Finite Mixture Models for an underlying zero-one inflated Beta Distribution

Jang SCHILTZ (University of Luxembourg)

joint work with

Cédric NOEL(University of Lorraine & University of Luxembourg)

Deep Learning and Clustering Workshop
December 9, 2024

Outline

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.

The basic model (Nagin 2005)

Consider a population of size N and a variable of interest Y.
Let $Y_i = y_{i_1}, y_{i_2}, ..., y_{i_T}$ be T measures of the variable, taken at times $t_1, ...t_T$ 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_{i_t} over the T periods of measurements.

The density f of Y is given by

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

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.

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

Outline

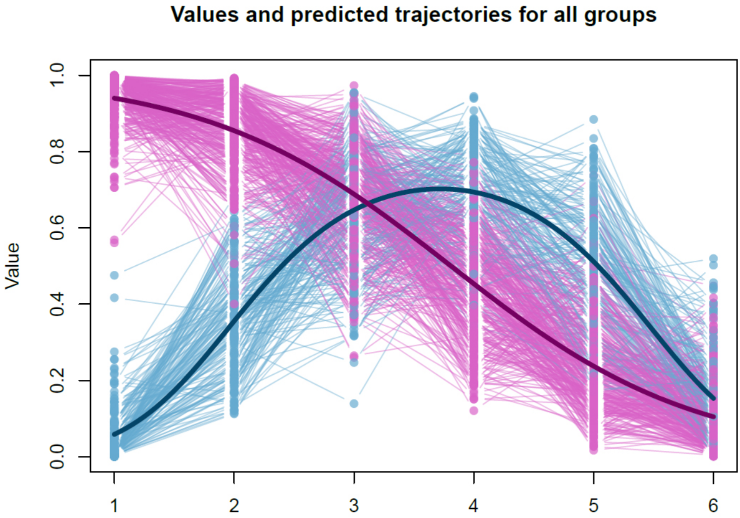
Function signature

```
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

Density of the Beta distribution

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 $\phi > 0$.

Finite mixture models for an 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:

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

Problem with the Beta distribution

Computationally, there are problems with data values of 0 or 1.

- Nagin's model: Data need to be strictly between 0 and 1.
- Our model: transformation of y into $(y \cdot (n-1) + 0.5)/n$ (Smithson and Verkuilen, 2006)

Outline

The zero-one inflated Beta distribution

$$P\left(Y_{it}=y_{it}|C_i=k\right)=\begin{cases}\rho_{ikt}^{01}\left(1-\rho_{ikt}^{1}\right) & \text{if } y_{it}=0\\ \rho_{ikt}^{01}\rho_{ikt}^{1} & \text{if } y_{it}=1\\ \left(1-\rho_{ikt}^{01}\right)\text{Beta}(\mu_{ikt},\phi_{ikt}) & \text{if } 0<y_{it}<1\end{cases}$$

where ρ_{ikt}^{01} is the probability that the variable takes the value 0 or 1, and ρ_{ikt}^{1} is the conditional probability of inflation one

Linking formulae

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

where $A_{it} = (1, a_{it}, a_{it}^2, \cdots, a_{it}^{n_\beta - 1})^t$, $A_{\zeta,it} = (1, a_{it}, a_{it}^2, \cdots, a_{it}^{n_\zeta - 1})^t$, $W_{it} = (w_{i1}, \cdots, w_{in_\delta})^t$, $\beta_k = (\beta_{k1}, \cdots, \beta_{kn_\beta})$ and $\zeta_k = (\zeta_{k1}, \cdots, \zeta_{kn_\zeta})$.

The other parameters are modeled as follows:

$$\log\left(\frac{\rho_{ikt}^{01}}{1 - \rho_{ikt}^{01}}\right) = \nu_k^{01} A_{it} \tag{7}$$

and

$$\log\left(\frac{\rho_{ikt}^{1}}{1 - \rho_{ikt}^{1}}\right) = \nu_k^{1} A_{it} \tag{8}$$

Log-Likelihood Function

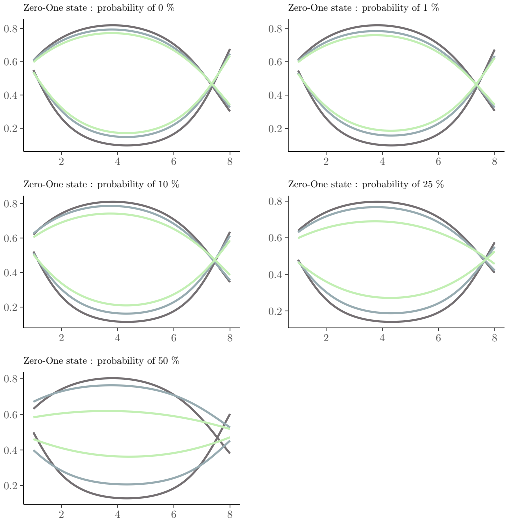
$$l(\psi; y) = \sum_{i=1}^{n} \log \left(\sum_{k=1}^{K} \pi_k g_k(y_i; \beta_k, \delta_k, \zeta_k, \nu_k^{01}, \nu_k^{1}) \right), \tag{9}$$

where

$$g_k(y_i; \beta_k, \delta_k, \zeta_k, \nu_k^{01}, \nu_k^{1}) = \prod_{y_{it}=0} \rho_{ikt}^{01}(1 - \rho_{ikt}^{1}) \prod_{y_{it}=1} \rho_{ikt}^{01}\rho_{ikt}^{1} \tag{10}$$

$$\times \prod_{0<y_{it}<1} \left(1 - \rho_{ikt}^{01}\right) \frac{\Gamma(\phi_{ikt})}{\Gamma(\mu_{ikt}\phi_{ikt})\Gamma((1-\mu_{ikt})\phi_{ikt})} y_{it}^{\mu_{ikt}\phi_{ikt}-1}(1 - y_{it})^{(1-\mu_{ikt})\phi_{ikt}-1}. \tag{11}$$

Numerical Simulation



Real Data Example - data

De la Sablonnière et al. (2020) studied the quality of sleep among Canadians during the COVID-19 pandemic.

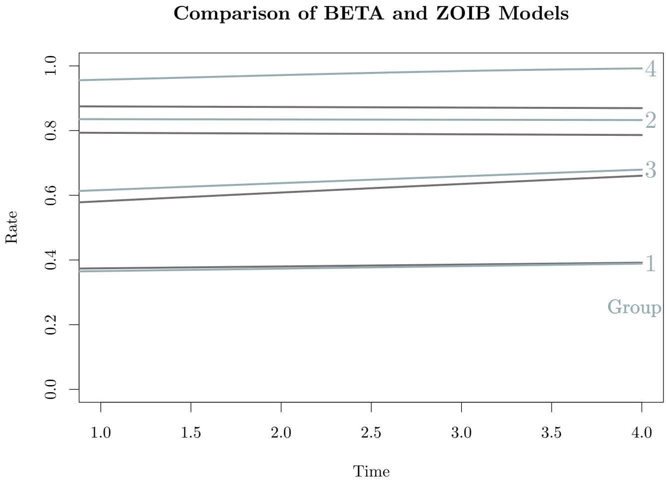
In our study, we used a random subsample of 375 respondents from periods 1 to 4 (Caron-Diotte et al.).

The sleep quality score was measured on a scale from 0 to 10, which we then scaled to a range of 0 to 1.

Since many participants reported high sleep quality, resulting in a significant number of scores of 1, the use of a standard beta regression model was not recommended. Instead, we used the zero-one inflated beta regression model, which is more suitable for this type of data distribution.

Following the work of de la Sablonnière et al. (2020) we calibrated a 4 group solution with linear typical trajectories.

Real data example - comparison



Real Data Example - results

	π_1	π_2	π_3	π_4
ZOIB	13.25	34.18	47.27	5.29
BETA	13.61	25.10	36.02	25.27

	β_{10}	β_{11}	β_{20}	β_{21}	β_{30}	β_{31}	β_{40}	β_{41}
ZOIB	-0.58	0.03	1.44	-0.01	0.29	0.10	-1.44	0.87
BETA	-0.54	0.02	1.36	-0.01	0.22	0.11	1.96	-0.02

	ϕ_1	ϕ_2	ϕ_3	ϕ_4
ZOIB	1.64	3.18	1.93	4.77
BETA	1.62	3.21	2.01	0.65

Conclusion

- For values close to one (group 4), the BETA and ZOIB models diverge significantly. The BETA model overestimates the size of this group compared to the ZOIB model and increases the dispersion parameter ϕ, thus allowing for greater variability within group 4 to capture all individuals.
- The ZOIB model provides supplementary and informative details, particularly regarding the probabilities of belonging to the excess states, which are $\pi_1^{01} = 0\%$, $\pi_2^{01} = 14.34\%$, $\pi_3^{01} = 5.02\%$, and $\pi_4^{01} = 93.32\%$.
- The conditional probabilities of being in the excess one state, given that an individual is in the zero-one state, are $\pi_1^1 = 99.95\%$, $\pi_2^1 = 100\%$, $\pi_3^1 = 100\%$, and $\pi_4^1 = 100\%$. These probabilities are extremely high because the value 0 does not appear in the data set.

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.
- Elmer, J., Jones, B.L.,& Nagin, D.S. 2018: Using the beta distribution in group-based trajectory models. *MC Med. Res. Methodol.*, 18-1.
- Noel, C & Schiltz, J. 2022: TrajeR - an R package for finite mixture models. SSRN paper 4054519.
- Noel, C & Schiltz, J. 2024: Finite Mixture Models for an underlying Beta distribution with an application to COVID-19 data.
- de la Sablonnière and ali (2020: *COVID-19 Canada: The end of the world as we know it?*. Technical Report Université de Montréal.