# **Supplementary material**

Longitudinal data, trajectories and telemonitoring: how to analyze them?

Example of sleep data

## Codes R et codes Python

Mettre un lien gitlab ~~ou mettre le code R puis le code Python en 2 parties ou mettre les documents d’origine (.R et .py) en SM dans des fichiers autres.~~

## ~~Codes Python (ou les mettre en doc .py à part)~~

## Additional information on models

(C) LCA model

*Description* – The LCA model is a semi-parametric, finite mixture modeling model for cross-sectional categorical data (non-longitudinal version of the LTA model). This model determines unobserved categorical variables by dividing the population into mutually exclusive and collectively exhaustive latent classes. It is particularly suitable for an exploratory approach. Parameters are estimated by maximum likelihood or Bayesian methods. The probability of class membership is evaluated, enabling each individual to be assigned to a class. The contribution of each covariate can be modeled within each class, and the probability of class membership depends on the values or categories of the covariates. It is preferable to recode variables with as few categories as possible when the data originally had too many categories. In the case of very large data sets, where the LCA model requires significant statistical prerequisites, it is preferable to subdivide the data into random sub-samples to determine the optimum number of classes. The BIC criterion and the elbow method are used to determine this optimal number. Then, once this number has been obtained, the LCA model is rerun on the total population, setting certain parameters such as the maximum number of iterations (N = 1000) or the number of times the model is estimated with different probabilities of conditional class responses (N = 25). A training/testing (90%) and validation (10%) method can also be run. Missing data are also taken into account, using deletion or multiple imputation methods, for example. Finally, a χ² test for categorical variables and a Kruskal-Wallis test for continuous variables are applied to study differences between classes. These latent variables can then be imputed in regression models, for example, to study associations between an outcome and these populations, its characteristics.

*Process –* The LCA model follows an iterative process based on a priori decisions. These decisions depend on the researcher's knowledge of the field of study. Implementation requires a few steps:

1. Model characterization: decide on additional covariates and their coding, and on participant selection (sample size, number of values per participant);
2. Choice of number of latent covariate classes: based on results of hypothesis testing;
3. Model estimation: prior to analysis, the model should be defined using maximum likelihood (EM algorithm) or Bayesian methods (MCMC algorithm). When the sample size is less than 300 or between 300 and 500 individuals, Monte Carlo simulation is recommended;
4. Evaluate the optimal number of classes for the selected model: The AIC, BIC, sample size-adjusted BIC, Lo-Mendell-Rubin and Vuong-Lo-Mendell-Rubin criteria are tested to determine the best final model. A low criterion value corresponds to the best model. Class interpretation is also an important criterion;
5. Evaluation of the final model: various tests are carried out, such as entropy calculations to measure class separation. An entropy value close to 1, greater than 0.8, corresponds to the best model. Or calculating the average a posteriori probability of latent classes. Once the model is well defined, covariates can be added;
6. Interpret the final model: the aim is to prove the robustness of the model. Random starting points can be implemented, class verification or the reproducibility of analyses for other data can be analyzed.

Note: before starting the analyses, it is essential to check the data (check for outliers or implausible values, transform the data: when the categorical variable has a category with less than 10% of values, this variable must be excluded or recoded, standardize, evaluate the correlation of variables (>0.5, requires special checking) to avoid bias and misclassification of individuals).

*Example (continued) –*

The item-response probabilities were:

|  |  |  |  |
| --- | --- | --- | --- |
|  | **[0h; 2h[** | **[2h; 4h[** | **≥4h** |
| **T1** | | | |
| Cluster 1 | 0.10 | 0.58 | 0.32 |
| Cluster 2 | 0.04 | 0.30 | 0.67 |
| **T2** | | | |
| Cluster 1 | 0.15 | 0.42 | 0.43 |
| Cluster 2 | 0.13 | 0.29 | 0.57 |
| **T3** | | | |
| Cluster 1 | 0.24 | 4.40e-10 | 0.76 |
| Cluster 2 | 0.05 | 0.57 | 0.38 |
| **T4** | | | |
| Cluster 1 | 0.13 | 0.09 | 0.77 |
| Cluster 2 | 0.05 | 0.64 | 0.31 |
| **T5** | | | |
| Cluster 1 | 0.30 | 6.54e-94 | 0.70 |
| Cluster 2 | 0.05 | 0.54 | 0.41 |

(E) GMM model

*Description* – This parametric model is a finite mixture model that analyzes both continuous and categorical longitudinal tracking data. It provides an alternative representation of trajectory changes often modeled via polynomial models, and is particularly effective in representing complex trajectory shapes. The model describes different longitudinal trajectory changes by defining data subgroups and defining changes within subgroups. These subgroups correspond to classes, and this class variable is latent or unobserved. More precisely, the GMM model estimates the average trajectory of the curves for each class, and allows for variations between individuals in the same class. For each trajectory, the model estimates the origin coordinates, slope and variance of the trajectory data by maximizing the log-likelihood function. The random origin coordinates capture each individual's initial level on the measured data, while the random slope captures each individual change in these repeated measurements as a function of time. Then, when assigning the names of each latent class, it is necessary to take into account the unique differences of each pattern (different trajectory shapes, different levels: high, low, medium, different scales). Once the trajectory is well defined, it can be used as a dependent or non-dependent variable to explore predictors of the health trajectory and explore their contribution to future health outcomes.

There are, however, a few important elements to consider, such as the fact that the final trajectory depends on the initial parameters. Random effects can thus be introduced into the GMM model, and their distribution as well as fixed effects can be specified for each class. Covariates (time-varying or not) are accepted to explore the relationship between the original characteristics and the trajectory pattern, but only once the final model has been defined. They must not modify the profile of the selected model, but will influence the final model. Predictors can also be added to determine the latent variable.

*Process* – The GMM model follows an iterative process based on a priori decisions:

1. Hypothesize the expected number of latent classes and the shape of the trajectory curve for each class (linear, quadratic, cubic, ...);
2. Characterize the model: make decisions on the curve's trajectory parameters (intercept, variance and slope covariance);
3. Model estimation: by maximum likelihood (EM algorithm) or Bayesian methods (MCMC algorithm), where the unobserved class classification variable is treated as a latent variable. These algorithms use an iterative process to obtain parameter estimates and a posteriori estimates of the probability of each individual belonging to each possible class with respect to the observed data. Each individual belongs to the class for which it has the highest probability of membership;
4. Model selection: choose the model giving the best representation of the observed data by various tests (Lo-Mendell Rubin adjusted likelihood ratio test (LRT, p < 0.05), Parametric bootstrapped likelihood ratio (BLRT, p < 0. 05), BIC (minimum BIC corresponds to the best model), AIC (minimum AIC corresponds to the best model), convergence and ability to distinguish the different classes (entropy close to 1.0)) and validation of model parameters (proportion of the population in each class (> 5%), probability of belonging to a class (close to 1.0), frugality and usefulness of latent classes for the study).
5. Mixed model

*Description* – The mixed model can be linear or more generalized. In the case of the linear mixed model, longitudinal or cluster data are analyzed. It estimates the relationship between dependent variables and the fixed and random effects of independent variables, considering both means and covariances. If the number of predictors is too large, inference and prediction become too complex and infeasible. Fixed effects estimates can be penalized and random effects selected. For example, the estimates of the fixed and random effects and the covariance structure of the random effects are chosen simultaneously in a model with a penalty function. However, for high dimensions, it is necessary to reduce the data dimension before using the model, as the EM algorithm will not be efficient, plausible, fast and will converge slowly. It is also possible to calculate the maximum likelihood function and then use this algorithm. A second example is the selection of important fixed and random effects, independently in 2 separate models. For this, the use of Proxy matrices enables the unknown variance-covariance structure of the random effects to be taken into account during the selections, and the penalized likelihood equation for the parameters of the fixed effects to be minimized. The random effects are then selected using Bayesian methods. For high-dimensional data, the dimension of the fixed effects must be smaller than the sample size before the model is used. A final example estimates and selects the fixed and random effects and the covariance structure of the selected random effects simultaneously in a linear mixed-effects model using 2 penalty functions (by a modified log-likelihood using REML (Restricted Maximum Likelihood)).

The generalized model is a hybrid between the linear mixed model and the generalized linear model. It can include repeated-measures data for which the assumption of independence of observations is unverifiable, non-Normal dependent variables measured more than once, mixed data (continuous and categorical) that are independent thanks to the construction of a pseudo-dependent variable encompassing all measurements of each dependent variable, and the construction of a pseudo-independent variable to identify the origin of each instance of the pseudo-dependent variable.

1. GBTM model

*Description* – The GBTM model is a finite mixture model for continuous or categorical longitudinal data with a discrete distribution such as a Bernouilli, Poisson or Normal distribution of its population. This model assumes that there is no variation between individuals in the same class (no intra-class variance), and allows classes to be distinguished within a homogeneous population. It also assumes that the residual variance is constant over time and the same for each class. For each dependent variable, there are 2 or more groups with moderate intra-group variation and high inter-group variation. These groups cannot be directly created with precision from an independent variable, so a categorical latent class variable is introduced to the model to determine group membership. At each level of the latent variable, a trajectory specific to the dependent variable is estimated. Overall, the proportion of the population belonging to each class is estimated, the probability of belonging to the various classes is calculated for each individual (posterior probability of the group) and then the individuals are assigned to a class, the one for which they have the highest probability of belonging. For this, the parameters are estimated by maximum likelihood. This probability of membership can then be used for prediction. Time-dependent and non-time-dependent covariates can be added to the model.

*Process* – The GBTM model follows an iterative process based on a priori decisions:

1. Hypothesize the expected number of latent classes and the shape of the trajectory curve for each class (linear, quadratic, cubic, ...);
2. Model characterization: initially, it is necessary to test a model with 1 class, then progressively adjust the maximum number of subgroups. This number must be greater than the expected number of classes. To determine the optimum number, compare models for 1 trajectory (for 1 class) with a quadratic curve and a linear curve. Then, if the quadratic model is good, compare 2 trajectories with quadratic components. To test the models, the BIC criterion is used (differences between 2 adjacent models > 10), while the average a posteriori probability (> 70%) and relative entropy (> 0.8) are evaluated to test the correct number of individual trajectories among the identified classes;
3. Model estimation: by maximum likelihood or Bayesian methods;
4. Model selection: knowledge of the study domain is the first step in model selection. It is necessary to opt for a useful and parsimonious model, and to have a close correspondence between the estimated probability of each class and the proportion of individuals attributed to each class according to the rule of maximum membership probability. Specifically, the average of the a posteriori membership probabilities must be greater than or equal to 0.7, and there must be at least 5% of individuals in each class. There must also be a reasonable confidence interval and a BIC difference between 2 models for a different number of class trajectories.
5. LTA model

*Description* –

*Process* – The LTA model follows an iterative process based on a priori decisions based on knowledges of the studied field:

1. Choice of the latent class number based on the hypothesis test results;
2. Model characterization: decide on the temporal invariance of the probability of item-response, on the invariance of the measures for the transition probability and on the covariates add;
3. Model estimation: the choice of the model must be made before the models’ application. The parameters are estimated by maximum likelihood (EM algorithm) or by Bayeian approach (MCMC algorithm).
4. Model selection: by AIC or BIC criteria.

*Example (continued)*  –

The conditional response probabilities:

|  |  |  |
| --- | --- | --- |
| **State** | **1** | **2** |
| **Category** |  |  |
| **[0h;2h[** | 0.03 | 0.22 |
| **[2h;4h[** | 0.33 | 0.59 |
| **≥4h** | 0.63 | 0.19 |

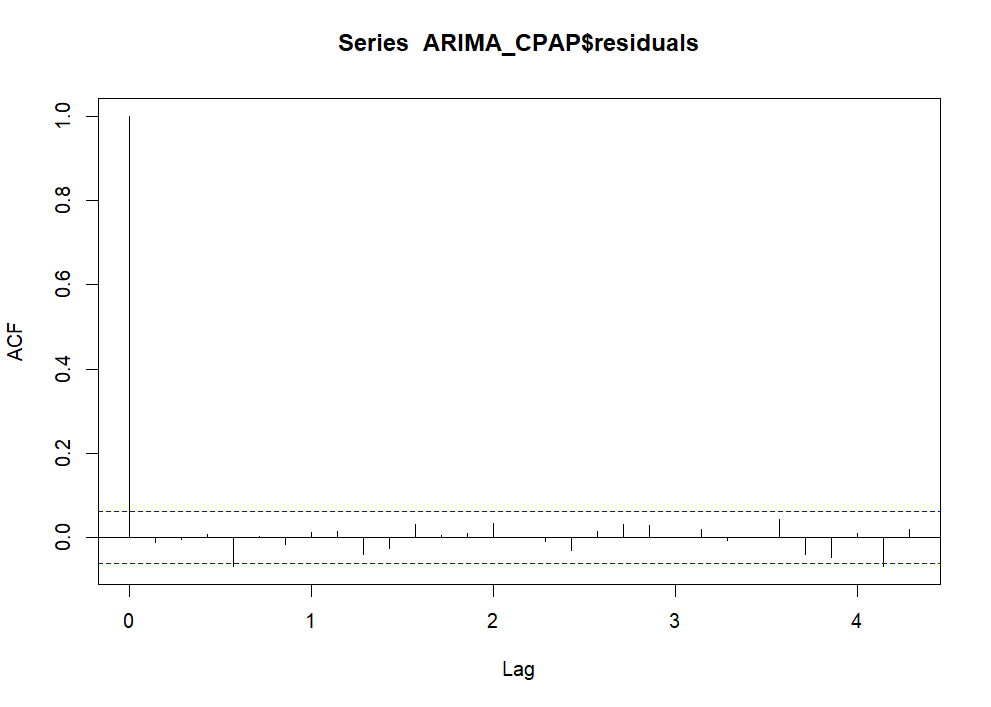
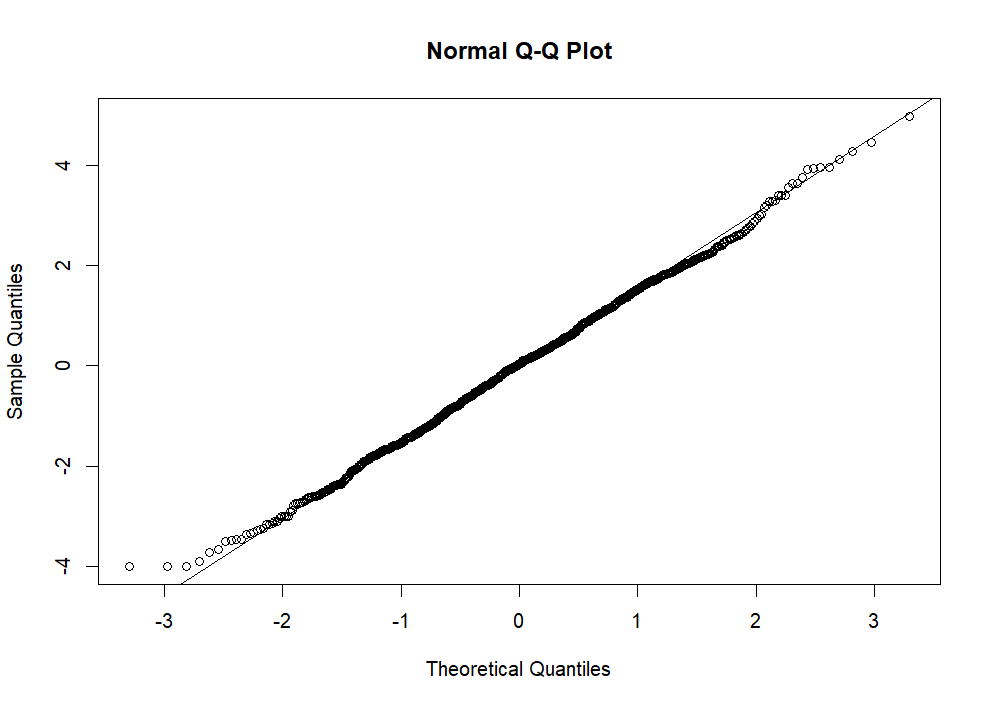
1. ARIMA & CCF model

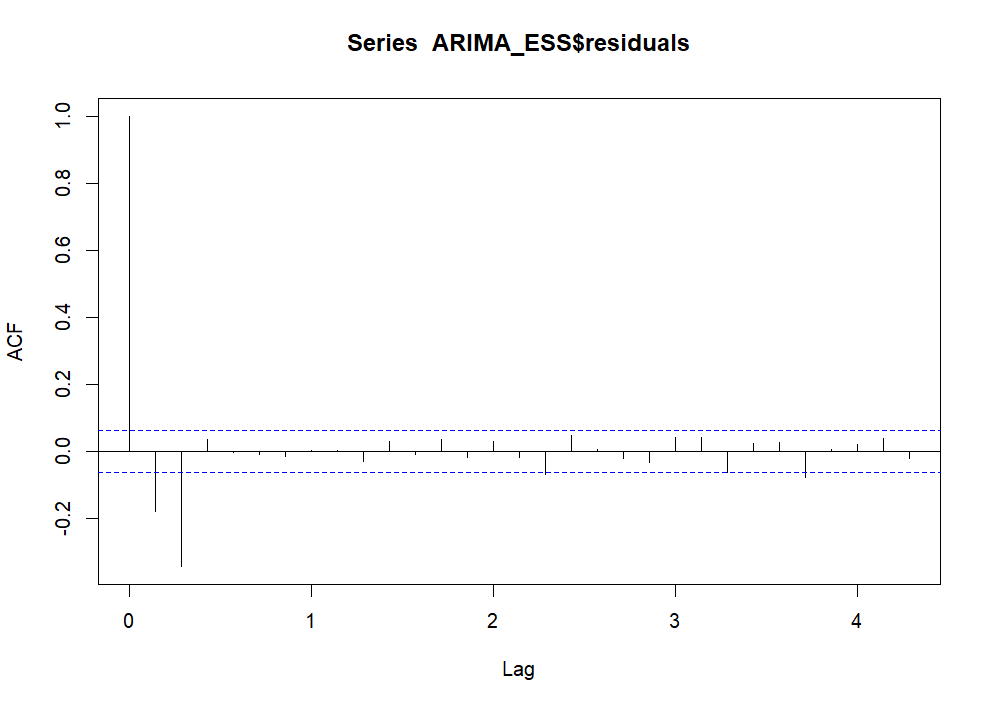
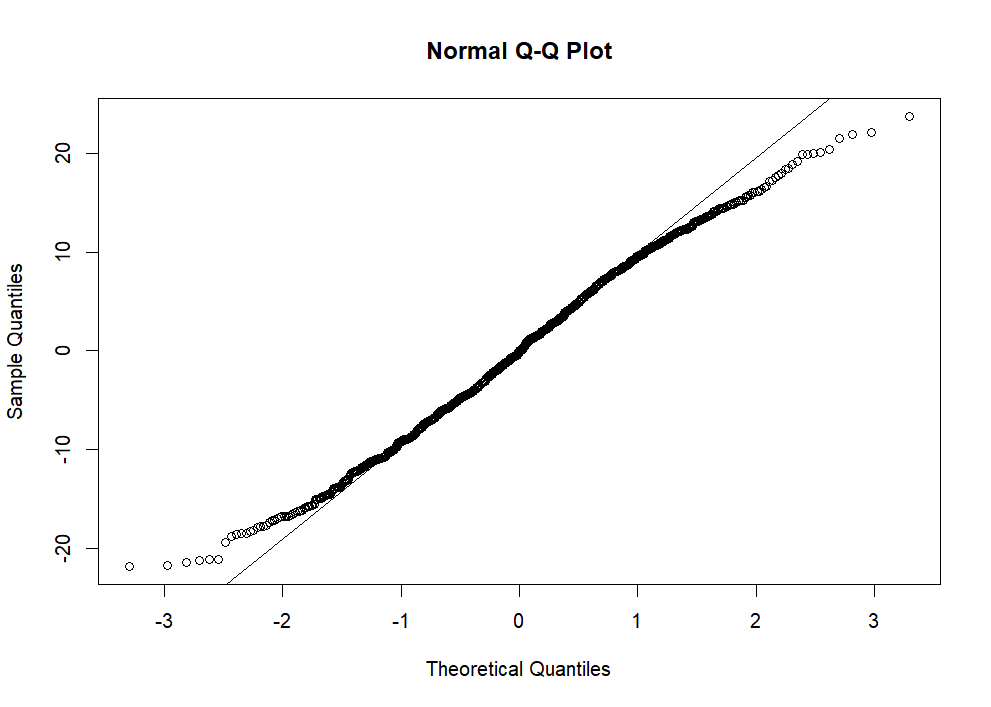
*Description* –

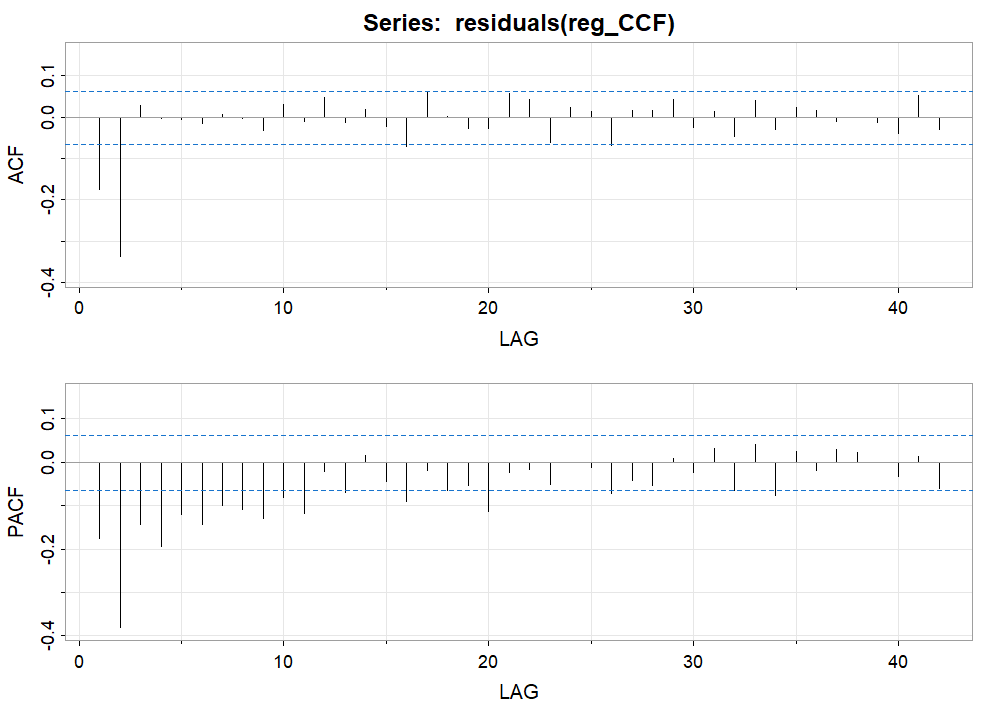
*Process* – First, the ARIMA model is defined and then the Cross-Correlation function is applied:

1. Analyze the ARIMA model: adjust the ARIMA model with time series, i.e. analyze the stability and the seasonality of the data (ACF, PACF) then estimate and test the model (AIC, BIC, Box Ljung statistic);
2. Construction and validation of the ARIMA model: divide the data into 2 parts to perform the method of training/test and validation. The 1st part is used to define the model and the 2nd to validate it;
3. Model application: validation of the prediction and the execution of the model by RMSE. The trajectory of the time series evolution can be analyzed and predicted;
4. Evolution of the 2 time series: perform cross-correlation functions.

*Example (continued)* –

CPAP adherence validation:

ESS score validation:



1. DTW model

*Description* –

*Process* – The DTW model follows 3 steps:

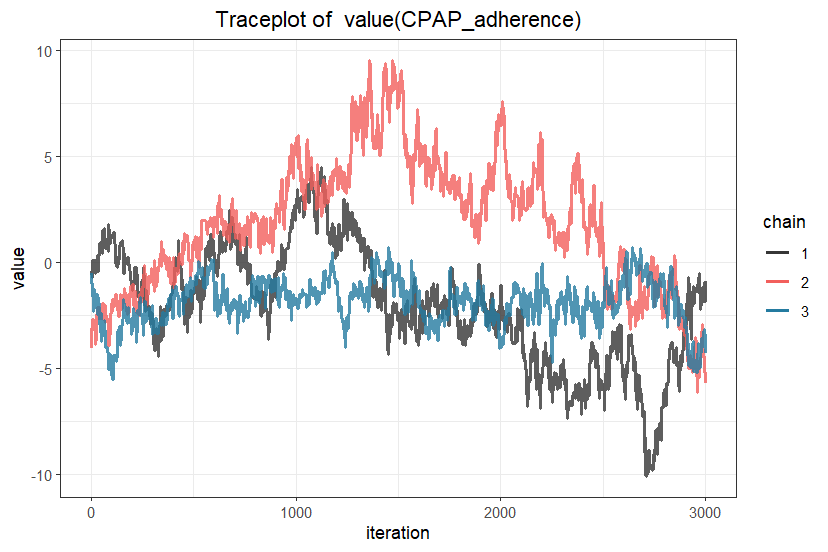
1. Define dissimilarities between clusters: determined by the matrix of individual dissimilarity by paire (or distance matrix) as, for example, those of means, completes or those of Ward;
2. Choose the number of clusters: using the index of variance coefficient which is an accuracy measure;
3. Determine centroids: for that, this model uses the calculation of the DTW barycenter mean which is a strategy of calculation of the global mean and which provide centroids for groups of time series well adapted to the DTW dissimilarity. This calculation depends on the initial step, the initial data.
4. Joint model

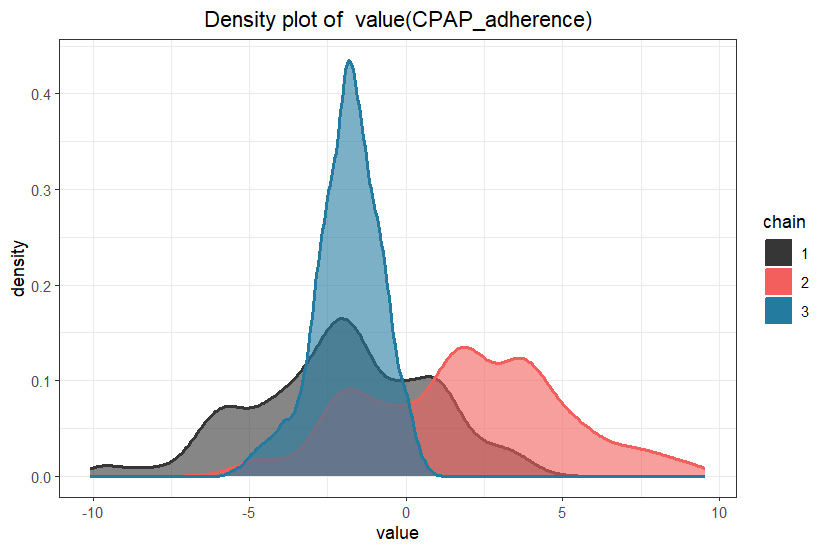
*Description* –

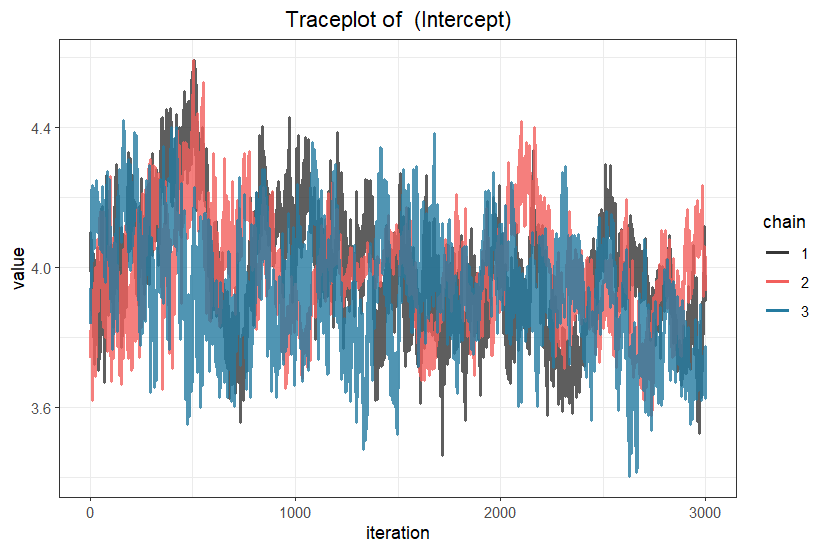
*Process* – The joint model follows 3 overall steps:

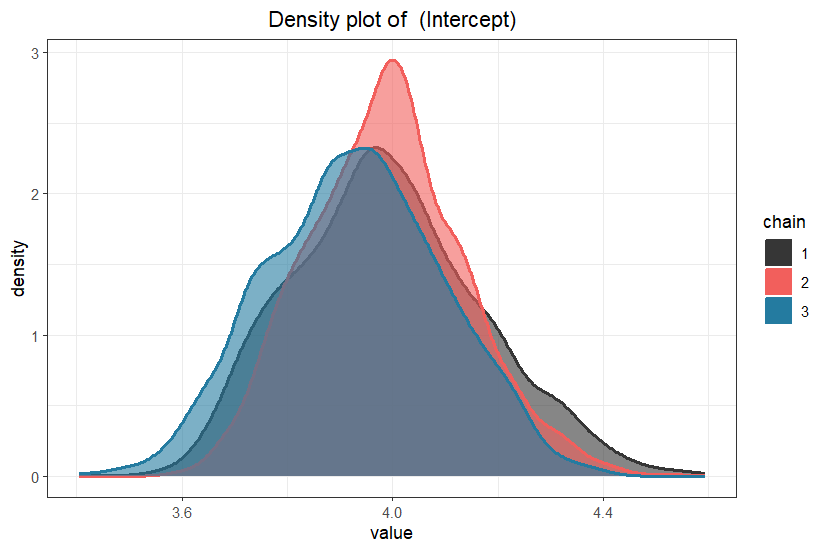
1. Define a mixed regression model;
2. Define a Cox, survival model;
3. Model the evolution of the 2 models at the same time.

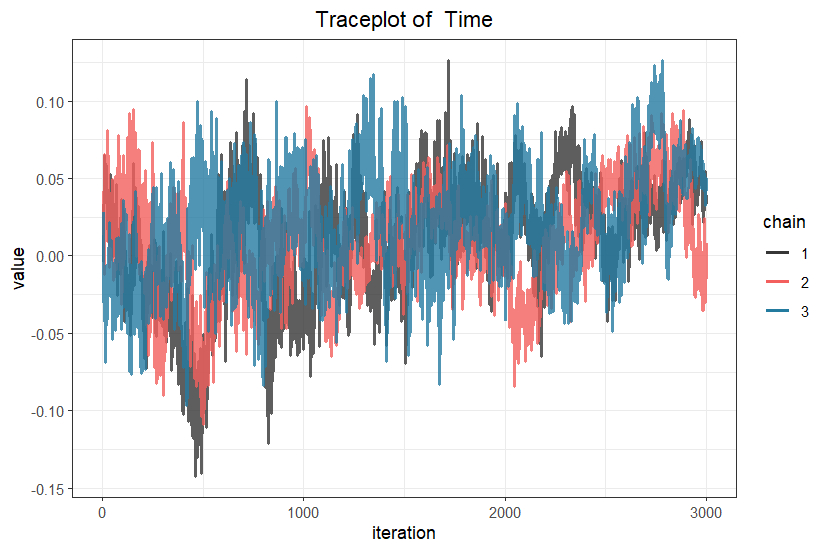
*Example (continued)* –

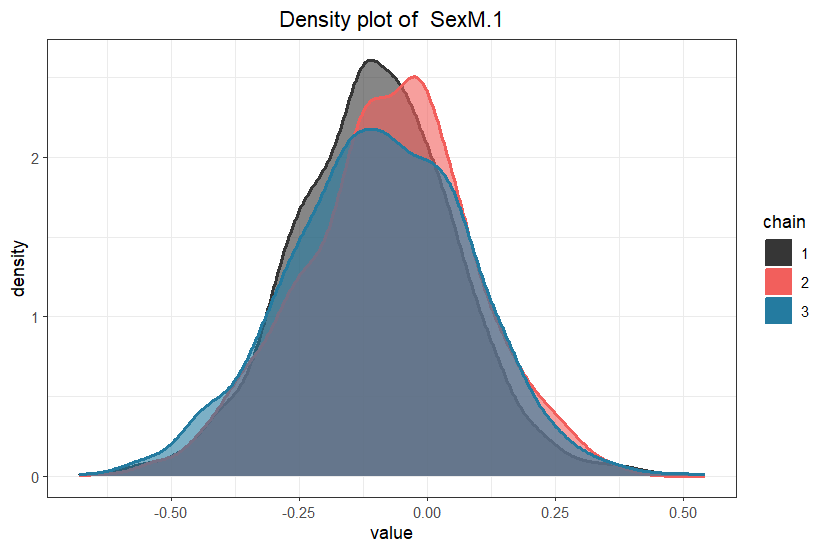
Validation of the joint model (density; sampling behavior, mixing across chains and convergence):

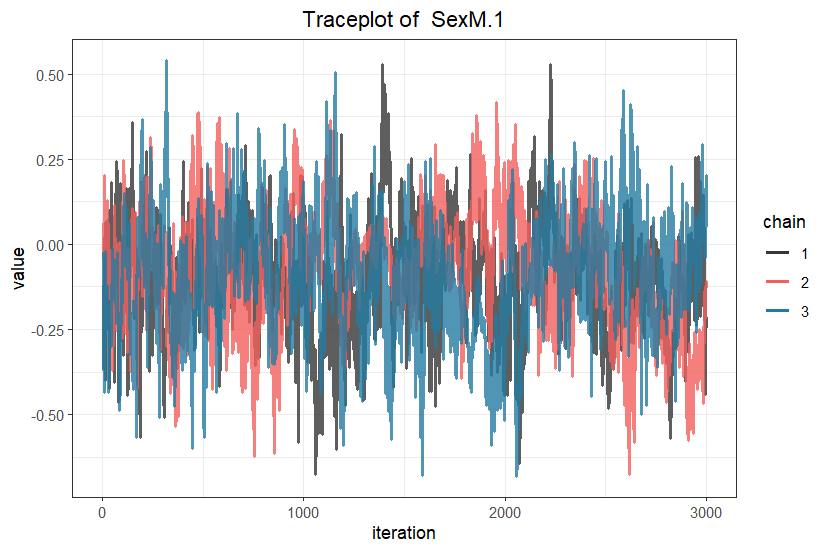


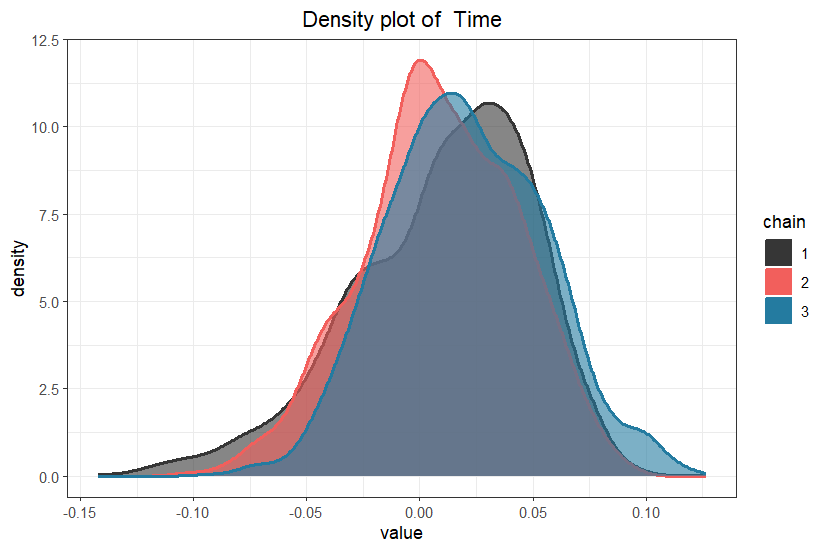












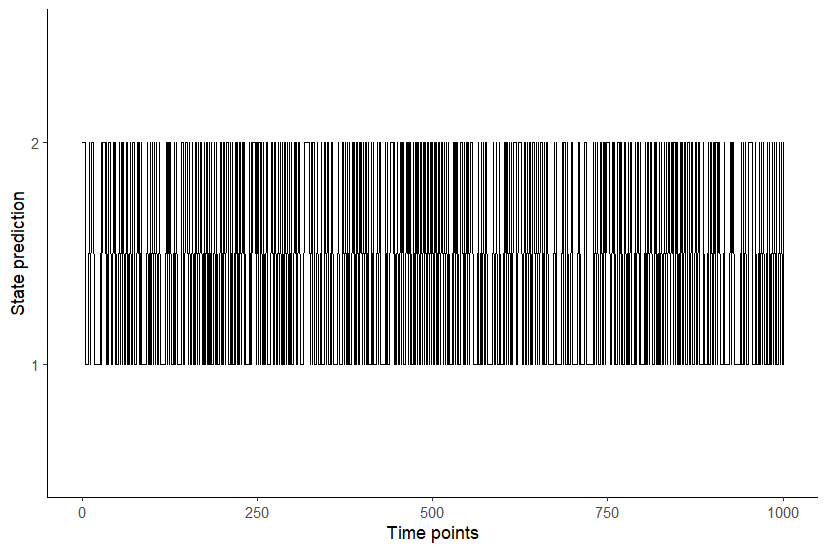
1. Hidden Markov model

*Description* –

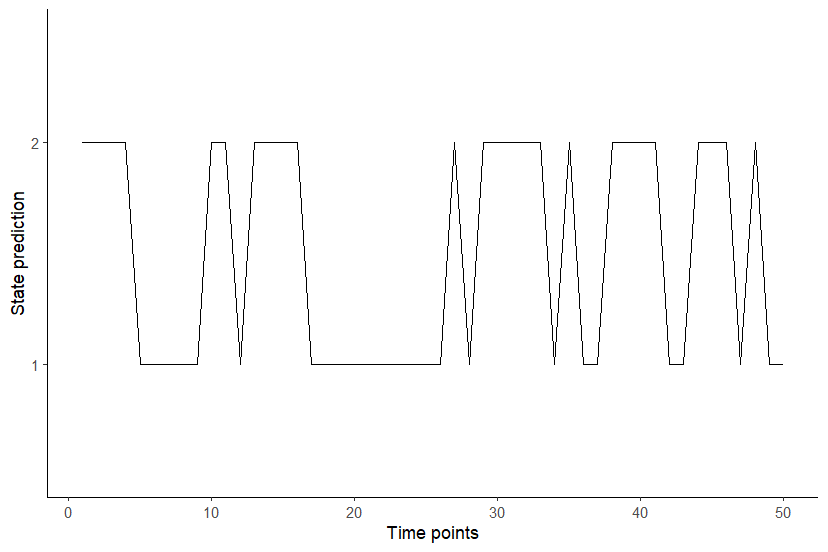
*Process* – The Hidden Markov model follows 3 steps:

1. Markov hypotheses: the transition probability from one state to another has not to depend on previous visited states or on spent time among the actual state and it supposes that the transition rate have to be constant over time and common to all individuals;
2. Estimation of the model parameters: estimate the rate of transitions of a transition probability which are directly estimable according to the data. A Bayesian method using MCMC algorithm could be a solution;
3. Random effect and individuals’ covariates inclusion: assuming that the transition rates are not fixed but modifiable;
4. Model the unobserved responses: several scenarios are possible according to the missing data (MAR, MCAR, MNAR, LOCF-like missingness) and the lost of follow-up.

*Example (continued)* –

State prediction: For all time points

To be more readable, we selected the first 50 time points:



## Results with Python

1. ANOVA model
2. χ² model
3. LCA model
4. K-means model
5. GMM model
6. Mixed model
7. GBTM model
8. LTA model
9. ARIMA & CCF model
10. DTW model
11. Joint model
12. Hidden Markov model