



OTTAWA 2023

64TH WORLD STATISTICS CONGRESS



CPS 46

Impact of covid II

Cédric Noel & Jang Schiltz
University of Luxembourg
Tuesday 18 July, 4:00PM - 5:25PM

Finite Mixture Models For An Underlying Beta Distribution With An Application To COVID-19 Data

Jang SCHILTZ (University of Luxembourg)

joint work with

Cédric NOEL (University of Lorraine & University of Luxembourg)

- Finite Mixture Models
- Finite Mixture Models for underlying Beta distributions
- The R package trajeR
- Application to COVID-19 data

- **Finite Mixture Models**
- Finite Mixture Models for underlying Beta distributions
- The R package `trajeR`
- Application to COVID-19 data

The basic model (Nagin 2005)



Consider a population of size N and a variable of interest Y .

Let $Y_i = y_{i1}, y_{i2}, \dots, y_{iT}$ be T measures of the variable, taken at times t_1, \dots, t_T for subject number i and π_k the probability of a given subject to belong to group number k

For a given group G_k , we suppose conditional independence for the sequential realizations of the elements y_{it} over the T periods of measurements.

The density f of Y is given by

$$f(y_i; \psi) = \sum_{k=1}^K \pi_k g^k(y_i; \Theta_k), \quad (1)$$

where $g^k(\cdot)$ denotes the distribution of y_{it} conditional on membership in group k and the role of the parameters Θ_k is to describe the shape of the trajectories in group k .

Possible data distributions



- Poisson distribution
- Binary logit distribution
- (Censored) normal distribution
- Beta distribution (Noel & S. 2023)

Predictors of trajectory group membership



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

Multinomial logit model:

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

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

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

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

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

- Finite Mixture Models
- Finite Mixture Models for underlying Beta distributions
- The R package trajeR
- Application to COVID-19 data

The Beta distribution



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

Density of the Beta distribution



Let Y be a random variable following a Beta distribution with mean μ .

Consider the parameter ϕ defined by

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

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 $\phi > 0$.

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

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

Likelihood of the data:

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

- Finite Mixture Models
- Finite Mixture Models for underlying Beta distributions
- The R package `trajeR`
- Application to COVID-19 data

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 results

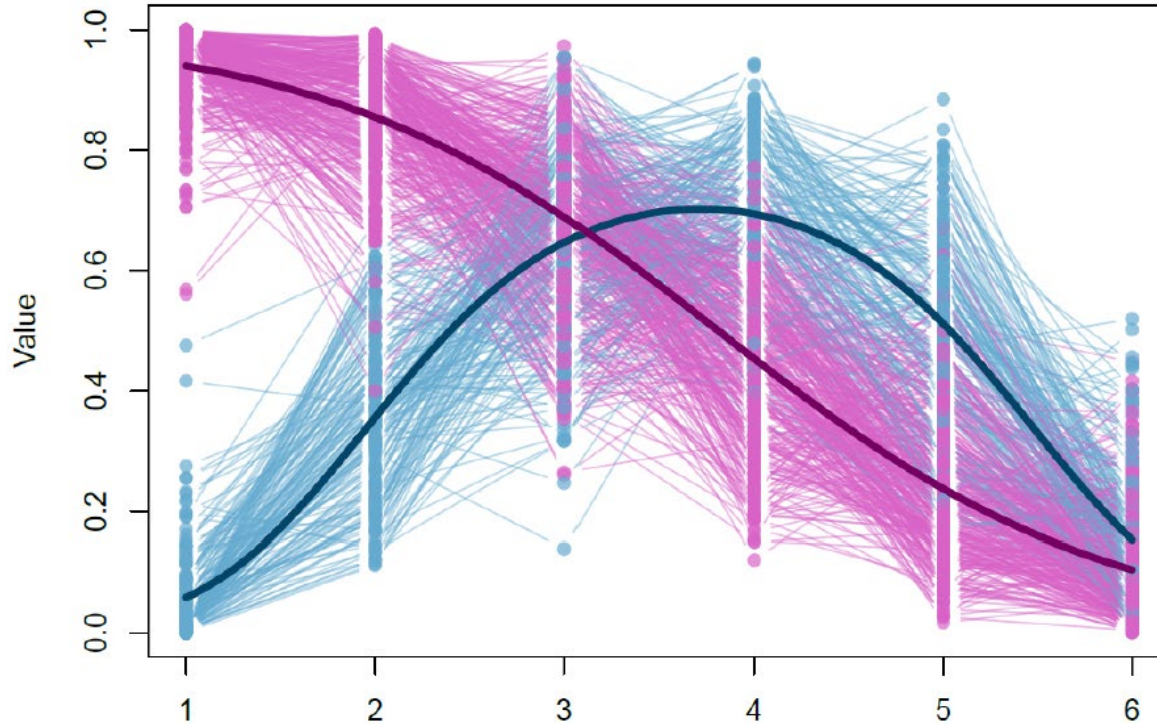


```
## Model : Beta
## Method : Likelihood
##
##   group   Parameter   Estimate   Std. Error   T for H0:   Prob>|T|
##                                     param.=0
## -----
##   mean
##     1   Intercept   -5.95316   0.1281     -46.4734     0
##           Linear     3.66558   0.07649     47.92297     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



Values and predicted trajectories for all groups



- Finite Mixture Models
- Finite Mixture Models for underlying Beta distributions
- The R package trajeR
- Application to COVID-19 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

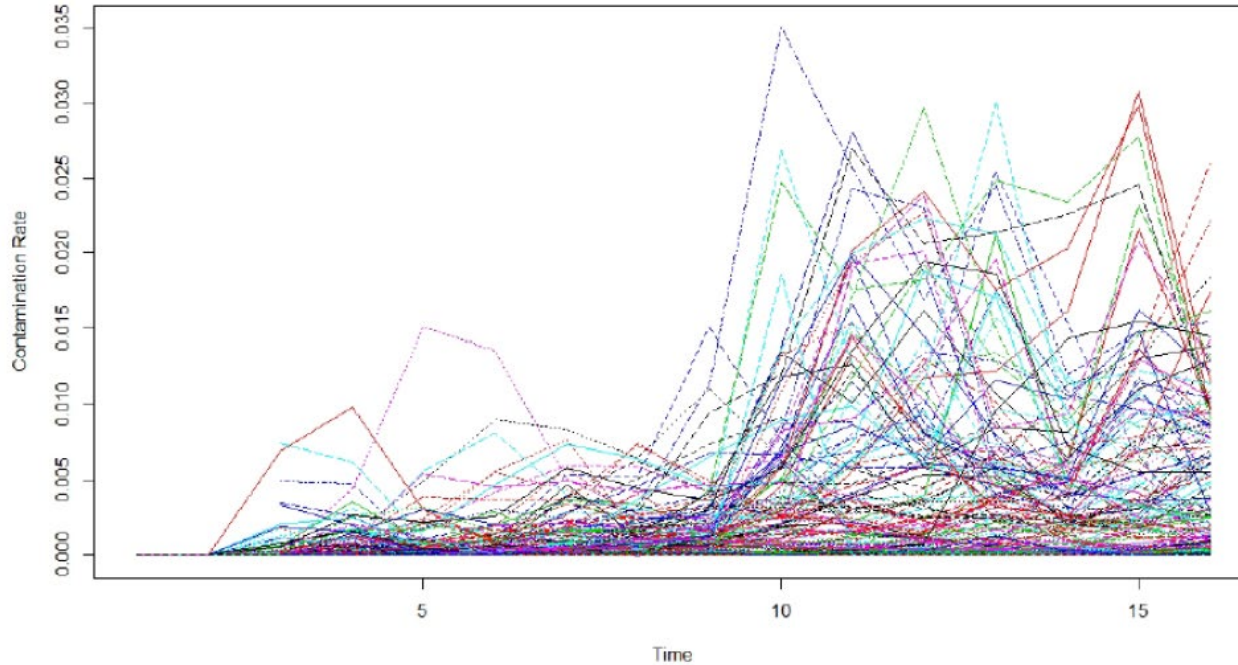


Figure 2 – *Contamination rates for all countries.*

Model selection



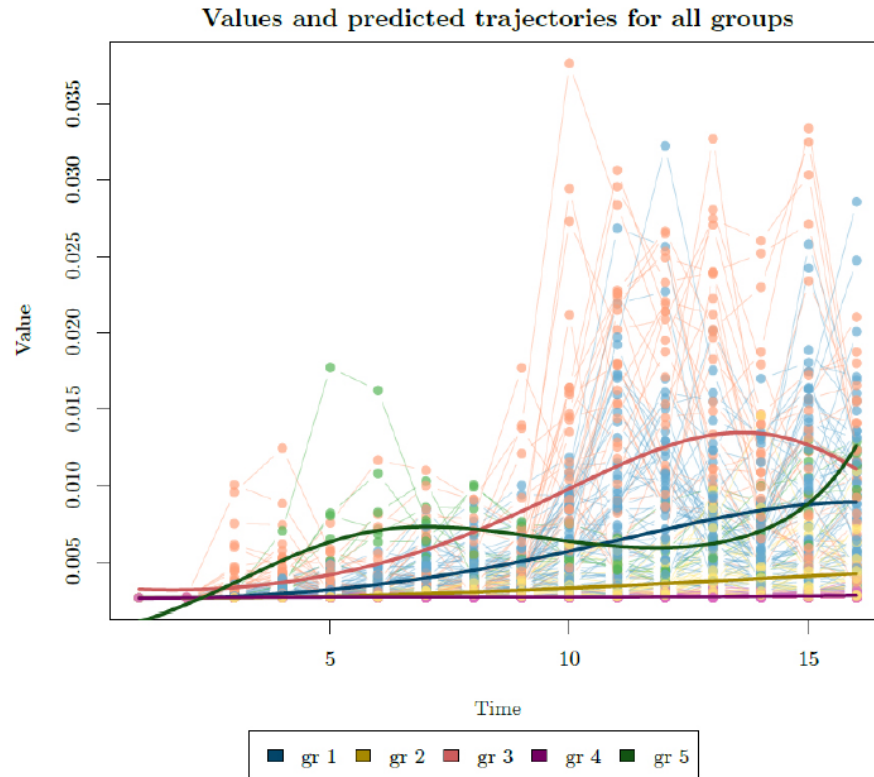
Kass and Wasserman's criterion: Let p_k be the probability that a model with k groups is the correct model. They show that p_k can be approximated by

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

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



Map of the different groups

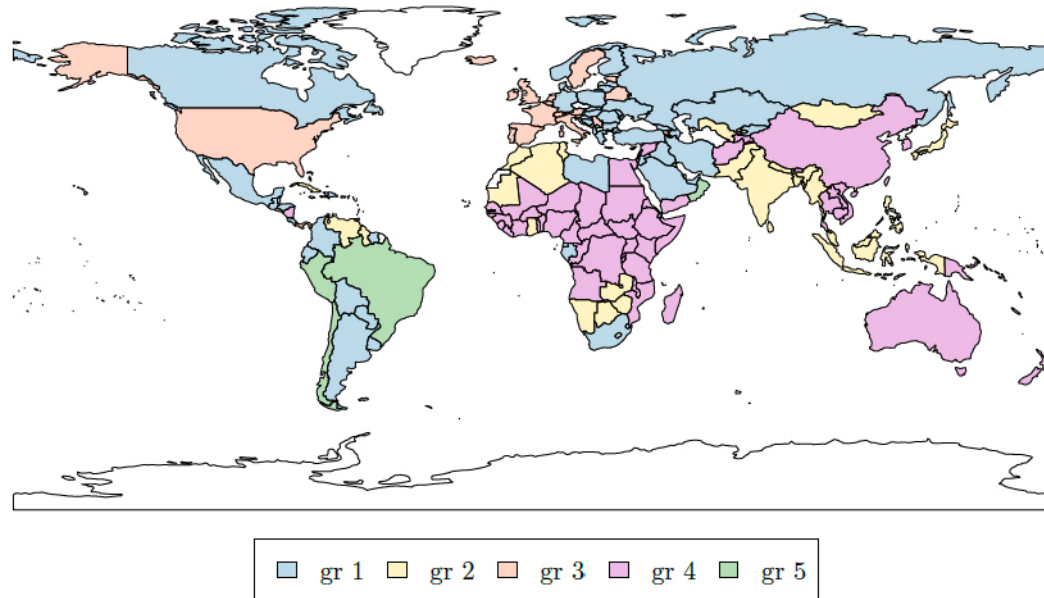


Figure 5 – *World map with the geographic distribution of the five groups*

Description of the groups



Table 3 – Means and standard deviations of the descriptive variables for each group

Variables	Group 1 mean (sd)	Group 2 mean (sd)	Group 3 mean (sd)	Group 4 mean (sd)	Group 5 mean (sd)
population size	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 and older	10.54 (6.22)	6.75 (4.69)	17.08 (5.38)	4.83 (3.3)	5.27 (3.7)
aged 70 and 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)
cardiovascular 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	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)
Stringency index	64.75 (16.98)	64.1 (15.74)	58.01 (18.17)	49.9 (16.22)	68.72 (22.02)

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	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	Delta 3			0.156	0.030	0.00
Beta 2			-1.710	0.125	0.000	0.010	0.001	0.000	Prob. 4		
-5.927	0.003	0.000	0.061	0.006	0.000	Delta 4			0.301	0.035	0.00
-0.014	0.004	0.000	Phi 3			0.000	0.000	0.000	Prob. 5		
0.005	0.001	0.000	9.624	0.369	0.000	Delta 5			0.040	0.016	0.01
0.000	0.000	0.001	-0.521	0.097	0.000	0.004	0.001	0.004			
Beta 3			0.016	0.005	0.003						
-5.602	0.117	0.000	Phi 4								
-0.421	0.070	0.000	12.887	0.372	0.000						
0.076	0.009	0.000	0.148	0.085	0.082						
-0.003	0.000	0.000	-0.015	0.004	0.000						
Beta 4			Phi 5								
-5.972	0.012	0.000	7.384	0.137	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 5 – parameters of the final model with time dependent covariates.



OTTAWA 2023

64TH WORLD STATISTICS CONGRESS



THANK YOU.