

REVUE DE STATISTIQUE APPLIQUÉE

ALI GANNOUN

STÉPHANE GIRARD

CHRISTIANE GUINOT

JÉRÔME SARACCO

Trois méthodes non paramétriques pour l'estimation de courbes de référence - application à l'analyse de propriétés biophysiques de la peau

Revue de statistique appliquée, tome 50, n° 1 (2002), p. 65-89

http://www.numdam.org/item?id=RSA_2002__50_1_65_0

© Société française de statistique, 2002, tous droits réservés.

L'accès aux archives de la revue « *Revue de statistique appliquée* » (<http://www.sfds.asso.fr/publicat/rsa.htm>) implique l'accord avec les conditions générales d'utilisation (<http://www.numdam.org/conditions>). Toute utilisation commerciale ou impression systématique est constitutive d'une infraction pénale. Toute copie ou impression de ce fichier doit contenir la présente mention de copyright.

NUMDAM

Article numérisé dans le cadre du programme
Numérisation de documents anciens mathématiques

<http://www.numdam.org/>

TROIS MÉTHODES NON PARAMÉTRIQUES POUR L'ESTIMATION DE COURBES DE RÉFÉRENCE – APPLICATION À L'ANALYSE DE PROPRIÉTÉS BIOPHYSIQUES DE LA PEAU

Ali GANNOUN⁽¹⁾, Stéphane GIRARD⁽¹⁾,
Christiane GUINOT⁽²⁾ et Jérôme SARACCO⁽¹⁾

⁽¹⁾ *Laboratoire de Probabilités et Statistique, CC 051, Université Montpellier II,
Place Eugène Bataillon, 34095 Montpellier Cedex 5, France
Téléphone : +33-4-67-14-3505; Fax : +33-4-67-14-4974
e-mail : {gannoun, girard, saracco}@stat.math.univ-montp2.fr*

⁽²⁾ *CE.R.I.E.S, 20, Rue Victor Noir, 92 521 Neuilly sur Seine Cedex, France
Téléphone : +33-1-46-43-43-59; Fax : +33-1-46-43-46-00
e-mail : christiane.guinot@ceries-lab.com*

RÉSUMÉ

La construction de courbes de référence est un sujet fréquemment rencontré dans les domaines biomédicaux ou industriels. Ces courbes sont classiquement estimées par des méthodes paramétriques nécessitant des hypothèses contraignantes. Pour pallier cet inconvénient, des méthodes non paramétriques, fondées sur l'estimation de quantiles conditionnels, ont été développées. Dans cet article, nous présentons la mise en œuvre complète de trois d'entre elles. Leurs performances sont tout d'abord illustrées sur des simulations, puis par comparaison à une méthode paramétrique sur des données réelles concernant des propriétés biophysiques de la peau de femmes japonaises saines.

Mots-clés : *Quantiles Conditionnels, Estimateurs à Noyau, Intervalles de Référence, Courbes de Référence.*

ABSTRACT

Reference curves are required in various biomedical or industrial problems. These curves are classically estimated through parametric methods demanding strong assumptions. To overcome this limitation, nonparametric methods based on conditional quantile estimation have been developed. In this paper, we present the complete implementation of three of them. Their performances are both illustrated by a simulation study and by comparison with a specific parametric method on a data set collected to establish reference curves for biophysical properties of the skin of healthy japanese women.

Keywords : *Conditional Quantiles, Kernel Estimators, Reference Ranges, Reference Curves.*

1. Introduction

De nombreuses expérimentations, en particulier dans le cadre d'études biomédicales, sont conduites pour établir des intervalles de valeurs qui sont prises «normalement» par une variable d'intérêt, que l'on notera Y , dans une population cible. Ici, le terme «normalement» fait référence aux 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 (les sujets de référence). Ces intervalles sont souvent appelés intervalles de référence et les valeurs correspondantes sont appelées valeurs de référence. Par exemple, on peut s'intéresser à un intervalle excluant les 5% d'observations les plus grandes et les 5% d'observations les plus petites. Ainsi, la construction d'intervalles de référence repose sur le calcul de quantiles.

Par ailleurs, il arrive régulièrement que, sur la population cible, l'on dispose simultanément, avec la variable d'intérêt Y , d'une information complémentaire sous la forme d'une covariable X . Très souvent, X représente l'âge du sujet. Pour une valeur donnée x de X , on peut construire un intervalle de référence. Lorsque x varie, on obtient alors des courbes de référence. Dans ce cadre-là, il est nécessaire de travailler avec les quantiles conditionnels de Y sachant X . Le tracé de courbes de référence sur le nuage des valeurs prises par le couple (X, Y) pour les sujets de référence donne un résumé graphique très utile et interprétable. Ainsi, un individu i représenté par le point (x_i, y_i) pourra être comparé à la population de référence. En d'autres termes, une «anormalité» de cet individu sera suspectée si ce point se situe en dessous de la courbe de référence inférieure ou au dessus de la courbe de référence supérieure. On trouve dans la littérature de nombreux articles sur le sujet comme ceux de Healy, Rasbash et Yang (1988), Cole (1988), Goldstein et Pan (1992) ou Royston et Altman (1992) dans les domaines biomédical et biométrique, ou encore Magee, Burbidge et Robb (1991), Hendricks et Koenker (1992) dans le domaine industriel.

• Quantiles conditionnels et courbes de référence

Plus précisément, le quantile conditionnel d'ordre $\alpha \in]0,5, 1[$ de la variable Y sachant que $X = x$ est défini de la manière suivante :

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

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

Une caractérisation alternative de $q_\alpha(x)$ est obtenue sous forme d'un problème d'optimisation (\mathcal{P}) :

$$q_\alpha(x) = \arg \min_{\theta \in \mathbb{R}} E \{ \rho_\alpha(Y - \theta) | X = x \},$$

où ρ_α est la fonction définie par :

$$\rho_\alpha(z) = \alpha z \mathbb{I}_{(0, \infty)}(z) - (1 - \alpha) z \mathbb{I}_{(-\infty, 0)}(z).$$

La théorie sous-jacente est décrite par exemple dans Basset et Koenker (1982), Jones et Hall (1990), Chaudhuri (1991) et Yu (1997).

Pour une valeur x donnée, l'intervalle de référence contenant $100(2\alpha - 1)\%$ des sujets de référence est ensuite défini par

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

Les courbes de référence sont alors les ensembles de points $\{(x, q_{1-\alpha}(x))\}$ et $\{(x, q_\alpha(x))\}$ lorsque x varie.

Soit $q_{n,\alpha}(x)$ un estimateur de $q_\alpha(x)$ à partir de l'échantillon $\{(X_i, Y_i), i = 1, \dots, n\}$ de n réalisations indépendantes du couple de variables aléatoires (X, Y) . L'estimateur correspondant de $I_\alpha(x)$ est défini par

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

En pratique, pour obtenir les courbes de référence à 90%, α est choisi égal à 0.95.

Deux types d'approches ont été développées pour l'estimation des quantiles conditionnels.

• Approche paramétrique

Quand l'échantillon est de petite taille, des hypothèses paramétriques sont habituellement imposées afin de réduire le nombre de paramètres à estimer. En particulier, la fonction de répartition conditionnelle est supposée gaussienne. Ainsi, un estimateur du quantile conditionnel $q_\alpha(x)$, est donné par $\hat{m}_n(x) + z_\alpha \hat{\sigma}_n(x)$, où $\hat{m}_n(x)$ et $\hat{\sigma}_n^2(x)$ désignent respectivement les estimateurs de l'espérance conditionnelle $m(x) = E[Y|X = x]$ et de la variance conditionnelle $\sigma^2(x) = V(Y|X = x)$, et où z_α est le quantile d'ordre α de la loi normale centrée réduite $N(0, 1)$.

Un modèle linéaire ou polynomial (voir Royston (1991), Royston et Wright (1998)) associé à la méthode des moindres carrés est généralement utilisé pour estimer $m(x)$ et $\sigma^2(x)$. Cependant ces estimations sont très sensibles aux valeurs aberrantes de Y . De plus, il peut être nécessaire de transformer les données de départ dans l'espoir d'obtenir des résidus normalement distribués avec cette nouvelle échelle (voir, par exemple, Cole (1988), Tango (1998), Wright et Royston (1997)). C'est très souvent le cas avec des données biologiques. L'existence d'une telle transformation n'est nullement garantie. Il est aussi connu que les hypothèses paramétriques sont restrictives et peuvent rarement être faites avec certitude (voir Cole (1988)).

Ainsi, l'approche paramétrique peut être mal adaptée à la réalité des données en particulier biologiques. Une approche non paramétrique du problème a alors été développée afin de pallier ces problèmes d'hypothèses et de modélisation paramétriques.

• Approche non paramétrique

De nombreux travaux récents ont été menés pour l'estimation non paramétrique des quantiles conditionnels aussi bien dans un cadre théorique que sur le plan des applications (voir, par exemple, Bhattacharya et Gangopadhyay (1990), Ducharme *et al* (1995), Jones et Hall (1990), Poiraud-Casanova et Thomas-Agnan (1998) ou Poiraud-Casanova (2000)). Ces méthodes de régression ne nécessitent pas d'hypothèse sur la nature de la distribution.

A titre d'exemple, Horn *et al* (1998) proposent une approche robuste permettant d'obtenir des courbes de référence. Cette méthode est basée sur une partition du support de X en intervalles. Pour chaque intervalle, les deux quantiles $q_\alpha(x)$ et $q_{1-\alpha}(x)$ sont estimés de façon usuelle. L'ensemble des quantiles estimés est alors lissé selon x pour obtenir les courbes de référence.

L'approche non paramétrique que nous exposons ici ne nécessite pas de partition du support de X . De plus, elle est robuste dans le sens où les courbes de référence calculées sont peu sensibles à la présence de points aberrants.

Dans cet article, trois méthodes non paramétriques d'estimation des quantiles conditionnels sont présentées :

- une méthode d'estimation par noyau,
- la méthode de la constante locale (« local constant kernel estimation »),
- une méthode d'estimation par noyau produit (« double kernel estimation »).

Ces méthodes sont ensuite comparées entre elles ainsi qu'à la méthode paramétrique de Royston (1991). La suite de cet article est organisée de la façon suivante. Nous présentons dans la partie 2 les estimateurs non paramétriques des quantiles conditionnels mentionnés ci-dessus. Nous précisons également le choix des différents paramètres de lissage intervenant dans ces estimateurs. La partie 3 est consacrée à une étude comparative des différents estimateurs non paramétriques. Dans la partie 4, nous exposons une application visant à établir des courbes de référence, en fonction de l'âge, des propriétés biophysiques de la peau de femmes japonaises sur deux zones du visage et une zone de l'avant-bras. Nous comparons l'approche non paramétrique préconisée à l'approche paramétrique de Royston (1991). Enfin, dans une dernière partie, nous donnons quelques conclusions et extensions possibles de ce travail.

2. Méthodes non paramétriques d'estimation des quantiles conditionnels

Par souci de simplicité, nous supposons désormais que X est une variable aléatoire réelle de support compact C .

Nous décrivons ci-dessous les trois méthodes d'estimation des quantiles conditionnels envisagées.

2.1. Méthode d'estimation par noyau

Définissons tout d'abord un estimateur non paramétrique de la fonction de répartition conditionnelle de Y sachant $X = x$, pour $y \in \mathbb{R}$:

$$\tilde{F}_n(y|x) = \sum_{i=1}^n K\{(x - X_i)/h_n\} \mathbb{I}_{\{Y_i \leq y\}} \Big/ \sum_{i=1}^n K\{(x - X_i)/h_n\} \quad (2)$$

La fonction K , appelée noyau, est mesurable sur \mathbb{R} et d'intégrale égale à un. Le paramètre h_n permet de contrôler le lissage appliqué aux données. Son choix en pratique est discuté dans le paragraphe 2.4.

Il est alors naturel d'estimer le quantile conditionnel $q_\alpha(x)$ par $\tilde{q}_{\alpha,n}(x)$ défini de la manière suivante :

$$\tilde{q}_{\alpha,n}(x) = \tilde{F}_n^{-1}(\alpha|x) = \inf\{y|\tilde{F}_n(y|x) \geq \alpha\}.$$

Sous certaines conditions de régularité, Stute (1986), Samanta (1989) et Gannoun (1990) ont montré que :

$$\tilde{q}_{\alpha,n}(\cdot) \rightarrow q_\alpha(\cdot),$$

simplement et uniformément sur C lorsque $n \rightarrow \infty$.

2.2. Méthode de la constante locale

Une approche linéaire locale a été développée pour résoudre le problème (\mathcal{P}). Le quantile inconnu est approché par une fonction linéaire $q_\alpha(z) \approx q_\alpha(x) + q'_\alpha(x)(z - x) \equiv a + b(z - x)$ pour z dans un voisinage de x , la notation q'_α désignant la dérivée de q_α . Localement, estimer $q_\alpha(x)$ est alors équivalent à estimer le coefficient a , et estimer $q'_\alpha(x)$ revient à estimer b .

Ainsi, de façon tout à fait naturelle, on peut définir des estimateurs de $q_\alpha(x)$ et $q'_\alpha(x)$ en minimisant par rapport à a et b la quantité

$$\sum_{i=1}^n \rho_\alpha(Y_i - a - b(X_i - x))K\left(\frac{x - X_i}{h_n}\right),$$

où h_n et K désignent la fenêtre et le noyau mentionnés précédemment.

Si $b = 0$, on définit la méthode dite de la constante locale, et on obtient, pour $q_\alpha(x)$, l'estimateur suivant :

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

Cette méthode directe d'estimation présente en particulier l'avantage d'un bon comportement face aux effets de bords (voir Chaudhuri (1991), Fan *et al* (1994), Koenker *et al* (1992) et Mint el Mouvit (2000)). De plus, sous des conditions générales, la convergence de cet estimateur est obtenue sans étudier préalablement la convergence de l'estimateur de la fonction de répartition conditionnelle. En particulier, d'après Stone (1977) et Yu et Jones (1998), on a

$$\bar{q}_{\alpha,n}(\cdot) \rightarrow q_\alpha(\cdot),$$

en probabilité et en moyenne quadratique lorsque $n \rightarrow \infty$. De plus, en adaptant un résultat de Berline *et al* (2001), on obtient la convergence uniforme sur C .

2.3. Méthode d'estimation par noyau produit

Une version plus « lisse » de l'estimateur de la fonction de répartition conditionnelle définie en (2) peut être introduite en remplaçant la fonction indicatrice par une nouvelle densité symétrique ω . L'estimateur correspondant, appelé estimateur par noyau produit est défini comme suit pour $x \in C$ et $y \in \mathbb{R}$:

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

où Ω est la fonction de répartition associée à ω . Cet estimateur peut également être vu comme l'intégrale de l'estimateur à noyau de la densité conditionnelle (voir par exemple Berlinet *et al* (2001)). Il en découle naturellement un estimateur $\hat{q}_{\alpha,n}(x)$ du quantile conditionnel défini par

$$\hat{q}_{\alpha,n}(x) = \hat{F}_n^{-1}(\alpha|x) = \inf\{y \mid \hat{F}_n(y|x) \geq \alpha\}.$$

Cette approche est attractive mais nécessite le choix de deux paramètres de lissage $h_{1,n}$ et $h_{2,n}$. Il apparaît en pratique que cet estimateur est extrêmement sensible au choix de ces deux paramètres. Une méthode empirique pour choisir $h_{1,n}$ et $h_{2,n}$ a été proposée par Yu et Jones (1998), et sera décrite dans le paragraphe 2.4. Du point de vue théorique, Roussas (1991) et Berlinet *et al* (2001) ont montré, sous des conditions de régularité, que

$$\hat{q}_{\alpha,n}(\cdot) \rightarrow q_{\alpha}(\cdot),$$

simplement et uniformément sur C lorsque $n \rightarrow \infty$.

2.4. Choix des noyaux et des paramètres de lissage

Nous indiquons dans ce paragraphe des choix possibles de noyaux et de fenêtres.

• Choix des noyaux

La qualité des estimateurs n'étant pas très affectée par le choix des noyaux (voir par exemple Deheuvels (1977)), nous utiliserons dans toutes les applications, les noyaux K et ω suivants :

- le noyau normal : $K(u) = \phi(u) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{u^2}{2}\right)$, $u \in \mathbb{R}$,
- le noyau uniforme : $\omega(v) = \frac{1}{2} \mathbb{I}\{|v| \leq 1\}$, $v \in \mathbb{R}$.

• Choix des paramètres de lissage

Le choix de la fenêtre est quant à lui crucial. La qualité des estimateurs non paramétriques basés sur les noyaux y est étroitement liée. Une importante littérature est consacrée à ce sujet, et en particulier aux méthodes de sélection automatique par minimisation d'un critère. La méthode de validation croisée entre dans ce cadre, voir

par exemple Härdle (1990). D'autres propositions intéressantes dans ce domaine se trouvent dans l'article de Fan et Gijbels (1992).

Nous avons retenu les choix suivants pour les différentes fenêtres intervenant dans chacun des estimateurs.

- Pour l'estimateur $\tilde{q}_{\alpha,n}(x)$, une approche dérivée du critère de validation croisée est utilisée (voir Yao (1999)) :

$$h_n = \arg \min_{h>0} \sum_{t=1}^n \int_{\mathbb{R}} \{\mathbb{I}_{\{Y_t \leq y\}} - \tilde{F}_{n,-t}(y|x)\}^2 \omega(y) dy, \quad (3)$$

où $\tilde{F}_{n,-t}(\cdot|x)$ est l'estimateur de $F(\cdot|x)$ défini au paragraphe 2.1 mais calculé à partir de l'échantillon $\{(X_i, Y_i), i = 1, \dots, n\}$ privé de la t -ème observation.

- Pour l'estimateur $\bar{q}_{n,\alpha}(x)$, Yu et Jones (1998) proposent la règle empirique suivante,

$$h_n = h_{moy} \left(\frac{\alpha(1-\alpha)}{\phi(\Phi^{-1}(\alpha))^2} \right)^{1/5}, \quad (4)$$

où h_{moy} est la fenêtre obtenue par validation croisée dans le cadre de la régression à noyau de Y sur X . Cette règle repose sur l'hypothèse de normalité de la loi conditionnelle de Y sachant X . Les fonctions ϕ et Φ sont respectivement la densité et la fonction de répartition de la loi normale centrée réduite. Dans les cas où $\alpha = 0,05$ et $\alpha = 0,95$, la table 1 de Yu et Jones (1998) donne $h_n = 1,34h_{moy}$.

- Pour l'estimateur $\hat{q}_{\alpha,n}(x)$, le choix de deux fenêtres est nécessaire. Nous utilisons encore les méthodes proposées par Yu et Jones (1998). Le choix $h_{1,n}$ de la première fenêtre est identique à celui de la fenêtre h_n décrit par la formule (4). Pour la sélection de la fenêtre $h_{2,n}$, la règle suivante a été adoptée :

$$h_{2,n} = \begin{cases} \max \left(\frac{h_{0,n}^5}{h_{1,n}^3}, \frac{h_{1,n}}{10} \right) & \text{si } h_{0,n} < 1, \\ \frac{h_{0,n}^4}{h_{1,n}^3} & \text{sinon.} \end{cases},$$

où $h_{0,n} = h_{moy}(\pi/2)^{1/5}$ est l'évaluation de l'expression (4) lorsque $\alpha = 1/2$.

L'utilisation de telles règles empiriques présente l'avantage d'une mise en œuvre simple et rapide. Cependant cet avantage est acquis au prix d'une perte de généralité due à l'ajout d'une hypothèse de normalité. Dans la partie suivante, nous évaluons la robustesse vis-à-vis de cette hypothèse sur simulations avant de mettre en œuvre ces estimateurs sur des données réelles concernant des propriétés biophysiques de la peau de femmes japonaises.

3. Comparaison sur simulations des trois méthodes non paramétriques

Dans cette partie, nous évaluons les trois méthodes non paramétriques décrites précédemment sur des données simulées mais réalistes au vu de l'application sur données réelles considérées dans la partie 4. Plus précisément, les choix de fenêtre des estimateurs 2 et 3 étant basés sur des considérations de normalité, nous nous proposons d'évaluer leur comportement vis à vis de cette hypothèse.

3.1. Description du modèle simulé

Nous considérons le modèle suivant :

$$y = 16(x - 0.5)^2 + \varepsilon, \quad (5)$$

où le terme d'erreur ε est égal à $\varepsilon^* - a\lambda$ avec ε^* suivant une loi Gamma de paramètres positifs (λ, a) dont la densité est définie pour tout $x > 0$ par

$$f(x) = \frac{1}{\lambda^a} \frac{1}{\Gamma(a)} x^{a-1} \exp(-x/\lambda).$$

Rappelons que $E[\varepsilon^*] = a\lambda$, $V(\varepsilon^*) = a\lambda^2$ et $S(\varepsilon^*) = 2/\sqrt{a}$, S désignant le coefficient d'asymétrie, dit «Skewness» en anglais.

Dans la suite de la simulation, on choisit $a = 1/\lambda^2$ afin de travailler avec un terme d'erreur de variance unité. Dans ce cadre particulier, $\varepsilon \in [-1/\lambda, +\infty[$, l'espérance de ε^* étant égale à $1/\lambda$, le terme d'erreur ε est centré. De plus, le coefficient d'asymétrie est alors égal à 2λ . Ainsi, plus λ est grand, plus la distribution de l'erreur est dissymétrique et donc s'éloignera de la normalité. Nous étudierons les cas où $\lambda = 0.5, 1, 1.5, 2, 2.5$ et 3 . Les données réelles que nous avons étudiées dans la partie suivante, présentent, pour la plupart des variables d'intérêt, une répartition du nuage de points du même type que $\lambda = 1$ ou 1.5 dans notre modèle (5). Dans le cadre de ce modèle, le quantile conditionnel d'ordre α s'écrit ici explicitement sous la forme :

$$q_\alpha(x) = 16(x - 0.5)^2 + q_\alpha^{(\Gamma)} - 1/\lambda,$$

où $q_\alpha^{(\Gamma)}$ est le quantile d'ordre α de la loi Gamma de paramètres $(\lambda, 1/\lambda^2)$.

Pour chacune des valeurs de λ considérées, nous avons simulé $N = 50$ échantillons de taille $n = 200$ selon le modèle (5) avec la covariable X générée selon la loi uniforme discrète sur $\{j/30, j = 0, 1, \dots, 30\}$. Le choix de l'uniforme discrète se justifie par le fait que, sur les données réelles étudiées, la covariable est l'âge (arrondi à l'année) des individus. Pour chaque échantillon simulé, nous avons ensuite estimé $q_\alpha(z_j)$, pour $z_j = j/50, j = 0, 1, \dots, 50$, par chacune des trois méthodes non paramétriques. Afin d'évaluer les différentes méthodes, nous avons calculé pour chaque échantillon ($k = 1, \dots, 50$) l'Erreur Relative Médiane, notée

$ERM(k)$:

$$ERM(k) = \text{médiane} \left\{ \left| \frac{q_{\alpha,n}^{(k)}(z_j) - q_{\alpha}(z_j)}{q_{\alpha}(z_j)} \right|, j = 0, 1, \dots, 50 \right\},$$

où $q_{\alpha,n}^{(k)}(z_j)$ correspond à l'estimation de $q_{\alpha}(z_j)$ obtenue par la méthode non paramétrique considérée sur l'échantillon k . Le critère global pour mesurer l'erreur d'estimation utilisé est le suivant :

$$MERM = \text{médiane} \{ERM(k), k = 1, \dots, 50\}.$$

Nous avons préféré la médiane à la moyenne pour estimer la tendance centrale de l'erreur en raison de sa robustesse.

3.2. Résultats et commentaires

Un aperçu qualitatif du type de résultats obtenus pour les 3 estimateurs est présenté figure 1. Visuellement, lorsque la dissymétrie de la distribution de l'erreur augmente, les performances des estimateurs 2 et 3 se dégradent pour le quantile d'ordre 5% alors que celles de l'estimateur 1 restent correctes. Ces conclusions sont inversées en ce qui concerne le quantile d'ordre 95%.

La figure 2 permet une vision plus quantitative de ce phénomène en montrant l'évolution des $MERM$ obtenues pour les différents estimateurs en fonction du degré d'asymétrie du bruit. La figure 3 complète ces résultats en présentant les distributions des ERM sous forme de boîtes à moustaches. La dégradation des performances des estimateurs 2 et 3 avec l'augmentation de l'asymétrie du bruit n'est pas surprenante, le choix des fenêtres étant basé sur une hypothèse de normalité. La détérioration des résultats est surtout sensible pour l'estimation des quantiles d'ordre 5%, c'est à dire pour l'estimation des quantiles proches de l'extrémité finie $(-1/\lambda)$ du support de la loi du bruit. Ceux-ci sont globalement sous-estimés, le plus souvent situés en deçà du support ($q_{200, 0.05} < -1/\lambda$). Pour éliminer ce problème, l'unique solution consisterait à changer de règle de choix de largeur de fenêtre. Nous proposons quelques possibilités en conclusion.

Par contre, l'hypothèse de normalité n'est pas requise pour le choix de la fenêtre lors de la mise en œuvre de l'estimateur 1. Cependant, même si la qualité de l'estimation du quantile d'ordre 5% reste satisfaisante, la qualité de l'estimation du quantile d'ordre 95% chute lorsque la dissymétrie augmente beaucoup. Ce phénomène s'explique par le très faible nombre d'observations dans la partie non bornée du support de la loi. L'estimation du quantile conditionnel, purement non paramétrique, souffre alors d'une grande sensibilité vis-à-vis de ces rares observations. Cette limitation peut être dépassée en effectuant un lissage des courbes de quantiles conditionnels selon des abscisses. Les effets de ce lissage sont illustrés sur données réelles dans la partie suivante.

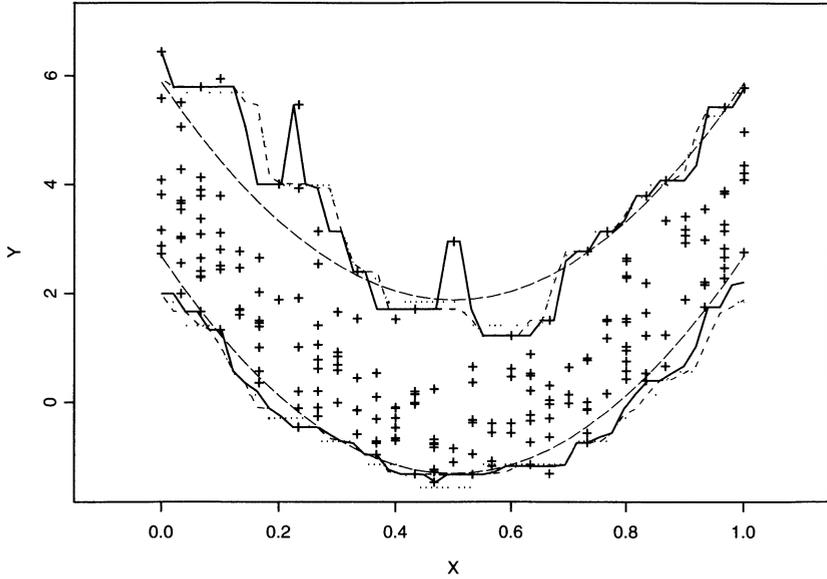
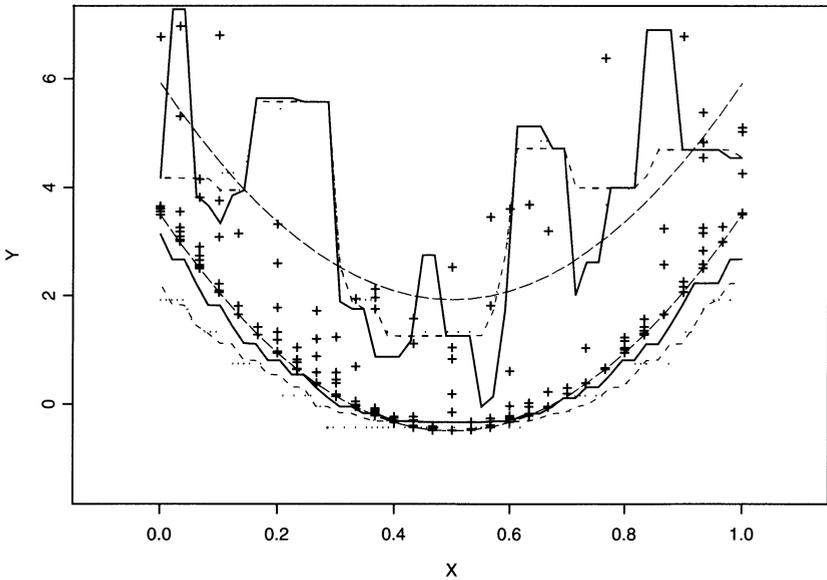
(a) Coefficient d'asymétrie $S = 1$ (b) Coefficient d'asymétrie $S = 4$

FIGURE 1

Courbes de référence à 5% et 95% obtenues. (trait continu : méthode d'estimation par noyau, pointillés : méthode de la constante locale, petits tirets : méthode d'estimation par noyau produit, grands tirets : quantiles théoriques.)

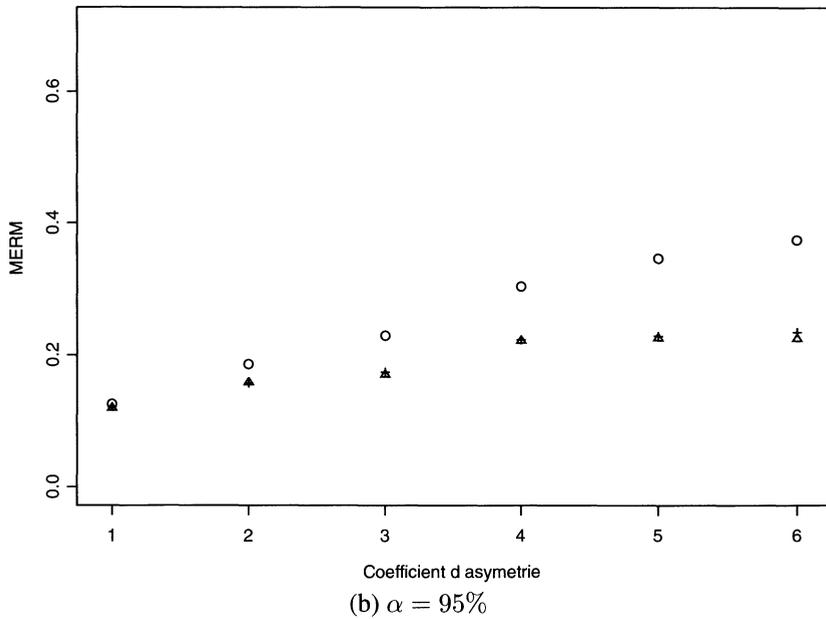
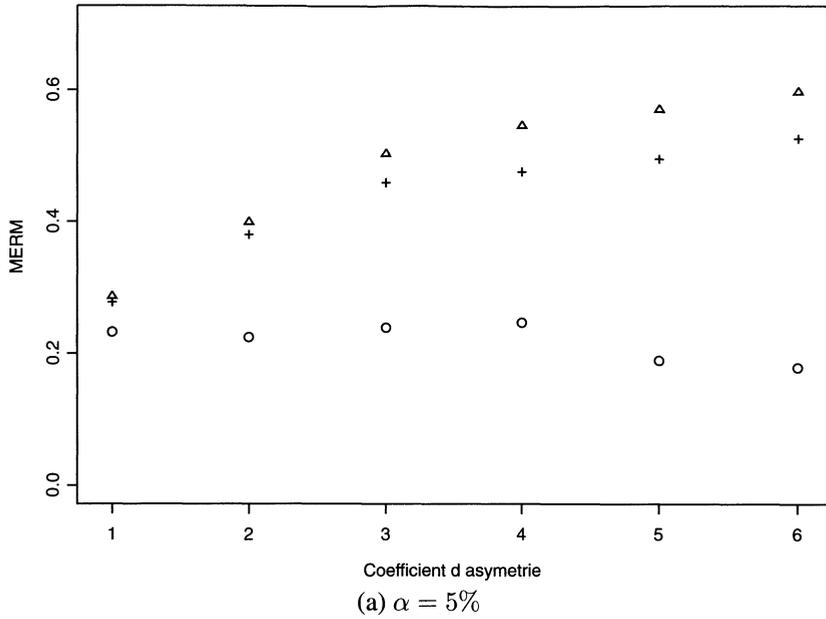


FIGURE 2

Evolution de la Médiane de l'Erreur Relative Médiane en fonction de l'asymétrie de l'erreur (ronds : méthode d'estimation par noyau, triangles : méthode de la constante locale, croix : méthode d'estimation par noyau produit.)

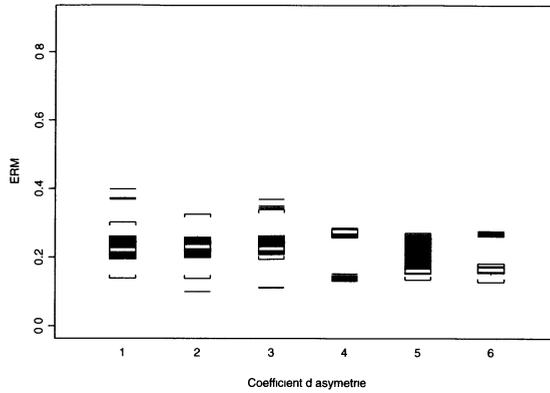
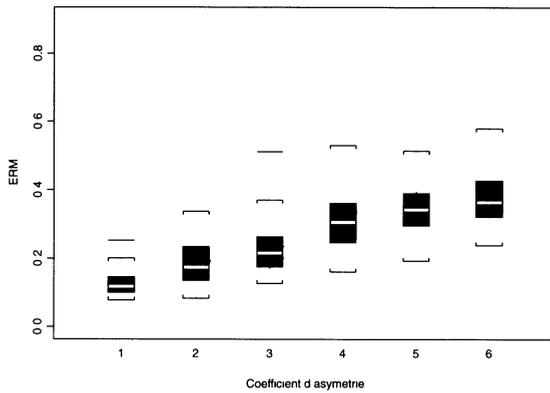
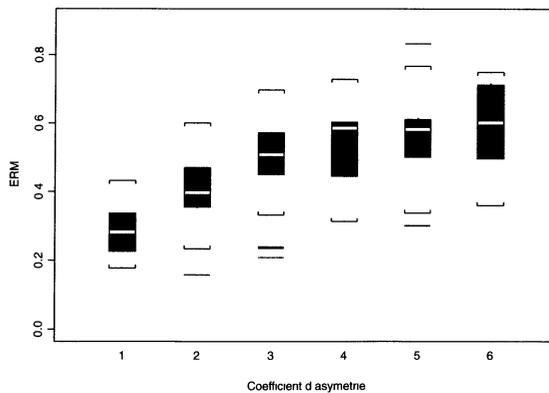
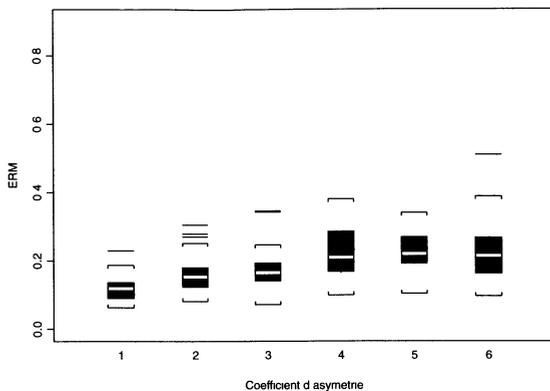
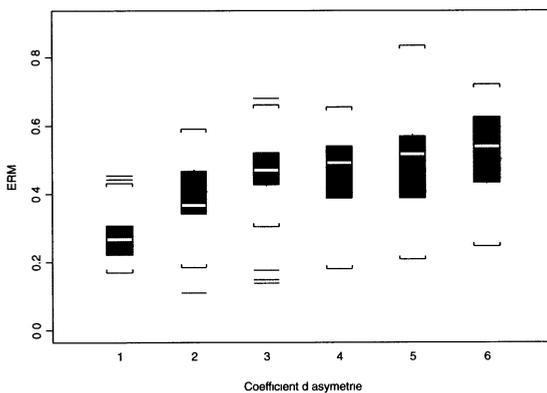
(a) Estimateur 1, $\alpha = 5\%$ (b) Estimateur 1, $\alpha = 95\%$ (c) Estimateur 2, $\alpha = 5\%$

FIGURE 3

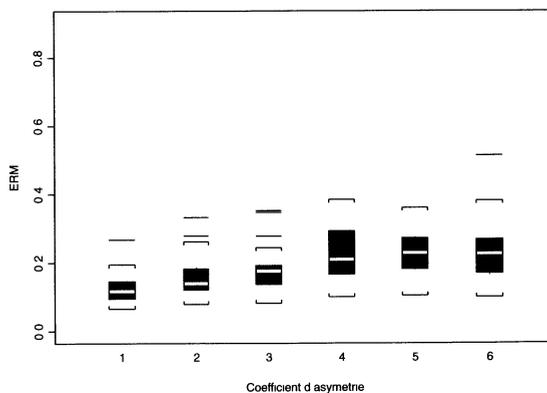
Boîtes à moustaches pour les différentes Erreurs Relatives Médianes



(d) Estimateur 2, $\alpha = 95\%$



(e) Estimateur 3, $\alpha = 5\%$



(f) Estimateur 3, $\alpha = 95\%$

FIGURE 3

Boîtes à moustaches pour les différentes Erreurs Relatives Médianes

4. Application à des données réelles

Les connaissances relatives aux courbes de référence ne sont documentées que par très peu d'études souvent réalisées avec des méthodologies non appropriées. Notre objectif est d'établir des courbes de référence à 90% en fonction de l'âge pour les propriétés biophysiques de la peau de femmes japonaises saines sur deux zones du visage et une zone de l'avant-bras.

4.1. Les données

Une étude réalisée par le CE.R.I.E.S a été conduite entre le 15 décembre 1998 et le 15 avril 1999 à Sendai (Japon) sur $n = 120$ femmes japonaises âgées de 20 à 80 ans présentant une peau apparemment saine (c'est-à-dire sans aucun signe de dermatose en cours ou de maladie générale avec manifestations cutanées avérées). Chaque volontaire a été examinée en atmosphère contrôlée (température de $23 \pm 1^\circ\text{C}$ et humidité relative de $50 \pm 5\%$). Cette étude comportait des questionnaires sur les habitudes de vie, un interrogatoire et un examen médical cutané, ainsi qu'une évaluation des propriétés biophysiques cutanées. Les propriétés biophysiques de la peau incluaient : le taux de sécrétion de sébum (taux instantané de lipides ($\mu\text{g}/\text{m}^2$), SM810PC Sebumeter©, Courage et Khazaka Electronic GmbH, Köln, Germany), la température cutanée (exprimée en degrés Celsius, thermomètre différentiel PT200, IMPO Electronics, Denmark), la perte insensible en eau (TWL ($\text{g}/\text{m}^2\text{h}$), EP1 Evaporimeter©, Servomed, Kinna, Sweden), le pH cutané (pHmeter©, Courage et Khazaka Electronic GmbH, Koln, Germany), l'hydratation de la peau par estimée par la capacitance (Corneometer© CM820 et CM825, Courage et Khazaka Electronic GmbH, Koln, Germany), et la conductance (Skicon©, I.B.S., Hamamatsu-shi, Japan), et la couleur de la peau (Chromameter© CR-300, Minolta, Osaka, Japan). La couleur a été exprimée à l'aide des trois paramètres du système $L^*a^*b^*$ CIE 1976, où L^* exprime la luminosité, a^* les coordonnées de chromacité rouge/vert et b^* les coordonnées de chromacité jaune/bleu. Les deux derniers paramètres sont la saturation (C) et l'angle de teinte (H), exprimé en degrés. L'angle typologique individuel (ITA exprimé en degrés) a été calculé pour chaque volontaire et chaque zone (voir Chardon *et al* (1991)). L'évaluation des paramètres biophysiques a été effectuée sur deux zones du visage (front et joue) et sur la face antérieure de l'avant-bras gauche, sauf pour le taux de sébum mesuré uniquement sur les deux zones du visage.

En résumé, les variables d'intérêt sont au nombre de 11 sur la face antérieure de l'avant-bras gauche et de 12 sur le front et la joue. La covariable est l'âge des volontaires.

4.2. Estimation paramétrique

La méthodologie utilisée pour l'établissement de valeurs de référence des mesures biophysiques cutanées en fonction de l'âge des sujets découle de celle proposée par Royston (1991). Cette méthodologie comporte six étapes légèrement modifiées afin de l'adapter aux données. Nous les décrivons brièvement ci-dessous :

- Représentation graphique de la variable d'intérêt en fonction de l'âge : elle permet de déceler d'éventuels sujets extrêmes, c'est-à-dire des sujets présentant des valeurs supposées être trop élevées ou trop faibles, ainsi que des regroupements de sujets.
- Sélection du modèle par régression polynomiale ascendante.
- Vérification de l'ajustement du modèle aux données grâce aux valeurs résiduelles.
- Transformation éventuelle des données pour normaliser les valeurs résiduelles.
- Calcul des valeurs de référence et de l'intervalle de référence en tenant compte d'une éventuelle hétéroscédasticité de la mesure en fonction de l'âge.
- Vérification de l'acceptabilité des courbes de référence obtenues avec le modèle final (voir au paragraphe 4.4 les critères d'acceptabilité des courbes de référence).

4.3. Estimation non paramétrique

La covariable AGE ne prend que des valeurs entières allant de 20 à 80 ans. Notons $\{z_t, t = 1, \dots, T\}$ ces T valeurs distinctes. Pour chacun des estimateurs non paramétriques, le quantile conditionnel $q_\alpha(x)$ a été évalué en ces valeurs. On obtient donc l'ensemble des points $\{(z_t, \widehat{q}_\alpha(z_t)), t = 1, \dots, T\}$. Pour la représentation graphique des courbes de référence, une approche basique consiste à réaliser une interpolation linéaire entre ces différents points. Cependant les courbes obtenues avec cette approche présentent un aspect visuel non "lisse". Ainsi, pour pallier ce défaut, nous avons opté pour un lissage par la méthode du noyau (de type Nadaraya-Watson) de ces points, le noyau choisi étant le noyau normal et la fenêtre utilisée étant obtenue par validation croisée.

Les trois estimateurs non paramétriques ont été programmés en langage C pour des raisons de temps d'exécution, les étapes de minimisation (\mathcal{P}) ainsi que d'optimisation des critères de validation croisée étant assez lourdes. Les différents graphiques, quant à eux, sont réalisés avec Splus.

4.4. Résultats

Nous précisons tout d'abord les critères utilisés pour accepter ou rejeter *a posteriori* les courbes de référence obtenues. Puis, nous résumons les résultats obtenus par la méthode paramétrique et les méthodes non paramétriques. Nous décrivons ensuite l'étude de deux variables particulières pour lesquelles la modélisation paramétrique a échoué. Nous remarquons enfin que, même lorsqu'il existe un modèle paramétrique et que les courbes de référence sont acceptées, l'approche non paramétrique fonctionne bien.

• Critère d'acceptabilité des courbes de références

Les courbes de référence (obtenues par les méthodes paramétrique ou non paramétriques) sont considérées comme acceptables si elles satisfont les trois conditions suivantes :

- (a) Elles n'incluent pas de valeurs impossibles pour Y (*i.e.* par exemple, des valeurs nulles ou négatives alors que la variable Y ne peut prendre en réalité que des valeurs strictement positives).
- (b) Elles contiennent le pourcentage désiré d'individus à savoir $100(2\alpha - 1)\%$.
- (c) Les valeurs individuelles qui se trouvent en dehors des limites des courbes de référence sont réparties de façon uniforme en fonction de la covariable AGE et aucun regroupement de valeurs individuelles n'apparaît.

• Synthèse des résultats obtenus par la méthode paramétrique

Pour un certain nombre des variables étudiées, nous avons pu établir des courbes de référence en fonction de l'âge par la méthode paramétrique. Le tableau 1 résume les résultats ainsi obtenus. Remarquons que le modèle paramétrique est acceptable lorsque les résidus sont normalement distribués (après transformation au préalable des données si nécessaire). Si le modèle paramétrique n'est pas accepté, les courbes de référence ne peuvent alors être construites.

TABLEAU 1
Récapitulatif des résultats obtenus avec la méthode paramétrique

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

• Synthèse des résultats obtenus par les méthodes non paramétriques.

Les courbes de référence en fonction de l'âge ont été établies avec succès pour tous les paramètres biophysiques analysés par l'estimateur à noyau (méthode 1). Pour la presque totalité des paramètres, la méthode de la constante locale (méthode 2) donne des courbes de référence acceptables. Par contre, les courbes de référence obtenues par la méthode du noyau produit (méthode 3) sont moins satisfaisantes. Il apparaît clairement que la règle empirique pour le choix de $h_{2,n}$ est mal adaptée et entraîne des courbes de référence sur-lissées. Pour remédier à ce problème déjà constaté dans la partie 3 sur les simulations, d'autres méthodes de choix ne reposant pas sur une hypothèse de normalité doivent être envisagées. Le tableau 2 résume l'ensemble des résultats obtenus avec les trois méthodes non paramétriques.

TABLEAU 2

Récapitulatif des résultats obtenus avec les méthodes non paramétriques
(méthode 1 : méthode d'estimation par noyau, méthode 2 : méthode de la constante locale, méthode 3 : méthode d'estimation par noyau produit)

Zones	Joue	Front	Avant-bras
Nombre de variables	12	12	11
Méthode 1 : nombre de courbes de référence acceptées	12	12	11
Méthode 2 : nombre de courbes de référence acceptées	12	11	11
Méthode 3 : nombre de courbes de référence acceptées	3	3	2

• Etude détaillée de trois variables

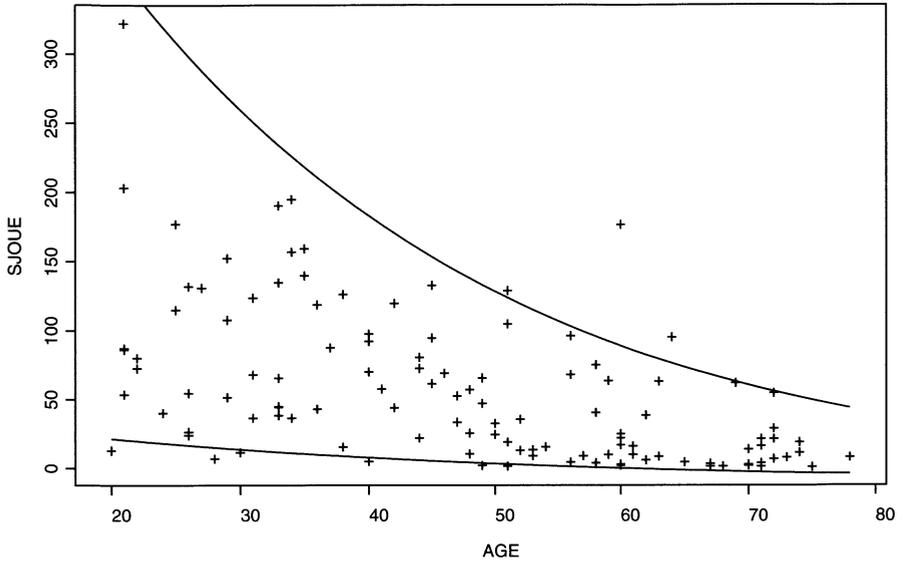
Nous présentons ci-après l'investigation relative aux variables de production de sébum. Elles correspondent aux taux instantanés de sébum mesurés sur le front (variable SFRONT sur les graphiques) et sur la joue (variable SJOUE). Pour ces deux variables, l'approche paramétrique a échoué.

Dans le cas de SJOUE, un modèle linéaire a été estimé après une transformation logarithmique de la forme $\log_e(SJOUE + \lambda)$ avec $\lambda = 10$ (pour accepter la normalité des résidus). Cependant, la courbe de référence estimée de niveau 5% passe sous la valeur zéro, atteignant des valeurs négatives qui ne sont pas physiologiquement acceptables (voir la figure 4(a)).

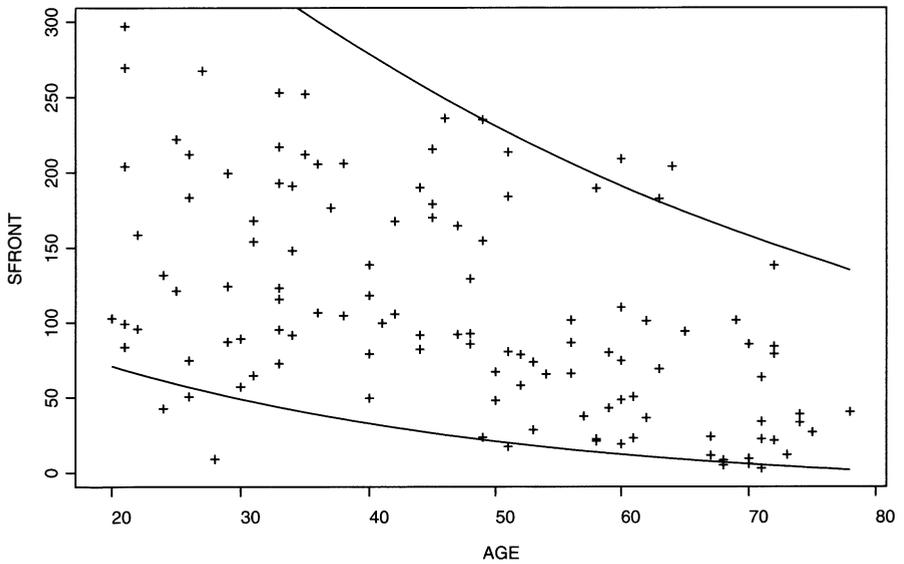
Pour la variable SFRONT, en utilisant un modèle linéaire, les résidus sont apparus normalement distribués après une transformation de la forme $\log_e(SFRONT + \lambda)$ avec $\lambda = 12$. Cependant ces courbes de référence à 90% ne sont pas totalement satisfaisantes, bien que 9% des observations se trouvent à l'extérieur des courbes de référence. En effet, celles-ci ne sont pas réparties de façon homogène le long de l'étendue de l'âge. De plus, la courbe de référence supérieure croît beaucoup trop rapidement pour les faibles valeurs de l'âge (voir la figure 4(b)). La valeur maximale observée à 20 ans sur l'échantillon est de $300\mu g/cm^2$, or la courbe de référence atteint une valeur supérieure à $400\mu g/cm^2$ pour cet âge. Les courbes de référence ne sont donc pas acceptables.

Pour établir des courbes de référence lorsque ces problèmes apparaissent, les méthodes non paramétriques s'avèrent nécessaires.

Les figures 5(a) et 5(b) représentent les critères de validation croisée dont ont été déduites les fenêtres h_n et h_{moy} (utilisées respectivement dans le premier et les deux derniers estimateurs) pour la variable SFRONT. Les valeurs permettant d'atteindre le minimum du critère sont respectivement $h_n = 7,9$ et $h_{moy} = 8,6$ dans ce cas. Des graphiques similaires sont obtenus pour les autres variables.



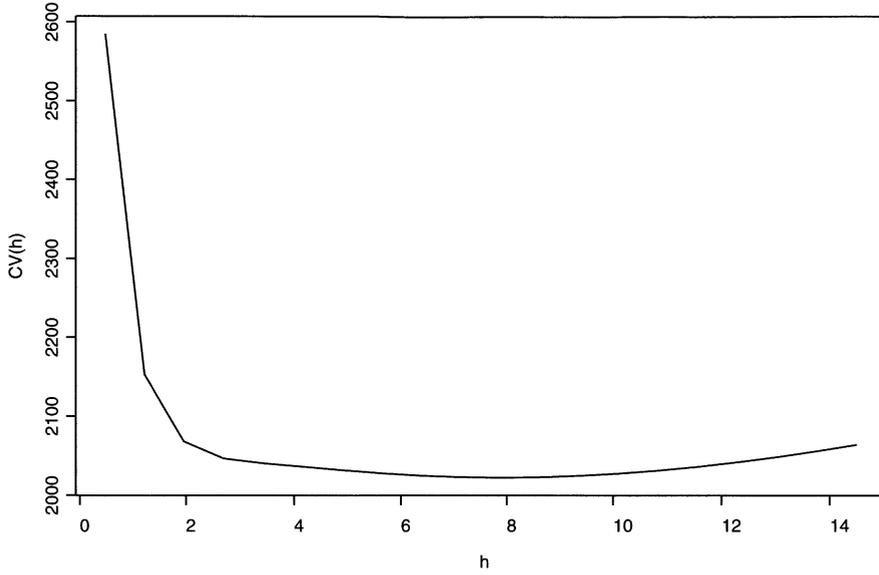
(a) Taux instantanés de sébum mesurés sur la joue (variable SJOUE)



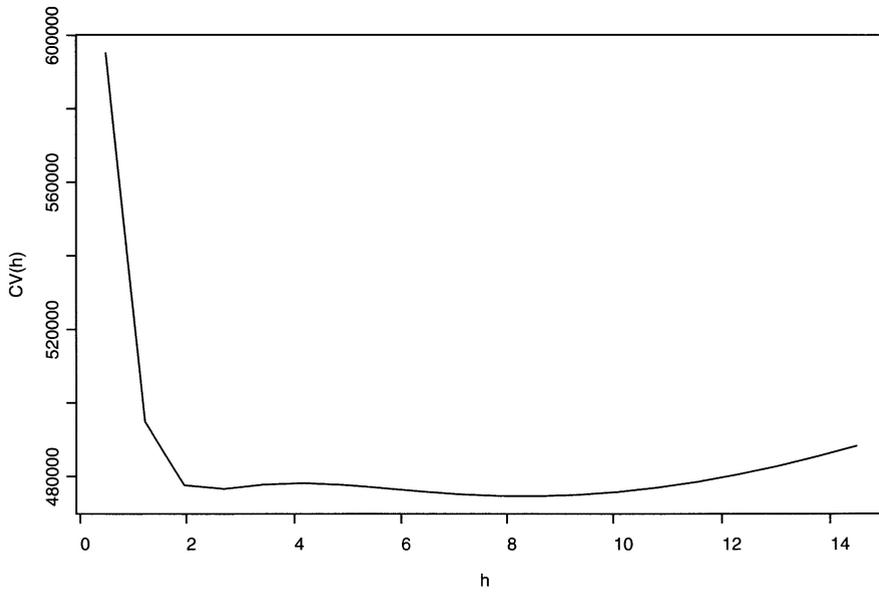
(b) Taux instantanés de sébum mesurés sur le front (variable SFRONT)

FIGURE 4

Courbes de référence à 5% et 95% obtenues avec la méthode paramétrique pour les variables SJOUE et SFRONT



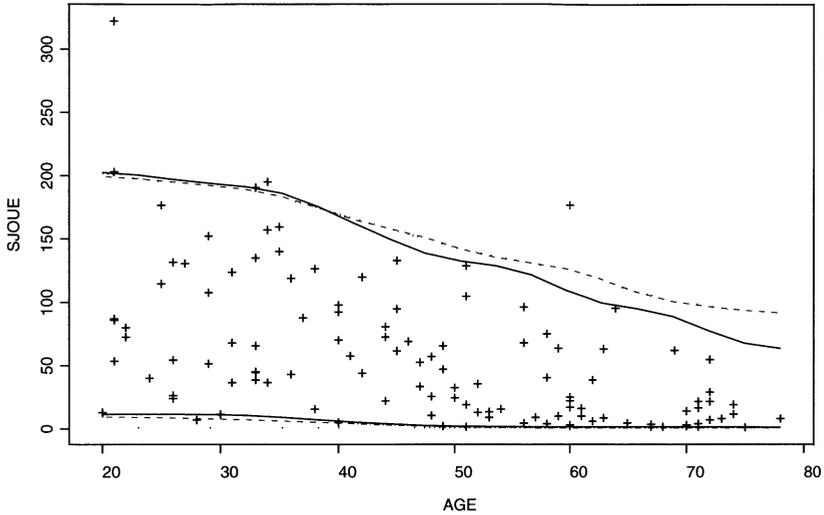
(a) Choix de h_n



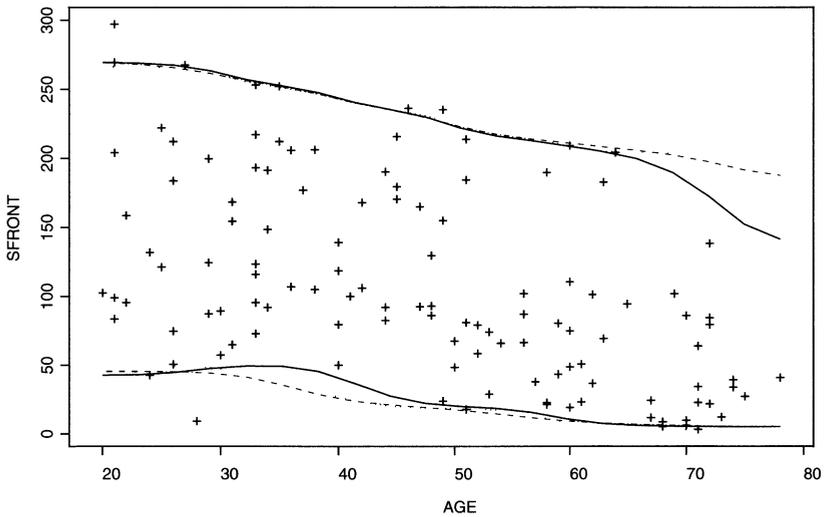
(b) Choix de h_{moy}

FIGURE 5

Critères de validation croisée déterminant le choix des fenêtres des estimateurs non paramétriques pour la variable SFRONT



(a) Taux instantanés de sébum mesurés sur la joue (variable SJOUE)



(b) Taux instantanés de sébum mesurés sur le front (variable SFRONT)

FIGURE 6

Courbes de référence à 5% et 95% obtenues avec les méthodes non paramétriques pour les variables SJOUE et SFRONT (trait continu : méthode d'estimation par noyau, pointillés : méthode de la constante locale, tirets : méthode d'estimation par noyau produit)

Les graphiques 6(a) et 6(b) montrent les courbes de référence à 90% qui ont été construites par les trois méthodes non paramétriques pour les variables SJOUE et SFRONT.

En ce qui concerne la variable SJOUE, les courbes de référence supérieures obtenues avec les trois estimateurs non paramétriques correspondent bien à ce que l'on s'attend à observer d'un point de vue biologique (décroissance du taux instantané de sébum avec l'âge), seul l'aspect légèrement « ondulé » n'est pas totalement conforme. Pour « lisser » un peu plus ces courbes, il serait possible de prendre une fenêtre légèrement plus large que celle sélectionnée par la méthode de validation croisée.

Pour ce qui est de la variable SFRONT, les courbes de référence supérieures se comportent de manière semblable jusqu'à l'âge de 65 ans. Ensuite, le premier estimateur décroît plus rapidement que les deux autres ce qui correspond bien à une diminution attendue du taux instantané de sébum avec l'âge. En ce sens, cette courbe de référence pourrait être préférée aux deux autres.

Notons enfin que même lorsqu'un modèle paramétrique a pu être construit et que les courbes de référence ont été déclarées acceptables, l'approche non paramétrique donne des résultats semblables. Par exemple, sur la figure 7 les courbes de référence obtenues avec la méthode paramétrique peuvent être comparées à celles obtenues avec l'estimateur à noyau (méthode 1) pour l'angle typologique individuel du front (variable ITAFRONT), les deux séries de courbes évoluant de façon comparable sur la fourchette d'âges étudiée.

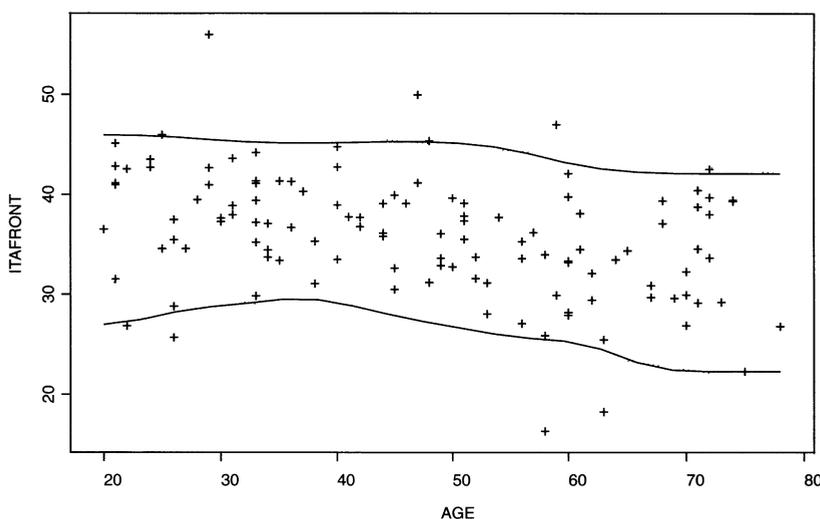


FIGURE 7

Courbes de référence à 5% et 95% de l'angle typologique individuel sur le front (variable ITAFRONT) obtenues avec la méthode paramétrique (pointillés) et le premier estimateur non paramétrique (trait continu)

5. Conclusion et extensions

Cet article présente une alternative à l'approche paramétrique de Royston (1991) pour l'estimation des courbes de référence. L'approche non paramétrique proposée est robuste, et les courbes de référence sont ainsi déterminées sans détection préalable des points aberrants. Dans les études empiriques que nous avons réalisées, il apparaît que deux des estimateurs non paramétriques fournissent des courbes de référence acceptables pour la presque totalité des variables étudiées. Cette analyse statistique peut ainsi être très utile pour la détermination de courbes de référence à partir de données d'une fiabilité médiocre. Le troisième estimateur (méthode d'estimation par noyau produit) pose encore quelques problèmes pratiques provenant du choix des fenêtres, d'autres règles doivent être étudiées. Nous envisageons deux possibilités : une méthode de type validation croisée, ne nécessitant pas d'hypothèse sous-jacente mais numériquement coûteuse, ou le développement d'une règle empirique plus adaptée à la nature du bruit observé.

Cependant, bien qu'extrêmement utiles, ces estimateurs ne portent que sur une seule variable d'intérêt et ne prennent en compte qu'une seule covariable (l'âge dans notre application). Il serait intéressant de les adapter à des situations plus générales.

- D'une part, il est parfois nécessaire d'utiliser plusieurs covariables quantitatives (le poids et l'âge par exemple) pour augmenter la précision des courbes de référence. Un indice $X'\beta$ peut alors être intégré à cette fin pour résumer la partie explicative d'une covariable vectorielle X . On peut aussi envisager la présence de plus d'un indice.
- D'autre part, dans certains domaines, en particulier sur la peau, la construction de « régions » de référence multivariées est utile. En effet, bien que la valeur des paramètres de certains individus soient dans les limites de référence des paramètres pris isolément, lorsque les valeurs de ces mêmes individus sont examinées grâce à des méthodes multivariées, ils peuvent apparaître en dehors de la « région » de référence. Cette apparente contradiction est généralement due aux corrélations entre les différents paramètres biophysiques qui ne sont pas prises en compte avec les courbes de référence (univariées). Ces corrélations pourraient être prises en compte dans la construction de « régions » de référence multivariées, ce qui en fait leur intérêt majeur. Par exemple, lorsque pour un sujet la corrélation entre les paramètres n'est pas conforme à celle attendue, compte tenu des informations apportées par la totalité de l'échantillon, une anomalie multivariée apparaît, anomalie qui ne pourrait pas être détectée par l'examen isolé de chacun des paramètres. La révélation de cette anomalie permet l'identification de sujets particuliers, et de procéder à un examen minutieux des données de ces individus.

L'estimation des quantiles conditionnels multivariés peut permettre d'élargir notre approche non paramétrique à ce contexte.

Plus généralement, lorsque la variable d'intérêt et la covariable sont toutes les deux multidimensionnelles, une extension naturelle est la combinaison de ces deux directions. Ces thèmes sont actuellement en cours d'étude.

Remerciements

Les auteurs remercient le Pr. E. Tschachler pour ses encouragements, le Pr. H. Tagami, le Dr K. Numagami et toute l'équipe du CE.R.I.E.S. pour leur contribution aux données, en particulier J. Latreille, I. Le Fur et le Dr F. Morizot.

Références bibliographiques

- BASSET, G. & KOENKER, R. (1982). An empirical quantile function for linear models with iid errors. *Journal of the American Statistical Association*, **77**, 407-415.
- BERLINET, A., GANNOUN, A. & MATZNER-LØBER, E. (2001). Asymptotic normality of convergent estimates of conditional quantiles. *Statistics*, **35**, 139-169.
- BHATTACHARYA, P.K. & GANGOPADHYAY, A.K. (1990). Kernel and nearest-neighbor estimation of a conditional quantile. *The Annals of Statistics*, **18**, 1400-1415.
- CHARDON, A., CRETOIS, I. & HOURSEAU, C. (1991). Skin colour typology and suntanning pathways. *International Journal of Cosmetic Science*, **13**, 191-208.
- CHAUDHURI, P. (1991). Nonparametric estimates of regression quantiles and their local bahadur representation. *The Annals of Statistics*, **19**, 760-777.
- COLE, T.J. (1988). Fitting smoothed centile curves to reference data. *Journal of Royal Statistical Society, Series A*, **151**, 385-418.
- DEHEUVELS, P. (1977). Estimation non paramétrique de la densité par histogramme généralisé. *La Revue de Statistique Appliquée*, **35**, 5-42.
- DUCHARME, G.R., GANNOUN, A., GUERTIN, M.C. & JÉQUIER, J.C. (1995). Reference values obtained by kernel-based estimation of quantile regression. *Biometrics*, **51**, 1105-1116.
- FAN, J. & GJBELS, I. (1992). Variable bandwidth and local linear regression smoothers. *The Annals of Statistics*, **20**, 2008-2036.
- FAN, J., HU, T.C., & TRUONG, Y.K. (1994). Robust nonparametric function estimation. *Scandinavian Journal of Statistics*, **21**, 433-446.
- GANNOUN, A. (1990). Estimation non paramétrique de la médiane conditionnelle : médianogramme et méthode du noyau. *Annales de l'I.S.U.P.*, **XXXXV**, 11-22.
- GOLDSTEIN, H. & PAN, H. (1992). Percentile smoothing using piecewise polynomials, with covariates. *Biometrics*, **48**, 1057-1068.
- HÄRDLE, W. (1990). *Applied nonparametric regression*, Cambridge University Press, Cambridge.
- HEALY, M.J.R., RASBASH, J., & YANG, M. (1988). Distribution-free estimation of age-related Centiles. *Annals of Human Biology*, **15**, 17-22.

- HENDRICKS, W. & KOENKER, R. (1992). Hierarchical spline models for conditional quantiles and the demand for electricity. *Journal of the American Statistical Association*, **99**, 58-68.
- HORN, P.S., PESCE, A.J., & COPELAND, B.E. (1998). A robust approach to reference interval estimation and evaluation. *Clinical Chemistry*, **44**, 622-631.
- JONES, M.C. & HALL, P. (1990). Mean squared error properties of kernel estimates of regression quantiles. *Statistic and Probability Letters*, **10**, 283-289.
- KOENKER, R., PORTNOY, S., & NG, P. (1992). Nonparametric estimation of conditional quantile functions. *L₁- statistical analysis and related methods*, ed Y. Dodge, Elsevier : Amsterdam, 217-229.
- MAGEE, L., BURBIDGE, J.B., & ROBB, A.L. (1991). Computing kernel-smoothed Conditional Quantiles from Many Observations. *Journal of the American Statistical Association*, **86**, 673-677.
- MINT EL MOUVIT, L. (2000). Sur l'estimateur linéaire local de la fonction de répartition conditionnelle. *Thèse de doctorat, Université Montpellier 2*.
- POIRAUD-CASANOVA, S. & THOMAS-AGNAN, C. (1998). Quantiles conditionnels. *Journal de la Société Française de Statistique*, **139**(4), 31-44.
- POIRAUD-CASANOVA, S. (2000). Estimation non paramétrique des quantiles conditionnels. *Thèse de doctorat, Université Toulouse 1*.
- ROUSSAS, G.R. (1991). Estimation of transition distribution function and its quantiles in Markov processes : strong consistency and asymptotic normality. *Non-parametric Functional Estimation and Related Topics*, Kluwer Academic Publishers : Netherlands, 443-462.
- ROYSTON, P. (1991). Constructing time-specific reference ranges. *Statistics in Medicine*, **10**, 675-690.
- ROYSTON, P. & ALTMAN, D.G. (1992). Regression using fractional polynomials of continuous covariates : parsimonious parametric modelling (with discussion). *Applied Statistics*, **43**, 429-467.
- ROYSTON, P. & WRIGHT, E.M. (1998). How to construct normal ranges for fetal variables. *Ultrasound Obstet Gynecol*, **11**, 30-38.
- SAMANTA, T. (1989). Non-parametric estimation of conditional quantiles. *Statistic and Probability Letters*, **7**, 407-412.
- STONE, C.J. (1977). Consistent nonparametric regression (with discussion). *The Annals of Statistics*, **5**, 595-645.
- STUTE, W. (1986). Conditional empirical processes. *The Annals of Statistics*, **14**, 638-647.
- TANGO, T. (1998). Estimation of age-specific reference ranges via smoother AVAS. *Statistics in Medicine*, **17**, 1231-1243.
- WRIGHT, E.M. & ROYSTON, P. (1997). A Comparaison of Statistical Methods for Age-Related Reference Intervals. *J. R. Statist. Soc. Series A*, **160**, 47-69.
- YAO, Q. (1999). Conditional predictive regions for stochastic processes. *Technical report*, University of Kent at Canterbury, UK.

YU, K. (1997). *Smooth regression quantile estimation*. Ph.D. thesis, The open University, UK.

YU, K. & JONES, M.C. (1998). Local linear quantile regression. *Journal of the American Statistical Association*, **93**, 228-237.