KmL3D: K-means pour données longitudinales jointes

Christophe Genolini[1,2,⋆], **Jean-Baptiste Pingault**[3,4] **and Bruno Falissard**[4,5]

1. UMR 1027, INSERM, Université Paul Sabatier, Toulouse III, France
2. CeRSM (EA 2931), UFR STAPS, Université de Paris Ouest-Nanterre-La Défense, France
3. Research Unit on Children's Psychosocial Maladjustment, University of Montreal and Sainte-Justine Hospital, Montreal, Quebec, Canada
4. UI669 INSERM, Paris, France
5. University Paris-Sud and University Descartes, Paris, France
⋆Contact author: <christophe.genolini@u-paris10.fr>

Mots clefs : K-means, trajectoires jointes, partitionnement, interface graphique, graphes 3D dynamiques.

Les études longitudinales sont des études dans lesquelles les mêmes variables sont mesurées de manière répétée au cours du temps. Chaque suite de mesures, appelée variable-trajectoire, reflète l'évolution d'un phénomène. Ces études touchent des domaines variés (médecine, épidémiologie, économie, sociologie, informatique, physique,...) et sont de plus en plus nombreuses.
Ces dernières années ont vu se développer de nouveaux outils pour l'analyse de ces évolutions. Les plus largement utilisés sont les techniques de partitionnement (comme Proc Traj [1] ou KmL [2,3]). Elles consistent à grouper ensemble les individus dont les trajectoires se ressemblent et ainsi à définir des « trajectoires types » qui reflètent le comportement « moyen » des individus d'un même sous-groupe.
Les études longitudinales travaillent généralement non pas sur une mais sur de nombreuses variables-trajectoires. Se pose alors la question du partitionnement de plusieurs variables-trajectoires (appelées « trajectoires jointes »)

Si on note m le nombre de variables-trajectoires à analyser, la méthode d'analyse classique consiste à partitionner les m variables-trajectoires indépendamment les unes des autres, à obtenir ainsi m partitions P_i et de considérer comme partition finale la partition croisée $P = \prod_{1 \leq i \leq m} P_i$.

Figure 1: Graphe 3D dynamique. Cliquer sur le graphe avec le bouton gauche, puis faire bouger la souris pour changer de point de vue.

Or, de même que dans le cas classique les variables sont souvent corrélées, il est très probable que des variables-trajectoires évoluent conjointement. De plus, la richesse d'information contenue dans les trajectoires permet d'envisager des modes d'interaction bien plus complexes qu'une simple corrélation, ou qu'une monotonie conjointe. Malheureusement, la méthode des partitions croisées ne permet pas de détecter ce genre d'interactions complexes.
Une solution à ce problème consiste à partitionner simultanément les m trajectoires jointes. Pour cela, nous avons considéré un espace vectoriel de dimension $m+1$. Sur le premier axe, nous avons placé le temps. Chacun des m autres axes correspond à une variable-trajectoire. Nous avons ensuite défini une distance entre trajectoires jointes dans cet espace vectoriel. Au final, cela nous a permis d'appliquer k-means, un algorithme de partitionnement classique, aux trajectoires jointes. Un exemple en dimension 3 ($m = 2$ variables-trajectoires) est donné figure 1.
La procédure a été publiée dans [4], utilisée dans [5], puis programmée et mise à disposition de la communauté scientifique sous forme d'un package R, le package KmL3D [6]. Disponible sur le site du CRAN, il est dédié au partitionnement des trajectoires jointes. Comme le package KmL, il propose à l'utilisateur un certain nombre de solutions face aux problèmes posés par le partitionnement des données longitudinales. 12 méthodes d'imputation des manquantes sont proposées ; des exécutions multiples de k-means, en variant les conditions initiales et / ou le nombre de groupes considéré, sont gérés automatiquement ; une interface graphique conviviale et interactive permet à l'utilisateur de représenter graphiquement les partitions obtenues. Enfin, dans le cas de deux trajectoires jointes (représentation graphique dans $\mathbf{R}^3$), il est possible d'exporter des graphiques « 3D dynamiques » vers des fichiers pdf. Ces graphes dynamiques offrent à l'utilisateur la possibilité de changer de point de vue, permettant ainsi une meilleure visualisation de la troisième dimension (voir figure 1).

Références

[1] Jones, B. L., Nagin, D. S., & Roeder, K. (2001). A SAS procedure based on mixture models for estimating developmental trajectories. *Sociological Methods & Research*, 29(3), 374-393.
[2] Genolini, C., & Falissard, B. (2010). KmL: k-means for longitudinal data. *Computational Statistics*, 25(2), 317-328.
[3] Genolini, C., & Falissard, B. (2011). KmL: A package to cluster longitudinal data. *Computer Methods and Programs in Biomedicine*, 104(3), 112-121.
[4] Genolini, C., Pingault, J. B., Driss, T., Côté, S., Tremblay, R. E., Vitaro, F., ... & Falissard, B. (2012). KmL3D: A non-parametric algorithm for clustering joint trajectories. *Computer methods and programs in biomedicine*.
[5] Pingault, J. B., Côté, S. M., Galéra, C., Genolini, C., Falissard, B., Vitaro, F., & Tremblay, R. E. (2012). Childhood trajectories of inattention, hyperactivity and oppositional behaviors and prediction of substance abuse/dependence: a 15-year longitudinal population-based study. *Molecular Psychiatry*.
[6] Christophe Genolini (2012). kml3d: K-means for joint Longitudinal data. *R package version 2.1.2*. http://CRAN.R-project.org/package=kml3d