Article

SMMR
STATISTICAL METHODS IN MEDICAL RESEARCH

Statistical Methods in Medical Research
0(0) 1–18

DOI: 10.1177/0962280214549040
smm.sagepub.com
SAGE

An alternative classification to mixture modeling for longitudinal counts or binary measures

Fabien Subtil,[1,2,3,4] Olayidé Boussari,[1,2,3,4,5] Mathieu Bastard,[6] Jean-François Etard,[6,7] René Ecochard[1,2,3,4] and Christophe Génolini[8,9,10]

Abstract

Classifying patients according to longitudinal measures, or trajectory classification, has become frequent in clinical research. The k-means algorithm is increasingly used for this task in case of continuous variables with standard deviations that do not depend on the mean. One feature of count and binary data modeled by Poisson or logistic regression is that the variance depends on the mean; hence, the within-group variability changes from one group to another depending on the mean trajectory level. Mixture modeling could be used here for classification though its main purpose is to model the data. The results obtained may change according to the main objective. This article presents an extension of the k-means algorithm that takes into account the features of count and binary data by using the deviance as distance metric. This approach is justified by its analogy with the classification likelihood. Two applications are presented with binary and count data to show the differences between the classifications obtained with the usual Euclidean distance versus the deviance distance.

Keywords

Longitudinal data, k-means, cluster analysis, likelihood, binary data, count data

[1]Université de Lyon, Lyon, France
[2]Université Lyon 1, Villeurbanne, France
[3]CNRS, UMR5558, Laboratoire de Biométrie et Biologie Evolutive, Villeurbanne, France
[4]Hospices Civils de Lyon, Service de Biostatistique, Lyon, France
[5]International Chair in Mathematical Physics and Applications, Université d'Abomey-Calavi, Abomey-Calavi, Bénin
[6]Epicentre, Paris, France
[7]UMI 233 TransVIHMI, Institut de Recherche pour le Développement, Université Montpellier 1, Montpellier, France
[8]INSERM, UMR 1027, Research Unit on Perinatal Epidemiology and Childhood Disabilities, Adolescent Health, Toulouse, France
[9]Université Paul Sabatier, UMR 1027, Toulouse, France
[10]CeRSM (EA 2931), UFR STAPS, Université de Paris Ouest-Nanterre-La Défense, France

Corresponding author:
Fabien Subtil, Hospices Civils de Lyon, Service de Biostatistique, 162 avenue Lacassagne, F-69003 Lyon, France.
Email: fabien.subtil@chu-lyon.fr

1 Introduction

An ever-increasing number of studies in epidemiology and clinical research are now carried out on repeated or longitudinal measures. For example: (i) the detection of prostate cancer recurrence relies on prostate specific antigen (PSA) levels which are longitudinally monitored after the first radiotherapy;[1] (ii) the detection of osteoporosis by regular measures of bone mineral density; (iii) the analysis of the long-term immune recovery of HIV-infected and treated patients through recurrent CD4 cell counts;[2] (iv) the identification of profiles of juvenile delinquency by the use of teacher reports of physical aggression by pupils aged 6–15.

Longitudinal data are not specific to the medical field. For example, in zoology, a better understanding of differences between species can be obtained by comparing the changes in some parameters over the life of the animals (e.g. the height). All these examples rely on the analysis of time-dependent curves or trajectories. Examples of time-independent curves are mass spectra that represent peptide-ion expression activities in function of their mass-to-charge. In oncology, these curves are used to identify subtypes of tumors.

Two main reasons may motivate the analysis of longitudinal data. The first is the need for modeling these longitudinal data for prediction purposes or to assess the impact of various factors. The second (we focus on here) is an attempt to identify and describe patterns of change over time, which requires classifying the patients into groups with similar changes over time (or similar ''trajectories'').

Various methods dedicated to non-longitudinal data classification have been applied to trajectory classification; e.g. support vector machines algorithms.[3] Another example is the k-means algorithm,[4] that was used to classify trajectories using different distances or dissimilarity metrics between curves.[5] The k-means algorithm corresponds to a non-parametric classification method: it searches the classification of the data that minimizes a specific within-group distance metric. There is no restriction on the distance metric to be used: depending on the objectives of the classification and the nature of the data, some distance metrics may be more relevant than others; however, the Euclidean distance metric is the most frequently used. *kml*, an R package, is an implementation of k-means to longitudinal data.[6] It enables the users to explore and exploit several distance metrics such as Euclidean, Manhattan, or Fréchet metric.[7] (As a reminder, Fréchet distance between two trajectories may be seen as the minimum length of a leach that would separate a master from his dog, each travelling along separate trajectories at different speeds.) Graphical representations of the solutions are offered as well as different methods for choosing the optimal number of groups. K-means have been used in various medical contexts for trajectory classification, for example: (i) to identify metabolite trajectories after nutritional challenges and analyze the interactions with some obesity risk genotypes;[8] (ii) to identify groups of phenotypes of children with similar respiratory/allergic symptom trajectories from birth to four years old and analyze the associated risk factors;[9] (iii) to identify trajectory groups of inattention and hyperactivity to predict educational attainment in early adulthood.[10] There are also applications in psychology: the method was used to identify trajectories of conflict levels over 28 group-counseling sessions and analyze the degree of correspondence between a developmental group counseling theory and the conflict levels reported by group members.[11] One limitation of k-means with longitudinal data is that it requires measures in all individuals at all-time points.

Other methods, called sometimes functional data classification, have been developed specifically for trajectory classification. They rely, for example, on filtering the data before classification to reduce their dimension. The *fclust* R-package projects each curve onto a finite dimensional basis, such as a natural cubic spline basis, and then clusters the basis coefficients.[12] Functional principal component analysis is another way of filtering curves before performing the classification according

to the principal components (*Funclust* and *FunHDCC* R packages).[13,14] A detailed review on functional data classification can be found in a recent article by Jacques and Preda.[14]

However, the aforementioned methods are generally applied to continuous data. For example, with k-means, the usual Euclidean distance metric is well-suited for continuous measures—even if it is subject to assumptions detailed later in the article—but not for binary outcomes or count data, cases in which the variance depends on the mean. Regarding count data, and assuming a Poisson distribution, high count measures have high variances; hence, groups with globally high count values are naturally more heterogeneous than groups with low count values. This fact cannot be taken into account by common distance metrics. This problem matters not only for trajectory classification, but also for classification of individuals on the basis of several non-continuous variables measured at a single time point. For non-continuous data, either longitudinal or multi-dimensional, the classification is often performed by mixture modeling.[15–18] In the case of trajectory classification, each individual trajectory is modeled by a mixture of a finite number of polynomials or spline functions, the mixing proportions varying from one individual to another. Some methods assume that there is no intra-group heterogeneity (e.g. proc Traj of SAS),[19] whereas others allow for this heterogeneity.[20] However, the first objective of mixture modeling is not to classify data, but to model them; for example, for inference purposes. There is a conceptual difference: classification approaches, like k-means, try to classify individuals into groups that could really exist whereas the mixture modeling approaches try to model trajectories by a mixture of group trajectories, the group trajectories being only theoretical constructs defined to take into account the heterogeneity in the data. These two different concepts do not necessarily lead to the same results.[21,22] Methods for classification of non-continuous data are thus needed.

This article takes one advantage of mixture modeling methods (i.e. the use of the likelihood) to extend k-means algorithms to the cases of binary or count data with specific application to trajectory classification. The deviance, which is proportional to minus twice the log-likelihood, is used as a distance metric to take into account the features of such data. This approach is justified by its connection with the classification likelihood. The implementation of this approach to count and binary data is presented here, as well as the classification expectation maximization (CEM) algorithm used to optimize the process. The article presents two examples that prove the advantage of the deviance over the usual Euclidean distance metric: one example is about the observance of antiretroviral treatments for HIV and the other about mosquito counts for malaria control.

2 Theoretical background

The theoretical backgrounds about k-means and classification likelihood are presented within the context of trajectory clustering even though they are not restricted to this context.

2.1 K-means

K-means[4,23] is an unsupervised learning algorithm that classifies a dataset into a fixed and a priori defined number of groups, k. Let $\boldsymbol{y}_i$ denote the l measures of a variable over time for an individual i (y_{ij}, $j=1,\ldots,l$). Within the context of trajectory clustering, k-means will define k centroids, or k mean trajectories $\mathbf{x}_h$ (x_{hj}, $j=1,\ldots,l$; $h=1,\ldots,k$) and a partition $P=(P_1,\ldots,P_k)$ of the n individuals into k groups so that $\sum_{h=1}^{k}\sum_{\mathbf{y}_i\in P_h} \mathrm{d}\left(\mathbf{y}_i,\mathbf{x}_{g_i}\right)$ is minimum. d is a distance metric; for example, the square function in the case of the Euclidean distance metric. K-means provides an

optimal partition in terms of mean distance between each individual trajectory and the mean trajectory of the group to which the individual belongs. Each mean trajectory is characterized by l parameters x_{hj}, $j=1,\ldots,l$, which correspond to the means of the measures of the hth group at various time points.

The algorithm starts from an initial random partition (several possibilities are available, see Genolini et al.[6]), calculates the mean trajectory per group, and updates the partition according to the distance between each individual trajectory and each mean trajectory. This process is repeated until there are no more changes in the partition. This is the most frequently used process; however, more efficient k-mean versions have been also proposed.[24]

2.2 Mixture likelihood

In mixture modeling, instead of assigning each trajectory to a group, each trajectory is modeled by a mixture of the mean trajectories associated with the groups.[21] The likelihood of an individual trajectory is then:

$$L(\mathbf{y}_i) = \sum_{h=1}^{k} \pi_h f(\mathbf{y}_i, \boldsymbol{\theta}_h)$$

$\boldsymbol{\theta}_h$ is the set of parameters that characterizes the hth mean trajectory—for example, parameters of polynomials or parameters of covariates that affect the mean trajectory—and $f(\mathbf{y}_i, \boldsymbol{\theta}_h)$ is the likelihood of trajectory $\mathbf{y}_i$ under the hypothesis of belonging to the hth group. The likelihood reflects the natural dispersion of individual trajectories around a mean trajectory. With continuous data, it is the multivariate Gaussian distribution that is often chosen. The π_h values are mixing proportions; i.e. π_1 represents the general influence of the first mean trajectory on each individual trajectory. Parameters $(\boldsymbol{\theta}_h, \pi_h)$ are estimated by maximizing the likelihood (called mixture likelihood) or the log-likelihood over all individuals:

$$\ln L(\mathbf{y}) = \sum_{i=1}^{n} \ln\left(\sum_{h=1}^{k} \pi_h f(\mathbf{y}_i, \boldsymbol{\theta}_h)\right) \tag{1}$$

In this formula, an individual contributes to the likelihood of each of the groups. In return, the predicted trajectory for an individual is a mixture of the mean trajectories of the groups.

In the calculation of f, Nagin[16] considers that, conditional to the group membership, the observations relative to each individual are independent. This assumption is similar to the one considered in the standard random effect model that assumes that observations relative to each individual are independent conditional on the individual's random effect. So, the assumption of conditional independence is not made at the same level. On the contrary, Muthén and Shedden[20] uses an additional random effect model within the groups to take into account the remaining serial dependence of the measures over time, resulting in what is called a "growth mixture model". The assumption of conditional independence in Nagin's method is stronger than in Muthén's method. However, the former has the advantage of not making the strong assumption that the random effects are independently and identically distributed according to the normal distribution.

2.3 Classification likelihood

In the classification likelihood approach, trajectories are assigned to groups.[25] Let $z_{ih} = 1$ when the ith trajectory belongs to the hth group and zero otherwise. The likelihood of an individual trajectory i belonging to the hth group is $L(\mathbf{y}_i) = f(\mathbf{y}_i, \boldsymbol{\theta}_h)$. Over all individuals, the log-classification-likelihood is given by:

$$\ln CL(\mathbf{y}) = \sum_{i=1}^{n} \sum_{h=1}^{k} z_{ih} \ln\left(f(\mathbf{y}_i, \boldsymbol{\theta}_h)\right) \tag{2}$$

In this formula, an individual contributes to the likelihood of only one group, group h to which he belongs ($z_{ij} = 0$). In return, the predicted trajectory for an individual is the mean trajectory of the group to which he belongs. Parameters ($\boldsymbol{\theta}_h$) as well as the indicator vectors $\mathbf{z}_1, \ldots, \mathbf{z}_n$ are estimated by maximizing the aforementioned log-classification-likelihood.

In equation (2), proportions π_h do not appear; they are implicitly supposed to be equal. This constraint tends to yield groups with similar sample sizes.[25] A more general version of classification likelihood has been proposed to relieve this constraint:[25]

$$\ln CL(\mathbf{y}) = \sum_{i=1}^{n} \sum_{h=1}^{k} z_{ih}\left(\ln(\pi_h) + \ln\left(f(\mathbf{y}_i, \boldsymbol{\theta}_h)\right)\right)$$

Still, in equation (2), f can be calculated by assuming that the measures relative to each individual are independent conditional to the group membership or by adding random effects to achieve the assumption of conditional independence making thus the assumption that the random effects are independently and identically distributed according to the normal distribution. In the following sections, only the first approach will be used.

Within this context, the Euclidean distance being proportional to minus twice the log-likelihood, maximizing the classification likelihood for trajectory classification is equivalent to using k-means with Euclidean distance in case of continuous data, when: (1) a multivariate Gaussian distribution with an identity covariance matrix is used for the classification likelihood, (2) the proportions π_h are constrained to be equal for all groups.[26] By adding these proportions to the optimization function of the k-means algorithm, the second constraint is no more necessary. Thus, there is a clear connection between classification likelihood and k-means when the assumption of conditional independence is made at the level of the group. To generalize this statement, one has to recognize that the deviance is proportional to minus twice the log-likelihood. Hence, the use of the deviance as distance metric in k-means leads to results identical to those obtained by classification likelihood, the assumption of conditional independence being made at the level of the group.

One natural extension of k-means algorithms is the use of the deviance as distance metric, providing thus an alternative to mixture modeling in case of non-continuous data.

3 k-Means and the deviance distance metric for count and binary data

This section focuses on the theoretical benefit of using the deviance as distance metric in k-means for count and binary data as well as the specific implementation of the k-means algorithm within this context.

3.1 Features of count and binary data

Contrarily to Gaussian measures, binary measures have a variance that, under the classical binomial distribution assumption, depends on the mean. The variance of a binary event of mean θ being equal to $\theta(1-\theta)$, groups with proportions often close to 0.5 allow naturally more heterogeneity than groups with proportions constantly close to zero or to one. For count measures, assuming a Poisson distribution, the variance is equal to the mean θ. Hence, groups with high mean counts allow naturally more heterogeneity than groups with mean counts close to zero. This has consequences on trajectory clustering.[27]

Let us consider two trajectories and let d be the Euclidean distance between them. Even when d is high, these two trajectories may stem from the same group if the intra-group variance is high in this region. On the contrary, even when d is very low, the trajectories may not stem from the same group if the intra-group variance is very low in this region; e.g. the mean count values are close to 0. The natural heterogeneity, or intra-group variance, and its variation according to the mean should then be considered in the clustering process. This is not done when one uses k-means with the Euclidean distance metric because this metric gives the same weight to equal distances between trajectories whatever the mean value of the measures.

To tackle this issue, the deviance distance metric will be used in k-means for binary or count data trajectories. To be consistent with the k-means algorithm with continuous data, each mean trajectory will be characterized by l parameters θ_{hj}, one for each time point.

3.2 Deviance and binary or count measures

Binary data are classically supposed to have a binomial distribution. Hence, by using the binomial deviance as metric, the distance between the individual trajectory of the ith individual and the mean trajectory of the hth group is proportional to:

$$\sum_{j=1}^{l} \left(y_{ij} \ln \theta_{hj} + \left(1 - y_{ij}\right) \ln\left(1 - \theta_{hj}\right) \right)$$

where θ_{hj} is the proportion of events in group h at time j.

Now, let us consider the simple case of binary trajectories with only one measure per individual and a group mean value $\theta = \theta_1$. Assuming that an individual belongs to this group, the deviance of his measure $\theta = \theta_2$ can be calculated for different values of θ_1 and different Euclidean distances d between the trajectory of this individual and the mean trajectory. The results are shown in Figure 1a. When θ_1 is close to 0.5, the deviance is insensitive to the variations of d; i.e. the Euclidean distance does not greatly influence group membership whereas it varies greatly when θ_1 gets close to zero or to one. Hence, with binary data, minimizing the deviance provides trajectories with very different distances to the mean trajectory within the same group when the mean trajectory is close to 0.5 but favors small distances when the mean trajectory is close to zero or to one. This is in agreement with the sampling variability of binary data.

The standard distribution for count data is the Poisson distribution. In this case, by using the Poisson deviance, the distance between the individual trajectory of the ith individual and the mean trajectory of the hth group is proportional to:

$$\sum_{j=1}^{l} \left(y_{ij} \ln\left(\theta_{hj}\right) - \theta_{hj} - \ln\left(y_{ij}!\right) \right)$$

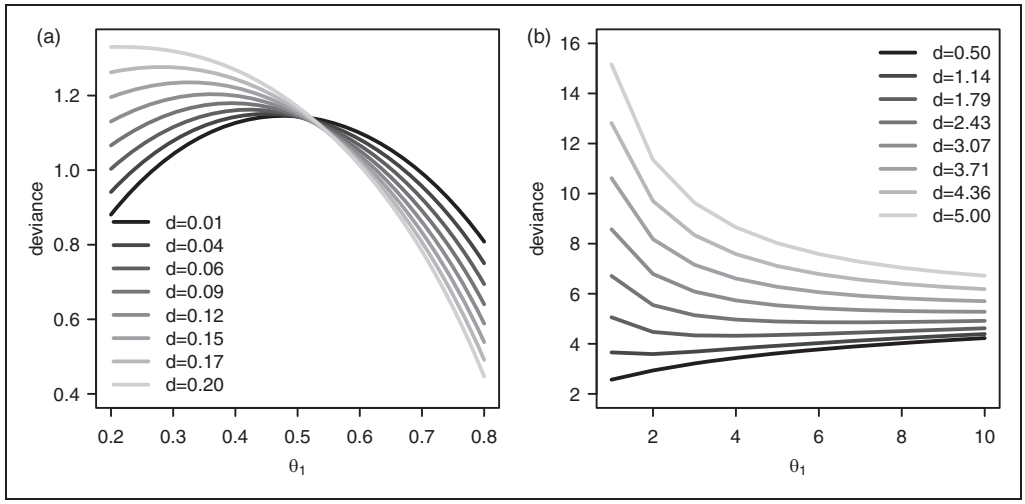

Figure 1. Deviance of measures assuming that they stem from a group centered on θ_1, for different Euclidean distances d between the measure and the mean of the group. (a) Binomial likelihood. (b) Poisson deviance.

where θ_{hj} is the mean count in group h at time j. Assuming that an individual belongs to the group with mean trajectory $\theta = \theta_1$, Figure 1b represents the deviance of his observation $\theta = \theta_2$ for different Euclidean distances d between this individual and the mean trajectory. The deviance is less sensitive to changes in d when the mean trajectory is high than when it is close to zero. Here again, the deviance distance performs the clustering in agreement with the intra-group variance of count data.

This justifies the use of the deviance as distance metric for k-means. For simplicity, "deviance distance metric" will be used herein even when proportions π_h are not constrained.

3.3 The CEM algorithm

The partition of the data and the parameter estimates that are solutions of the classification optimization problem are the ones that maximize the sum of the deviance, or maximize the log-likelihood, over all individuals. Maximizing the classification likelihood can be a difficult task due to the high dimension of the space to be explored (parameter space and all possible partitions of the individuals). The CEM is an EM-type algorithm that solves iteratively the problem by alternating between two phases: computing the estimates $(\boldsymbol{\theta}_h, \pi_h)$ given the partition and finding the partition of the individuals given the parameter estimates.[28] It adds a classification step to the EM algorithm, which is the difference with mixture modeling.

An initial partition is given to initialize the algorithm. Then, the mth iteration of the algorithm occurs as follows:

- E-step—for each individual, computes the posterior probability to belong to groups (p_{ih}^m) given the parameter estimates:

$$p_{ih}^m = \frac{\pi_h^{m-1} f\left(\mathbf{y}_i, \theta_h^{m-1}\right)}{\sum_{h'=1}^{k} \pi_{h'}^{m-1} f\left(\mathbf{y}_i, \theta_{h'}^{m-1}\right)}$$

- C-step—assigns each individual to the group with the maximum posterior probability.

- M-step—estimates the parameters by maximum likelihood in the partitions given by C-step.

With the family of exponential distributions, the maximum likelihood estimators correspond to moment estimators.[29] Hence, θ_{hj} is estimated by the mean of the observations of the hth group at time j whatever the type of data (continuous, binary, count). π_h is estimated by the mean of the z_{ih} values over all individuals. The algorithm stops when there is no more change in the groups between two successive iterations.

Celeux and Govaert[26] have shown that the CEM algorithm converges at a linear rate to a local maximum when the initial partition is in the neighborhood of this local maximum (and when the Hessian of the classification likelihood is negative at the local maximum). However, this does not guarantee the convergence to the global maximum and is one of several pitfalls shared by the k-means and the CEM algorithm. The result depends on the initial partition. One solution is to run the algorithm on several initial partitions and keep the best solution.[6] In the present work, the best solution is defined as the one that gives the highest classification likelihood value or the smallest sum of deviances.

Within the context of trajectory classification, the CEM algorithm is very similar to the classical k-means algorithm when the assumption of conditional independence is made at the level of the group and even equivalent when proportions π_h are constrained to be equal. The benefit of the deviance over a simple Euclidean distance metric is not obvious, especially with count data. The results obtained with the two distance metrics will be compared herein by simulation and on two examples.

4 Simulations

4.1 Design

Simulations were performed to assess the accuracy of the proposed deviance distance metric in terms of classification and compare it to the traditional Euclidean distance.

Data were simulated considering three groups at 15 time points, with binary or count outcomes. Two different simulation designs were used for binary data as well as for count data. The details on the way the data were simulated are given in Appendix 1. The mean profiles of the three groups are shown in Figure 2.

Eighty individual trajectories were simulated in each group. These were then classified using k-means, first with the deviance distance metric then with the Euclidean distance metric. Also, for each individual, the likelihood that he belongs to each group was calculated according to the true group parameters used to simulate the data. Each individual was then classified into the group with the highest likelihood. This classification was considered as the target classification; i.e. the best achievable given the way the data were generated. Indeed, even if a trajectory was simulated from one group, the obtained simulated trajectory may be closer to another group than to the true group due to heterogeneity. Hence, because of measure heterogeneity, even a perfect classification method would not achieve a 100% correct classification. This is why the percentages of correct classifications using k-means with the deviance or the Euclidean distance metric were compared with the percentage of correct classifications regarding the target classification. For each of the four simulation designs (2 for binary data and 2 for count data), 80 trajectories were simulated 50 times and the derived trajectories classified according to each of the three classification methods. Summary measures of accuracy were calculated per simulation design over the 50 repetitions.

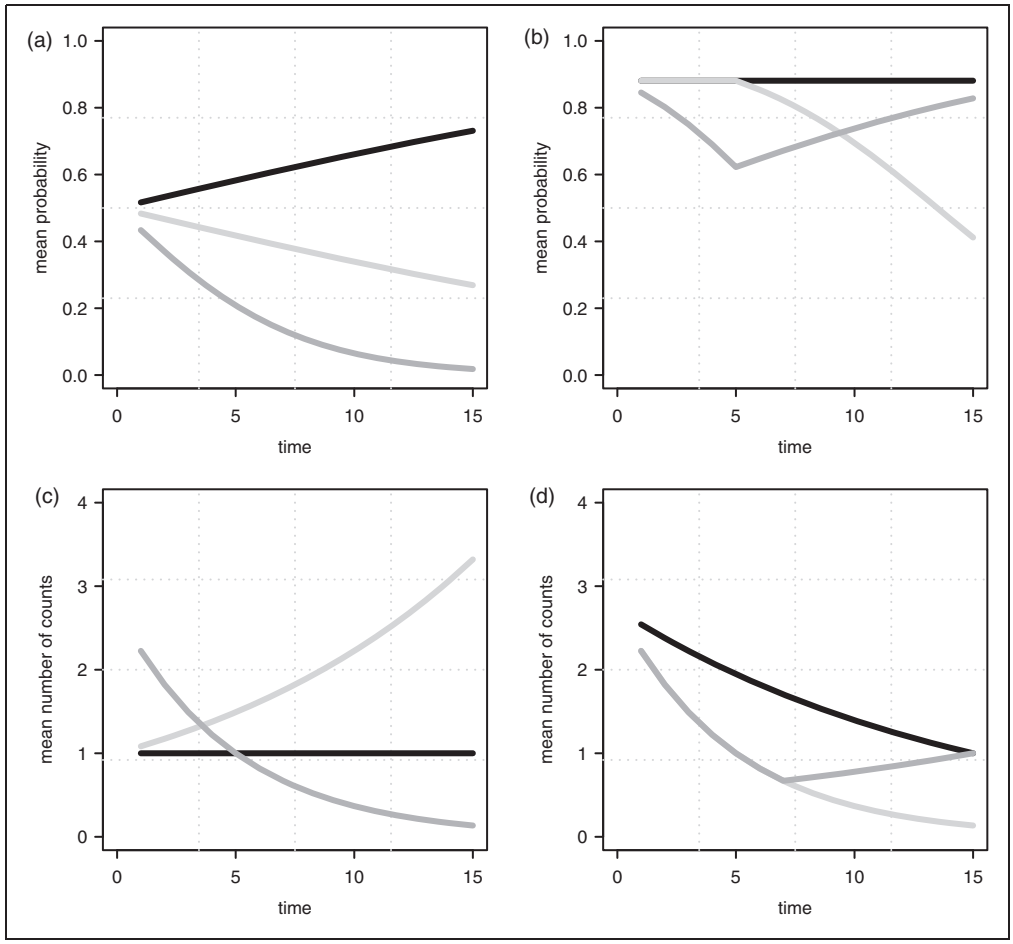

Figure 2. Mean trajectories used for the simulations. (a) Binary design 1. (b) Binary design 2. (c) Count design 1. (d) Count design 2.

4.2 Results

Table 1 presents, for each simulation design, summarized distributions of the differences in percentages of correct classifications between classification methods.

For binary data, k-means with the deviance distance metric provided slightly better percentages of correct classifications than k-means with the Euclidean distance metric (the differences between the percentages were more often greater than zero; i.e. in favor of the first method).

For count data, k-means with the deviance distance metric provided results close to the target classification; the differences between the percentages of correct classifications were at most 5%. The disagreement was higher between the target classification and the classification obtained with k-means and the Euclidean distance metric. Consequently, k-means with the deviance distance metric provided better classification results than k-means with the Euclidean distance metric with differences ranging from 3% to 10% according to the simulation design.

Table 1. Simulation results: summaries of the distributions of the differences in the percentages of correct classifications between methods.

Design and methods	1st quartile	Mean	3rd quartile
Binary design 1			
Deviance/Target	−0.133	−0.1164	−0.092
Euclidean/Target	−0.219	−0.178	−0.145
Deviance/Euclidean	0.021	0.062	0.116
Binary design 2			
Deviance/Target	−0.185	−0.144	−0.113
Euclidean/Target	−0.210	−0.171	−0.130
Deviance/Euclidean	−0.003	0.028	0.054
Count design 1			
Deviance/Target	−0.017	−0.007	0.000
Euclidean/Target	−0.054	−0.040	−0.026
Deviance/Euclidean	0.018	0.032	0.050
Count design 2			
Deviance/Target	−0.054	−0.031	−0.012
Euclidean/Target	−0.211	−0.177	−0.046
Deviance/Euclidean	0.108	0.146	0.180

5 Applications

5.1 Classification of African villages according to mosquito counts over the year

Malaria is still a major public health issue in sub-Saharan Africa. *Anopheles gambiae* (M and S forms) is one of the main malaria vectors in this region. The fight against malaria aims to decrease the transmission of *Plasmodium spp* parasites to humans by the mosquito vectors. A cluster-randomized controlled trial has compared the efficacies of various strategies of decreasing this transmission in 28 villages of Southern Benin.[30] Mosquitoes were collected by human landing catches every six weeks between January and December 2009 (eight surveys of two successive nights at four sites per village). The study did not show differences between the strategies but high variations in the density of malaria vectors were observed over time and space.

Figure 3a shows the counts of *A. gambiae* (M and S forms) in the 28 villages over all surveys. These counts were low at the beginning and at the end of the year, many villages having zero mosquitoes collected, but there was an explosion during the rainfall season (Survey 3 to Survey 5).

The 28 villages were clustered into two groups using k-means with the Euclidean and Poisson deviance distance metrics. The mean trajectories obtained with the two metrics as well as the percentage of villages in each group are shown in Figure 3c and d. Whatever the distance metric, there was one group of a few villages with a high increase in mosquito count during the rainfall season (group 1) and another group with a moderate increase (group 2). Two villages changed group with the change of distance metric (Figure 3b). These villages had a moderate increase in mosquito count during the rainfall season but small counts at the beginning of the year. Their trajectories were too distant from the mean trajectory of group 1 during the rainfall season; they were consequently classified in group 2 by the Euclidean distance. However, with the Poisson deviance, the variance was equal to the expected value. During the rainfall season, despite moderate increases in mosquito counts, both villages could switch to group 1 because of the increase in heterogeneity allowed by the

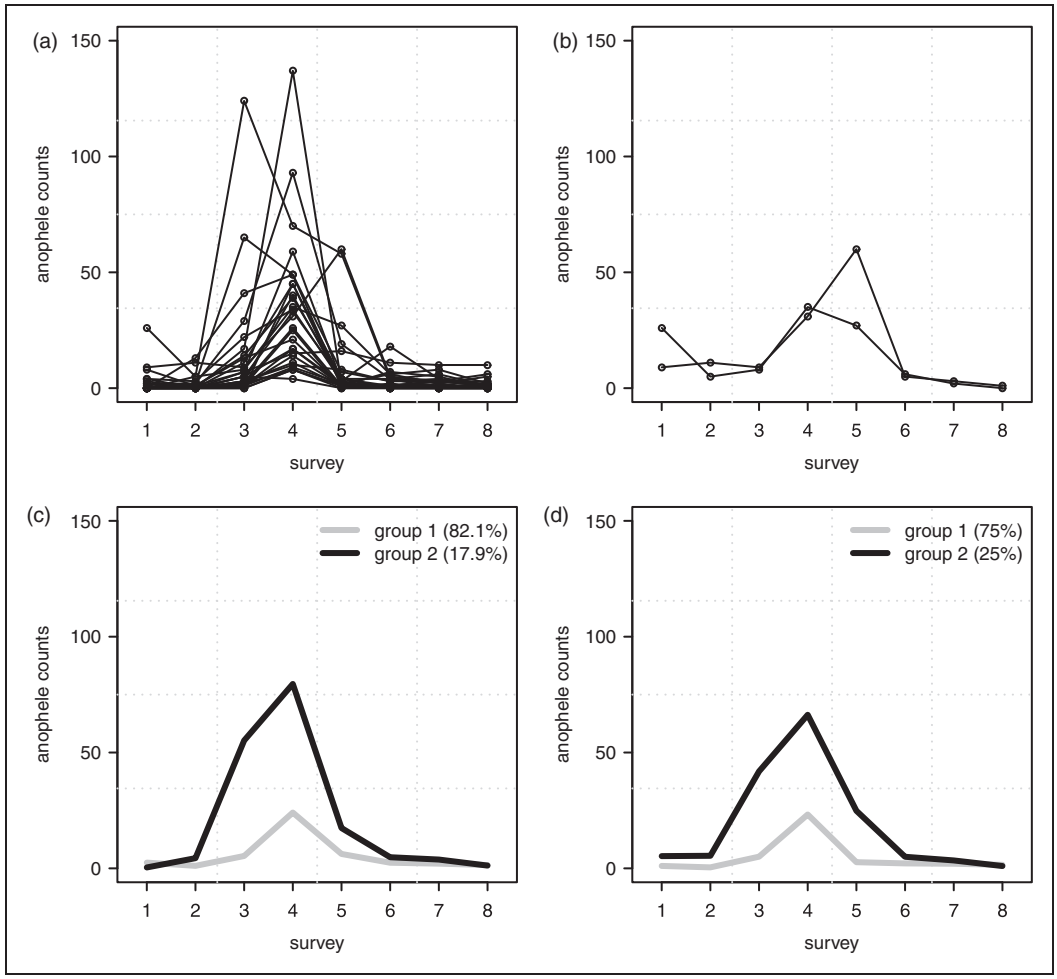

Figure 3. (a) *Anopheles* counts in the different villages over the surveys. (b) Villages that changed group according to the distance metric (Euclidean or Poisson deviance). (c) Mean trajectories using the Euclidean distance. (d) Mean trajectories using Poisson deviance.

Poisson deviance. At the beginning of the year, a very low heterogeneity was allowed in group 2 by the Poisson deviance due to a high number of villages with very low mosquito counts, which excluded the two villages from this group. This is why the villages switched from group 2 to group 1 with the use of the Poisson deviance.

The above results were obtained without constraining a priori the proportions of individuals in the groups to be equal; however, identical results were obtained with this constraint.

5.2 Classification of patients according to the adherence to the antiretroviral therapy

The ISAARV project was launched in Senegal to provide highly active antiretroviral therapy to HIV-positive patients.[31,32] In this project, 404 patients were followed at least every two months to

assess their adherence to the treatment and determine the causes of non-adherence. Here, adherence was defined as the ratio of the number of tablets considered as taken to the number of prescribed tablets. In this application, adherence was analyzed up to 91 months (the median follow-up of the cohort). Data were not recorded for 74 patients, mainly because of early death. Besides, 51 additional patients were discarded from the analyses because they have participated less than 10 assessments. Adherence was then averaged over a 3-month period. Missing values for a period were imputed using the copyMean method.[6,33] A "good adherence" during each period was defined as a mean adherence over 95%; this led to binary data.

The 279 patients were clustered into three groups using k-means with the Euclidean distance and the binomial deviance. The mean trajectories with the two metrics as well as the percentages of patients in each group are shown in Figure 4a and b. The results of the cross-classification using the two distance metrics are shown in Table 2. Overall, the mean trajectories seemed similar with the two distance metrics but the third mean trajectory was lower at the end of the follow-up with the binomial deviance.

Eighteen patients switched from group 3 to group 1. The mean trajectory of these 18 patients is shown in Figure 4c. Overall, their adherence was similar to that of group 1 or group 3 at the beginning of the follow-up but, after four years, it was about 0.5. Using the Euclidean distance, after four years, the third mean trajectory was the mean of some trajectories below 0.35 and some others above 0.35, this included the 18 trajectories. During the classification step, patients below 0.35 were considered to be as distant from the mean trajectory as patients above 0.35. However, using the binomial deviance, the variance of the observations below 0.35 was smaller than that above 0.35 and less heterogeneity was allowed below than above 0.35. Consequently, the third mean trajectory was attracted by patients below 0.35; this led to the switch of the 18 patients toward group 1.

The Euclidean distance gives the same weight to equal distances between trajectories whatever the levels of these trajectories whereas the binomial deviance takes into account the fact that groups of trajectories with probabilities close to zero or one are more stable than groups with probabilities close to 0.5.

Figure 5 shows the results obtained when the proportions of individuals in the groups were a priori supposed to be equal. These proportions had closer values between them than those shown in Figure 4, which was expected according to the theoretical backgrounds. In comparison with the analysis where the proportions were not a priori constrained to be equal, patients switched from group 1 to group 3, which led to a higher third mean trajectory after four years. There was less discrepancy between the Euclidean distance and the binomial deviance than in Figure 4. Hence, the a priori use of a constraint on the proportions of patients in the groups may affect the groups and the trajectories obtained.

The removal of 51 patients from the analysis due to insufficient number of measures (less than 10) may have influenced the shape of the mean trajectories or the proportions of the patients in each group. Hence, the aforementioned results may be different from those that would be obtained by an analysis performed on the whole cohort. The figures and proportions presented should be analyzed with caution.

6 Discussion

The present work intended to cluster count or binary measures using the deviance as a measure of similarity, with a specific application to trajectory classification. The deviance considers that the intra-group variance of such data changes naturally depending on the mean. The intra-group variance is an essential point in clustering because the distance between two trajectories has to be

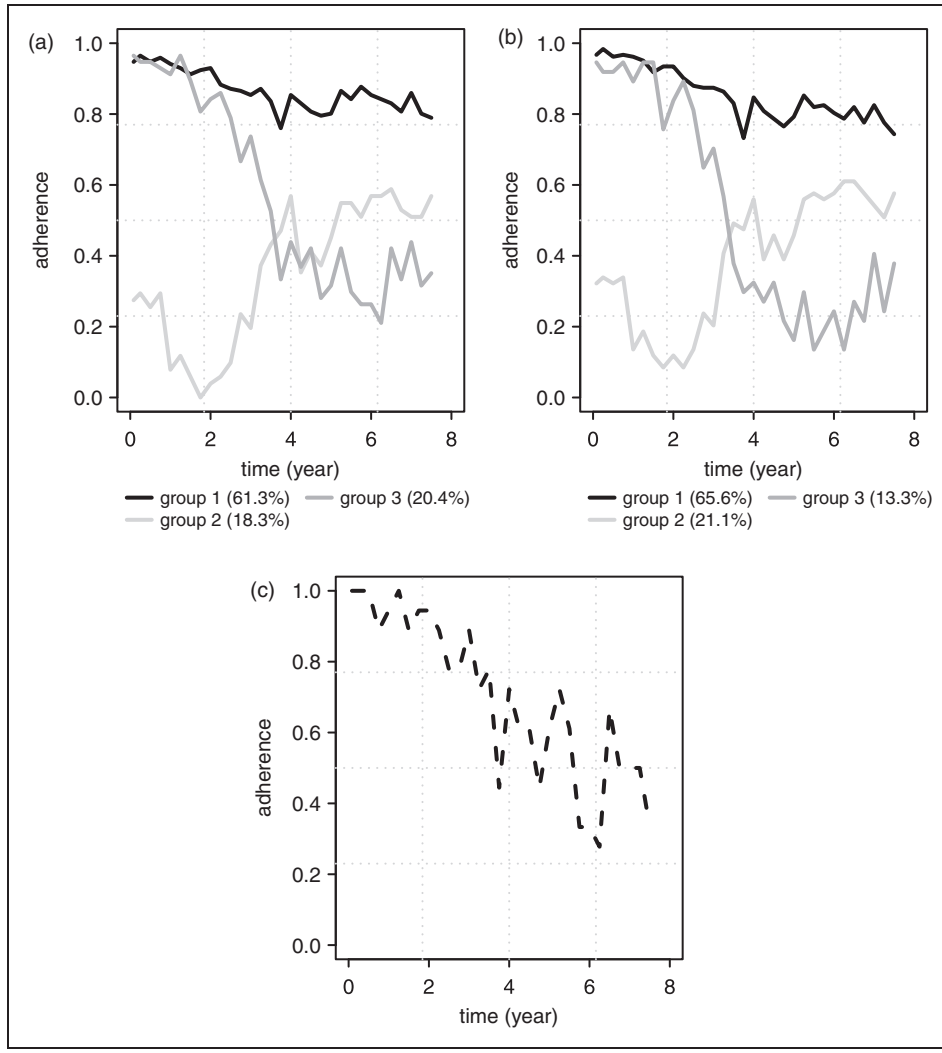

Figure 4. Mean trajectories of adherence to treatment over time. (a) Euclidean distance metric. (b) Binomial deviance metric. (c) Mean trajectory of patients that moved from group 3 to group 1.

Table 2. Cross-classification obtained using the Euclidean distance and the binomial deviance with data on adherence to antiretroviral therapy over time.

	Binomial deviance		
	Group 1	Group 2	Group 3
Euclidean distance			
Group 1	165	6	0
Group 2	0	51	0
Group 3	18	2	37

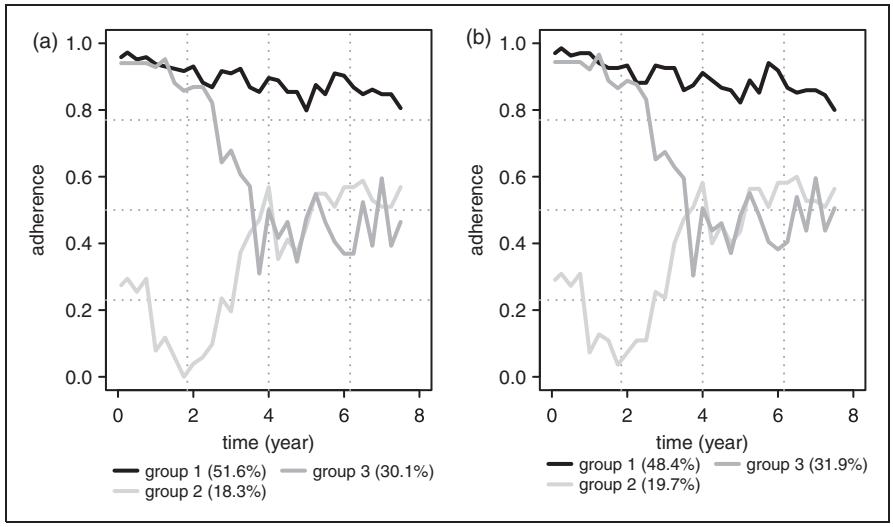

Figure 5. Mean trajectories of adherence to treatment over time when the proportions of patients in the groups are assumed a priori equal. (a) Euclidean distance. (b) Binomial deviance.

assessed with regard to the sampling variability. This issue is not often mentioned because clustering is often performed on measures that are considered Gaussian with constant intra-group variance. In this case, the k-means algorithm with the Euclidean distance may be appropriate to cluster the trajectories, under the assumption of conditional independence between measures within each group. However, caution should be taken because continuous measures do not always follow Gaussian distributions. A gamma or an exponential distribution may be more appropriate, in which case the variance is proportional to the square of the mean and the issue of intra-group variance is raised again. Hence, the use of the deviance as a distance metric may be useful in some cases of continuous data. In the two examples given here, the use of the deviance led to substantial differences in the groups obtained versus those obtained with the simple Euclidean distance; the differences were explained by the concept of intra-group variance.

With the family of exponential distributions, the CEM algorithm is equivalent to the k-means algorithm when the proportions of individuals within groups are a priori assumed to be equal. Thus, the present work about count and binary data can be extended to all generalized linear models, especially gamma and exponential models. The deviance distance metric presented in this article was implemented in the R package kmlCov.

Simulation results showed that, with count and binary data, the deviance distance metric tends to give better classification results than the Euclidean distance though the difference is only moderate. The variance of binary data can range from 0 to 0.25, whereas the variance of count data can range from 0 to infinity depending on the mean. This is why the difference in classification performance between the deviance distance metric and the Euclidean distance metric is greater with count data than with binary data. It is also agreed that, in some cases—depending on the mean trajectories, the number of individuals per group, or the intra-group variability—there would be no differences between the Euclidean and the deviance distance metrics.

The use of the deviance as a distance metric with k-means is justified by its analogy with the classification likelihood method. In fact, Govaert and Nadif[34] have already proposed to cluster binary data using the classification likelihood methodology; the present article extends their work

to the case of longitudinal data. However, longitudinal data present the problem of serial dependence between measures of the same individual. The validity of our approach in this context relies on the assumption of conditional independence between measures within each group. Hence, the classification of individuals into different subgroups is itself used to remove the serial dependence between the measures of the same individual. This hypothesis was already made by Nagin in mixture modeling of trajectory data. We recognize that, in some cases, the classification itself may not remove all the serial dependence between the measures over time. Various methods may be proposed to take into account the remaining serial dependence when the assumption of conditional independence is made at the group level. One method consists in allowing for the autocorrelation using ARMA autocorrelation functions. Another method consists in adding random effects inside each group; it was proposed by Muthén in mixture modeling. However, the benefit of taking into account this serial dependence has not been analyzed yet in terms of classification performance; this would be a useful work. Another useful work would be to compare, on longitudinal data, the results of classification models with those of mixture models in terms of classification. Regarding the constraint on the proportions of individuals in the groups, assuming a priori that these proportions are equal is often difficult to justify, but this is what is done by the k-means algorithm. In the adherence to treatment application, making this assumption led to groups with more even sizes than without making it. Hence, on the basis of the present work, a particular attention should be paid to this assumption. However, a slight extension of the optimization function in the k-means algorithm allowed relieving this hypothesis. Celeux and Govaert[21] argued that approaches assuming equal proportions are more parsimonious and less initial-position dependent; however, the unrestricted approach is preferred when the mixing proportions are extreme and the sample size is not very large.[34]

The choice of the number of groups was not the purpose of the present work and remains a difficult task. This choice should be guided by a good knowledge of the specific study context. Several summary indices have been proposed and their respective performances compared.[35,36] However, most of these indices were developed for mixture models whose major aim is not classification. The integrated classification likelihood-Bayesian information criterion (ICL-BIC),[37] (based on the completed likelihood) or the normalized entropy criterion[38] are classification-based information criteria, and might be used in this context. The choice of the number of groups should not rely entirely on statistical indices; there should be also a clinical rationale, especially in classification likelihood where the latent groups are not theoretical constructs but should reflect reality. Some articles or books give application guidance on this subject.[16,39,40]

The present article focuses on trajectory clustering. Other methods have been developed within the context of mixture modeling of trajectories, with a lot of program implementations with various features.[19,41–43] The two approaches are similar; however, the former maximizes the classification likelihood whereas the second maximizes the mixture likelihood. Comparing the results, it has been shown, with Gaussian data, that classification likelihood does not generally lead to the same parameter estimates and that the bias in parameter estimation persists asymptotically.[44] The distinction is mainly conceptual: mixture approaches aim to model the data whereas classification approaches aim to cluster the data.

The connection between k-means algorithms dedicated to trajectory clustering and classification likelihood will allow a lot of extensions. For the time being, *kml* is non-parametric over time and calculates the mean of each group at different time points. This imposes having measures for all patients at all fixed time points, which is exceptional. In the ISAARV study, this constraint led to the removal from the analysis of 51 patients, which may have influenced the mean trajectories or the proportions of patients in the groups. Some flexible functions may be used to model the biomarker

change over time; e.g. polynomials with parameters estimated during the M-step of the CEM algorithm by likelihood maximization. The use of parametric functions would allow dealing with different numbers of measures at different time points and with values missing completely at random or missing at random. Moreover, covariates may be added to the model and their effects estimated during the M-step. For example, in the adherence to HIV treatment application, the effect of sex could have been estimated.

In the field of mixture likelihood, a growing number of programs are developed to deal with more features: consideration of heteroscedasticity,[27] introduction of covariates that can change the trajectories or influence group membership, etc. The deviance distance metric presented here will be implemented in the kmlCov package. By putting k-means algorithms dedicated to trajectory clustering into a likelihood context, the present work prepares for the incorporation of such features.

Acknowledgments

The authors would like to thank Armel Djènontin and Vincent Corbel who supervised the data collection and the project used for the *Anopheles gambiae* example. They also thank the ANRS 1215 Study Group for the data about the adherence to the antiretroviral therapy. The authors are also grateful to Dr Jean Iwaz whose suggestions improved significantly the clarity of the final manuscript.

The French Ministère des Affaires Etrangères et Européennes (MAEE) supported project FSP/REFS N°2006-22 that generated the data about *Anopheles gambiae*.

References

1. Proust-Lima C, Séne M, Taylor JM, et al. Joint latent class models for longitudinal and time-to-event data: a review. *Stat Methods Med Res* 2014; **23**(1): 74–90.
2. Bastard M, Soulinphumy K, Phimmasone P, et al. Women experience a better long-term immune recovery and a better survival on HAART in Lao People's Democratic Republic. *BMC Infect Dis* 2013; **22**(13): 27.
3. Schmah T. Comparing classification methods for longitudinal fMRI studies. *Neurol Comput* 2010; **22**(11): 2729–2762.
4. MacQueen JB. Some methods for classification and analysis of multivariate observations. In: *Proceedings of 5-th Berkeley symposium on mathematical statistics and probability*, Berkeley, University of California Press, 1967.
5. Jacques J and Preda C. *Functional data clustering: a survey*. Lille: INRIA, 2013.
6. Genolini C and Falissard B. KmL: a package to cluster longitudinal data. *Comput Methods Prog Biomed* 2011; **104**: e112–e121.
7. Alt H and Godau M. Computing the Fréchet distance between two polygonal curves. *Int J Comput Geom Appl* 1995; **5**: 75–91.
8. Wahl S, Krug S, Then C, et al. Comparative analysis of plasma metabolomics response to metabolic challenge tests in healthy subjects and influence of the FTO obesity risk allele. *Metabolomics* 2014; **10**(3): 386–401.
9. Rancière F, Nikasinovic L, Bousquet J, et al. Onset and persistence of respiratory/allergic symptoms in preschoolers: new insights from the PARIS birth cohort. *Allergy* 2013; **68**(9): 1158–1167.
10. Pingault JB, Tremblay RE, Vitaro F, et al. Childhood trajectories of inattention and hyperactivity and prediction of educational attainment in early adulthood: a 16-year longitudinal population-based study. *Am J Psychiatr* 2011; **168**: 1164–1170.
11. Brossart DF, Parker R and Willson VL. A comparison of two methods for analyzing longitudinal data: tuckerized growth curves and an application of K means analysis. *Learn Individ Differ* 1998; **10**(2): 121–136.
12. James GM and Sugar CA. Clustering for sparsely sampled functional data. *J Am Stat Assoc* 2003; **98**(462): 397–408.
13. Bouveyron C, Girard S and Schmid C. High dimensional data clustering. *Comput Stat Data Anal* 2007; **52**: 502–519.
14. Jacques J and Preda C. Model-based clustering for multivariate functional data. *Comput Stat Data Anal* 2014; **71**: 92–106.
15. Chu M-KM and Koval JJ. Trajectory modelling of longitudinal binary data: application of the EM algorithm for mixture models. *Commun Stat Simul Comput* 2014; **43**(3): 495–519.
16. Nagin DS. *Group-based modeling of development*. London: Harvard University Press, 2005.
17. Pickles A and Croudace T. Latent mixture models for multivariate and longitudinal outcomes. *Stat Methods Med Res* 2009; **19**(3): 271–289.
18. Vrbik I and McNicholas PD. Parsimonious skew mixture models for model-based clustering and classification. *Comput Stat Data Anal* 2014; **71**: 196–210.
19. Jones BL and Nagin DS. Advances in group-based trajectory modeling and an SAS procedure for estimating them. *Sociol Methods Res* 2007; **35**: 542–571.
20. Muthén BO and Shedden K. Finite mixture modeling with mixture outcomes using the EM algorithm. *Biometrics* 1999; **55**(2): 463–469.
21. Celeux G and Govaert G. Comparison of the mixture and the classification maximum likelihood. *J Stat Comput Simul* 1993; **47**: 127–146.
22. Hathaway RJ. Another interpretation of the EM algorithm for mixture distributions. *Stat Probab Lett* 1986; **4**: 53–56.

23. Hartigan JA and Wong MA. Algorithm AS 136: a K-means clustering algorithm. *J R Stat Soc Ser C Appl Stat* 1979; **28**(1): 100–108.
24. Kanungo T, Mount DM, Netanyahu NS, et al. An efficient k-means clustering algorithm: analysis and implementation. *IEEE Trans Pattern Anal Mach Intell* 2002; **24**(7): 881–892.
25. Symons MJ. Clustering criteria and multivariate normal mixtures. *Biometrics* 1981; **37**: 35–43.
26. Celeux G and Govaert G. A classification EM algorithm for clustering and two stochastic versions. *Comput Stat Data Anal* 1992; **14**: 315–332.
27. Elsensohn MH, Klich A, Ecochard R, et al. A graphical method to assess distribution assumption in group-based trajectory models. *Stat Methods Med Res* Epub ahead of print 2013.
28. Celeux G and Govaert G. Clustering criteria for discrete data and latent class models. *J Classif* 1991; **8**: 157–176.
29. McCullagh P and Nelder JA. *Generalized linear models*, 2nd ed. Boca Raton: Chapman and Hall/CRC, 1989.
30. Corbel V, Akogbeto M, Damien GB, et al. Combination of malaria vector control interventions in pyrethroid resistance area in Benin: a cluster randomised controlled trial. *Lancet Infect Dis* 2012; **12**(8): 617–626.
31. Desclaux A, Lanièce I, Ndoye I, et al. *The Senegalese antiretroviral drug access initiative. An economic, social, behavioural and biomedical analysis*. Paris: ANRS, UNAIDS, WHO, 2004.
32. Bastard M, Fall MB, Lanièce I, et al. Revisiting long-term adherence to highly active antiretroviral therapy in Senegal using latent class analysis. *J Acquir Immune Defic Syndr* 2011; **57**(1): 55–61.
33. Genolini C, Ecochard R and Jacqmin-Gadda H. Copy mean: a new method to impute intermittent missing values in longitudinal studies. *Open J Stat* 2013; **3**: 26–40.
34. Govaert G and Nadif M. Comparison of the mixture and the classification maximum likelihood in cluster analysis with binary data. *Comput Stat Data Anal* 1996; **23**: 65–81.
35. Biernacki C and Govaert G. Using the classification likelihood to choose the number of clusters. *Comput Sci Stat* 1997; **29**(2): 451–457.
36. McLachlan G and Peel D. *Finite mixture models*. New York, NY: Wiley, 2000.
37. Biernacki C, Celeux G and Govaert G. *Assessing a mixture model for clustering with the integrated classification likelihood*. Grenoble: IS2 (INRIA Rhône Alpes), 1998.
38. Biernacki C and Govaert G. Choosing models in model-based clustering and discriminant analysis. *J Stat Comput Simul* 1999; **64**: 49–71.
39. Nagin DS and Odgers CL. Group-based trajectory modeling in clinical research. *Annu Rev Clin Psychol* 2010; **6**: 109–138.
40. Nylund KL, Asparouhov T and Muthén BO. Deciding on the number of classes in latent class analysis and growth mixture modeling: a Monte Carlo simulation study. *Struct Equ Modeling* 2007; **4**(4): 535–569.
41. Muthén LK Muthén BO Mplus User's Guide. 7th ed. Los Angeles: CA Muthén & Muthén, 2012.
42. Proust C and Jacqmin-Gadda H. Estimation of linear mixed models with a mixture of distribution for the random effects. *Comput Methods Prog Biomed* 2005; **78**(2): 165–173.
43. Proust-Lima C, Amieva H and Jacqmin-Gadda H. Analysis of multivariate mixed longitudinal data: a flexible latent process approach. *Br J Math Stat Psychol* 2013; **66**(3): 470–487.
44. Bryant P and Williamson JA. Asymptotic behaviour of classification maximum likelihood estimates. *Biometrika* 1978; **65**(2): 273–281.

Appendix 1

Binary design 1

The data were generated according to a binomial distribution with the following mean values for the different groups and time points:

$$\text{Group 1} : \exp(time/15 + u_i)$$
$$\text{Group 2} : \exp(-time/15 + u_i)$$
$$\text{Group 3} : \exp(-4 \times time/15 + u_i)$$

time is the time point of the measurement (from 1 to 15) and u_i is a random intercept for individual i, sampled from a normal distribution with mean 0 and standard deviation 0.05 whatever the group.

Binary design 2

Data were generated according to a binomial distribution with the following mean values for the different groups and time points:

$$\text{Group 1} : \exp(2 + u_i)$$
$$\text{Group 2} : \exp(2 - 0.23 \times (time - 5) \times I(time \geq 5) + u_i)$$
$$\text{Group 3} : \exp(1 - 4/5 \times time + 0.41 \times (time - 5) \times I(time \geq 5) + u_i)$$

time is the time point of the measurement (from 1 to 15) and u_i is a random intercept for individual i, sampled from a normal distribution with mean 0 and standard deviation 0.1 whatever the group.

Count design 1

Data were generated according to a Poisson distribution with the following mean values for the different groups and time points:

$$\text{Group 1} : \exp(u_i)$$
$$\text{Group 2} : \exp(0.08 \times time + u_i)$$
$$\text{Group 3} : \exp(-0.2 \times time + u_i)$$

time is the time point of the measurement (from 1 to 15) and u_i is a random intercept for individual i, sampled from a normal distribution with mean 0 and standard deviation 0.1 whatever the group.

Count design 2

Data were generated according to a Poisson distribution with the following mean values for the different groups and time points:

$$\text{Group 1} : \exp(1 - time/15 + u_i)$$
$$\text{Group 2} : \exp(1 - 0.2 \times time + u_i)$$
$$\text{Group 3} : \exp(1 - 0.2 \times time + 0.25 \times (time - 7) \times I(time \geq 7) + u_i)$$

time is the time point of the measurement (from 1 to 15) and u_i is a random intercept for individual i, sampled from a normal distribution with mean 0 and standard deviation 0.1 whatever the group.