Analyse multiple de données longitudinales dans le cadre des modèles de mélanges

Cédric Noel, Jang Schiltz

Analyse multiple de données longitudinales dans le cadre des modèles de mélanges.

Cédric NOEL[1] Jang SCHILTZ[2]

cedric.noel@univ-lorraine.fr jang.schiltz@uni.lu

[1] IUT de Thionville-Yutz, Université de Lorraine
Faculty of Science, Technology and Medicine, University of Luxembourg

[2] Université du Luxembourg
Department of Finance, Luxembourg School of Finance

Thèmes – *Biologie - Santé - Économie - Informatique - Mathématiques*

Résumé – *Les modèles de mélanges appliqués à des données longitudinales peuvent servir à dégager des groupes qui suivent une trajectoire commune. Plusieurs modèles peuvent servir à modéliser la loi de probabilité suivie par les individus. Leur trajectoire est modélisée par une régression linéaire en la variable temporelle. Cet article présentera la cas où la variable d'intérêt est multidimensionnelle. Un individu sera représenté par plusieurs trajectoires qui représentent plusieurs mesures au cours du temps. Le modèle présenté cherchera à trouver des groupes d'individus suivants des trajectoires communes pour chaque mesure et cherchera à comprendre l'intensité des liens entre eux.*

Mots-clés – *Modèle de mélanges, Données longitudinales, Cluster, R, Modélisation.*

1 Introduction

On appelle données longitudinales des mesures répétées au cours du temps. Par exemple, en finance on peut mesurer le salaire de différents employés au cours d'une période donnée en criminologie on peut mesurer un indice d'agression physique au cours du temps ou encore en médecine, par exemple, on peut mesurer l'électroencéphalogramme de malades durant une période de temps donnée afin de prévoir les chances de survie. Pour un individu de l'étude, on mesure différentes valeurs de la variable d'intérêt au cours du temps. Elles peuvent être représentées par des tendances temporelles ou trajectoires, et elles peuvent dépendre d'une ou plusieurs variables. La complexité et la variabilité des mesures peuvent être modélisées en introduisant plusieurs groupes. On introduit ainsi K groupes et l'assignement des individus à l'intérieur d'un groupe est basé sur un degré de similarité de la trajectoire des individus. Ce modèle ne suppose pas de variabilités à l'intérieur d'un groupe (par définition, un groupe contient des individus ayant le même comportement) et la variabilité est supposée la même dans chaque groupe. Dans certains cas, les individus de l'étude peuvent être mesurés à travers plusieurs grandeurs d'intérêts, par exemple l'évolution temporelle de la pression sanguine et celle de la température. Dans ce cas, les trajectoires de ces grandeurs peuvent être liées et il faut donc les étudier conjointement.

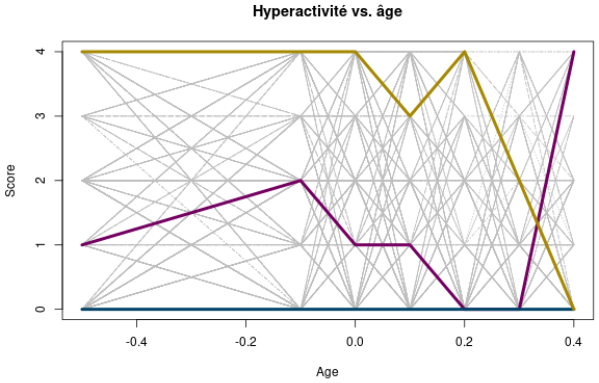

FIGURE 1 – Trajectoires des scores d'hyperactivité en fonction de l'âge.

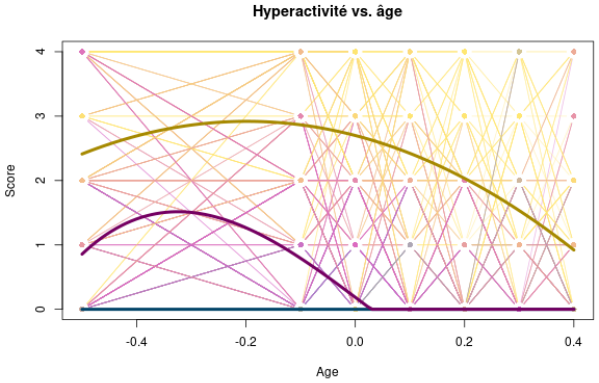

FIGURE 2 – Trajectoires des scores d'hyperactivité en fonction de l'âge après classification.

2 Modèle simple

2.1 Exemple introductif

L'étude "Montreal Longitudinal Study" [1] étudie le développement des comportements antisociaux de la maternelle au lycée pour des enfants. On considèrera dans cet article un score d'hyperactivité (échelle de 0 à 4) mesuré au cours du temps pour chaque participant. Ainsi pour un individu, les scores forment des trajectoires temporelles. La figure 1 montre les scores de chaque participant et en couleurs certaines trajectoires particulières. Il est légitime de penser que ces trajectoires ne sont pas les mêmes pour chaque individu (par exemple les individus colorés sur la figure 1) mais que, par contre, il existe plusieurs schémas évolutifs suivis approximativement par les individus. A partir des données, comment trouver le nombre de groupes et la forme des trajectoires pour partitionner l'ensemble des individus ?

2.2 Modèle

Soit Y_i une variable longitudinale appartenant à une population de taille N. Soit, $Y_i = y_{i_1}, \cdots, y_{i_T}$ T mesures d'une variable Y prises à des temps $t_1, \cdots, t_T$, dans notre exemple il s'agit des scores d'hyperactivité. On suppose que la population est divisée en K sous-populations homogènes et que, étant donné un groupe k, les réalisations y_{i_t} sont indépendantes sur les T périodes de mesure. Nagin [2] propose d'écrire la densité f de Y sous la forme d'un modèle de mélanges $f(y_i;\psi) = \sum_{k=1}^{K} \pi_k g_k(y_i;\Theta_k)$ où π_k est la probabilité pour un individu i d'appartenir au groupe k, $g_k(\cdot)$ est la distribution de y_i conditionnelle à l'appartenance à un groupe k et Θ_k sont des paramètres décrivant la forme des trajectoires suivies par les individus d'un groupe k.
La vraisemblance s'écrit

$$\prod_{i=1}^{N} \sum_{k=1}^{K} \frac{e^{x_i\theta_k}}{\sum_{k=}^{K} e^{x_i\theta_k}} \prod_{t=1}^{T} g_k\left(y_{it}|\Theta_k\right)$$

où $g^k(\cdot|\Theta_k)$ désigne également la densité conditionnelle de y_{it} sachant le temps et éventuellement d'autres covariables. Elle peut suivre différentes lois : Poisson, Logit, Normale ou encore Beta.

Les paramètres du modèle sont calculés à l'aide du package R, `trajeR` (github.com/gitedric/trajeR) en utilisant des lois normales pour les densités. La figure 2 affiche le résultat de la recherche de trois groupes. On distingue trois comportements différents : le premier groupe a un score nul tout le temps, le second a un score qui décroît rapidement tandis que le dernier a un score qui décroît lentement.

3 Modèle multiple

3.1 Exemple introductif

Dans l'étude de Montréal, un score d'opposition aux règles a été évalué au cours du développement des enfants. C'est une échelle comprise entre 0 et 10. On peut supposer que les scores d'hyperactivité et d'opposition sont liés et que les groupes s'influencent mutuellement. La méthode présentée ici permet de généraliser le modèle précédent au cas multiple.

3.2 Modèle

Dans ce cas, on suit J variables $Y^1,\dots,Y^J$ concernant les mêmes N individus. Un individu i est donc représenté par J trajectoires qui peuvent être indépendantes entre elles ou non. On suppose, étant donné un groupe, la condition d'indépendance conditionnelle pour les réalisations séquentielles des $\left(Y^j\right)_{1\leq j\leq J}$ sur les T périodes de mesure.

3.3 Modèle dépendant

On suppose que les variables $\left(Y^j\right)_{1\leq j\leq J}$ peuvent être dépendantes entre elles. Dans ce cas, la probabilité jointe conditionnelle aux temps ou autres variables ne peut plus s'écrire comme auparavant, elle change pour chaque variable j. On a alors $P(Y_i^1,\dots,Y_i^J)=\sum_{(k_1,\dots,k_J)\in K_1\times\cdots\times K_J}\pi_{k_1\dots k_J}\prod_{j=1}^J P^{k_j}(Y_i^j|\Theta_{k_j}^j)$. Avec l'aide des probabilités conditionnelles, la probabilité jointe s'écrit alors sous la forme $P(Y_i^1,\dots,Y_i^J)=\sum_{(k_1,\dots,k_J)\in K_1\times\cdots\times K_J}\pi_{k_J|k_1\dots k_{J-1}}\times\dots\times\pi_{k_2|k_1}\times\pi_{k_1}\prod_{j=1}^J\prod_{t=1}^{T^j}p^{k_j}(y_{it}^j|\Theta_{k_j}^j)$ où $\pi_{k_J|k_1\dots k_{J-1}}$ est la probabilité pour un individu i d'appartenir au groupe k_J pour la variable J sachant l'appartenance aux groupes $k_1,\cdots,k_{J-1}$ pour les variables $1,\cdots,J-1$. En adaptant une approche utilisée par Bel [3] dans le cadre d'un modèle multinomial de choix, nous pouvons réécrire la probabilité pour un individu d'appartenir à un groupe pour une variable et par la suite les probabilités jointes. Soit $Z_i=\left(Z_i^1,\dots,Z_i^J\right)$ une variable qui indique le groupe dans lequel est l'individu i pour chaque variable j, $P\left(Z_i^j=k|z_i^h\ for\ h\neq j,X_i^j\right)=e^{B_{ik}^j}/\sum_{l=1}^{K_j}e^{B_{il}^j}$ où $B_{ik}^j=\theta_k^jX_i^j+\sum_{h\neq j}\psi_{kz_i^h}^{jh}$, θ_{k0}^j est l'interception, θ_k^j est un vecteur correspondant à la variable X_i^j, z_i^h est le groupe dans lequel est l'individu i pour la h-ème mesure et ψ_{kl}^{jh} peut être interprété comme un paramètre mesurant l'association entre le fait d'appartenir au groupe k pour la variable j et celle d'appartenir au groupe l pour la variable h. Si le paramètre est positif alors la probabilité que z_i^j et z_i^h varient semble est plus grande que celle qu'ils varient séparément. Pour des raisons d'identifiabilité, un groupe est choisi comme référence. Il est intéressant de noter que si les paramètres ψ sont tous nuls alors on retrouve bien la définition de π_k.

Concrètement, cette écriture permet de prendre en compte d'éventuelles variables qui pourraient influencer les probabilités pour chaque variable mais aussi de prendre en compte les relations entre les différents groupes des différentes variables.

On peut montrer que la vraisemblance peut s'écrire sous une forme ressemblante à celle du cas univariable

$$\prod_{i=1}^N\sum_{(k_1,\dots,k_J)\in K_1\times\cdots\times K_J}\frac{e^{\mu_{z_i}}}{\sum_{z_i\in S}e^{\mu_{z_i}}}\prod_{j=1}^J\prod_{t=1}^{T^j}g^{k_j}(y_{it}^j;\Theta_{k_j}^j)$$

où $\mu_{z_i}=\sum_{j=1}^J\left(\theta_{z_i^j}^jX_i^j+\sum_{h<j}\psi_{z_i^hz_i^j}^{hj}\right)$.

Les paramètres du modèle sont évalués à l'aide du package R `trajeR` en utilisant des lois normales pour les deux variables et en choisissant de prendre trois groupes pour le premier score et quatre pour le second. Par exemple, les premiers paramètres d'association sont $\psi_{22}^{12}\simeq 19,97949$, $\psi_{23}^{12}\simeq 10,51569$ et $\psi_{24}^{12}\simeq 8,94481$. Ils indiquent que le groupe 2 de la variable 1 varie plus souvent avec le groupe 2 de la variable 2 qu'avec les autres groupes. L'analyse de l'ensemble des paramètres ψ indique que les groupes 1, 2 et 3 de chaque série varient respectivement plus souvent ensemble, alors que le groupe 4 de la série 2 varie plus souvent avec le groupe 3 de la série 1. Les probabilités jointes d'appartenance à chaque groupe peuvent être facilement déduites.

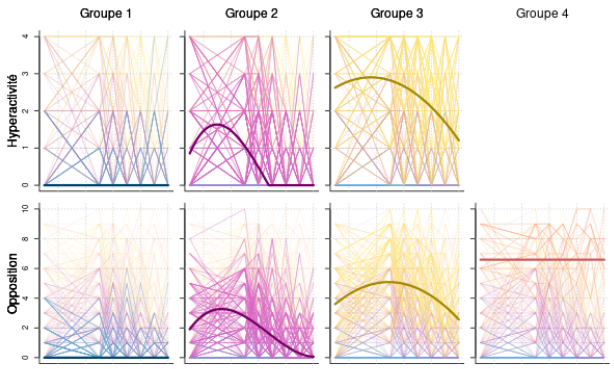

FIGURE 3 – Groupes pour les scores d'hyperactivité et d'opposition.

Références

[1] Richard E. Tremblay, Frank Vitaro, Daniel Nagin, Linda Pagani, and Jean R. Séguin. *The Montreal Longitudinal and Experimental Study*, pages 205–254. Springer US, Boston, MA, 2003.

[2] Daniel S. Nagin. *Group-Based Modeling of Development*. Harvard University Press, 2005.

[3] Koen Bel and Richard Paap. A multivariate model for multinomial choices. Technical report, 2014.