



trajeR

Un package R : clustering de données longitudinales

Cédric NOEL

Université de Lorraine/Université du Luxembourg

✉ cedric.noel@univ-lorraine.fr

Travail commun avec Jang SCHILTZ, université du Luxembourg

Présentation du modèle

On considère des trajectoires de différents individus dont on suppose qu'elles se répartissent en plusieurs groupes homogènes.

Présentation du modèle

On considère des trajectoires de différents individus dont on suppose qu'elles se répartissent en plusieurs groupes homogènes.

LCMM : latent class mixed-effect models (Muthén (1999)).

GBTM : (group-based trajectory modeling) recherche ces groupes ainsi que la trajectoire moyenne de chaque groupe (Nagin (2005)).

Kml : basé sur la méthode des **k-means** (Genolini (2011)).

Exemple illustratif

Taux de contamination du Covid sur la période : janvier 2020 - avril 2021
(<https://ourworldindata.org/>)

Exemple illustratif

Taux de contamination du Covid sur la période : janvier 2020 - avril 2021

(<https://ourworldindata.org/>)

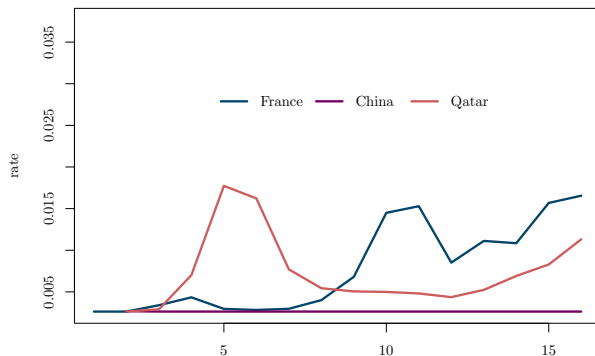
219 pays présents dans le fichier et un certain nombre de covariables.

Exemple illustratif

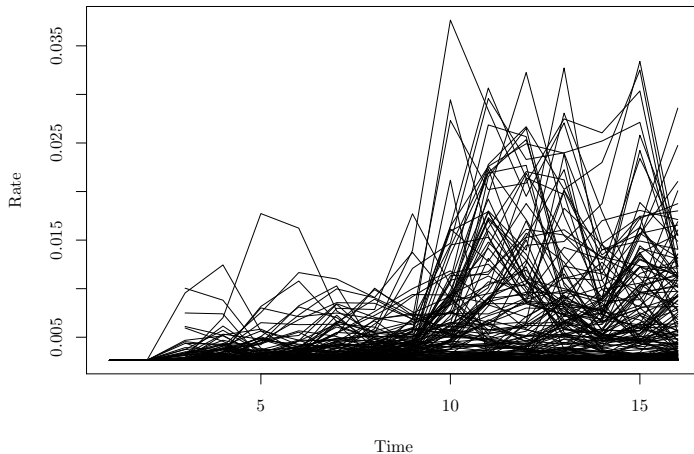
Taux de contamination du Covid sur la période : janvier 2020 - avril 2021

(<https://ourworldindata.org/>)

219 pays présents dans le fichier et un certain nombre de covariables.



Exemple illustratif



Présentation du modèle

Soit $Y_i = \{y_{i_1}, y_{i_2}, \dots, y_{i_T}\}$, T mesures de la variable Y_i , à différents temps $t_1, \dots, t_T$ pour un individu i et $Y = (Y_1, \dots, Y_n)$.

Soit π_k la probabilité pour un individu donné d'appartenir au groupe k parmi K . La probabilité jointe de Y est

$$P(Y) = \prod_{i=1}^n \sum_{k=1}^K \pi_k \prod_{t=1}^T g_k(y_{it}; \Theta_k) \quad (1)$$

But : estimer les paramètres $\Omega = \{K, \pi_k, \Theta_k; k = 1, \dots, K\}$.

Présentation du modèle

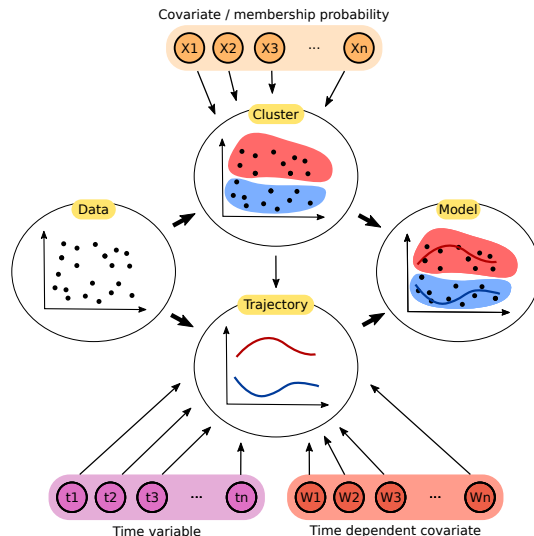
La densité g_k est influencée par une fonction du temps A_i qui est reliée à la trajectoire par les paramètres β_k et parfois par une autre variable temporelle W_i avec les paramètres δ_k .

La **probabilité** d'appartenance à un groupe peut être influencée par une variable X_i avec des paramètres θ_k .

Loi multinationale

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

Présentation du modèle



Présentation du modèle

5 distributions pour modéliser Y :

- la loi **logistique** ;
- la loi de **Poisson** ;
- la loi **ZIP** (Zero Inflated Poisson) ;
- la loi **normale censurée** ;
- la loi **beta**.

Loi logistique

Soit $\rho_{ikt} = P(Y_{it} = 1 | W_i = w_i, C_i = k)$ la probabilité que $y_{it} = 1$ étant donné l'appartenance à un groupe k .

$$\rho_{ikt} = \frac{e^{\beta_k A_{it} + \delta_k W_{it}}}{1 + e^{\beta_k A_{it} + \delta_k W_{it}}} \quad (2)$$

où $A_{it} = (1, a_{it}, a_{it}^2, \dots, a_{it}^{n_\beta - 1})^t$, $W_t = (w_{i1}, \dots, w_{in_\delta})^t$, $\beta_k = (\beta_{k1}, \dots, \beta_{kn_\beta})$ and $\delta_k = (\delta_{k1}, \dots, \delta_{kn_\delta})$

Loi Zero Inflated Poisson

La loi ZIP utilise deux processus différents : une loi binaire qui génère les zéros en excès, et une distribution de Poisson qui génère le comptage.

On a

$$P(Y_{it} = y_{it} | W_i = w_i, C_i = c_i) = \begin{cases} \rho_{ikt} + (1 - \rho_{ikt})e^{-\lambda_{ikt}}, & y_{it} = 0 \\ (1 - \rho_{ikt}) \frac{\lambda_{ikt}^{y_{it}} e^{-\lambda_{ikt}}}{y_{it}!}, & y_{it} > 0 \end{cases} \quad (3)$$

où $\lambda_{ikt} = e^{\beta_k A_{it} + \delta_k W_t}$ et $\rho_{ikt} = \frac{e^{\nu_k A_{it}}}{1 + e^{\nu_k A_{it}}}$

Pour la loi de Poisson $\rho_{ikt} = 0$.

Loi Censored Normal Model

On considère que la variable Y_{it} est censurée et une variable Y_{it}^* qui suit une loi normale telle que

$$y_{it}^* = f(a_{it}; \beta_k, \delta_k) + \epsilon_{it} = \beta_k A_{it} + \delta_k W_t + \epsilon_{it} \quad (4)$$

où $\epsilon_{it} \sim \mathcal{N}(0; \sigma)$ et donc $y_{it}^* \sim \mathcal{N}(\beta_k A_{it} + \delta_k W_t; \sigma)$

On peut lier y_{it}^* aux données observées et censurées y_{it} .

$$y_{it} = y_{\min} \text{ if } y_{it}^* < y_{\min} \quad (5)$$

$$y_{it} = y_{it}^* \text{ if } y_{\min} \leq y_{it}^* \leq y_{\max} \quad (6)$$

$$y_{it} = y_{\max} \text{ if } y_{it}^* > y_{\max} \quad (7)$$

Remarque - dans le cadre de la loi normale, la fonction f peut être choisie non polynomiale.

Loi Beta

On considère que Y_{it} suit une loi beta que l'on exprime à l'aide de la moyenne (Ferrari et Cribari-Neato (2004)). Sa densité, conditionnellement à un groupe k est

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

où

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

Vraisemblance

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

trajeR

Nagin & Jones ont programmé une procédure SAS et Stata : traj.

On propose un package R avec 2 méthodes :

Likelihood (méthode quasi Newton BFGS) and **EM**.



 **CRAN** : <https://CRAN.R-project.org/package=trajeR>

 **GitHub** : <https://github.com/gitedric/trajeR>

trajeR

Nagin & Jones ont programmé une procédure SAS et Stata : traj.

On propose un package R avec 2 méthodes :

Likelihood (méthode quasi Newton BFGS) and **EM**.



 **CRAN** : <https://CRAN.R-project.org/package=trajeR>

 **GitHub** : <https://github.com/gitedric/trajeR>

Fonction principale :

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

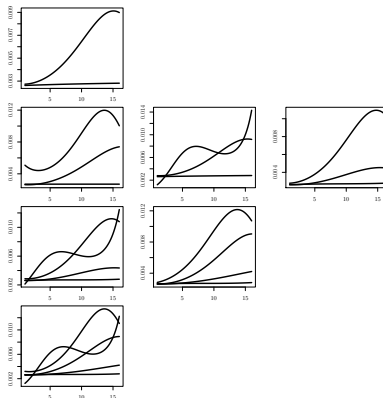
$$\text{Modèle : } \mu_{ikt} = \frac{e^{\beta_k A_{it}}}{1 + e^{\beta_k A_{it}}}$$

$$\text{Modèle : } \mu_{ikt} = \frac{e^{\beta_k A_{it}}}{1 + e^{\beta_k A_{it}}}$$

Recherche du
nombre de
groupes et des
degrés des
polynômes (β_k et
 ζ_k).

Plusieurs valeurs
de départ.

BIC, AIC.



Number of
groups

	AIC	BIC
2	29851.99	14902.64
3	30341.00	15142.28
3	29945.96	14936.64
3	30777.14	15352.23
4	30839.69	15370.52
4	31192.78	15547.06
5	31241.46	15558.41

Call TrajeR with 5 groups and a 3,3,3,3,3 degrees of polynomial shape of trajectory.

Model : Beta

Method : Likelihood

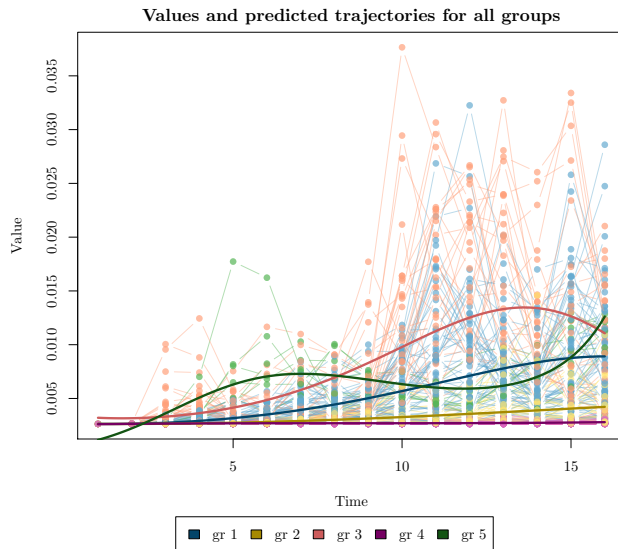
group	Parameter	Estimate	Std. Error	T for H0: param.=0	Prob> T

mean					
1	Intercept	-5.90246	0.01886	-312.96208	0
	Linear	-0.05213	0.01309	-3.98256	7e-05
	Quadratic	0.0204	0.0024	8.48551	0
	Cubic	-0.00078	0.00011	-6.83871	0
var.					
1	Intercept	14.64818	0.30861	47.46525	0
	Linear	-1.25279	0.07249	-17.28236	0
	Quadratic	0.04463	0.00378	11.8035	0
...					
...					
mean					
5	Intercept	-7.62854	0.33788	-22.57763	0
	Linear	0.96226	0.1349	7.13322	0
	Quadratic	-0.10905	0.0169	-6.45322	0
	Cubic	0.00385	0.00063	6.16391	0
var.					
5	Intercept	7.42213	0.1463	50.73339	0

1	pi1	0.30152	0.03404	0	0
2	pi2	0.18327	0.0284	-17.52907	0
3	pi3	0.15938	0.02766	-23.05472	0
4	pi4	0.31945	0.03389	1.7038	0
5	pi5	0.03639	0.01376	-153.62932	0.00825

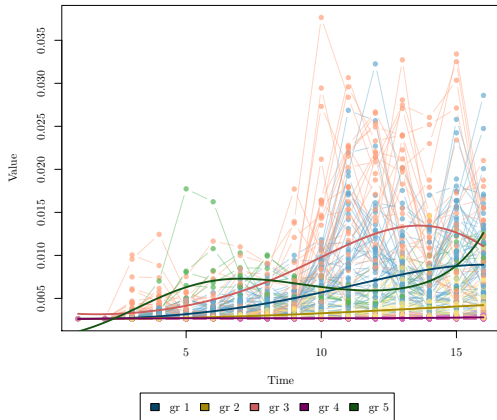
Likelihood : 15659.63

Exemple illustratif

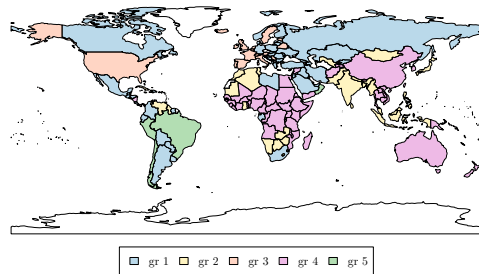


Exemple illustratif

Values and predicted trajectories for all groups



Map of the different groups



GroupProfiles(...)

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)

Variables liées (groupes)

Ajout de variables liées pour l'appartenance aux groupes : l'âge médian, la densité de population, le taux de personnes de plus de 65 ans, le PIB, l'espérance de vie, la moyenne et l'écart-type d'un indicateur de rigueur des mesures prises par les pays.

Variables liées (groupes)

Ajout de variables liées pour l'appartenance aux groupes : l'âge médian, la densité de population, le taux de personnes de plus de 65 ans, le PIB, l'espérance de vie, la moyenne et l'écart-type d'un indicateur de rigueur des mesures prises par les pays.

```
trajeR(  
  Y = rate, A = mattime[1:nrow(rate), ],  
  Risk = dd,  
  degre = c(3,3,3,3,3),  
  degre.phi = c(2,2,2,2,0),  
  paraminit = paraminit,  
  Model = "BETA", Method = "L"  
)
```

Variables liées (groupes)

Ajout de variables liées pour l'appartenance aux groupes : l'âge médian, la densité de population, le taux de personnes de plus de 65 ans, le PIB, l'espérance de vie, la moyenne et l'écart-type d'un indicateur de rigueur des mesures prises par les pays.

```
trajeR(  
Y = rate, A = mattime[1:nrow(rate), ],  
Risk = dd,  
degre = c(3,3,3,3,3),  
degre.phi = c(2,2,2,2,0),  
paraminit = paraminit,  
Model = "BETA", Method = "L"  
)
```

	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

Variables liées (trajectoires)

Est-ce qu'une ou plusieurs variables peuvent influencer la forme des trajectoires ?

Variables liées (trajectoires)

Est-ce qu'une ou plusieurs variables peuvent influencer la forme des trajectoires ?

Modèle : $\mu_{ikt} = \frac{e^{\beta_k A_{it} + \delta_k W_{it}}}{1 + e^{\beta_k A_{it} + \delta_k W_{it}}}$ où W_{it} est composé d'un indice qui indique la sévérité des mesures du pays i au temps t .

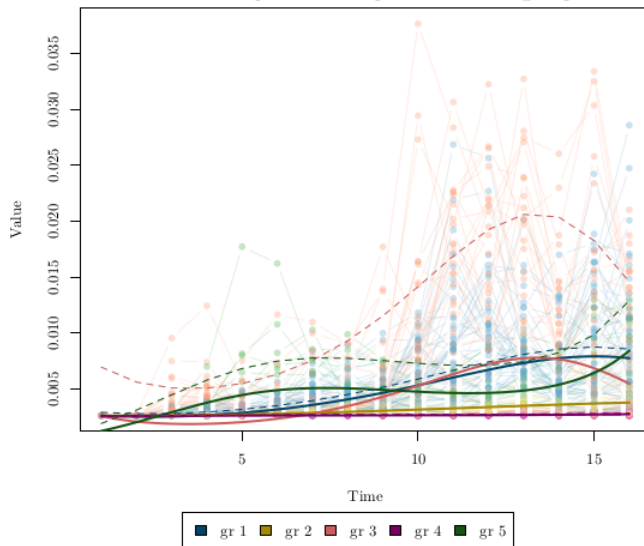
Variables liées (trajectoires)

Est-ce qu'une ou plusieurs variables peuvent influencer la forme des trajectoires ?

Modèle : $\mu_{ikt} = \frac{e^{\beta_k A_{it} + \delta_k W_{it}}}{1 + e^{\beta_k A_{it} + \delta_k W_{it}}}$ où W_{it} est composé d'un indice qui indique la sévérité des mesures du pays i au temps t .

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				0.301	0.035	0.00
-0.014	0.004	0.000	Phi 3			Delta 4			Prob. 5		
0.005	0.001	0.000	9.624	0.369	0.000	0.000	0.000	0.000	0.040	0.016	0.01
0.000	0.000	0.001	-0.521	0.097	0.000	Delta 5					
			0.016	0.005	0.003	0.004	0.001	0.004			
Beta 3			Phi 4								
-5.602	0.117	0.000	12.887	0.372	0.000						
-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									

Values and predicted trajectories for all groups



Merci de votre attention

 **CRAN** : <https://CRAN.R-project.org/package=trajeR>

 **Github** : <https://github.com/gitedric/trajeR>

cedric.noel@univ-lorraine.fr