# **Supplementary material**

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

**Example of sleep data**

**Title for CHEST (50 characters): How to analyze longitudinal data? Sleep modeling**

Authors

Affiliations

## e-Appendix 1. Codes in R and Python

https://gitlab.com/opheliecoiffier/methodo\_stat\_donnees\_long

## e-Appendix 2. Additional information on statistical methods

(C) LCA method

*Description* – The LCA method is a semi-parametric, finite mixture modeling model for cross-sectional categorical data (non-longitudinal version of the LTA method). 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 method 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 method 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 method 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 (additional information) –*

A seed was implemented for the reproducibility of the analysis. The optimal number of clusters were found using BIC and AIC criterion.

|  |  |  |  |
| --- | --- | --- | --- |
|  | **2 clusters** | **3 clusters** | **5 clusters** |
| **AIC** | 493.4 | 497.6 | 517.3 |
| **BIC** | 533.5 | 558.8 | 620.5 |

Finally, the best model was consisted of 2 clusters with a probability of, respectively, 0.37 and 0.63 to belong to the 1st and 2nd cluster.

The item-response probabilities were:

|  |  |  |  |
| --- | --- | --- | --- |
|  | **[0h; 2h[** | **[2h; 4h[** | **≥4h** |
| **T1** | | | |
| Cluster 1 | 0.19 | 0.58 | 0.22 |
| Cluster 2 | 0.03 | 0.35 | 0.62 |
| **T2** | | | |
| Cluster 1 | 0.21 | <0.01 | 0.79 |
| Cluster 2 | 0.12 | 0.43 | 0.45 |
| **T3** | | | |
| Cluster 1 | 0.27 | 0.16 | 0.57 |
| Cluster 2 | 0.08 | 0.41 | 0.51 |
| **T4** | | | |
| Cluster 1 | <0.01 | <0.01 | 1.00 |
| Cluster 2 | 0.10 | 0.55 | 0.35 |
| **T5** | | | |
| Cluster 1 | <0.01 | <0.01 | 1.00 |
| Cluster 2 | 0.18 | 0.43 | 0.40 |

(D) K-means method

*Example (additional information) –*

The first step was to transform the long format data into ClusterLongData format. Then, the kml function performed the clustering. In this example, 15 redrawings for each of the clusters and a test with 2 to 6 clusters were implemented. Finally, according to the Calinski-Harabatz score, the model with 2 clusters was the best model.

(E) LTA model

*Description* –The LTA model is a finite, semi-parametric mixture model of longitudinal data that analyzes changes in categorical variables, contingency tables, measured over time. The LTA model is therefore based on latent class theory, where homogeneous subgroups of individuals can be identified by their response to repeated measurements over time. It uses observed data from a categorical dataset to define a latent variable at each time point. However, when the categorical variable contains too many categories, it is recommended to reduce the number of categories as much as possible, and it is also recommended to use this model only with a number of time points less than or equal to 6. In addition, this model allows each individual to change classes over time. Thus, it studies the probability of an individual's transition from one class at one time point, to another class at the next time point. To do this, it studies the transition probability matrix between two consecutive time points. Estimation of the model therefore provides us with information on the probability of belonging to the latent status at the 1st time point, the proportion of the population within each latent class at each time point, the conditional probability of moving from one latent status to another over time, the conditional probability of belonging to the latent status item, and, at any time point, the probability of the classes a posteriori. Model parameters are estimated by maximizing the likelihood function using the EM algorithm, or by a Bayesian method using the MCMC algorithm. Please note, however, that the MCMC algorithm is not recommended when the sample size is too small or when the item response probability is far from 0 or 1.

For this model, covariates are accepted, but the number of classes must be known before they are added, to avoid potential changes in the number of classes in the final model with and without covariates.

*Process* – The LTA model follows an iterative process based on a priori decisions based on knowledge of the domain under study:

1) Choice of the latent class number based on the results of the hypothesis test;

2) Model characterization: decision on the time invariance of item response probability, on the invariance of transition probability measures and on the addition of covariates;

3) Model estimation: the choice of model must be made before the models are applied. Parameters are estimated by maximum likelihood (EM algorithm) or by the Bayesian approach (MCMC algorithm).

4) Model selection: by AIC or BIC criteria.

*Example (additional information)* –

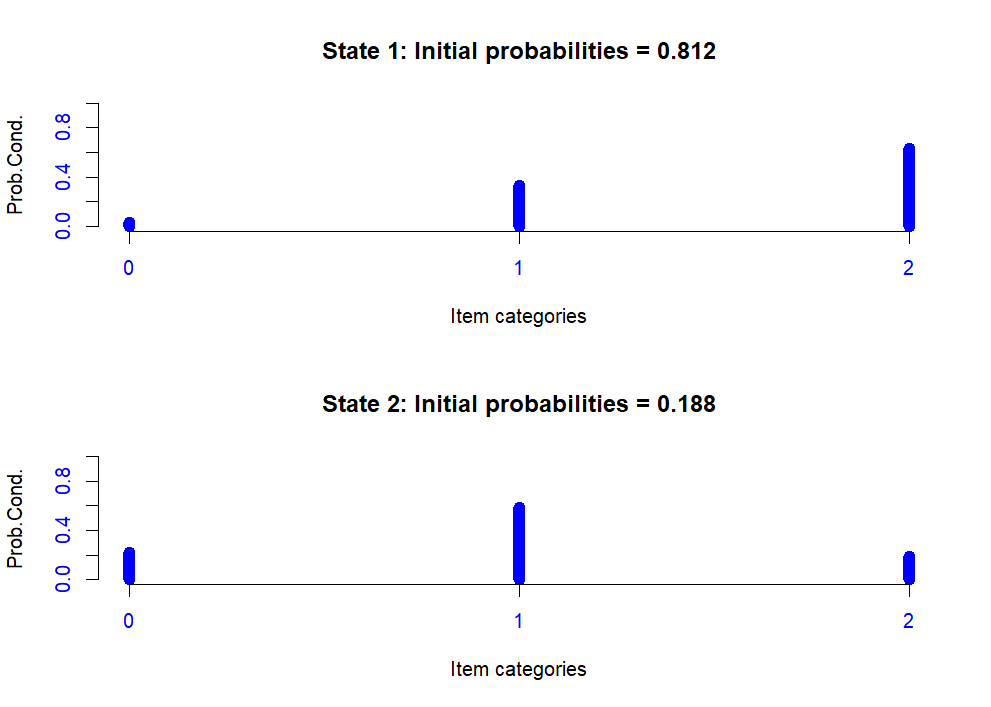
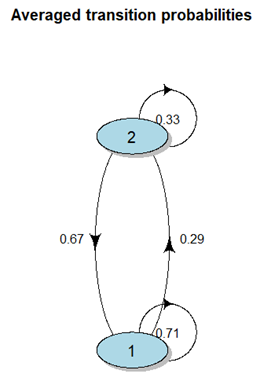
The model used long format data. For the reproducibility of the analysis, a seed was implemented. We test the model for 2, 3 and 4 clusters. According to the Loglikelihood and the BIC or AIC criteria, the best model created 2 clusters.

|  |  |  |  |
| --- | --- | --- | --- |
|  | **2 clusters** | **3 clusters** | **4 clusters** |
| **AIC** | 46461.5 | 46469.2 | 46487.1 |
| **BIC** | 46474.9 | 46495.9 | 46531.1 |
| **Loglikelihood** | -23223.8 | -23220.6 | -23220.6 |

The conditional response probabilities were described below.

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

**~~Figure x: Averaged transition probability and conditional response probabilities to belong to clusters using LTA model.~~**



1. GBTM method

*Description* – The GBTM method 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 method 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 method 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.

*Example (additional information) –*

First, we performed GBTM method with linear curve, quadratic curve and cubic curve with 2 clusters to find the best type of curve. The chosen model was the quadratic curve, according to BIC (difference should be > 10) and loglikelihood criteria. Then, tests on number of clusters were applied with quadratic curve for 2, 3 or 4 clusters. According to BIC, Average Posterior Probability (better if ≥ 0.7) and Proportion of assignment parameters, the model with 2 clusters was the best model.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | **Linear curve – 2 clusters** | **Quadratic curve – 2 clusters** | **Cubic curve – 2 clusters** | **Quadratic curve – 3 clusters** | **Quadratic curve – 4 clusters** |
| **BIC** | 913.5 | 918.1 | 922.3 | 932.4 | 945.2 |
| **Loglikelihood** | -443.0 | -441.4 | -439.6 | -438.8 | -435.4 |
| **Average Posterior Probability** |  | Cluster 1 = 0.84  Cluster 2 = 0.82 |  | Cluster 1 = 0.78  Cluster 2 = 0.86  Cluster 3 = 0.75 | Cluster 1 = 0.83  Cluster 2 = 0.76  Cluster 3 = 0.93  Cluster 4 = 0.82 |
| **Proportion of assignment** |  | Cluster 1 = 0.58  Cluster 2 = 0.42 |  | Cluster 1 = 0.10  Cluster 2 = 0.40  Cluster 3 = 0.50 | Cluster 1 = 0.22  Cluster 2 = 0.10  Cluster 3 = 0.06  Cluster 4 = 0.62 |
| **Estimate probabilities** |  | Cluster 1 = 0.56  Cluster 2 = 0.44 |  | Cluster 1 = 0.13  Cluster 2 = 0.44  Cluster 3 = 0.43 | Cluster 1 = 0.26  Cluster 2 = 0.13  Cluster 3 = 0.06  Cluster 4 = 0.55 |

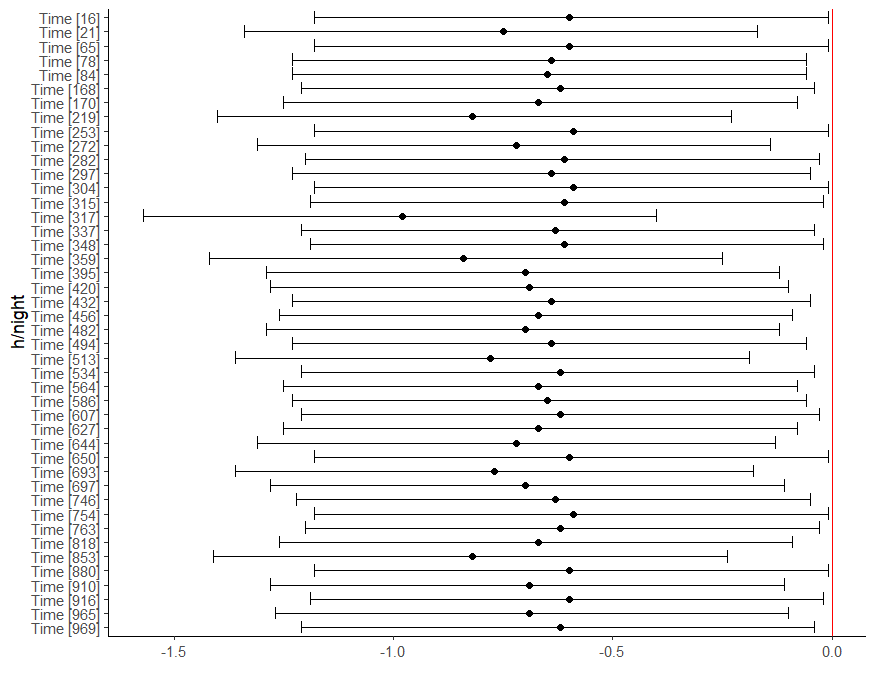
1. Mixed model

*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.

*Example (additional information) –*

The model used long format data. The marginal R² was 0.02. All significant variables were plotted below. According to the results, CPAP adherence were negatively associated with some time points and ESS baseline was not significantly associated with CPAP adherence.



(H) 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).

*Example (additional information) –*

The model was defined by long format data, a random intercept and slope and a mixture parameter on the time variable. A seed was implemented for the reproducibility. Four tests were performed to find the best number of clusters (1, 2, 3 or 4 clusters). According to the BIC criteria, the model with 2 clusters was the best model. Moreover, the distribution of patients in the clusters was fairly equally distributed including 64% in the 1st cluster and 36% in the 2nd cluster.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | **Loglik** | **BIC** | **%cluster 1** | **%cluster 2** | **%cluster 3** | **%cluster 4** |
| **1 cluster** | -440.17 | 962.49 | 100 |  |  |  |
| **2 clusters** | -428.44 | 966.42 | 64 | 36 |  |  |
| **3 clusters** | -417.10 | 971.13 | 28 | 66 | 6 |  |
| **4 clusters** | -410.26 | 984.82 | 24 | 6 | 32 | 38 |

The mean of posterior probability in each cluster was, for this final model:

|  |  |  |
| --- | --- | --- |
|  | **Prob1** | **Prob2** |
| **Cluster 1** | 0.99 | 0.01 |
| **Cluster 2** | 0.01 | 0.99 |

The probability to belong to a cluster was also available.

1. ARIMA model & CCF method

*Description* – The ARIMA model is a powerful tool for analyzing non-stationary time series containing seasonal and ordinary trends, to capture periodicity, trend and randomness in the data. It can also be used to add factors with lags to become an ARIMAX model (ARIMA with exogenous variables). To assess the correlation between two time series, cross-correlation can be used. It can also be used to study the correlation between 2 time-varying stimuli or events, which may or may not coincide over time intervals. A measurement vector is selected from each time series so that both vectors contain the same number of occasions, then the Pearson product-moment correlation is calculated for these two vectors. These vectors may or may not start at the same time. A lag is then introduced, i.e. a time difference separating the first measurements of the 2 time series. In general, observed data are considered to come from stationary processes, i.e. means and variances are constant over time. For each time window, a windowed cross-correlation can be run to examine the strength and lags of associations between the 2 time series as they evolve over time. Several methods are derived from this cross-correlation, such as DCCA, which analyzes the power laws of cross-correlations between non-stationary time series by removing local trends, thus ensuring that the results are not affected by these trends. Or the DPCCA method, an approach using partial correlation techniques, a hybrid method between DCCA and PCCA (Partial cross-correlation analysis).

The ARIMA model can also be used to predict the evolution of time series trajectories.

*Process* – The ARIMA model is first defined, then the cross-correlation function is applied:

1) ARIMA model analysis: fit the ARIMA model to the time series, i.e. analyze the stability and seasonality of the data (ACF, PACF), then estimate and test the model (AIC, BIC, Box Ljung statistic);

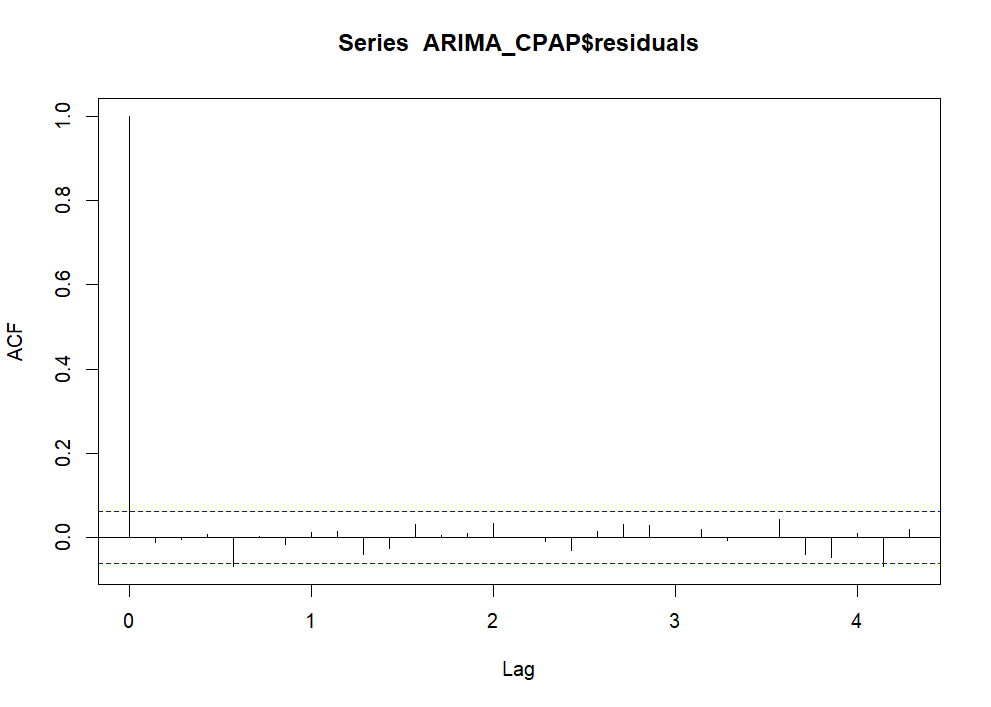
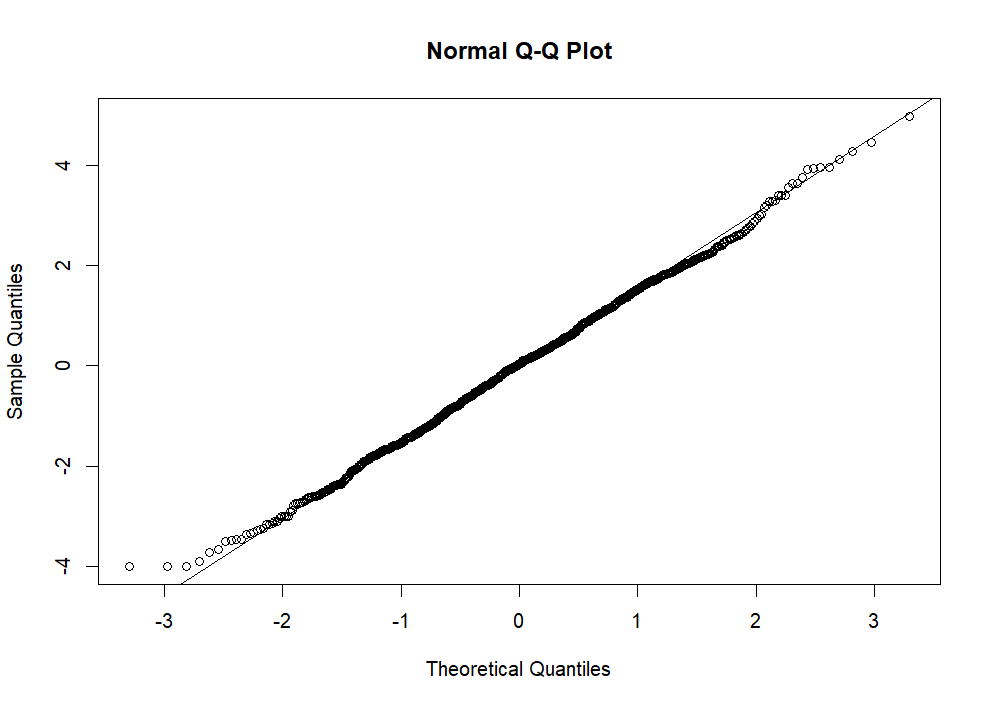
2) ARIMA model construction and validation: divide the data into 2 parts to perform the training/testing and validation method. The first part is used to define the model and the second to validate it;

3) Model application: validation of model prediction and execution by RMSE. The trajectory of the time series evolution can be analyzed and predicted;

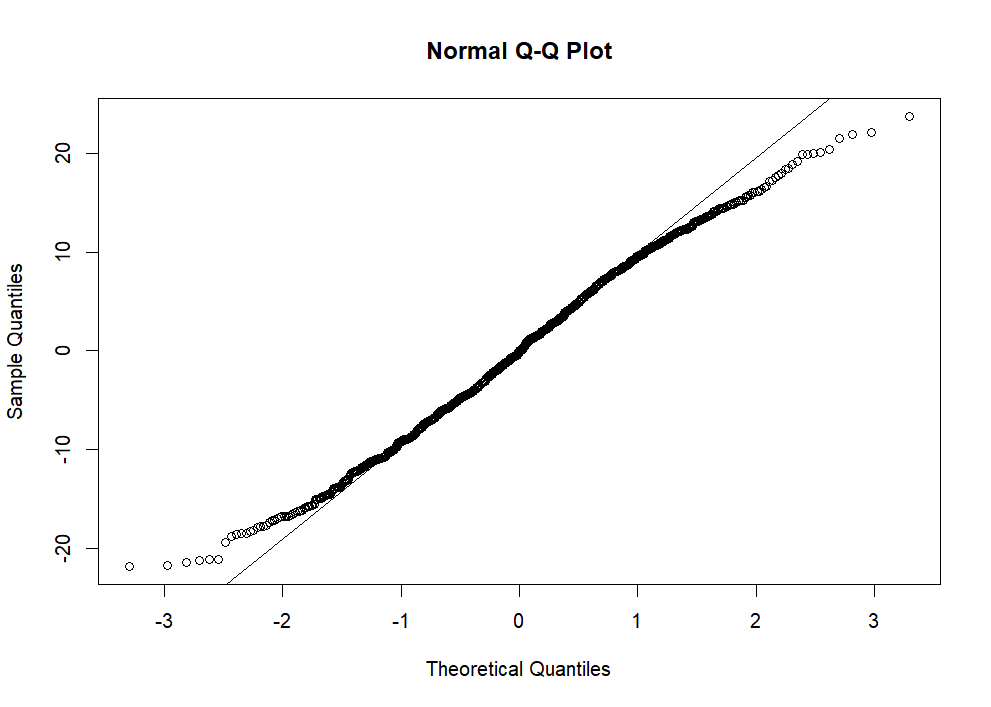
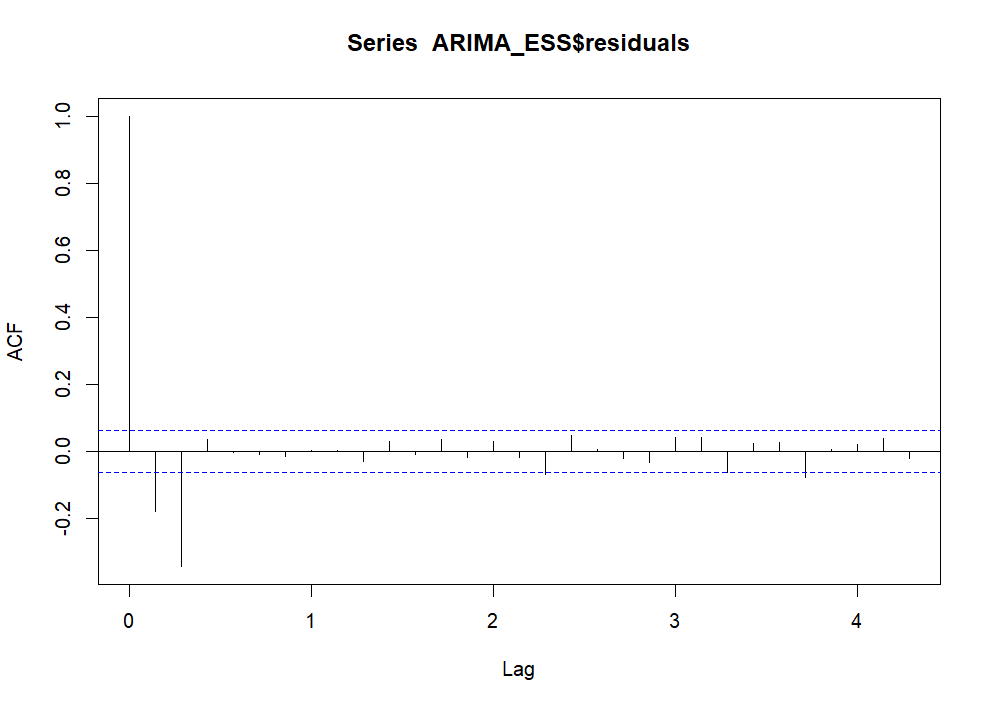
4) Evolution of the 2 time series: perform cross-correlation functions.

*Example (additional information)* –

For CPAP adherence validation, the p-value for Box-Ljung test were 0.70 showing that there was no pattern in the residuals.

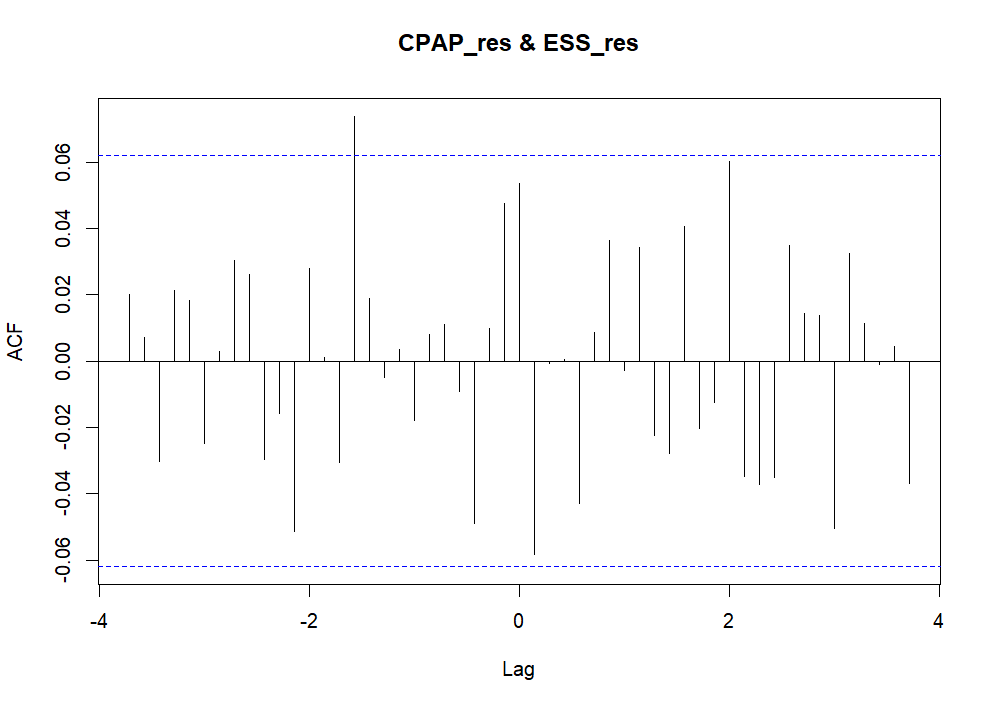


For ESS score validation, the p-value for Box-Ljung test were < 0.05 showing that there was pattern in the residuals. According to the ACF plot, there were correlations between the 1st time point and the other values and between the 2nd time point and the other values.



A cross-correlation function was performed to compare the correlation between these two time series (the variables detrend by the ARIMA model).

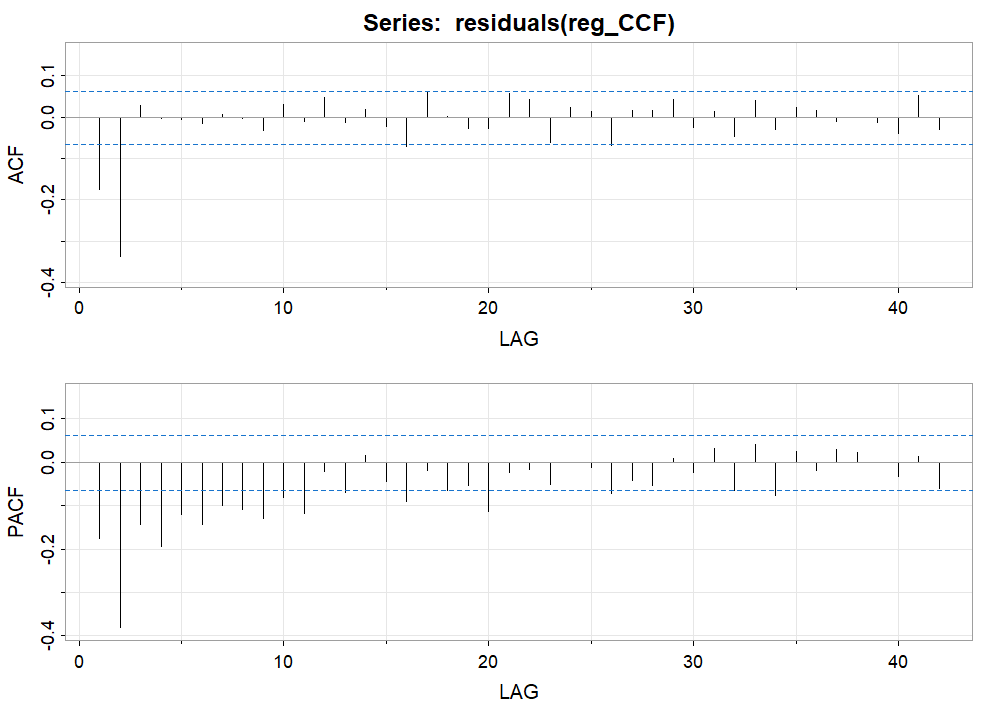
**~~Figure x: Autocorrelation (ACF) of the cross-correlation function (CCF) between continuous positive airway pressure (CPAP) adherence and the Epworth sleepiness scale (ESS) score.~~**



These lags could be implemented to a regression to study the association of the ESS score and the CPAP adherence at different lags.

These lags could be implemented to a regression to study the association of the ESS score and the CPAP adherence at different lags. For the example, 3-time lags were chosen: a lag of -11, a lag of 1 and a lag of 14. The equation was:

Validation of the regression by ACF and PACF plots: residuals were not correlated with lag.



Consequently, we found the same result, i.e. the increase of CPAP adherence increased, with 0.43 points for an increase of 1h of CPAP, the ESS score with a lag of 11 time points. (In SM, ACF and PACF for residuals(reg\_CCF)).

|  |  |  |  |
| --- | --- | --- | --- |
|  | **β** | **SE** | **P-value** |
| **CPAP lag -11** | 0.43 | 0.19 | *0.02* |
| **CPAP lag 1** | -0.33 | 0.19 | 0.08 |
| **CPAP lag 14** | 0.36 | 0.19 | 0.06 |

1. Joint model

*Description* – The joint model models the processes of repeated observations and measurements by applying a longitudinal joint model and a Survival model. The 2 processes are linked by the individual-specific sharing of random effects. The parameters of the models are estimated by an EM algorithm, treating random effects as missing data, and those of the Cox model are estimated by a 1-step Newton-Rasphson algorithm. In this way, standard errors can be approximated after convergence of the EM algorithm. For the Survival model, conditional on random effects, the model for each observation over time is a proportional hazards model. The model is defined using maximum likelihood. Moreover, it is not possible to define a single likelihood, as this introduces the distribution of random effects. However, methods such as Gaussian quadrature or Monte Carlo integration may provide a solution. To validate the joint model, the observation process must be at least random. This model can easily be generalized to multiple random effects, to different flexible parametric formulations of the basis risk for the recurrent event model, and to include other results. The bias induced by the informative observation process can be adjusted by an extension of the IPW method (IVW) using marginal regressions, giving each observation a weight based on the inverse of the probability of each observed measurement. The result is a pseudo population in which the observation process is static and can be ignored. The weights are estimated by a regression model including all the covariates informative about the observation process and the last values of each covariate that could impact the visit process. However, to avoid adding bias, covariates can be linked to the observation process and must be included in the weighting model. Two adjustments have to be made, as the last observation of each individual represents the end of the study follow-up, each weight is shifted by one time point and, given that each individual has at least one observation, a weight of 1 is assigned to the 1st observation of each individual.

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

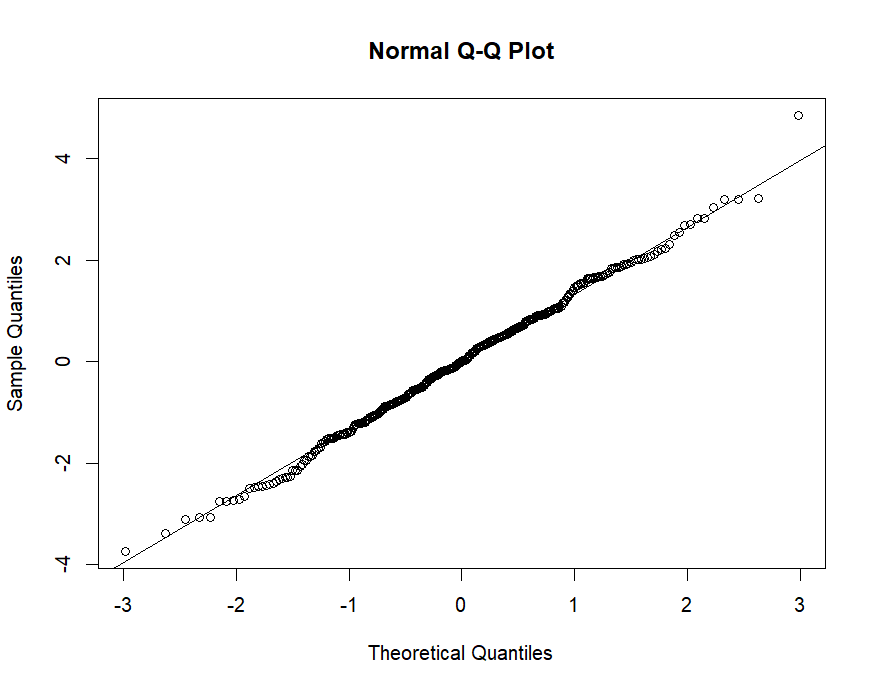
1) Define a mixed regression model;

2) Define a Cox survival model;

3) Model the evolution of both models at the same time.

*Example (additional information)* –

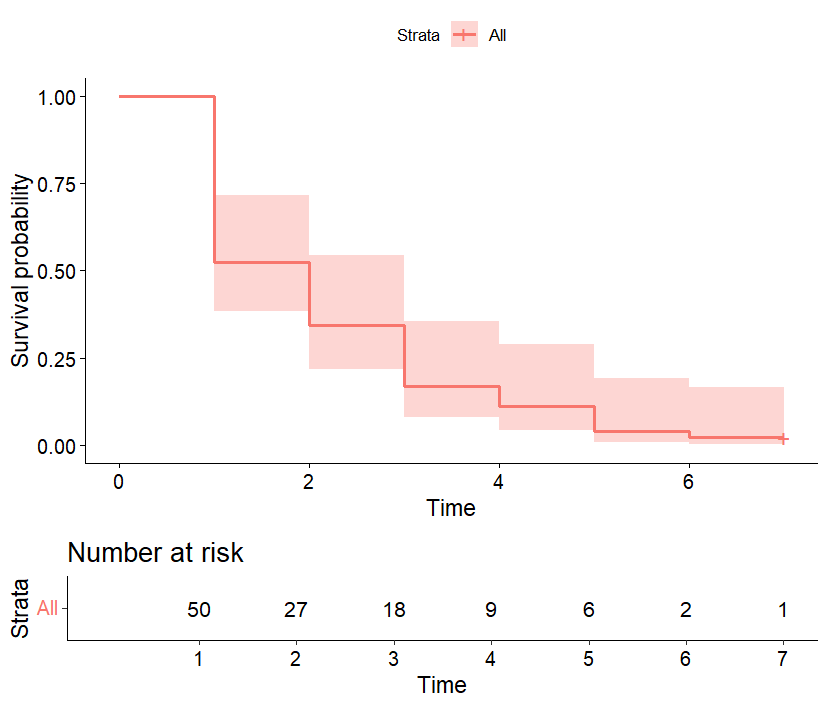
The mixed model was performed using continuous CPAP adherence and a random intercept and slope on patient. A maximum of 100 iterations was used for the lme optimization algorithm and for the optimization step inside the lme optimization and the ‘optim’ value for the optimizer parameter. The model was fitted by Maximum Likelihood. The normality of the model was validated using the QQ plot.

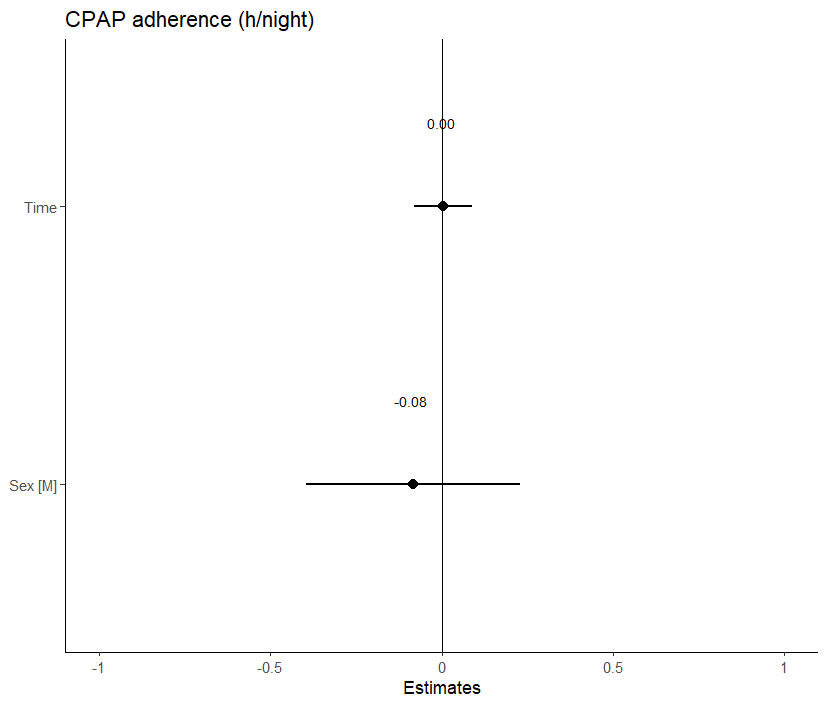


Fixed effects showed no significant association.

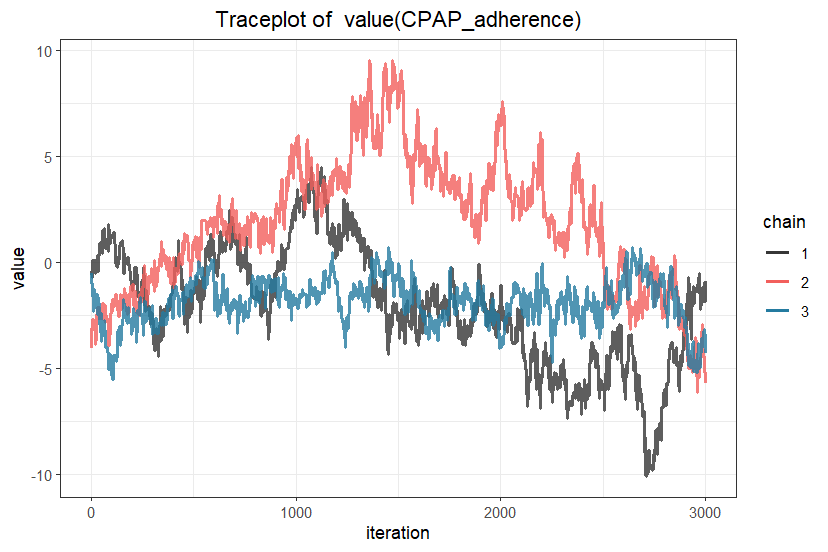
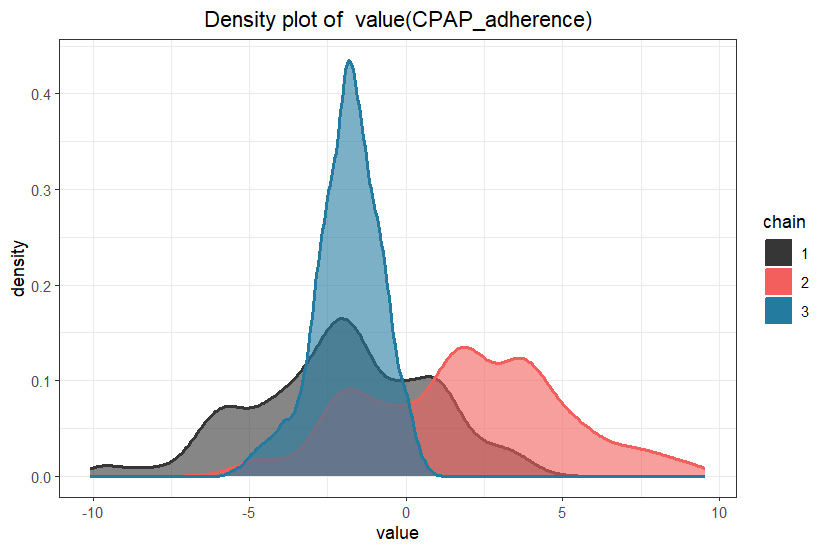
|  |  |  |
| --- | --- | --- |
|  | **β (CI 95%)** | **P-value** |
| **Time** | 0.00 (-0.08 ; 0.09) | 0.94 |
| **Sex - Men** | -0.08 (-0.39 ; 0.22) | 0.60 |

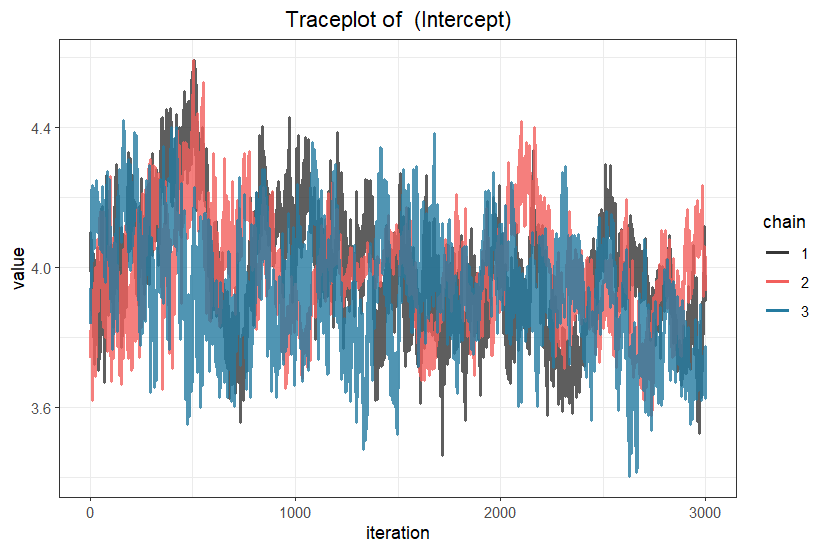
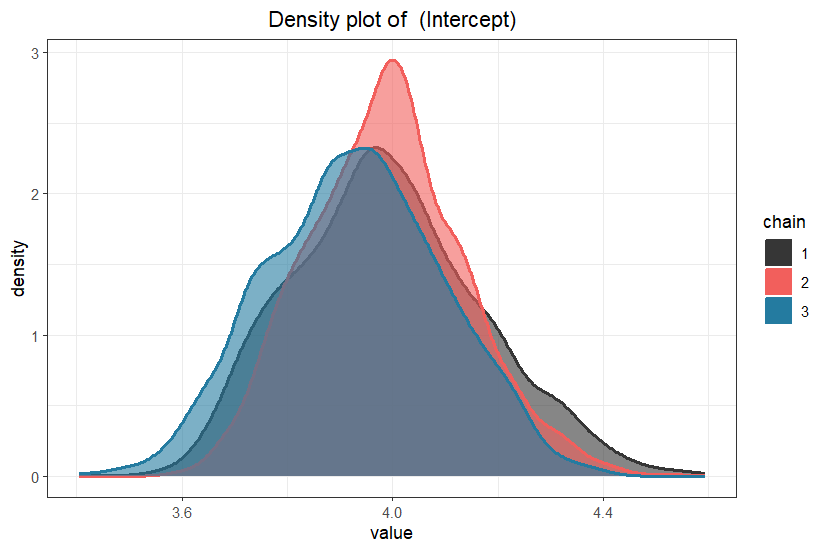
The Cox model was run using the ESS score and clustered by patient. The validation of the model was measured using the relative risk proportion test. The p-value of this test was 0.66 > 0.05, consequently the model was usable for the analyses. However, the likelihood ratio (p-value = 0.7) and score tests (p-value = 0.7) assumed independence of observations within a cluster, the Wald (p-value = 0.6) and robust score tests (p-value = 0.6) did not. According to the results, the sex did not significantly influence the survival curve (p-value = 0.64 > 0.05).

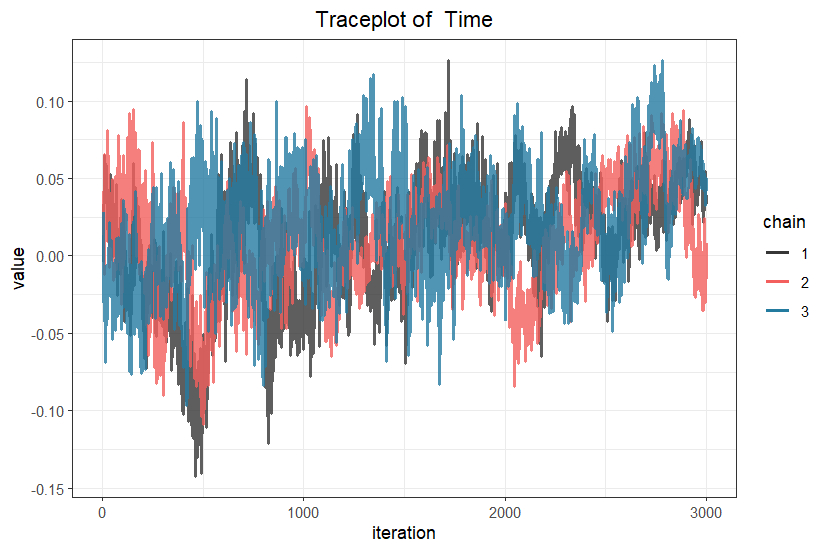
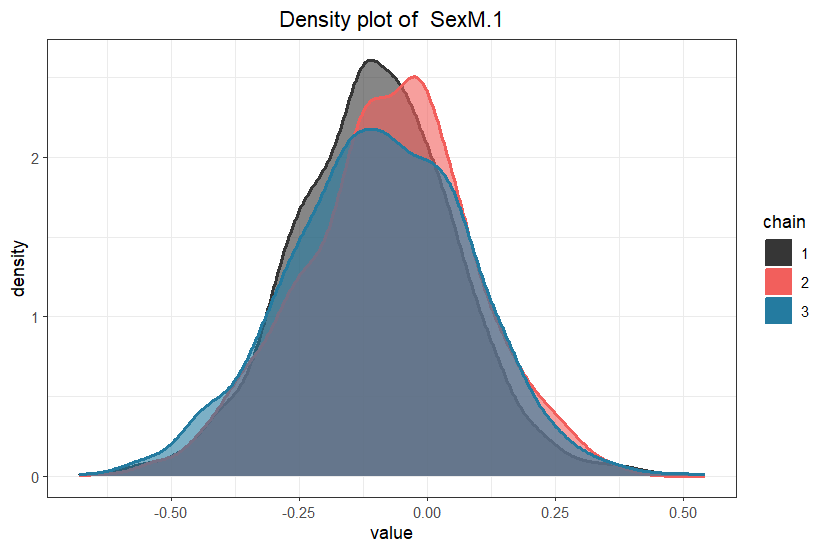


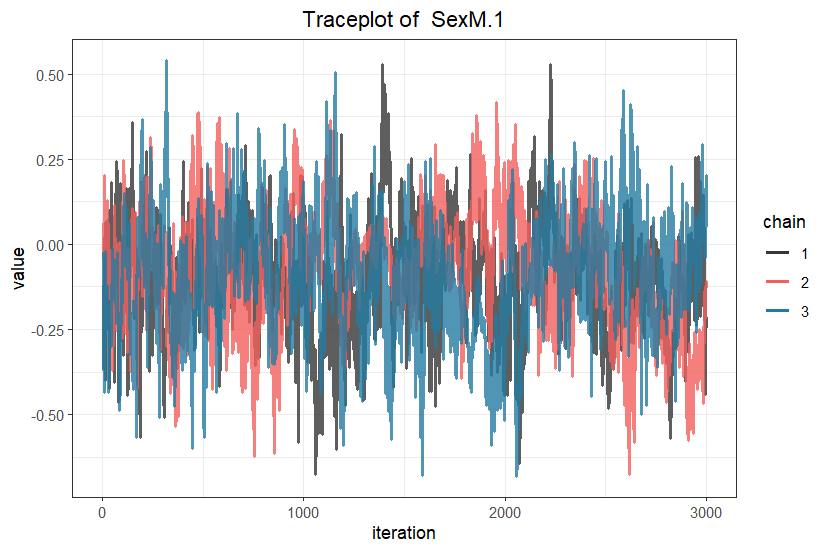
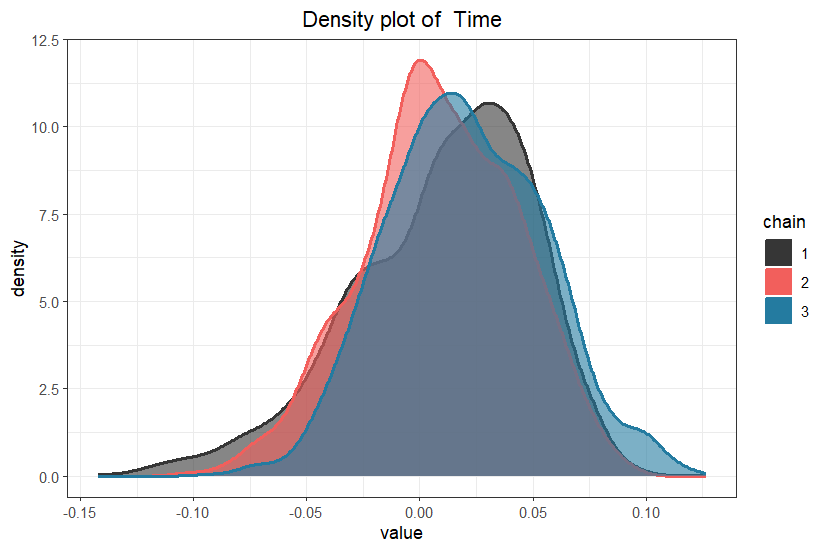


The joint model was validated by density, sampling behavior, mixing across chains and convergence graphs:









1. Hidden Markov model

*Description* – The Hidden Markov Model is a generative statistics model used to model time-varying stochastic processes. The model is defined by an initial probability vector, a transition matrix and emission densities of the hidden states. This method models the interaction between 2 sequences: 1) the observed sequence corresponding to the observed time series and 2) the unobserved sequence corresponding to the discrete Markov chain, the hidden state. In other words, the model is characterized by a longitudinal response following a known distribution with a density that is parameterized by a function of covariates and hidden state-dependent “regression” coefficients. For the first-order Markov chain, each value depends exclusively on the preceding values. From the set of observable sequences, the aim is to identify the best set of parameters. From the sequence of observations and the set of hidden Markov model parameters, the most probable state sequence is estimated using the Viterbi algorithm. This model also distinguishes between a number of latent classes that differ from each other in terms of model parameters, while within each class no individual differences are allowed. The choice of the number of latent classes can be defined a priori and then evaluated as the most reasonable, interpretable number of classes, or with criteria such as AIC and BIC. The transition matrix is made up of the proportions of transitions from one state to another. Inter-state transitions are described by the Hidden Markov Process, and the Markov Chain is able to take into account different types of longitudinal data that are collected regularly, irregularly or continuously. In other words, an individual can move from one class to another over time. These dynamics correspond to trajectories. For this purpose, a sequence of discrete latent variables rather than a single latent variable is associated with each individual, giving rise to a hidden process assumed to follow a Markov chain. Finally, once this matrix is known, it is applied to a K-Means method. After estimating the model by maximum likelihood (EM algorithm), the standard error of the estimated parameters can be obtained on the basis of the observed information matrix. The final model must be a good compromise between quality and complexity. Latent variables must take into account unobserved heterogeneity between individuals that cannot be explained by observed covariates. A training/test and validation method can be used to determine the final model. Once this model has been defined, clustering is performed using the maximum a posteriori (MAP) rule.

More specifically, the Latent Markov (LM) model is used for univariate and multivariate longitudinal data analysis based on repeated observations of a sample over time. This model assumes that conditional response probabilities are homogeneous over time.

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

1) Markov assumptions: the probability of transition from one state to another must not depend on previously visited states or time spent in the current state, and it assumes that the transition rate must be constant over time and common to all individuals;

2) Estimation of model parameters: estimation of the transition rate from a transition probability that can be directly estimated from the data. A Bayesian method using the MCMC algorithm could be a solution;

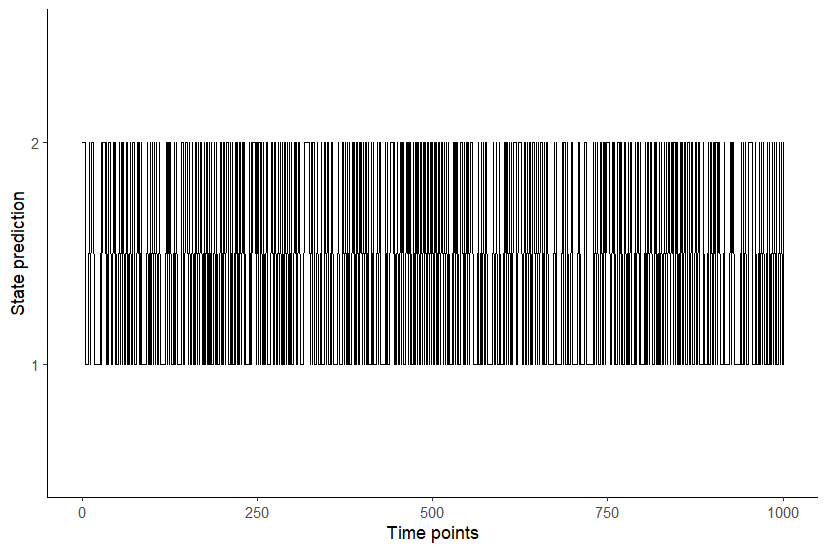
3) Inclusion of random effects and individual covariates: assume that transition rates are not fixed but modifiable;

4) Modeling unobserved responses: several scenarios are possible, depending on missing data (MAR, MCAR, MNAR, LOCF-like missingness) and loss to follow-up.

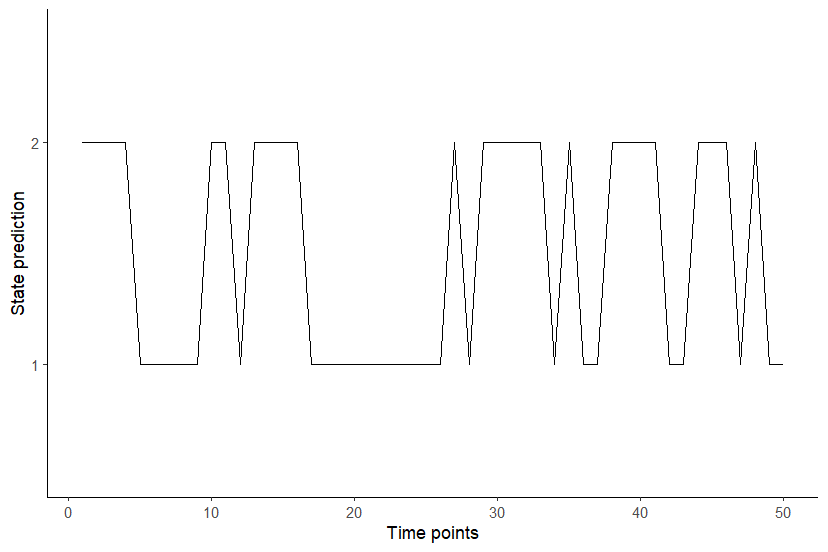
*Example (additional information)* –

This model fitted EM algorithm and used the multinomial family for the categorial observations. The independence of observation variable from all covariates were added to the model. A seed was used for the reproducibility. A test for other number of hidden states can be applied and comparisons using BIC, AIC and loglikelihood criteria could be performed. However, the interpretation of hidden states must stay possible.

The prediction of the patient’s state at each time point was predicted and plotted in the graph below.



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



The states prediction included 59.1% of time points in the 1st state and 40.9% in the 2nd state.

## e-Appendix 3. Table comparaison R and Python

Tableau avec packages R et Python ou existe en R et/ou Python