Modélisation statistique et réduction de la dimension : une application à l'estimation non-paramétrique et semi-paramétrique de courbes de référence

Marie Chavent, Jérôme Saracco

Université de Bordeaux,
Institut de Mathématiques de Bordeaux (IMB),
Equipe CQFD, INRIA Bordeaux Sud Ouest

Rencontre CEA - Equipe CQFD, INRIA Bordeaux Sud Ouest - Jeudi 23 octobre 2008

Plan

1. Introduction générale
2. Introduction sur les courbes de référence
3. Caractérisation des quantiles
4. Estimation non-paramétrique des courbes de référence
5. Application à la construction de courbes de référence pour des propriétés biophysiques de la peau
6. Estimation semi-paramétrique des courbes de référence

Plan

1. Introduction générale
2. Introduction sur les courbes de référence
3. Caractérisation des quantiles
4. Estimation non-paramétrique des courbes de référence
5. Application à la construction de courbes de référence pour des propriétés biophysiques de la peau
6. Estimation semi-paramétrique des courbes de référence

Modèle paramétrique, non-paramétrique ou semi-paramétrique : cas de la régression

- Objectif de la régression : étudier les relations entre une variable à expliquer Y et une variable explicative X (unidimensionnelle ou multidimensionnelle).
- Souvent, un modèle **paramétrique** est utilisé.

  Modèle paramétrique général : $Y = r_\theta(X) + \epsilon$, où $\theta \in \mathbb{R}^d$.

  Objectif : estimer θ.

  Exemple : modèle de régression linéaire $Y = \beta_0 + \beta_1 X + \epsilon$ avec $\epsilon \sim N(0, \sigma^2) \rightarrow$ estimer $\theta = (\beta_0, \beta_1)$.

  Méthodes standards d'estimation : moindres carrés, maximum de vraisemblance, etc..

  Problème : Choix d'une "bonne" famille de modèle paramétrique.

- Un modèle **non paramètrique** de régression apparaît alors comme une alternative raisonnable.

  Thème commun de la régression non paramétrique : lissage local qui explore les propriétés de continuité et de dérivabilité de la fonction de régression.

  Exemple de modèle non paramétrique : $Y = r(X) + \epsilon$, où $r \in R = \{\phi \text{ continue de } \mathbb{R} \text{ dans } \mathbb{R}\}$.

  Objectif : estimer $r(x_0)$ pour une valeur x_0 donnée.

  Méthodes standards d'estimation : estimateurs à noyau, splines de lissage, ondelettes, etc...

  Problème : le fléau (ou malédiction) de la dimension ("curse of dimensionnality"). En pratique, il faut avoir suffisamment d'observations autour du point d'intérêt, ce qui pose un problème dès que la dimension de x augmente.

Comment surmonter ce problème de dimension ?

Eléments de réponse : modifier le modèle

▷ passer à un modèle additif : $y = \sum_{j=1}^{p} f_j(x_j) + \varepsilon$ ("Projection Pursuit")

▷ "retrouver" les caractéristiques "intéressantes" des données de grande dimension ($x \in \mathbb{R}^d$) sur des sous-espaces de dimension plus faible (souvent par des projections)

Idée : utiliser non plus x mais des combinaisons linéaires $x'\beta_k$.

$\hookrightarrow$ **modèle semi-paramétrique**

Intérêt d'une modélisation semi-paramétrique incluant une réduction de dimension :

Bon compromis entre la modélisation paramétrique (interprétabilité) et la modélisation non-paramétrique (souplesse)

Exemple de modèle semi-paramétrique de régression (Li, 1991) :

$$Y = g(X'B, \epsilon)$$

- $Y \in \mathbb{R}$ = variable à expliquer.
- $X \in \mathbb{R}^p$ = variable explicative.
- $B = [\beta_1, ..., \beta_K]$ avec $\beta_k \in \mathbb{R}^p$ = paramètres euclidiens inconnues ,

  NB : quand $K << p$, le but de réduction de dimension est atteint.
- g = fonction de $\mathbb{R}^{K+1}$ dans $\mathbb{R}$ inconnue et arbitraire.
- ϵ = terme d'erreur aléatoire de loi arbitraire et inconnue, indépendante de x.

Objectif : estimer le paramètre euclidien B et le paramètre fonctionnel g.

Une méthode d'estimation de B : méthode SIR (Sliced Inverse Regression)

Plan

1 Introduction générale

2 Introduction sur les courbes de référence

3 Caractérisation des quantiles

4 Estimation non-paramétrique des courbes de référence

5 Application à la construction de courbes de référence pour des propriétés biophysiques de la peau

6 Estimation semi-paramétrique des courbes de référence

- Intervalles et courbes de référence utilisés dans les études biomédicales ou biométriques, les applications industrielles, ...
- Intervalle de référence = intervalle de valeurs qui sont prises "normalement" par une variable d'intérêt Y dans une population cible

  Valeurs "normalement" prises = valeurs que l'on est susceptible d'observer avec une probabilité donnée, dans des conditions normales et pour des individus-types présumés en bonne santé ou sans défaut (les sujets de référence)
- *Exemple :* intervalle contenant 90% des observations "centrales" de Y, i.e. excluant les 5% d'observations les plus grandes et les 5% d'observations les plus petites.

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

Régulièrement, sur la population cible, on dispose simultanément de
- la variable d'intérêt $Y \in \mathbb{R}$,
- une covariable $X \in \mathbb{R}^p$ (Ex. : X = âge, cond. expérimentales)

$\hookrightarrow$ Pour une valeur donnée x de $X \longrightarrow$ un intervalle de référence.

$\hookrightarrow$ Lorsque x varie $\longrightarrow$ des "courbes" de référence.

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

• **Définition de l'intervalle de référence à** $100(2\,\alpha\,-\,1)\%$ (pour $X = x$ et $\alpha \in]0.5, 1[$) :

$$I_\alpha(x) \;=\; [\; q_{1-\alpha}(x)\;;\; q_\alpha(x)\;],$$

où $q_\alpha(x)$ est le quantile conditionnel d'ordre α de Y sachant que $X = x$.

• **Définition des courbes de référence :**

$$\{(x, q_{1-\alpha}(x)\} \quad \text{et} \quad \{(x, q_\alpha(x)\} \quad \text{lorsque } x \text{ varie}.$$

• *Exemple :* pour $\alpha = 95\%$, courbes de référence à 90%

$$\{(x, q_{5\%}(x)\} \quad \text{et} \quad \{(x, q_{95\%}(x)\} \quad \text{lorsque } x \text{ varie}.$$

• **Une "utilisation"** *a posteriori* **des courbes de référence** : comparer un individu i à la population de référence $\longrightarrow$ détecter si cet individu est "hors-norme"

Plan

1 Introduction générale

2 Introduction sur les courbes de référence

3 Caractérisation des quantiles

4 Estimation non-paramétrique des courbes de référence

5 Application à la construction de courbes de référence pour des propriétés biophysiques de la peau

6 Estimation semi-paramétrique des courbes de référence

Soit Y une variable aléatoire réelle.

Deux caractérisations de la médiane :

- $\mu = F^{-1}(1/2)$ où F est la fonction de répartition de Y : $F(y) = P(Y \leq y)$
- $\mu = \arg\min_{\theta \in \mathbb{R}} \ \mathbb{E}\left[|Y - \theta|\right]$

Soit $\alpha \in (0, 1)$.

Deux caractérisations du quantile d'ordre α de Y, noté μ_α :

- $\mu_\alpha = F^{-1}(\alpha)$
- $\mu_\alpha = \arg\min_{\theta \in \mathbb{R}} \mathbb{E}\left[L_\alpha(Y - \theta)\right],$

où $L_\alpha(v) = |v| + (2\alpha - 1)v$ définit une "fonction de perte" sur $\mathbb{R}$.

Estimateur du quantile μ_α *:* $\{Y_1, \ldots, Y_n\}$, n observations de Y

- ▷ Approche indirecte : $\mu_{\alpha,n} = F_n^{-1}(\alpha)$
- ▷ Approche directe : $\mu_{\alpha,n} = \arg\min_{\theta \in \mathbb{R}} \sum_{i=1}^{n} L_\alpha(Y_i - \theta)$

Soit $\alpha \in (0,1)$. Soit X un vecteur aléatoire à valeurs dans $\mathbb{R}^p$ (avec $p \geq 1$).
Soit $x \in \mathbb{R}^p$.

Deux caractérisations du quantile conditionnel d'ordre α de Y sachant que $X = x$, noté $q_\alpha(x)$:

▷ Approche indirecte : $q_\alpha(x) = F^{-1}(\alpha|x)$

où $F(.|x)$ désigne la fonction de répartition conditionnelle de Y sachant que $X = x$: $F(y|x) = P(Y \leq y|X = x)$.

▷ Approche directe : $q_\alpha(x) = \arg\min_{\theta \in \mathbb{R}} \mathbb{E}\left[L_\alpha(Y - \theta) \mid X = x\right]$.

Comment estimer $q_\alpha(x)$ à partir de $\{(X_i, Y_i),\ i = 1, \ldots, n\}$, n réalisations de (X, Y) ?

↪ méthode paramétrique

↪ méthode non-paramétrique

↪ méthode semi-paramétrique

Plan

1 Introduction générale

2 Introduction sur les courbes de référence

3 Caractérisation des quantiles

4 Estimation non-paramétrique des courbes de référence

5 Application à la construction de courbes de référence pour des propriétés biophysiques de la peau

6 Estimation semi-paramétrique des courbes de référence

Un exemple d'une méthode paramétrique d'estimation de $q_\alpha(x)$

(approche paramétrique utilisée par le CERIES, centre de recherche sur la peau humaine financé par CHANEL)

- *Hypothèse :* $F(y|x)$ est supposée gaussienne,
  i.e. $Y|X = x \sim N(m(x), \sigma^2(x))$ avec $m(x) = E(Y|X = x)$, $\sigma^2(x) = \text{var}(Y|X = x)$
- *Quantile conditionnel "vrai" :* $q_\alpha(x) = m(x) + z_\alpha \sigma(x)$

  où z_α est le quantile d'ordre α de $N(0, 1)$.
- Méthode de Royston (1991) : modélisation polynomiale de $m(x)$ et de $\sigma(x)$ avec une transformation éventuelle des données pour normaliser les valeurs résiduelles $\longrightarrow$ $m_n(x)$ et de $\sigma_n(x)$
- *Quantile conditionnel estimé :* $q_{\alpha,n}(x) = m_n(x) + z_\alpha \sigma_n(x)$.
- *Critiques* :

  - hypothèses restrictives et mal adaptées à la réalité des données biologiques (aucune garantie de l'existence d'une transf. permettant de normaliser les résidus),

  - sensibilité aux valeurs aberrantes, ...

$\Longrightarrow$ Développement d'une **approche non-paramétrique** permettant de pallier ces problèmes d'hypothèses et de modélisation paramétriques

Avantages : - pas d'hypothèse sur la nature de la distribution,

- robustesse à la présence de points aberrants, ...

Trois méthodes non-paramétriques d'estimation de $q_\alpha(x)$

- Deux méthodes indirectes : estimation préalable de la fonction de répartition conditionnelle

  $\hookrightarrow$ **Méthode d'estimation par noyau** ("Kernel estimator")

  $\hookrightarrow$ **Méthode d'estimation par noyau produit** ("double kernel estimation")

- Une méthode directe :

  $\hookrightarrow$ **Méthode d'estimation linéaire locale par noyau** ("local linear kernel estimation")

Rapides notions sur les estimateurs à noyau de la régression

Modèle : $Y = r(X) + \varepsilon$ Echantillon : $\{(X_i, Y_i), i = 1, \ldots, n\}$

Estimateur de $r(x_0)$: $r_n(x_0) = \sum_{i=1}^{n} \left(\frac{K\{(x_0 - X_i)/h_n\}}{\sum_{j=1}^{n} K\{(x_0 - X_j)/h_n\}} \right) Y_i$

avec : K= noyau (fonction), h=largeur de fenêtre (réel)

Exemples de noyau : $K(x) = (1 - |x|)\mathbb{I}\{x \in [-1, 1]\}$ (noyau triangulaire), $K(x)$ = densité de N(0,1)

Idée : donner plus de poids aux observations telles que X_i est proche de x_0.

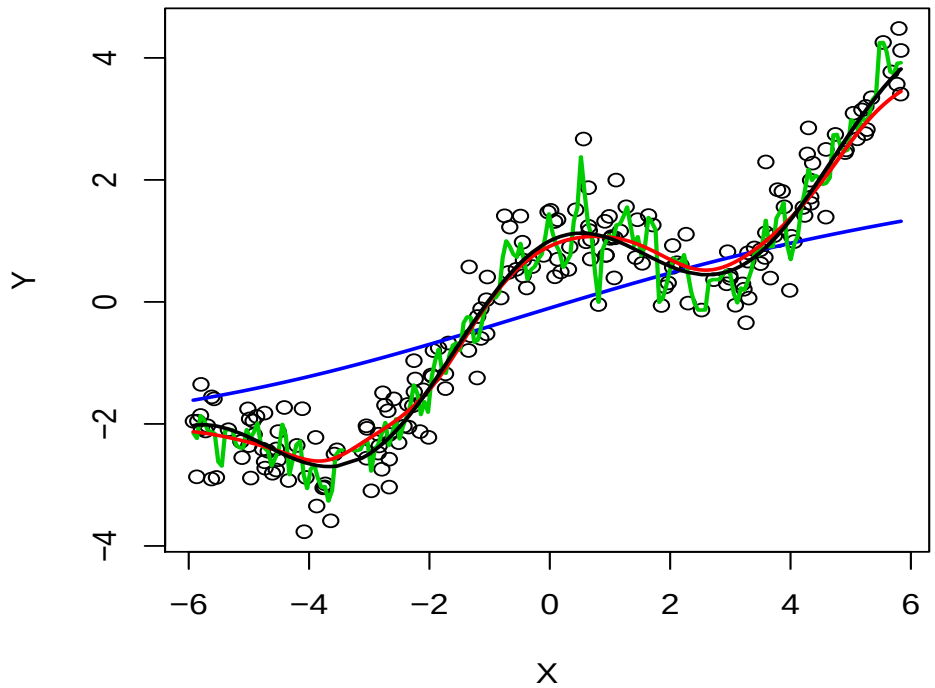
Remarque : estimation peu sensible au choix du noyau K

Largeur de fenêtre h : paramètre de lissage ← son choix est primordial

- plus h est petit, moins on va donner de poids aux observations éloignées de x_0 → sous-lissage
- plus h est grand, plus on va donner de poids à des observations éloignées de x_0 → sur-lissage
- h optimal ? (validation croisée, ...)

Vrai modèle : $Y = \cos(X) + 0.5X + \varepsilon$ (courbe noire)
Echantillon de taille $n = 200$

↪ Courbe noire : tracé de la fonction $r(x) = \cos(x) + 0.5x \longrightarrow$ vraie courbe
↪ Courbe verte : estimateur à noyau avec $h = 0.1 \longrightarrow$ sous-lissage
↪ Courbe rouge : estimateur à noyau avec $h = 1.5$ (proche du h_n optimal au sens de la validation croisée)
↪ Courbe bleue : estimateur à noyau avec $h = 10 \longrightarrow$ sur-lissage



(1) Méthode d'estimation par noyau (de la fonction de répartition conditionnelle)

En posant $Y^\star = \mathbb{I}_{\{Y \leq y\}}$, on a :
$F(y|x) = P(Y \leq y|X = x) = P(Y^\star = 1|X = x) = E(Y^\star|X = x)$.

- *Estimation de la fonction de répartition conditionnelle :*

$$F_n(y|x) = \sum_{i=1}^{n} \left(\frac{K\{(x - X_i)/h_n\}}{\sum_{j=1}^{n} K\{(x - X_j)/h_n\}} \right) \mathbb{I}_{\{Y_i \leq y\}}$$

- *Estimation des quantiles conditionnels :*

$$q_{\alpha,n}(x) = F_n^{-1}(\alpha|x) = \inf\{y|F_n(y|x) \geq \alpha\}$$

- *Choix des paramètres de lissage :*

- Noyau normal : $K(u) = \phi(u) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{u^2}{2}\right)$, $u \in \mathbb{R}$,

- Largeur de fenêtre : littérature très abondante $\longrightarrow$ approche dérivée du critère de validation croisée (voir Yao (1999)) :

$$h_n = \arg\min_{h>0} \frac{1}{n} \sum_{t=1}^{n} \sum_{j=1}^{n} \left(\mathbb{I}_{\{Y_t \leq Y_j\}} - \tilde{F}_{n,-t}(Y_j|x) \right)^2,$$

où $\tilde{F}_{n,-t}(.|x)$ est l'estimateur de $F(.|x)$ calculé à partir de l'échantillon $\{(X_i, Y_i),\ i = 1, \ldots, n\}$ privé de la t-ème observation.

- *Propriétés mathématiques* (Stute (1986), Samanta (1989), Gannoun (1990)) :

$q_{\alpha,n}(.)$ converge vers $q_\alpha(.)$ simplement et uniformément sur C lorsque $n \longrightarrow \infty$.

(2) Méthode d'estimation par noyau produit

- *Estimation de la fonction de répartition conditionnelle* (*intégrale de la densité conditionnelle*) :

$$\hat{F}_n(y|x) = \sum_{i=1}^{n} \left(\frac{K\{(x - X_i)/h_n\}}{\sum_{j=1}^{n} K\{(x - X_j)/h_n\}} \right) \Omega \left(\frac{y - Y_i}{h_{2,n}} \right)$$

- *Estimation des quantiles conditionnels :* $\hat{q}_{\alpha,n}(x) = \hat{F}_n^{-1}(\alpha|x)$
- *Choix des paramètres de lissage :*

- Choix des noyaux : K=densité de $N(0,1)$, Ω=fct de répartition de $U[0,1]$.
- Largeur de fenêtres : $h_{1,n}$ et $h_{2,n}$ choisies par la règle empirique de Yu et Jones (1998).

- *Propriétés mathématiques* (Roussas (1991) et Berlinet *et al* (2000)) :

$\hat{q}_{\alpha,n}(.)$ converge vers $q_\alpha(.)$ simplement et uniformément sur C lorsque $n \longrightarrow \infty$.

(3) Méthode d'estimation par noyau de la constante locale

- *Estimation des quantiles conditionnels :*

$$q_{\alpha,n}(x) = \arg\min_{a\in\mathbb{R}} \sum_{i=1}^{n} L_\alpha(Y_i - a) K\left(\frac{x - X_i}{h_n^*}\right).$$

Avantage : bon comportement face aux effets de bords.

- *Choix des paramètres de lissage :*

- Choix du noyau : K=densité de la loi $N(0,1)$

- largeur de fenêtres : h_n^* choisie par la règle empirique de Yu et Jones (1998).

- *Propriétés mathématiques* (Stone (1977), Yu and Jones (1998), Berlinet et *al* (2000)) :

$q_{\alpha,n}(.)$ converge vers $q_\alpha(.)$ en probabilité, en moyenne quadratique et uniformément sur C.

Plan

1 Introduction générale

2 Introduction sur les courbes de référence

3 Caractérisation des quantiles

4 Estimation non-paramétrique des courbes de référence

5 Application à la construction de courbes de référence pour des propriétés biophysiques de la peau

6 Estimation semi-paramétrique des courbes de référence

Objectif de l'étude : établir des courbes de référence à 90% en fonction de l'âge pour de 35 propriétés biophysiques (variables d'intérêt) de la peau de femmes japonaises sur deux zones du visage (front et joue) et sur la face antérieure de l'avant-bras gauche.

- Etude réalisée par le CE.R.I.E.S. à Sendai (Japon) sur $n = 120$ femmes japonaises.
- *Propriétés biophysiques de la peau mesurées (11 ou 12 variables d'intérêt selon la zone considérée) :*
  - le taux de sécrétion de sébum,
  - la température cutanée,
  - la perte insensible en eau,
  - le pH cutané,
  - l'hydratation de la peau par capacitance et conductance,
  - la couleur de la peau,
  - l'angle typologique individuel.
- Covariable = âge des volontaires.

Méthodes : méthode paramétrique de Royston, méthodes non-paramétriques

Critère d'acceptabilité utilisé pour accepter ou rejeter *a posteriori* les courbes de référence obtenues

$\hookrightarrow$ 3 conditions à vérifier :

- Pas de valeurs impossibles pour Y (*i.e.* par exemple, des valeurs nulles ou négatives alors que Y ne peut prendre en réalité que des valeurs strictement positives).
- Pourcentage calculé d'individus entre les courbes $\simeq 90\%$
- Répartition uniforme en fonction de la covariable AGE des individus se trouvant en dehors des limites des courbes de référence (Pas de regroupement de valeurs individuelles).

Résultats obtenus :

- Résumés des résultats obtenus
  - par la méthode paramétrique,
  - les trois méthodes non paramétriques.
- Etude de deux variables particulières.

Très brève synthèse des résultats obtenus par la méthode paramétrique

▷ modèle paramétrique accepté = résidus normalement distribués (après transformation au préalable des données si nécessaire)

↪ construction des courbes de référence (acceptées ou pas)

▷ modèle paramétrique non accepté ⟶ pas de courbes de référence

Zones	Joue	Front	Avant-bras
Nombre de variables	12	12	11
Nombre de modèles acceptés	9	6	4
Nombre de courbes de référence acceptées	8	4	4

Très brève synthèse des résultats obtenus par les méthodes non paramétriques.

Méthode (1) = méthode d'estimation par noyau
Méthode (2) = méthode d'estimation par noyau produit
Méthode (3) = méthode d'estimation par noyau de la constante locale

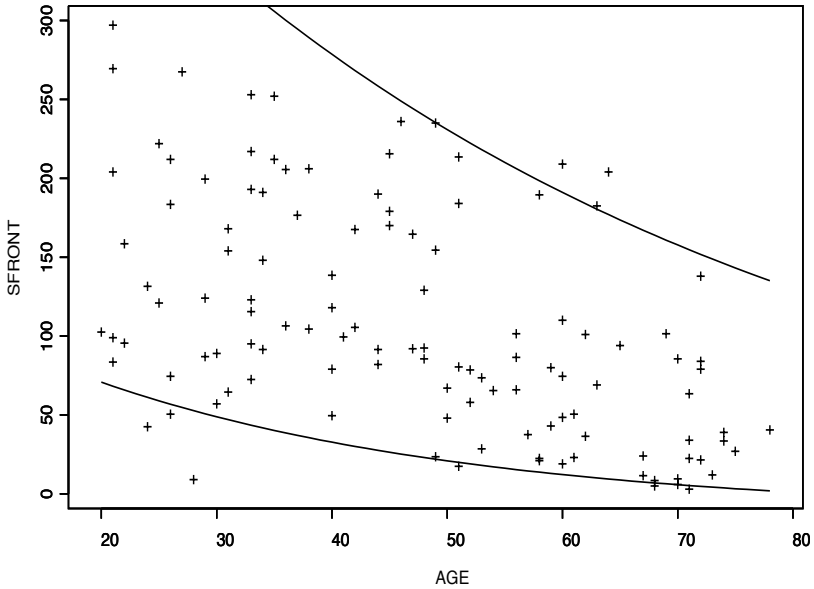
Zones	Joue	Front	Avant-bras
Nombre de variables	12	12	11
Méthode (1) : nbre de courbes de réf. acceptées	12	12	11
Méthode (2) : nbre de courbes de réf. acceptées	3	3	2
Méthode (3) : nbre de courbes de réf. acceptées	12	11	11

Problème avec la méthode (2) : règle empirique pour le choix de $h_{2,n}$ mal adaptée.

Brève synthèse globale des résultats obtenus :

- Méthode paramétrique de Royston : échec pour 54% des variables.
- Méthode non-paramétrique (1) : **aucun échec**.
- Méthode non-paramétrique (2) : échec pour 77% des variables (règle empirique de choix de $h_{2,n}$ mal adaptée).
- Méthode non-paramétrique (3) : échec pour 3% des variables.

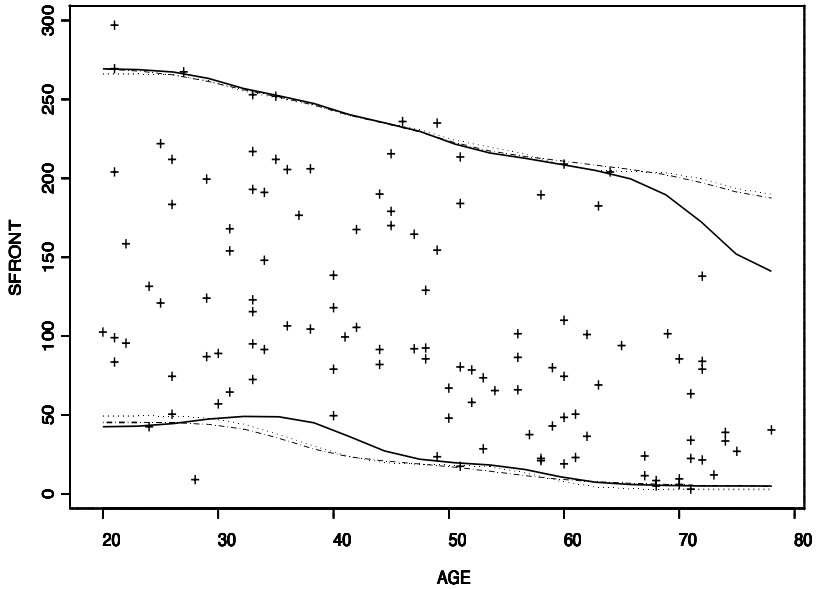
Etude de la variable "taux instantané de sébum" mesuré sur le front



Méthode paramétrique

courbes de réf. "physiologiquement" **non acceptables**

↪ valeurs négatives pour la courbe de réf. inférieure
↪ croissance trop rapide pour la courbe de réf. supérieure



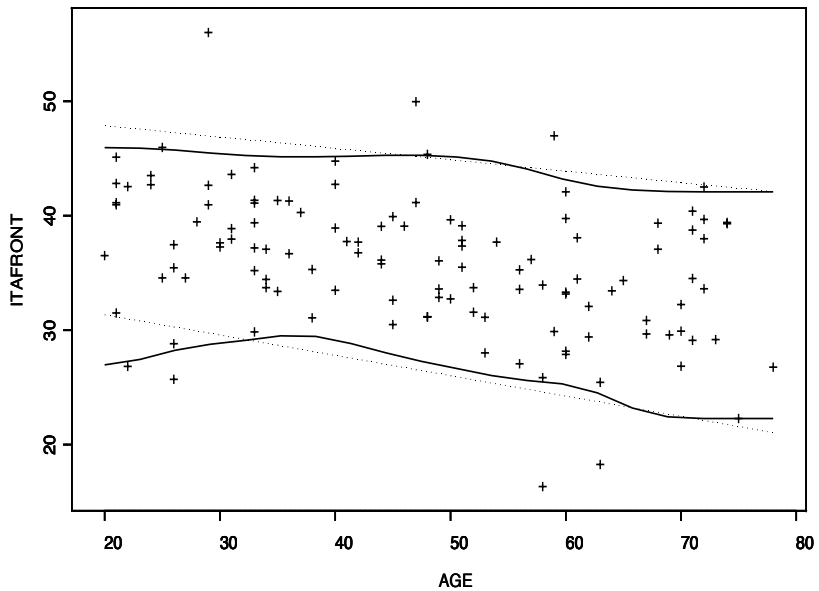
Méthodes non-paramétriques

courbes de réf. "physiologiquement" **acceptables**
—— : estimation par noyau
- - - : estimation par noyau produit
..... : estimation par noyau de la cste locale

Commentaire rapide :
- jusqu'à l'âge de 65 ans, comportement semblable des 3 courbes de référence
- après l'âge de 65 ans, décroissance plus rapide du premier estimateur : conforme à une diminution attendue du taux instantané de sébum avec l'âge, courbe de référence correspondante préférée aux deux autres.

Etude de la variable “angle typologique individuel du front”

$\Longrightarrow$ Approche paramétrique : pas de problèmes.
$\Longrightarrow$ Approches non paramétriques (méthodes (1) et (3)) : pas de problèmes.



- pointillés : méthode paramétrique,

- trait continu : premier estimateur non paramétrique.

Remarques sur l'approche non-paramétrique

Extension au cas où la dimension p de X augmente,

▷ pas de problème théorique,

▷ sur le plan pratique,

- fléau de la dimension
- "courbes" de référence = hyper-surfaces de $\mathbb{R}^{p+1}$ $\begin{cases} p = 2 & \longrightarrow \text{ graphique en 3D} \\ p > 2 & \longrightarrow \text{ pas de graphique} \end{cases}$

$\Longrightarrow$ **Idée :** réduire la dimension de la partie explicative

(1) sans perte d'information sur la distribution conditionnelle de Y sachant X
(2) sans spécifier un modèle paramétrique sous-jacent

$\hookrightarrow$ Développement d'une **approche semiparamétrique** avec une étape de réduction de la dimension de X au moyen d'indices $X'\beta_k,\ k = 1, \ldots, d_X$

Avantages : si la réduction de dimension est "efficace" ($p \rightarrow d_X << p$)

- possibilité de faire à nouveau des graphiques (pour $d_X = 1$ ou 2)
- estimation non paramétrique des quantiles conditionnels plus "aisée"

Plan

1 Introduction générale

2 Introduction sur les courbes de référence

3 Caractérisation des quantiles

4 Estimation non-paramétrique des courbes de référence

5 Application à la construction de courbes de référence pour des propriétés biophysiques de la peau

6 Estimation semi-paramétrique des courbes de référence

Contexte théorique de réduction de la dimension

- **Hypothèse :** Il existe une matrice $\beta = [\beta_1, \ldots, \beta_{d_X}]$ de dimension $p \times d_X$ ($d_X \leq p$) telle que

$$Y \perp X \mid \beta' X$$

Remarques :

- Une telle matrice existe toujours (cas trivial $\beta = I_p$)
- Si $d_X < p$, la réduction de dimension est effective.
- *Remarque sur l'identifiabilité de* β : β peut être remplacée par n'importe quelle matrice dont les colonnes forment une base de $S(\beta) = \text{Vect}(\beta)$
  Vocabulaire : $S(\beta) =$ "espace EDR (Effective Dimension Reduction)"

- **Conséquence directe :**

$$F(y|x) = F(y|\beta' x) \quad \Longrightarrow \quad q_\alpha(x) = q_\alpha(\beta' x)$$

Caractérisation de l'espace EDR : méthode SIR (**S**liced **I**nverse **R**egression)

- *Inverse* $\longrightarrow$ utilisation de propriétés géométriques des moments ("inverses") de X sachant Y :

$$\mathbb{E}[X|Y] \quad \text{ou/et} \quad \mathbb{V}[X|Y]$$

- *Sliced* $\longrightarrow$ discrétisation (ou "tranchage") de Y qui va permettre de simplifier l'écriture (et l'estimation) des moments intervenant dans les propriétés géométriques.

Avantages de la méthode SIR :

i) estimation des indices EDR sans avoir à estimer la fonction de lien
ii) calculs numériques très rapide (calculs matriciels + décomposition aux valeurs propres)

Remarque : une hypothèse raisonnable sur X

Procédure d'estimation des courbes de référence

Echantillon $= \{(X_i, Y_i),\ i = 1, \ldots, n\}$

- **Etape 1** : Estimation de l'espace EDR $\longleftarrow$ méthode SIR

$$\hookrightarrow \quad \text{base estimée } \left\{\hat{b}_k\right\}_{k=1}^{d_X}$$

Remarques : en pratique,
- d_X est estimée par $\hat{d}_X$.
- Simplification éventuelle des indices EDR afin d'en avoir une interprétation plus aisée.

- **Etape 2** : à partir de l'échantillon $= \{(\hat{b}'X_i, Y_i), i = 1, \ldots, n\}$ (ici pour $d_X = 1$)

$\triangleright$ Estimation non-paramétrique des quantiles conditionnels

$$\hookrightarrow \quad q_{\alpha,n}(\hat{b}'x) = F_n^{-1}(\alpha \mid \hat{b}'x)$$

▷ Courbes de référence estimées :

$$\left\{\left(x, q_{1-\alpha,n}(\hat{b}'x)\right)\right\} \quad \text{et} \quad \left\{\left(x, q_{\alpha,n}(\hat{b}'x)\right)\right\} \quad \text{lorsque } x \text{ varie.}$$

Propriétés asymptotiques : Sous des hypothèses techniques "usuelles", pour x de $\mathbb{R}^p$ fixé,

$$\sup_{y\in\mathbb{R}} \mid F_n(y \mid \hat{b}'x) - F(y \mid x) \mid \longrightarrow_{proba} 0$$

$$\Longrightarrow \qquad q_{\alpha,n}(\hat{b}'x) \longrightarrow_{proba} q_\alpha(x)$$

Extension au cas où Y est multidimensionnelle

Modèle sous-jacent : $Y \perp X \mid \beta' X$ avec $Y \in \mathbb{R}^q$

Approche semi-paramétrique directe :

- Etape 1 : Caractérisation de l'espace EDR $\longleftarrow$ méthode SIR multivariée
- Etape 2 : Evaluation des quantiles conditionnels **spatiaux** $\longleftarrow$ estimation par noyau

  Difficultés : estimation, interprétation et représentation graphique lorsque q est grand

$\Longrightarrow$ **Idée :** réduire simultanément la dimension de X et de Y au moyen d'indices.

$\hookrightarrow$ Nouvelle méthode semi-paramétrique :

- Etape 1 : Alternating SIR $\longrightarrow \left\{ \begin{array}{lcl} b_k' X & = & \text{indice EDR} \\ a_k' Y & = & \text{indice MP} \end{array} \right.$
- Etape 2 : Estimation par noyau des quantiles conditionnels unidimensionnels ou spatiaux

Exemple de définition de quantiles conditionnels spatiaux

Pour x de $\mathbb{R}^p$, on définit une fonction vectorielle de θ ($\theta \in \mathbb{R}^q$) par

$$\phi(\theta, x) = \mathbb{E}[||Y - \theta||_{2,\alpha} - ||Y||_{2,\alpha} \mid X = x] = \int_{\mathbb{R}^q} \left(||y - \theta||_{2,\alpha} - ||y||_{2,\alpha} \right) Q(dy|X = x)$$

où $||Z||_{2,\alpha} = ||(z_1, \ldots, z_q)||_{2,\alpha} = \left|\left| \left(\frac{|z_1| + (2\alpha - 1)z_1}{2}, \ldots, \frac{|z_q| + (2\alpha - 1)z_q}{2} \right) \right|\right|_2$.

De Gooijer *et al.* (2002) définissent le quantile spatial conditionnel d'ordre α par tout $\theta_\alpha(x)$ tel que

$$\phi(\theta_\alpha(x), x) = \inf_{\theta \in \mathbb{R}^q} \phi(\theta, x).$$

Application à l'analyse de propriétés biophysiques de la peau

Objectif de l'étude : établir des courbes de référence à 90% pour des propriétés biophysiques de la peau mesurées sur une zone du visage (front) en fonction de l'âge, de la taille et du poids de l'individu, ainsi que des conditions expérimentales.

Les données : Nouvelle étude conduite par le CE.R.I.E.S. sur $n = 316$ femmes françaises "caucasiennes" présentant une peau apparemment saine.

Variables d'intérêt ($q = 10$) :

- taux de sécrétion de sébum $\longrightarrow$ SFOREHEAD
- température cutanée $\longrightarrow$ TFOREHEAD
- perte insensible en eau $\longrightarrow$ FOREHEAD1
- pH cutané $\longrightarrow$ PFOREHEAD
- hydratation de la peau estimée $\longrightarrow$ C2FOREHEAD
- couleur de la peau $\longrightarrow$ AFOREHEAD, LFOREHEAD, BFOREHEAD
- conductance de la peau $\longrightarrow$ KFOREHEAD
- rugosité de la peau $\longrightarrow$ RFOREHEAD

Covariables ($p = 5$) :

- âge des volontaires $\longrightarrow$ AGE
- poids et taille des volontaires $\longrightarrow$ WEIGHT, HEIGHT
- conditions expérimentales $\longrightarrow$ TEMP, HYGRO

Etape 1* : Estimation (après sélection de variables) des indices EDR et MP

Une seule direction EDR et une seule direction MP ont été retenues dans la modélisation.

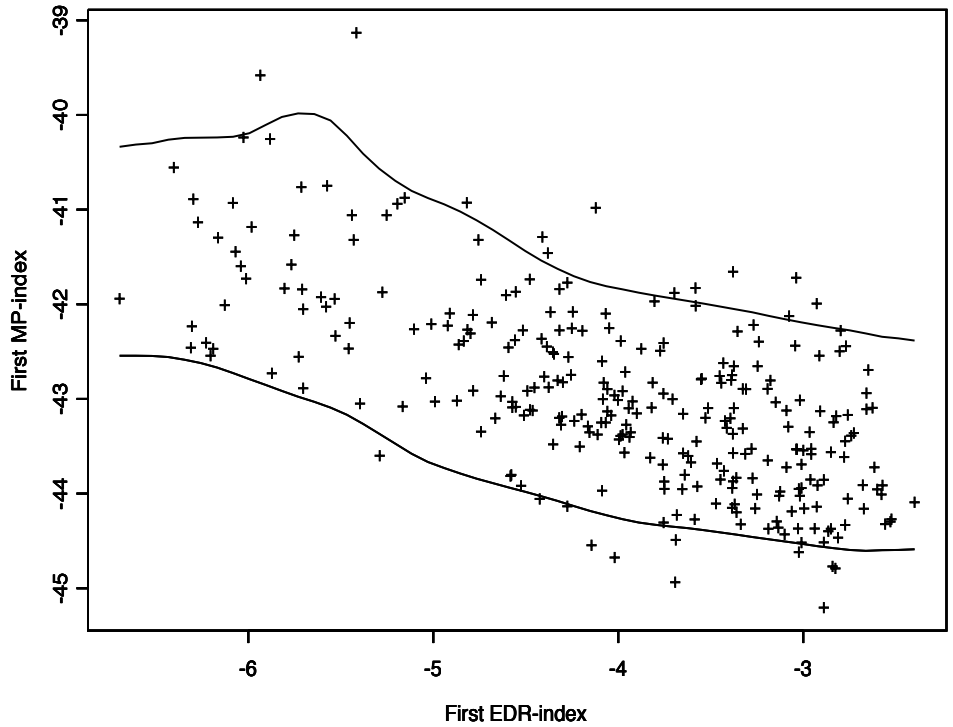
Selected covariates	EDR direction
AGE	-0.062
WEIGHT	-0.024

Selected variables of interest	MP direction
LFOREHEAD	-0.340
TFOREHEAD	-0.420
RFOREHEAD	+0.041
BFOREHEAD	-0.289
AFOREHEAD	-0.225
SFOREHEAD	-0.005
FOREHEAD1	-0.083
C2FOREHEAD	+0.008

- *Quelques rapides commentaires :*

- Seules 2 covariables ont été retenues dans l'indice EDR : les conditions expérimentales n'ont pas d'influence, idem pour la taille de l'individu
- Seules 8 propriétés biophysiques de la peau ont été retenues dans l'indice MP : le pH et la conductance de la peau n'interviennent pas dans l'indice.
- Les indices EDR et MP sont physiologiquement interprétables.

Etape 2 : Estimation des courbes de référence à 90%



Remarques finales

- Utilisation des quantiles conditionnels pour faire de la prévision (ici, on a vu un aspect "modélisation") : construction d'intervalle de prédiction I (à l'horizon H) tel que

$$P(\{Y_{t+H} \in I | \text{passé de } Y_t, \text{covariables}) = 90\%$$

$\hookrightarrow$ estimation de quantiles conditionnels d'ordres 5% et 95%.