KMLCOV : K-MEANS LONGITUDINAL AVEC AJUSTEMENT SUR COVARIABLES

Fabien Subtil[1] & René Ecochard[2] & Christophe Genolini[3]

[1] Hospices Civils de Lyon, Service de Biostatistique, CNRS UMR 5558, Laboratoire de Biométrie et Biologie Evolutive, 162 avenue Lacassagne, 69003 Lyon <fabien.subtil@chu-lyon.fr>
[2] Hospices Civils de Lyon, Service de Biostatistique, CNRS UMR 5558, Laboratoire de Biométrie et Biologie Evolutive, 162 avenue Lacassagne, 69003 Lyon <rene.ecochard@chu-lyon.fr>
[3] INSERM UMR 1027, 37 allées Jules Guesde, 31000 Toulouse <christophe.genolini@u-paris10.fr>

Résumé. La classification de données longitudinales est une méthode de plus en plus fréquemment utilisée en médecine. Les modèles de mélange permettent d'obtenir un compromis entre la classification et l'adéquation aux données. K-means est un algorithme dédié à la classification pure, mais il ne permet pas d'introduire des covariables dans la modélisation des trajectoires, et suppose que les mesures sont réalisées toutes au même temps. Le package R kmlCov propose une extension de l'algorithme du k-means en utilisant la vraisemblance comme métrique de distance. Ceci permet d'introduire des covariables dans la modélisation des trajectoires (à effet fixe ou variable d'un groupe à l'autre), de prendre en compte différentes natures de variables trajectoire (continue, binaire, comptage), et d'autoriser des mesures à des temps différents d'un individu à l'autre.

Mots-clés. K-means, données longitudinales, classification, covariable

Abstract. Classification of longitudinal measurements is more and more frequent in medicine. Mixture models can be used for this purpose, but they result in a compromise between classification and good data modeling. K-means is an algorithm dedicated solely to classification, but it does not allow to take into account covariates in trajectory modeling, and it supposes that measurements are taken at the same times for all subjects. The R package kmlCov is an extension of the k-means algorithm that uses the opposite of the likelihood as distance metric. It allows to introduce covariates in trajectory modeling (with fixed or varying effects depending on the group), to classify different kinds of longitudinal data (continuous, binary or count data), and to have varying measurement times from one individual to another one.

Keywords. K-means, longidutinal data, classification, covariable

1 Contexte

L'analyse de mesures longitudinales est de plus en plus fréquente dans le domaine médical. Par exemple, les dosages réguliers d'un biomarqueur peuvent permettre le diagnostic ou le pronostic de pathologies. Un des intérêts de l'analyse de ces données longitudinales est d'identifier des groupes d'individus ayant des trajectoires similaires ; les groupes obtenus peuvent dans certains cas être reliés au pronostic. Ils permettent également de créer une typologie des individus.

La classification de données longitudinales est régulièrement effectuée par des modèles de mélanges (Nagin et Odgers, 2010 ; Pickles et Croudace, 2010). Il est supposé que la trajectoire de chaque individu peut être reproduite par un mélange d'un petit nombre de trajectoires typiques, avec des poids différents donnés à chacune de ces trajectoires d'un individu à l'autre. Pour un

individu et une trajectoire typique donnée, ce poids correspond à la probabilité a posteriori que l'individu soit issu du groupe en question. La particularité de ces modèles de mélange est qu'ils ne classent pas les individus pour déterminer les trajectoires typiques : la partition des individus dans les groupes ne fait pas partie des paramètres du modèle. Chaque individu participe donc, plus ou moins, à l'estimation des paramètres de toutes les trajectoires typiques. La signification de ces trajectoires typiques est difficile à établir puisqu'elles ne représentent pas des individus particuliers. Ainsi, les trajectoires typiques ne sont que des construits utilisés pour mieux modéliser les données en tenant compte d'une hétérogénéité inter-individuelle latente. Une fois les trajectoires typiques estimées, il reste néanmoins possible de classer les individus en les affectant au groupe pour lequel la probabilité d'appartenance a posteriori est maximale (MAP).

Une approche alternative aux modèles de mélange est l'utilisation d'algorithmes de type k-means (MacQueen, 1967). L'algorithme vise à déterminer la partition des individus et les trajectoires typiques de sorte à minimiser une fonction de perte :

$$\sum_{h=1}^{k} \sum_{\boldsymbol{y}_i \in P_h} d(\boldsymbol{y}_i, \boldsymbol{\theta}_h)$$

où d est une métrique de distance (par exemple euclidienne), P_h représente l'ensemble des trajectoires classées dans le groupe h, $\boldsymbol{y}_i$ est le vecteur des observations pour l'individu i, et $\boldsymbol{\theta}_h$ est l'ensemble des paramètres caractérisant la trajectoire typique h. L'algorithme du k-means est une méthode de classification très géométrique. Elle nécessite que les individus aient des mesures effectuées aux même temps. De plus, il ne permet pas de tenir compte de covariables pouvant influencer les trajectoires individuelles. Par exemple, l'analyse des trajectoires d'un biomarqueur mesurant la réponse à un traitement peut faire apparaître des groupes de réponse : bons répondeurs et mauvais répondeurs. Mais une variable extérieure, telle que l'âge, peut entraîner une baisse du niveau du biomarqueur chez les personnes âgées quel que soit le groupe. Ne pas tenir compte de cette variable dans la classification peut entraîner une mauvaise prise en compte de l'hétérogénéité inter-individuelle, et faire apparaître des groupes qui ne sont pas forcément reliés au phénomène d'intérêt (ici la réponse au traitement).

Le package R kmlCov est une extension de l'algorithme k-means, permettant de prendre en compte des covariables dans la modélisation des trajectoires, de modéliser des variables réponses non forcément continues (binaire, comptage) et de tenir compte de la variabilité des temps de mesure d'un individu à l'autre.

2 Vraisemblance classifiante et K-means

L'extension de l'algorithme k-means nécessite de passer d'une vision géométrique de la classification à une version probabiliste. Celeux et Govaert (1993) ont défini la notion de vraisemblance classifiante pour des données non longitudinales, par exemple pour classer des individus selon des mesures de différents marqueurs. La vraisemblance classifiante correspond à la vraisemblance augmentée utilisée dans les algorithmes EM pour l'estimation des paramètres d'un modèle de mélange :

$$CL(\boldsymbol{y}; \boldsymbol{\theta}, \boldsymbol{\pi}) = \prod_{h=1}^{k} \prod_{i=1}^{n} \left(\pi_h f_h(\boldsymbol{y}_i, \boldsymbol{\theta}_h)\right)^{z_{ih}}$$

z_{ih} = 1 si l'individu i est affecté au groupe h, 0 sinon. f_h correspond à la vraisemblance des mesures de l'individu i sous l'hypothèse qu'il appartient au groupe h.

La notion de vraisemblance classifiante peut être étendue au cas de données longitudinales. Pour cela, il est supposé que les mesures de l'individu i sont indépendantes les unes des autres conditionnellement au fait d'appartenir au groupe h. Cette hypothèse est équivalente à celle

effectuée par Nagin dans les modèles de mélange de type latent class growth analysis (2005). Ainsi :

$$f_h(\boldsymbol{y}_i, \boldsymbol{\theta}_h) = \prod_{j=1}^{n_i} g_h(y_{ij}, \boldsymbol{\theta}_h)$$

où g_h est la loi de probabilité associée à la $j^{\text{ème}}$ mesure de l'individu i (y_{ij}), par exemple une loi normale ou Gamma pour des données continues, de Bernoulli pour des données binaires, ou de Poisson pour des données de comptage. L'hypothèse d'indépendance des mesures d'un même individu conditionnellement à l'appartenance au groupe peut être assouplie en introduisant une structure de covariance entre les observations.

Contrairement aux modèles de mélange, les paramètres à estimer sont ici à la fois les paramètres des trajectoires typiques et la partition des individus dans les groupes. Ainsi, l'estimation des paramètres des trajectoires typiques est basée sur de vrais individus affectés aux groupes. Les groupes ne sont plus des construits théoriques, ils ont une signification réelle en termes d'individus.

Les termes π_h représentent l'influence globale d'un groupe dans la classification. Dans certains cas, ces termes sont omis, ce qui consiste à supposer a priori que tous les groupes ont une influence similaire.

La comparaison de la fonction à minimiser pour k-means et de la vraisemblance classifiante à maximiser montre que ces deux méthodes sont équivalentes lorsque :

- la distance euclidienne est utilisée pour K-means, et la loi normale est utilisée pour la vraisemblance classifiante ;
- les variances résiduelles sont supposées identiques dans les différents groupes pour la vraisemblance classifiante ;
- les proportions π_h sont supposées égales a priori (fixées à $1/k$).

De façon plus générale, en prenant comme métrique de distance dans k-means l'opposé de la vraisemblance, il est possible d'obtenir des résultats équivalents à ceux obtenus par vraisemblance classifiante. Ceci a pour avantage :

- d'utiliser des lois de probabilité qui reflètent la vraie nature des données, par exemple une loi gamma lorsque la variabilité des données augmente avec la moyenne ;
- de pouvoir prendre en compte des covariables dans la modélisation des trajectoires, en les rajoutant dans l'espérance de la variable aléatoire modélisée ; ces covariables peuvent avoir un effet identique ou variable d'un groupe à l'autre.

Par ailleurs, l'espérance de y_{ij} sachant l'appartenance au groupe h peut être modélisée par une fonction paramétrique du temps, et non plus par un paramètre à chaque pas de temps comme l'exige k-means. Des polynômes orthogonaux, fractionnaires ou bien des fonctions non linéaires plus complexes comme celles employées en pharmacocinétique peuvent être utilisés. Cela permet de s'affranchir des mesures à temps fixe pour chaque individu.

L'algorithme du k-means est donc généralisable à une large gamme de problèmes de classification en utilisant comme métrique de distance l'opposé de la vraisemblance.

Bibliographie

[1] Nagin, D. S. et Odgers, C. L. (2010), Group-based trajectory modeling in clinical research, *Annual Revue of Clinical Psychology*, 6, 109-138.
[2] Pickles, A. et Croudace, T (2010), Latent mixture models for multivariate and longitudinal

outcomes, *Statistical Methods in Medical Research*, 19, 271-89.
[3] MacQueen, J. B. (1967), Some methods for classification and analysis of multivariate observations, *Proceedings of 5-th Berkeley Symposium on Mathematical Statistics and Probability*, University of California Press, Berkeley.
[4] Celeux G. et Govaert G. (1993), Comparison of the mixture and the classification maximum likelihood, *Journal of Statistical Computation and Simulation*, 47, 27-146.
[5] Nagin, D. S. (2005), *Group-based modeling of development*, Harvard University Press, London.
Noteur, U. N. (2003), Sur l'intérêt des résumés, *Revue des 0rganisateurs de Congrès*, 34, 67–89.