------------------|-------------------|-------------------|
| optimum ($\hat{C}_n = \mathbf{M}_F^{-1}$) | (6.2, 0.8) | (3.1, 0.2) | (3.1, 0.2) |
| LS | (8.8, 1.2) | (13.6, 3.6) | (9.1, 2.0) |
| ME | (9.2, 1.25) | (3.8, 0.4) | (4.9, 0.4) |

q aberrations ajoutées aux $n = 100$ observations régulières,

$$q \sim \mathcal{N}(10, 4)$$

$$(10^{-3} \text{trace}(\hat{C}_n), 10^{-2} \det(\hat{C}_n)), f = \exp$$

| q | 20 | 40 | 60 | 80 |
|-----------------|---------------|----------------|----------------|---------------|
| LS | (84.25, 42.8) | (146.25, 67.0) | (184.45, 58.2) | (208.7, 58.5) |
| MHD | (4.25, 0.9) | (12.7, 2.25) | (23.4, 4.5) | (56.95, 19.8) |
| ME ₁ | (4.0, 0.5) | (9.8, 1.7) | (6.0, 0.9) | (6.7, 13.6) |
| ME ₂ | (3.95, 0.5) | (8.4, 1.7) | (5.5, 0.9) | (10.9, 39.7) |

ME₁ = estimateur type Shannon

ME₂ = estimateur type Ahmad (empirique)

Propriétés asymptotiques

Modèle de position

$$Y_i = \bar{\theta} + \varepsilon_i, \quad \mathbf{f \text{ symétrique en } 0}$$

Problème avec adaptation [Bickel,82]

Régression non-linéaire

$$Y_i = \bar{\beta} + \eta(\bar{\theta}, X_i) + \varepsilon_i, \quad \mathbf{f \text{ symétrique en } 0}$$

Problème avec adaptation [Manski,84]

Résultats pour le modèle de position

- Adaptativité d'une approche en 2 étapes:

Séparation des données

$$\begin{aligned}(Y_1, \dots, Y_m) &\Rightarrow \hat{\theta}_1^m \Rightarrow e_i(\hat{\theta}_1^m) \text{ and } \hat{f}_m \\(Y_{m+1}, \dots, Y_n) &\Rightarrow \hat{H}_n(\theta) = -\frac{1}{n-m} \sum_{i=m+1}^n \log \hat{f}_m(e_i(\theta))\end{aligned}$$

Un pas de Newton-Raphson $\hat{\theta}_1^n = \hat{\theta}_1^n - [M_n(\hat{\theta}_1^n)]^{-1} \nabla \hat{H}_n(\hat{\theta}_1^n)$

$$\text{avec } \left\{ \begin{array}{l} \rho_m = (\hat{f}_m)' / \hat{f}_m U_n(u) \\ M_n(\theta) \text{ approximation de } \nabla^2 \hat{H}_n(\theta) \\ \nabla \hat{H}_n(\hat{\theta}_1^n) = -\frac{1}{n-m} \sum_{i=m+1}^n \rho_m(e_i(\theta)) \nabla \eta(\theta, x_i) \end{array} \right.$$

Si $h_m \rightarrow 0$, $a_m \rightarrow \infty$, $b_m \rightarrow 0$, $c_m \rightarrow \infty$, et $m^{-1} a_m h_m^{-3} \rightarrow 0$, $h_m c_m \rightarrow 0$,

$m \rightarrow \infty$, $m/n \rightarrow 0$ alors **Adaptation [Bickel,82]**

• Propriétés asymptotiques d'une approche directe:

$$\hat{\theta}^n = \arg \min_{\theta \in \Theta} -\frac{1}{n} \sum_{i=1}^n \log \hat{f}_{n,i}^{\theta}(e_i(\theta)) U_n(e_i(\theta))$$

$$\hat{f}_{n,i}^{\theta}(x) = \frac{1}{2(n-1)h_n} \sum_{j=1, j \neq i} \left[K\left(\frac{x-e_j(\theta)}{h_n}\right) + K\left(\frac{x+e_j(\theta)}{h_n}\right) \right]$$

a) $\hat{H}_n(\theta) \xrightarrow{\theta, p} H(\theta)$, avec $H(\bar{\theta}) < H(\theta) \forall \theta \neq \bar{\theta} \Rightarrow \hat{\theta}^n \xrightarrow{p} \bar{\theta}$

b) $\nabla^2 \hat{H}_n(\theta) \xrightarrow{\theta, p} \nabla^2 H(\theta) \Rightarrow \nabla^2 \hat{H}_n(\hat{\theta}^n) \xrightarrow{p} \mathbf{M}_2 = \nabla^2 H(\bar{\theta}) \succ 0$

c) si $\sqrt{n} \nabla \hat{H}_n(\bar{\theta}) \xrightarrow{d} \mathcal{N}(\mathbf{0}, \mathbf{M}_1)$, alors $\sqrt{n}(\hat{\theta}^n - \bar{\theta}) \xrightarrow{d} \mathcal{N}(\mathbf{0}, \mathbf{M}_2^{-1} \mathbf{M}_1 \mathbf{M}_2^{-1})$.

\Rightarrow Si $[\mathbf{M}_2^{-1} \mathbf{M}_1 \mathbf{M}_2^{-1}] =$ inverse de la matrice d'info de Fisher] alors **adaptativité** !

Densités multivariées

- **Problème des méthodes à noyaux en dimension supérieure:**
fléau de la dimension \Rightarrow sélection du lissage h difficile ($\dim d > 3$)
- **Pas de lissage optimal** pour des noyaux variables, calculs lourds car répétés (ou alternatives contraignantes).

Une alternative: la méthode des k^e plus proches voisins (kPPV)

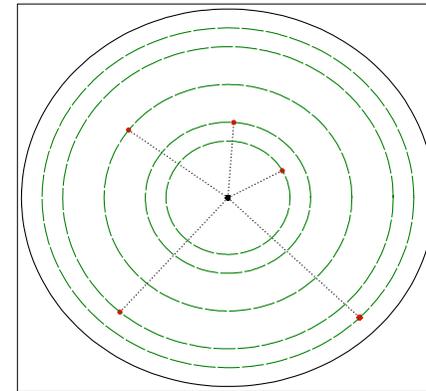
- Noyaux (variables) uniformes multidimensionnels \Leftrightarrow **approche robuste**
- Pas de fonction de lissage h
- Toutes dimensions
- Discontinuités de gradient du critère \rightarrow problèmes de recherche exhaustive ($\theta \in$ **ensemble fini**), e.g. traitement d'images

[Goria, Leonenko, Mergel, Novi Inverardi, 2005]

- L'estimateur par kPPV minimise (données de dimension d):

$$H_{k,n}(\theta) = d \log \bar{\rho}_k(\theta) + \log(n-1) - \psi(k) + \log c_1(d)$$

- $\bar{\rho}_k(\theta) = \left(\prod_{i=1}^n \rho_{i,k}(\theta) \right)^{1/n}$
- $\psi(k) = \Gamma'(k)/\Gamma(k)$ fonction digamma
- $c_1(d) = 2\Pi^{d/2}/(d\Gamma(d/2))$ vol. boule unité de \mathbb{R}^d
- $k/n \rightarrow 0$, et $k \rightarrow \infty, n \rightarrow \infty$ (ex $k = \sqrt{n}$)



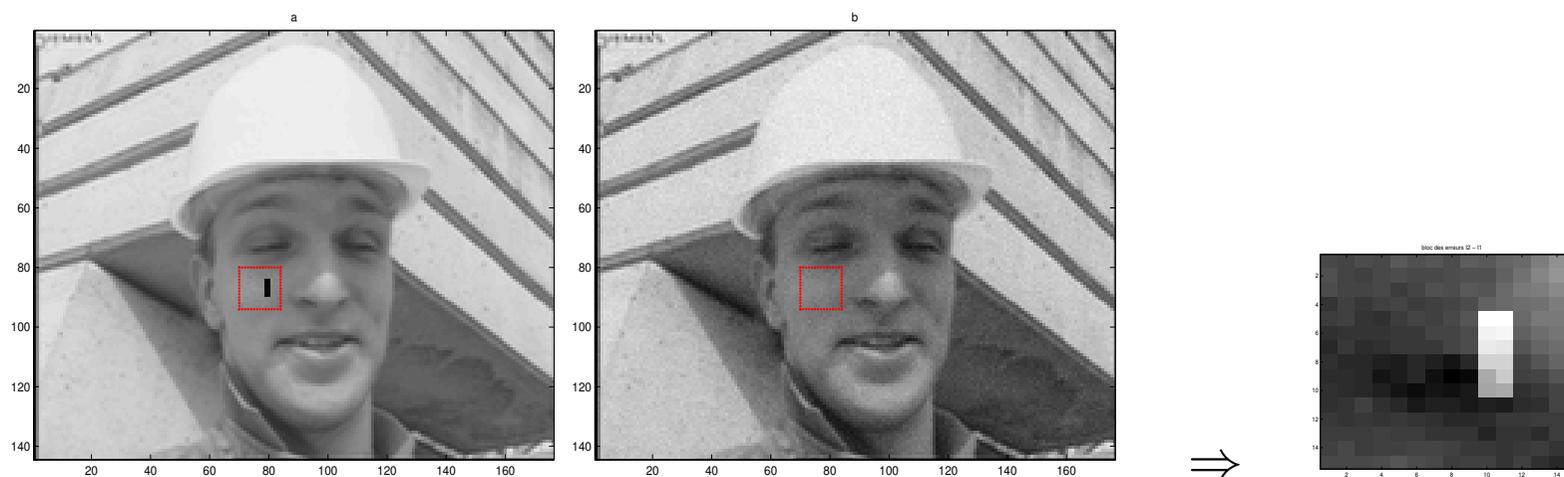
- Remarque: correspond à un estimateur de l'entropie par substitution

$$\tilde{H}_n(\theta) = -\frac{1}{n} \sum_{i=1}^n \log \left(\frac{k}{n (\rho_k(x))^d c_1(d)} \right) + b(n, k)$$

(estimateur de densité par kPPV converge p.s.)

Application en traitement d'images

- 2 copies **a** et **b** bruitées et altérées d'une même image, d'intensités lumineuses \neq
- observations = le seul bloc rouge (optimal) pris à l'intérieur de **a**
- $\bar{\theta} \Leftrightarrow$ coordonnées de ce bloc dans l'image **a**
- on parcourt **b** par blocs; on considère leur différence avec le bloc pris dans **a**



Difficultés: nature du bruit inconnue, images altérées par endroits, intensités \neq

Intérêts de l'entropie:

- entropie minimale \Leftrightarrow taux de compression maximal
- entropie invariante et \neq intensités lumineuses \Rightarrow **pas de symétrisation**

Exemple 1: image N&B ($\dim(e_i) = 1$)

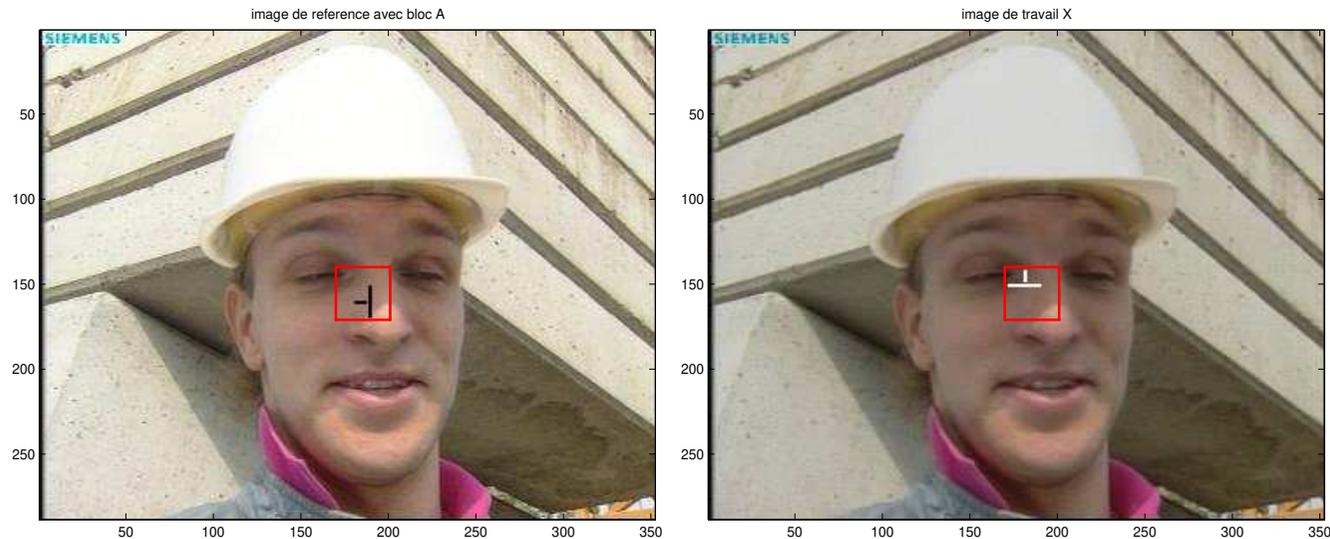
- $\varepsilon \sim \mathcal{N}(0, 10)$
- différence d'intensités $I_1 - I_2 = 10$
- blocs 15×15 , 2×6 aberrations (5.33%)

moyennes des estimées sur 100 répétitions

| | $\bar{\theta}$ | MEs | kPPV | Hellinger | MVA | MC |
|---------|----------------|-------|-------|-----------|-------|-------|
| sym | 80 | 81.11 | 80.00 | 74.64 | 82.15 | 86.29 |
| | 70 | 64.98 | 69.99 | 84.48 | 64.45 | 65.87 |
| non sym | 80 | 80.04 | 80.01 | 74.09 | 81.89 | 86.34 |
| | 70 | 71.47 | 70.09 | 86.00 | 64.43 | 65.84 |

Exemple 2: image couleur ($\dim(e_i) = 3$)

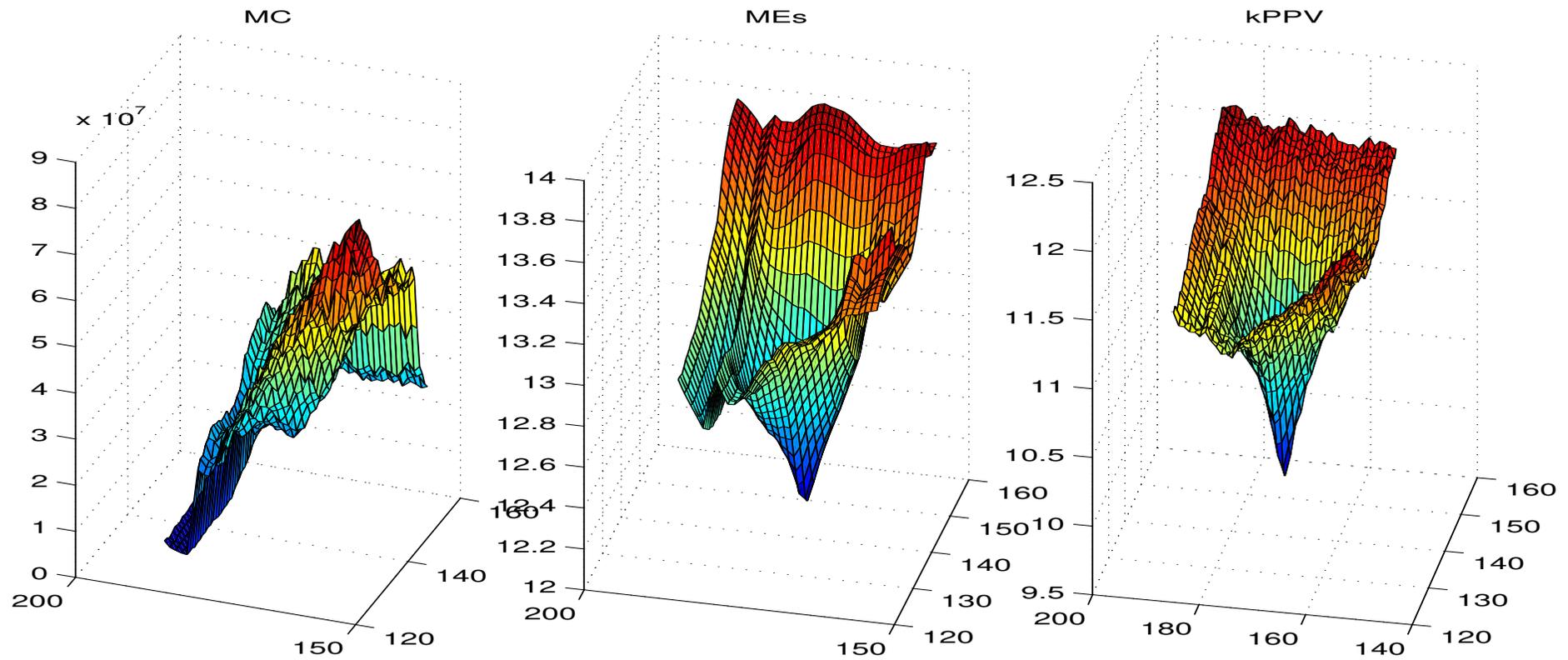
- $\varepsilon \sim \mathcal{N}(0, 10)$
- différence d'intensités $I_1 - I_2 = 40$
- blocs 32×32 , $2 \times 9 + 7 \times 2$ aberrations (3.1%)



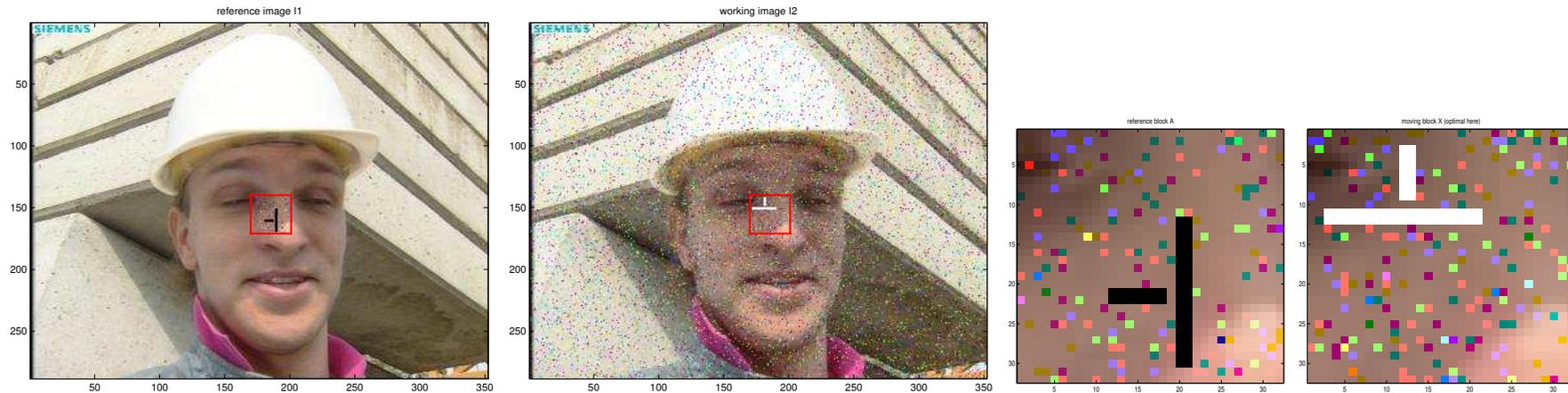
Estimées pour 1 expérience

| $\bar{\theta}$ | MC | kPPV | MEs |
|----------------|-----|------|-----|
| 140 | 152 | 140 | 140 |
| 170 | 157 | 170 | 170 |

Critères vs θ (image couleur, bruit gaussien $\mathcal{N}(0, 10)$ avec aberrations)



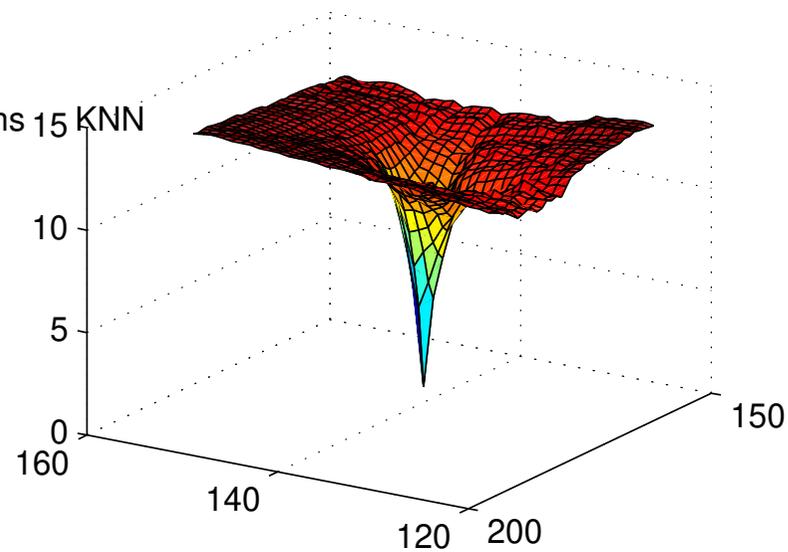
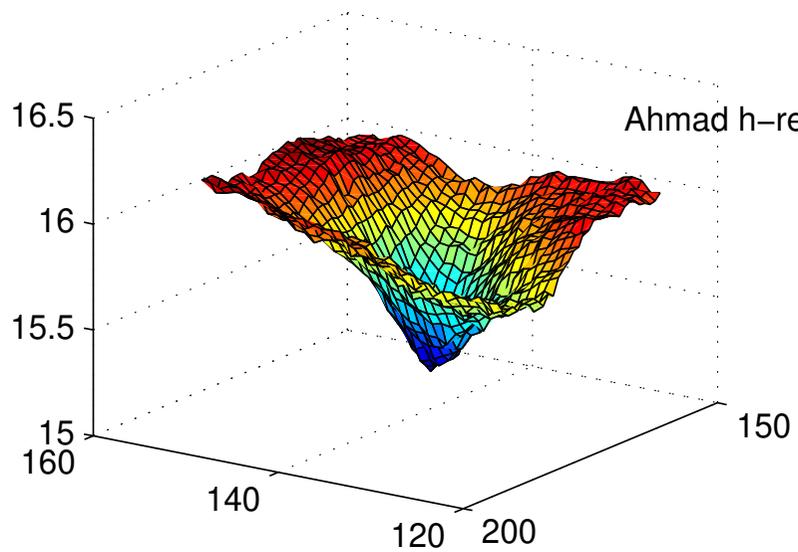
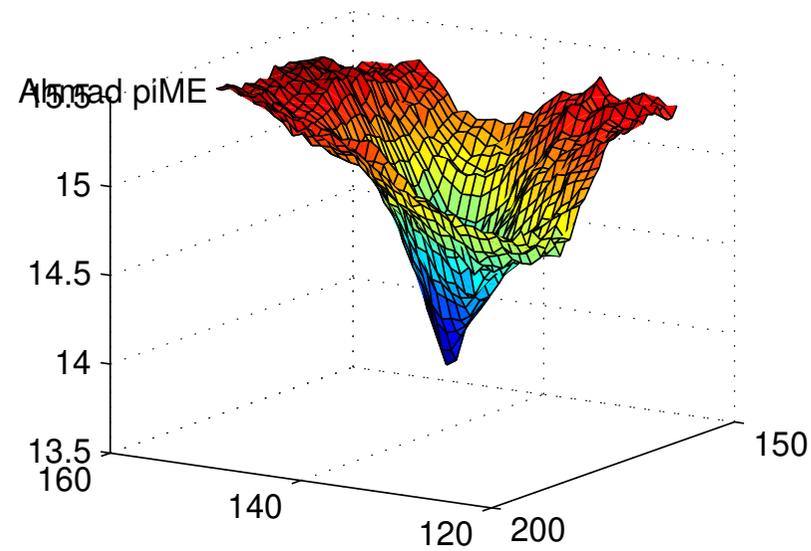
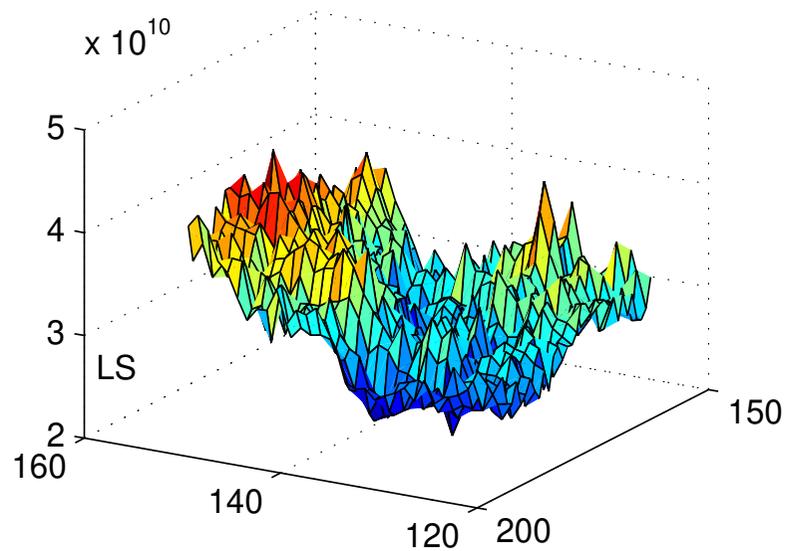
Exemple 3: couleur, ajout de bruit *poivre et sel* sur l'exemple 2
 (valeurs de pixels ← max ou min, pixels aberrants uniformément répartis).



Estimées pour 1 expérience

| $\bar{\theta}$ | MC | kPPV | MEs |
|----------------|-----|------|-----|
| 140 | 152 | 140 | 140 |
| 170 | 161 | 170 | 170 |

Bruit poivre et sel sur bruit gaussien, couleur



Exemple 4: bruit non-symétrique lognormal, var. 10, \neq d'intensités $I_1 - I_2 = 10$

- Image N&B

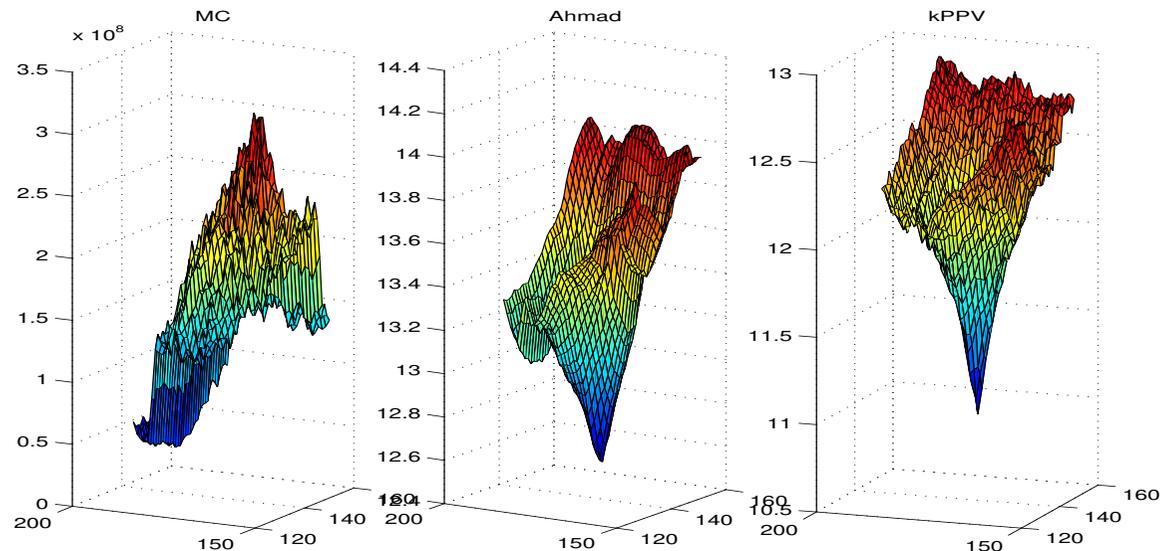
Moyennes des estimées sur 100 répétitions

| | $\bar{\theta}$ | MEs | kPPV | Hellinger | M-estim. | MC |
|---------|----------------|-------|-------|-----------|----------|-------|
| non sym | 80 | 80.54 | 80.01 | 74.88 | 81.61 | 85.86 |
| | 70 | 71.36 | 70.27 | 85.73 | 64.33 | 65.83 |

- Image couleur

Estimées pour 1 expérience

| $\bar{\theta}$ | MC | kPPV | MEs |
|----------------|-----|------|-----|
| 140 | 152 | 140 | 140 |
| 170 | 167 | 170 | 170 |



Perspectives

- Applications pratiques en image (compression vidéo, recalibration, ...): premiers résultats encourageants en compression vidéo
- Applications en traitement du signal en dim supérieure, modèles AR, ...
- Etude des propriétés d'alternatives à l'estimateur par substitution (dont kNN)
- Le MV coïncide avec la divergence de Kullback-Leibler $KL(f, f_e^\theta)$, dont on tire l'entropie de Shannon.
D'autres divergences permettent-elles l'utilisation d'autres formes d'entropies (ex Renyi) ?

Quelques références

- [BeirlantDGM97] J. Beirlant, E.J. Dudewicz, L. Györfi , and E.C. van der Meulen, *Nonparametric entropy estimation; an overview*, Intern. J. Math. Stat. Sci., 6(1), p.17–39, 1997.
- [BerlinetD94] A. Berlinet and L. Devroye, *A comparison of kernel density estimates*, Publications de l'institut de statistique de l'Université de Paris, 38, p.3-59, 1994.
- [Bickel82] P.J. Bickel, *On adaptive estimation*, Annals of Statistics, 10, p.647–671, 1982.
- [GoriaLMNI05] M.N. Goria, N.N. Leonenko, V.V. Mergel and P.L. Novi Inverardi, *A new class of random vector entropy estimators and its applications in testing statistical hypotheses*, Journal of Nonparametric Statistics, 2005.
- [Manski84] C. Manski, *Adaptive estimation of nonlinear regression models*, Econometric Reviews, 3(2), p.145-194, 1984.
- [Scott92] D.W. Scott. *Multivariate Density Estimation: Theory, Practice, and Visualization*. Wiley, 1992.
- [WTP05] E. Wolsztynski, E. Thierry, and L. Pronzato, *Minimum-entropy estimation in semi-parametric models*, Signal Processing, Special Issue on Information Theoretic Signal Processing, 2005.