Finite Mixture Models for an underlying Beta distribution with an application to COVID-19 data

Cédric NOEL Jang SCHILTZ

May 4, 2022

Abstract

We introduce an extension of Nagin's finite mixture model to underlying Beta distributions and present our R package trajeR which allows to calibrate the model. In a second part of the paper, we use this model to analyse the efficiency of the sanitary measures taken by the different countries during the first part of the COVID-19 pandemics.

1 Introduction

Finite Mixture Models in the sense of Nagin (2005) are fuzzy logic cluster analysis models for time series. Starting from a sample of trajectories, the aim is to detect a number of subgroups of the sample, so that subjects in the same group exhibit quite similar data trajectories, whereas two subjects from two different groups have trajectories that differ in some sense. These models are part of a larger strand of models that analyze latent evolutions in longitudinal data.

Model-based approaches comprise latent class growth analysis, growth mixture modeling and group-based trajectory models (van der Nest et ali. 2020). The objective of all these approaches is to capture information about interindividual differences in intraindividual change over time (Nesselroade 1991).

There are a host of model-based techniques for analysing outcome development over time (see Verbeke et ali. 2014). According to van der Nest et ali (2020), they can be divided into models comprising the estimation of one latent class, such as growth curve models (Laursen and Hopf, 2006), or more than one latent class, such as latent class growth analysis (Berlin, Parra and Williams 2014), examples of which includer growth mixture models (Muthén and Shedden 1999) and group-based trajectory models (Nagin 2005).

Growth curve models are not categorizing subjects, but try to explain the relationship between explanatory variables and the trajectories of interest as a whole. Most modern econometric techniques fall into this category (see Greene 2018).

In latent class growth analysis, the relationship between the explanatory variables and the trajectories of interest is group-specific. Growth mixture modeling, introduced by Muthen & Shedden (1999), is a very suitable framework to handle the issue of unobserved heterogeneity. In growth mixture modeling, there are random effects influencing all the parameters defining the typical group trajectories for all subjects in the sample. This has the advantage that fewer groups are required to specify a satisfactory model, but the groups are not always clearly defined and group cross-over effects may exist. Group-based trajectory modeling, also called finite mixture models (Nagin 2005) on the other hand has no random effects to capture individual differences in a continuous way. All individual deviations from the typical trajectory of a group are treated as residual errors. It was originally introduced by Nagin & Land (1993) and is actually specifically

designed to detect the presence of distinct subgroups among a set of trajectories. Thus, it represents an interesting compromise between analysis around a single mean trajectory and case studies (VonEye & Bergman). Compared to subjective classification methods, the nonparametric mixed model has the advantage of providing a formal framework for testing the existence of distinct groups of trajectories. Furthermore, in finite mixture models, the error variance is supposed to remain constant over all groups and the complete time line. This implies that far fewer parameters need to be estimated and these kind of models remain useful for small samples. Schiltz (2015) introduced a generalization of this model in which the error variance remains constant inside a given group, but is allowed to differ across groups, which is a far more realistic assumption in a lot of applications. While the conceptual aim of the analysis is to identify clusters of individuals with similar trajectories, the model's estimated parameters are not the result of a classical cluster analysis, but of maximum likelihood estimation (Nagin 2005). Let's also point out that finite mixture models are actually fuzzy logic cluster analysis of time series models and should not be confounded with finite mixture models in the sense of McLachlan & Peel (2000), which refer to mixtures of probability distributions.
This paper extends finite mixture models to the case of an underlying Beta distribution and presents the R package trajeR which allows to calibrate the model. The package can be seen as an extension of the **SAS** procedure **Proc Traj** developed by Jones, Nagin & Roeder (2001), Jones & Nagin (2007) and their **Stata** version (Jones & Nagin 2012) to the generalized finite mixture model, with the additional advantage that, as an **R** package, it is open source software and can be adapted to the needs of the users.

There a hundreds of papers that have been published on research about the COVID-19 pandemics over the last two years and an excellent overview about the different strands of research thematic and the corresponding papers can be found in Masters & Spiegelhalter (2021). The use of a cluster analysis methodology in this context is however much rarer. Chandu (2020) analyzes 89 countries by means of a K-means clustering algorithm and identifies two groups of countries, showing that a high COVID-19 case fatality rate, higher proportion of positive test results, higher percentage of GDP spent as public health expenditure and a greater percentage of elderly people are related. Zarikas et ali. (2020) use hierarchical cluster analysis to analyze active cases, active cases per population and active cases per population and per area based on John Hopkins epidemiocolgical data and they identify four different shapes in the evolution rate of COVID-19. Alvares et ali. (2020) use hierarchical cluster analysis methods on time series with the aim of identifying groups of countries with a similar spread of the coronavirus. They show that there are groups of countries with differentiated contagion dynamics and conclude that the actions taken by the countries, the speed at which they were taken and the number of tests carried out may explain part of the differences in these dynamics. Rahman et ali (2020) propose a data-driven dynamic clustering framework for moderating the adverse economic impact of COVID-19 flare-up. They model lockdown as a clustering problem and design a dynamic clustering algorithm for localized lockdown by taking into account the pandemic, economic and mobility aspects. James and Menzies (2020) propose a hierarchical and K-means cluster-based method to analyze COVID-19 pandemic cases and death counts. They study the change in both quantities over time and identify anomalous countries in the progression from COVID-19 cases to deaths. This analysis can aid in highlighting the most and least significant public policies in minimizing a country's COVID-19 mortality rate. Kinnunen et ali. (2021) identify the groups of countries with similar Covid strategies on the one hand and groups with similar performance success on the other hand, and construct a composite Covid Mitigation Index for comparative purposes, thus, implying how to redesign the strategic policies. They use dynamic clustering with a Gaussian Mixture Model in the sense of McLachlan on 2 month periods in order to achieve this.

In the second part of this paper, we use a generalized finite mixture with underlying Beta distribution to study the rate of COVID-19 contamination during the first part of the pandemic. The remainder of this paper is structured as follows. In section two, we present the generalized finite mixture model for an underlying Beta distribution. Section three introduces our R package **trajeR** and section four shows how the model introduced in section two can be calibrated with the help **trajeR**.

2 The generalized finite mixture model for an underlying Beta distribution

2.1 Finite mixture models

In group based trajectory modeling, we consider a population of size N divided into K latent classes. The assignment of the individuals into classes is based on the degree of similarity of the developmental trajectories.

More precisely, consider a time-varying variable of interest Y defined in a population Ω of size N. Let $Y_i = y_{i_1}, \cdots, y_{i_T}$ be T measures of the variable Y, taken at times $t_1, \cdots, t_T$ for subject number i belonging to a sample of size n.

The aim of the analysis is to divide the population into K sub-populations $G_1, ...G_K$, which are homogeneous in the sense that two subjects in the same group have similar trajectories for the variable of interest Y and two subjects in different groups have quite different trajectories for the variable of interest Y.

Let $P^k(Y_i)$ be the probability of Y_i given membership in group G_k and $P(Y_i)$ the unconditional probability of observing the realization Y_i of Y. Furthermore, for a given group G_k, we suppose conditional independence for the sequential realizations of the elements y_{i_t} over the T periods of measurements. Then,

$$P(Y_i) = \sum_{k=1}^{K} P(G_i = k) P^k(Y_i). \tag{1}$$

By definition of a finite mixture model (Nagin 2005), the density f of Y is given by

$$f(y_i; \psi) = \sum_{k=1}^{K} \pi_k g_k(y_i; \Theta_k). \tag{2}$$

The role of the parameters Θ_k is to describe the shape of the trajectories in group k.

This models supposes no structural between-subject variability within a class, hence the error variance is assumed to be constant inside a given class. Moreover, the group size $\pi_k > 0$ denotes the probability of a given subject to belong to group number k and thus

$$\sum_{k=1}^{K} \pi_k = 1.$$

Since in practice, it is difficult to constraint the π_k to be numbers between 0 and 1, we link the π_k to a set of parameters $\theta_1, ..., \theta_K$ such that

$$\pi_k = \frac{e^{\theta_k}}{\sum_{k=1}^{K} e^{\theta_k}}.$$

The model depends thus on the parameter set $\psi = (K, \theta_1, \cdots, \theta_{K-1}, \Theta_1, \cdots, \Theta_K)$.

If we suppose moreover that the trajectories of Y are influenced by a static set of R risk variables $X = (X_1 \cdots X_R)$, as well as by a time-dependent covariate W which is independent of X, the conditional density of Y given X and W is given by

$$f(y_i|x_i, w_i) = \sum_{k=1}^{K} \left(P(G_i = k|X_i = x_i) \prod_{t=1}^{T} P\left(Y_{i_t} = y_{i_t}|X_i = x_i, W_i = w_i, G_i = k\right) \right), \tag{3}$$

which can be written as

$$f(y_i|x_i, w_i) = \sum_{k=1}^{K} \left(\sum_{j=1}^{R} \frac{e^{x_i^j \theta_k^j}}{1 + e^{x_i^j \theta_k^j}} \prod_{t=1}^{T} P\left(Y_{i_t} = y_{i_t}|W_i = w_i, G_i = k\right) \right). \tag{4}$$

Nagin (2005) defined this model for underlying logit, (censored) normal and zero inflated Poisson distributions. In this paper, we extend the model to an underlying Beta distribution.

2.2 The Beta distribution

The Beta distribution is quite useful for modelizing percentages and proportions, since it takes values between 0 and 1. Its density depends on the two positive shape parameters α and β. The primary advantage of this distribution is the flexibility of the shape of its density, as can be seen in Figure 1.

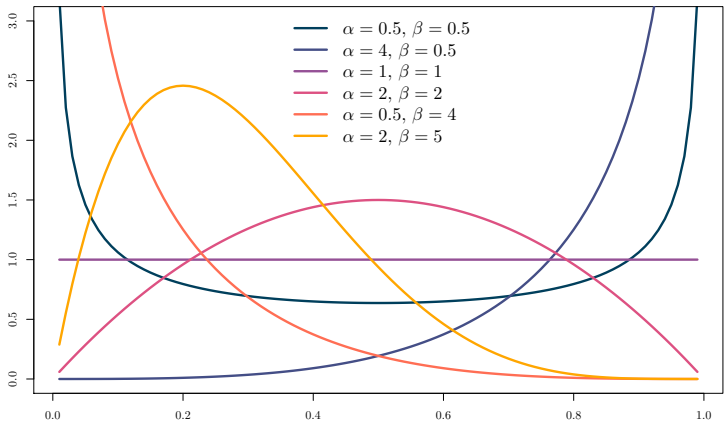


Figure 1 – *Example of different shapes of the Beta density for some parameters.*

In a regression context,another parametrization of the density is commonly used (Ferrari and Cribari-Neto 2004). Let Y be a random variable following a Beta distribution with mean μ. Consider the parameter ϕ defined by

$$var(Y) = \frac{\mu(1-\mu)}{1+\phi}.$$

ϕ can be interpreted as a precision parameter, in the sense that a large value of ϕ implies a small variance of Y. The density f of Y can be written as

$$f(y;\mu;\phi) = \frac{\Gamma(\phi)}{\Gamma(\mu\phi)\Gamma((1-\mu)\phi)} y^{\mu\phi-1}(1-y)^{(1-\mu)\phi-1},$$

where $0 < \mu < 1$ and $\phi > 0$.

2.3 Finite mixture models for the Beta distribution

For the finite mixture model with underlying Beta distribution, called here the BETA model, we consider a latent variable y_{it}^* such that

$$y_{it}^* = f(a_{it};\beta_k,\delta_k) + \varepsilon_{it} = \beta_k A_{it} + \delta_k W_t + \varepsilon_{it}, \tag{5}$$

where $\varepsilon_{it} \sim \mathcal{N}(0,\ \sigma_k)$, $A_{it} = (1, a_{it}, a_{it}^2, \cdots, a_{it}^{n_\beta - 1})^t$, $W_t = (w_{i1}, \cdots, w_{in_\delta})^t$, $\beta_k = (\beta_{k1}, \cdots, \beta_{kn_\beta})$ and $\delta_k = (\delta_{k1}, \cdots, \delta_{kn_\delta})$. n_δ denotes the dimension of the covariate W and n_β the number of measurements of Y for every individual. Here, and in the rest of this paper, we take into account that in some applications the data are actually given as a function of the age a_{it} of the subjects at the time of measurement.
The usual assumption is that $y_{it} = 1$ if $y_{it}^* > 0$ and $y_{it} = 0$ if $y_{it}^* \leq 0$.

The density of y_{it} conditional to membership in group C_k can then be written as

$$g_k(y_{it};\mu_{ikt},\phi_{ikt}) = \frac{\Gamma(\phi_{ikt})}{\Gamma(\mu_{ikt}\phi_{ikt})\Gamma((1-\mu_{ikt})\phi_{ikt})} y_{it}^{\mu_{ikt}\phi_{ikt}-1}(1-y_{it})^{(1-\mu_{ikt})\phi_{ikt}-1},$$

with

$$\mu_{ikt} = \frac{e^{\beta_k A_{it} + \delta_k W_{it}}}{1 + e^{\beta_k A_{it} + \delta_k W_{it}}} \text{ and } \phi_{ikt} = \zeta_k A_{it}. \tag{6}$$

One difficulty with the Beta distribution is that for some values of the parameters the density may converge to infinite in the neighborhood of 0 and 1. In case a big part of the data has values close to 0, it is advisable to transform the raw data y into $(y \cdot (n-1) + 0.5)/n$, where n is the sample size (Smithson and Verkuilen 2006).

To fit the model, we use the quasi Newton method BFGS, which calculates an iterative approximation of the inverse of the Hessian matrix, to maximize the log-likelihood of the data.

$$l = \sum_{i=1}^{n} \log\left(\sum_{k=1}^{K} \pi_k \prod_{t=1}^{T} \frac{\Gamma(\phi_{ikt})}{\Gamma(\mu_{ikt}\phi_{ikt})\Gamma((1-\mu_{ikt})\phi_{ikt})} y_{it}^{\mu_{ikt}\phi_{ikt}-1}(1-y_{it})^{(1-\mu_{ikt})\phi_{ikt}-1} \right). \tag{7}$$

3 The R package trajeR

The **R** package **trajeR** allows to calibrate finite mixture models for different types of densities (Noel and Schiltz 2022). The proofs of all used algorithms inside this package can be found in Noel's PhD thesis (Noel 2023). For an underlying Beta distribution, **trajeR** maximizes the log-likelihood by direct optimization of its derivative. The package **trajeR** is written in **R** and **C++**. The linkage between **R** and **C++** is achieved through the packages **Rcpp** (Eddelbuettel & François 2011) and **RcppArmadillo** (Eddelbuettel & Sanderson 2014).
Missing numbers are supposed to be missing at random by following Rubin's missing data

mechanisms (Rubin 1976).

The package **trajeR** is built around the core function `trajeR` which fits the model and computes its parameters for given degrees of the polynomial trajectories in the different groups. The function signature for `trajeR` is

```
R> trajeR(Y, A, Risk = NULL, TCOV = NULL, degre, degre.nu = 0,
+         Model, Method = "L",
+         ssigma = FALSE, ymax = max(Y) + 1, ymin = min(Y) - 1,
+         hessian = TRUE, itermax = 100, paraminit = NULL,
+         ProbIRLS = TRUE, refgr = 1,
+         fct = NULL, diffct = NULL, nbvar = NULL, nls.lmiter = 50)
```

Some of these arguments are mandatory others optional.

The mandatory arguments are the main data matrices `Y`, `A`, as well as `degre`, `Model` and `Method`.

Here `Y` is the matrix containing the values of the variable of interest and `A` is the matrix containing the age or time variable. In most applications, this matrix just contains times of measurement that are the same for each individual in the sample, implying that all lines of the matrix `A` are equal, but this is not necessarily the case. `A` can for instance contain the age of the different individuals at the times of measurement, which is generally different for each individual in the sample.

`degre` is a vector indicating the degree of the polynomials describing the typical trajectories in the different groups. Implicitly, the dimension of this vector also determines the number of groups into which we want to divide the population,

`Model` is a string defining the underlying distribution used in the model. The possible choices are LOGIT for the Logistic Regression Mixture Model, CNORM for the Censored Normal Mixture Model, ZIP for the Zero Inflated Poisson Mixture model and BETA for the BETA model.

`Method`, finally, is a string to decide which algorithm is used for estimating the model parameters. In case of the BETA model, only `L` for direct optimization is possible.

The optional arguments are `Risk`, `TCOV`, `degre.nu`, `ssigma`, `ymax`, `ymin`, `hessian`, `itermax`, `paraminit`, `ProbIRLS`, `refgr`, `fct`, `diffct`, `nls.lmiter`, `ng.nl` and `nbvar`.

`Risk` is a data matrix that contains the values of the covariate X modifying the group membership probability. By default, there is no such variable and `Risk` is a one-column matrix with value 1.

`ProbIRLS` allows to decide which method is used to compute the predictor probabilities. If its value is TRUE (default setting) we use the IRLS method and if it is FALSE we use the optimization method.

`TCOV` is an optional data matrix containing a time-dependent covariate W that influences the trajectories themselves. By default its value is NULL.

To ensure the identifiability of the parameters of the predictor, we have to fix a reference group. This can be done by the `refgr` command. It's default value is 1.

`hessian` indicates if we want to calculate the Hessian matrix, the default value being FALSE. If the method used is direct optimization, the Hessian matrix is computed by inverting the Fisher Information Matrix.

`itermax` gives the maximal number of iterations for the `optim` function or for the EM algorithm. The choice of the initial parameters is very important in optimization problems. We can specify these initial parameters by `paraminit`. By default `trajeR` calculates the initial value based on

the range or the standard deviation of the data (for the details, see Noel (2023)).

The output of `trajeR` for the BETA model is an object of class `Trajectory.BETA`, as described in following sections.

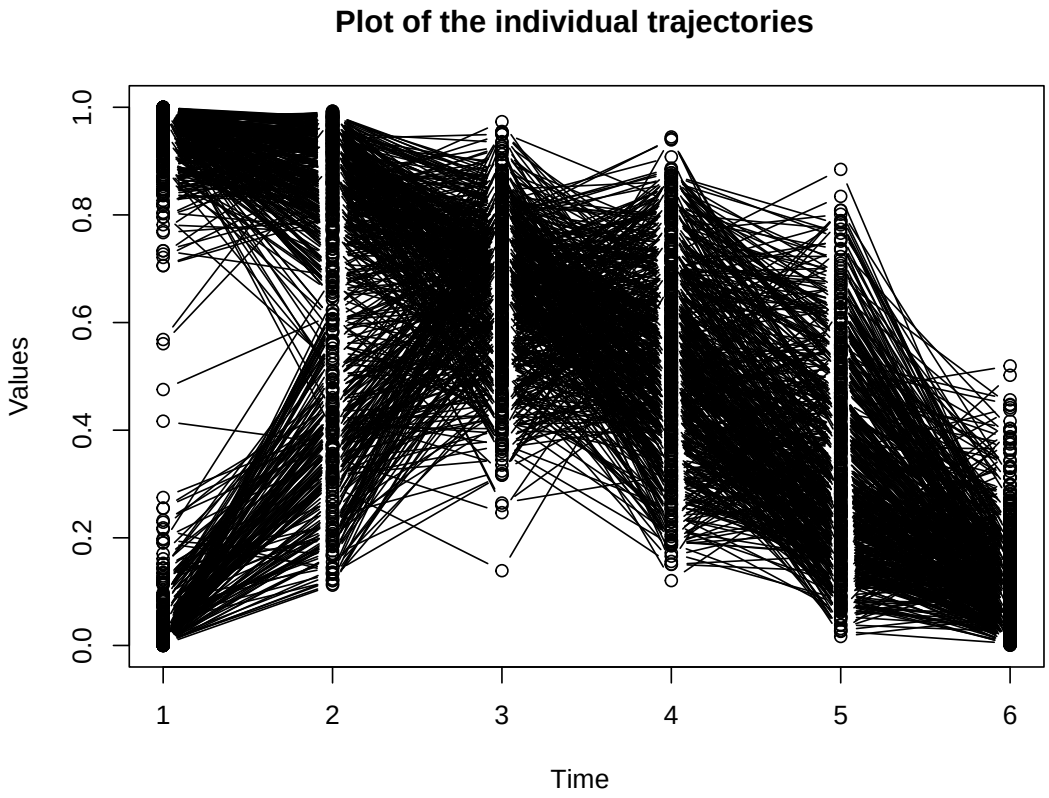
4 An example with simulated data

We use the simulated data set `BETA\_data01` that comes with installing the package from the CRAN repository. The sample consists of 500 trajectories with 6 time-points each, the values of the variable Y being between 0 and 1. The data set consists in a simulated 2 group solution, with respective group sizes are $\pi_1 = 0.65$ and $\pi_2 = 0.35$. In group one, the typical trajectory is a polynomial of degree 2 with parameters $\beta_1 = (-6, 3.7, -0.5)$, whereas in group 2, the typical trajectory is a polynomial of degree 1 with parameters are the parameters $\beta_2 = (3.8, -1)$.

The variable of interest Y_i is contained in `data[,2:7]`, the time variable A_i in `data[,8:13]`, the time-dependent covariate W, which could for instance indicate the presence of a characteristics of the individual, in `data[,15:20]` and a covariate X influencing group membership probabilty in `data[,14]`. Hence,

- `data[,2:7]` is a matrix with values between 0 and 1.
- `data[,8:13]` is a matrix with timepoints from 1 to 6.
- `data[,14]` is a column with real numbers.
- `data[,15:20]` is a matrix with real numbers.

We first plot the trajectories of Y to get a first impression of our sample.



To fit the model, we specify the number of group to two, ng=2, and fix the degree of the polynomial shape of the trajectories. Here we choose a line and a cubic polynomial, so degree is the vector $(1, 2)$.
We specify `hessian=TRUE` to ask the computation of the Hessian matrix.
In case of a Beta distributions, the computations are sensible to the starting values of the algorithm. By default, the starting value for the dispersion parameter is set to 5 for each group, which can give rise to numerical errors. So we use the `paraminit` parameter to manually set more sensible values.

Finally, to use the Likelihood method we call `trajeR` with option `Method ="L"`.

```
> paraminit = c(0, -0.9292991, 0, 0, 0.904195 , 0, 1, 0, 1, 0)
> solL = trajeR(Y = data[,2:7], A = data[,8:13],
+               param=paraminit,
+               degre=c(2,1), degre.phi = c(1,1),
+               Model = "BETA", Method = "L", hessian = T)
```

The basic results are contained in the object solL.

```
> solL

## trajeR with 2 groups with trajectories of degrees 2 and 1.
```

```
## Model : Beta
## Method : Likelihood
##
##   group   Parameter   Estimate   Std. Error   T for H0:   Prob>|T|
##                                                param.=0
## ------------------------------------------------------------------------
##    mean
##       1   Intercept   -5.95316      0.1281    -46.4734           0
##              Linear    3.66558      0.07649    47.92297          0
##           Quadratic   -0.49316      0.01027   -48.04232          0
##    var.
##       1   Intercept    2.26533      0.0993    22.81197           0
##              Linear   -0.00558      0.02466    -0.22636    0.82094
##
##    mean
##       2   Intercept    3.73504      0.04525    82.53444          0
##              Linear   -0.98061      0.01144   -85.70519          0
##    var.
##       2   Intercept    2.35458      0.07128    33.03302          0
##              Linear   -0.00144      0.01771    -0.08113    0.93534
## ------------------------------------------------------------------------
##       1         pi1      0.344      0.02069           0          0
##       2         pi2      0.656      0.02069    31.19708          0
## ------------------------------------------------------------------------
## Likelihood : 2516.737
```

We find that...

Next, we add a covariate that influences the group membership probabilities probability to the model. By default, the first group is the refernce group, meaning that the influence of the covariate is compared to the first group, but we can change this setting with the argument **refgr**.

```
> paraminit = c(0, 0, -0.9292991, 0, 0, 0.904195 , 0, 1, 0, 1, 0)
> solLRisk = trajeR(Y = data[,2:7], A = data[,8:13], Risk = data[,14],
+                   param=paraminit,
+                   degre=c(2,1), degre.phi = c(1,1),
+                   Model = "BETA", Method = "L", hessian = T)
```

```
> solLRisk We get the following results.

## Call TrajeR with 2 groups and a 2,1 degrees of polynomial shape of trajectory.
## Model : Beta
## Method : Likelihood
##
##   group   Parameter   Estimate   Std. Error   T for H0:   Prob>|T|
##                                                param.=0
## ------------------------------------------------------------------------
```

```
##    mean
##       1   Intercept   -5.95316      0.12845   -46.34479          0
##              Linear    3.66558      0.07666    47.81836          0
##           Quadratic   -0.49316      0.01031   -47.82021          0
##    var.
##       1   Intercept    2.26532      0.09986    22.68433          0
##              Linear   -0.00558      0.02483    -0.22484    0.82212
##
##    mean
##       2   Intercept    3.73504      0.04676     79.8723          0
##              Linear   -0.98061       0.0117   -83.81714          0
##    var.
##       2   Intercept    2.35458      0.07022    33.53206          0
##              Linear   -0.00144      0.01749    -0.08219     0.9345
## ------------------------------------------------------------------------
##       1    Baseline          0           NA          NA         NA
##
##       2   Intercept    0.64783      0.09442     6.86114          0
##                   X   -0.09072      0.05239     -1.7316    0.08345
## ------------------------------------------------------------------------
## Likelihood : 2518.219
```

We find...

Finally, we add a time-dependent covariate that influences the shape of the trajectories directly. We introduce their effects by using the option `TCOV` in the command `trajeR`.

```
> paraminit = c(0, -0.9292991, 0, 0, 0.904195 , 0, 1, 0, 1, 0, 0, 0)
> solLTCOV =  trajeR(Y = data[,2:7], A = data[,8:13], TCOV = data[,15:20],
+                    param=paraminit,
+                    degre=c(2,1), degre.phi = c(1,1),
+                    Model = "BETA", Method = "L", hessian = T)
```

We get the following results.

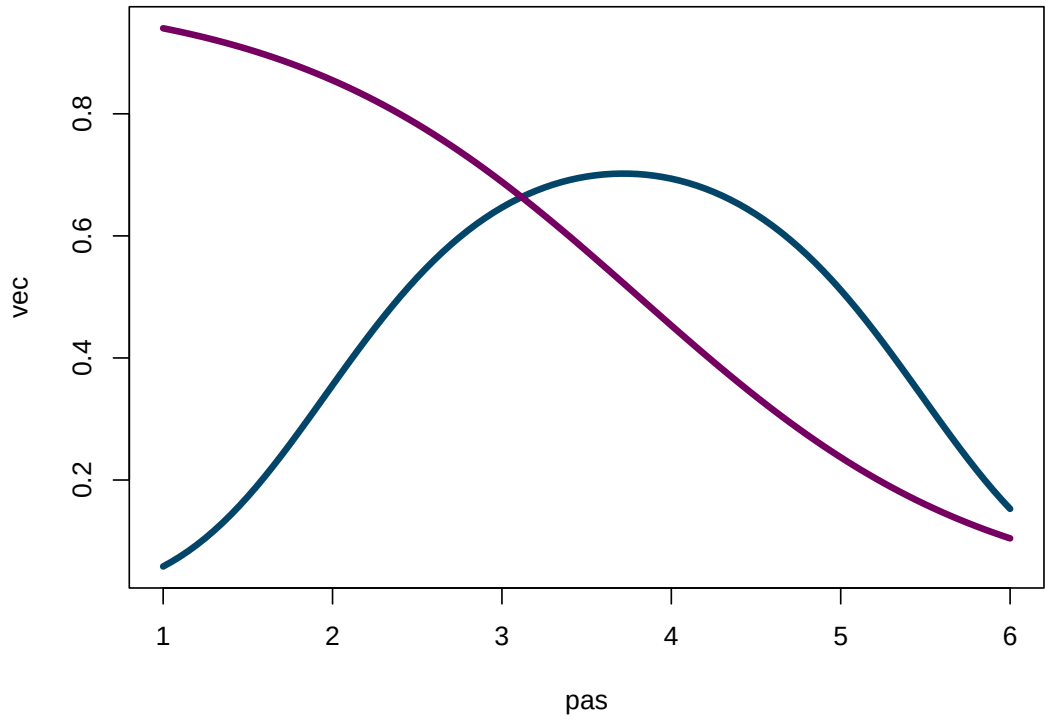
```
> solLTCOV

## Call TrajeR with 2 groups and a 2,1 degrees of polynomial shape of trajectory.
## Model : Beta
## Method : Likelihood
##
##   group   Parameter   Estimate   Std. Error   T for H0:   Prob>|T|
##                                                param.=0
## ------------------------------------------------------------------------
##    mean
##       1   Intercept   -5.95198       0.1294   -45.99705          0
##              Linear    3.66416       0.0771    47.52598          0
##           Quadratic   -0.49294      0.01036   -47.57542          0
##               TCOV1    -0.0081      0.01256    -0.64495    0.51901
##    var.
```

```
##       1   Intercept    2.26489      0.10084    22.46055          0
##              Linear   -0.00542      0.02503    -0.21647    0.82863
##
##    mean
##       2   Intercept    3.73513      0.04761    78.45198          0
##              Linear   -0.98071      0.01179    -83.1569          0
##               TCOV1   -0.00324      0.00919     -0.3524    0.72456
##    var.
##       2   Intercept     2.3544      0.07265    32.40833          0
##              Linear   -0.00137      0.01818    -0.07554    0.93979
## ------------------------------------------------------------------
##       1         pi1      0.344      0.02161           0          0
##       2         pi2      0.656      0.02161    29.86858          0
## ------------------------------------------------------------------
## Likelihood : 2517.012
```

We find...

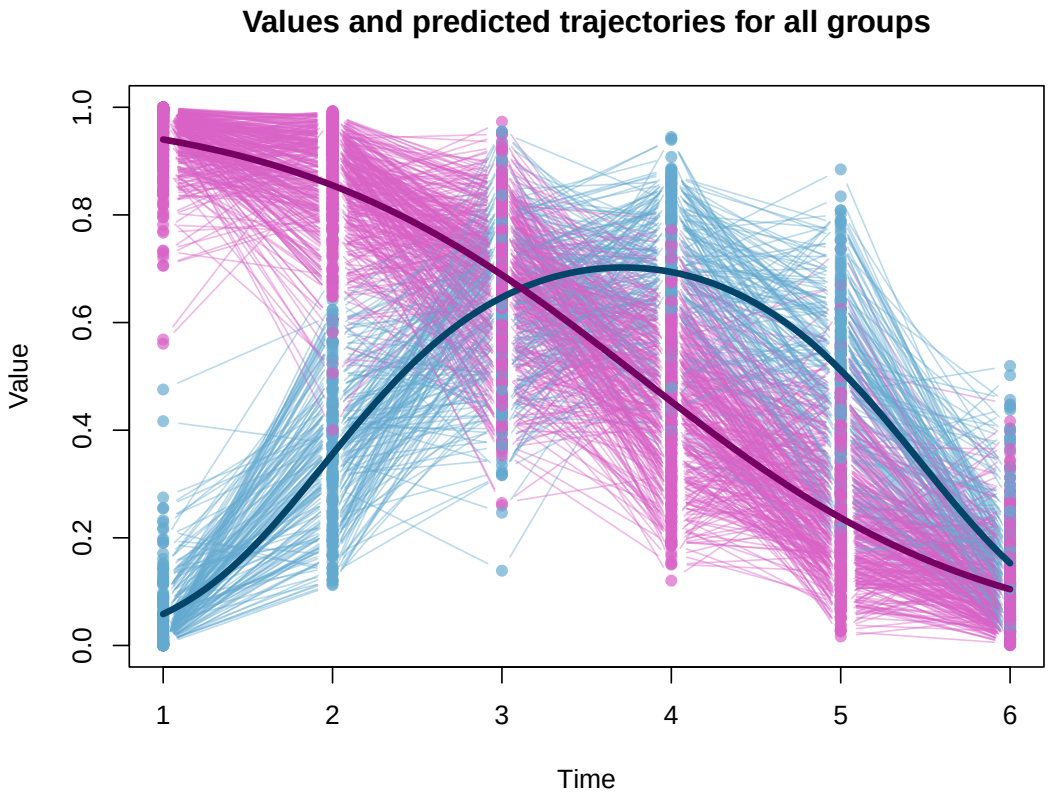
There are different possibilities to present the results graphically. The basic graph consists in plotting the typical trajectories of the different groups, which can be done with the command `plotrajeR` included in our **trajeR** package, applied to an object of class `Trajectory`. By default colors are gray scale, but we can specify any colors we want, for instance pink and light blue for the two groups in our example.

```
> plotrajeR(solLTCOV, col =
vcol)
```



But the graph looks much nicer if we add the initial longitudinal data. In order to achieve this, we have to specify the variables Y and A in the function `plot()`.

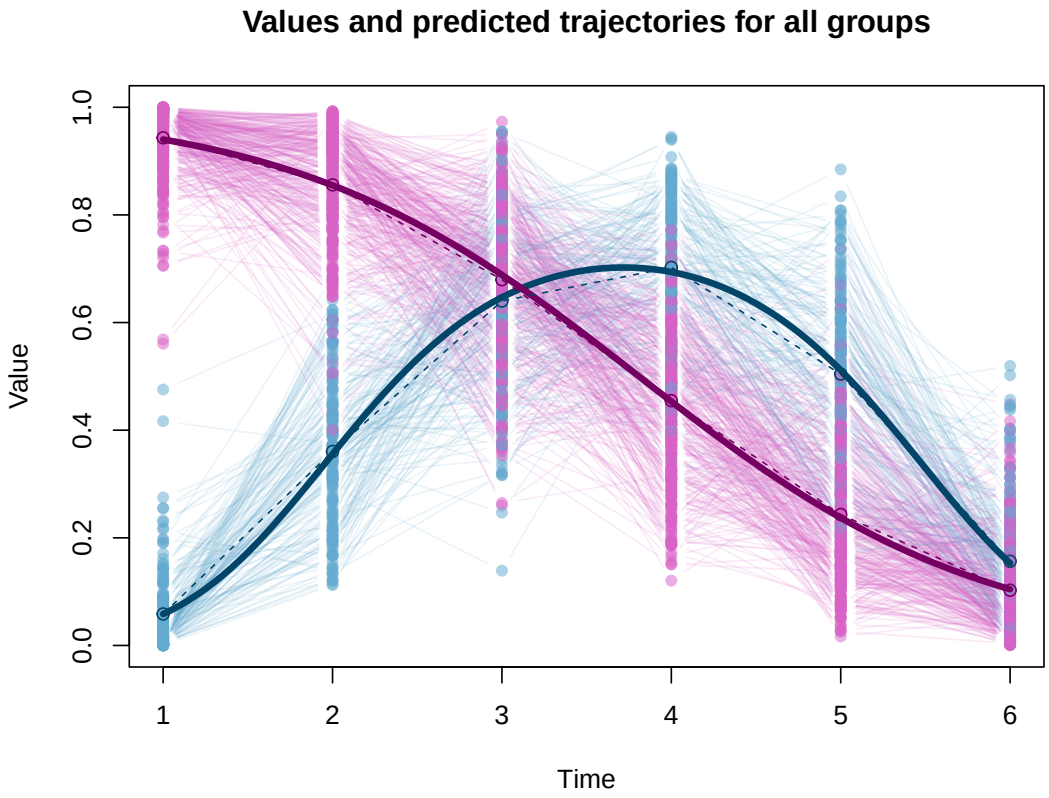
```
> # Defining the colours
> trans = "70"
> col1 = "#034569"
> col1.1 = paste0("#64AAD0", trans)
> col2 = "#750062"
> col2.1 = paste0("#D962C7", trans)
> cols1 = c(col1.1, col2.1)
> cols2 = c(col1, col2)
> vcol = c(cols1, cols2)
>
> plotrajeR(solLRisk, Y = data[,2:7], A = data[,8:13], col = vcol)
```



We can add the average of the data points on the plot for each time values. For each group and each time the mean of the data are calculated and add to the plot.

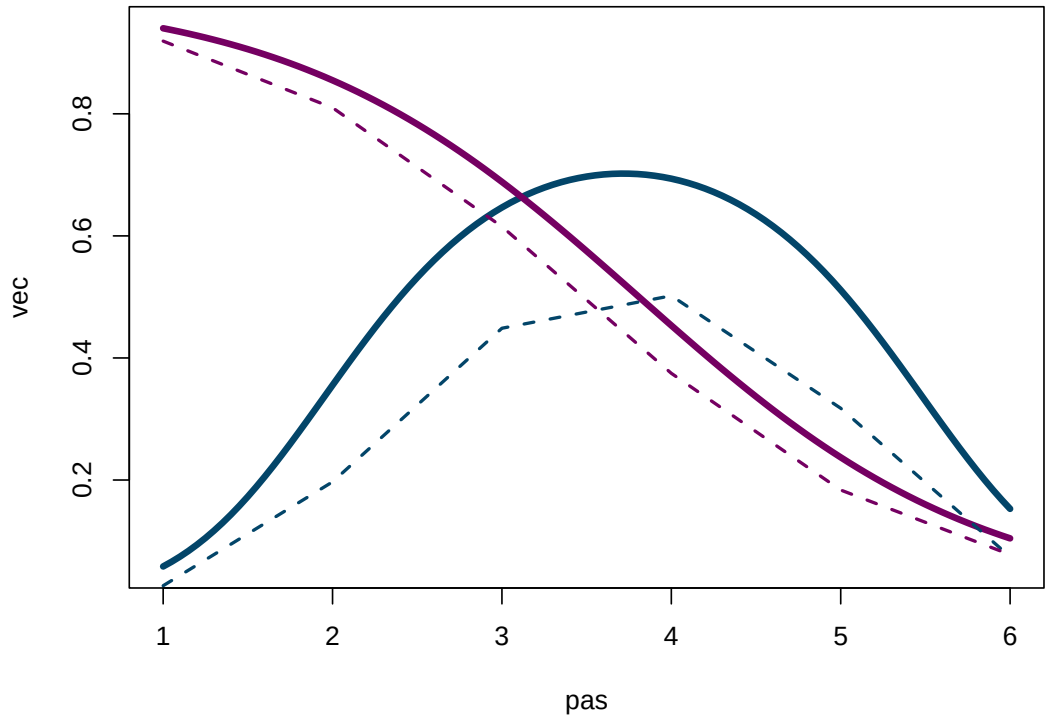
We use option `mean =TRUE`.

```
> # colour's defintion
> trans = "70"
> col1 = "#034569"
> col1.1 = paste0("#64AAD0", trans)
> col2 = "#750062"
> col2.1 = paste0("#D962C7", trans)
> cols1 = c(col1.1, col2.1)
> cols2 = c(col1, col2)
> vcol = c(cols1, cols2)
>
> plotrajeR(solLRisk, Y = data[,2:7], A = data[,8:13],
+       col = vcol, mean = TRUE, alpha = 0.3)
```



If we want show the impact of a particular value of the time covariate in the trajectory, we can add this to the plot by `plotcov` option.
The fill line is the trajectory with the time covariate matrix and th dashed one show the impact on this trajectory of a particular value.

```
> plotrajeR(solLTCOV, col = vcol, plotcov = c(100,100,100,100,100,100))
```



5 An application to COVID-19 data

5.1 Data

The data used in this article comes from the site "Our World In data". A complete description of this data can be found in Hasell and ali. (2020).
The data concerns 219 countries or groups of countries around the world. We remove this group, like North America, Europe, Africa, to keep only individual countries. We obtain 190 different countries. For each country, we are interested in new cases, population for million, total cases per million, media age, population's density, number of people over 65, an index of stringency, gpd per capita, an index of extreme poverty, the rate of cardiovascular death, the rate of prevalence of diabetes, an index of handwashing facilities, the rate of hospital's beds per thousand, the life expectancy and an index of human development.
We summarize all this covariates by month by taking the sum of new cases, the mean of the population by millions and by taking the last of the values for all others.
Some covariates contain missing values. We completed it : the median age, the density of the

population, the over 65[1], the gdp per capita[2] and the life expectancy[3]. Finally we replace the value 8???? for Luxembourg that is negative and we replace it by 0.

So we created an 16-period panel, starting at month January 2020 and finishing at month April 2021. Alvarez, Brida and Limas (2020), who studies longitudinal COVID data too, have created period panel by considering as start value the first case (or the tenth) of COVID in each country. Their choice is motivating by the fact that their time period is shorter than our (they consider daily data too and not monthly on a period of 100 days) and a shift can produce bad results. Our choice is to consider the same date as staring value. First our period is long, 14 months. Then an eventual shift will have less consequences on the cluster and the data are grouped by month, that limit the shift. Furthermore, in the data group by month, we van remark that the rate increase rapidly for each countries in some epoch. Second, it is very difficult to determine the beginning of the epidemic and in the data we can differences the NA values, there are no positive case or we are not make test. Moreover, the missing data are considering as missing at random, not as 0.

We observed the rate of contamination for a country i at time t, Y_{it} for $n = 190$ subjects at $T = 16$ time periods. $Y_i = (Y_{i1}, \cdots, Y_{iT})$ denotes the rate history for the country i. The figure 2 shows the longitudinal trajectories of rate for each of the 190 countries.

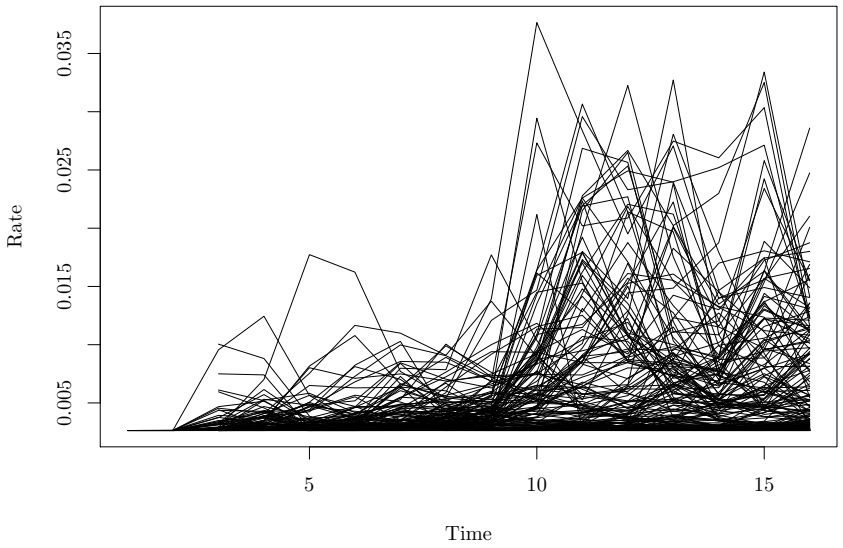


Figure 2 – *Rate of contamination for each countries.*

[1]https://www.nationmaster.com
[2]https://tradingeconomics.com
https://www.macrotrends.net
https://georank.org
[3]https://data.worldbank.org

5.2 Choice of the model

To find the model, we have to find first the number of groups and second, the parameters. Indeed, the number of groups is a parameter that is unknown. At the beginning, we don't know how many groups to take. To make the choice between different models, we use BIC criteria or AIC criteria or use a metric to compare more than 2 models provided by Kass and Wasserman (1995). Let p_k the probability that a model with k groups be the correct model from a set of K different models. They show that p_k is approximated by

$$\frac{e^{BIC_k - BIC_{max}}}{\sum_k e^{BIC_k - BIC_{max}}}$$

Furthermore, Nagin (2005) proposes different methods to check the accuracy of the model and in particular the Average Posterior Probability (AvePP). Ideally it would be 1, and Nagin considered that it should be at least 0.7 for all groups. Anyway, a model with group with AvePP 0 is a bad model.

We follow Nagin's procedure (Nagin 2005) to find the best model. First step : choice of the group, second step : choice of the degree of the polynomial shape. We could have followed the method resumed in (Nguena Nguefack and ali. 2020) too. It begins with 2 groups and finds the best model with BIC and adds group until it doesn't increase BIC. The problem here is the choice of starting point that encourages us to the first method. So, we fix the number of the polynomial for the mean to degree 3 to take account of, eventually, two points of inflection and for the precision parameter to degree 2.

Optimization methods suffer from choice of starting value. If it is bad value, the algorithm may not converge or converges to a local extremum. Unfortunately we have no method to choose this starting point. In a first time, our strategy is to choose, for starting value, k lines parallel to the axis x, i.e. the parameters $\beta_k = (\beta_0, 0, 0, \dots)$ and $\zeta_k = (\zeta_0, 0, 0, \dots)$ for $k = 2$ groups to 10. In a second time, we try some affine functions to take into account the growth of the data during the time period. Most of tries converge to 2 groups (row 1 of figure 4) and some to 3 groups (row 2 of figure 4). In a last time, we use the different groups found in the previous step for starting values to test 4 and 5 groups (rows 3 and 4 of figure 4).
Before using BIC or AIC to choose the model, we look if the models found in the previous step are well defined by using average posterior probability. We removed the model which some groups contains 0 individuals.

Finally, we find 7 models with 2 until 5, plot in the figure 4. To choose between them, we compute BIC, AIC and the probability p_k, see table 1. The greatest values of BIC and AIC are calculated for the model with 5 groups, BIC = 15558.41 and AIC = 31241.46. Then the probability for the 5 model is the greatest, 0.99999. The table 2 show the parameters of this model. We can see that for some parameters ϕ, the probabilities are zero and they doesn't be significant. So we can remove them. The final model is given in table 3. We can note that the trajectory of the parameters ϕ for the group 5 is a line parallel to the abscissa and so, for this group the precision is constant throughout times, while for the precision in this group, its trajectory becomes more heterogeneous during time.

5.3 Analysis of the model

The chosen model has 5 groups. Group 1, which can be called "growth" and contains 57 countries; group 2, "moderate growth" and contains 35 countries; group 3, "high growth" and contains 30

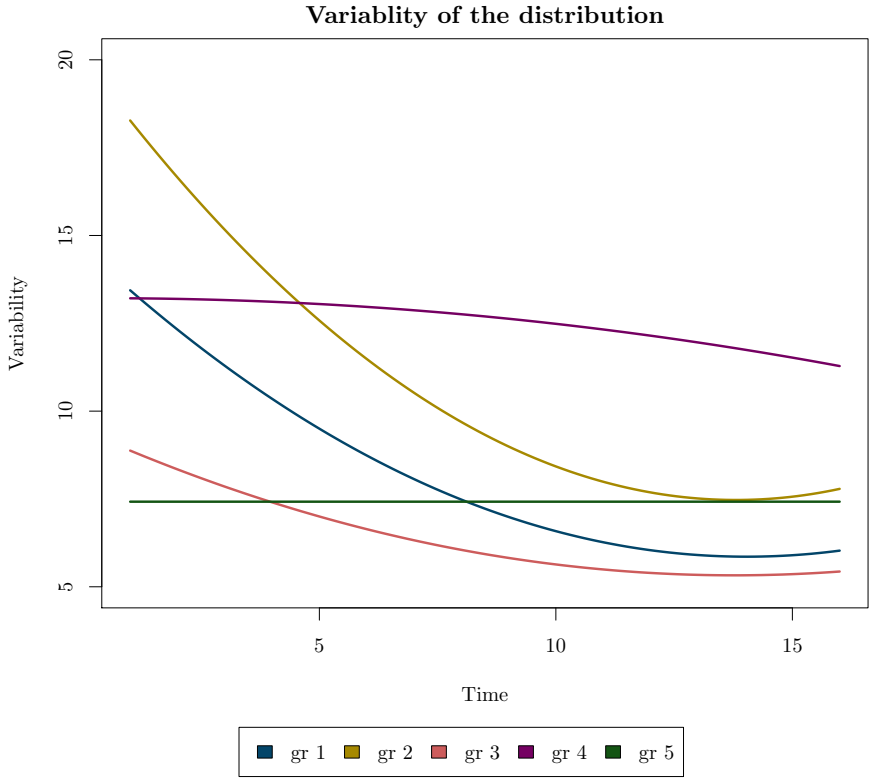


Figure 3 – *Growth of the precision parameters for each group.*

Number of groups	AIC	BIC	Prob
2	29851.99	14902.64	0.00000
3	30341.00	15142.28	0.00000
3	29945.96	14936.64	0.00000
3	30777.14	15352.23	0.00000
4	30839.69	15370.52	0.00000
4	31192.78	15547.06	0.00001
5	31241.46	15558.41	0.99999

Table 1 – *Some criteria computed to choose the right model*

countries; group 4, ”no contamination” and contains 61 countries and group 5, ”mountain growth” and contains 7 countries. The names of these countries can be found in table 4. Figure 5 shows the data divided in 5 groups following their own trajectory. These groups are presented on a world map, see figure 6. The model, in group 5, finds the particular behaviors of countries like Chile or Brazil which see the rate of contamination increase then decrease and finally increase again, certainly because of the presence of new variant described in Hojo de Souza et ali. (2021). Group 1, ”growth”, contained countries rate increase across time. Group 2, ”moderate growth”,

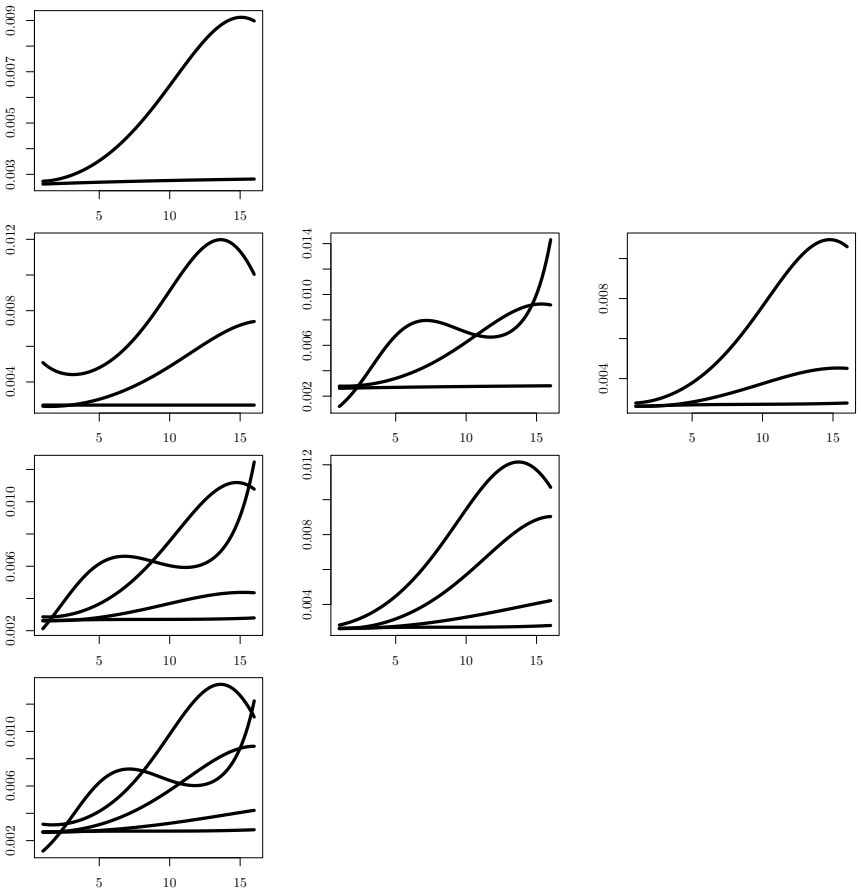


Figure 4 – *Different models found with some choice of starting values and number of groups.*

contained countries which increase slowly. Group 3, "high growth", contained countries rate increase quickly. Group 4, "no contamination", contained countries rate are almost zero.
The table 4 shows the principal characteristics of the different groups. The median age grows with the rate of contamination. The "no contamination" group has the smallest median age (23.49) while the "high growth" group has the highest median age (41.67). The same note can be make to the covariates rate of people aged 65 older or 70 older and life expectancy. On the other hand, we can note that the variability of the measures of stringency are almost the same for each group and the mean of these measures do not seem to correspond to a particular behavior linked to a group. For example, the "no contamination" group and the "high growth" are the lower (resp. 49.9 and 58.01), although the first one is further from the others than the second one.

Carrillo-Larco and Castillo-Cara (2020) use k-means to search some clusters but not taking account time series. They find 5 or 6 clusters that divide the data considering selected diseases, socio-economic status, air pollution and health system. Their groups show us a difference between the number of confirmed cases and we can see a lot of countries classify in the same group as us. Therefore we have more information as them with shape of trajectory of the rate.

Param.	sd	Test
Beta 1		
-5.902	0.018	0.000
-0.052	0.013	0.000
0.020	0.002	0.000
-0.001	0.000	0.000
Beta 2		
-5.927	0.003	0.000
-0.015	0.003	0.000
0.005	0.001	0.000
0.000	0.000	0.005
Beta 3		
-5.659	0.133	0.000
-0.119	0.066	0.071
0.040	0.009	0.000
-0.002	0.000	0.000
Beta 4		
-5.962	0.011	0.000
0.018	0.005	0.000
-0.002	0.001	0.000
0.000	0.000	0.000
Beta 5		
-7.511	0.370	0.000
0.911	0.143	0.000
-0.102	0.016	0.000
0.004	0.001	0.000

Param.	sd	Test
Phi 1		
14.648	0.299	0.000
-1.253	0.071	0.000
0.045	0.004	0.000
Phi 2		
20.026	0.354	0.000
-1.818	0.089	0.000
0.066	0.005	0.000
Phi 3		
9.461	0.377	0.000
-0.601	0.098	0.000
0.022	0.005	0.000
Phi 4		
13.212	0.367	0.000
0.007	0.086	0.935
-0.008	0.004	0.063
Phi 5		
7.713	1.021	0.000
-0.136	0.245	0.579
0.010	0.013	0.445

Param.	sd	Test
Probability 1		
0.301	0.035	0.000
Probability 2		
0.183	0.030	0.000
Probability 3		
0.160	0.028	0.000
Probability 4		
0.319	0.035	0.000
Probability 5		
0.036	0.013	0.005

Table 2 – *parameters of the 5 groups model*

Param.	sd	Test
Beta 1		
-5.902	0.018	0.000
-0.052	0.013	0.000
0.020	0.002	0.000
-0.001	0.000	0.000
Beta 2		
-5.927	0.003	0.000
-0.015	0.003	0.000
0.005	0.001	0.000
0.000	0.000	0.005
Beta 3		
-5.659	0.133	0.000
-0.119	0.066	0.071
0.040	0.009	0.000
-0.002	0.000	0.000
Beta 4		
-5.962	0.011	0.000
0.018	0.005	0.000
-0.002	0.001	0.000
0.000	0.000	0.000
Beta 5		
-7.511	0.370	0.000
0.911	0.143	0.000
-0.102	0.016	0.000
0.004	0.001	0.000

Param.	sd	Test
Phi 1		
14.648	0.309	0.000
-1.253	0.072	0.000
0.045	0.004	0.000
Phi 2		
20.026	0.280	0.000
-1.818	0.088	0.000
0.066	0.005	0.000
Phi 3		
9.454	0.374	0.000
-0.600	0.099	0.000
0.022	0.005	0.000
Phi 4		
13.212	0.360	0.000
0.007	0.085	0.934
-0.008	0.004	0.061
Phi 5		
7.422	0.146	0.000

Param.	sd	Test
Probability 1		
0.302	0.034	0.000
Probability 2		
0.183	0.028	0.000
Probability 3		
0.159	0.028	0.000
Probability 4		
0.319	0.034	0.000
Probability 5		
0.036	0.014	0.008

Table 3 – *parameters of the final model*

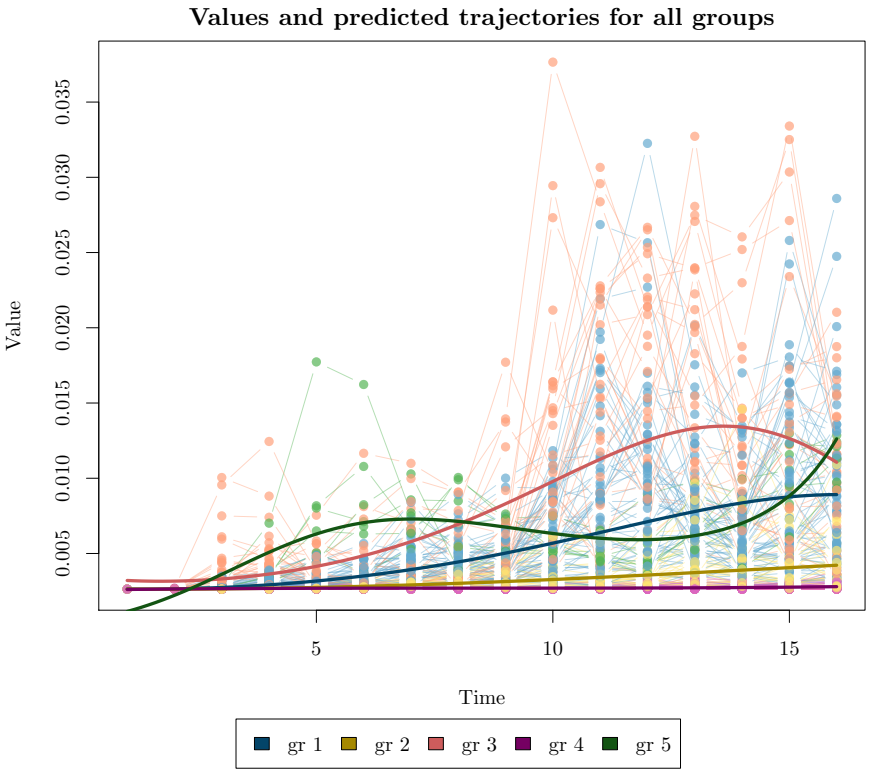


Figure 5 – *Curve of the trajectories of the 5 groups for the rate of contamination. The group 1 is called ?growth?, the group 2 is called ?moderate growth?, the group 3 is called ?high growth?, the group 4 is called ?no contamination? and the group 5 is called ?mountain growth?.*

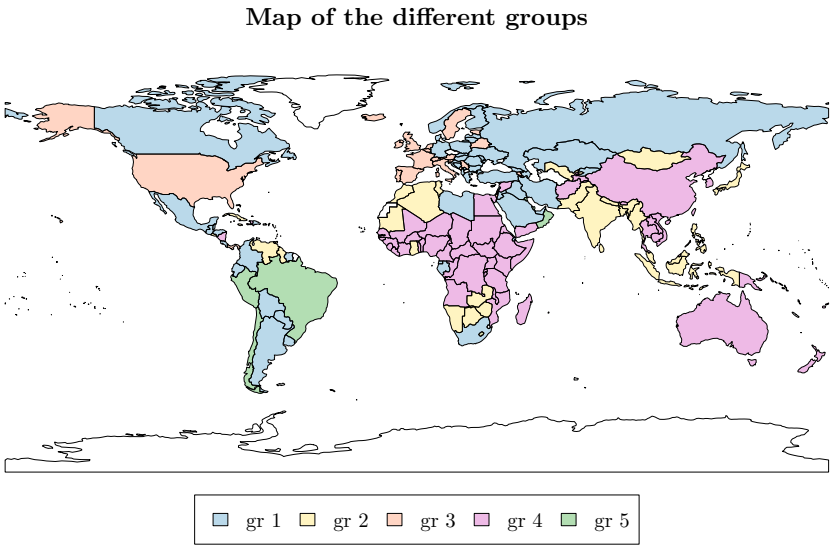


Figure 6 – *Colored countries by groups. The group 1 is called ?growth?, the group 2 is called ?moderate growth?, the group 3 is called ?high growth?, the group 4 is called ?no contamination? and the group 5 is called ?mountain growth?.*

Group 1	Albania, Argentina, Azerbaijan, Bahamas, Belize, Bolivia, Bosnia and Herzegovina, Bulgaria, Canada, Cape Verde, Colombia, Costa Rica, Croatia, Cyprus, Denmark, Djibouti, Dominican Republic, Ecuador, Equatorial Guinea, Finland, Gabon, Georgia, Germany, Greece, Guatemala, Honduras, Hungary, Iran, Iraq, Jordan, Kazakhstan, Kosovo, Kyrgyzstan, Latvia, Lebanon, Libya, Lithuania, Malta, Mexico, Moldova, North Macedonia, Norway, Palestine, Paraguay, Poland, Romania, Russia, Sao Tome and Principe, Saudi Arabia, Seychelles, Slovakia, South Africa, Suriname, Turkey, Ukraine, United Arab Emirates, Uruguay
Group 2	Algeria, Antigua and Barbuda, Bangladesh, Barbados, Botswana, Comoros, Cuba, El Salvador, Eswatini, Gambia, Ghana, Guyana, India, Indonesia, Jamaica, Japan, Lesotho, Malaysia, Mauritania, Mongolia, Morocco, Myanmar, Namibia, Nepal, Pakistan, Philippines, Saint Lucia, Saint Vincent and the Grenadines, Sri Lanka, Trinidad and Tobago, Tunisia, Uzbekistan, Venezuela, Zambia, Zimbabwe
Group 3	Andorra, Armenia, Austria, Bahrain, Belarus, Belgium, Czechia, Estonia, France, Iceland, Ireland, Israel, Italy, Liechtenstein, Luxembourg, Monaco, Montenegro, Netherlands, Panama, Portugal, San Marino, Serbia, Singapore, Slovenia, Spain, Sweden, Switzerland, United Kingdom, United States, Vatican
Group 4	Afghanistan, Angola, Australia, Benin, Bhutan, Brunei, Burkina Faso, Burundi, Cambodia, Cameroon, Central African Republic, Chad, China, Congo, Cote d'Ivoire, Democratic Republic of Congo, Dominica, Egypt, Eritrea, Ethiopia, Fiji, Grenada, Guinea, Guinea-Bissau, Haiti, Kenya, Laos, Liberia, Madagascar, Malawi, Mali, Marshall Islands, Mauritius, Micronesia (country), Mozambique, New Zealand, Nicaragua, Niger, Nigeria, Papua New Guinea, Rwanda, Saint Kitts and Nevis, Samoa, Senegal, Sierra Leone, Solomon Islands, Somalia, South Korea, South Sudan, Sudan, Syria, Taiwan, Tajikistan, Tanzania, Thailand, Timor, Togo, Uganda, Vanuatu, Vietnam, Yemen
Group 5	Brazil, Chile, Kuwait, Maldives, Oman, Peru, Qatar

Table 4 – *Names of countries belonging to each group.*

Table 5 – *Characteristic of the variable of each group*

Variables	Group 1 mean (sd)	Group 2 mean (sd)	Group 3 mean (sd)	Group 4 mean (sd)	Group 5 mean (sd)
population per million	19.97 (31.3)	76.41 (235.85)	23.38 (61.44)	48.49 (184.5)	39.64 (77.13)
median age	33.67 (7.76)	28.24 (6.88)	41.67 (5.29)	23.49 (7.51)	32.13 (2.19)
population density	122.67 (220.76)	207.31 (242.47)	1148.75 (3729.09)	126.64 (153.86)	286.19 (524.35)
aged 65 older	10.54 (6.22)	6.75 (4.69)	17.08 (5.38)	4.83 (3.3)	5.27 (3.7)
aged 70 older	6.87 (4.25)	4.16 (3.21)	10.76 (3.43)	2.81 (2.13)	3.23 (2.34)
gdp per capita	20583.62 (14511.5)	10267.22 (8325.13)	50138.17 (41050.34)	7655.52 (12468.09)	40673.95 (38570.57)
extreme poverty	3.63 (6.68)	12.96 (16.69)	0.81 (0.67)	31.28 (24.1)	2.73 (1.24)
cardiovasc death rate	259.68 (110.33)	285.53 (113.24)	172.99 (108)	306.11 (121.94)	161.7 (56.6)
diabetes prevalence	8.35 (3.3)	7.75 (3.63)	6.93 (2.84)	7.72 (5.59)	10.95 (4.08)
handwashing facilities	72.84 (22.95)	55.53 (28.91)	95.88 (2.6)	32.5 (27)	96.6 (1.13)
hospital beds per thousand	3.63 (2.18)	2.47 (2.5)	4.45 (2.67)	1.87 (2.1)	1.79 (0.38)
life expectancy	75.2 (4.56)	70.73 (6.42)	80.97 (3.16)	67.18 (7.41)	77.9 (1.95)
human development index	0.79 (0.09)	0.68 (0.1)	0.9 (0.05)	0.59 (0.14)	0.8 (0.04)
Mean of stringency	64.75 (10.3)	64.1 (10.66)	58.01 (11.2)	49.9 (15.28)	68.72 (2.95)
St. dev. of stringency	16.98 (5.82)	15.74 (7.22)	18.17 (5.2)	16.22 (7.07)	22.02 (4.96)

5.4 Explanatory variables

We tested several covariates to understand if some of them influenced the group's membership probability. We tested the median age, the population density, the rate of people aged 65 older, the gdp per capita, the life expectancy and to catch the effect of the measures of stringency, the mean of the mean and the standard deviation of stringency during time period for each country inside a group. The mean of the standard deviation is used to catch some variations around the mean and some different country's strategy of stringency inside a group. Classically, in a multinational regression, we have to use one group as basis. We chose to use the group 4, **?**no contamination**?** to analyze another one. After having fit the model, we note that the GPD is linked to any group. So, we removed it to the analysis.

Table 6 resumes the estimation of the parameters fit by the model and the probability.
The median age is linked to groups "growth" ($p = 0.02445$) and "up and down" ($p = 0.04142$), the density is very lightly bound with the group "growth" ($p = 0.09272$), the 65 old is linked to "up and down" group ($p = 0.00188$), the life expectancy with groups "high" ($p = 0.00388$) and "up and down" ($p = 0.02254$)and the mean of stringency is linked to all groups (in order $p = 0,\ 0.00008,\ 0.00058,\ 0.00228$).

For each group, the mean of stringency differs from the base group. To understand if it differs between groups, we use a χ^2-based test of multiple contrasts. The degree of freedom of this test equals the number of equality constraints being tested or the number of different coefficients being tested minus 1. Here, the degree of freedom is 3.

We tested if each coefficient, noted θ_i^{st}, of mean of stringency differs from one group to another. Let

$$H_0 :\ \theta_2^{st} = \theta_3^{st} \text{ and } \theta_2^{st} = \theta_4^{st} \text{ and } \theta_2^{st} = \theta_5^{st} \tag{8}$$
$$H_1 :\ \theta_2^{st} \neq \theta_3^{st} \text{ or } \theta_2^{st} \neq \theta_4^{st} \text{ or } \theta_2^{st} \neq \theta_5^{st} \tag{9}$$

Let denoted

$$\theta = \left(\theta_2^{st}, \theta_3^{st}, \theta_4^{st}, \theta_5^{st}\right)'$$

We tested the hypothesis above in the form

$$H_0 :\ H\theta = q \tag{10}$$
$$H_1 :\ H\theta \neq 0 \tag{11}$$

where $q = (0,0,0)'$ and $H = \begin{pmatrix} 1 & -1 & 0 & 0 \\ 1 & 0 & -1 & 0 \\ 1 & 0 & 0 & -1 \end{pmatrix}$.

The χ^2 test is computed by

$$\chi^2 = \left(H\hat{\theta} - q\right)' \left(HV_{\hat{\theta}}H'\right)^{-1} \left(H\hat{\theta} - q\right) \tag{12}$$

where $\hat{\theta}$ is the counterpart vector of estimations and $V_{\hat{\theta}}$ is the matrix variance/covariance of the parameter estimated in $\hat{\theta}$.

By applying this formula to the estimations of the mean of stringency, we obtain $\chi^2 = 5.62$ with 3 degree of freedom which is far short of significance. This result implies that the hypothesis the mean of stringency has no differential impact across the rate of contamination trajectory is supported. Thus, while the results in table 6 show that high stringency is linked to the rate of contamination, the χ^2 based statistic implies that high stringency does not distinguish the specific developmental course of rate. Presumably, we can suppose that the mean of stringency is a consequence of the rise of the rate of contamination, these countries try to stem the pandemic and it is not responsible for the shape of the rate.
The groups "high" and "up and down" are linked to life expectancy which is the two oldest ages (80.97 and 77.9) and the group "growth" is linked to the median age which is the second oldest (33.67). Table 6 shows the boxplot of the median age for each group. We can note that the median age is relatively important for almost all the individuals in this group. We can conclude that the shape of the rate is link to the age variable, and more particularly, older is the age, higher is the level of contamination.

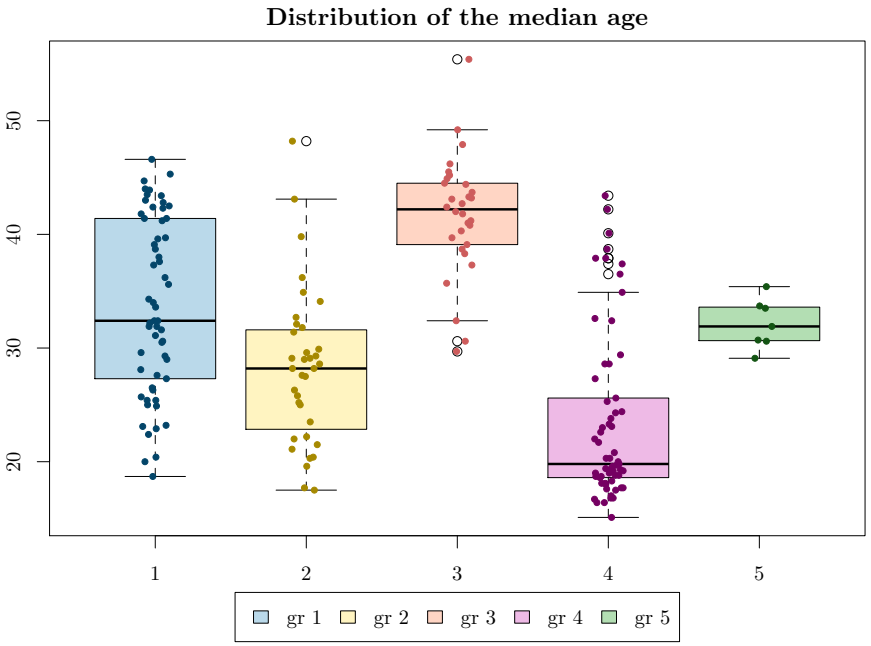


Figure 7 – *Distribution of the median age of each group.*

5.5 Stringency as time dependent covariate

GBTM allows us to use some covariates which depend on time, as explanatory variables that can modify the shape of the trajectory. It is the part W_{it} in equation (6). We can wonder if the different measures of stringency, taken by the different countries, have influenced the shape of the trajectory of the rate of the contamination and, in particular, if this curve was flatten by these measures.
Thus, W_{it} is composed by the measures at time t, for $1 \leq t \leq 16$, of the stringency for the country

	Group 1			Group 2		
	Estimate	Std. Error	Prob>|T|	Estimate	Std. Error	Prob>|T|
intercept	-16.812	4.681	0	-4.805	3.422	0.16
median age	0.193	0.086	**0.024**	0.172	0.101	**0.088**
population density	-0.003	0.002	0.093	0.000	0.001	0.869
aged 65 older	-0.021	0.132	0.871	-0.060	0.126	0.631
life expectancy	0.073	0.080	0.364	-0.073	0.071	0.304
mean of stringency	0.112	0.023	**0**	0.092	0.023	**0**
	Group 3			Group 5		
	Estimate	Std. Error	Prob>|T|	Estimate	Std. Error	Prob>|T|
intercept	-67.733	19.400	0	-73.689	23.469	0.002
median age	0.129	0.158	0.412	0.418	0.205	**0.041**
population density	0.000	0.001	0.784	0.000	0.001	0.926
aged 65 older	0.109	0.178	0.542	-0.640	0.206	**0.002**
life expectancy	0.646	0.223	**0.004**	0.646	0.283	**0.023**
mean of stringency	0.185	0.054	**0.001**	0.228	0.075	**0.002**

Table 6 – *Estimation of the parameters of the covariate.*

i. We removed from the data the countries with no stringency value [4] and we completed the eventual missing values for the others by linear interpolation.

The table (7) shows the parameters of the model. Their interpretation is the same as above except for delta which refers to the stringency time covariate. For the groups with high rate, the values of delta parameters are positive while, for the other, they are null or almost. Therefore, a high value of stringency implies a raise of the trajectory.

[4] Antigua and Barbuda , Armenia , Comoros , Equatorial Guinea , Grenada , Guinea-Bissau , Liechtenstein , Maldives , Marshall Islands , Micronesia (country) , Montenegro , North Macedonia , Saint Kitts and Nevis , Saint Lucia , Saint Vincent and the Grenadines , Samoa , Sao Tome and Principe , Vatican.

Param.	sd	Test	Param.	sd	Test	Param.	sd	Test	Param.	sd	Test
Beta 1			**Phi 1**			**Delta 1**			**Prob. 1**		
-5.843	0.026	0.000	14.337	0.317	0.000	0.001	0.000	0.001	0.328	0.039	0.00
-0.120	0.024	0.000	-1.164	0.076	0.000				**Prob. 2**		
0.029	0.004	0.000	0.040	0.004	0.000	**Delta 2**			0.175	0.030	0.00
-0.001	0.000	0.000	**Phi 2**			0.000	0.000	0.955	**Prob. 3**		
			19.866	0.570	0.000				0.156	0.030	0.00
Beta 2			-1.710	0.125	0.000	**Delta 3**			**Prob. 4**		
-5.927	0.003	0.000	0.061	0.006	0.000	0.010	0.001	0.000	0.301	0.035	0.00
-0.014	0.004	0.000	**Phi 3**						**Prob. 5**		
0.005	0.001	0.000	9.624	0.369	0.000	**Delta 4**			0.040	0.016	0.01
0.000	0.000	0.001	-0.521	0.097	0.000	0.000	0.000	0.000			
			0.016	0.005	0.003						
Beta 3			**Phi 4**			**Delta 5**					
-5.602	0.117	0.000	12.887	0.372	0.000	0.004	0.001	0.004			
-0.421	0.070	0.000	0.148	0.085	0.082						
0.076	0.009	0.000	-0.015	0.004	0.000						
-0.003	0.000	0.000									
			Phi 5								
Beta 4			7.384	0.137	0.000						
-5.972	0.012	0.000									
0.012	0.005	0.018									
-0.001	0.001	0.043									
0.000	0.000	0.027									
Beta 5											
-7.304	0.366	0.000									
0.701	0.147	0.000									
-0.078	0.017	0.000									
0.003	0.001	0.000									

Table 7 – *parameters of the final model with time dependent covariates.*

In figure (8), we plot an example of the influence of the stringency. The plain line is the trajectory when the stringency is zero during all time periods while for the dotted line, the stringency is 100, the maximal value. We can see, for the 3 groups with high contamination rate, that highest is the stringency, highest is the trajectory. For the group with low contamination rate, the shape is unchanged.

We retrieve the conclusion that stringency is linked to the contamination rate and that it is a consequence of the raise of the contamination.

6 Conclusion

In this paper, we presented an extension of finite mixture models to Beta distributed data and showed how to calibrate this kind of models with our R package trajeR. Further extensions of the package will allow to handle multitrajectory models, as well as incorporate different strategies to handle missing data, which is quite important in some research areas like medicine, where missing data often cannot be assumed to happen randomly (there are for instance patient dropouts, due to the death of patients).
In a second part of the paper, we used the before introduced model to discover and validate five groups of countries that follow distinct rate of contamination trajectories during the first phase of the COVID-19 pandemics. We found a significant difference in the treatment between countries with no contamination, and the ones that can be interpreted like a response to the

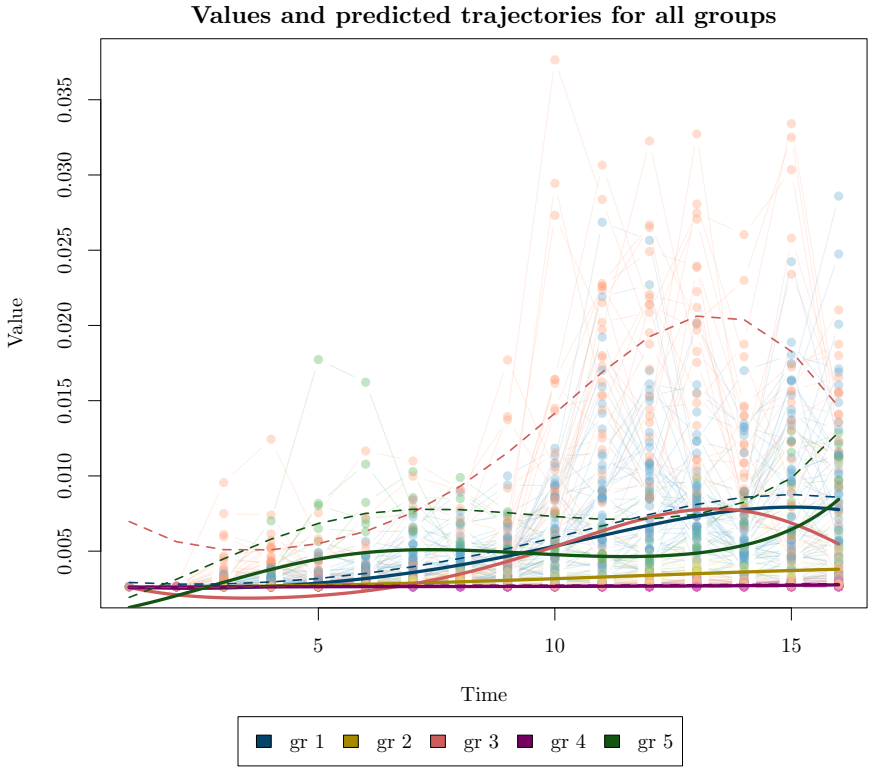


Figure 8 – *Curve of the trajectories of the 5 groups for the rate of contamination with stringency as time dependent covariate. The dotted line is the trajectory when the stringency is maximal, i.e. 100, while the plain line is the trajectory when the stringency is zero.*

growth of the pandemic and a way to contain it. We also found a significant link between the age of the population and the rate of contamination. In conclusion, this study seems to show that among the analyzed factors the only which influences the rate of contamination is the age of the population. Specifically, the level of stringency of the different measures adapted by the countries does not influence the contamination rate. In the same way, in their meta analysis, Herby, Jonung and Hanke (2022) conclude too that the lockdowns have not a large effect on the mortality rate.
Of course, this conclusion must be seen cautiously. First of all, the group-based model trajectory method is more accurate with a lot of individuals and the number of countries is low. Second, the beta distribution is sensible to values which are close to 0 and 1, which is the case here. Nevertheless, it remains a means of trying to understand the evolution of the epidemic.

In further works, it would be interesting to include other variable in these studies, like, for example, the presence of variants. It would also be interesting to study more finely the impact of the measures of stringency on the trajectory, particularly by using its longitudinal reporting and the dual trajectory model (METTRE REF VERS article dual). Finally, it would by interesting to study the second wave of the pandemic with several new variants, and by including the rate of vaccination in each countries.

References

[1] Alvarez, E. Brida, J.G. & Limas, E. (2020). Comparisons of COVID-19 dynamics in the different countries of the World using Time-Series clustering. medRxiv preprint.

[2] Berlin, K.S., Parra, G.R. & Williams, N.A. (2014). *An introduction to latent variable mixture modeling*. Journal of Pedritic Psychology, **39 (2)**, 174-203.

[3] Carrillo-Larco, R.M. Castillo-Cara, M. (2020). Using country-level variables to classify countries according to the number of confirmed COVID-19 cases: An unsupervised machine learning approach. Wellcome Open Research, **5**, 56.

[4] Chandu, V.C. (2020). Identification of spatial variations in covid-19 epidemiological data using k-means clustering algorithm: a global perspective. medRxiv preprint. t doi:https://doi.org/10.1101/2020.06.03.20121194.

[5] Ferrari, S.L.P. & Cribari-Neto, F. (2004). *Beta Regression for Modelling Rates and Proportions*. Journal of Applied Statistics. **31 - 7**, 799-815.

[6] Greene, W.H. (2018). *Econometric Analysis*. 8th edition. New York: Pearson.

[7] Hasell, J. Mathieu, E. Beltekian, D. Macdonald, B. Giattino, C., Ortiz-Ospina, E. Roser, M. & Ritchie, H. (2020). *A cross-country database of COVID-19 testing*. Scientific Data, **7**, 345.

[8] Herby, J. Jonung, L. & Hanke, S.H. (2022). *A Literature Review and Meta-Analysis of the Effects of Lockdowns on COVID-19 Mortality*. Studies in Applied Economics, **200**. John Hopkins Institute for Applied Economics.

[9] Hojo de Souza, F.S. Satchiko Hojo-Souza, N. Maciel da Silva C. & Guidoni, D.L. (2021). *Second wave of COVID-19 in Brazil: younger at higher risk*. European Journal of Epidemiology, **36**, 441-443.

[10] James, N. & Menzies, M. (2020). *Cluster-based dual evolution for multivariate time series: Analyzing COVID-19*. Chaos, **30**, 061108.

[11] Jones, B.L. & Nagin, D.S. (2007). *Advances in Group-Based Trajectory Modeling and an SAS Procedure for Estimating Them*. Sociological Methods & Research, **Vol. 35 No.4**, 542-571.

[12] Jones, B.L. Nagin, (2012). *A Stata Plugin for Estimating Group-Based Trajectory Models*. Heinz College Research Working Paper.

[13] Jones, B.L. Nagin, D.S. & Roeder, K. (2001). *A SAS Procedure Based on mixture Models for Estimating Developmental Trajectories*. Sociological Methods & Research, **Vol. 29 No.3**, 374-393.

[14] Kass, R.E. & Wasserman, L. (1995). *A Reference Bayesian Test for Nested Hypotheses and Its Relationship to the Schwarz Criterion*. Journal of the American Statistical Association, **90 No. 431**, 928-934.

[15] Kinnunen, J. Georgescu, I. Hosseini, Z. & Androniceanu, A.M. (2021). *Dynamic indexing and clustering of government strategies to mitigate Covid-19*. Entrepreneurial Business and Economics Review, **9 No.2**,

[16] Laursen, B.P. & Hopf, E. (2006). *Person-centered and variable centered approaches to longtitudinal data*. Merrill-Palmer Quaterly, **52 (3)**, 377-389.

[17] McLachlan, G. & Peel, D. (2000). *Finite mixture models*. New York: Wiley.

[18] Muthén, B.O., & Shedden, K. (1999). Finite mixture modeling with mixture outcomes using the EM algorithm. *Biomectrics*, 55, 463-469.

[19] Nagin, D.S. (1999). *Analyzing Developmental Trajectories: Semi-parametric. Groupe-based Approach*. Psychological Method, **4**, 139-157.

[20] Nagin, D.S.(2002). *Analyse des trajectoires de développement: vue d'ensemble d'une méthode semiparamétrique fondée sur le groupement*. Recueuil du Symposium de Statistique Canada Modélisation des données d'enquête pour la recherche sociale et économique.

[21] Nagin, D.S. (2005). *Group-Based Modeling of Development*. Cambridge. Massachusets: Harvard University Press.

[22] Nagin, D.S. & Land, K.C. (1993). *Age, criminal careers and population heterogeneity: Specifiction and estimation of a nonparametric, mixed Poisson model*. Criminology, **31**, 327-362.

[23] Nagin, D.S. & Odgers, C.L. (oldstylenums2010). *Group-Based Trajectory Modeling (Nearly) Two Decades Later*. Journal of Quantitative Criminology, **26**, 445-453.

[24] Nesselroade, J.R. (1991). Interindividual differences in intraindividual change. In L.A. Collins & J.L. Horn (Eds.), *Best Methods for the Analysis of Change* (pp.92-106). Washington, DC: American Psychological Association.

[25] Nguena Nguefack, H.L., Pagé, M.G. Katz, J. Choinière, M. Vanasse, A. Dorais, M. Mallé Samb, O. & Lacasse, A. (2020). *Trajectory Modelling Techniques Useful to Epidemiological Research: A Comparative Narrative Review of Approaches*. Clinical Epidemiology, **12**, 1205-1222.

[26] Noel, C. (2023). *On a generalisation of Nagin's finite mixture model*. PhD Thesis. University of Luxembourg. Luxembourg.

[27] Noel, C. & Schiltz, J. (2022) *trajeR, an R package for cluster analysis of time series*. Working Paper. University of Luxembourg. Luxembourg.

[28] Rahman, M.A., Zaman, N., Asyhari, A.T., Al-Turjman, F., Alam Bhuiyan, M.Z. & Zolkipli M.F. (2020). *Data-driven dynamic clustering framework for mitigating the adverse economic impact of Covid-19 lockdown practices*. Sustainable Cities and Society, **62**, 102372.

[29] Rubin, D.B. (1976). *Inference and Missing Data*. Biometrika, **63-3**, 581-592.

[30] Schiltz, J. (2015). *A generalization of Nagin's finite mixture model*. In: M. Stemmler, A. Von Eye & W. Wiedermann (eds.) *Dependent Data in Social Sciences Research*. New York: Springer. 107-126.

[31] Smithson, M. & Verkuilen, J. (2006). *A Better Lemon Squeezer? Maximum-Likelihood Regression With Beta-Distributed Dependent Variables*. Psychological Methods, **11 No.1**, 54-71.

[32] Spiegelhalter, D. & Masters, A. (2021). Covid by Numbers. UK: Penguin.

[33] van der Nest, G. Lima Passos, V., Candel, M.J.J.M. & van Breukelen, G.J.P. (2020). *An overview of mixture modelling for latent evolutions in longitudinal data: Modelling approaches, fit statistics and software*. Advances in Life Course Research, **43**, 100323.

[34] Verbeken, G. Fieuws, S. & Molenberghs, G. (2014). *The analysis of multivariate longitudinal data: a review*. Statistical Methods in Medical Research, **2014 (1)**, 42-59.

[35] Zarikas, V. Poulopoulos, S.G., Gareiou, Z. & Zervas, E. (2020). *Clustering analysis of countries using the covid-19 cases dataset*. Data in Brief, **31**, 105787.