Contributions à l'inférence statistique semi- et non-paramétrique

Stéphane Girard

SMS/LMC-IMAG, Université Grenoble 1

Habilitation à Diriger des Recherches

Thèmes abordés

1. Estimation de quantiles extrêmes.
2. Estimation de frontière (de support).
3. Réduction de dimension et analyse d'images.
4. Estimation de courbes de référence.
5. Perpectives.

1. Estimation de quantiles extrêmes

En collaboration avec :

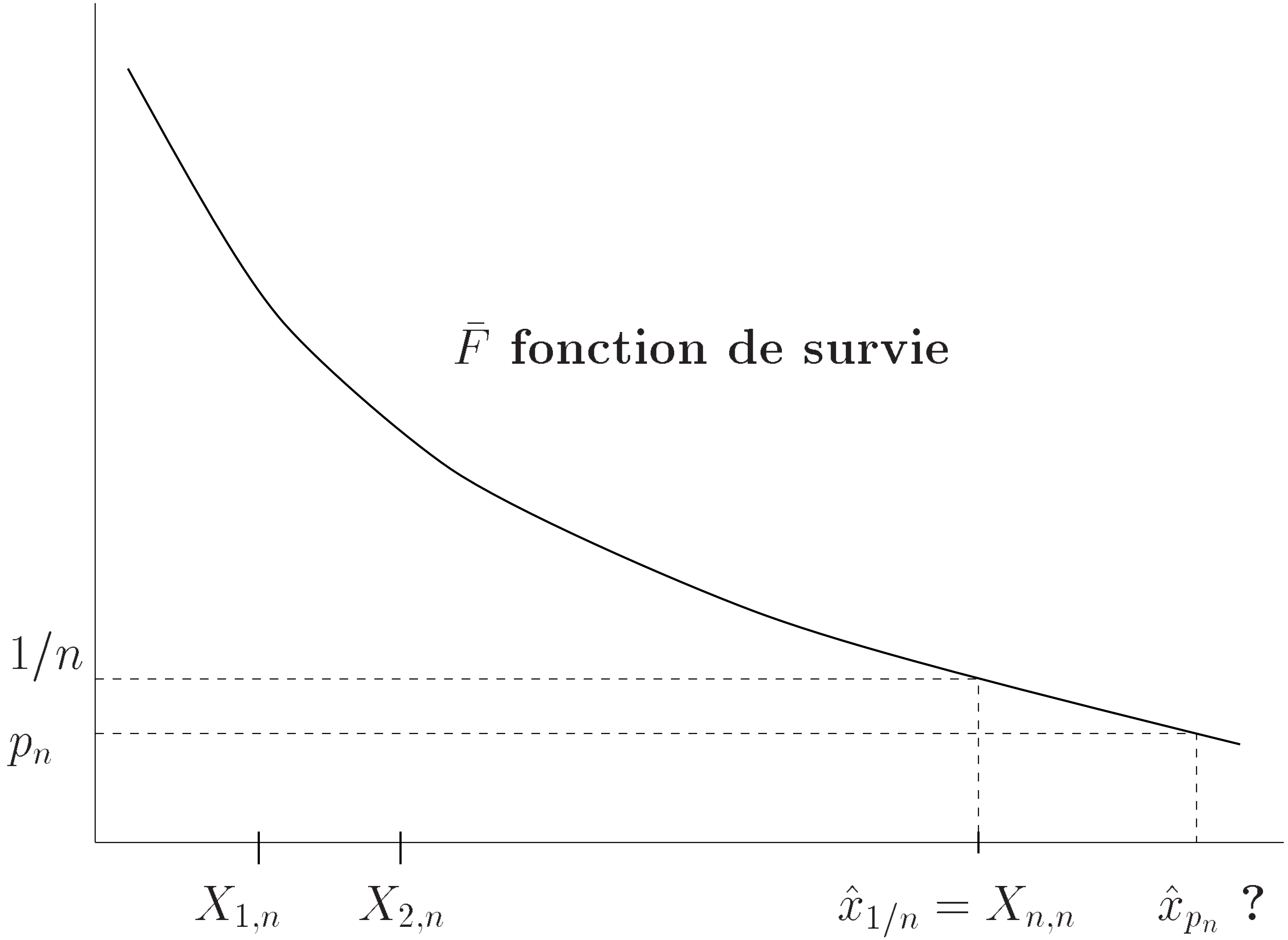
- Jean Diebolt (CNRS, Université Marne-la-Vallée),
- Jérôme Ecarnot (INRIA Rhône-Alpes),
- Laurent Gardes (INPG),
- Mhamed El-Aroui (ISG Tunis),
- Myriam Garrido (Université Toulouse III),
- Dominique Lagrange (EDF).

Dans les cadres suivants :

- Contrat de recherche EDF/INRIA Rhône-Alpes (1997–2002),
- Thèse de Myriam Garrido (1999–2002).

1.1. Le problème

- n variables aléatoires $X_1, \ldots, X_n$ indépendantes de même loi que X.
- Quantile x_{p_n} d'ordre p_n de X : $P(X > x_{p_n}) = \bar{F}(x_{p_n}) = 1 - F(x_{p_n}) = p_n$.
- Si $p_n < 1/n$, alors ce quantile est dit extrême.
  Il est "en général" supérieur à l'observation maximale $X_{n,n} = \max\{X_1, \ldots, X_n\}$.
  Si $np_n \to 0$, alors $P(x_{p_n} > X_{n,n}) \to 1$ quand $n \to \infty$.
- Nécessité de méthodes d'extrapolation au delà de l'observation maximale sans hypothèse paramétrique sur $\bar{F}$.


$\bar{F}$ **fonction de survie**
$1/n$
p_n
$X_{1,n}$
$X_{2,n}$
$\hat{x}_{1/n} = X_{n,n}$
$\hat{x}_{p_n}$?

1.2. Théorie des valeurs extrêmes

- *Théorème des valeurs extrêmes.* (Gnedenko, 1943)
  Il existe $\xi \in \mathbb{R}$ et $(\alpha_n)_{n\geq 1}$ et $(\beta_n)_{n\geq 1}$ tels que

$$\lim_{n\to\infty} P(\beta_n^{-1}(X_{n,n} - \alpha_n) \leq x) = \lim_{n\to\infty} F^n(\alpha_n + \beta_n x) = H_\xi(x),$$

  où H_ξ est la fonction de répartition de la loi des valeurs extrêmes (EVD) :

$$H_\xi(x) = \begin{cases} \exp\left[-(1+\xi x)_+^{-1/\xi}\right] & \text{si } \xi \neq 0, \quad \text{où } y_+ = \max(0, y). \\ \exp(-\exp(-x)) & \text{si } \xi = 0. \end{cases}$$

  ξ est l'indice des valeurs extrêmes.

- *Domaines d'attraction.*
  - Si $\xi < 0$, $F \in$ DA(Weibull), point terminal fini.
  - Si $\xi = 0$, $F \in$ DA(Gumbel), queue de type exponentiel.
  - Si $\xi > 0$, $F \in$ DA(Fréchet), queue lourde de type puissance.

- *Théorème de Pickands.* (Pickands, 1975)
  Soit Y l'excès de X au delà du seuil u défini par $Y = X - u$ quand $X > u$.
  Sa fonction de survie est donnée par $\bar{F}_u(x) = P(X - u > x | X > u) = \bar{F}(u + x)/\bar{F}(u)$.
  Il existe une fonction σ telle que

$$\lim_{u \to x_F} \sup_{0 < x < x_F - u} \left| \bar{F}_u(x) - \bar{F}^{\text{GPD}}_{\xi,\sigma(u)}(x) \right| = 0,$$

où $\bar{F}^{\text{GPD}}_{\xi,\sigma}$ est la fonction de survie de la loi de Pareto généralisée (GPD) :

$$\bar{F}^{\text{GPD}}_{\xi,\sigma}(x) = \begin{cases} (1 + \xi x/\sigma)^{-1/\xi} & \text{si } \xi \neq 0, \\ \exp(-x/\sigma) & \text{si } \xi = 0, \end{cases}$$

définie pour $x \geq 0$ si $\xi \geq 0$ et $0 \leq x \leq -\sigma/\xi$ sinon.

1.3. Exemple d'application à l'estimation de quantiles extrêmes

- *La méthode des excès.* – Peaks Other Threshold (POT)
  Approximation de $\bar{F}_u$ basée sur le théorème de Pickands, puis estimation de (u, σ, ξ).
  - $u = u_n$ est un quantile classique défini par $\bar{F}(u_n) = c_n$, avec $1/n \le c_n < 1$ et estimé par $\hat{u}_n = X_{n-k_n+1,n}$ où $k_n = nc_n$.
  - $\sigma(u_n)$ et ξ sont estimés à partir des excès ordonnés $\{X_{n-i+1,n} - X_{n-k_n+1,n},\ i = 1, \dots, k_n - 1\}$ par $\hat{\sigma}_n$ et $\hat{\xi}_n$.

$$\hat{x}^{\text{POT}}_{p_n} = \hat{u}_n + \left(\bar{F}^{\text{GPD}}_{\hat{\xi}_n, \hat{\sigma}_n}\right)^{-1}(p_n/c_n) = X_{n-k_n+1,n} - \frac{\hat{\sigma}_n}{\hat{\xi}_n}\left(1 - (c_n/p_n)^{\hat{\xi}_n}\right).$$

- *Cas particulier : la méthode ET.* On suppose $F \in \text{DA}(\text{Gumbel})$ et on choisit $\hat{\xi}_n = 0$.

$$\hat{x}^{\text{ET}}_{p_n} = X_{n-k_n+1,n} + \hat{\sigma}_n \log(c_n/p_n),\ \hat{\sigma}_n = \frac{1}{k_n - 1}\sum_{i=1}^{k_n-1}(X_{n-i+1,n} - X_{n-k_n+1,n}).$$

(Breiman *et al*, 1990)

1.4. Nos contributions

- *Estimation de l'indice des valeurs extrêmes ξ.*
  - Introduction d'estimateurs bayésiens $\hat{\xi}^{\text{BAYES}}$ pour le cas $F \in \text{DA(Fréchet)}$ $(\xi > 0)$.
  - Introduction d'un nouvel estimateur $\hat{\xi}_n^{\text{G}}$ pour le cas $\xi \in \mathbb{R}$.
- *Estimation des quantiles extrêmes.*
  - Etude d'un estimateur existant : normalité asymptotique de $\hat{x}_{p_n}^{\text{ET}}$.
  - Introduction d'un nouvel estimateur pour le cas F à queue de type Weibull (sous cas de $F \in \text{DA(Gumbel)}$).
  - Introduction des estimateurs déduits de $\hat{\xi}^{\text{BAYES}}$.
- *Introduction de tests d'adéquation à la queue de distribution.*
- *Développement du logiciel* EXTREMES.

1.5. Un nouvel estimateur des quantiles extrêmes pour les lois à queue de type Weibull

- *Définition.* Lois à queue de type Weibull :

$$\bar{F}(x) = \exp(-H(x)),\ H^{-1}(t) = t^{\theta}\ell(t),$$

où ℓ est une fonction à variations lentes, $\ell(\lambda x)/\ell(x) \to 1$, $x \to \infty$, $\lambda > 0$, et θ est l'indice de queue de Weibull.

- *Principe.* Similaire à la méthode des excès. L'approximation

$$\frac{(\bar{F})^{-1}(t)}{(\bar{F})^{-1}(s)} = \frac{H^{-1}(\log(1/t))}{H^{-1}(\log(1/s))} \simeq \left(\frac{\log(1/t)}{\log(1/s)}\right)^{\theta},$$

permet de se ramener à l'estimation d'un quantile classique u_n et de l'indice de queue θ

$$\hat{x}_{p_n} = X_{n-k_n+1,n}\left(\frac{\log(1/p_n)}{\log(1/c_n)}\right)^{\hat{\theta}_n},$$

- *Exemples d'estimateurs de l'indice de queue de Weibull.*

$$\hat{\theta}_n^{\text{G}} = \sum_{i=1}^{k_n-1} (\log(X_{n-i+1,n}) - \log(X_{n-k_n+1,n})) \Big/ \sum_{i=1}^{k_n-1} (\log\log(n/i) - \log\log(n/k_n)) \,,$$

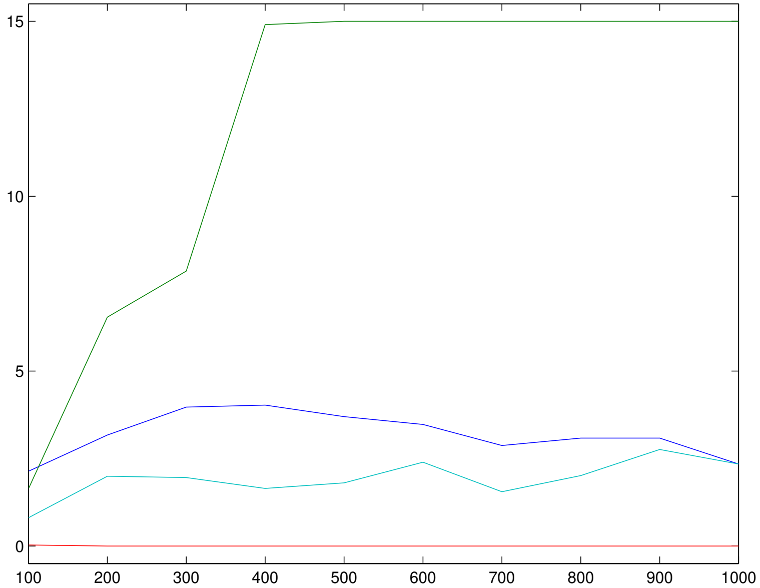
$$\hat{\theta}_n^{\text{BBTV}} = \frac{\log(n/k_n)}{X_{n-k_n+1,n}} \frac{1}{k_n - 1} \sum_{i=1}^{k_n-1} (X_{n-i+1,n} - X_{n-k_n+1,n}) \ \text{(Beirlant } \textit{et al}\text{, 1995).}$$

- *Normalité asymptotique.* Sous certaines conditions, on montre que si $\hat{\theta}_n = \hat{\theta}_n^{\text{G}}$ ou $\hat{\theta}_n = \hat{\theta}_n^{\text{BBTV}}$ alors

$$\frac{\log(1/c_n) k_n^{1/2}}{\log(c_n/p_n)} \left(\frac{\hat{x}_{p_n}}{x_{p_n}} - 1 \right) \xrightarrow{d} \mathcal{N}(0, \theta^2).$$

- *Simulations.* Le choix $\hat{\theta}_n = \hat{\theta}_n^{\text{G}}$ donne de meilleurs résultats que $\hat{\theta}_n = \hat{\theta}_n^{\text{BBTV}}$.

- *Exemple.* Pouvoir d'extrapolation (El-Aroui & Diebolt, 2002) en fonction de n pour une loi de Weibull $\mathcal{W}(0.7, 1)$ et une loi normale $\mathcal{N}(4, 1)$.



vert : estimateur précédent avec $\hat{\theta}_n = \hat{\theta}_n^{\mathrm{G}}$,
rouge : estimateur précédent avec $\hat{\theta}_n = \hat{\theta}_n^{\mathrm{BBTV}}$,
bleu : $\hat{x}_{p_n}^{\mathrm{ET}}$ estimateur ET,
cyan : $\hat{x}_{p_n}^{\mathrm{POT}}$ estimateur POT.

1.6. Perspectives

- *Estimation de l'indice des valeurs extrêmes ξ.*
  - Correction de biais pour l'estimateur $\hat{\xi}_n^{\text{G}}$ pour le cas $\xi \in \mathbb{R}$.
  - Choix adaptatif du paramètre k_n.
- *Estimation des quantiles extrêmes.*
  - Même travail pour l'estimateur dédié au cas F à queue de type Weibull.
  - Estimateur des quantiles extrêmes déduit de $\hat{\xi}_n^{\text{G}}$.
- *Amélioration des tests d'adéquation à la queue de distribution.*
- *Poursuite du développement du logiciel* EXTREMES.
- *Vers l'estimation de frontière, de quantiles conditionnels extrêmes ...*

2. Estimation de frontière

En collaboration avec :

- Guillaume Bouchard (INRIA Rhône-Alpes),
- Jean Geffroy,
- Anatoli Iouditski (Université Grenoble I),
- Pierre Jacob, Ludovic Menneteau (Université Montpellier II),
- Alexander Nazin (Institute of Control Sciences, Moscou).

Dans les cadres suivants :

- Thèse de Laurent Gardes (2000–2003),
- IAP (Interuniversity Attraction Pole) Network (2002–2006).

2.1. Le problème

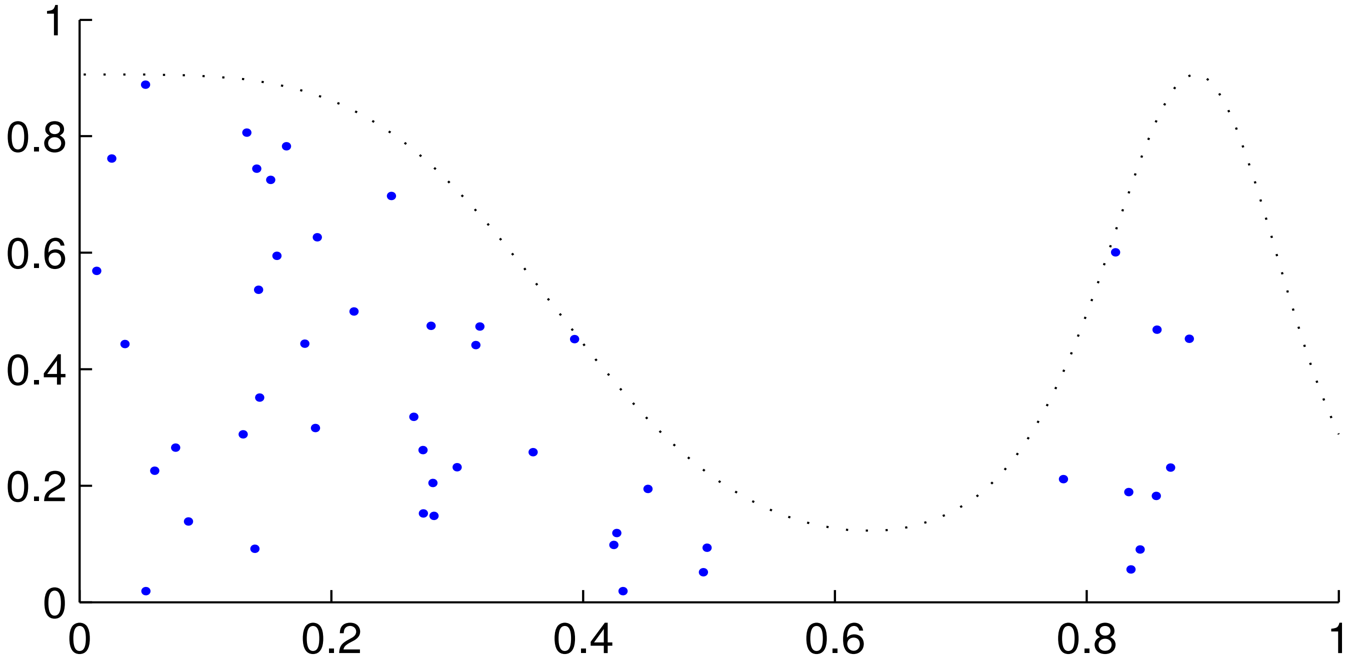
- Estimation d'un ensemble D borné de $\mathbb{R}^{d+1}$, $d \geq 1$, de la forme

$$D = \{(x, y),\ x \in E,\ 0 \leq y \leq f(x)\}$$

  où $E \subset \mathbb{R}^d$ est connu et f est une fonction de E dans $\mathbb{R}^+$ inconnue.

- Données : ensemble Σ_n de points (X_i, Y_i), $i \geq 1$ disposés aléatoirement dans D.
- Le problème se réduit à l'estimation de f.
- De nombreuses déclinaisons selon :
  - les hypothèses faites sur f (monotonicité, ...),
  - les hypothèses faites sur Σ_n (échantillon, processus de Poisson, uniformité ou non de la distribution des points sur D, ...),
  - la dimension d.

Exemple : Σ_n échantillon de $n = 50$ points répartis uniformément sur D, $E = [0, 1]$, $d = 1$.



2.2. L'estimateur de Geffroy et ses extensions

- *Hypothèses :* Σ_n échantillon de n points de densité ϕ sur D, $E = [0, 1]$, $d = 1$.
- *Estimateur constant par morceaux :* (Geffroy, 1964)
  $I_{n,r}$, $r = 1, \dots, k_n$, partition de l'intervalle $[0, 1]$, $Y^{\star}_{n,r} = \max\{Y_i : X_i \in I_{n,r}\}$.

$$\hat{f}^0_n(x) = \sum_{r=1}^{k_n} \mathbb{1}\{x \in I_{n,r}\} Y^{\star}_{n,r}.$$

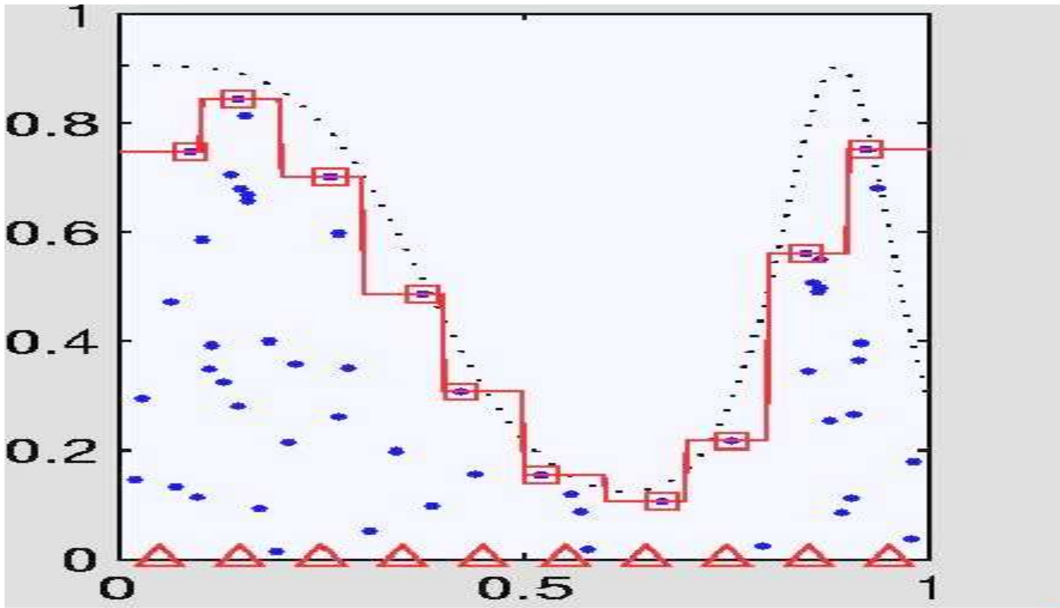
- *Estimateur polynômial par morceaux* (Korostelev & Tsybakov, 1993)

$$\hat{f}^\theta_n(x) = \sum_{r=1}^{k_n} \mathbb{1}\{x \in I_{n,r}\} P^\theta_{n,r}(x).$$

  Sur $I_{n,r}$, $P^\theta_{n,r}$ polynôme de degré θ englobant tous les points et d'aire minimum :

$$\min \int_{I_{n,r}} P^\theta_{n,r}(x)dx \text{ s.c. } P^\theta_{n,r}(X_i) \geq Y_i, \;\; X_i \in I_{n,r}.$$

- *Théoriquement :* Sous certaines conditions (f α-lipschtizienne, ϕ constante, k_n bien choisi,...) $\hat{f}_n^0(x)$ est minimax pour les normes L_1 et L_∞.
- *En pratique :* Estimateur non lisse, biaisé, choix de la partition et en particulier de k_n.



2.3. Estimateurs basés sur une partition, lissés avec correction de biais

- *Cadre :* Σ_n processus de Poisson d'intensité moyenne $E(N_n) = nc\nu \otimes \lambda \mathbb{1}_D$,
  $I_{n,r}$, $r = 1, \dots, k_n$ une partition de $E \subset \mathbb{R}^d$,
  $D_{n,r} = \{(x, y),\ x \in I_{n,r},\ 0 \leq y \leq f(x)\}$, $r = 1, \dots, k_n$ la partition de D associée.
- *Famille d'estimateurs :*

$$\hat{f}_n(x) = \sum_{r=1}^{k_n} \nu(I_{n,r}) \kappa_{n,r}(x) Y_{n,r}^{\star\star}$$

  - $\kappa_{n,r} : E \to \mathbb{R}$ est un noyau généralisé,
  - $Y_{n,r}^{\star\star}$ est la valeur extrême $Y_{n,r}^{*}$ incluant une correction de biais.
- *Exemples de corrections de biais :*
  - correction locale $Y_{n,r}^{\star\star} = Y_{n,r}^{*}(1 + 1/N_n(D_{n,r}))$,
  - correction globale $Y_{n,r}^{\star\star} = Y_{n,r}^{*} + \sum_{s=1}^{k_n} \nu(I_{n,s}) Z_{n,s}^{\star}$, où $Z_{n,s}^{\star} = \min\{Y_i : X_i \in I_{n,s}\}$.

- *Exemples de noyaux généralisés.* $E = [0,1]$, $d = 1$, $(I_{n,r})$ partition régulière.
  - Noyaux de type Dirichlet : $\kappa_{n,r}(x) = K_{b_n}(x, x_r)$, x_r centre de $I_{n,r}$,
  $$K_{b_n}(x,y) = \sum_{\ell=0}^{b_n} e_\ell(x) e_\ell(y),$$
  où $(e_\ell)_{\ell \geq 0}$ est une base orthogonale ou non (trigonométrique, Haar, Faber-Shauder).
  - Noyaux de type Parzen-Rosenblatt : $\kappa_{n,r}(x) = K_{h_n}(x - x_r)$,
  $$K_{h_n}(x-y) = \frac{1}{h_n} K\left(\frac{x-y}{h_n}\right),$$
  K noyau de Parzen-Rosenblatt, h_n fenêtre de lissage.
  - Noyaux intégrés : par exemple,
  $$\kappa_{n,r}(x) = k_n \int_{I_{n,r}} K_{h_n}(x-t)dt,$$
  conduit à une convolution de l'estimateur de Geffroy par la fonction noyau K_{h_n}.

- *Exemple de résultats asymptotiques.*
  - $E = [0,1]^d$, ν mesure de Lebesgue sur E, $(I_{n,r})$ partition régulière,
  - $\kappa_{n,r}$ noyau de Parzen-Rosenblatt intégré, correction de biais locale,

L'estimateur s'écrit

$$\hat{f}_n(x) = \sum_{r=1}^{k_n} \left(\int_{I_{n,r}} K_{h_n}(x-t)dt \right) \left(1 + \frac{1}{N_n(D_{n,r})} \right) Y^{\star}_{n,r}.$$

Si de plus,

- f est α-lipschitzienne,
- $h_n = n^{-\frac{1}{\alpha+d}}$, $k_n = n^{\frac{d}{\alpha+d}} u_n^2$, avec $u_n \to \infty$,

alors pour tout $(x_1, \ldots, x_p) \subset (0,1)^d$,

$$\left\{ n^{\frac{\alpha}{\alpha+d}} u_n^{-1} c \left(\widehat{f}_n(x_j) - f(x_j) \right) : 1 \le j \le p \right\} \xrightarrow{d} \mathcal{N}\left(0, \|K\|_2^2 \, I_p\right).$$

2.4. Estimateurs à base de programmation linéaire, lissés

- *Cadre :* Σ_n $n-$échantillon de loi uniforme sur D, $E = [0, 1]$, $d = 1$.
- *Famille d'estimateurs :*

$$\hat{f}_n(x) = \sum_{i=1}^{n} \alpha_i K_{h_n}(x - X_i),$$

  - Le paramètre α_i détermine l'influence du point (X_i, Y_i) dans l'estimation, $i = 1, \ldots, n$.
  - Déterminés par un problème d'optimisation linéaire sous contraintes linéaires.
- *Programme linéaire :*

$$\min \int_{\mathbb{R}} \hat{f}_n(x)dx = \min \sum_{i=1}^{n} \alpha_i$$

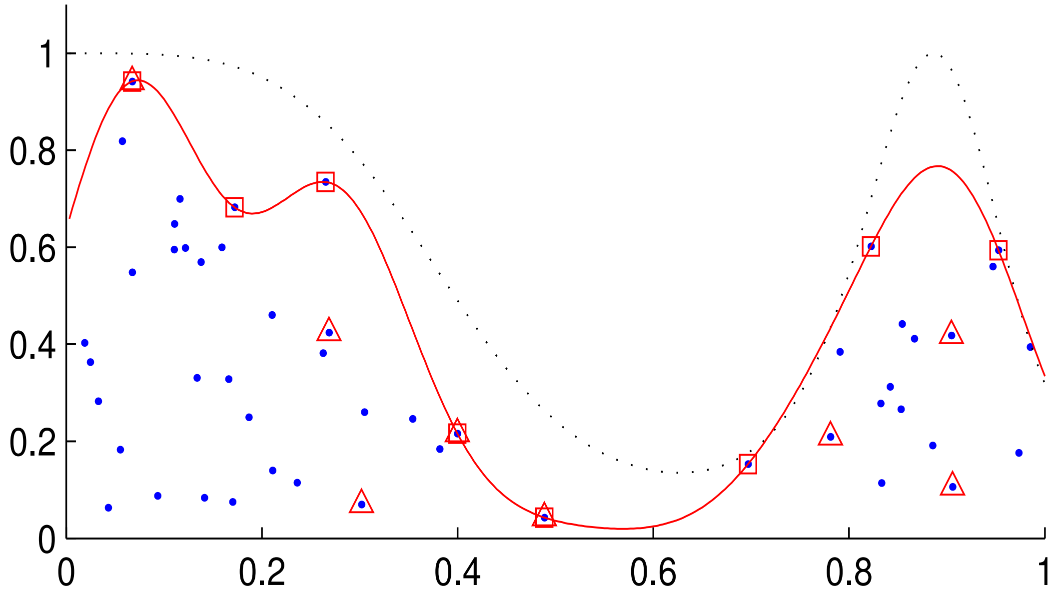
sous les $2n$ contraintes $\hat{f}_n(X_i) \geq Y_i$, $\alpha_i \geq 0$, $i = 1, \ldots, n$.

- *Premier résultat asymptotique.* Si f est 1-lipschitzienne, $h_n \to 0$, $\log n/(nh_n) \to 0$ alors

$$\limsup_{n\to\infty} \varepsilon_n^{-1} \int_{\mathbb{R}} |\hat{f}_n(x) - f(x)|dx \leq C < \infty \quad \text{p.s.}$$

avec $\varepsilon_n = \max\left\{h_n, \sqrt{\log n/(nh_n)}\right\}$, et où K est un noyau "bien choisi".

- *Exemple de résultat.*



2.5. Perspectives

- *Estimateurs basés sur une partition.*
  - Déterminer, pour une régularité de frontière f donnée, la meilleure famille de noyaux $(\kappa_{n,r})$.
  - Vitesse de convergence en norme L_1 et L_∞ pour comparaison avec les vitesses minimax.
- *Estimateurs à base de programmation linéaire.*
  - Amélioration de l'estimateur proposé par introduction d'une contrainte de régularité.
  - Nouvelle vitesse de convergence pour comparaison à la vitesse minimax.
  - Loi asymptotique.
- *Perspectives communes.*
  - Choix adaptatif du paramètre de lissage.
  - Extension à des distributions de points non uniformes, courbes de quantiles conditionnels extrêmes (utilisation de la théorie des valeurs extrêmes).

3. Réduction de dimension et analyse d'images

En collaboration avec :

- Bernard Chalmond (ENS Cachan, Université Cergy-Pontoise),
- Jean-Marc Dinten (CEA Grenoble),
- Serge Iovleff (Université Lille I),
- Charles Bouveyron, Cordelia Schmid (INRIA Rhône-Alpes),
- Philippe Guérin, Henri Maître, Michel Roux (ENST Paris).

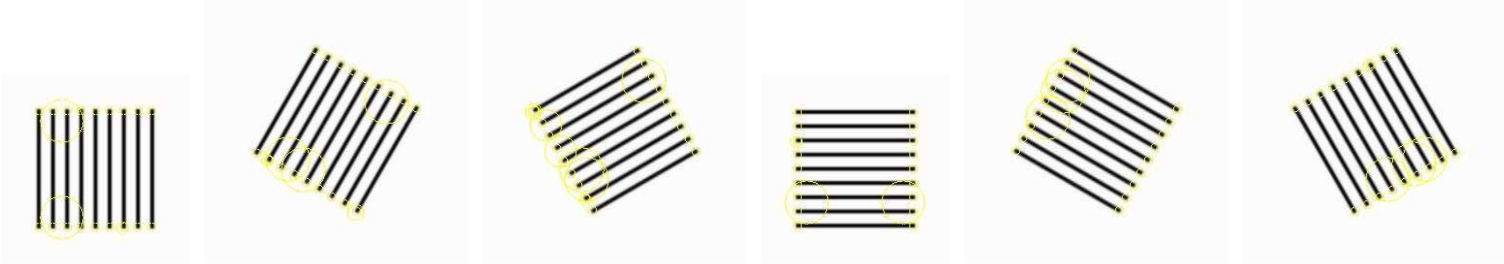
Dans les cadres suivants :

- Projet européen Verbonds (1993–1996),
- Projet européen Impact (1996–1999),
- ACI Masse de données (2003–2006),
- Thèse de Charles Bouveyron (2003–2006).

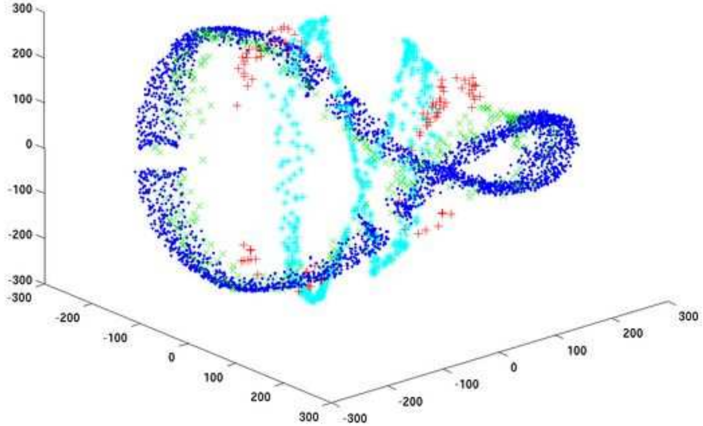
3.1. Motivation.

- Base d'images I_i, $i = 1, \ldots, m$ de taille $M \times M$ en niveaux de gris.
- Plusieurs types de représentation selon le type d'application. Pour $i = 1, \ldots, m$
  - 1 image $I_i \longrightarrow$ 1 vecteur X_i de $\mathbb{R}^p$.
    Au final, $n = m$ vecteurs de dimension $p = M^2$.
  - 1 image $I_i \longrightarrow n_i$ descripteurs $X_{i,j}$, $j = 1, \ldots, n_i$ dans $\mathbb{R}^{128}$, par exemple.
    Au final, $n = \sum_{i=1}^m n_i$ vecteurs de dimension $p = 128$.
- Dans les deux cas, nécessité de réduire la dimension.
- Difficulté : structure non linéaire du nuage de points obtenus.
- Exemple.
  - $m = 180$ images obtenues par rotations successives de 1 degré d'un même objet.
  - $n = 3800$ descripteurs SIFT (Lowe, 2004) en dimension 128.
  - Projection sur les 3 premiers axes obtenus par Analyse en Composantes Principales (ACP) pour visualisation.

Extrait de la base de 180 images.



Projection sur les 3 premiers axes obtenus par ACP.



3.2. L'Analyse en Composantes Principales (ACP).

- *Principe.* Construction d'un modèle linéaire de dimension d pour X vecteur aléatoire de $\mathbb{R}^p$ (on suppose X centré)

$$X = \sum_{j=1}^{d} Y_j a^j + R^d.$$

- *Algorithme itératif.*
  - Pour $j = 0$, on pose $R^0 = X$.
  - Pour $j = 1, \dots, d$:

[A] Déterminer $a^j = \arg\max_{x \in \mathbb{R}^p} \mathbb{E}\left(\left\langle x, R^{j-1}\right\rangle^2\right)$ s.c. $\|x\| = 1$, $\left\langle x, a^k \right\rangle = 0$, $1 \leq k < j$.

[P] Calculer $Y_j = \left\langle a^j, R^{j-1}\right\rangle$.

[R] Déterminer $b^j = \arg\min_{x \in \mathbb{R}^p} \mathbb{E}\left(\left\| R^{j-1} - Y_j x\right\|^2\right)$ s.c. $\left\langle x, a^j \right\rangle = 1$.

On trouve $b^j = a^j$ et on pose $s^j(Y_j) = Y_j b^j$

[M] Calculer $R^j = R^{j-1} - s^j(Y_j)$.

- *Propriétés.*
  - Les variables principales Y_j, $j = 1, \dots, d$ sont décorrélées et centrées.
  - Le résidu R^j est centré et orthogonal aux axes principaux précédents : $\langle R^j, a^k \rangle = 0$ p.s. $k = 1, \dots, j$.
  - Les résidus sont presque sûrement décroissants $\|R^j\| \le \|R^{j-1}\|$ p.s.
  - Le modèle peut être réécrit $F(X) = R^d$ avec $F : \mathbb{R}^p \to \mathbb{R}^p$ définie par

$$F(x) = \sum_{j=1}^{d} \langle x, a^j \rangle a^j .$$

    L'équation $F(x) = 0$ définit un sous-espace vectoriel de dimension d dans $\mathbb{R}^p$.
  - Pour $d = p$, le modèle est exact : $R^p = 0$ p.s.
- *Notre contribution.* Modification de l'algorithme itératif de l'ACP de manière à :
  - obtenir un modèle non-linéaire de X,
  - conserver l'essentiel des propriétés précédentes.

3.3. Les modèles Auto-Associatifs (AA).

- *Principe.* Construction d'un modèle de dimension d pour X vecteur aléatoire de $\mathbb{R}^p$

$$X = \sum_{j=1}^{d} s^j(Y_j) + R^d.$$

- *Algorithme itératif.*
  - Pour $j = 0$, on pose $R^0 = X$.
  - Pour $j = 1, \dots, d$:

[A] Déterminer $a^j = \arg\max_{x \in \mathbb{R}^p} I\left(\left\langle x, R^{j-1}\right\rangle\right)$ s.c. $\|x\| = 1, \left\langle a^k, x\right\rangle = 0,\ 1 \leq k < j$.

[P] Calculer $Y_j = \left\langle a^j, R^{j-1}\right\rangle$.

[R] Choisir $s^j \in \arg\min_{s \in \mathcal{S}(\mathbb{R},\mathbb{R}^p)} \mathbb{E}\left(\left\|R^{j-1} - s(Y_j)\right\|^2\right)$ s.c. $P_{a^j} \circ s = \mathbb{I}$.

[M] Calculer $R^j = R^{j-1} - s^j(Y_j)$.

- *Propriétés.*
  - Les variables principales Y_j, $j = 1, \ldots, d$ sont centrées.
  - Les fonctions explicatrices sont orthogonales aux axes principaux précédents : $\left\langle s^j(Y_j), a^k \right\rangle = 0$ p.s. $k = 1, \ldots, j-1$.
  - Le résidu R^j est centré et orthogonal aux axes principaux précédents : $\left\langle R^j, a^k \right\rangle = 0$ p.s. $k = 1, \ldots, j$.
  - Les résidus sont décroissants $\|R^j\| \leq \|R^{j-1}\|$ p.s.
  - Le modèle peut être réécrit $F(X) = R^d$ avec $F : \mathbb{R}^p \to \mathbb{R}^p$ définie par

$$F(x) = \left(\mathbb{I}_{\mathbb{R}^p} - s^d \circ P_{a^d}\right) \circ \ldots \circ \left(\mathbb{I}_{\mathbb{R}^p} - s^1 \circ P_{a^1}\right)(x) = \coprod_{k=d}^{1} \left(\mathbb{I}_{\mathbb{R}^p} - s^k \circ P_{a^k}\right)(x).$$

    L'équation $F(x) = 0$ définit une variété de dimension d dans $\mathbb{R}^p$.
  - Pour $d = p$, le modèle est exact : $R^p = 0$ p.s.

3.4. Exemple des modèles Auto-Associatifs de Régression (AAR).

- Etape [R] : En choisissant $\mathcal{S}(\mathbb{R},\mathbb{R}^p)=L_2$, on obtient $s^j(t)=\mathbb{E}(R^{j-1}|Y_j=t)$. Estimation de la fonction de régression s^j par une méthode non-paramétrique classique (noyau, splines ...)
- Etape [A] : Problème de Poursuite de Projection.
  - Etant donné les résidus de la $(j-1)$ème itération $R_1^{j-1},\dots,R_n^{j-1}$, l'axe a^j est obtenu par maximisation en x de l'index :

$$I\left(\langle x,R^{j-1}\rangle\right)=\sum_{i=1}^n\left\langle x,R_i^{j-1}\right\rangle^2\bigg/\sum_{k=1}^n\sum_{\ell=1}^n m_{k\ell}^j\left\langle x,R_k^{j-1}-R_\ell^{j-1}\right\rangle^2,$$

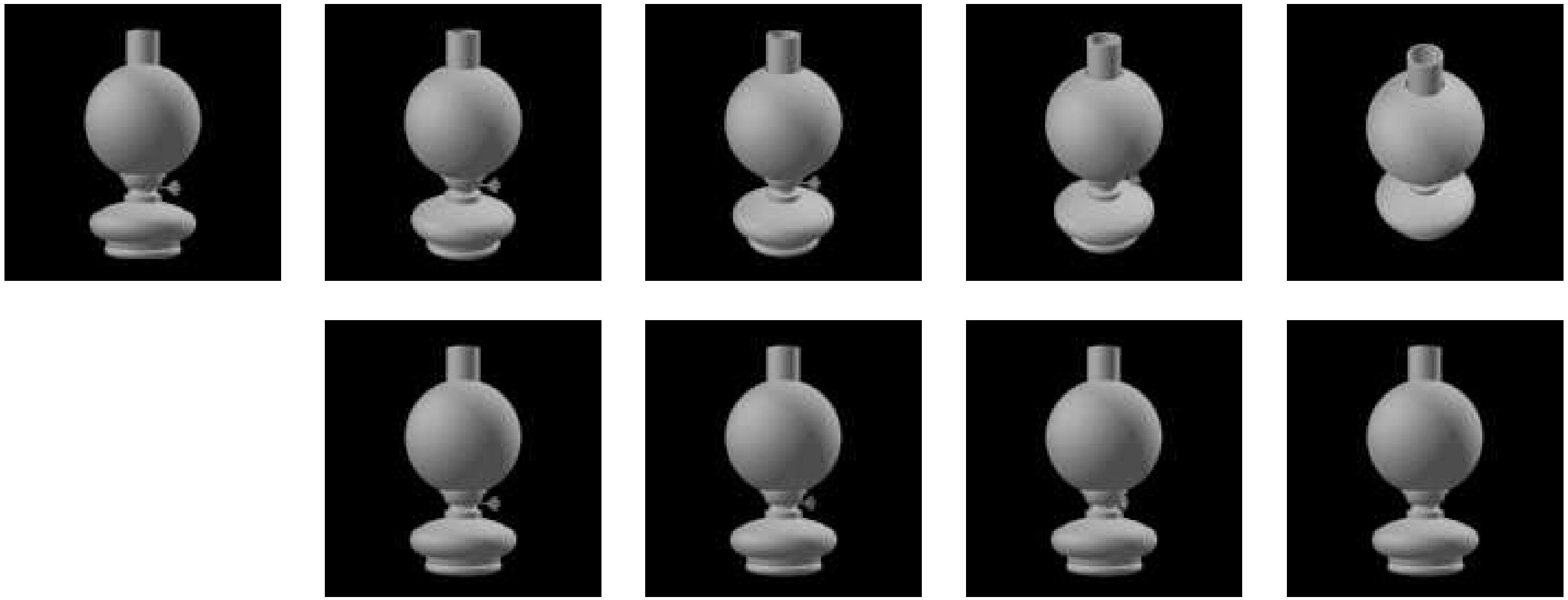
  où $m_{k\ell}^j$ est le coefficient (k,ℓ) de la matrice de contiguité $m_{k\ell}^j=\mathbb{I}\{R_k^{j-1}\text{ ppv }R_\ell^{j-1}\}$.
  - Solution explicite : a^j vecteur propre de la matrice

$$\left(\sum_{k=1}^n\sum_{\ell=1}^n m_{k\ell}^j\,{}^t(R_k^{j-1}-R_\ell^{j-1})(R_k^{j-1}-R_\ell^{j-1})\right)\left(\sum_{k=1}^n {}^tR_k^{j-1}R_k^{j-1}\right)^{-1},$$

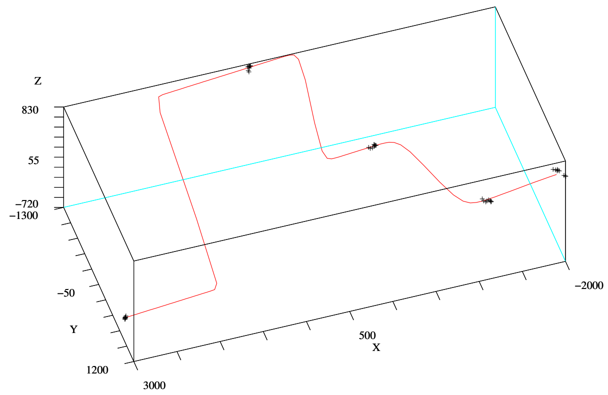
  associé à la plus grande valeur propre.

3.5. Illustration.

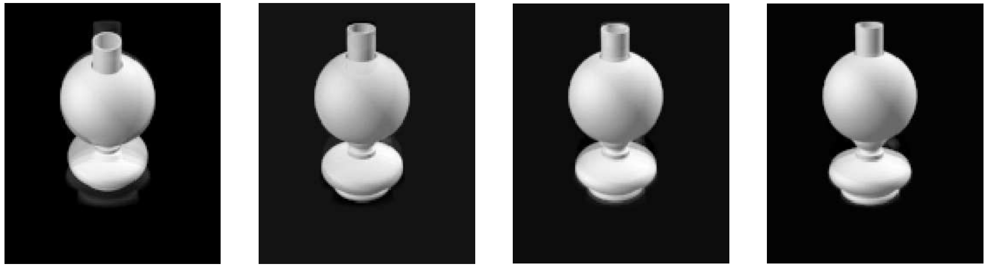
- Base de 45 images de synthèse de taille 256×256.
- Représentation par un nuage de $n = 45$ points en dimension 256^2, puis par changement de repère en dimension $p = n - 1 = 44$.
- Extrait de la base :



Variété de dimension 1 estimée dans le repère formé des trois premiers axes principaux.



Simulation de 4 images par le modèle AAR de dimension 1.



La variable Y_1 est simulée uniformément sur l'intervalle $[\min_i Y_{1,i}, \max_i Y_{1,i}]$.

3.5. Perspectives

- *Modèles Auto-Associatifs.*
  - Propriétés asymptotiques des estimateurs.
  - Extension à des formes de variétés plus générales.
  - Lois de mélange.
- *Application à l'analyse d'image.*
  - Prise en compte de la nature spatiale des données.
  - Application à un problème concret.

4. Estimation de courbes de référence

En collaboration avec :

- Ali Gannoun, Jérôme Saracco (Université Montpellier II),
- Christiane Guinot (CERIES).

Dans le cadre suivant :

- Contrat de recherche CERIES/Université Montpellier II.

4.1. Introduction

- *Intervalles de référence.* Intervalles de valeurs qui sont prises "normalement" par une variable d'intérêt Y, dans une population cible (par exemple, intervalle excluant les 5% d'observations les plus grandes et les 5% d'observations les plus petites)

Construction d'intervalles de référence $\longrightarrow$ Calcul de quantiles de Y.

- *Courbes de référence.* On dispose avec la variable d'intérêt Y, d'une information complémentaire sous la forme d'une covariable X (par exemple l'âge du sujet).
  Pour une valeur donnée x de X, on peut construire un intervalle de référence.
  Lorsque x varie, on obtient des "courbes de référence" (si $X \in \mathbb{R}$) ou des hypersurfaces (si $X \in \mathbb{R}^p$).

Construction de courbes de référence $\longrightarrow$ Calcul de quantiles conditionnels de Y sachant X.

4.2. Covariable X unidimensionnelle

- *Intérêt.* Comparaison d'un individu i représenté par le point (x_i, y_i) à la population de référence.
- *Caractérisations.*
  - Quantile conditionnel d'ordre $\alpha \in]0.5, 1[$ de Y sachant $X = x$:
  $$q_\alpha(x) = F^{-1}(\alpha | x),$$
  avec $F(.|x)$ f.d.r. conditionnelle de Y sachant $X = x$. De façon équivalente :
  $$q_\alpha(x) = \arg\min_{\theta \in \mathbb{R}} \mathbb{E}(\rho_\alpha(Y - \theta) | X = x),$$
  avec ρ_α est la fonction de perte définie par $\rho_\alpha(z) = \alpha z \mathbb{I}_{[0,\infty)}(z) - (1-\alpha) z \mathbb{I}_{(-\infty,0)}(z)$.
  - Invervalle de confiance contenant $100(2\alpha - 1)\%$ des sujets de référence :
  $$I_\alpha(x) = [q_{1-\alpha}(x),\ q_\alpha(x)].$$
  - Courbes de référence : ensembles de points
  $$\{(x, q_{1-\alpha}(x)\} \text{ et } \{(x, q_\alpha(x)\}$$
  lorsque x varie.

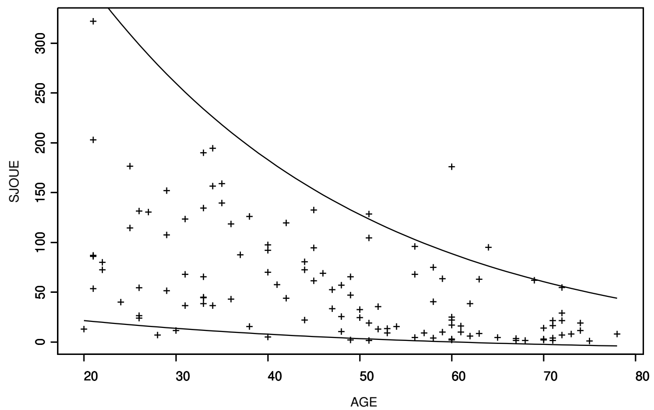
- *Comparaison de 5 méthodes d'estimation.*
  - 1 méthode paramétrique. Hypothèse de gaussianité de $F(y|x)$ (Royston, 1991).
  - 1 méthode semi-paramétrique. Recherche d'une transformation des valeurs observées de Y permettant de se ramener au cas précédent (Cole, 1988).
  - 3 méthodes non-paramétriques.

$$\hat{F}_{1,n}(y|x) = \sum_{i=1}^{n} K\left(\frac{x-X_i}{h_{1,n}}\right)\mathbb{I}\{Y_i \leq y\} \Bigg/ \sum_{i=1}^{n} K\left(\frac{x-X_i}{h_{1,n}}\right) \rightarrow \hat{q}_{\alpha,1,n}(x) = \hat{F}_{1,n}^{-1}(\alpha|x),$$

$$\hat{q}_{\alpha,2,n}(x) = \arg\min_{a\in\mathbb{R}} \sum_{i=1}^{n} \rho_\alpha(Y_i - a) K\left(\frac{x-X_i}{h_{2,n}}\right),$$

$$\hat{F}_{3,n}(y|x) = \sum_{i=1}^{n} K\left(\frac{x-X_i}{h_{3,n}}\right)\Omega\left(\frac{y-Y_i}{h_{4,n}}\right) \Bigg/ \sum_{i=1}^{n} K\left(\frac{x-X_i}{h_{3,n}}\right) \rightarrow \hat{q}_{\alpha,3,n}(x) = \hat{F}_{3,n}^{-1}(\alpha|x).$$

- *Application à des données réelles.*
  - Objectif : établir des courbes de référence à 90% en fonction de l'âge (covariable) pour 38 propriétés biophysiques (variables d'intérêt) de la peau de femmes de type caucasien mesurées sur le front, la joue et la face antérieur du bras gauche.
  - Heuristique de choix des paramètres de lissage.
    $h_{1,n}$: critère de type validation croisée dérivé de (Yao, 1999),
    $h_{2,n}, \ldots, h_{4,n}$: règle empirique (Yu & Jones, 1998).
  - Implantation en C des méthodes.
  - Exemple de courbes de références non acceptables.



Synthèse des résultats :

	Joue	Front	Avant-bras
Nombre de variables	13	13	12
Méthode paramétrique			
Nombre de courbes de référence acceptées	6	9	6
Méthode semi paramétrique			
Nombre de courbes de référence acceptées	11	11	12
Méthodes non paramétriques			
Méthode 1 : nombre de courbes de référence acceptées	13	13	12
Méthode 2 : nombre de courbes de référence acceptées	12	12	12
Méthode 3 : nombre de courbes de référence acceptées	6	5	3

4.3. Covariable X multidimensionnelle

Y est une variable aléatoire, X est un vecteur aléatoire de $\mathbb{R}^p$, $p > 1$.

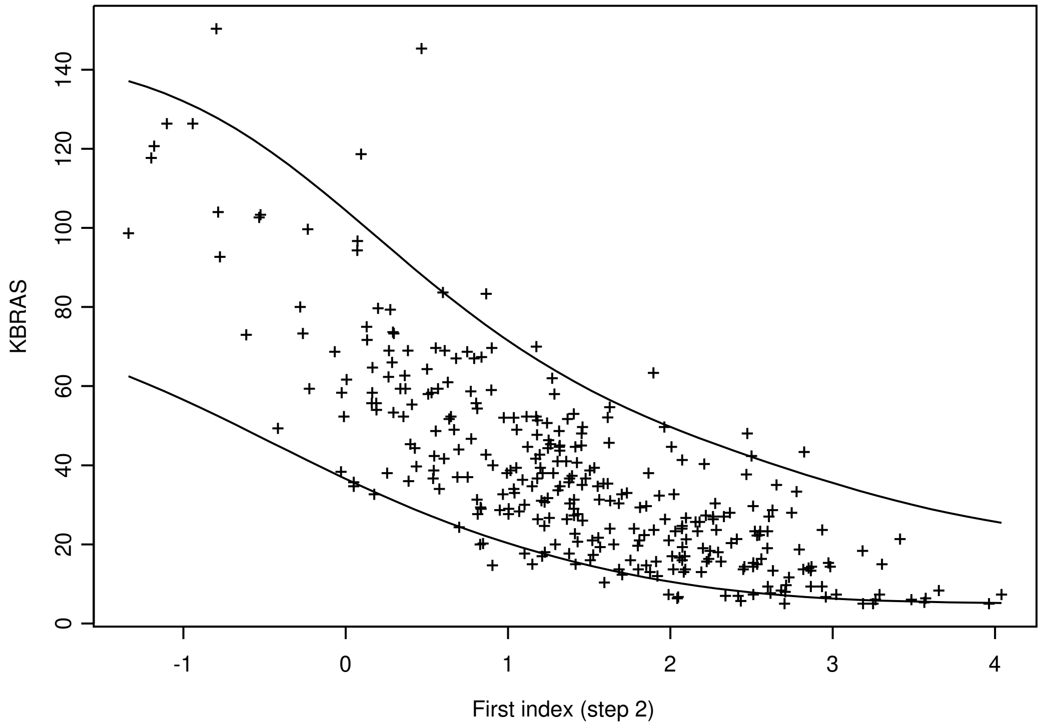
- *En théorie.* Pas de problème de définition d'estimateurs de $q_\alpha(x)$, $x \in \mathbb{R}^p$.
- *En pratique.*
  - Fléau de la dimension,
  - Problème de visualisation des hyper-surfaces de référence pour $p > 2$.
- *Notre approche.* Réduction de la dimension de X, sans perte d'information, sans introduire de modèle paramétrique.
  - Hypothèse : il existe une matrice β, de taille $p \times d$, de rang minimal, telle que $F(y|x) = F(y|\,{}^t\beta x)$.
  - Conséquence : $q_\alpha(x) = q_\alpha({}^t\beta x)$.
  - Objectif : Estimer une base de $S(\beta)$, sous-espace vectoriel engendré par β (espace EDR "Effective Dimension Reduction").
  - Outil : méthode SIR "Sliced Inverse Regression".

- *Principe de la méthode SIR.* (Li, 1991)
  - Hypothèse : Pour tout $b \in \mathbb{R}^p$, $\mathbb{E}({}^t bX|\, {}^t\beta X)$ est linéaire en ${}^t\beta X$.
  - Conséquence : L'estimation de l'espace EDR se ramène au calcul des vecteurs propres $b_1, \dots, b_d$ de M la matrice de covariance de $\mathbb{E}(X|\tilde{Y})$, où $\tilde{Y}$ est la v.a. obtenue en partitionnant le support de Y en tranches.
- *Estimation des quantiles conditionnels.*
  (pour simplifier, on suppose $d = 1$, et on note $b = b_1$.)
  - $\hat{F}_n(y|\, {}^t bx) = \sum_{i=1}^{n} K\left(\frac{{}^t bx - {}^t bX_i}{h_n}\right) \mathbb{I}\{Y_i \le y\} \Big/ \sum_{i=1}^{n} K\left(\frac{{}^t bx - {}^t bX_i}{h_n}\right)$
  - $\hat{q}_{\alpha,n}({}^t bx) = \hat{F}_n^{-1}(\alpha|{}^t bx)$.
- *Hyper-surfaces de référence.* Ensembles de points

$$\{(x, \hat{q}_{1-\alpha,n}({}^t bx)\} \text{ et } \{(x, \hat{q}_{\alpha,n}({}^t bx)\}$$

lorsque x varie. Lorsque $d = 1$, elles sont invariantes selon $b^\perp$. Visualisation possible par projection sur b.

Exemple. Projection des hyper-surfaces de référence à 90% estimées pour la variable KBRAS (conductance mesurée sur l'avant-bras).



4.4. Prolongements

- *Travaux en cours.*
  (George Bonney, Ali Gannoun, Christiane Guinot, Jérôme Saracco, Wolfang Urfer)
  - Test de comparaison d'espaces EDR obtenus sur des échantillons différents.
  - Extension au cas où Y est aussi multidimensionnelle.
  - Adaptation au cas de données censurées.
- *Perspectives.*
  - Convergence en loi de $\hat{q}_{\alpha,n}({}^t bx) = \hat{F}_n^{-1}(\alpha|{}^t bx)$ afin d'apprécier théoriquement le gain apporté par la réduction de dimension.

5. Perspectives de recherche

- Etude des extrêmes multidimensionnels.
  Outils : Théorie des valeurs extrêmes, copules.
- Estimation de frontières de support de formes plus générales.
  Outils : Théorie des valeurs extrêmes, programmation linéaire.