trajeR

An R package for the clustering of longitudinal data

Cédric NOEL[1], Jang SCHILTZ[2]

[1] University of Luxembourg and IUT of Thionville-Yutz, University of Lorraine, Espace Cormontaigne Impasse Alfred Kastler F-57970 Yutz, France
[2]Department of Finance, University of Luxembourg, 6, rue Richard Coudenhove-Kalergi L-1359 Luxembourg, Luxembourg

Model's presentation

Starting from a collection of individual trajectories, the aim of group-based trajectory modeling (GBTM) is to divide the population into a number of homogeneous sub-populations and to estimate, at the same time, a typical trajectory for each sub-population

Model's presentation

Let $Y_i = y_{i_1}, y_{i_2}, ..., y_{i_T}$ be T measures of the variable Y, taken at times $t_1, ..., t_T$ for subject number i.
Denote the probability of a given subject to belong to group number k by π_k for a number of groups K.

$$P(Y_i) = \sum_{k=1}^{K} \pi_k g_k(y_i; \Theta_k). \tag{1}$$

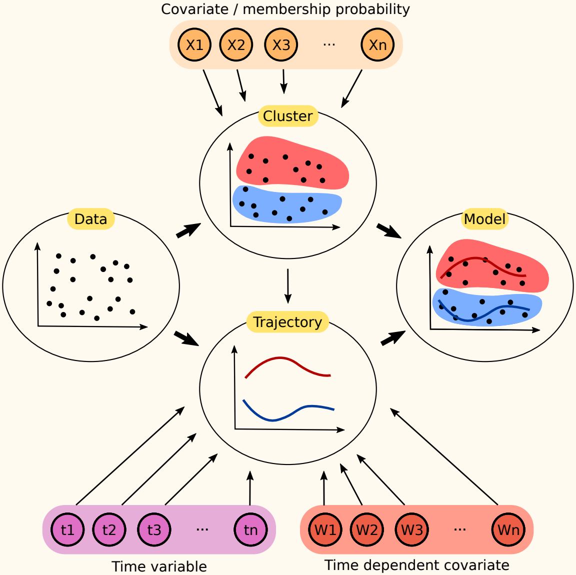
Aim : estimate a set of parameters $\Omega = \{K, \pi_k, \Theta_k; k = 1, ..., K\}$.

Model's presentation

The density g_k is influenced by a function of time denoted A_i that are related to the shape of the trajectories with parameters β_k and sometimes by some covariate denoted W_i with parameters δ_k.

The membership probability can be influenced by some covariate denoted X_i with parameters θ_k.

Model’s presentation



Model's presentation

Nagin proposes 3 different distributions to modelize Y :

- the censored normal distribution ;
- the logistic distribution ;
- the ZIP (Zero Inflated Poisson) distribution.

To fit his model Nagin & Jones have programmed a SAS and Stata procedure : `traj`.

trajeR

We propose a R package to fit this model, like `traj` for SAS.

with 2 methods : Likelihood and EM.

Main function :

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.lmiter)

trajeR - Censored Normal Model

We suppose that the variable Y_{it} is a censored variable.

We consider a variable Y_{it}^* that is normally distributed and such

$$y_{it}^* = f(a_{it}; \beta_k, \delta_k) + \epsilon_{it} = \beta_k A_{it} + \delta_k W_t + \epsilon_{it} \tag{2}$$

where $\epsilon_{it} \sim \mathcal{N}(0;\ \sigma)$, $A_{it} = (1, a_{it}, a_{it}^2, \cdots, a_{it}^{n_\beta - 1})^t$, $W_t = (w_{i1}, \cdots, w_{in_\delta})^t$, $\beta_k = (\beta_{k1}, \cdots, \beta_{kn_\beta})$ and $\delta_k = (\delta_{k1}, \cdots, \delta_{kn_\delta})$

Furthermore, we can link y_{it}^* to the observed and censored data y_{it}.

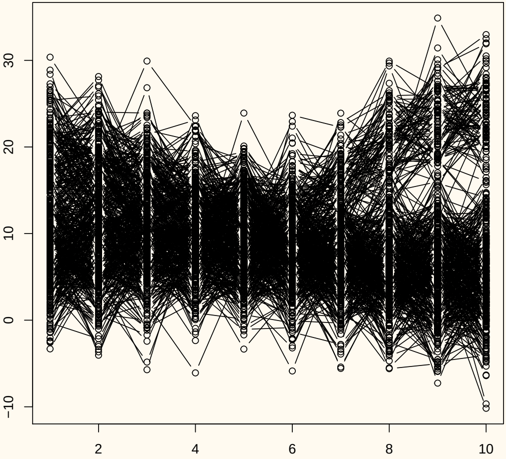
$$y_{it} = y_{min} \text{ if } y_{it}^* < y_{min} \tag{3}$$

$$y_{it} = y_{it}^* \text{ if } y_{min} \leq y_{it}^* \leq y_{max} \tag{4}$$

$$y_{it} = y_{max} \text{ if } y_{it}^* > y_{max} \tag{5}$$

trajeR - Censored Normal Model

we simulate 500 trajectories to fit this model



trajeR - Censored Normal Model

We suppose $K = 3$ and we search polynomial shapes with degree 0, 3 and 4.

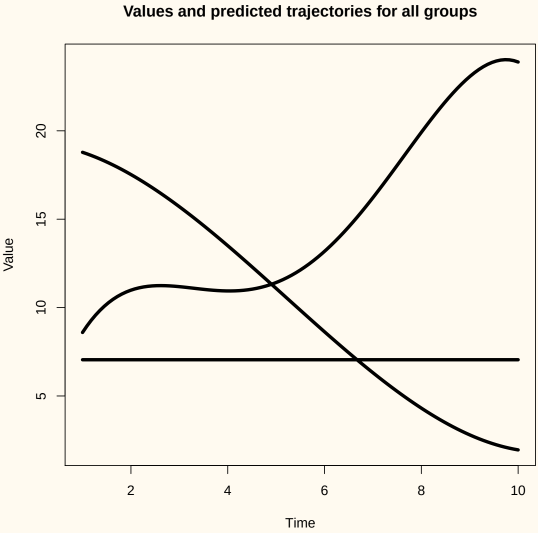
```
solL = trajeR(Y = data[,2:11], A = data[,12:21], ng = 3, degre =
c(0,3,4), Model = "CNORM", Method = "L", hessian = TRUE,
ssigma = FALSE)
```

trajeR - Censored Normal Model

```
Call TrajeR with 3 groups and a 0,3,4 degrees of polynomial shape of trajectory.
Model : Censored Normal
Method : Likelihood

  group   Parameter   Estimate   Std. Error   T for H0:   Prob>|T|
                                               param.=0
--------------------------------------------------------------------
      1   Intercept    7.04989      0.08425    83.67725          0

      2   Intercept    19.3042      0.08322   231.96629          0
             Linear   -0.09337      0.10609    -0.88009    0.37885
          Quadratic   -0.45605      0.03793   -12.02404          0
              Cubic    0.02918      0.00287    10.16725          0

      3   Intercept    1.66272      0.01264   131.55795          0
             Linear   10.12613      0.02797   362.03353          0
          Quadratic   -3.71001      0.04415   -84.02376          0
              Cubic    0.53801      0.01303    41.28388          0
            Quartic    -0.0246      0.00087   -28.35329          0
--------------------------------------------------------------------
      1      sigma1     3.9581      0.01522   260.00927          0
      2      sigma2    4.11023      0.01714    239.8676          0
      3      sigma3    4.00186      0.02338   171.14613          0
--------------------------------------------------------------------
      1         pi1    0.45891       0.0616    22.42831          0
      2         pi2    0.34895      0.06515    17.00024          0
      3         pi3    0.19214      0.07596     6.72576          0
--------------------------------------------------------------------
Likelihood : -14564.35
```

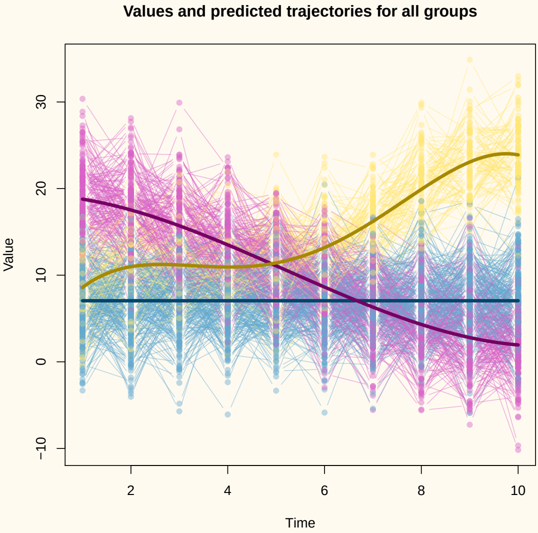
trajeR - Censored Normal Model

```
plot(solL)
```



trajeR - Censored Normal Model

```
plot(solL, Y = data[,2:11], A = data[,12:21], col = vcol)
```



trajeR - Censored Normal Model

adding a covariate influencing group membership

```
solLRisks = trajeR(Y = data[,2:11], A = data[,12:21], Risk =
data[,42:43], ng = 3, degre = c(0,3,4), Model = "CNORM",
Method = "L", ssigma = TRUE, hessian = FALSE)
```

trajeR - Censored Normal Model

```
Call TrajeR with 3 groups and a 0,3,4 degrees of polynomial shape of trajectory.
Model : Censored Normal
Method : Likelihood

  group   Parameter   Estimate   Std. Error   T for H0:   Prob>|T|
                                               param.=0
-------------------------------------------------------------------
      1   Intercept    7.04921      0.04182   168.56007          0

      2   Intercept   19.30527      0.11619   166.14851          0
             Linear   -0.09488      0.12002    -0.79059    0.42922
          Quadratic   -0.45575      0.04882     -9.3359          0
              Cubic    0.02917      0.03002     0.97159     0.3313

      3   Intercept    1.68068      0.03135    53.61591          0
             Linear   10.10606      0.08431   119.86617          0
          Quadratic   -3.70319      0.08324   -44.48605          0
              Cubic    0.53712      0.10611     5.06184          0
            Quartic   -0.02456      0.03794     -0.6475    0.51734
-------------------------------------------------------------------
      1      sigma1    3.95758      0.00287   1378.46434          0
      2      sigma2    4.11102      0.01273   322.89742          0
      3      sigma3    4.00154      0.02823   141.74385          0
-------------------------------------------------------------------
      1    Baseline          0      0.04437    36.35294          0

      2   Intercept    -0.6561      0.01524    62.79053          0
                V42    0.19568      0.01714    63.87678          0
                V43    0.57614      0.02341    51.61321          0

      3   Intercept   -1.18333       0.0404    10.63963          0
                V42    0.10695      0.13102     7.67879          0
                V43    0.52682      0.13634     8.50175          0
-------------------------------------------------------------------
Likelihood : -14562.59
```

Adding time dependent covariate

solLTCOV2 = **trajeR**(Y = data[,2:11], A = data[,12:21], TCOV = data[,22:41], ng = 3, degre = **c**(0,3,4), Model = "CNORM", Method = "L", ssigma = FALSE, hessian = FALSE)

trajeR - Censored Normal Model

```
Call TrajeR with 3 groups and a 0,3,4 degrees of polynomial shape of trajectory.
Model : Censored Normal
Method : Likelihood

  group   Parameter   Estimate   Std. Error   T for H0:    Prob>|T|
                                               param.=0
-------------------------------------------------------------------
      1   Intercept    6.75745      0.08309    81.32696           0
              TCOV1   -0.00358      0.16289    -0.02195     0.98249
              TCOV2    0.58151      0.06557     8.86839           0

      2   Intercept   19.45552      0.07795   249.59559           0
             Linear   -0.09592      0.10136    -0.94628     0.34405
          Quadratic   -0.45507      0.03622   -12.56502           0
              Cubic     0.0291      0.00277    10.51302           0
              TCOV1   -0.17051      0.19635     -0.8684     0.38522
              TCOV2   -0.13303      0.05521    -2.40954     0.01601

      3   Intercept    1.79828      0.06183    29.08577           0
             Linear   10.13362      0.03745   270.57751           0
          Quadratic   -3.71451      0.07286   -50.98126           0
              Cubic    0.53887      0.01732    31.10904           0
            Quartic   -0.02465      0.00106   -23.18069           0
              TCOV1    0.08064      0.21207     0.38024     0.70378
              TCOV2    -0.3598        0.038    -9.46729           0
-------------------------------------------------------------------
      1      sigma1    3.95415      0.01531   258.26618           0
      2      sigma2    4.10997      0.01714   239.80629           0
      3      sigma3    4.00024      0.02353   169.98647           0
-------------------------------------------------------------------
      1         pi1    0.45881       0.0617    22.38634           0
      2         pi2    0.34911      0.06516    17.00593           0
      3         pi3    0.19207       0.0761     6.70953           0
-------------------------------------------------------------------
Likelihood : -14561.48
```

The ZIP employ two different process : a binary distribution that generate structural zero, i.e. the excess zeros, and and poisson distribution that generate the counts.

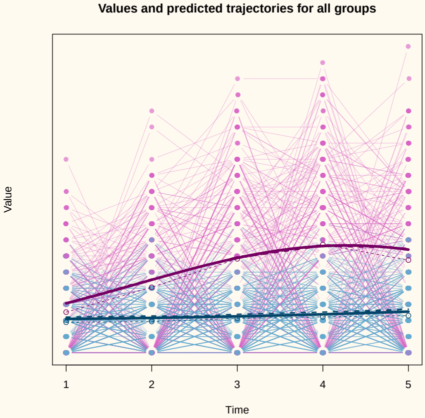
We have

$$P\left(Y_{it}=y_{it}|W_i=wi,C_i=ci\right)=\begin{cases}\rho_{ikt}+(1-\rho_{ikt})e^{-\lambda_{ikt}}, & y_{it}=0\\ (1-\rho_{ikt})\frac{\lambda_{ikt}^{y_{it}}e^{-\lambda_{ikt}}}{y_{it}!}, & y_{it}>0\end{cases} \tag{6}$$

trajeR - Zero Inflated Poisson

```
paraminit = c(1,1, -3.066653,0,0, -0.046659,0,0, -3,0, -3,0)

solL = trajeR(Y = data[,2:6], A = data[,7:11], ng = 2,
degre=c(2,2), degre.nu = c(1,1), Model="ZIP", Method = "L",
hessian = TRUE, paraminit = paraminit)
```

```
Call TrajeR with 2 groups and a 2,2 degrees of polynomial shape of trajec
tory.
Model : Zero Inflated Poisson
Method : Likelihood

  group   Parameter   Estimate   Std. Error   T for H0:   Prob>|T|
                                               param.=0
---------------------------------------------------------------------
      1   Intercept    1.04873      0.09415    11.13945          0
             Linear    0.61506      0.06423     9.57663          0
          Quadratic   -0.07773      0.00998    -7.79183          0
               Nu11    0.03161      0.16749     0.18871    0.85034
               Nu12   -0.16776      0.05183    -3.23661    0.00123

      2   Intercept    0.90787      0.09589      9.4679          0
             Linear    0.01869      0.07032      0.2658    0.79042
          Quadratic    0.00619      0.01131     0.54724    0.58426
               Nu21   -1.08139      0.16018    -6.75111          0
               Nu22    0.00452      0.04689     0.09632    0.92327
---------------------------------------------------------------------
      1         pi1    0.34869      0.04936    13.93138          0
      2         pi2    0.65131      0.04936    26.59044          0
---------------------------------------------------------------------
Likelihood : -5162.009
```



Let $\rho_{ikt} = P(Y_{it} = 1 | W_i = w_i, C_i = k)$ the probability of $y_{it} = 1$ given membership in group k.

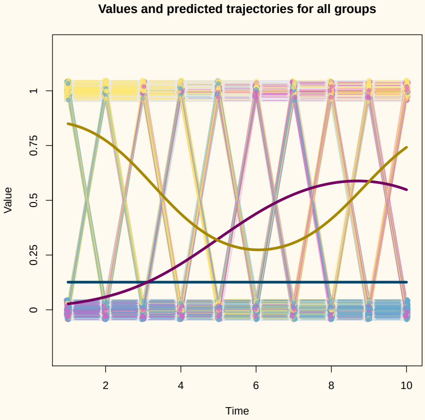
$$\rho_{ikt} = \frac{e^{\beta_k A_{it} + \delta_k W_{it}}}{1 + e^{\beta_k A_{it} + \delta_k W_{it}}} \tag{7}$$

trajeR - Logit

```
solL = trajeR(Y = data[,2:11], A = data[,12:21], ng = 3, degre = c(0,3,4), Model = "LOGIT", Method = "L", hessian = TRUE, itermax = 300, paraminit = paraminit)
```

trajeR - Logit

```
Call TrajeR with 3 groups and a 0,3,4 degrees of polynomial shape of traj
ectory.
Model : Logit
Method : Likelihood

  group   Parameter   Estimate   Std. Error   T for H0:   Prob>|T|
                                              param.=0
---------------------------------------------------------------------
      1   Intercept   -1.94482      0.10059   -19.33387          0

      2   Intercept   -1.77021      0.03535   -50.07501          0
             Linear   -0.44043          NaN         NaN        NaN
          Quadratic    0.18687          NaN         NaN        NaN
              Cubic   -0.01232          NaN         NaN        NaN

      3   Intercept    1.74626          NaN         NaN        NaN
             Linear     0.6443      0.07247     8.89028          0
          Quadratic    -0.5554      0.02475   -22.44466          0
              Cubic    0.08204      0.00195    42.07317          0
          Quadratic   -0.00336          NaN         NaN        NaN
---------------------------------------------------------------------
      1         pi1    0.41816      0.10875     2.41645    0.01571
      2         pi2    0.21915      0.11423    -3.35532      8e-04
      3         pi3    0.36269       0.0623     1.93388    0.05318
---------------------------------------------------------------------
Likelihood : -2852.749
```



We suppose that the variable Y_{it} is defined by

$$y_{it} = f(a_{it}; \beta_k) + \epsilon_{it} \tag{8}$$

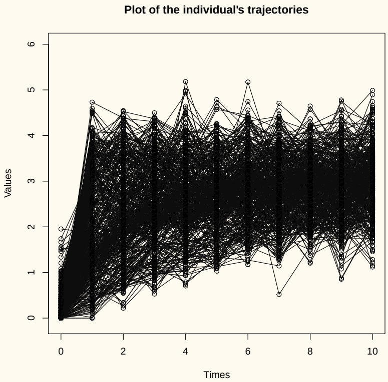
where $\epsilon_{it} \sim \mathcal{N}(0;\ \sigma_k)$.
In this case, the function f_k is not linear in the parameters β_k and

$$E(Y_{it} = y_{it} | W_i = w_i, C_i = k) = f(a_{it}; \beta_k) \tag{9}$$

trajeR - Non Linear

Example with

$$f(t;\beta_k) = \frac{\beta_{k1} t}{\beta_{k2} + t}$$



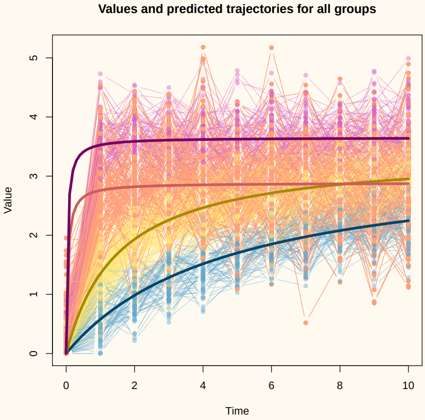
By analysing the graph above, we fit initial parmaters as paraminit=c(0.25,0.25,0.25,0.25,2, 0.1,2.4,0.1,2.8,0.1,3,0.1,0.2,0.2,0.2,0.2)

trajeR - Non Linear

```
fct <- function(t, betak, TCOV){ return(
(betak[1]*t)/(betak[2]+t) )
diffct <- function(t, betak, TCOV){ return(c( t/(betak[2]+t),
-(betak[1]*t)/(betak[2]+t)**2 ))

solEM = trajeR(Y = data[,2:12], A = data[,13:23], ng = 4, nbvar
= 2, Method = "EM", Model = "NL", hessian = TRUE, fct =
fct, diffct = diffct, paraminit = paraminit)
```

trajeR - Non Linear

```
Call TrajeR with 4 groups and a 1,1,1,1 degrees of polynomial shape of tr
ajectory.
Model : Non Linear
Method : Expectation-maximization

  group   Parameter   Estimate   Std. Error   T for H0:    Prob>|T|
                                               param.=0
---------------------------------------------------------------------
      1   Intercept    3.30542     0.18506    17.86117           0
             Linear    4.71859     0.60508     7.79825           0

      2   Intercept    3.17397     0.06338    50.07605           0
             Linear    1.29635     0.12496    10.37439           0

      3   Intercept    3.07645     0.06425    47.88329           0
             Linear    0.60537     0.11135     5.43645           0

      4   Intercept    3.34255     0.02515   132.91026           0
             Linear    0.04503      0.0236     1.90772     0.05648
---------------------------------------------------------------------
      1      sigma1    0.63465         NaN         NaN         NaN
      2      sigma2    0.56821      0.0036   157.93961           0
      3      sigma3    0.60047     0.02216     27.0945           0
      4      sigma4    0.71747     0.01667     43.0461           0
---------------------------------------------------------------------
      1         pi1    0.15427     0.01646      9.3702           0
      2         pi2    0.23712     0.02962     8.00592           0
      3         pi3    0.25955         NaN         NaN         NaN
      4         pi4    0.34906         NaN         NaN         NaN
---------------------------------------------------------------------
Likelihood : -4891.536
```



trajeR - Some functions

- Membership's probability – `GroupProb(...)`
- Profiles of group – `GroupProfiles(...)`

	Gr 1	Gr 2	Gr 3	Gr 4
X1	1.8467	3.2709	6.0562	6.5621
X2	7.7916	-0.9212	-4.9915	5.3515
X3	56.1620	74.0841	41.0704	32.8957
X4	0.3636	0.2895	0.4972	0.3333
X5	0.2545	0.3289	0.4525	0.3421

trajeR - Model performance

- Average Posterior Probability – `AvePP(...)`
- Odds of Correct Classification – `OCC(...)`
- Estimated group probabilities versus proportion of the sample assigned to the group – `propAssign(...)`
- Confidence interval – `ConfIntT(...)`
- Summary – `adequacy(...)`

	1	2	3	4
Prob. est.	0.10227620	0.3118053	0.2247159	0.3612026
CI inf.	0.08165381	0.2720132	0.1916078	0.3161885
CI sup.	0.12689314	0.3542225	0.2625223	0.4086869
Prop.	0.09800000	0.3180000	0.2220000	0.3620000
AvePP	0.89837432	0.9395957	0.9843333	0.9973927
OCC	77.59454723	34.3317995	216.7625509	676.5346954

trajeR - Model selection

- AIC – `trajeRAIC(...)`
- BIC – `trajeRBIC(...)`
- Slope Heuristics – `trajeRSH(...)`