Estimation de courbes de référence pour l'analyse de propriétés biophysiques

JÉRÔME SARACCO[1], ALI GANNOUN[1,2] et CHRISTIANE GUINOT[3]

[1] Laboratoire de Probabilités et Statistique, Université Montpellier II.

[2] Statistical Genetics and Bioinformatics Unit, National Genome Center, Howard University, Washington D. C., U.S.A.

[3] CEntre de Recherches et d'Investigations Epidermiques et Sensoriels (CE.R.I.E.S), Neuilly-sur-Seine.

1 Problématique

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 dans une population cible. Cette variable sera notée Y par la suite. 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, ces derniers sont 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.

D'autre part, 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, nous sommes amenés à travailler avec les quantiles conditionnels de Y sachant X. Pour les sujets de référence, le tracé de courbes de référence sur le nuage des valeurs prises par le couple (X, Y) 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.

Plus précisément, pour une valeur x donnée et pour $\alpha \in]0.5, 1[$, l'intervalle de référence contenant $100(2\alpha - 1)\%$ des sujets de référence est défini par

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

où $q_\alpha(x)$ est le quantile conditionnel d'ordre α de la variable Y sachant que $X = x$. Il est défini de la manière suivante :

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

$F(.|x)$ désignant la fonction de répartition conditionnelle de Y sachant que $X = x$. Les courbes de référence inférieure et supérieure sont alors les ensembles de points

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

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

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

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

Au moins deux types d'approches, l'une paramétrique et l'autre non paramétrique, ont été développés pour l'estimation des quantiles conditionnels et par voie de conséquence des courbes de référence.

L'approche paramétrique repose sur le choix d'une classe paramétrée de distributions. L'estimation des paramètres permet alors de retenir l'une des distributions de cette classe. On a donc forcé la solution à appartenir à une classe de distribution donnée. Ainsi, cette approche paramétrique nécessitant donc des hypothèses restrictives peut être mal adaptée à la réalité des données en particulier biologiques.

L'approche non paramétrique a alors été développée afin de pallier ces problèmes d'hypothèses et de modélisation paramétriques. Les méthodes non paramétriques ne nécessitent en effet pas d'hypothèse sur la nature de la distribution. Elles sont de plus robustes car elles sont déterminées sans détection préalable de points aberrants. En conséquence, une analyse statistique utilisant l'estimation non paramétrique des courbes de référence peut être faite à partir de données d'une fiabilité médiocre.

Dans la suite, nous précisons tout d'abord le cadre et les données de l'étude. Nous détaillons ensuite la méthode non paramétrique d'estimation par noyau des quantiles conditionnels et nous appliquons cette méthode pour construire les courbes de référence à 90% d'un paramètre biophysique de la peau. Enfin, nous terminons en mentionnant quelques extensions de ce travail, en particulier, au cadre multidimensionnel et à d'autres estimateurs non paramétriques des quantiles conditionnels.

2 Les données

En vue de cibler des produits cosmétiques sur le marché japonais, Chanel a demandé au CE.R.I.E.S (centre de recherche sur la peau humaine financé par Chanel et situé à Neuilly-sur-Seine) de faire une étude sur les propriétés biophysiques de la peau de femmes japonaises. La Statistique, au moyen de l'estimation de courbes de référence en fonction de l'âge pour ces propriétés biophysiques, va ainsi servir d'aide à la décision pour adapter au mieux les produits à ce nouveau marché asiatique.

L'objectif de l'étude (partiellement présentée ici) réalisée par le CE.R.I.E.S était d'établir donc ces courbes de référence à 90% pour les propriétés biophysiques (mesurées sur deux zones du visage et une zone de l'avant-bras).

Les données utilisées ont été recueillis par le CE.R.I.E.S entre le 15 décembre 1998 et le 15 avril 1999 à Sendai (Japon) sur $n = 120$ femmes japonaises 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 $23 \pm 1°C$ 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. 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. Les paramètres biophysiques (variables d'intérêt) mesurés ou calculés sont les suivants : le taux de sécrétion de sébum (mesuré uniquement sur le front et la joue), 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 (exprimée à l'aide de trois grandeurs), l'angle typologique individuel, la saturation et l'angle de teinte. La covariable est l'âge des volontaires.

Dans la suite de ce document, nous nous limiterons uniquement à l'étude de la variable mesurant le sébum instantané de la joue (en g/cm^2), notée SJOUE, en fonction de l'âge.

3 Un peu de modélisation mathématique

Rappelons tout d'abord qu'un estimateur de la fonction de répartition conditionnelle de Y sachant $X = x$, est défini,pour $y \in \mathbb{R}$, par :

$$F_n(y|x) = \begin{cases} \displaystyle\sum_{i|X_i=x} \left(\frac{\mathbf{1}}{\mathbf{n_x}}\right) \mathbb{I}_{\{Y_i \leq y\}} & \text{si } \; n_x > 0, \\ 0 & \text{sinon,} \end{cases}$$

où n_x désigne le nombre d'individus de l'échantillon tels que $X_i = x$. Il apparaît clairement que, pour faire une estimation correcte de cette fonction de répartition, il faut disposer d'un certain nombre d'observations Y_i telles que $X_i = x$, ce qui est rarement le cas en pratique. De plus, cet estimateur n'est pas "lisse" (continue) en x. Il est donc préférable d'introduire un estimateur qui permet de contourner ces problèmes.

Définissons alors maintenant un estimateur non paramétrique ("lisse" en x) de la fonction de répartition conditionnelle de Y sachant $X = x$, pour $y \in \mathbb{R}$:

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

La fonction K, appelée noyau[1], permet de faire intervenir dans le calcul de l'estimateur tous les points de l'échantillon affectés de poids d'autant plus grand que X_i est voisin

1. fonction que l'on suppose généralement positive, symétrique et maximale en zéro (très souvent, il s'agit d'une densité de probabilité)

de x. On prend généralement pour K la densité de la loi normale centrée réduite. Le paramètre h_n, appelée fenêtre, permet de contrôler le lissage appliqué aux données : plus h_n est grand, plus l'estimateur va prendre en compte un nombre important d'observations et donc plus le lissage sera important ; à l'inverse plus h_n est petit, et moins l'estimateur sera "lisse". Son choix est crucial en pratique car il faut éviter de faire du sur-lissage ou du sous-lissage. Un choix "optimal" pour h_n peut être obtenu de manière automatique à partir des données au moyen d'une méthode de validation croisée [2].

De l'estimateur de la fonction de répartition conditionnelle $\tilde{F}_n(y|x)$, il est alors naturel d'estimer le quantile conditionnel $q_\alpha(x)$ par $\tilde{q}_{\alpha,n}(x)$ défini de la manière suivante :

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

l'inversion étant faite de manière numérique.

4 Application à la variable SJOUE

Avant de fournir les résultats obtenus pour cette variable d'intérêt, on précise que les courbes de référence obtenues 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 ici 90%.

(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.

Trois méthodes non paramétriques d'estimation des quantiles conditionnels et donc des courbes de référence ont été mise en oeuvre dans cette étude : la méthode d'estimation par noyau décrite précédemment, ainsi qu'une méthode d'estimation par noyau dite de la constante locale et une méthode d'estimation par noyau produit[2]). La Figure 1 nous donne le nuage des points croisant les variables AGE et SJOUE sur lequel les courbes de référence à 90% obtenues avec les différentes méthodes non paramétriques ont été superposées.

On peut constater que ces courbes de référence sont physiologiquement acceptables. En particulier 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. Il serait cependant "techniquement" possible de "lisser" un peu plus ces courbes en prenant des fenêtres légèrement plus larges que celles sélectionnées automatiquement par la méthode de validation croisée.

2. Pour plus de détails et des références sur ces différentes méthodes, nous renvoyons le lecteur aux deux articles (1) et (2) mentionnés en bibliographie.

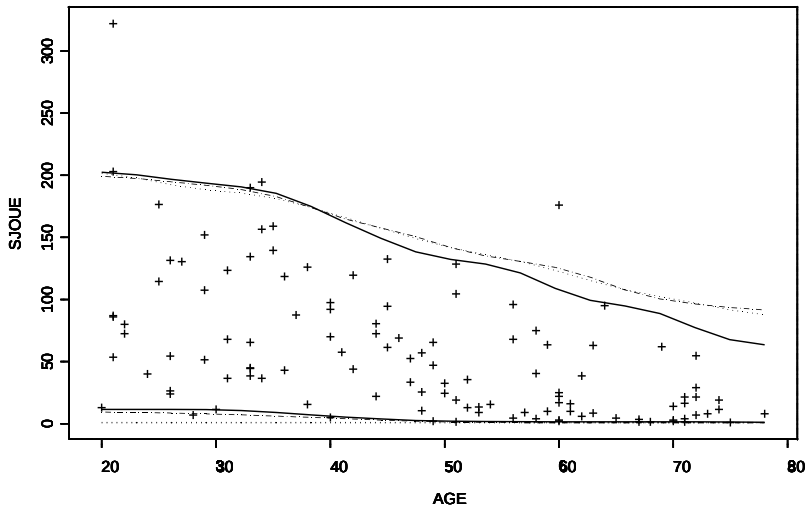

FIG. 1 – *Courbes de référence à 90% obtenues avec des méthodes non paramétriques pour la variable SJOUE (trait continu : méthode d'estimation par noyau, pointillés : méthode de la constante locale, tirets : méthode d'estimation par noyau produit).*

5 Compléments et extensions

Bien qu'extrêmement utiles, les estimateurs non paramétriques mentionnés ici 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 (l'âge, le poids et les conditions expérimentales par exemple) pour augmenter la précision des courbes de référence. Le fait que la covariable devienne multidimensionnelle, n'est pas un obstacle au développement de la théorie. Cependant les méthodes non paramétriques souffrent d'un point de vue pratique du "fléau de la dimension" [3]. Se pose aussi le problème de la représention graphique des données et des estimateurs. Malgré l'évolution des logiciels, il est impossible de faire des représentations graphiques sérieuses quand la dimension de l'espace qui contient les variables est supérieure à 3. Ceci prive l'analyste de faire des constatations vraisemblables et d'avancer des conclusions plausibles à la seule lecture des graphiques.

  Nous avons alors proposé une nouvelle méthodologie fondée sur une étape de réduction de la dimension de la covariable, suivie d'une étape d'estimation non paramétrique de quantiles conditionnels. Cette approche semiparamétrique combine la méthode SIR (Sliced Inverse Regression) et l'estimation à noyau de quantiles conditionnels. L'étape de réduction de la dimension permet d'obtenir un ou plusieurs

3. Cela décrit le fait qu'il devient de plus en plus difficile de faire une estimation de bonne qualité vu la dispersion de plus en plus grande des données dans un espace de grande dimension.

indices $X'\beta$ résumant la partie explicative de la covariable vectorielle X. Notons que cette technique de réduction de la dimension au moyen d'indices entraîne non seulement l'amélioration de l'estimation du quantile conditionnel (disparition ou forte diminution du fléau de la dimension), mais donne aussi aux analystes la possibilité d'interpréter et de quantifier le rôle joué par chaque covariable dans l'étude. Nous avons illustré cette méthode en établissant des courbes de référence pour des propriétés biophysiques de la peau de femmes françaises "saines" en fonction de l'âge, des conditions expérimentales et d'autres propriétés biophysiques.

- D'autre part, le cadre où la variable d'intérêt est multidimensionnelle intéresse aussi les praticiens, en particulier ceux travaillant sur la peau. On parle alors non plus de courbes de référence mais de "régions" de référence multivariées. Par exemple, la variable d'intérêt multidimensionnelle aurait pour composantes toutes les paramètres biophysiques évalués sur une zone particulière (la joue, le front ou l'avant-bras).

  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 ainsi être introduites 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 permet d'élargir l'approche non paramétrique à ce contexte multidimensionnel.

- 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. Ce thème est actuellement en cours d'étude.

Références bibliographiques

Pour plus de détails théoriques et des résultats plus détaillés concernant cette étude et une étude similaire faite sur un échantillon de femmes françaises pour lesquelles les courbes de référence à 90% ne sont établies qu'en fonction de l'âge, le lecteur pourra se référer aux deux premiers articles cités ci-après. En ce qui concerne la construction de courbes de référence utilisant une covariable multidimensionnelle, nous renvoyons le lecteur au troisème article.

(1) Gannoun, A., Girard, S., Guinot, C. & Saracco, J. (2002). "Trois méthodes non paramétriques pour l'estimation de courbes de référence - Application à l'analyse

de propriétés biophysiques de la peau". A paraître dans la *Revue de Statistique Appliquée*.

(2) Gannoun, A., Girard, S., Guinot, C. & Saracco, J. (2002). "Reference curves based on nonparametric quantile regression". A paraître dans *Statistics in Medicine*.

(3) Gannoun, A., Girard, S., Guinot, C. & Saracco, J. (2002). "Dimension reduction in reference curves estimation". Soumis pour publication.