trajeR - une nouvelle librairie R pour les modèles de mélanges pour données longitudinales.

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 – *Statistique - Informatique - Santé - Économie - Sociologie - Biologie*

Résumé – *Cet article présente une nouvelle librairie R qui permet de travailler avec des modèles de mélanges pour données longitudinales. Le but de ces méthodes est de trouver des groupes homogènes de sujets qui seront caractérisés par une trajectoire type. Notre librairie permet de calibrer ce genre de modèles pour 3 types de lois statistiques sous-jacentes, les lois normales tronquées, les lois ZIP (Zero Inflated Poisson) et les lois de Bernoulli. L'influence du temps sur les trajectoires types est modélisée par un modèle de régression linéaire. Le modèle permet également d'utiliser d'autres variables qui peuvent influencer les trajectoires types ou l'appartenance aux groupes.*

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

1 Introduction

Des données longitudinales se rencontrent souvent en pratique. Pour un individu de l'étude, on mesure différentes valeurs de la variable d'intérêt au cours du temps. Par exemple, en finance on peut mesurer le salaire de différents employés au cours d'une période donnée [1], en criminologie on peut mesurer un indice d'agression physique au cours du temps [2] 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 [3]. La famille des modèles en équations structurelles des courbes de croissance (en anglais LGM - Latent Growth Model -) étudie les variations inter-individuelles, c'est-à-dire entre des groupes d'individus et les variations intra-individuelles, c'est-à-dire entre les individus à travers le temps. Ces variations peuvent être représentées par des tendances temporelles ou trajectoires et elles peuvent dépendre d'une ou de plusieurs variables. La complexité des mesures est modélisée en introduisant K groupes d'individus. L'affectation des individus aux différents groupes est basée sur un degré de similarité de la trajectoire des individus. Comme, par définition, un groupe contient des individus ayant le même comportement, la variabilité à l'intérieur d'un groupe donnée est modélisée par une loi normale de moyenne zéro et d'écart-type constant pour un groupe donné.

2 Modèle

On considère une variable Y qui dépend du temps qui est définie pour une population de taille N. Plus précisément, on suppose qu'on a T mesures $Y_i = y_{i_1}, \cdots, y_{i_T}$ de la variable Y prises à des temps $t_1, \cdots, t_T$ et appartenant à un échantillon de taille n. Pour un groupe donné, on suppose l'indépendance conditionnelle pour les réalisations successives sur l'intervalle de temps considéré. Cela veut dire que la distribution théorique au temps t ne dépend pas des réalisations antérieures, ce qui implique que y_{i_t} ne dépend pas des valeurs passées $y_{i_{t'}}$, $t' < t$.
On suppose que la population est divisée en K sous-populations homogènes $C_1, ..., C_K$. Soit $P^k(Y_i)$ la probabilité de Y_i sachant que l'individu i appartient un groupe C_k et $P(Y_i)$ la probabilité de la réalisation Y_i de Y. Alors, $P(Y_i) = \sum_{k=1}^{K} P(\omega_i \in C_k) P^k(Y_i)$ et la densité f de Y_i est décrite sous la forme d'un modèle de mélanges, comme dans [2], par $f(y_i; \psi) = \sum_{k=1}^{K} \pi_k g_k(y_i; \Theta_k)$.

Ici π_k est la probabilité pour un individu i d'appartenir au groupe k. Le modèle dépend des paramètres $\psi = (K, \pi_1, \cdots, \pi_{K-1}, \Theta_1, \cdots, \Theta_K)$. Θ_k est composé des paramètres β_k qui décrivent la forme de la trajectoire du groupe k (supposée polynomiale) en fonction du temps et éventuellement des paramètres δ_k qui mesurent l'effet d'une variable temporelle W_i. La probabilité π_k peut être, quant à elle, influencée par un ensemble de variables X_i et elle sera modélisée à l'aide d'une loi multinomiale. Dans la suite,

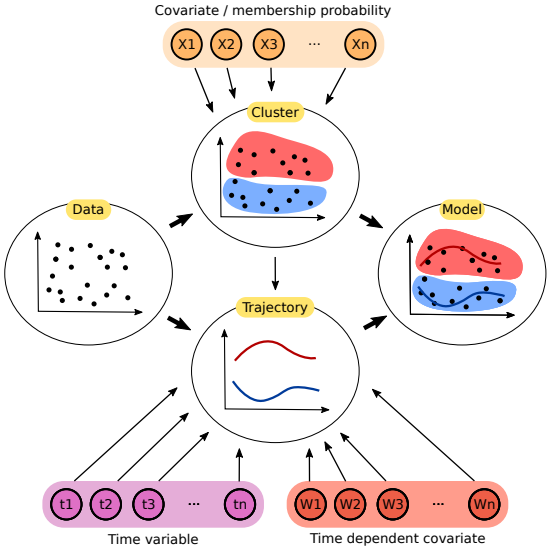

FIGURE 1 – Représentation du modèle

on note $A_{it} = \left(1, a_{it}, a_{it}^2, \cdots, a_{it}^{n_\beta - 1}\right)^t$ la variable de temps, $W_{it} = (w_{i1}, \cdots, w_{in_\delta})^t$, $\beta_k = (\beta_{k1}, \cdots, \beta_{kn_\beta})$, $\delta_k = (\delta_{k1}, \cdots, \delta_{kn_\delta})$ et $\varepsilon_{it} \sim \mathcal{N}(0;\ \sigma_k)$.

2.1 LOGIT

Le modèle LOGIT est utile pour modéliser des variables temporelles dont les réponses au cours du temps sont 0 ou 1. On utilisera un modèle logistique classique. Soit y_{it}^* tel que $y_{it}^* = \beta_k A_{it} + \delta_k W_{it} + \varepsilon_{it}$. On définit la variable binaire $y_{it} = 1$ si $y_{it}^* > 0$ et $y_{it} = 0$ si $y_{it}^* \leq 0$. Notons $\rho_{ikt} = P(Y_{it} = 1 | W_i = w_i, C_i = k)$ la probabilité que $y_{it} = 1$ sachant que l'individu i est dans le groupe k. Alors, $\rho_{ikt} = \frac{e^{\beta_k A_{it} + \delta_k W_{it}}}{1 + e^{\beta_k A_{it} + \delta_k W_{it}}}$.

2.2 Loi normale tronquée

Une loi normale tronquée intervient lorsque les mesures sur les données, au delà ou en deçà, d'une certaine borne sont concentrées sur une unique valeur. On suppose que la variable Y_{it} suit une loi normale tronquée, c'est-à-dire que ses valeurs sont comprises entre 2 bornes y_{min} et y_{max}.
On considère Y_{it}^* qui suit une loi normale telle que $y_{it}^* = \beta_k A_{it} + \delta_k W_{it} + \varepsilon_{k,it}$ et on peut lier y_{it}^* aux données censurées observées y_{it} de la façon suivante : $y_{it} = y_{min}$ si $y_{it}^* < y_{min}$, $y_{it} = y_{it}^*$ si $y_{min} \leq y_{it}^* \leq y_{max}$ et $y_{it} = y_{max}$ si $y_{it}^* > y_{max}$.

Si aucune valeur de y_{it}^* ne dépasse les bornes, alors y_{it}^* coïncide avec y_{it} et on retrouve une loi normale classique.

2.3 Modèle ZIP

Le modèle ZIP (Zero Inflated Poisson) est utilisé pour modéliser des situations de comptage dans lesquelles le nombre de zéros est trop important pour être modélisé par une loi de Poisson. Il utilise 2 différents processus : une distribution binaire qui génère les éventuels zéros en excès, et une distribution de Poisson qui génère le comptage.

Par conséquent, les zéros peuvent survenir pour deux raisons. La première est lors du comptage par la loi de Poisson avec une probabilité $P(Y_{it}=0)$ pour $Y_{it} \sim \mathcal{P}(\lambda_{ikt})$ et la seconde est le fait que la distribution binaire produit des zéros avec une probabilité ρ_{ikt}.

On a $P^k(Y_{it}=y_{it}) = \begin{cases} \rho_{ikt} + (1-\rho_{ikt})e^{-\lambda_{ikt}}, & y_{it}=0 \\ (1-\rho_{ikt})\frac{\lambda_{ikt}^{y_{it}} e^{-\lambda_{ikt}}}{y_{it}!}, & y_{it}>0 \end{cases}$

où λ_{ikt} est lié au temps par $\log(\lambda_{ikt}) = \beta_k A_{it} + \delta_k W_{it}$ et ρ_{ikt} par $\log\left(\frac{\rho_{ikt}}{1-\rho_{ikt}}\right) = \nu_k A_{it}$ où $\nu_k = (\nu_{k1}, \cdots, \nu_{kn_\nu})$.

3 Estimation des paramètres

Chacun des modèles précédents peut être estimé grâce à la méthode du maximum de vraisemblance. La vraisemblance étant une fonction très complexe, on utilise des méthodes d'optimisation de quasi-Newton pour rechercher les maximums et en particulier la méthode BFGS (Broyden–Fletcher–Goldfarb–Shanno). Cette méthode à l'avantage de fournir des propriétés asymptotiques des estimateurs et de pouvoir calculer leurs écart-types.
Une deuxième méthode utilisée pour estimer les paramètres est la méthode EM (Expectation Maximization) qui est une méthode itérative qui converge vers le maximum de vraisemblance. Elle alterne entre 2 étapes : une première qui consiste à calculer l'espérance de la vraisemblance à partir des derniers paramètres estimés (E) et une seconde (M) qui consiste à maximiser la vraisemblance trouvée à l'étape (E). Les écart-types des paramètres sont calculés à l'aide de la méthode de Louis.

4 Évaluation du modèle

Le modèle peut être évalué à l'aide des méthodes BIC et AIC. Toutefois, deux étapes sont nécessaires. Une première, pour déterminer le nombre de groupes et une seconde, pour déterminer le degré des polynômes à utiliser.

5 Librairie R

R est un langage de programmation et un logiciel open source utilisé pour les calculs statistiques et l'analyse de données. Une librairie a été développée pour estimer de tels modèles : `trajeR`. D'autres logiciels existent, comme la librairie pour SAS ou S de Nagin & Jones [4] ou alors Mplus, mais ils ne sont pas libres et leur code n'est pas consultable. Ce sont des "boîtes noires" qui donnent un résultat sans savoir précisément comment il est calculé. La fonction principale de notre librairie est `trajeR()` qui a plusieurs paramètres pour contrôler le modèle. Par exemple, dans la suite, on considère un ensemble de données longitudinales suivant une loi normale et on suppose l'existence de plusieurs groupes dont les individus possèdent un développement similaire au cours du temps. A titre d'illustration, on décide de rechercher 3 groupes dont les membres suivent des trajectoires polynomiales de degré 0, 3 et 4 afin de capter les changements de ces trajectoires. Ces choix peuvent être dictés par la connaissance à-priori d'experts ou en utilisant un critère de sélection comme le critère BIC par exemple.

```
> trajeR(Y = data[,2 :11], A = data[,12 :21], degre = c(0,3,4), Model = "CNORM", Method = "L", hessian = TRUE, ssigma = FALSE)
```

On note les paramètres sous la forme $(\beta|\pi|\sigma)$.

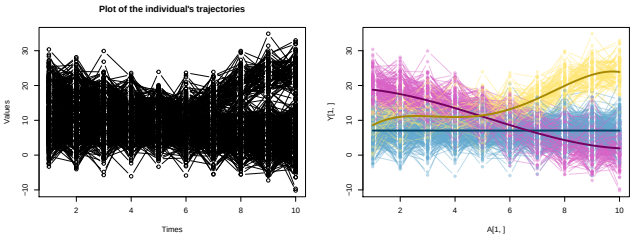

FIGURE 2 – Données initiales à gauche et données partagées en 3 groupes avec leur trajectoire.

Pour le groupe 1, les paramètres théoriques sont $(2.797, 8.809, -3.201, 0.463, -0.021|0.32|4)$ et les paramètres du modèle sont $(1.67, 10.118, -3.707, 0.538, -0.025|0.192|4.002)$, pour le groupe 2, on a $(7|0.54|4)$ et $(7.049|0.459|3.958)$ et pour le groupe 3 $(19.545, -0.297, -0.407, 0.026|0.14|4)$ et $(19.305, -0.093, -0.456, 0.029|0.349|4.111)$.

6 Conclusion

Cet article a permis une brève introduction aux modèles de mélanges pour des données longitudinales dont le but était de trouver des groupes qui suivent une même trajectoire au cours du temps. Les modèles utilisés sont LOGIT, normal censoré et ZIP. Enfin, une librairie pour le logiciel R a été présentée, montrant comment résoudre de telles situations.

Références

[1] Jang Schiltz. A generalized finite mixture model. Proceedings of the 60th ISI World conference (2015) p.1562-1567.

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

[3] Jonathan Elmer, Bobby L. Jones, Vladimir I. Zadorozhny, Juan Carlos Puyana, Kate L. Flickinger, Clifton W. Callaway, and Daniel Nagin. A novel methodological framework for multimodality, trajectory model-based prognostication. 137 :197–204.

[4] Bobby L. Jones and Daniel S. Nagin. Advances in group-based trajectory modeling and an sas procedure for estimating them. *Sociological Methods & Research*, 35(4) :542–571, 2007.