*Original Article*

SMMR
STATISTICAL METHODS IN MEDICAL RESEARCH

Statistical Methods in Medical Research
0(0) 1–15

DOI: 10.1177/0962280213475643
smm.sagepub.com
SAGE

# A graphical method to assess distribution assumption in group-based trajectory models

**Mad-Hélénie Elsensohn,[1,2] Amna Klich,[1,2] René Ecochard,[1,2] Mathieu Bastard,[3] Christophe Genolini,[4] Jean-François Etard[5] and Marie-Paule Gustin[1,6,7]**

## Abstract

Group-based trajectory models had a rapid development in the analysis of longitudinal data in clinical research. In these models, the assumption of homoscedasticity of the residuals is frequently made but this assumption is not always met. We developed here an easy-to-perform graphical method to assess the assumption of homoscedasticity of the residuals to apply especially in group-based trajectory models. The method is based on drawing an envelope to visualize the local dispersion of the residuals around each typical trajectory. Its efficiency is demonstrated using data on CD4 lymphocyte counts in patients with human immunodeficiency virus put on antiretroviral therapy. Four distinct distributions that take into account increasing parts of the variability of the observed data are presented. Significant differences in group structures and trajectory patterns were found according to the chosen distribution. These differences might have large impacts on the final trajectories and their characteristics; thus on potential medical decisions. With a single glance, the graphical criteria allow the choice of the distribution that best capture data variability and help dealing with a potential heteroscedasticity problem.

## Keywords

Checking assumptions, grouped-based trajectory models, homoscedasticity, longitudinal data, model adequacy

[1]Hospices Civils de Lyon, Service de Biostatistique, Lyon, France; Université de Lyon, Lyon, France; Université Lyon 1, Villeurbanne, France
[2]CNRS, UMR5558, Laboratoire de Biométrie et Biologie Evolutive, Equipe Biostatistique-Santé, Villeurbanne, France
[3]Epicentre, F-75011 Paris, France
[4]UMR U1027, INSERM, Université Paul Sabatier, Toulouse III; CeRSME (EA 2931), UFR STAPS, Université de Paris Ouest-Nanterre-La Défense
[5]UMI 233 "TransVIHMI" Institut de Recherche pour le Développement, Université Montpellier 1, F-34394 Montpellier, France
[6]Département de santé publique, Institut des Sciences Pharmaceutiques et Biologiques (ISPB), Université de Lyon, Université Lyon 1, Lyon, France
[7]Equipe d'Accueil Mixte 4173; Université de Lyon, Université Lyon 1, Hôpital Nord-Ouest Villefranche-sur-Saône, Lyon, France

**Corresponding author**:
Pr René Ecochard, Service de Biostatistique – Hospices Civils de Lyon, 162 Avenue Lacassagne, F-69424 – Lyon Cedex 03.
Email: rene.ecochard@chu-lyon.fr

# 1 Introduction

Although each person has certain characteristics that make him or her unique, a collection of persons may share one or several characteristics that allow sorting them into categories or groups. In the context of time-series data, clustering individual serial data allows defining group-trajectories; i.e., typical trajectories for certain groups.[1–3] Each individual has then a given probability of belonging to each typical trajectory; however, these groups cannot be directly identified within the population; they are thus called ''latent groups''.

Despite such clustering, a certain degree of heterogeneity remains within groups and between groups. At present, all statistical models do not allow for all kinds of heterogeneity. For example, growth mixture models allow for intra-group variation of individual trajectories whereas group-based trajectory models (also named ''latent class growth models'') assume no such variation and, currently, statistical packages are specific for each model.[4–6] In the latter model, a deviation from the group trajectory is attributed to measurement errors or to time-dependent random variations. As these errors and variations are very frequent in biology and medicine, we will restrict our study here to the group-based trajectory models whose use in clinical research has increased over the last decade.[7]

In group-based trajectory models, the measurements may be continuous, binary or ordinal. Continuous measurements are generally considered as normally distributed, either directly or after logarithmic transformation. However, because biological variables are seldom normally distributed, a gamma distribution may be more suitable. The generalized linear model framework provides a wide range of solutions by allowing a linear combination of explanatory variables to be related to the response variable via a link function and allowing the variance of each measurement to be a function of its predicted value.[8] The model takes thus into account the main part of the fluctuating variance through a variance function and the residual part, supposed to be time-invariant, through a term of dispersion; i.e., an assumption of homoscedasticity of the residuals. Thus, in many cases, the observed distribution deviates from the modelled one; i.e., the variance of the observed measurements is not proportional to the variance estimated by the model. This generates an intra-group heteroscedasticity of the residuals. Moreover, the variance may differ between groups. For example, the variance of a measurement in diseased subjects is generally greater than in healthy subjects because of the heterogeneity of disease severity. This generates an inter-group heteroscedasticity of the residuals.

It is therefore not only useful but necessary to build models that examine the homoscedasticity of intra- and inter-group residuals. These checks avoid violating of the assumption of homogeneity of the variance because this violation may lead to erroneous interpretations of the group-trajectory modelling results.

Within this context, we focused on designing a graphical tool for quick checks of distribution adequacy and homoscedasticity of the residuals. For illustration, we used data on CD4 T lymphocyte (CD4) counts in patients with human immunodeficiency virus (HIV) receiving antiretroviral treatment (ART).

# 2 The generalized linear group-based trajectory model

## 2.1 The group-based trajectory model

A two-level group-based trajectory model may be described as a hierarchical model with a group level (represented by the typical trajectories) and a measurement level. The group-based trajectory model estimates a few typical trajectories and gives each subject a set of probabilities of belonging to each of these trajectories (Figure 1). The mixed model, commonly used in growth modelling,

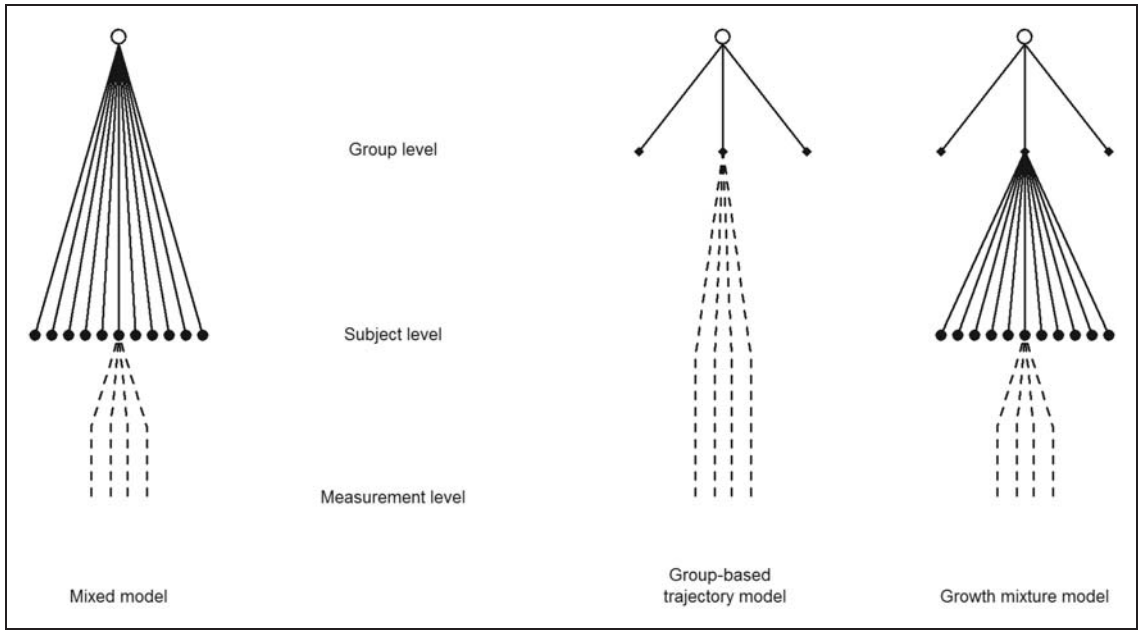


**Figure 1.** On the left, the mixed model considers two levels: a measurement level (dashed lines) and a subject level (continuous lines); i.e., each subject contributes several measurements. In the middle, the group-based trajectory model considers two levels: a measurement level and a group level (three groups in this figure); i.e., each group contributes several measurements whatever the contributing subject. On the right, the growth mixture model considers three levels: measurement, subject and group.

estimates a mean trajectory for the whole population and shows the way each individual trajectory deviates from this mean trajectory. In the former model, the residuals are the differences between the values that form the typical trajectory (group level) and the observed values (measurement level). The growth mixture model takes into account the potential heterogeneity between subjects within the same group, which was not the focus of the present study.

In the context of group-based trajectory modelling, let $Y$ be the matrix of the random variable, let $Y_{\mathrm{i}}$ be the submatrix of the $i^{\mathrm{th}}$ subject, and let $y_{\mathrm{it}}$ be the value relative to the $i^{\mathrm{th}}$ subject at time $t$. Similarly, let $X$ be the matrix of the explanatory variables (time, sex, age, etc.), let $X_{\mathrm{i}}$ be the submatrix of the $i^{\mathrm{th}}$ subject, and let $x_{\mathrm{it}}$ be the matrix line relative to $X_{\mathrm{i}}$ at time $t$ in case of time-dependent variables. If $g = 1, 2, \ldots, G$ where $G$ is the number of groups (or typical trajectories) to reach by clustering, the general form of the model will be:

$$\left.\begin{array}{l} y_{\mathrm{it}} = \mu_{\mathrm{it}} + \varepsilon_{\mathrm{it}} \\ \text{and } \mu_{\mathrm{it}} = x_{\mathrm{it}}\beta_{\mathrm{g}} \end{array}\right\} \text{when } i \in g \tag{1}$$

$\mu_{\mathrm{it}}$ represents each typical trajectory, $\beta_{\mathrm{g}}$ the vector of parameters relative to the effects of covariates $x$ for subjects belonging to $g$, and $\varepsilon_{\mathrm{it}}$ the model's residual error for the $i^{\mathrm{th}}$ subject at time $t$. When the trajectories are modelled using polynomials of time, these polynomials are included in the vector of covariates. To simplify the application, we will present a case without covariates; then, $\mu_{\mathrm{it}} = \mu_{\mathrm{gt}}$ when $i \in g$. Besides, herein, prior probability is not considered in a Bayesian context (where "prior" is the information about the most probable values of the parameters) but in an empirical Bayesian

context where "prior" is applied to the whole sample and "posterior" to each individual of the sample.

The probability of observing a given series of values $Y_i = (y_{i1}, y_{i2}, \ldots, y_{iT})'$ relative to subject $i$ is then:

$$P(Y_i) = \sum_{g}^{G} \pi_g P(Y_i | g)$$

where $\pi_g$ is the prior probability for a subject to belong to group $g$; i.e., the proportion of subjects forming each group, and $P(Y_i | g)$ the probability of observing $Y_i$ values of the $i^{th}$ subject knowing that he belongs to a given group. In other words, $P(Y_i | g)$ corresponds to the likelihood of $Y_i$ when subject $i$ belongs to group $g$.

The posterior probability is defined as the probability of belonging to a given group knowing the data; that is, $P(g|Y_i)$. According to Bayes theorem, the posterior probability for the $i^{th}$ subject to belong to a given typical trajectory $g$ is:

$$P(g|Y_i) = \frac{P(Y_i | g)\pi_g}{\sum_{1}^{G} \pi_g P(Y_i | g)}, g = 1, 2, \ldots, G$$

## 2.2 The intra-group variance

When the distribution of the measurements belongs to the exponential family, using a notation close to that of McCullagh and Nelder[8] the intra-group variance of the measurements can be written as:

$$Var(y_{it}) = \phi_g \, V(\mu_{gt}), \; i \in g$$

where $\phi_g$ is the time-invariant dispersion within group $g$ and $V(\mu_{gt})$ the variance function.

The variance of each measurement is proportional to a function of its predicted value $\mu_{gt}$. Let $V(\mu_{gt}) = \mu_{gt}(1 - \mu_{gt})$ be the expression of this function with a binomial distribution of the variable, $V(\mu_{gt}) = \mu_{gt}$ its expression with a Poisson distribution, $V(\mu_{gt}) = \mu_{gt}^2$ its expression with a gamma distribution, etc. This function describes also naturally some inter-group variance; e.g., variance between group $g$ and $g'$, because, generally $V(\mu_{gt}) \neq V(\mu_{g't})$.

The factor of proportionality, i.e., the dispersion parameter $\phi_g$, is time-invariant but may differ between groups, which provides another way of modelling inter-group heteroscedasticity: the difference in intra-group variance between group $g$ and $g'$ may be expressed using the dispersion parameter, $\phi_g \neq \phi_{g'}$. Figure 2 illustrates these two aspects of intra- and inter-group variance.

## 2.3 Choice of the intra-group distribution

### 2.3.1 Normal distribution and homoscedasticity

Under the assumption of a normal distribution of the measurements, equation (1) becomes:

$$\left.\begin{array}{l} y_{it} = \mu_{gt} + \varepsilon_{it} \\ \text{and } \varepsilon_{it} \sim N(0, \sigma_g^2) \end{array}\right\} \text{when } i \in g$$

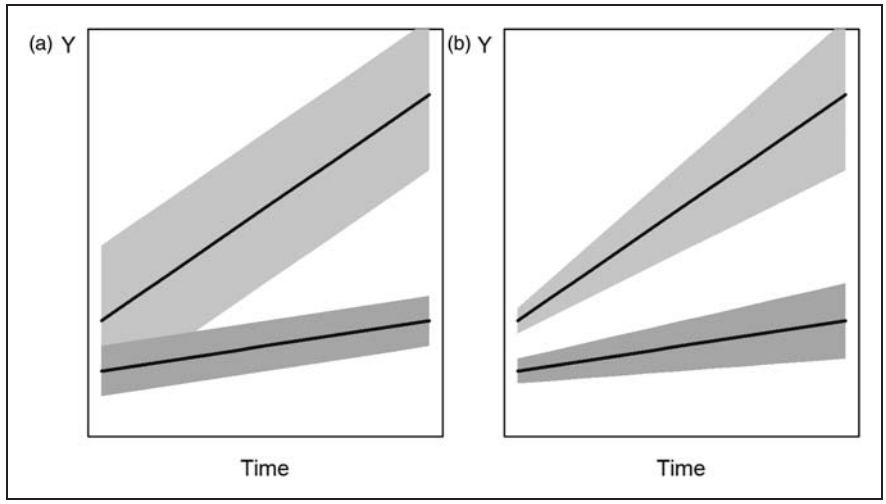


**Figure 2.** The solid black lines represent two typical trajectories. The grey intervals represent the predicted variability around these trajectories. (a) The variance of the residuals is supposed to be time-invariant but may differ between groups; (b) the variance of the residuals is supposed to vary according to the expected mean.

$\mu_{gt} = x_{it}\beta_g$ is the mean relative to group $g$ at time $t$ and $\sigma_g^2$ the variance. Then, $Var(y_{it}) = \sigma_g^2 = \phi_g$ and $V(\mu_{gt}) = 1$. Within each group, the variance is supposed to be constant, equal to the time-invariant dispersion $\phi$, which is supposed to be the same for all groups: $\phi_1 = \phi_2 = \cdots = \phi_G$.

### 2.3.2 *Alternative assumptions*

Let us consider the case of two typical latent trajectories: one followed mainly by healthy subjects and the other followed mainly by diseased subjects. In this case, the individual trajectories of the diseased subjects may be very dispersed whereas those of the healthy subjects are often clustered. This inter-group heteroscedasticity should then be taken into account.

Under the assumption of normal distribution of the variable of interest within each group, the inter-group heteroscedasticity (Figure 2(a)) is taken into account by the fact that $\sigma_g^2$ is different between groups. Moreover, it is well-known that the intra-group variability of the measurements around their expected values are sometimes naturally dependent on these expected values, which themselves evolve over time. For example, the variance of a biomarker values is frequently proportional to their expected values; this is a situation where the intra-group heteroscedasticity may be modelled using a non-normal distribution.

One alternative to the normal distribution is the gamma distribution which is defined for positive values (the case of most biomarkers) and characterized by a constant coefficient of variation, the variance being equal to the square of the mean. The probability density of a gamma distribution is:

$$f(y) = \frac{1}{\Gamma(\alpha)\theta^{\alpha}} y^{\alpha-1} \exp\left(\frac{-y}{\theta}\right)$$

For a subject $i$ belonging to group $g$ at time $t$, $y_{it} \sim \text{gamma}(\alpha, \theta_{gt})$. Then:

$$\left.\begin{array}{l} E(y_{it}) = \alpha\theta_{gt} = \mu_{gt} \\ \text{and } Var(y_{it}) = \frac{1}{\alpha}\mu_{gt}^2 \end{array}\right\} \text{when } i \in g$$

Thus, under the assumption of a gamma distribution of a set of measurements, the dispersion parameter $\phi$ equals the inverse of the shape parameter $\alpha$, and $V(\mu_{gt}) = \mu_{gt}^2$; i.e., within each group, the variability of the measurements is supposed to increase rapidly along with their expected values but the coefficient of variation remains stable. The gamma distribution was used to account for intra-group variability (Figure 2(b)). When the measurement levels differ between groups, the variance function affects intra- and inter-group variances. When the shape parameter $\alpha_g$ is specific to each group $g$, it is also possible to introduce an additional inter-group heteroscedasticity.

Another alternative to the normal distribution is the negative binomial (NB) distribution that may be considered as a mixture of Poisson distributions whose means follow a gamma distribution. Within this context of parameterization, considering the gamma distribution and parameters $\alpha$ and $\theta_{gt}$, the probability density of a negative binomial distribution is[8]:

$$f(Y = y; \alpha; \theta_{gt}) = \frac{\Gamma(y+\alpha)}{\Gamma(y+1)\Gamma(\alpha)} \left(\frac{1}{\theta+1}\right)^{\alpha} \left(\frac{\theta}{\theta+1}\right)^{y}$$

For a subject $i$ belonging to group $g$ at time $t$, $y_{it} \sim \mathrm{NB}(\alpha, \theta_{gt})$. Then:

$$\left.\begin{array}{l} E(y_{it}) = \alpha\theta_{gt} = \mu_{gt} \\ \text{and } Var(y_{it}) = \left(\mu_{gt} + \frac{1}{\alpha}\mu_{gt}^2\right) \end{array}\right\} \text{when } i \in g$$

Just as a gamma distribution, a negative binomial distribution introduces naturally into the model a variability that is a function of the height of the typical trajectory with $V(\mu_{gt}) = \mu_{gt} + \frac{1}{\alpha}\mu_{gt}^2$. Another inter-group heteroscedasticity; i.e., a group-specific dispersion, may be also introduced using a shape parameter $\alpha_g$ for each group $g$.

## 3 Time-specific and group-specific dispersion of the residuals

The variance of the measurements around the typical trajectories may be modelled according to the positions of these trajectories; i.e., as time- and group-specific trajectories. Nevertheless, the appropriateness of the model should be checked. The observed variance may or may not be correctly described by the fitted model. If not, this results in a heteroscedasticity of residuals and model inadequacy.

The observed time- and group-specific variance of the residuals may be calculated and a Pearson estimation of the "local" dispersion computed by dividing this observed variance by the variance function; i.e., the expected variance. A stable dispersion over time will be in favour of an assumption of homoscedasticity of the residuals. If this local dispersion is time- and/or group-dependent, the model will be considered inadequate.

The situation is complicated by the fact that each subject has a specific probability of belonging to each group. Thus, an estimation of the local variance may be obtained by weighting appropriately the participation of each individual to the estimation of the local variance of each group.

### 3.1 Local dispersion of the residuals

The local dispersion of the residuals was estimated using the following steps for each typical trajectory $g$:

(1) Calculation of the observed residuals $y_{it} - \hat{\mu}_{gt}$, for each subject, whatever the group the subject belongs to.

(2) Calculation of the group- and time-specific weighted local variance of the residuals, $\hat{\sigma}^2_{gt}$: these are obtained as time-moving averages of the observed squared residuals, the weights being the posterior probabilities of each measurement $y_{it}$ to belong to a particular group $g$.
(3) Calculation of the local dispersion of the residuals: $\frac{\hat{\sigma}^2_{gt}}{V(\hat{\mu}_{gt})}$

## 3.2 Envelope plot of the local standard deviation and the local dispersion of the residuals

An envelope plot is obtained by drawing two lines, above and below a predicted trajectory. The width of this envelope is either the amplitude of the local standard deviation or that of the local dispersion of the residuals. When this envelope is too narrow, a magnification factor $k$ may be added to improve the graph readability.

The boundaries of the envelope are thus:

$\hat{\mu}_{gt} \pm k\hat{\sigma}_{g_{\hat{\sigma}^2_{gt}}}$ for the local standard deviation of the residuals and
$\hat{\mu}_{gt} \pm k\frac{\hat{\sigma}^2_{gt}}{V(\hat{\mu}_{gt})}$ for the local dispersion of the residuals.

These two variability intervals are equal with a normal distribution of the variable of interest because the variance function equals one. These variability intervals show the dispersion of the measurements around the typical trajectories not the confidence intervals.

During model building, the assumptions of normal distribution of the measurements and homoscedasticity of the residuals may be used to obtain a first group-based trajectory model. The shape of the envelope of the local standard deviation of the residuals may then indicate the appropriateness of the adopted distribution: if the boundaries of the envelopes are parallel but the widths of the intervals differ between groups, a normal distribution with a group-specific coefficient of dispersion should be preferred. If, on the contrary, the boundaries are not parallel and the envelope enlarges with the increase of the expected measurement, then a gamma distribution or a negative binomial distribution should be chosen. At the final stage, having chosen the model, the parallelism of the boundaries of the variability interval may be considered as a criterion for model appropriateness: a departure from the parallelism should raise doubts about the appropriateness of the model.

## 3.3 Statistical inference

The models presented here were applied to the HIV data using various algorithms: (1) *proc traj* in SAS[9,10] for normal distribution and homoscedasticity; (2) *mmlcr* library in *R*[11] for heteroscedastic normal and negative binomial distributions with group-specific shape parameters and (3) *allvc* function from *npmlreg* library in *R*[12] for gamma distribution without additional inter-group heterogeneity.

Smooth curves for the variability intervals were calculated using *supsmu* function from *stats* library in *R*.

# 4 Analysis of HIV data

After several years, an infection with the HIV may induce an *acquired immune deficiency syndrome (AIDS)*.[13,14] Today's fight against this disease includes prevention and ART.

The latter does not eliminate the virus but can significantly slow the infection by reducing the amount of virus in the blood. Actually, the life expectancy of people on ART has greatly increased in recent years. However, ART remains inaccessible for many patients, especially in underdeveloped countries.[15]

To halt the spread of AIDS, the Senegalese government launched in 1998 the ISAARV Initiative.[16,17] A part of this program is the ANRS 1215 cohort from which the present study data were extracted. The main objective of establishing this cohort was evaluating the short- and long-term impacts of ART. The ANRS 1215 cohort study collected data on 404 patients enrolled between 1998 and 2002 (of whom 246 patients are still followed-up). All received ART upon entry into the cohort. Moreover, all had clinical and biological check ups at inclusion then regular monitoring visits.

Various tests were carried out during patient visits, especially CD4-cell counts in blood (3043 measurements in 403 patients; one invalid set of counts). This variable was used to explore typical CD4-cell-count trajectories. Thus, we monitored the change in CD4 counts over time under different assumptions and considered three trajectory groups because patients put on ART respond generally in one of three ways: (1) recovery of a ''normal'' CD4 count after a few months or years; (2) no progress or (3) some intermediate progress.

In addition, we considered the explanatory variable ''time'' under a polynomial form of order 4.

## 4.1 Case of normal distribution and homoscedasticity

Figure 3 shows the results obtained under the assumptions of normal distribution of the variable and homoscedasticity of the residuals. The subjects were assigned to one of the trajectory groups according to the highest posterior probability of belonging to each group. Figure 3(a) shows that group 3 (the highest trajectory) includes only 2% of the cohort. Figure 3(b) shows the typical trajectories and their local standard deviation envelopes. The width of these envelopes vary along time between groups and within each group showing the importance of examining inter- and intra-group heterogeneity. With these data and model, the variability along time of the highest trajectory may be due to the small number of subjects.

## 4.2 Case of normal distribution and heteroscedasticity

Under the assumption of heteroscedasticity, the variances estimated for groups 1, 2 and 3 were different (10505, 14503 and 59577, respectively). Unlike the previous model, the high variability of group 3 made it attract 19% of the cohort. The typical trajectory of group 3 became logically closer to the other trajectories than before (Figure 4). Though, the envelopes were still not parallel. The variability intervals increased along with the average under the assumption of either homoscedasticity (Figure 3(b)) or heteroscedasticity (Figure 4(b)). It becomes then more appropriate to consider another distribution than the normal one.

## 4.3 Case of a gamma distribution

When the distribution of the CD4-cell counts was assumed to follow a gamma distribution, the additional variability taken into account changed the number of subjects within each group. Indeed, the highest group now includes 22% of the cohort (vs. 2% and 19% with the normal homoscedastic and the normal heteroscedastic distribution, respectively) (Figure 5). As expected, the local standard deviation envelopes are no more parallel (Figure 5(b)), but the local dispersion envelopes may be considered parallel (Figure 5(c)).

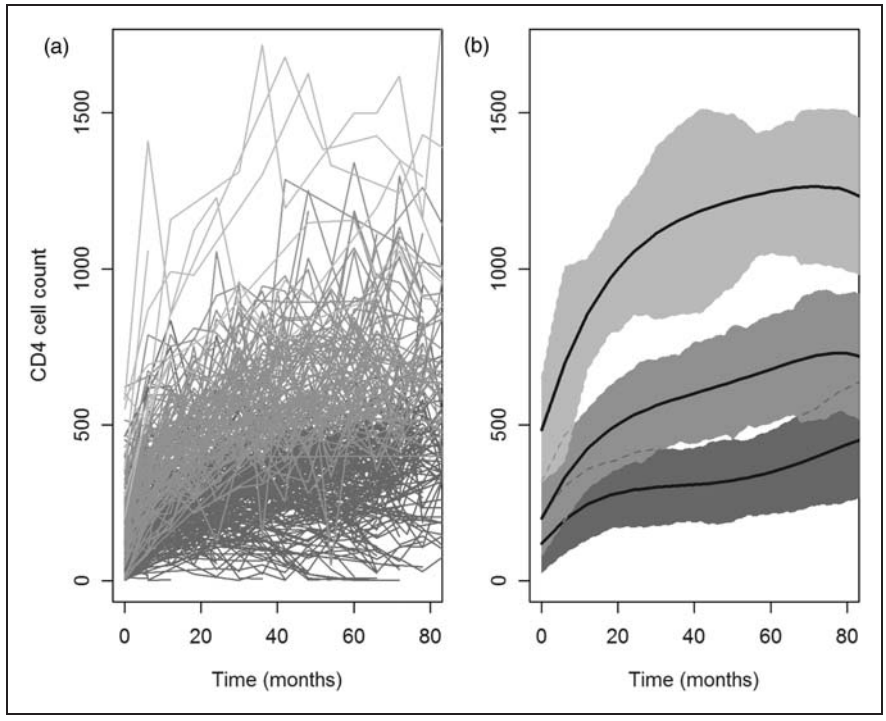


**Figure 3.** Assumption of normality and homoscedasticity. (a) On the left, each subject was assigned to one of the groups according to the highest posterior probability; the shades of grey identify the groups. (b) On the right, the three typical trajectories with their standard deviation envelopes (magnification factor 1). The prior probabilities, from top to bottom, are 2%, 33% and 65%.

## 4.4 Case of a negative binomial distribution with group-specific shape parameter

When the response variable was assumed to follow a negative binomial distribution with different group-specific shape parameters, the three groups included almost equivalent proportions of subjects. The widths of the local standard deviation envelopes increased with the average in groups 1 and 3 (Figure 6(b)). The boundaries of the local dispersion envelopes could be considered parallel and their widths were similar (Figure 6(c)).

## 4.5 Comparison of the four-case results

Table 1 shows the estimations of the variances and the group sizes. It shows models in which the dispersion is supposed identical in all groups (normal distribution with homoscedasticity and gamma distribution with same shape factor for all groups) and models in which the dispersion was allowed to vary between groups (normal distribution with heteroscedasticity and negative binomial distribution with group-specific shape factors). The proportions of the cohort within the groups differ dramatically.

Table 2 presents two criteria for the likelihood of the data: the Bayesian Information Criterion (BIC) and entropy. The BIC criterion is an indicator of model adequacy. Here, the two smallest BIC

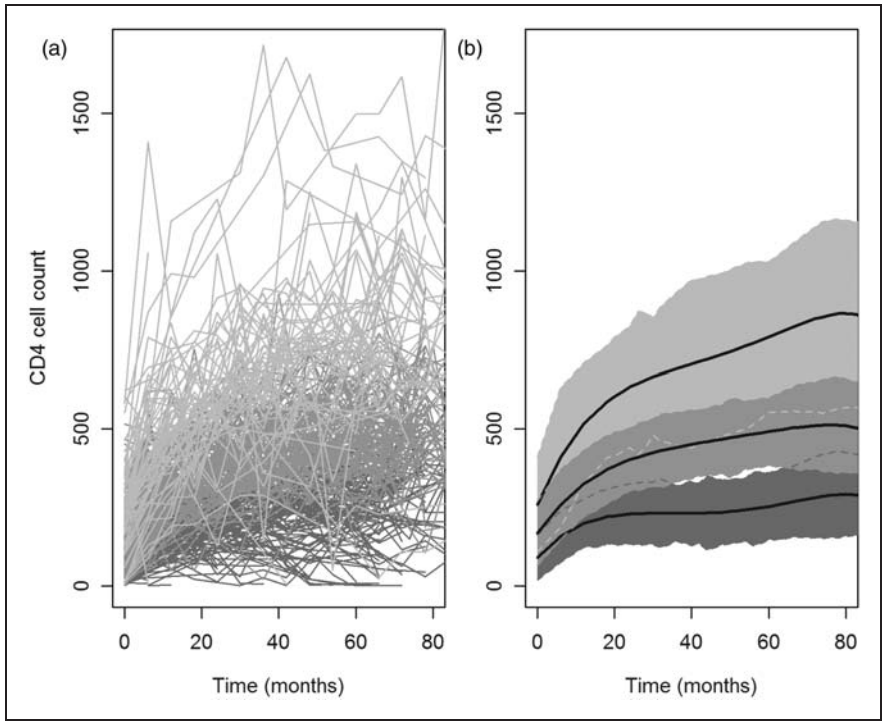


**Figure 4.** Assumption of normality and heteroscedasticity. (a) On the left, each subject was assigned to one of the groups according to the highest posterior probability; the shades of grey identify the groups; (b) on the right, the three typical trajectories with their standard deviation envelopes (magnification factor 1). The prior probabilities, from top to bottom, are 19%, 41% and 40%.

values were observed with heteroscedastic normal as well as with negative binomial distributions; they differed by 172 units in favour of the heteroscedastic normal distribution. Entropy is an indicator of model prediction ability or the inability of a model to predict with certainty the group each subject belongs to (the lowest entropy value corresponds to the best predicting model). The model that assumed a homoscedastic normal distribution had the smallest entropy (70 units) and the one that assumed a negative binomial distribution had a slightly higher entropy (76 units). The model that assumed a gamma distribution showed the highest BIC and the highest entropy.

## 5 Discussion

A first step of checking intra- and inter-group homoscedasticity of the residuals is crucial for building group-based trajectory models in medicine. We developed here a new user-friendly graphical method to use in group-based trajectory model building to check efficiently the assumption of homoscedasticity of the residuals. The method consists in using the residual and fitted values to draw a local dispersion envelope around each typical trajectory. This graphical tool allows a quick check of whether a given model succeeds or fails in capturing data heterogeneity so as to implement modelling corrections or refine the interpretation of

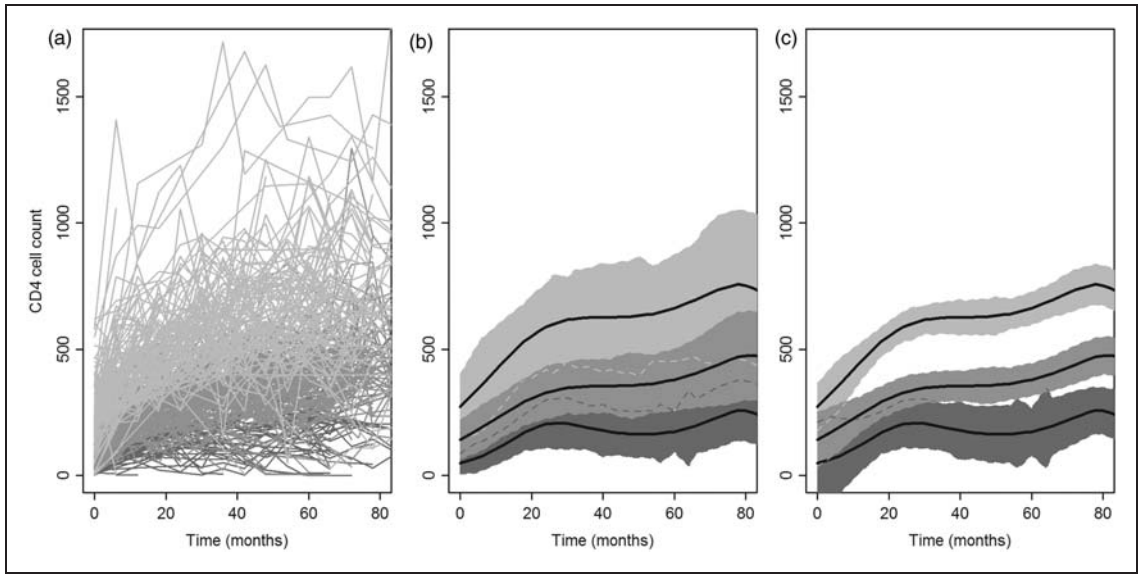


**Figure 5.** Assumption of gamma distribution. (a) On the left, each subject was assigned to one of the groups according to the highest posterior probability; the shades of grey identify the groups. (b) The three typical trajectories with their standard deviation envelopes (magnification factor 1) and (c) their local dispersion envelopes (magnification factor 200, to show larger envelopes). The prior probabilities, from top to bottom, are 22%, 33% and 45%.

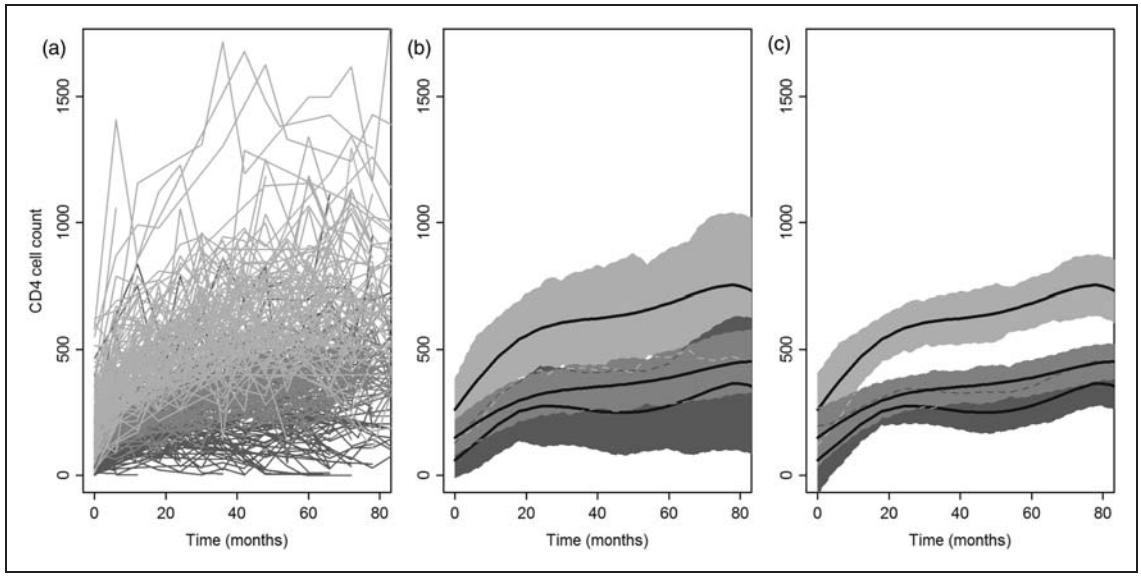


**Figure 6.** Assumption of negative binomial distribution with group-specific shape parameter. (a) On the left, each subject was assigned to one of the groups according to the highest posterior probability; the shades of grey identify the groups. (b) The three typical trajectories with their standard deviation envelopes (magnification factor 1) and (c) their local dispersion envelopes (magnification factor 100, to show larger envelopes). The prior probabilities, from top to bottom, are 32.5%, 35% and 32.5%.

**Table 1.** Estimation of the parameters of each distribution used to determine the group trajectories.

| | Homoscedastic normal distribution | | Heteroscedastic normal distribution | | Gamma distribution | | Negative binomial distribution with group-specific shape parameter | |
|---|---|---|---|---|---|---|---|---|
| Group trajectory | $\sigma_g^{2a}$ ($10^3$) | $\pi_g$ (%)[b] | $\sigma_g^{2a}$ ($10^3$) | $\pi_g$ (%) | $\alpha^c$ | $\pi_g$ (%) | $\alpha_g{}^c$ | $\pi_g$ (%) |
| Group 1 | 21 | 65 | 11 | 40 | 5.2 | 45 | 6.7 | 32 |
| Group 2 | 21 | 33 | 15 | 41 | 5.2 | 33 | 1.5 | 36 |
| Group 3 | 21 | 2 | 60 | 19 | 5.2 | 22 | 10.0 | 32 |

[a]The variance of the normal distribution; i.e., the dispersion.
[b]Prior probabilities: Proportion of subjects within each group.
[c]Shape parameter of the gamma or of the negative binomial distribution.

**Table 2.** Bayesian information criterion (BIC) and entropy values with each distribution studied.

| | Number of parameters[a] | BIC | Entropy |
|---|---|---|---|
| Homoscedastic normal distribution | 18 | 39572 | 70 |
| Heteroscedastic normal distribution | 22 | 39117 | 108 |
| Gamma distribution | 19 | 40319 | 451 |
| Negative binomial distribution with group-specific shape parameter | 22 | 39289 | 76 |

[a]The number of parameters was the number of degrees of freedom plus one minus the number of groups.

the trajectories. Whenever it is not possible to find a model that respects the assumption of homoscedasticity, still carrying out the analysis may lead to erroneous results and interpretations. In such a case, given the current absence of specific functions for latent class models, the problem should be mentioned and its consequences discussed with the appropriate limits.

For the theoretical application, we chose four distribution types from the generalized linear model framework that take progressively into account the variability of the data. Indeed, homoscedastic normal distribution ignores completely intra- and inter-group variabilities, the heteroscedastic normal distribution deals only with inter-group variability, the gamma distribution allows for intra-group variability because of the linear variance function of the mean, and, finally, the negative binomial distribution deals more properly with intra- and inter-group variabilities through a complex quadratic variance function of the mean and a group-specific shape parameter. Not allowing the maximum amount of variability in a model affects the results of the analysis. We noticed that the higher the variability taken into account by the model, the more balanced the sizes of the groups.

The assumption of homoscedasticity of the residuals was checked in the four models using the graphical tool we developed. The graphical visualization of the local standard deviation and the local dispersion envelopes may be considered as validity criteria for the chosen model. With normal distributions, the local standard deviation envelope is sufficient because the widths of the two variability intervals are equal. Whatever the model, the assumption about intra- and inter-group variabilities can be validated whenever the variability intervals are parallel and have equal widths for all group. In this case, the model captures correctly the variability of the data.

Whenever necessary, smoothing the standard deviation of the residuals over time results in an easier-to-read curve and compensates for small and uneven sample sizes at each time point. This smoothing does not interfere with between-group or model comparisons because it is carried out on all groups in all models.

Once the model assumptions checked, other features may be also examined: the model likelihood (using the BIC criterion) or the model's ability to identify well-separated groups (using entropy). As already shown, the assumption of parallel boundaries (i.e., uniform envelope width) is met with gamma or negative binomial distributions. Comparison of BICs and entropies indicated that the model that assumed a negative binomial distribution was the best.

Choosing a model according to the aspects of its variability intervals is essential. Indeed, the present results show that BIC or entropy, alone, would not have favoured the homoscedastic models; i.e., models with gamma or negative binomial distributions. In our example, ignoring this first step of checking the assumption of homoscedasticity of the residuals would have led to consider the normal heteroscedastic model as the best one and to erroneous conclusions. Indeed, there are substantial differences in the aspects of the typical trajectories between this model and the best one that assumes a negative binomial distribution.

For the practical application, the method explored CD4-cell count trajectories in HIV patients put on ART. Such data are well-suited to demonstrate the benefit of the method because of their high variability (daily fluctuation, measurement errors, etc.). In this illustration, three trajectory groups were decided according to the three types of patient responses to ART. In other settings, when the number of groups cannot be preselected on clinical or other considerations, an optimal number of groups may be obtained using the BIC criterion, the Calinski-Harabasz criterion, or else.[18,19] Moreover, in our example, no covariates, such as the baseline CD4-cell count, the clinical stage, sex or age were introduced into the model though some are well-known predictors of CD4-cell count trajectories; this was to simplify the clinical setting and focus on the graphical tool. If one wishes to introduce covariates, the method may be applied without major changes. Indeed, the introduction of covariates does not change the calculation of the standard deviation; thus, the calculation of the local dispersion of the residuals. The introduction of covariates decreases frequently the dispersion of the residuals but does not suppress them; thus, the assumption of homoscedasticity should still be checked.

Here the method was applied to a continuous variable. The extension of this method to a binary outcome is possible and could be made using predicted values per subject and quantifying the time-dependent dispersion using specific formulas.[8]

This work has shown the importance of the source and amount of variability in defining trajectory groups. This variability is expressed on at least two levels: subject level and measurement level. The former is important because the closeness of a given subject trajectory to a typical trajectory does not mean this subject belongs strictly to this trajectory. In this framework, the intra-group variability was explained only by the measurement variability because the normal distribution can be regarded as a marginal distribution with a normal subject variability and a normal measurement variability. Similarly, the negative binomial distribution can be regarded as the result of a subject variability (described by a gamma distribution) and a measurement variability (described by a Poisson distribution). This two-level model (group level and measurement level) should not discard another hierarchical model, the growth mixture model, that considers an additional subject level laying between the group level and the measurement level. This is an interesting objective for further work.

The proposed graphical representation gave an idea about the progression over time of the standard deviation of the residuals of each trajectory group. It becomes then possible to examine

rapidly both the temporal trend of the residual standard deviation and its difference between groups. This examination of the heteroscedasticity of the residuals allows choosing the model that fits best the data. This graphical information is richer than that given by other standard statistical tests that compare several variances. Indeed, it takes into account both, the residual variability over time and the posterior probability of each subject to belong to each group. To our knowledge, this is the first time such a method is adapted and applied to group-based trajectory models. However, the method requires some time to develop the adequate programs and graphics but it avoids false result interpretations. It would be, thus, very useful to develop a specific software package. Another limit of this work is that the generalized linear framework we chose restricted the choice of the distributions, but the method may be extended to more general models such as the general non-linear mixture of curves that allows the modelling of any conceivable mixture of distributions.

In the recent years, group-based trajectory models have been increasingly used in clinical research. The present innovative method is based on the calculation of local dispersion of the residuals and allowed checking the assumption of homoscedasticity of these residuals through a graphical diagnostic tool that shows the local dispersion of the residuals over time. This first step is crucial for building reliable group-based trajectory models and should precede the use of other criteria such as BIC or entropy to avoid choosing an inadequate model and drawing erroneous conclusions. However, one limitation of this method is that group-based trajectory models allow a limited choice of data distributions. Thus, it would be interesting to try extending the method to non-linear models and growth mixture models that allow individual variability within typical trajectories and the use of any conceivable mixture of distributions.

## Acknowledgements

The authors are grateful for Eric Delaporte, Ibra Ndoye, and the ANRS 1215 study group for providing the dataset. Project ANRS 1215 was funded by the Agence Nationale de Recherche sur le Sida (ANRS) and the Institut de Recherche pour le Développement (IRD).

## Funding

This research received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors.

## References

1. Everitt BS, Landau S and Leese M. *Cluster analysis*, 5th edn. Wiley: Chichester, 2010.
2. Nagin DS and Tremblay RE. Parental and early childhood predictors of persistent physical aggression in boys from kindergarten to high school. *Arch Gen Psychiatry* 2001; **58**: 389–394.
3. Nagin DS and Tremblay RE. Analyzing developmental trajectories of distinct but related behaviors: a group-based method. *Psychol Methods* 2001; **6**: 18–34.
4. Muthén B and Shedden K. Finite mixture modeling with mixture outcomes using the EM algorithm. *Biometrics* 1999; **55**: 463–469.
5. Muthén B. Mplus, http://www.statmodel.com/ (accessed 02 February 2011).
6. Nagin DS. Analyzing developmental trajectories: a semiparametric, group-based approach. *Psychol Method* 1999; **4**: 139–157.
7. Nagin DS and Odgers CL. Group-based trajectory modeling in clinical research. *Ann Rev Clin Psychol* 2010; **6**: 109–138.
8. McCullagh P and Nelder JA. *Generalized linear models*, 2nd edn. London: Chapman and Hall, 1989.
9. Jones BL, Nagin DS and Roeder K. A SAS procedure based on mixture models for estimating developmental trajectories. *Sociol Methods Res* 2001; **29**: 374–393.
10. 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.
11. Buyske S. The mmlcr Package, http://212.219.56.146/sites/lib.stat.cmu.edu/R/CRAN/doc/packages/mmlcr.pdf (2006, accessed 20 January 2011).
12. Einbeck J, Darnell R and Hinde J. The npmlreg Package, http://212.219.56.146/sites/lib.stat.cmu.edu/R/CRAN/doc/packages/npmlreg.pdf (accessed 20 January 2011).

13. OMS. VIH/SIDA, http://www.who.int/topics/hiv_aids/fr/ (accessed 5 November 2010).
14. Centers for Disease Control and Prevention. HIV/AIDS, http://www.cdc.gov/hiv/default.htm (accessed 5 November 2010).
15. ONUSIDA. AIDS Epidemic update. 2009.
16. 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, p.230.
17. Etard JF, Ndiaye I, Thierry-Mieg M, et al. Mortality and causes of death in adults receiving HAART in Senegal: a 7-year cohort study. *AIDS* 2006; **20**: 1181–1189.
18. Schwarz G. Estimating the dimension of a model. *Ann Stat* 1978; **6**: 461–464.
19. Calinski T and Harabasz J. A dendrite method for cluster analysis. *Commun Stat* 1974; **3**: 1–27.