A new R package for Finite Mixture Models with an application to clustering countries with respect to COVID data

Jang SCHILTZ (University of Luxembourg)

joint work with

Cédric NOEL(University of Lorraine & University of Luxembourg)

SMTDA 2022
June 10, 2022

Outline

Outline

Outline

Outline

General description of Finite Mixture models

We have a collection of individual trajectories.

We try to divide the population into a number of homogenous sub-populations and to estimate, at the same time, a typical trajectory for each sub-population. (Nagin 2005, Schiltz 2015)

This model can be interpreted as functional fuzzy cluster analysis.

Aim of the Analysis (underlying normal distribution)

Find K groups of trajectories of a given kind, for instance polynomials of degree 4, $P(t) = \beta_0 + \beta_1 t + \beta_2 t^2 + \beta_3 t^3 + \beta_4 t^4$.

Statistical Model:

$$y_{i_t} = \beta_0^k + \beta_1^k t + \beta_2^k t^2 + \beta_3^k t^3 + \beta_4^k t^4 + \varepsilon_{i_t}^k, \qquad (1)$$

where $\varepsilon_{i_t}^k \sim \mathcal{N}(0, \sigma_k)$, σ_k being the standard deviation, constant inside group k.

We try to estimate a set of parameters $\Omega = \left\{\beta_0^k, \beta_1^k, \beta_2^k, \beta_3^k, \beta_4^k, \pi_k, \sigma_k\right\}$ which allow to maximize the probability of the measured data.

Predictors of trajectory group membership

x : vector of variables potentially associated with group membership (measured before t_1).

Multinomial logit model:

$$\pi_k(x_i) = \frac{e^{x_i \theta_k}}{\sum_{k=1}^{K} e^{x_i \theta_k}}, \tag{2}$$

where θ_k denotes the effect of x_i on the probability of group membership for group k.

$$L = \prod_{i=1}^{N} \sum_{k=1}^{K} \frac{e^{x_i \theta_k}}{\sum_{k=1}^{K} e^{x_i \theta_k}} \prod_{t=1}^{T} p^k(y_{it}), \tag{3}$$

where $p^k(\cdot)$ denotes the distribution of y_{it} conditional on membership in group k.

Adding covariates to the trajectories

Let W be a vector of covariates potentially influencing Y.

The likelihood then becomes

$$L = \prod_{i=1}^{N} \sum_{k=1}^{K} \frac{e^{x_i \theta_k}}{\sum_{k=1}^{K} e^{x_i \theta_k}} \prod_{t=1}^{T} p^k(y_{it} | A_i, W_i, \Theta_k).$$

Possible data distributions

- Poisson distribution
- Binary logit distribution
- Censored normal distribution
- Beta distribution

The Beta distribution

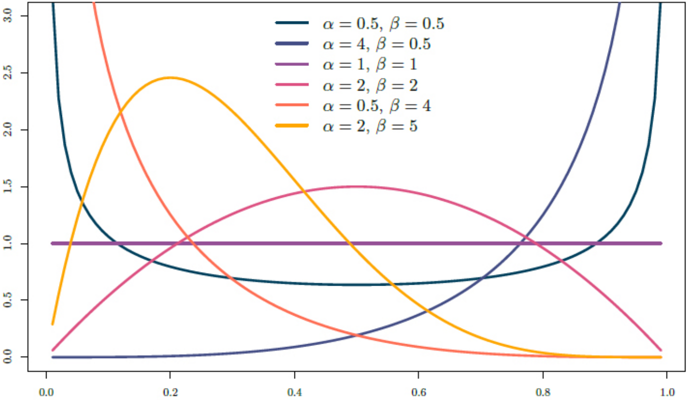


Figure 1 – *Example of different shapes of the Beta density for some parameters.*

Underlying Beta distribution

Density of y_{it} conditional to membership in group C_k :

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

with

$$\mu_{kit} = \frac{e^{\beta_k A_{it}+\delta_k W_{it}}}{1+e^{\beta_k A_{it}+\delta_k W_{it}}} \text{ and } \phi_{kit} = \zeta_k A_{it}. \tag{4}$$

Likelihood of the data:

$$L = e^{\prod_{i=1}^{n}\left(\sum_{k=1}^{K}\pi_k \prod_{t=1}^{T} \frac{\Gamma(\phi_{kit})}{\Gamma(\mu_{kit}\phi_{kit})\Gamma((1-\mu_{kit})\phi_{kit})} y_{it}^{\mu_{kit}\phi_{kit}-1}(1-y_{it})^{(1-\mu_{kit})\phi_{kit}-1}\right)}. \tag{5}$$

Outline

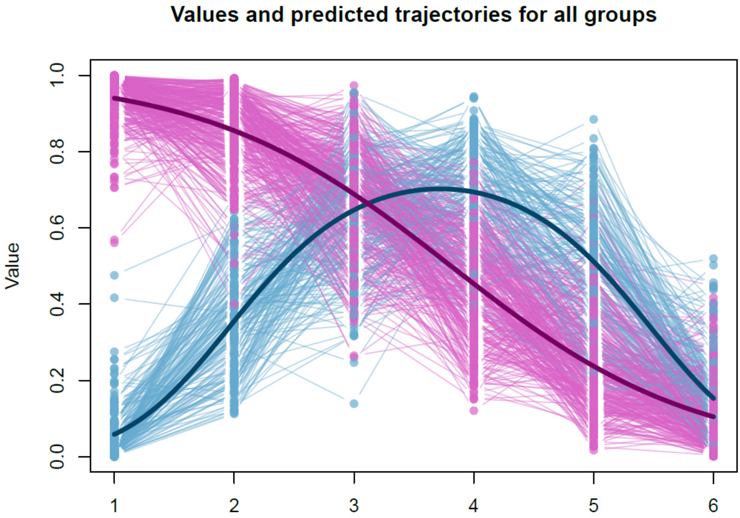
Function signature

```
R> trajeR(Y, A, Risk = NULL, TCOV = NULL, degre, degre.phi = 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,
```

Output of result

```
## 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
##    zeta
##       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
##    zeta
##       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
```

Graphical illustration of result



Outline

Data

Data from 190 countries from "Our World In Data".

Main variable of interest: **contamination rate**. We create a panel with monthly data from January 2020 till April 2021.

Covariates: new cases, population size (in million inhabitants), total cases per million people, median age of the population, population density, number of inhabitants over 65 (in million inhabitants), government response stringency index, GDP per capita, extreme poverty index, cardiovascular death rate, diabetes prevalence rate, index of handwashing facilities, rate of hospital beds per thousand inhabitants, life expectancy, index of human development and stringency index.

The nine metrics used to calculate the **stringency index** are: school closures; workplace closures; cancellation of public events; restrictions on public gatherings; closures of public transport; stay-at-home requirements; public information campaigns; restrictions on internal movements; and international travel controls.

Individual trajectories

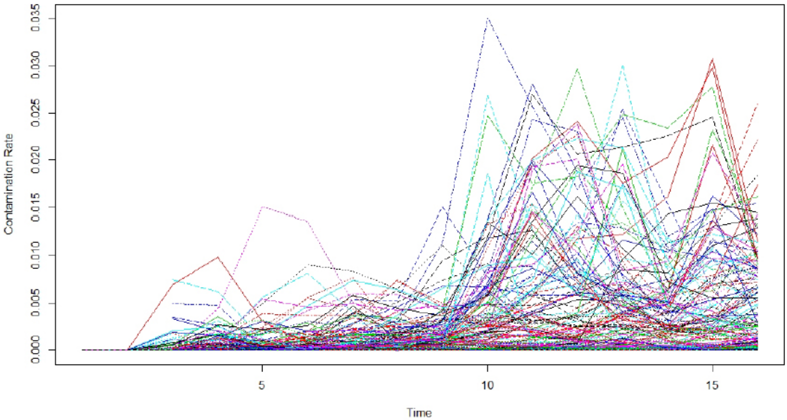


Figure 2 – *Contamination rates for all countries.*

Model selection

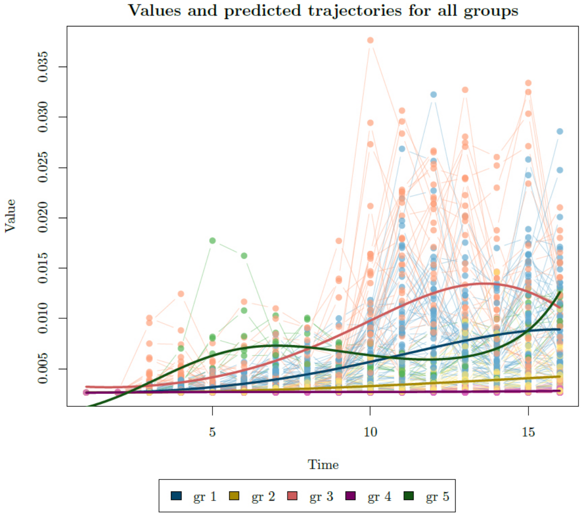
Kass and Wasserman's crierion: Let p_k be the probability that a model with k groups is the correct model. They show that p_k can be approximated by

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

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 – *Model selection criteria*

Typical trajectories



World Map with the five clusters

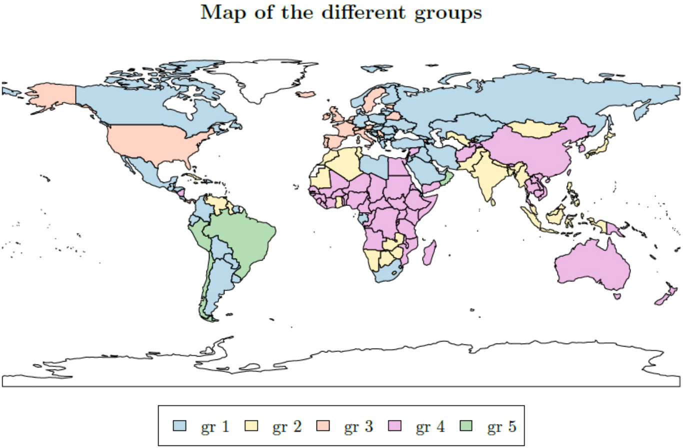


Figure 5 – *World map with the geographic distribution of the five groups*

Predictors of group membership

	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 4 – *Predictors of group membership.*

Stringency index as time dependent covariate

Param.	sd	Test
Beta 1		
-5.843	0.026	0.000
-0.120	0.024	0.000
0.029	0.004	0.000
-0.001	0.000	0.000
Beta 2		
-5.927	0.003	0.000
-0.014	0.004	0.000
0.005	0.001	0.000
0.000	0.000	0.001
Beta 3		
-5.602	0.117	0.000
-0.421	0.070	0.000
0.076	0.009	0.000
-0.003	0.000	0.000
Beta 4		
-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

Param.	sd	Test
Phi 1		
14.337	0.317	0.000
-1.164	0.076	0.000
0.040	0.004	0.000
Phi 2		
19.866	0.570	0.000
-1.710	0.125	0.000
0.061	0.006	0.000
Phi 3		
9.624	0.369	0.000
-0.521	0.097	0.000
0.016	0.005	0.003
Phi 4		
12.887	0.372	0.000
0.148	0.085	0.082
-0.015	0.004	0.000
Phi 5		
7.384	0.137	0.000

Param.	sd	Test
Delta 1		
0.001	0.000	0.001
Delta 2		
0.000	0.000	0.955
Delta 3		
0.010	0.001	0.000
Delta 4		
0.000	0.000	0.000
Delta 5		
0.004	0.001	0.004

Param.	sd	Test
Prob. 1		
0.328	0.039	0.00
Prob. 2		
0.175	0.030	0.00
Prob. 3		
0.156	0.030	0.00
Prob. 4		
0.301	0.035	0.00
Prob. 5		
0.040	0.016	0.01

Table 5 – *parameters of the final model with time dependent covariates.*

Bibliography

- Nagin, D.S. 2005: *Group-based Modeling of Development*. Cambridge, MA.: Harvard University Press.
- Schiltz, J. 2015: A generalization of Nagin's finite mixture model. In: Dependent data in social sciences research: Forms, issues, and methods of analysis' Mark Stemmler, Alexander von Eye & Wolfgang Wiedermann (Eds.). Springer 2015.
- Nagin, D.S., Jones, B.L., Lima Passos, V. & Tremblay, R.E. 2018: Group-based multi-trajectory modeling. *Statistical Methods in Medical Research*, 27-7.
- Noel, C & Schiltz, J. 2022: TrajeR - an R package for finite mixture models. SSRN paper 4054519.
- Noel, C & Schiltz, J. 2022: Finite Mixture Models for an underlying Beta distribution with an application to COVID-19 data.