KmL: K-means for longitudinal data

KmL: K-means for Longitudinal Data

Christophe Genolini[1,2,3,*] Bruno Falissard[1,3,4]

Accepted in *Computational Statistics* as [1]

1. Inserm, U669, Paris, France
2. Univ Paris-Sud and Univ Paris Descartes, UMR-S0669, Paris, France
3. Modal'X,Univ Paris-Nanterre, UMR-S0669, Paris, France
4. AP-HP, Hôpital de Bicêtre, Service de Psychiatrie, Le Kremlin-Bicêtre, France
* Contact author: <christophe.genolini@u-paris10.fr>

Abstract

Cohort studies are becoming essential tools in epidemiological research. In these studies, measurements are not restricted to single variables but can be seen as trajectories. Statistical methods used to determine homogeneous patient trajectories can be separated into two families: model-based methods (like Proc Traj) and partitional clustering (non-parametric algorithms like k-means).

KmL is a new implementation of k-means designed to work specifically on longitudinal data. It provides scope for dealing with missing values and runs the algorithm several times, varying the starting conditions and/or the number of clusters sought; its graphical interface helps the user to choose the appropriate number of clusters when the classic criterion is not efficient.

To check KmL efficiency, we compare its performances to Proc Traj both on artificial and real data. The two techniques give very close clustering when trajectories follow polynomial curves. KmL gives much better results on non-polynomial trajectories.

1 Introduction

Cohort studies are becoming essential tools in epidemiological research. In these studies, measurements are not restricted to single variables but can be seen as trajectories. As for regular variables, statistical methods can be used to determine homogeneous patient trajectories [2, 3, 4, 5]. The field of functional cluster analysis can be separated into two families. The first comprises model-based methods. These are related to mixture modelling techniques or latent class analysis [6, 7].The second family relates to the more classical algorithmic approaches to cluster analysis, such as hierarchical or partitional clustering [8, 9, 10]. The pros and cons of both approaches are regularly discussed [11, 9], even if there is at present little data to show which method is preferable in which situation. In favour of mixture modelling or model-based methods more generally: 1/ formal tests can be used to check the validity of the partitioning; 2/ results are invariant in linear transformation, so there is no need to standardize variables (this will not be an issue on longitudinal data since all measurements are performed on the same scale), 3/ if the model is realistic, inferences about the data-generating process may be possible. On the other hand, traditional algorithmic methods can also have some potential advantages: 1/ they do not require any normality or parametric assumptions within clusters (they might be more efficient under a given assumption, but they do not require one; this can be of great interest when the task is to cluster data on which no prior information is available); 2/ they are likely to be more robust as regards numerical convergence; 3/ in the particular context of longitudinal data, they do not require any assumption regarding the shape of the trajectory (this is likely to be an important point: clustering of longitudinal data is basically an exploratory approach), 4/ also in the longitudinal context, they are independent from time-scaling. Even if both methods have been extensively studied, they still present considerable weaknesses, and first of all the difficulty in finding the exact number of clusters. [12, 13, 14, 15] provide examples of criteria used to solve this problem. [16, 17, 18, 19] compare them using artificial data. Even if criteria perform unequally, all of them fail on a significant proportion of data. Moreover, no study compares criteria specifically on longitudinal data. The problem of cluster selection is indeed an important issue for longitudinal data. More information about clustering longitudinal data can be found in [20]. Regarding software, longitudinal mixture modeling analysis has been implemented by B. John and D. Nagin [21, 22, 23, 24] in a procedure called Proc Traj on the SAS platform. It has already be extensively used in research on various topics [22, 25, 26, 27]. On the R platform [28], S.G.Buyske has proposed the mmlcr package, but the statistical background of this routine is not fully documented. Mplus [29] is also statistical software that provides a general framework that can deal with mixture modeling on longitudinal data. It can be noted that these three procedures are model-based. For the non-parametric solutions, numerous versions of k-means exist, whether strict [30, 31] or with variation [32, 33, 34, 35, 36, 37], but they have considerable drawbacks: 1/ they are not able to deal with missing values; 2/ since the determination of the number of clusters is still an open issue, they require the user to manually re-run k-means several times. In simulation, numerous authors use k-means to compare the different criteria used to find the best cluster number. But the performance of k-means has never been compared to parametric algorithms on longitudinal data.

The rest of this paper is organized as follows: section 2 presents KmL, a package implementing k-means (Lloyd version, [38]) Our package is designed for R platform and is available at [39]. It is able to deal with missing values; it also provides an easy way to run the algorithm several times, varying the starting conditions and/or the number of clusters looked for; its graphical interface helps the user to choose the appropriate number of clusters when the classic criterion is not efficient. Section 3 presents simulations on both artificial and real data. Performances of k-means on longitudinal data are compared to Proc Traj results (this appears as the fully dedicated statistical tool that is the most widely used in the literature). Section 4 is the discussion.

2 Algorithm

2.1 Introduction to K-means

K-means is a hill-climbing algorithm [9] belonging to the EM class (Expectation-Maximization) [31]. EM algorithms work as follows: Initially, each observation is assigned to a cluster. Then the optimal clustering is reached by alternating two phases. During the *Expectation* phase, the centers of each cluster (called seeds) are computed. Then the *Maximisation* phase consists in assigning each observation to its “nearest cluster”. The alternation of the two phases is repeated until no further changes occur in the clusters.

More precisely, consider a set S of n subjects. For each subject, an outcome variable Y at t different times is measured. The value of Y for subject i at time k is noted as y_{ik}. For subject i, the sequence y_{ik} is called a trajectory, it is noted $y_i = (y_{i1}, y_{i2}, ..., y_{it})$. The aim of the clustering is to divide S into g homogeneous sub-groups. Traditionally, k-means can be run using several distances. KmL can use the Euclidean distance $Dist(y_i, y_j) = \sqrt{\frac{1}{t}\sum_{k=1}^{t}(y_{ik} - y_{jk})^2}$ and the Manathan distance $Dist_M(y_i, y_j) = \frac{1}{t}\sum_{k=1}^{t}|y_{ik} - y_{jk}|$ (more robust towards outliers [30]).

2.2 Choosing an optimal number of clusters

To chose the optimal number of clusters, KmL uses the Calinski and Harabasz criterion $C(g)$ [40]. It has interesting properties, as shown by several authors [16, 17]. Let n_m be the number of trajectories in cluster m; $\overline{y_m}$ the mean trajectory of cluster m; $\overline{y}$ the mean trajectory of the whole set S and v' denotes the transposition of vector v. Then the between-variance matrix is $\mathbf{B} = \sum_{m=1}^{g} n_m(\overline{y_m} - \overline{y})(\overline{y_m} - \overline{y})'$; the trace of the between-variance is the sum of its diagonal coefficients. High between-variance denotes well separated clusters, low between-variance means groups close to each other. The within-variance is $\mathbf{W} = \sum_{m=1}^{g}\sum_{k=1}^{n_m}(y_{mk} - \overline{y_m})(y_{mk} - y_m)'$. Low within-variance denotes compact groups, high within-variance denotes heterogeneous groups (more details on between and within variance in [9]). The Calinski and Harabazt criterion combines the within and between matrices to evaluate clustering quality. The optimal number of clusters corresponds to the value of g that maximizes $C(g) = \frac{Trace(B)}{Trace(W)} \cdot \frac{n-g}{g-1}$ where $\mathbf{B}$ is the between-matrix and $\mathbf{W}$ the within-matrix.

2.3 Avoiding local maxima

One major weakness of hill-climbing algorithms is that they may converge to a local maximum that does not correspond to the best possible clustering in terms of homogeneity. To overcome this problem, different solutions have been proposed. [41, 42] suggest choosing the initial clusters. [35] run a “wavelet” k-means process, modifying the result of a computation and using it as the starting point for the next computation. [15, 43] suggest running the algorithm several times, and retaining the best solution. It is this approach that has been chosen here. As for the cluster number, the “best” solution is the one

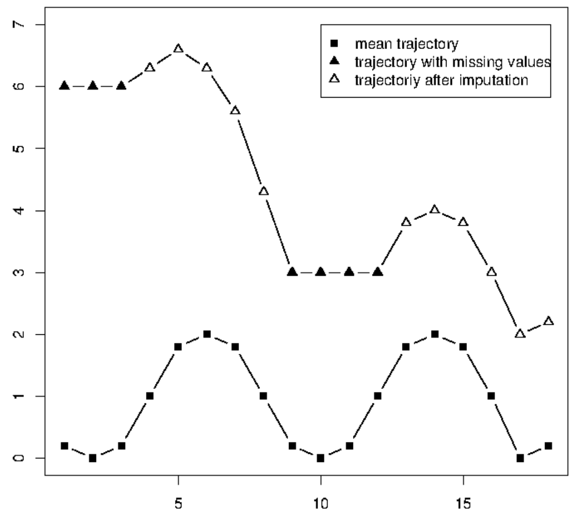


Figure 1: Example of *mean shape copying* imputation.

that maximizes the between-matrix variance and minimizes the within-variance. Once more, we use the Calinski and Harabatz criterion.

2.4 Dealing with missing value

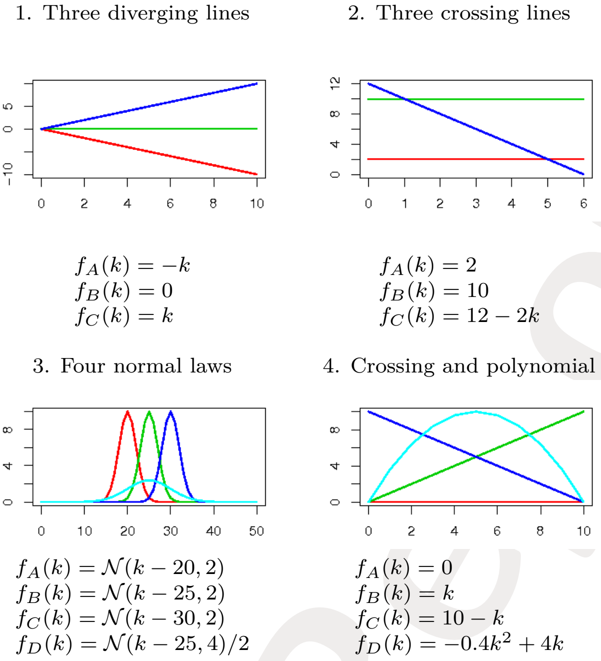
There are very few studies that try to cluster data assuming missing values [44]. The simplest way to handle missing data is to exclude trajectories for which certain data are missing. This can severely reduce the sample size, and longitudinal data are especially concerned and subject to missing values (missing values are more likely when an individual is asked to complete certain variables every week than when subjects are asked to complete data only once). In addition, having missing values can be a characteristic that defines a particular cluster, for example an "early drop-out" group.

A different approach has been used here. There is a need to deal with missing data at two different stages. First, during clustering, it is necessary to calculate the distance between two trajectories. Instead of using classic distances as defined in section 2.1, we use distances with Gower adjustment [45]: Given y_i and y_j, let w_{ijk} be 0 if y_{ik} or y_{jk} or both are missing, and 1 otherwise; the Euclidian distance with Gower adjustment between y_i and y_j is $Dist_{Gower}(y_i, y_j) = \sqrt{\frac{1}{\sum w_{ijk}} \sum_{k=1}^{t} (y_{ik} - y_{jk})^2 . w_{ijk}}$.

The second problematic step is the calculation of $C(g)$ which helps in the determination of the optimal clustering. At this stage, missing values need to be imputed. We use the following rules (called *mean shape copying*): if y_{ik} is missing, let y_{ia} and y_{ib} be the closest preceding and following non-missing values of y_{ik}; let $\overline{y_m} = (\overline{y_{m1}}, ..., \overline{y_{mt}})$ denote the mean trajectory of y_i cluster. Then $y_{ik} = y_{ia} + (\overline{y_{mk}} - \overline{y_{ma}}) \times \frac{y_{ib} - y_{ia}}{\overline{y_{mb}} - \overline{y_{ma}}}$. If first values are missing, let y_{ib} be the first non-missing value. Then $y_{ik} = y_{ib} + (\overline{y_{mk}} - \overline{y_{mb}})$. If last values are missing, let y_{ia} be the last non-missing value. Then $y_{ik} = y_{ia} + (\overline{y_{mk}} - \overline{y_{ma}})$. Figure 1 gives an example of mean shape copying imputation.

2.5 Implementation of the package

The k-means algorithm used is the Lloyd version [38]. Most of KmL code is written in R using S4 objects [46]. The critical part of the programme, clustering, is implemented in two different ways. The first, written in R, provides several options: it can display a graphical representation of the cluster during the convergence of the algorithm; it also lets the user define a distance function that KmL can use to cluster the data. The second, in C (compiled), does not offer any option but is optimized: the C procedure is around 20 times faster than the R procedure. Note that the user does not have to choose between the two functions: KmL automatically selects the fast one when possible, otherwise the slow one.



1. Three diverging lines

$f_A(k) = -k$
$f_B(k) = 0$
$f_C(k) = k$

2. Three crossing lines

$f_A(k) = 2$
$f_B(k) = 10$
$f_C(k) = 12 - 2k$

3. Four normal laws

$f_A(k) = \mathcal{N}(k - 20, 2)$
$f_B(k) = \mathcal{N}(k - 25, 2)$
$f_C(k) = \mathcal{N}(k - 30, 2)$
$f_D(k) = \mathcal{N}(k - 25, 4)/2$

4. Crossing and polynomial

$f_A(k) = 0$
$f_B(k) = k$
$f_C(k) = 10 - k$
$f_D(k) = -0.4k^2 + 4k$

Figure 2: Trajectory shapes.

3 Simulations and applications to real data

3.1 Construction of artificial data sets

To compare the efficiency of Proc Traj and KmL, simulated data were used. We worked on 5600 data sets defined as follow: a data set is the mixture of several sub-groups. A subgroup m is defined by a function $f_m(k)$ called the *theoretical trajectory*. Each subject i of a sub-group follows the theoretical trajectory of its subgroup plus a personal variation $\epsilon_i(k)$. The mixture of the different theoretical trajectories is called the *data set shape*. The 5600 data sets were formed varying the data set shape, the number of subjects in each cluster and the personal variations. We defined four data set shapes (presented figure 2).

1. "Three diverging lines" is defined by $f_A(k) = -k$; $f_B(k) = 0$; $f_C(k) = k$ with k in $[0 : 10]$.

2. "Three crossing lines" is defined by $f_A(k) = 2$; $f_B(k) = 10$; $f_C(k) = 12 - 2k$ with k in $[0 : 6]$.

3. "Four normal laws" is defined by $f_A(k) = \mathcal{N}(k-20, 2)$; $f_B(k) = \mathcal{N}(k-25, 2)$; $f_C(k) = \mathcal{N}(k-30, 2)$; $f_D(k) = \mathcal{N}(k - 25, 4)/2$ with k in $[0 : 50]$ and $\mathcal{N}(m, \sigma)$ denote the normal law with a mean of m and a standard deviation of σ.

4. "Crossing and polynomial" is defined by $f_A(k) = 0$; $f_B(k) = k$; $f_C(k) = 10 - k$; $f_D(k) = -0.4k^2 + 4k$ with k in $[0 : 10]$.

They were chosen either to correspond to three clearly identifiable clusters (set 1), to present a complex structure (every trajectory intersecting all the others, set 4) or to copy real data ([47] and data presented in section 3.3, sets 2 and 3). Personal variations $\epsilon_i(k)$ are randomised and follow the normal law $\mathcal{N}(0, \sigma)$. Standard deviations increase from $\sigma = 1$ to $\sigma = 8$ (by steps of 0.01). Since the distance between two theoretical trajectories is around 10, $\sigma = 1$ provides "easily identifiable and distinct clusters" whereas $\sigma = 8$ gives "markedly overlapping groups". The number of subjects in each cluster is set at either 50 or 200. Overall, 4 (data set shape) x 700 (variance) x 2 (number of subjects) = 5600 data sets were created. In a specific data set, the trajectories y_{ik} of an individual belonging to group g is defined by $y_{ik} = fg(k) + \epsilon_i(k)$, with $\epsilon_i(k)\ \mathcal{N}(0, \sigma^2)$. For the analyses using Proc Traj and KmL, the appropriate number of groups was entered. In addition, the analyses using Proc Traj required the degrees of polynomials that best fitted the trajectories.

3.2 Comparison of KmL and Proc Traj on artificial data sets

Evaluation of KmL and Proc Traj efficiency was performed by measuring two criteria on each clustering C that they found. Firstly, on the artificial data set, the real clustering R is known (the clusters in which

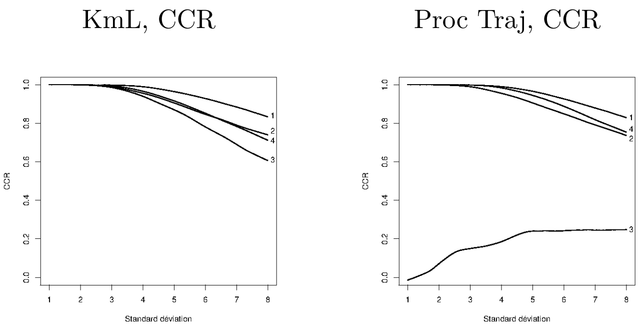


Figure 3: Comparison of *Correct Classification Rate* between KmL and Proc Traj.

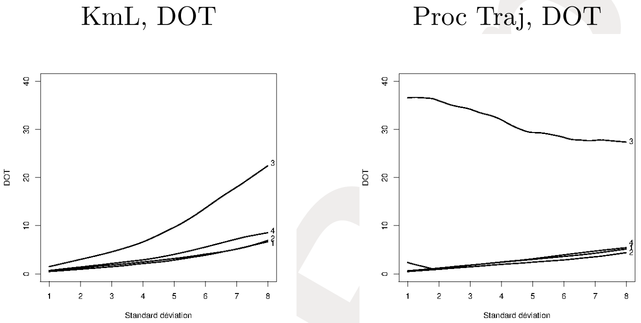


Figure 4: Comparison of *Distance Observed - Theoretical trajectories* between KmL and Proc Traj.

each subject should be). The *Correct Classification Rate* (CCR) is the percentage of trajectories that are in the same cluster in C and R [48], that is the percentage of subjects for whom an algorithm makes the right decision. Secondly, working on C, it is possible to evaluate the mean trajectory of each cluster (called the observed trajectory of a cluster). Observed trajectories are an estimation of the theoretical trajectory $f_A(k)$, $f_B(k)$, $f_C(k)$ and $f_D(k)$. An efficient algorithm will find observed trajectories close to the theoretical trajectories. Thus the second criterion, DOT, is the average *Distance between Observed and Theoretical trajectories*. Figures 3 and 4 present the results of the simulations. The graphs present the CCR (resp. the DOT) according to the standard deviation. Table 1 shows the average CCR (resp. the average DOT) for each data set shape.

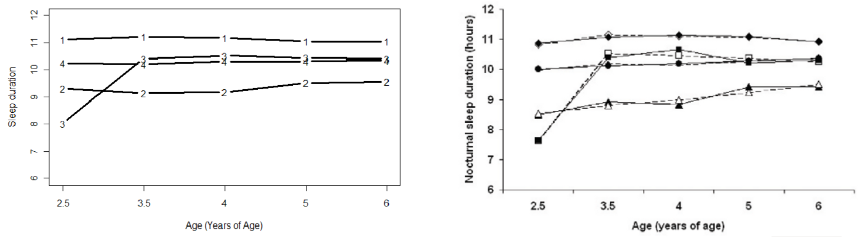
On dataset shape for 1, 2 and 4, KmL and Proc Traj give very close results whether on CCR or on DOT. In example 3: “Four normal laws”, Proc Traj does not converge, or finds results very far removed from the real clusters. KmL performances are as relevant as those obtained on examples 1, 2 and 4.

3.3 Application to real data

The first real example is derived from [49]. This study was conducted as part of the Quebec Longitudinal Study of Child Development (Canada) initiated by the Quebec Institute of Statistics. The aim of the study was to investigate the associations between longitudinal sleep duration patterns and behavioral/cognitive functioning at school entry. 1492 families participated in the study until the children were 6 years old. Nocturnal sleep duration was measured at 2.5, 3.5, 4, 5, and 6 years of age by an open question on the Self-Administered Questionnaire for the Mother (SAQM). In the original article, a semiparametric model was used to identify subgroups of children who followed different developmental trajectories. They

Average CCR				Average DOT		
Data set	KmL	Proc Traj		Data set	KmL	Proc Traj
1	0.95	0.95		1	3.17	3.02
2	0.91	0.91		2	3.04	2.48
3	0.86	0.20		3	9.66	34.28
4	0.91	0.91		4	4.24	3.79

Table 1: Comparison of average *DOT* and average *CCR* between KmL and Proc Traj.



Trajectories found by KmL

Trajectories found by Proc Traj

Figure 5: Sleep duration, means trajectories found by KmL and Proc Traj

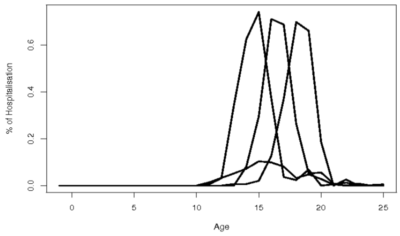


Figure 6: Hospitalisation length, mean trajectories found by KmL

obtained 4 sleep duration patterns, as illustrated in Figure 5: a persistent short pattern composed of children sleeping less than 10 hours per night until age 6; a increasing short pattern composed of children who slept fewer hours in early childhood but whose sleep duration increased around 41 months of age, a 10-hour persistent pattern composed of children who slept persistently approximately 10 hours per night; and an 11-hour persistent pattern composed of children who slept persistently around 11 hours per night.

On this data, KmL finds an optimal solution for a partition into four clusters (as does PROC TRAJ). The trajectories found by both methods are very close (see figure 5). The average distance between observed trajectories found by Proc Traj and by KmL is 0.31, which is rather small considering the range of the data (0;12).

The second real example is from a study on the *Trajectories of adolescents hospitalized for Anorexia Nervosa and their social integration in adulthood*, by Hubert, Genolini and Godart (submitted). This study is being conducted at the Institut Mutualiste Montsouris. The authors investigate the relation between adolescent hospitalization for anorexia and their social integration in adulthood. 311 anorexic subjects were included in the study. They were followed from age 0 to 26. The outcome considered here is the annual hospitalisation length, as a percentage. KmL found an optimal solution for a partition into four clusters. The trajectories found by KmL are shown in figure 6. Depending on the number of clusters specified in the program, Proc Traj either stated a “false convergence” or gave incoherent results.

4 Discussion

In this article, we present KmL, a new package implementing k-means. The advantage of KmL over the existing procedures (“cluster”, “clusterSim”, “flexclust” or “mclust”) is that it is designed to work specifically on longitudinal data. It provides scope for dealing with missing values; it runs the algorithm several times, varying the starting conditions and/or the number of clusters sought; its graphical interface helps the user to choose the appropriate number of clusters when the classic criterion is not efficient. We also present simulations, and we compare k-means to the latent class model Proc Traj. According to simulations and analysis of real data, k-means seems as efficient as the existing parametric algorithm on polynomial data, and potentially more efficient on non-polynomial data.

4.1 Limitations

The limitations of KmL are inherent in all clustering algorithms. These techniques are mainly exploratory, they cannot statistically test the reality of cluster existence. Moreover, the determination of the optimal cluster number is still an unsettled issue and EM-algorithms can be particularly sensitive to the problem of the local maximum. KmL attempts to deal with these two points by iterating an optimisation process with different initial seeds. Finally, KmL is not model-based, which can be an advantage (non-parametric, more flexible) but also a disadvantage (no scope for testing goodness of fit).

4.2 Advantages

KmL presents some improvement compared to the existing procedures. Since it is a non-parametric algorithm, it does not need any prior information and consequently avoids the issues related to model selection, a frequent concern reported with existing model-based procedures ([27] page 65). KmL enables the clustering of trajectories that do not follow polynomial trajectories. Thus, it can deal with a larger set of data (such as Hubert's hospitalization time in anorexics which follows a normal distribution).

The simulations have shown overall that KmL (like Proc Traj) gives acceptable results for all polynomial examples, even with high levels of noise. A major interest of KmL is that it can work in conjunction with Proc Traj. Finding the number of clusters and the shape of the trajectories (the degree of the polynomial) is still a long and difficult task for Proc Traj users. Running KmL first can give information on both these parameters. In addition, even if Proc Traj has already proved to be an efficient tool in many situations, there is a need to confirm the results, which are mainly of an exploratory nature. When the two algorithms yield similar results, it reinforces confidence in the results.

4.3 Perspectives

A number of unsolved problems need investigation. The optimization of cluster number is a long-standing and important question. Perhaps the particular situation of univariate longitudinal data could yield an efficient solution not yet found in the general context of cluster analysis.

Another interesting point is the generalisation of KmL to problems of higher dimension. At this time, KmL deals only with longitudinal trajectories for a single variable. It would be interesting to develop it for multidimensional trajectories, considering several facets of a patient jointly.

As a last perspective, present algorithms agglomerate trajectories with similar global shape. Thus two trajectories that may be identical in a time translation (one starting early, the other starting late but with the same evolution) will be allocated to two different clusters. One may however consider that the "starting time" is not really important and that the local shape (the evolution of the trajectory) should be given more emphasis than the overall shape. In this perspective, two individuals with the same development, one starting early and one starting later, would be considered as belonging to the same cluster.

Conflict of interest statement None

Acknowledgements Thanks to Evelyne Touchette, Tamara Hubert and Nathalie Godart for allowing us to use their data. Thanks to Lionel Riou França, Laurent Orsi and Evelyne Touchette for their helpful advices in programming.

References

[1] C. Genolini and B. Falissard, "Kml: k-means for longitudinal data," *Computational Statistics*, vol. 25, no. 2, pp. 317–328, 2010.

[2] T. Tarpey and K. Kinateder, "Clustering functional data," *Journal of classification*, vol. 20, no. 1, pp. 93–114, 2003.

[3] F. Rossi, B. Conan-Guez, and A. E. Golli, "Clustering functional data with the SOM algorithm," in *Proceedings of ESANN*, pp. 305–312, 2004.

[4] C. Abraham, P. Cornillon, E. Matzner-Lober, and N. Molinari, "Unsupervised Curve Clustering using B-Splines," *Scandinavian Journal of Statistics*, vol. 30, no. 3, pp. 581–595, 2003.

[5] G. James and C. Sugar, "Clustering for Sparsely Sampled Functional Data," *Journal of the American Statistical Association*, vol. 98, no. 462, pp. 397–408, 2003.

[6] J. Boik, R. Newman, and R. Boik, "Quantifying synergism/antagonism using nonlinear mixed-effects modeling: A simulation study.," *Statistics in Medicine*, 2007.

[7] N. Atienza, J. García-Heras, J. Muñoz-Pichardo, and R. Villa, "An application of mixture distributions in modelization of length of hospital stay.," *Statistics in Medicine*, 2007.

[8] H. Goldstein, *Multilevel Statistical Models.* London: Edwar Arnold, 2nd ed., 1995.

[9] B. S. Everitt, S. Landau, and M. Leese, *Cluster Analysis.* A Hodder Arnold Publication, 4th ed., 6 2001.

[10] L. Ryan, "Combining data from multiple sources, with applications to environmental risk assessment.," *Statistics in Medicine*, vol. 27, no. 5, pp. 698–710, 2008.

[11] J. Magidson and J. K. Vermunt, "Latent class models for clustering: A comparison with K-means," *Canadian Journal of Marketing Research*, vol. 20, p. 37, 2002.

[12] H. Akaike, "A new look at the statistical model identification," *Automatic Control, IEEE Transactions on*, vol. 19, no. 6, pp. 716–723, 1974.

[13] J. Bezdek and N. Pal, "Some new indexes of cluster validity," *Systems, Man and Cybernetics, Part B, IEEE Transactions on*, vol. 28, no. 3, pp. 301–315, 1998.

[14] G. Schwarz, "Estimating the Dimension of a Model," *The Annals of Statistics*, vol. 6, no. 2, pp. 461–464, 1978.

[15] C. Sugar and G. James, "Finding the Number of Clusters in a Dataset: An Information-Theoretic Approach.," *Journal of the American Statistical Association*, vol. 98, no. 463, pp. 750–764, 2003.

[16] G. W. Milligan and M. C. Cooper, "An examination of procedures for determining the number of clusters in a data set," *Psychometrika*, vol. 50, no. 2, pp. 159–179, 1985.

[17] Y. Shim, J. Chung, and I. Choi, "A Comparison Study of Cluster Validity Indices Using a Nonhierarchical Clustering Algorithm," in *Proceedings of CIMCA-IAWTIC'05-Volume 01*, pp. 199–204, IEEE Computer Society Washington, DC, USA, 2005.

[18] U. Maulik and S. Bandyopadhyay, "Performance Evaluation of Some Clustering Algorithms and Validity Indices," *IEEE Transactions on Pattern Analysis and Machine Iintelligence*, pp. 1650–1654, 2002.

[19] K. Košmelj and V. Batagelj, "Cross-sectional approach for clustering time varying data," *Journal of Classification*, vol. 7, no. 1, pp. 99–109, 1990.

[20] T. Warren-Liao, "Clustering of time series data-a survey," *Pattern Recognition*, vol. 38, no. 11, pp. 1857–1874, 2005.

[21] B. L. Jones, D. S. Nagin, and K. Roeder, "A SAS Procedure Based on Mixture Models for Estimating Developmental Trajectories," *Sociological Methods & Research*, vol. 29, no. 3, p. 374, 2001.

[22] B. L. Jones and D. S. Nagin, "Advances in Group-Based Trajectory Modeling and an SAS Procedure for Estimating Them," *Sociological Methods & Research*, vol. 35, no. 4, p. 542, 2007.

[23] D. S. Nagin and R. E. Tremblay, "Analyzing developmental trajectories of distinct but related behaviors: A group-based method," *Psychological methods*, vol. 6, no. 1, pp. 18–34, 2001.

[24] B. L. Jones, "Proc traj." http://www.andrew.cmu.edu/user/bjones/, 2001.

[25] D. Clark, B. Jones, D. Wood, and J. Cornelius, "Substance use disorder trajectory classes: Diachronic integration of onset age, severity, and course," *Addictive Behaviors*, vol. 31, no. 6, pp. 995–1009, 2006.

[26] C. Conklin, K. Perkins, A. Sheidow, B. Jones, M. Levine, and M. Marcus, "The return to smoking: 1-year relapse trajectories among female smokers," *Nicotine & Tobacco Research*, vol. 7, no. 4, pp. 533–540, 2005.

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

[28] R Development Core Team, *R: A Language and Environment for Statistical Computing.* R Foundation for Statistical Computing, Vienna, Austria, 2009. ISBN 3-900051-07-0.

[29] L. Muthén and B. Muthén, "Mplus user's guide," *Los Angeles, CA: Muthén & Muthén*, vol. 2006, 1998.

[30] Kaufman and Rousseeuw, *Finding Groups in Data: An Introduction to Cluster Analysis.* Wiley, 1990.

[31] G. Celeux and G. Govaert, "A classification EM algorithm for clustering and two stochastic versions," *Computational Statistics and Data Analysis*, vol. 14, no. 3, pp. 315–332, 1992.

[32] S. Tokushige, H. Yadohisa, and K. Inada, "Crisp and fuzzy k-means clustering algorithms for multivariate functional data," *Computational Statistics*, vol. 22, no. 1, pp. 1–16, 2007.

[33] T. Tarpey, "Linear Transformations and the k-Means Clustering Algorithm: Applications to Clustering Curves," *The American statistician*, vol. 61, no. 1, p. 34, 2007.

[34] L. A. García-Escudero and A. Gordaliza, "A Proposal for Robust Curve Clustering," *Journal of Classification*, vol. 22, no. 2, pp. 185–201, 2005.

[35] M. Vlachos, J. Lin, E. Keogh, and D. Gunopulos, "A Wavelet-Based Anytime Algorithm for K-Means Clustering of Time Series," in *3rd SIAM International Conference on Data Mining. San Francisco, CA. May 1-3, 2003, Workshop on Clustering High Dimensionality Data and Its Applications*, 2003.

[36] P. D Urso, "Fuzzy C-Means Clustering Models for Multivariate Time-Varying Data: Different Approaches," *International Journal of Uncertainy Fuzziness and Knowledge Base Systems.*, vol. 12, no. 3, pp. 287–326, 2004.

[37] Y. Lu, S. Lu, F. Fotouhi, Y. Deng, and S. J. Brown, "Incremental genetic K-means algorithm and its application in gene expression data analysis," *BMC Bioinformatics*, vol. 5, 2004.

[38] S. Lloyd, "Least squares quantization in PCM," *IEEE Transactions on Information Theory*, vol. 28, no. 2, pp. 129–137, 1982.

[39] C. Genolini, "Kml." http://christophe.genolini.free.fr/kml/, 2008.

[40] T. Calinski and J. Harabasz, "A dendrite method for cluster analysis," *Communications in Statistics*, vol. 3, no. 1, pp. 1–27, 1974.

[41] J. Hartigan, *Clustering Algorithms.* John Wiley & Sons, Inc. New York, NY, USA, 1975.

[42] J. T.-L. Tou and R. C. Gonzalez, *Pattern recognition principles.* Addison-Wesley, 1974.

[43] D. Hand and W. Krzanowski, "Optimising k-means clustering results with standard software packages," *Computational Statistics and Data Analysis*, vol. 49, no. 4, pp. 969–973, 2005.

[44] L. Hunt and M. Jorgensen, "Mixture model clustering for mixed data with missing information," *Computational Statistics and Data Analysis*, vol. 41, no. 3-4, pp. 429–440, 2003.

[45] J. Gower, "A General Coefficient of Similarity and Some of Its Properties," *Biometrics*, vol. 27, no. 4, pp. 857–871, 1971.

[46] C. Genolini, *A (Not So) Short Introduction to S4*, 2009.

[47] R. E. Tremblay, *Prévenir la violence dès la petite enfance.* Odile Jacob, 2008.

[48] T. P. Beauchaine and R. J. Beauchaine, "A Comparison of Maximum Covariance and K-Means Cluster Analysis in Classifying Cases Into Known Taxon Groups," *Psychological Methods*, vol. 7, no. 2, pp. 245–261, 2002.

[49] E. Touchette, D. Petit, J. Seguin, M. Boivin, R. Tremblay, and J. Montplaisir, "Associations between sleep duration patterns and behavioral/cognitive functioning at school entry.," *Sleep*, vol. 30, no. 9, pp. 1213–9, 2007.