**Methodological Group**
**Author: Chris Edwards (OU)**

# Group Based Trajectory Modelling: Guide for approach to GBTM in R

**July 2023**

Co-funded by the Erasmus+ Programme of the European Union

Co-funded by the Erasmus+ Programme of the European Union

**Author:**

Chris Edwards



**Project:**

The Complex Trajectories project is funded by European Union Erasmus+ grant agreement No. KA203-082842, 2020-2023.

**Consortium members:**



Co-funded by the Erasmus+ Programme of the European Union

# Contents





Co-funded by the
Erasmus+ Programme
of the European Union

# 1. Introduction

This document provides some guidance on how to approach Group Based Trajectory Modelling (GBTM) in R. My first observation is that, if you have access to STATA, it may be best to initially run the analysis within this, using the separate instructions provided. This approach will ensure your results are comparable with the majority of those reported in this project.

The Complex Trajectories project documentation (Sánchez-Gelabert, 2023) indicates there are several choices for the package to use in R: crimCV, LCMM (Latent Class Mixed Models), flexmix. I selected the lcmm package which appears well maintained and offers a wide range of functions and might therefore, prove a useful long-term package to adopt.

I will assume you are familiar with the basics of using R, including the installation of new packages. If you are not, there are many great resources available from those within the R community. Having begun working with R just five years ago, I found a great way in was to use RStudio and the Tidyverse suite of packages. The code included below reflects this. Within this GBTM analysis, this choice resulted in the occasional requirement to convert from tibbles to conventional R dataframes, for some functions to work. If you, like many, prefer base R there should be no such complication.

# 2. Preparing the data for GBTM

For this analysis, we used the Jointlcmm function in the lcmm package. The most useful guide to this package is the detailed paper written by its creators (Proust-Lima et al, 2017) which includes examples. It is this paper that informed the following analysis.

In our data, we use the amount of academic *credit* (measured in ECTS credits) attained per academic year as a measure of success. Students may study with different outcomes; one of which is passing. They may also choose to study more than one module at a time. These values are coded into a *resultcode* variable, which indicates the different levels of study; or interaction with a programme of study. In addition, we included a survival element, *event1*, which has values either 0 or 1 and indicates whether a student achieves sufficient credit to attain a degree. A final variable is the number of years to attain the credit for a degree, *tevent1*, from the start of study.

The data were then arranged in a table that contains one row per student per academic year. The values for *credit* and *resultcode* vary by year. The values for *event1* and *tevent1* are the same for all rows for each student.



Co-funded by the Erasmus+ Programme of the European Union

When we come to the summary tables for the final Complex Trajectories report, we also need to bring in other data, including demographic and field of study values but these are not necessary for the following functions.

## 3. Explore the heterogeneity

van der Nest et al (2020) recommend plotting all the data and random samples of different size to explore the heterogeneity. We do this below for two sample sizes (as proportions of n). Students' records are de-identified and each student is given a random id from 1 to n. Many such plots could be created, as needed:

```
R> datatable %>% ggplot(aes(acyr,Credit,group=id)) + geom_line()

R> datatable %>% sample_frac(0.006) %>%
ggplot(aes(acyr,Credit,group=id, colour=id)) + geom_line()
```

These plots are sufficient to show there are many different patterns of achievement, in terms of credit gained.

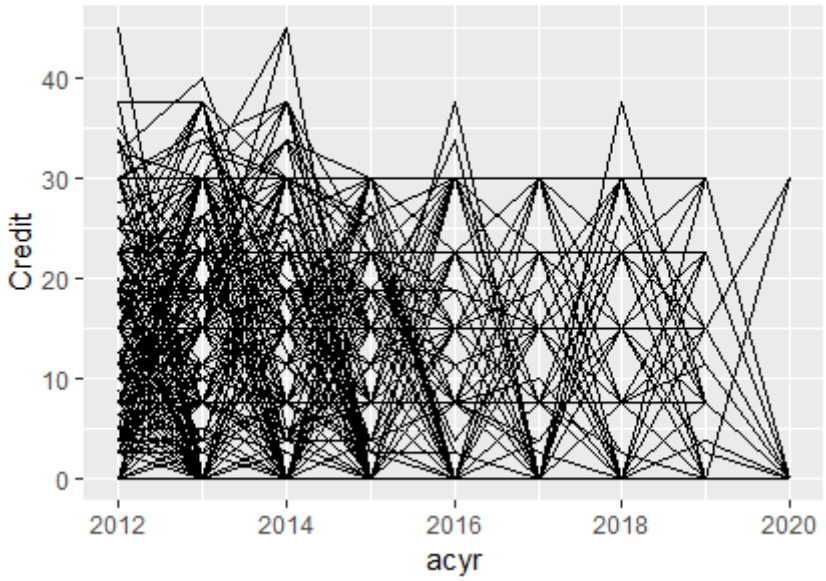


**Figure 1a**. Credit achieved by academic year for each student, for all students in the cohort.



Co-funded by the Erasmus+ Programme of the European Union

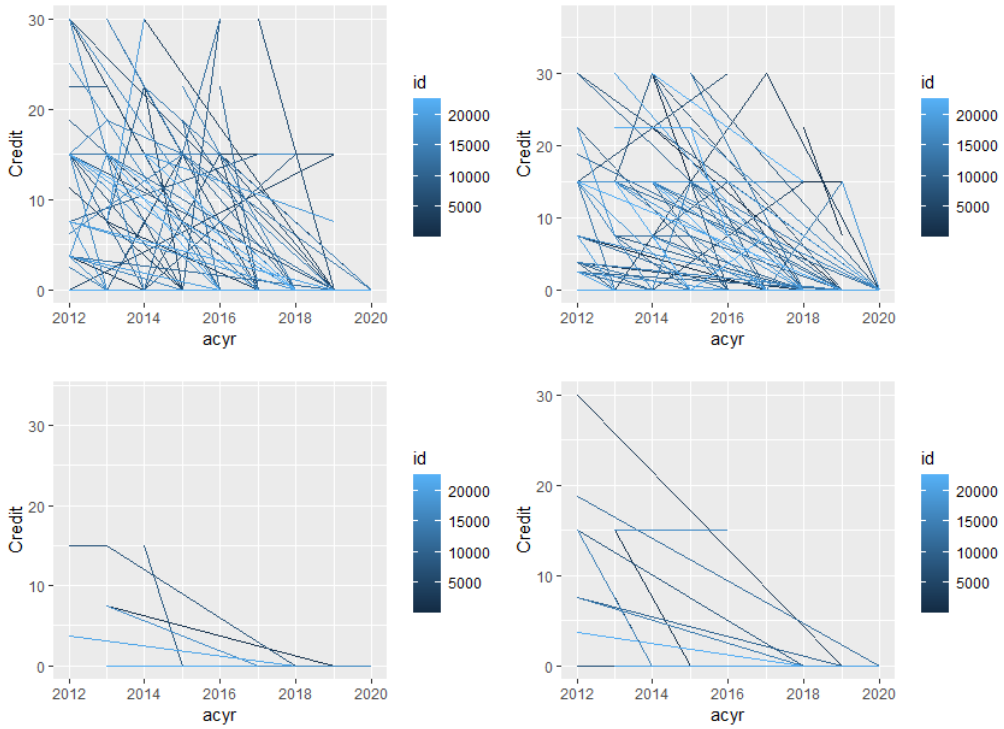


**Figure 1b**. Fiour plots of credit achieved by academic year for each student. The top pair two samples of 2% of the data, and the bottom row are two samples of 0.6% of the data.

# 4. Find the best lcmm function

In the Complex Trajectories guide for GBTM, Sánchez-Gelabert states (2023) that the

> '**optimal group** will be the one that has **the lowest BIC value**, **does not contain groups smaller than 5%** and, at the same time, the **trajectories are significant** according to their polynomial function.'

When setting up functions to test, it is important to ensure that the starting values are properly defined. Proust-Lima et al (2017) include a section setting out how to set up initial values. This is worth reading before you start testing different functions.

As well as running individual functions, it is possible to run a gridsearch. When using this approach, you define the maximum number of iterations you want the function to run and the number of random sets of initial values. This is an example of a gridsearch using the hlme function to find a 3 class (group) model of how cumulative credit varies with time (t). We did not go on to use cummulative credit in our final function.



Co-funded by the Erasmus+ Programme of the European Union

```
R> h13grid <- gridsearch(hlme(cumcred ~t, random = ~t, mixture = ~t,
ng = 3, subject = "id", data = datatable),rep = 30, maxiter = 15,
minit = h1)

R> summarytable(hc13grid)
```

**Table 1**. Summary table for the results of the gridsearch above

| | G | loglik | npm | BIC | %class1 | %class2 | %class3 |
|---|---|---|---|---|---|---|---|
| hc13grid | 3 | -686260 | 12 | 1372632 | 67.76057 | 11.27155 | 20.96788 |

After running many function variations, the one whose results best fit the criteria for OU data is the Jointlcmm function below. The summary table for *Results3* is given in Table 2.

```
Results1 <- Jointlcmm(credit ~t * years * resultcode,random = ~t,
subject = "id", survival = Surv(tevent1,event1) ~1, hazard =
"piecewise", logscale = TRUE, data = datatable, hazardtype = "PH", ng
= 1,   link="beta")

Results3 <- Jointlcmm(credit ~t * years * resultcode, random = ~t,
subject = "id", survival = Surv(tevent1,event1) ~1, hazard =
"piecewise", logscale = TRUE, data = testnewdata1, hazardtype = "PH",
ng = 3, B = Results1, mixture = ~t, link = "beta")
```

**Table 2**. The summary table for *Results3*

| | G | loglik | npm | BIC | %class1 | %class2 | %class3 |
|---|---|---|---|---|---|---|---|
| Results3 | 3 | 998312.9 | 26 | -1996383 | 21.11665 | 71.05977 | 7.823576 |

A fuller summary can be produced with this snippet. The output is presented in the Appendix.

```
R> summary(Results3)
```



Co-funded by the Erasmus+ Programme of the European Union

# 5. Plots

The residuals for the model can be plotted using this code snippet which produces the plot in Figure 2:

```
R > plot(Results3, which = "residuals", var.time = "t")
```

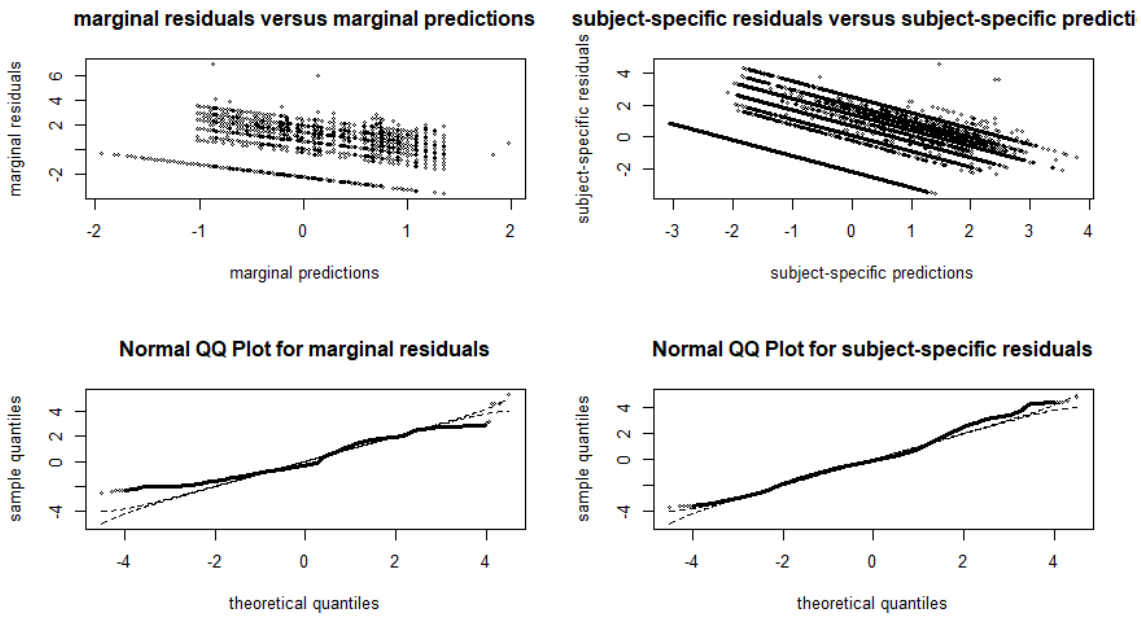


**Figure 2**. Plot of subject specific and marginal residuals for the 3 class model in *Results3*

# 6. Posterior classification

The most likely class for each student is recorded as the variable *class* in the table *pprob* within the *Results3* dataframe. The table can be saved as a tibble, or retained as a dataframe:

```
R> classification.table<-as_tibble(Results3$pprob)
```

The relationship between the class associated with each student and other factors, including demographic values can then be explored to produce the tables for the final report.

# 7. Conclusion

The lcmm package contains a powerful set of functions that can be deployed on the study data on a cohort of students. We have used it to identify three latant classes (groups) within the OU



Co-funded by the Erasmus+ Programme of the European Union

dataset, for those students who study beyond their first year. When all students within the cohort are considered, there appear to be four classes within the 2012/13 OU cohort.

Note: It was necessary to remove the group of students that did not return after the first year from this analysis. This was in order to meet the 5% of cohort minimum for each class. Because the Open University sets no entrance requirements for nearly all its courses, we find this group is larger than for most universities

# 8. Bibliography

Proust-Lima, C., Philipps V., Diakite, A., Liquet, B. (2023). Package 'lcmm': Extended Mixed Models Using Latent Classes and Latent Processes, *The Comprehensive R Archive Network (CRAN)*, https://cran.r-project.org/web/packages/lcmm/lcmm.pdf

Proust-Lima, C., Philipps, V., Liquet, B. (2017). Estimation of extended mixed models using latent classes and latent processes: The R package lcmm. *Journal of Statistical Software,* 78(2). https://doi.org/10.18637/jss.v078.i02.

Proust-Lima C, Séne M, Taylor JM, Jacqmin-Gadda H. (2014) Joint latent class models for longitudinal and time-to-event data: A review. *Statistical Methods in Medical Research*. 23(1):74-90. doi:10.1177/0962280212445839

Sánchez-Gelabert, A. (2023) Group Based Trajectory Modelling: Methodological Guide, *Complex Trajectories Project Document*

van der Nest, G., Lima Passos, V., Candel, M. J. J. M., & van Breukelen, G. J. P. (2020). An overview of mixture modelling for latent evolutions in longitudinal data: Modelling approaches, fit statistics and software. In *Advances in Life Course Research* (Vol. 43). Elsevier Ltd. https://doi.org/10.1016/j.alcr.2019.100323



Co-funded by the Erasmus+ Programme of the European Union

# Appendix

The output from the summary of *Results3*

```
R> summary(Results3)

Joint latent class model for quantitative outcome and competing risks
     fitted by maximum likelihood method

Jointlcmm(fixed = credit ~ t * years * resultcode, mixture = ~t,
    random = ~t, subject = "id", ng = 3, survival = Surv(tevent1,
        event1) ~ 1, hazard = "piecewise", hazardtype = "PH",
    link = "beta", data = Results1, logscale = TRUE)

Statistical Model:
     Dataset: datatable
     Number of subjects: 11427
     Number of observations: 176719
     Number of latent classes: 3
     Number of parameters: 26
     Event 1:
        Number of events: 8278
        Proportional hazards over latent classes and
        Piecewise constant baseline risk function with nodes
        0 8 9 9 9
     Link function for credit: Standardised Beta CdF

Iteration process:
     Maximum number of iteration reached without convergence
     Number of iterations:  100
     Convergence criteria: parameters= 2.6e-06
                         : likelihood= 15
                         : second derivatives= 1

Goodness-of-fit statistics:
     maximum log-likelihood: 526021.09
     AIC: -1051990.17
     BIC: -1051799.23


Maximum Likelihood Estimates:

Fixed effects in the class-membership model:
(the class of reference is the last class)

                       coef Se Wald
intercept class1  -0.71554
intercept class2   1.37266
                  p-value
intercept class1
intercept class2

Parameters in the proportional hazard model:

                              coef Se
event1 log(piecewise1) -10.04642
event1 log(piecewise2)  -7.94358
event1 log(piecewise3)  -1.38629
event1 log(piecewise4) 109.88880
event1 SurvPH class1     1.03102
event1 SurvPH class2     6.84790
```



Co-funded by the
Erasmus+ Programme
of the European Union

```
                        Wald p-value
event1 log(piecewise1)
event1 log(piecewise2)
event1 log(piecewise3)
event1 log(piecewise4)
event1 SurvPH class1
event1 SurvPH class2

Fixed effects in the longitudinal model:

                                      coef
intercept class1 (not estimated)         0
intercept class2                  -1.05978
intercept class3                  -0.11528
t class1                          -0.11878
t class2                          -0.21694
t class3                           0.06947
years                              0.13246
resultcode                         1.08206
t:years                           -0.00079
t:resultcode                      -0.25687
years:resultcode                  -0.11329
t:years:resultcode                 0.04030
                                 Se
intercept class1 (not estimated)
intercept class2
intercept class3
t class1
t class2
t class3
years
resultcode
t:years
t:resultcode
years:resultcode
t:years:resultcode
                                 Wald
intercept class1 (not estimated)
intercept class2
intercept class3
t class1
t class2
t class3
years
resultcode
t:years
t:resultcode
years:resultcode
t:years:resultcode
                                 p-value
intercept class1 (not estimated)
intercept class2
intercept class3
t class1
t class2
t class3
years
resultcode
t:years
t:resultcode
years:resultcode
t:years:resultcode
```



Co-funded by the Erasmus+ Programme of the European Union

```
Variance-covariance matrix of the random-effects:
          intercept       t
intercept   0.99072
t          -0.13348 0.01798

Residual standard error (not estimated) = 1

Parameters of the link function:

           coef Se Wald p-value
Beta1  -1.14968
Beta2   4.53990
Beta3   0.75472
Beta4   0.00241
```



Co-funded by the
Erasmus+ Programme
of the European Union