

kmlShape : un algorithme de partitionnement des données longitudinales basé sur la forme des trajectoires

C. Genolini^{a,b,*}, R. Ecochard^{c,d}, T. Driss^b, S. Andrieu^{a,e} et F. Subtil^{c,d}

^a Inserm UMR U1027, Université Paul Sabatier, Toulouse III, France

^b CeRSM (EA 2931), UFR STAPS Université Paris Ouest-Nanterre-La défense, France

^c Hospices Civils de Lyon, Service de Biostatistique, F-69003 Lyon, France

^d CNRS, UMR5558, Laboratoire de Biométrie et Biologie Evolutive, F-69100, Villeurbanne, France

^e Département d'épidémiologie et de santé publique, CHU Toulouse, Toulouse, France

* Contact : <christophe.genolini@u-paris10.fr>

Mots clefs : Partitionnement, données longitudinales, k-means, dynamic time warping, distance de Fréchet, Moyenne de Fréchet

Contexte

Les données longitudinales sont des données dans lesquelles chaque variable est mesurée à plusieurs reprises au cours du temps. Une possibilité pour analyser ce genre de données consiste à les partitionner. La majorité des méthodes de partitionnement classent deux individus ensemble si leurs trajectoires sont proches à chaque temps de mesure. Ce type de méthode prend en compte les ressemblances locales mais ne tient pas vraiment compte de la forme générale des trajectoires. Or, dans un certain nombre de situations, l'évolution d'un phénomène peut avoir plus d'importance que son moment d'apparition. On souhaite dans ces situations pouvoir partitionner en classant dans le même groupe des individus dont les trajectoires ont des formes similaires, indépendamment d'un décalage dans le temps. Cette situation est illustrée figure 1.

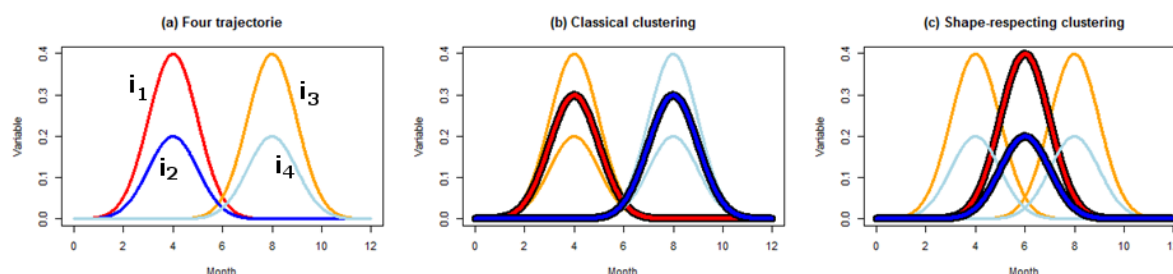


Figure 1: (a) 4 trajectoires (b) les méthodes de partitionnement classiques regroupent i_1 et i_2 dans un cluster (en rouge), i_3 et i_4 dans un autre (en bleu). (c) Un partitionnement basé sur les formes des trajectoires regroupe i_1 et i_3 dans un cluster (en rouge), i_2 et i_4 dans un autre (en bleu).

kmlShape : k-means pour distance silhouette

K-means [1] est un algorithme de partitionnement assez largement utilisé dans le cas de données longitudinales [2,3]. Son principe est d'alterner deux étapes : calcul des trajectoires moyennes de chaque groupe, puis calcul des distances entre chaque individu et la moyenne de son groupe (afin de pouvoir affecter un individu dans le groupe dont il est le plus proche). Pour appliquer k-means au cadre des partitionnements respectant les formes (partitionnement silhouette), il faut disposer d'une "distance silhouette" et d'une "moyenne silhouette". La distance de Fréchet fournit un tel outil.

Informellement, la distance de Fréchet entre deux trajectoires est souvent décrite en utilisant l'image d'une laisse : c'est la taille minimum que devrait faire une laisse reliant un homme et son chien qui parcourraient chacun une des deux trajectoires, à une vitesse libre (figure 2.a). Selon la distance de Fréchet, la courbe i_1 de la figure 1 est plus proche de la courbe i_3 que de la courbe i_2 . Du point de vue du partitionnement, cela signifie que i_1 et i_3 seront regroupés dans un même clusters, i_2 et i_4 dans un autre.

La moyenne de Fréchet entre deux trajectoires est simplement le milieu de la laisse reliant les deux trajectoires (trajectoire violette, figure 2.a). Cette définition peut se généraliser à n trajectoires en procédant par des regroupement deux par deux, un peu selon le principe d'un tournois.

L'ensemble de ces outils ont été intégrés dans la bibliothèque informatique **kmlShape**.

Exemple d'utilisation

La maladie d'Alzheimer est une maladie neuro-dégénérative qui entraîne un déclin des facultés cognitives et de la mémoire. Ictus [4] est une cohorte de 1380 patients atteints de la maladie d'Alzheimer suivis dans 12 pays d'Europe. L'évaluation semestrielle des patients comportait entre autres mesures le niveau d'autonomie dans la réalisation des activités (IADL). Une analyse classique des trajectoires d'IADL (en utilisant les modèles de mélange ou k-means) permet d'identifier les 4 groupes présentés figure 2.b. La caractéristique principale de ces groupes est de présenter un déclin régulier, assez proche les uns des autres. L'application de kmlShape appliqué sur quatre groupes identifie trois groupes similaires à ceux trouvés par les algorithmes classiques. Par contre, un quatrième groupe "déclin rapide" est maintenant détecté. La mise en évidence de ce groupe va permettre d'identifier des facteurs pronostics d'un déclin rapide et donc de mieux prévenir la maladie.

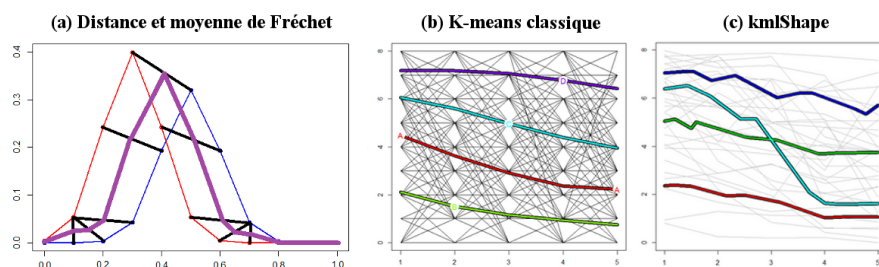


Figure 2: (a) La distance de Fréchet entre la courbe rouge et la courbe bleue est le plus long des segments noirs. La moyenne de Fréchet est la courbe passant par les milieux des segments noirs (trajectoire violette) (b) Trajectoires d'IADL partitionnées avec des méthodes classiques (k-means) (c) Trajectoires d'IADL partitionnées avec kmlShape

Références

- [1] Hartigan J, Wong M (1979) Algorithm as 136: A k-means clustering algorithm. In *Journal of the Royal Statistical Society Series C (Applied Statistics)* 28: 100-108.
- [2] Genolini C, Falissard B (2010) kml: k-means for longitudinal data. In *Computational Statistics* 25: 317-328.
- [3] Genolini C, Pingault JB, Driss T, Côté S, Tremblay RE, Vitaro F, Arnaud C, Falissard B. (2013) Kml3D: a non-parametric algorithm for clustering joint trajectories. In *Comput Methods Programs Biomed.* 109(1):104-11
- [4] Reynish E, Cortes F, Andrieu S, Cantet C, Olde Rikkert M, et al. (2007) The ictus study: A prospective longitudinal observational study of 1,380 ad patients in europe. In *Neuroepidemiology* 29: 29-38.