Découpage de courbes de densité : Application au dépistage du cancer

Fabrice MORLAIS[a,b] & Frédéric FERRATY[b] & Philippe VIEU[b]

[a] ERI3 INSERM 'Cancers & Populations', EA 3936 Université Caen, Faculté de médecine - avenue côte de nacre - 14032 Caen cedex
[b] Université Paul Sabatier, IMT UMR CNRS 5583, 118, Route de Narbonne, 31062 Toulouse Cedex, France

Résumé

Le dépistage actuel du cancer broncho-pulmonaire est effectué à l'aide d'une radiographie pulmonaire, d'un scanner thoracique et d'un examen cytologique des expectorations. La cytologie automatisée des expectorations est une méthode permettant l'analyse informatique des cellules d'un crachat sur la lame d'un microscope. Comme une personne est représentée par l'ensemble des cellules de sa lame, il nous a paru intéressant d'utiliser la densité de probabilité comme unité statistique. La modélisation fonctionnelle des données, méthode pour laquelle l'unité statistique est à valeurs dans un espace infini, répond bien à cette problématique statistique puisque, par définition, une densité de probabilité est une fonction. Lors de cet exposé nous présenterons la méthode de classification supervisée de courbes de densité que nous avons développée, pour discriminer des personnes ayant un cancer et des personnes saines, et nous vous donnerons quelques résultats issus de données réelles.

Abstract

Screening of bronchopulmonary cancer is currently performed using chest-X ray, chest CT scan and cytological examination of expectorated sputum. Automated cytology of expectorated sputum is a method which enables sputum cells to be analyzed on a microscope slide. It seems interesting to use density probability as statistical unit because a person is represented by all the cells of her slide. As in functional data analysis statistical unit takes values in an infinite dimensional space, we can use functional data analysis because density probability is a function. In this talk we will give the supervised classification of density curves that we have developed for discriminating persons having a cancer and control persons, and we will give some results that come from real data.

Introduction

En modélisation fonctionnelle (voir par exemple : Ferraty et Romain (2010), Ferraty et Vieu (2006) et Ramsay et Silverman (2005)) l'unité statistique n'est plus seulement représentée par un ensemble de n variables, à valeurs dans un espace $\mathbb{R}^n$, mais par une fonction, à valeurs dans un espace de dimensions infini. Une variable aléatoire χ est considérée comme fonctionnelle si elle prend des valeurs dans un espace infini, par exemple la variable fonctionnelle $\chi = \{X(t); t \in T\}$ avec $T \subset \mathbb{R}$ représente une courbe observée sur l'intervalle T de $\mathbb{R}$. La première étape à considérer dans une modélisation fonctionnelle est la transformation des données initialement discrétisées en une fonction continue. De nombreuses techniques statistiques existent pour réaliser cette transformation lorsque la fonction à estimer est une densité de probabilité (voir Wasserman (2007) pour la monographie la plus récente sur le sujet).
Bien que les méthodes statistiques classiques soient à peu près toutes développées dans le cadre fonctionnel, quelques difficultés surviennent pour les adapter lorsque les données fonctionnelles considérées sont des densités. Les méthodes statistiques listées ci-dessous sont des exemples permettant d'utiliser la densité de probabilité comme unité statistique. En statistique exploratoire, Kneip et Utikal (2001) ont développé une ACP fonctionnelle adaptée aux densités. En statistique supervisée, la méthode non paramétrique développée par Ferraty et Vieu (2003), dans le cadre de variables aléatoires fonctionnelles classiques, s'adapte aux densités puisqu'elle repose essentiellement sur la notion de distances entre courbes. Pour comparer globalement des courbes de densité, Delicado (2007) a développé une méthode modifiant l'ANOVA fonctionnelle de Cuevas *et al.* (2004) pour la rendre adaptable aux fonctions de densité de probabilité.
Nous avons privilégié l'approche de Ferraty et Vieu (2003) car elle correspondait parfaitement à notre problématique initiale qu'est la prévision. On trouvera dans le Chapitre 8 de Ferraty et Vieu (2006) et dans le Chapitre 10 de Ferraty et Romain (2010) des discussions bibliographiques plus complètes sur la classification de courbes. L'utilisation pratique de cette méthode nous a montré que, même dans le cadre de différences manifestes et visuelles entre groupes de densités, la discrimination au sens mathématique du terme n'était pas toujours retrouvée. Cela étant très probablement dû à certaines parties de la distribution perturbant l'analyse. Pour contourner ce problème nous avons développé une méthode statistique recherchant les morceaux de densité optimaux pour la discrimination. Cette méthode que nous nommerons par la suite 'Optimal cutting' sera présentée dans la partie 1. La partie 2 s'intéressera à la transformation de nos données initiales en fonctions, en utilisant un estimateur à noyau standard. La partie 3 décrira les données réelles sur lesquelles ont été utilisées nos méthodes.

1 Optimal cutting

Soit $(\mathbf{X}_i, \mathbf{Y}_i)_{i=1,\dots,n}$ un échantillon de paires indépendantes et identiquement distribuées de même loi $(\mathbf{X}, \mathbf{Y})$ à valeur dans $E \times \bar{G}$, avec $\mathbf{X}$ une variable aléatoire fonctionnelle définie sur un intervalle $[t_{min}, t_{max}] \subset \mathbb{R}$, (E, d) un espace vectoriel semi-métrique et $\mathbf{Y}$ une variable aléatoire catégorielle définie sur $\bar{G} = \{1, \dots, G\}$. Soit t_0 un point de l'intervalle $[t_{min}, t_{max}]$. A partir de notre variable aléatoire fonctionnelle initiale $\mathbf{X}_i$, on construit deux nouvelles variables aléatoires fonctionnelles en découpant $\mathbf{X}_i$ de la façon suivante :

$$\mathbf{X}_i^{1,t_0} = \{\mathbf{X}_i(t), t \in [t_{min}, t_0[\}$$
$$\mathbf{X}_i^{2,t_0} = \{\mathbf{X}_i(t), t \in [t_0, t_{max}]\}$$

Nous disposons donc d'un n-échantillon de triplets indépendants $(\mathbf{X}_i^{1,t_0}, \mathbf{X}_i^{2,t_0}, \mathbf{Y}_i)$ avec $\mathbf{X}_i^{1,t_0}$ et $\mathbf{X}_i^{2,t_0}$ des variables aléatoires fonctionnelles prenant des valeurs dans un espace infini borné de $\mathbb{R}$ et $\mathbf{Y}$ une variable aléatoire catégorielle à valeurs dans $\bar{G} = \{1, \dots, G\}$. A partir de ces morceaux de courbes indépendants, nous allons estimer pour chacun d'eux la probabilité a posteriori d'appartenance à G en fonction de t_0. Pour ce faire nous allons utiliser la classification supervisée non paramétrique de courbes fonctionnelles développée par Ferraty et Vieu (2003). Nous avons découpé notre échantillon initial en deux échantillons : un échantillon d'apprentissage (L) et un échantillon test (T). L'estimation de ces probabilités a posteriori se fera, pour chaque morceau de courbe X, de la façon suivante :

$$p_g(X) = P(\mathbf{Y} = g | \mathbf{X} = X), g \in \bar{G}$$

où $\mathbf{X}$ est une variable aléatoire fonctionnelle et X est une réalisation de cette variable aléatoire fonctionnelle. Une fois les G probabilités a posteriori estimées, nous affecterons à $\hat{Y}(x)$ le numéro de groupe de plus forte probabilité (classifieur bayésien):

$$\hat{Y}(X) = \arg\max_{g \in \bar{G}} \hat{p}_g(X)$$

Avant de définir notre estimateur à noyau de la probabilité a posteriori, remarquons que :

$$p_g(X) = P(\mathbf{Y} = g | \mathbf{X} = X) = E(\mathbb{1}_{[\mathbf{Y}=g]} | \mathbf{X} = X)$$

Cette probabilité peut donc être estimée en terme d'espérance conditionnelle et nous pouvons donc utiliser un estimateur de type noyau pour prédire cette espérance conditionnelle:

$$\hat{p}_g(X) = \hat{p}_{g,h}(X) = \frac{\sum_{i \in L} \mathbb{1}_{[\mathbf{Y}_i = g]} K\left(\frac{d(\mathbf{X}_i, X)}{h}\right)}{\sum_{i \in L} K\left(\frac{d(\mathbf{X}_i, X)}{h}\right)}$$

où K est un noyau asymétrique, h est la taille de fenêtre du noyau et d est une semi-métrique. L'expression ci-dessus correspond à la régression d'une variable dichotomique

sur une variable fonctionnelle. Le choix de la taille de fenêtre h optimale se fera en minimisant une fonction de coût du type :

$$h_L^{opt} = \arg\inf_h Loss_L(h)$$
$$Loss_L(h) = \sum_{j\in L}\sum_{g\in G}\left(\hat{p}_{g,h} - \mathbb{1}_{[Y_j=g]}\right)^2$$

Pour chaque valeur $t_0 = \{t_{min}, t_{min+1}, \ldots, t_{max},\}$ et pour chaque morceau de courbe X^1 et X^2, nous obtenons une taille de fenêtre minimisant la fonction $Loss$. Nous identifions ensuite la valeur t^{opt} et le morceau de courbe optimal pour lesquels la fonction $Loss$ atteint son minimum.
Ayant un effectif faible nous avons évalué la qualité de prédiction de la modélisation par validation croisée (Hatsie et al. (2009)) :

$$Misclass = \frac{1}{n}\sum_{i=1}^{n}\mathbb{1}_{[Y_i\neq\hat{Y}_i]}$$

où $\hat{Y}_i$ est la prédiction obtenue sur l'échantillon $L_i = L_{\{j,j\neq i\}}$ à partir du morceau de courbe optimal précédemment défini et Y_i est une réalisation de la variable aléatoire $\mathbf{Y}_i$.

2 Estimation de densité

Les variables fonctionnelles $\mathbf{X}_i$ sont des densités de probabilité provenant d'un échantillon de variables aléatoires indépendantes et identiquement distribuées $\mathbf{Z}_{i,1}, \ldots, \mathbf{Z}_{i,n_i}$. Pour estimer les densités de probabilité, nous avons utilisé un estimateur non paramétrique à noyau.

$$\hat{X}_i(t) = \frac{1}{n_i h_{d,i}}\sum_{j=1}^{n_i} K_d\left(\frac{t-\mathbf{Z}_{i,j}}{h_{d,i}}\right)$$

où K_d est un noyau symétrique tel que :

$$\int_{-\infty}^{+\infty} K_d(t)dt = 1$$

et $h_{d,i}$ la taille de la fenêtre. Deux procédures permettant la sélection automatique des fenêtres ont été implémentées. La première est standard et utilise la méthode 'Plug-in' de Sheather et Jones (1991). La seconde opère un choix adaptatif de la fenêtre lié au problème de discrimination. Du point de vue de la méthodologie, il suffit d'appliquer le découpage optimal aux densités obtenues avec le choix de fenêtre Plug-in. Pour le second choix de fenêtre, nous avons modifié l'optimal cutting. D'une part on simplifie le problème en prenant des fenêtres telles que :

$$h_{d,i} = h_{0,i} \times \alpha$$

où $h_{0,i}$ est la fenêtre obtenue par la la méthode Plug-in de Sheather et Jones (1991) et α est une constante. D'autre part, on sélectionne la taille de fenêtre $h_{d,i}$ ayant la fonction de coût *Loss* minimale. De plus, nous avons raffiné notre méthode en ajoutant un prétraitement à nos données. L'enregistrement de courbes (Kneip et Engel (1995), Kneip et Gasser (1992), Ramsay et Silverman (2005)) est utile lorsqu'une forme commune semble apparaître, et lorsque les réalisations individuelles de cette forme diffèrent en phase (variation horizontale) ou en amplitude (variation verticale). La non prise en compte de ces deux phénomènes lors de la modélisation peut amener à prendre en compte des déformations pouvant dégrader la qualité de prédiction. Dans notre étude nous ne nous sommes pas intéressés aux variations d'intensités (ou variation en amplitude, ou variation verticale) car nous souhaitions garder la caractéristique de densité de nos fonctions $\int_{t_{min}}^{t_{max}} X(t)dx = t$. Nous avons simplement aligné nos densités par rapport au mode principal.

3 Présentation des données réelles

Le dépistage actuel du cancer broncho-pulmonaire peut être effectué à l'aide d'une radiographie pulmonaire, d'un scanner thoracique et d'un examen cytologique des expectorations. En radiographie pulmonaire et en scanner thoracique, le praticien s'intéresse à la présence et à l'évolution de la taille des nodules (regroupement de cellules) suspects de cancer. En cytologie 'conventionnelle' des expectorations le pathologiste s'intéresse à la présence de cellules cancéreuses dans un crachat à l'aide d'un microscope optique. Quelques récents travaux (Belien *et al.* (1997), Doudkine *et al.* (1995), Palcic *et al.* (2002) et Payne *et al.* (1997)) ont montré l'intérêt d'une nouvelle technique cytologique des expectorations dans le dépistage précoce de cancers : la cytologie automatisée. La cytologie automatisée des expectorations est une méthode permettant l'analyse informatique des cellules d'un crachat sur la lame d'un microscope. Une caméra numérique reliée à un ordinateur découpe l'image de cette lame en petites images qui sont alors stockées dans l'ordinateur. Ces images sont ensuite traitées par un logiciel d'imagerie qui détecte les cellules du prélèvement, par une méthode de détection de contours, et qui les analyse. Ainsi pour chaque cellule, un certain nombre de paramètres de forme, de texture et d'intensité sont mesurés. Il est important de remarquer que même dans les cas de cancers, la grande majorité des cellules d'une lame est normale. Les cas de cancers présentent généralement très peu de cellules suspectes (ou malignes).
Nous avons comparé les distributions des cellules des individus sains et des individus ayant un cancer pour essayer de déterminer les caractéristiques ou groupes de caractéristiques cellulaires discriminants le mieux ces deux populations. Les résultats de notre méthode seront donnés sur ce jeu de données, puis comparés à des méthodes de classification supervisée classiques.

Bibliographie

Belien, J.A.M., Baak, J.P.A., van Diest, P.J., Misere, B.N.L.H.M., Meijer, G.A., Bergers, L. (1997) Prognostic value of image and flow cytometric DNA ploidy assessments in invasive breast cancer, *Electr. J. Pathol*, 3, 972-979

Cuevas, A., Febrero, M. et Fraiman, R. (2004) An anova test for functional data, *Computational Statistics and Data Analysis*, 44, 111–122.

Delicado, P. (2007) Functional k-sample problem when data are density functions, *Computational Statistics and Data Analysis*, 22, 391–440.

Doudkine, A., MacAulay, C., Poulin, N., Palcic, B. (1995) Nuclear texture measurements in image cytometry, *Pathologica*, 87, 286-299

Ferraty, F., Vieu, P. (2003) Curves discrimination : a non parametric functional approach, *Computational Statistics and Data Analysis*, 44, 161–173.

Ferraty, F., Vieu P. (2006) Nonparametric Functional Data Analysis, Springer.

Ferraty, F. Romain, Y. (2010). Handbook on functional data analysis and related topics. Oxford University Press, to appear.

Hastie, T., Tibshirani, E., Friedman, J. (2009) The Elements of Statistical Learning, Springer

Kneip, A., Gasser, T. (1992) Statistical tools to analyse data representing a sample of curves, *The Annals of Statistics*, 20, 1266-1305

Kneip, A., Engel, J. (1995) Model estimation in nonlinear regression under shape invariance, *The Annals of Statistics*, 23, 551-570

Kneip, A., Utikal, K.J. (2001) Inference for densities families unsing functional principal component analysis, *Journal of the American Statistical Association*, 96, 519–542.

Palcic, B., Garner, D.M., Beveridge, J., xiao Rong Sun, Doudkine, A., Macaulay, C., Lam, S.,Payne, P.W. (2002) Increase of sensitivity of sputum cytology using high-resolution image cytometry: field study results,*Cytometry*, 50, 168-176

Payne, P.W., Sbebo, T.J., Doudkine, A., Garner, D., MacAulay, C., Lam, S., LeRichie, J.C., Palcic, B. (1997), Sputum screening by quantitative microscopy : a reexamination of a portion of the National Cancer Institute Cooperative Early Lung Cancer Study, *Mayo Clin Proc*, 72, 697-704

Ramsay, J.O., Silverman, B.W. (2005) Functional Data Analysis, Springer.

Sheather, S.J., Jones, M.C. (1991) A reliable data-based bandwidth selection method for kernel density estimation, *Journal of the Royal Statistical Society. Series B (Methodological)*, 53, 983-990

Wasserman, L. (2007) All of Nonparametric Statistics, Springer.