Les données longitudinales, de trajectoire et de télémonitoring : qu’en fait-on ? Exemple des données du sommeil

Introduction

Nous cumulons de nombreuses données tout au long de notre vie, en particulier dans le domaine de la santé, concernant nos caractéristiques physiques, socioéconomiques, nos différentes maladies, traitements, hospitalisations, etc. Ce sont la base de notre trajectoire de données individuelle relatant tous nos changements et développement au cours de notre vie.

Lors des études cross-sectional nous ne prenons qu’un point de temps spécifique mais nous ne pouvons pas analyser les changements et trajectoires des individus au cours du temps. C’était pourtant le design d’étude le plus couramment utilisé. Une solution est donc d’évaluer les mesures de façon répétée dans le temps : ce sont les données longitudinales. Généralement, les données longitudinales sont centrées sur une période de temps précise avec des mesures prises à intervalles réguliers qui permet d’étudier par exemple, des facteurs de risques ou des effets de traitements1. Dans notre contexte de données du sommeil, l’utilisation de ces données longitudinales permet notamment d’étudier les données relatives à la PPC (Pression Positive Continue) comme la date de début du traitement du patients, ses arrêts et ses reprises du traitement au cours du temps, ses données manquantes ou extrêmes, ses changements de prestataires, ses sorties d’études, ses autres traitements, ses comorbidités, son décès).

Plus récemment, et particulièrement lors du COVID, l’arrivée du télémonitoring a permis une expansion du modèle de données longitudinales. Le télémonitoring est un suivi régulier des patients permettant de s’assurer de l’efficacité, de la tolérance et de l’observance des traitements2 (+ ref R.Tamisier 2020 ?). Ce modèle d’étude sera de plus en plus utilisé et particulièrement dans le domaine du sommeil avec l’analyse des données de PPC par exemple puisqu’il permet d’étudier l’observance et l’efficacité du traitement (IAH résiduel, fuites, …).

Ce design d’étude représente toutefois un défi pour les statisticiens et méthodologistes : la présence de données manquantes, la corrélation de certaines mesures répétées au sein d’un même cluster, la temporalité des observations : les données sont généralement ordonnées dans le temps mais les intervalles d’évaluation des mesures peuvent être inégaux, sont des contraintes que les modèles statistiques doivent prendre en compte1. De nouvelles méthodes sont développées pour analyser ces données tout en répondant au maximum à ces critères, mais comment choisir le bon modèle d’analyse statistique ?

Dans une première partie, nous ferons un tour non exhaustif des principaux modèles d’analyse dans le cadre des données du sommeil, puis nous définirons dans quel contexte utilisé ces modèles et enfin nous donnerons deux exemples d’analyse de données répétées dans le domaine du sommeil.

# I/ Les différents modèles statistiques

Pour chaque sous-partie, donner la description du modèle, ses avantages et ses limites

Latent class modelling (2, 3, 4): Latent variables = model a quantity that is not observed; include random variables that cannot be directly observed; individuals are assigned to latent trajectory subgroups on the basis of their observed symptoms or behaviors; highly flexible allowing for a variety of complexities including partially missing data, discretely scaled repeated measures or time-varying covariates

1. **GMM (Growth Mixture Modelling)**

Finite mixture model; parametric model for longitudinal data; estimates an average growth curve for each class and allows for variations between individuals of the same class; introducing random effects in the model, for each trajectory GMM estimates an intercept, slope and a growth parameter variance by maximizing the log-likelihood function; the probability of belonging to each class is estimated based on observed data; individuals are then assigned to subgroups based on their higher posterior group proba; contribution of covariates (vary or not over time) ok and can influence the model coeff; once trajectory membership has been established it can be used as a dependent or independent variable to explore predictors of health traj and their contribution to future health outcomes; type data = longitudinal data (continuous or categorical data); longitudinal or cross-sectional studies; repeated measures, time-varying covariates, partially missing data; process = iterative procedures, requires a priori decisions as well as stat inference: 1) hypothesized the expected nb of latent classes and shape of the curve for each class (linear, quadratic, cubic, …), 2) model specification: make decisions about growth parameters (intercept, slope variance and covariance), 3) model estimation: estimated by maximum likelihood or by Bayesian methods, 4) model selection: which of the model tested provides the best or most reasonable representation of the observed data, using Lo-Mendell-Rubin adjusted likelihood ratio test for nested models and/or bootstrapped likelihood ratio test and/or BIC (smaller BIC) + take into account convergence, ability of the model to provide well separated classes (entropy near 1.0), proportion of the sample in each traj (more than 5%), average posterior proba (near 1.0), parsimony and usefulness of the observed latent classes in practice; Advantages = handling missing data, allowing for correlated residuals, treating residuals in regressions and random effects in mixed effects models as latent variables, estimates a mean growth curve for each class and captures individual variation around these growth curves by the estimation of growth factor variances for each class; limits = GMM estimates many more parameters interpretation of the results can be complex; many parameters to decide a priori

lcmm (latent class mixture model) also called GMM for Gaussian longitudinal outcomes; population of heterogeneous and composed of a fixed nb of latent clusters of subjects characterized by their trajectories; trajectories modelled by a cluster-specific linear mixed model; estimation based on maximum likelihood and goodness-of-fit measures; highly flexible since the fixed effect and the distribution of the random effects can be specified for each cluster; predictors can be added to the model to find latent variable; result depend on the starting values.

GMM is used to examine unique longitudinal trajectories with repeated follow-up measures, Advantages = capture inter-and-intra individual differences over time, probabilistically identify homogenous subgroups within larger heterogeneous memberships and represent unobserved heterogeneity by inferring each individual's membership to latent classes from the growth model data; used to identify patterns of medication non-adherence; can include covariates to explore the relationship between baseline characteristics and adherence patterns; distinct, homogeneous and longitudinal trajectories with follow-up; comparison of different models according to parameters by Lo-Mendell Rubin adjusted likelihood ratio test (LRT, p < 0.05), Parametric bootstrapped likelihood ratio test (BLRT, p<0.05), BIC, AIC, convergence (entropy closet to 1.0).

Method for identifying multiple unobserved sub-population, describing longitudinal change within each unobserved sub-population and examining differences in change among unobserved sub-population. Growth Curve Modeling = describe and test hypotheses about interindividual differences in intraindividual change; great flexibility in characterizing nonlinear patterns or shapes of change over time; alternative representation of the change trajectories often modeled via polynomial models and is particularly useful for representing complex shaped trajectories in a parsimonious manner. Multiple-group Growth Curve Modeling: GCM can be extended to accommodate and examine differences in change among multiple groups of individuals; used to divide the longitudinal data into 2 subsets of data (sex for example). GMM = extension of the multiple-group growth model in which the grouping variable is latent or unobserved; be used to describe differences in how longitudinal change proceeds in defined subsets of data, the objective of the growth mixture model is also to describe differences in how longitudinal change proceeds in subsamples within the data; objective = describe in a post-hoc manner , possible sub-groups within the data and to describe group differences in longitudinal change between and within those unobserved groups; process = 1) problem definition: initial GMM hypotheses (nb of groups expected and how groups expected to differ in term of means, Covs and pattern for exemple) then other initiate parameters will be tested to choose the best model, 2) model specification: series of useful models can be specified and subsequently estimated, best model according to BIC and convergence, 3) model estimation: using maximum likelihood or Bayesian methods, current implementations an EM procedure or Markov Chain Monte Carlo procedures wherein the unobserved group classification variable is treated as a latent variable, use iterative procedures to obtain parameter estimates and posterior estimates of the proba of individual's membership in each of the possible group, 4) model selection and interpretation: sense for researchers, BIC, AIC, adjusted BIC, entropy = the accuracy or confidence with which individuals have been classified as belonging to 1 group or another (near 1.0 = better), likelihood ratio tests that quantify specific comparisons between the model of interest and a model with 1 fewer class (eg Vuong-Lo-Mendell-Rubin likelihood ratio test and Adjusted Lo-Mendell-Rubin likelihood ratio test and in brief apply a corrected likelihood ratio distribution to compare different models; missing data on the outcome variable are accommodated using full information maximum likelihood or Bayesian approach under MAR assumptions; limitations = GMM is a constrained exploratory technique that seeks out the story the data are trying to tell, real dangers of finding and identifying false groups => replicating the results with new data

GMM and LTA = well suited to model stability and change over time because of their ability to examine inter-individual differences in intra-individual processes; ideally placed to calls for more longitudinal, within-person research. GMM = aims at identifying subpopulations that follow different longitudinal growth trajectories over time, thereby being a mixture extension of latent growth or latent curve models; 1 or more variables is measured repeatedly and growth in the level of these variables across time is estimated via random intercept and slope(s) factors, random intercept factors capture each individual's initial level on the repeated measures, random slope factor capture each individual's change in those repeated measures as a function of time; being a mixture extension of the latent growth model, aims to identify subpopulations following different growth trajectories over time, similar to multi-group growth curve modelling where different growth models are tested for each group but in GMM the grouping variable is latent or unobserved, the latent subpopulations are allowed to differ regarding their average level on the growth factors (intercept and slope)or intercept and slope variances and covariances and even time-specific residuals and the subpopulations can also be allowed to follow a different functional form; Advantages = particularly interesting because it allows the change over time to be qualitatively different for different group of individuals. Limits = selecting the optimal number of latent profiles: models with an increasing number of latent profiles are tested after which the most optimal 1 is selected based on interpretability and theoretical conformity of the solution, statistical adequacy and statistical indicators (as BIC, sample-adjusted BIC, Consistent AIC, Bootstrap Likelihood Test, elbow plot); labeling of profiles: use of differences in the unique pattern (shape differences, level differences, scatter differences, ...); incorporated covariates: should only be included once the optimal unconditional profile solution is selected, include them in the final solution directly but has to not change the profile solution; multi-group invariance testing: whether profiles found in 1 sample generalize across known subpopulations, for LCA it has been tested with a 3 steps approach and it exists a 6 steps for LPA

1. **LTA (Latent Transition Analysis)**

Analyze changes in multiple categorical variables over time, contingency tables over time; semi-parametric finite mixture model for longitudinal data, observed data from a set of categorical variables to define latent variable for each time point; each individuals can change their class membership over time; objective = study the proba of transition of an ind from 1 class at 1 time point to another class at the next time point; change = matrix of transition proba between 2 consecutive time points; model estimates = latent status membership proba at Time 1, proportion of the pop in each latent class at each time point, the conditional proba of making a transition from 1 latent status to another over time, item-response proba conditional on latent stauts membership; at any time point, a posterior group proba can be predicted; parameters estimated by maximizing the likelihood function or by the Bayesian method; covariates can be added; requires the nb of classes before adding covariates principally to avoid a potential change in class number with and without covariates; set of categorical variables measured over time, when variables have too many categories recommended recoded into as few categories as possible, preferable to use it when the nb of time points is no larger than 6; process = iterative, requires a priori decisions based on field knowledge and requires several steps for its implementation, 1) choice of nb of latent class based on result of hypothesis test, 2) model specification: make a decision about the time invariance of item-response proba, the measurement invariance for transition proba and the addition of covariates, 3) model estimation: chosen before fitting the models, by maximum likelihood using the expectation maximization algo or Bayesian methods using Markov chain Monte Carlo algo, 4) AIC and BIC used to select the best model (smaller value = best model); Advantages = model a change over time and investigate predictors of this change, comparing different subgroups to test for treatment effects and for evaluating the contribution of different measures for each latent status; Limitations = requires large sample size because of numerous parameters to estimate (e.g. transition proba matrix) but each possible transition may be considered as a separate contingency table, this table often contains a large number of possible response patterns may be empty but larger the sample size lesser the likelihood of sparsity within the contingency table cells, nb of time points becomes high (>6) LTA more complex because of numerous parameters to estimate; LTA bears some similarities to Hidden Markov models

LTA = latent Markov model where stage membership at each time is unobserved but measured with a set of manifest items explained by the underlying categorical latent variable; maximum likelihood estimates for LTA are easily estimated by using the EM algo or Bayesian inference via Markov chain Monte Carlo but behave erratically when the sample size is small and many item-response proba are not close to 0 or 1; latent transition models based on latent class theory which posits that homogeneous subgroups of individuals can be identified based on their responses to manifest items; manifest items = measured repeatedly over time to identify latent classes at each occasion and the proba of transitions over time in latent class membership are estimated; to avoid boundary solution, Bayesian analysis via MCMC has been applied (vs. ML using an EM algo), Limits = the likelihood function of an LTA model may have unusual characteristics which can adversely affect inference so instead of standard testing methods for choosing an appropriate number of classes, are to choose penalized likelihood measures such as AIC, BIC, Lo-Mendell-Rubin, bootstrapping the LRT, Bayesian posterior predictive check distri; small sample size = problem with MCMC which generally used 'burn-in' period to eliminate dependence on the starting values or with label switching --> to handle: pre-assign 1 or more subject's class membership to break the symetry of the posterior distribution and dampen the posterior density over the nuisance modes.

LTA = longitudinal extension of LCA/LPA; can transition from 1 latent class to another over time; the latent classes refer to subgroup memberships at that particular point in time; for categorical indicators there are 3 sets of parameters estimated: 1) at each time point the proportion of individuals that is expected to belong to each latent status , referred to as the latent classe membership proba, 2) the transition proba capture the proba of transitioning from a specific latent status at time t to another latent status at time t+1, 3) item-response proba tap into the connection between latent status membership and the observed categorical indicators at each time point, provide information on the differentiation of the latent statuses. GMM and LTA = well suited to model stability and change over tiem because of their ability to examine inter-individual differences in intra-individual processes; ideally placed to calls for more longitudinal, within-person research. Limits = selecting the optimal number of latent profiles: models with an increasing number of latent profiles are tested after which the most optimal 1 is selected based on interpretability and theoretical conformity of the solution, statistical adequacy and statistical indicators (as BIC, sample-adjusted BIC, Consistent AIC, Bootstrap Likelihood Test, elbow plot); labeling of profiles: use of differences in the unique pattern (shape differences, level differences, scatter differences, ...); incorporated covariates: should only be included once the optimal unconditional profile solution is selected, include them in the final solution directly but has to not change the profile solution; multi-group invariance testing: whether profiles found in 1 sample generalize across known subpopulations, for LCA it has been tested with a 3 steps approach and it exists a 6 steps for LPA

1. **GBTM (Group-Based Trajectory Modelling)**

Finite mixture model, semi-parametric model for longitudinal data; discrete distribution of the population; assumes there is no variation between individuals in the same class (no within-class variance on the growth factors), distinguish in the pop subgroup/classes of homogeneous individuals, proportion of the population belonging to each of these subgroups is then estimated, determines for each ind the proba of belonging to 1 subgroup or another (posterior group proba); ind assigns to a subgroup based on their highest posterior group proba; parameters estimated by maximizing likelihood; covariates (vary or not over time) can be included in the model; longitudinal data (continuous and categorical data); Process = iterative, requires prior decisions, 1) hypothesized the expected nb of latent classes and shape of the curve for each class (linear, quadratic, cubic, …), 2) model specification: first test a 1-group model then gradually adjust the maximum logical nb of subgroup (greater than the expected nb of subgroups) = comparison between, quadratic vs. linear for 1 traj then if quadratic ok comparison for 2 traj wit quadratic component, BIC value, 3) model estimation: estimated by maximum likelihood or by Bayesian methods, 4) model selection: take into account field knowledge, preference for a useful and parsimonious model, close correspondence between the estimated proba of each subgroup and the prop of ind classified in such subgroup according to the rule of the max proba of belonging, average posterior proba of subgroup membership greater than or = to 0.7, sufficient nb of ind in each subgroup (more than 5%), reasonably narrow confidence intervals, difference of BICs between 2 models with different nb of traj subgroups; Advantages = simpler specification of the GMM, handling missing data, allowing for correlated residuals, fewer parameters to estimate than GMM, run faster with fewer errors than GMM, easier to interpret because less complex; Limitations = supposes that all individuals in a traj class have the same behavior (not within-class variation)

GBTM = for each dependent variable there exist 2 or more groups having modest within-group variation but extreme between-group variation, such groups cannot be accurately established directly from the independent variables, a latent categorical variable is introduced to define group membership, at each level of the latent variable a typical trajectory for the dependent variable is estimated; proba of membership in a group is expressed in terms of the independent variables; group membership can also be used to make predictions; Limits = missing data occur completely at random, creates groups that are so qualitatively different on the dependent variable that there is no question of whether clinically relevant differences exist between groups, there do not exist direct relationships between independent and dependent variables so that quantities like odds ratios and rate ratios are not routinely estimated; Advantages = the clinically relevant differences between groups as well as the typically modest numbers of groups chosen ensure that the results of GBTM will be highly amenable to visual interpretation

GBTM = availability of routines in standard statistical programs that provide easy-to-interpret visual summaries of the data; less computationally demanding, simpler to fit and easier to use in samples with smaller nb of obs than latent growth mixture modelling approach (most popular alternatives); Limits = strong assumptions about the distribution of traj --> create spurious finding if violations of the assumptions; BIC to select the number of subgroups and the order of the polynomial terms used to model the shapes of trajectories; average posterior probability (APP) and relative entropy to assess the adequacy of the classification of individual traj in the subgroups identifies; finite mixture model; observed repeated measures; model is estimated using a latent class formulation in which each subgroup has a specific sets of regression coefficients that corresponds to the variables indexing time; the model can be adapted for variables that follow Bernouilli, Poisson or Normal distribution; Individual posterior proba of subgroup membership can be derived from estimates of the mixing parameters to estimate regression coeff and posterior proba can be used to assign individuals to the subgroup for which their proba is the highest according to the maximum-proba assignment rule; optimal nb of traj subgroups is selected by estimating models with an increasing nb of traj using the BIC approximation to the Bays factor as long as the difference between 2 adjacent models was greater than 10; once the nb of traj is selected, the model is refitted using lower-order terms if the higher-order terms are not statistically significant at 5%; cubic polynomials generally used; the APP is calculated for each traj subgroup by averaging over the individual posterior proba of traj membership of individuals assigned to the subgroup, APP greater than 70% across traj subgroups = indicative of adequate classification; relative entropy measures the degree of classification accuracy of placing participants into a traj based on their posterior proba, greater than 0.80 = less classification uncertainty; mismatch = the difference between the estimated proba of subgroup membership and the proportion of individuals classified in that subgroup based on the highest posterior proba, the correspondence between these 2 proba decreases as assignment error increases, close to 0 = adequate fit; GBTM assumes that the residual variance is constant over time and similar across subgroups; LCGM outperforms GBTM in the presence of time point-specific overlap in the distri of the repeated measures (longitudinal data) and when there were no traj subgroups in the data; Limits = GBTM may create spurious traj in certain scenarios and APP for classification adequacy failed to detect spurious traj in most scenarios, GBTM overestimate the nb of traj when individual traj have the same shape and are distri on a continuum around the mean traj and overestimate the nb of subgroups in several scenarios when included subgroups.

1. **LCA (Latent Class Analysis)**

Underlying unobserved categorical variables that divide a pop into mutually exclusive and collectively exhaustive latent classes; semi-parametric model for categorical cross-sectional data (non-longitudinal version of LTA); parameters estimated by maximizing likelihood or by Bayesian method; contribution of covariates can be modelled in each class, proba of belonging to a class depends on the values or levels of the covariates; categorical variables measured in a cross-sectional way, when variables have too many categories it’s better to recode them into as few categories as possible; process = iterative, requires a priori decisions based on field knowledge and requires several steps for its implementation, 1) choice of nb of latent class based on result of hypothesis test, 2) model specification: make a decision about the addition of covariates, 3) model estimation: chosen before fitting the models, by maximum likelihood using the expectation maximization algo or Bayesian methods using Markov chain Monte Carlo algo, 4) AIC and BIC used to select the best model (smaller value = best model); Advantages = powerful tool for analyzing the structure of relationships among categorical variables, explore and interpret complex contingency tables, provides method for testing hypotheses regarding latent structure among cat variables, more suitable as an exploratory approach; Limitations = only applicable to cross-sectional nominal or ordinal data, analyzes cross-sectional data = cannot really be considered a “trajectory” modelling technique (LTA = LCA at each time point to determine classes)

LCA = to identify clusters of social risk factors using quartiles of the geographically informed social risk factors; due to large computationa requirements of performing LCA on large datasets, assessed the optimal number of clusters by sub-setting the data into 10 random subsamples of N = 5,000 participants and performing LCA using 1 through 5 clusters; BIC and the elbow method to determine the optimal nb of clusters; then re-run LCA in the complet dataset using only the solution, setting the maximum number of iterations through each estimation algo as 1,000 and the nb of times to estimate the model with different class-consitional response proba as 25; Cross-sectional associations were assessed using chi² test and unadjusted and adjusted linear/logistic regression with covariates; data were split into training/testing (90%) and validation (10%); for each cluster optimized 4 different ML methods (logistic regression, random forest, light gradient boost machine, extreme gradient boosting) and evaluated models with area under the ROC curve and area under the precision-recall curve using a 3-fold cross-validation design

LCA = clustering performed using LCA which groups participants into mutually exclusive clusters; BIC and AIC to choose the number of clusters; 4 endotypic traits were dichotomize by median into high- and low-value groups and the model be trained with also continuous variables using latent profile analysis with a choice of the nb of clusters by BIC and bootstrap likelihood ratio tests; examine the differences in demographic characteristics, etc between the enditype clusters using chi² tests for categorical variables and Kruskal-Wallis tests for continuous variabels; multivariate logistic regression models to examine the association between the 3 symptom and endotype clusters, post hoc pairwise comparison with Bonferroni adjustment performed to assess whether the symptom cluster was predictive of belonging to a specific endotype cluster; symptom proportions and ESS scores were standardized; adjustments with adjustment factors

LCA = finite mixture modelling most commonly used; to determine if unmeasured or unobserved groups exist within a population; unobserved or latent groups are inferred from patterns of the observed variables or indicators used in the modelling; observed indicators are all categorical vs. latent profile analysis where all indicators are numerical and continuous; but application to longitudinal indicators and continuous latent variables are also possible, mixed data; using a set of observed indicators, LCA models identify solutions that best describe these latent classes within which indicators follow the same distribution; using maximum likelihood estimates; LCA = probalistic method of unsupervised clustering; classes are homogeneous whithin but distinct from each other; once the model has been fitted the proba of class membership is estimated for each observation in the cohort then used to assign class; not assign individuals to latent class; Advantages = less misclassification rate than clusters methods as k-means, no choice of nb of clusters, ability to generate fit statistics which in turn allows stat inference, more statistically robust, generation of posterior proba allows quantitative assessment of uncertainty of class membership, allow the usage of mixed data types for the class-defining variables including different scaling; Limits = computationally demanding so limitation of the number of indicators according to processing power available, sensitive to extreme values; process = 1) Observed indicator selection: depend on the study question; 2) examine the data: check for extreme or implausible values, for continuous variables check the normality and if not tranform variables, for categorical with low frequencies (less than 10% = excluded from the analysis) are difficult to fit into a model if yes collapse categories together but this is subjective according to the data, standardization could be used too, within latent classes observed variables are independent of each other otherwise introduce bias and misclassification errors so examine the correlation matrix of the variables (>0.5 need to be examined), LCA = 'a large sample' method with sample sizes of greater than 500 models and fit statistics have been shown to consistently perform with high accuracy, smaller sample size = less reliable but for anlayses with less than 300 or between 300 and 500 recommend using Monte Carlo simulations to determine adequacy of power and power calculations to determine sufficient magnitude to be meaningfull, available with missing data: imputations for exemple deletion, multiple imputation and full information maximum likelihood, 3) fitting models to the data: the correct number of models to fit to the data be dictated by the sample size, number and quality of indicators used in the model and what an acceptable size may be for the smallest class, for each model parameters are estimated based on maximum likelihoods and numerous fit statistics + posterior proba for belonging to all the latent classes in the model for individual observations, 4) evaluating the models selecting optimal number of classes: with the fewest number of classes that best fits the data or BIC, sampli-size adjusted BIC, Lo-Mendell-Rubin and Vuong-Lo-Mendel-Rubin, the size of the smallest class, 5) evaluate the models: exemple calculate the entropy = a measure of class separation (p near 1), 6) interpreting the final model: validity of the model to consider its robustness, multiple random starts should be used to demonstrate sufficient replication of the maximum likelihood, inspect the classes, reproducibility in external datasets

LCA = statistical procedure used to odentify qualitatively differnet subgroups within populations that share certain outward characteristics; subgroups are referred to as latent groups; to detect the latent groups = participants' responses to categorical indicator varriables but when continuous = latent profile analysis; LCA = detect latent heterogeneity in samples; assumption: membership in unobserved classes can cause or explain patterns of scores across survey questions, assessment indicators or scales; proba of class membership are obtained not clear-cut assignments; can generate categorical classification variables for use in other analyses, limits = assigns indiv to classes based on their proba of being in classes given the pattern of scores they have on indicator variables (proper class assignment is not guaranteed), the exact nb of % of sample members within each class cannot be determined, have to make a number of statistical and theoretical decisions; process = 1) selection of participants: appropriate and theoretically justified, sample size: smaller samples may be adequate with simpler models and "well-separated" classes but problems could also be poor functioning fit indices, convergence failures and failure to uncover classes with low memberships, to determine sample size = Monte Carlo simulations, 2) selecting indicator variables: more indicator variables = better results but no N defined, having a strong theoretical rationale for using specific indicator variables makes the process of identifying the classes easier, helps with interpreting the results and results in class solutions that have clearer application to practice, codage of the indicators (recoding, collapsing, ...), 3) structuring the data set: according to statistics programs recoded missing data, ..., 4) estimators: decide which estimator to use depending on criteria such as sample size, number of variables, computation speed, the management of missing data, reporting conventions in researchers' discipline, 5) conducting LCA: conduct a sequence of models starting with a 1-class model and then specifying models with 1 additional class at a time, then compare the models until the best model is identified, 6) selecting a class solution: statistical criteria always be evaluated in conjunction with interpretability (class solution with superior statistics is not useful if it makes no sense theoretically) with multiple fit statistics should be used, BIC (or AIC or sample-size adjusted BIC or consistent AIC or likelihood tests, bootstrapped likelihood ratio test) and theoretical interpretability, diagnostic statistic as average latent class posterior probability or entropy (better if >0.8), 7) including covariates in LCA: it is possible, identify the final class model then add covariates, 8) interpretation and implication for practice: understand commonalities and differences across individuals, 9) validating the LCA model: determining if class assignments are related as expected to relevant outcomes, 10) reporting the results: detail study procedures and results with clarity and coherence (reporting the BIC and at least 2 additional fit indices, entropy, % and size of the smallest class, smallest off-diagonal value of the average latent class posterior proba matrix)

LCA and LPA = to identify subpopulations of people with those subpopulatins being characterized by distinct configurations of scores on a set of variables; model-based and prototypical which means they yield probabilistic rather than hard assignment; LCA = categorical indicators vs. LPA = continuous indicators; assuming that these indicators are normally distributed within each latent profile, the indicators are unrelated within each latent profile, LPA models the distri of observed scores on a set of indicators as a function of the proba of membership to the latent classes and each class'normal density, LPA model decomposes the variance of each indicator into 2 components: a between-profile component that captures how far the profile-specific means are from the general mean and a within-profile component containing the profile-specific variances; LPA can be used to estimate profiles differing in both means and variances and in which only the means are profile-specific. Limits = selecting the optimal number of latent profiles: models with an increasing number of latent profiles are tested after which the most optimal 1 is selected based on interpretability and theoretical conformity of the solution, statistical adequacy and statistical indicators (as BIC, sample-adjusted BIC, Consistent AIC, Bootstrap Likelihood Test, elbow plot); labeling of profiles: use of differences in the unique pattern (shape differences, level differences, scatter differences, ...); incorporated covariates: should only be included once the optimal unconditional profile solution is selected, include them in the final solution directly but has to not change the profile solution; multi-group invariance testing: whether profiles found in 1 sample generalize across known subpopulations, for LCA it has been tested with a 3 steps approach and it exists a 6 steps for LPA

1. **Modèle de Markov Caché**

HMM = generative stat models used for the modelling of stochastic time-varying proccesses; defined by a vector of initial proba, a transition matrix and emissions densities of the hidden states; model the interaction between 2 sequences: 1) an observable sequence corresponding to the observed time-series and referred to as the observation sequence and 2) an unobservable sequence the discrete-time Markov chain referred to as the 'hidden state'; in sequences described by a first-order Markov chain each value depends exclusively on the previous value; From the set of observable sequences the goal is to learn the best parameter-set then from the sequence of observations and the parameter-set of the HMM we can estimate the most probable sequence of states by using Viterbi algo; Limitations: to determine the number of states = to few states connot take into account occasional states whereas too many cause overfitting; fitting HMMs with different nb of states and then comparing the goodness-of-fit of the resulting HMMs (log-proba of each sequence using the forward algo); choice of emission distributionplot the estimated distribution against the empirical distribution shiw how well the models capture the behaviour of the data; extracted features - patient's transition matrices: with the x-state HMM enables us to compute the transition matrix built with the proportions of transitions from one state to another and features = the elements characterize the frequency of switches from 1 state to another; then the matrix of extracted features was applied to K-means

LM (latent Markov) models are designed for the analysis of univariate and multivariate longitudinal/panel data based on the repeated observation of a sample of units across time; to study the evolution of an individual characteristic of interest when this characteristic is not directly observable; the model at issue rely on a latent process following a Markov chain; to account for time-varying unobserved heterogeneity in addition to the effect of observable covariates on the response variables; the initial and transition proba of the latent process are allowed to vary across different latent subpopulations; conceived quite similarly to hidden Markov models for time-series data but they are tailored to longitudinal data where many individuals are observerd at only a few occasions (no more than 10); some R packages exist; allows for missing responses, drop-out, non-monotonic missingness under the MAR assumption; computationally efficient algo are implemented for estimation and prediction of the latent states; maximum likelihood estimation with covariates, log-likelihhod function can be maximized by the EM algo; after the model is estimated, standard errore for the parameter estimates may be obtained on the basis of the observed info matrix as the square root of the corresponding diagonal element of the inverse of the matrix; selecting the nb of latent states: inc ertain applications the nb of latent states can be a priori defined as in the univaraite case in which it is reasonable to fix the number to the nb of response categories, otherwise AIC and the BIC = final model is the best compromise between goodness-of-fit and complexity; LM model rules out individual covaraites and assumes that the conditional response proba are time homogeneous; can add individual covariates in the measurement model but the latent variables account for the unobserved heterogeneity between individuals that we cannot explain on the basis of the observable covariates; when covariates are included in the latent model we suppose that the response variables measure the indivdual characteristic of interest taht is represented by the latent variables, this characteristic is not directly observable and may evolve over time; mixed LM model = allowed the parameters of the latent process to vary in different latent subpopulations defined by an additional discrete latent variable

Markov model = analyzing incomplete dichotomous outcomes; offer an intuitive approach for modeling patients' transitions between a number of discrete states over time; can be either discrete time or continuous time; Continuous-time Markov models have both theoretical and practical advantages iver their discrete-time counterparts; analysis focused on the estimation of the transition rates between the different states rather than, proba of transition which depend on the elapsed time; provide an array of clinically interesting estimates, including additional unobserved states to accommodate assumptions regarding missingness mechanism and the outcomes of patients dropping out; Bayesian framework because it offers increased flexibility in modeling, MCMC techniques to fit all models but a frequentist approach can be also possible; the 3-state model = 3 different Markov states, 1) Markov assumptions: proba of a transition from a state to another does not depend on the previous states visited or on the time spent in current state and we assume that the transition rates are constant through time and are common for all patients randomized in the same treatment arm, model can be extended to include patient-level random effects in the transition rate to make the transition rates dependable on patient-level covariates or to account for time-dependency in the transition rates; 2) estimating the model parameters: estimate the transition rates from the transition proba that are directly estimable from the data, Bayesian framework by using MCMC software to estimate the parameters of the model (frequentist methods to maximize the likelihood can be use); 3) making inferences on relative treatment effects: inference on relative treatment effects can be made at any time point deemed to be clinically interesting, Advantages = proba for a patient to drop out due to inefficacy, the proba of a patient to drop out after responding, the expected time spent in each state can be a useful estimate as it summarizes the effect of the treatment in an easy-to-understand manner; 4) including random effects and patient-level covariates: the model can be extended to include random effects after assuming that the transition rates for patients randomized in a treatment arm are not fixed but exchangeable, that is, coming from a common distribution, MAR assumption of the dropout, dropout is dependent on observed data and can be dependent on (observed) patient characteristics, MCAR assumption exclude any covariates from the analysis so the dropout rate does not depend on whether or not a patient has responded to the treatment or any other observable or unobservable characteristics; 5) Modeling unobserved response: make inference about treatment effects in patients that drop out, extend the method we have presented so far by assuming that patients who drop out continue to undergo transitions between an unobserved nonresponse and an unobserved response state, becomes 4-state model with 8 transition rates but this transition rates cannot be directly estimated from the observed data and unobserved data (outcomes of dropout patients) would be needed for fitting this model, several scenarios are discussed: a) MCAR can be modeled by setting the assumption that the dropout rates do not depend on either observed or unobserved data and they are equal among responders and nonresponders and that the transitions between the unobserved response and nonresponse states follow the same pattern as the transitions of patients still in the study (dropout and response are 2 independent procedures and the transitions rates needs to be independent of any covariates); b) dropout rates being different across responders and nonresponders but to only depend on observed data (MAR); c) MNAR: the dropout rates depend on both observed and unobserved outcomes and the unobserved outcomes cannot be predicted solely by using the observed data, the transition rates between the unobserved states can be assumed to be equal to the corresponding observed ones; d) the LOCF-like missingness: all dropouts remain in the last observed state, estimation of transition rates and proba uses all available observations and not just the last observation from each patient, MCAR is necessary but not sufficient assumption.

HJMM = describe the relationship between 2 stochastic processes, an observable outcome process and an underlying hidden state process; the number of hidden state of an HMM is usually unknown, AIC and BIC can be used for selecting the number of hidden states but not justified for continuous-time HMM; the between-transition of states is described by a hidden Markov process instead of a Markov Chain is able to accommodate different types of lingitudinal data that are regularly, irregularly or continuously collected; maximize the observed-data log-likelihood function for estimation by using formula for differentiating and integrating exponential matrices --> resulting = maximum likelihood estimators; novel penalized method to estimate the nb of hidden states of HMM and establish the consistency of the resulting estimator (for continuous-time HMMs); Model specification: the longitudinal response follows a known distri with density which is parametrized by a function of covariates and hidden state dependent 'regression' coeff, allows for subject-specific generating matrix and initial values which elicit sunject-specific hidden traj, can be regarded as a generalization of GLMMs because it allows for outcome-vocariate association heterogeneity driven by transitions on Markovian hidden regimes; ML estimator: avoid the use of the EM algo, proposed estimation procedure directly maximizes the likelihood for sampling hidden states through the forward and backward algo and time-consuming iterations, implementation = simpler and computationally more efficient than the EM algo with faster convergence, with the ML estimated parameters we can recover the traj of the hidden Markov process for each observed subject through the 'Maximum a Posteriori' path, scatter plot of residuals = info about the goodness-of-fit of the posited model; Nb of hidden states: a knew penalized likelihood method to select the nb of hidden states of an HMM: 1) rank the state-specific parameters such that all the states are well labeled and the same states are adjoining, 2) select the penalty function to be smoothly clipped absolute deviation penalty, the proba in the proposed continuous-time HMM method and the elements of the generating matrix depend on covariate vectors and between-state transitions are associated with a multi-dimensional parameter vector (and based on the double penalized procedure), local quadratic approximation, minimizing BIC; Advantages = only 1 tuning parameter and computationally more efficient; Limits = proposed estimation procedure cannot be extended to accommodate non-homogeneous transitions in a straightforward manner and the associated inference become extremely challenging, not work well when the number of hidden states is large and/or covariates are of high dimension.

Latent Markov model = a latent state variable is related to observed data; time-canstant parameters and predictors but intensive longitudinal data lend themselves well to models involving time-varying parameters, ILD suitable for separate modeling of each individual person's data, mixed Markov modeling approach can borrow strength across persons, obatin estimates of the average population parameters and include predictors to investigate their relationship with the individual differences; 1st situation) Observed Markov Model: know the state that a person is in at each occasion, to investigate the temporal dynamics of the state-switching process, states may correspond to the categories of a discrete observed variable or may be created by discretizing continuous (or multivariate) data, analyze the observed state transitions over time and estimate the transition proba. 2nd situation) The LMM (latent Markov model): the state-switching process is not directly observed or is observed with methods that are expected to involve substantial measurements error, LMM = HMM; used to sstudy a latent state-switching process so needs to have a measurement (or conditional) model part which links the unobserved states of the system to the observed outcome variable(s) --> cause bias in the estimated transition proba, number of latent states based on theory or based on the fit of models with differing numbers of states; LMM does not always follow from the data as it does in the OMM, but the specification of the transition model id very similar excpt taht an LMM includes additional parameters representing the proba of starting out in the different states at the 1st occasion; LMM can be applied to univariate or multivariate and continuous or discrete data. To account for individual differences in dynamics = include subject-specific random effects by modeling the logits or estimating the proba directly (use regression to predict the logits), in a mixed LMM the latent state at the 1st time point is still modeled using fixed proba then the states are modeled using categorical distri where the transition proba depend on the individual (OMM and LMM); to predict some of the individual differences in switching proba we can add 1 or more predictor variables; the measurement model part in an LMM can also be specified to include random effect; Model estimation (Bayesian approach) if not Bayesian = frequentist estimation methods i.e. LMM with covariates, latent subgroups, random effects to account forr indiv diff but Advantages for Bayesian estimation = presents a highly flexible alternative approach and it is robust, more appropriate for small samples, Bayesian multiple imputation handling incomplete data without losing information or introducing bias (MAR), can extract the latent states from a fitted LMM without separate state decoding, Limits for Bayesian estimation = require the specification of prior distri for model param indicating the range of plausible values that they can take. Bayesian LMM = Limits = no criterion for model fit or model selection to choose the nb of latent state when that nb is not known a priori so estimate the nb of latent states by using reversible jump Markov Chain Monte Carlo estimation, label switching so have to impose order restrictions on the parameters for different states to enforce a unique labeling; to deal with between-person differences = 1st solution) include 1 or more covariates in the model wihtout allowing for residual unexplained individual variation, 2nd solution) mixture Markov model (= mixed Markov model) which tends to be computationally easire than the mixed Markov model; mixture Markov model distinguishes between a number of latent classes that differ from each other with regard to the model parameters <hile within each class no individual differences between persons are allowed; but assumption that there are a limited nb of homogeneous latent classes; choose the nb of latent subgroups using AIC, BIC in the frequentist framework, in the Bayesian apporach = difficult, Bayesian mixture model is computationally not easier than a mixed model; 3rd solution) inclusion of a continuous random effects distribution by specifying Markov model; there are alternative, semi-parametric or non-parametric approaches that involve less restrictive and more generalizable model specifications; allow for interpersonal variation in both parts of an LMM

HM = models time dependence in a flexible way and allow to perform a dynamic model-based clustering; the same individual is allowed to move between clusters across time and these dynamics are provided in terms of traj; a sequence of discrete latent vairables rather than a single latent variable is associated to every individual so give rise to a hidden process assumed to follow a Markov chain; the states of this chain = latent clusters or subpopulations of homogeneous individuals sharing the same latent characteristics; can handle individual covariates that affect the distri of the latent states and in particular the initial and the transition proba of the Markov chain; takes into account missing data (partially missing outcomes at a given time occasion, completely missing outcomes at 1 occasion without dropout from the sample of individual, dropout from the sample; for the 2 1st MAR assumption according to which the missing pattern is independent of the missing responses given the observed data and the 3rd dropout is not ignorable and specifying a model for the missing data mechanism is in order); ML approach to estimate the proposed model using extended EM algo based on suitable recursions and that relies on certain techniques used to estimate a finite mixture of Gaussian distri; selection of the number of components based on the AIC and BIC obtained through penalizations of the maximum log-likelihood, BIC preferred because AIC overestimate the nb of components; or cross-validated log-likelihood in which the data are repeatedly divided into 2 randomly chosen partitions (the training subset and the test subset), for every partition the model is estimated on the training subset and the corresponding log-likelihood is evaluated on the testing data, model with highest value of cross-validated log-likelihood is selected as the best one; finally on the basis of the estimation results, model-based clustering is performed with the maximum a posteriori (MAP) rule (with missing data, it is possible to perform a sort of multiple imputation taht allows us to predict the missing responses contitionally or unconditionally to the model component; Limits = label switching that is the invariance of the likelihood with respect to permutations of the first hidden states, the dropout state is instead not exchangeable with the other states; when covariates are included in the latent model, the interest is in modeling the effect of covariates on the distribution of the latent process; the inference (EM algo) is based on the solution corresponding to the largest value of the log-likelihood at convergence = the global maximum; once parameter estimates are computed for a given nb of latent states, the corresponding standard errors may be obtained on the basis of different methods (ex: bootstrap procedure but computational cost); model selection = BIC or good compromise between goodness-to-fit and interpretability of the resulting latent states; once nb of states selected, dynamic clustering is performed by assigning every unit to a latent state at each time occasion, the EM algo provides the estimated posterior proba

1. **Modèle mixte**

Linear mixed effects = used in the analysis of clustered or longitudinal data; estimate the relationship between the dependent variable and the fixed effects and random effects of independent variables by considering both means and covariances; too many predictors = inference and prediction becomes too complex and infeasible; can penalized estimation of fixed effects and select of random effects: Ex 1) estimates fixed effects, random effects and the covariance structure of the selected random effects simultaneously in a model with 1 penalty function but for high-dimensional it necessary to reduce the dimension of the data before using the method because of the use of the EM algo (not efficient, not be plausible and slow convergence rate and computational burden) --> maximizing the log-likelihood function and use EM algo / Ex 2) selects important fixed and random effects independently in 2 separate models use Proxy matrices to account for unknown variance-covariance structure of the random effects during the selections, minimized the penalized likelihood equation for the fixed effects parameter then the selection of the random effects is accomplished through Bayesian methods of deriving the restricted posterior distribution of the random effects and penalizing this solution of the restricted posterior mode, for high-dimensional data the dim of fixed effects must be lowered to below the sample size before using the methods, Ex 3) selects and estimates fixed effects, random effects and the covariance structure of the selected random effects simultaneously in a linear mixed effects model using 2 penalty functions, using a modified log-likelihood incorporating the REML (Restricted Maximum Likelihood)

GLMM = a sort of hybrid between linear mixed model and a generalized linear model; accomodates repeated measures data for which the usual assumption of independent observations is untenable, accomodates a non-normally distributed dependent variable; take into account multiple non-normally distributed dependent variables measured more than once each and a mix of continuous and categorical independent variables thanks to the artificial construction of a 'pseudo dependent variable' that encompasses all measurements of each dependent variable, a 'pseudo independent variable' is also created to identify the origin of each instance of the pseudo dependent variable; Advantages = GLMM coefficients interpretation: subject to the caveat that the data analyst is also controlling for random effects, estimated odds ratios or rate ratios, accomodate not just 2 but 3 or more dependent variables simultaneously since a given study may monitor more than 2 different drugs or health behaviors, missing data can be handled; Limits = the estimated dependent variable trajectories based on strata defined by the independent variables may not exhibit clinically relevant differences even if the differences are statistically significant, observations not recorded are assumed to be missing completely at random (such an assumption is generally implausible)

1. **Modèle joint**

Model the observation process and the repeated measures process using a joint longitudinal and survival model; conditional on random effects, the submodel for the time to each observation is a proportional hazards model with hazard for gap time; the 2 processes are linked together via the shared, individual-specific, random effect; model was fitted using maximum likelihood; the likelihood does not have a closed form as it is necessary to integrate out the distribution of the random effects, methods such as Gaussian quadrature and Monte Carlo integration can be used; for the joint model to be valid, the observation process has to be at least at random; this model can easily be extended to multiple random effects, to different parametric and flexible parametric baseline hazard formulations for the recurrent-events model and to include other outcomes; the bias induced by an informative observations process can be adjusted by using the IIVW method as an extension of the IPW method using a marginal regression by weighting each observation by the inverse of the proba of each measurement to be recorded; it creates a pseudopop in which the observation process is static and can be ignored; weights can be estimated by fitting a regression model including all covariates than inform the observation process and further stabilized to increase efficiency; weighting model could include current and past values of any covariate that may affect the visiting process but covariates might be related to the observation process, should be included in the weighting model otherwise bias will incur; 2 adjustments: because the last entry for each indiv represents the end of follow-up of the study, each weight is shifted by 1 time point and given that each indiv is observed at least once, a weight of 1 is assigned to the first observation of each individual; the marginal model for the longitudinal outcome is then fit using generalized estimating equations and including the normalized inverse intensity of visit weights as proba weights in the model; Advantages = mixed models performed worse than joint model, produce unbiased estimates of the regression coeff (overmodelling the observations process does not seem to introduce bias in the analysis), estimating the association between 2 outcomes, additional random effects could be introduced in the model to account for heterogeneity in the traj of the longitudinal outcome over time, the functional form of the effect of time could also be generalized by using fractional polynomials or splines (AIC, BIC), drop out could be add in the time-to-event submodel

Estimation of a joint model with a solitary time-to-event outcomen and multivaraite longitudinal data with a linear random effects structure, statistical estimation = quasi Monte Carlo that generate low discrepancy sequences, highly suitable to high-dimensonal generalized linear mixed models; submodel for the time-to-event outcome = hazard model; EM algo for fitting the model by tretaing the random effects as missing data, the parameters in the Cox proportional hazards submodel are estimated by a 1-step Newton-Rasphson or a quesi-Newton 1-step update that is an analogue of teh Gauss-Newton method, standard errors can be approximated after the EM algo has convergened using the empirical info matrix approximation allowing for Wald-like confidence intervals to be estimated, alternatively bootstrap estimation can be used but at increased computational expense; Monte Carlo integration is a probalistic representation of the integral (Ordinary Monte Carlo and Antithetic Monte Carlo are predicated on a probalistic interpretation, that is they use random sequences which ensure convergence, Quasi Monte Carlo use quasi-random sequences which are deterministic referred as low-order deterministic sequences); the motivation of QMC integration is to reduce the order of convergence and nb of nodes required by using nodes taht are scattered more uniformly than pseudo-random points which by virtue of independence often display clusterings; quasi-random sequences yield smaller errors than standard Monte Carlo integration methods; Limits of QMC method = do not permit the estimation of MC error; Advantages of QMC method = reduced the computational time required to fit multivariate joint models with a small sample size and was comparable to AMC for moderate sample size

1. **ARIMA + Cross Correlation**

ARIMA = Autoregression integrated moving average = useful tool for analyzing nonstationary time-series data containing ordinary or seasonal trends (climate-sensitive diseases suitable data for ARIMA); To establish whether a relationship exists between 2 variables observed over time = obvious approach is to compute correlation coefficients between 2 time series over a range of time lags: 1) fitted ARIMA models with the time series of the incidence of RRv, 2) checked the goodness-of-fit to the models for adequacy using both time-series (residual autocorrelation functions) and classic tools (the normality of residuals), 3) divided the data file into 2 data sets (model-building and validation set), 4) used the former to construct the ARIMA model and the latter to validate the model.

Can add factors with lag (CCF to evaluate the correlation between the meteo factors and mumps) 🡪 ARIMAX (ARIMA with exogenous variables)

ARIMA can capture the periodicity, trend and randomness of data, high prediction accuracy; ARIMAX: 1) time-delay correlation between variables by CCF, 2) forecast analysis using ARIMA model method, 3) add exogenous variables; 3 steps to modeling ARIMA model: 1) model recognition: stability and seasonality of data are analyzed and the difference method is used to stabilize the data (ADF, ACF, PACF), 2) parameter estimation and model test: min AIC and min SBC to determine the best model and Box-Ljung statistic to do white noise test on the residual sequence (if p<0,05, model ok for prediction), 3) prediction applications: selected models are used to fit the modeling data and to predict the values for a future period (RSME to measure the ability of fitting and forecasting, smaller = better)

cross-correlation = for estimating the association between events in 2 time series, the correlation between 2 time-varying stimuli or events over time intervals that may or may not be coincident. A vector of sequential occasions of measurement is selected from each time series such that both vectors contain the same number of occasions and then the Pearson product-moment correlation is calculated for these 2 vectors; vectors may or may not begin at the same occasion of measurement; lag or offset = interval of time separating the beginning of measurement for the 2 vectors; vector of sequential measurements sampled from a time series = window; most researchers assume the observed data come from stationary processes (means and variances = constant over time); Windowed cross-correlation = one way to examine how the strengths and lags of association between 2 time series are changing over time is to use only short intervals of data from each time series to estimate the association and then select these windows so that their starting points represent increasing elapsed time from the beginning of the experiment. Advantages = only making an assumption of local stationarity rather than assuming stationarity over the whole time series.

Time-lagged cross-correlation refers to the correlation between 2 time series shifted relatively in time, DCCA = to analyze power-law cross-correlations between nonstationary time series, by detrending local trends DCCA ensures that the results obtained are not affected by trends, can uncover more hidden correlation information than other analyses. Step 1: determine the profiles, 2 time series of equal length, Step 2: divide the profiles, into overlapping boxes, step 3: calculated time-lagged covariance of the residuals in each box, detrended walk is defined as the difference between the original walk and the local trend, the time-lagged covariance of the residuals in each box is calculated, Step 4: average over all boxes to obtain the detrended time-lagged covariance function, Step 5: compute time-lagged DCCA cross-correlation coeff

DCCA = to analyze power-law cross-correlations between nonstationary time series, DPCCA = coefficient is upgraded by combining partial-correlation technique, therefore it is expected to be useful in quantifying correlations of mmulti-signals in a complex system; can be used to diagnose 'intrinsic' relations of 2 nonstationary signals (with influences of other signals removed) on different time scles; based on DCCA but improved by including PCCA (partial cross-correlation analysis); advantages of both: robust results even when nonlinear trends are mixed in the data and further show relations between the 2 considered data on different time scles + the ability in investigating correlations when multi-signals are linked via inter-woven ties; in general, better performance in dealing with correlations in complex system; limits = multi-signals should have linear relationships with each other but this deficiency reduced by DCCA

1. **DTW (Dynamic Time Wrapping)**

*Package dtw / TimeShift*

Regarde la similarité entre les 2 séries temporelles

Regarde le meilleur alignement entre les 2 séries temporelles et estime la valeur de décalage temporelle entre les séries à chaque point de temps + estime la probabilité d’observé un décalage temporelle donné à chaque point de temps lorsque le décalage actuel entre les 2 séries temporelles est 0

Il vaut mieux avoir des séquences de même longueur mais de taille différente est possible ; déterminer une taille de fenêtre au préalable pour la reproductivité des résultats + little warping = good thing vs. Too much warping = bad thing => pour la précision de DTW (dépend de la taille de la fenêtre et de la taille des données) ; rapide ; select alignment combinations that preserve the order of the time series = min Euclidean distance ; effective in calculating time shift between 2 time series even when the proportion of noise is 20-30% of the total variance ; performs well for expression profiles containing both recurrent and non-recurrent changes and can estimate variation in the amplitude and direction of the time shift.

High calculation cost and need a specific averaging process to provide cluster centroids; can deal with trajectories having different nb of points; process = 1) linkage strategies: to the hierarchy of the partition to define the dissimilarities between clusters from the individual pairwise dissimilarity matrix: average, complete, Ward, 2) choosing the number of clusters, CVI, 3) Centroids: DTW barycenter averaging is a global averaging strategy providing centroids for groups of time series well adapted to the DTW dissimilarity, depend on an initialization step.

More flexible approaches for time series analysis, requires that test and reference timeseries are known in full to provide a distance. Based on the concept that the similarity between 2 time series should be computed by aligning significant patterns by locally deforming the time axis in order to minimize the cumulative difference between the aligned points; method is suitable for matching time series containing patterns that are qualitatively similar but have different lengths and paces. Set of constraints: start-point constraint, end-point constraint (mapping covers both the time series completely), monotonicity (no time loops), local slope constraint; optimal = minimizes the distance between the 2 warped time series; multivariate, nominal or mixed data; can handle incomplete input time series with complete references with OE-DTW (open-end DTW)

D’autres méthodes existent comme les méthodes de clustering (cross-sectional data, continuous, categorical, mixed data), les méthodes d’analyse de séquence (categorical longitudinal data to describe the sequence of events), … (Lore et al., 2020), K-means for longitudinal data, change indices + PCA + k-means (Verboon, 2022), Factor Mixture Analysis (FMA), k-means, Mixture Regression Analysis, Configural Frequency Analysis (Hofmans, 2020), latent growth curve modeling (Charnigo, 2011) => Mais pas forcément idéales pour les données longitudinales/télémonitoring/de trajectoire

# II/ Comment s’en servir ? Quel choix faire ?

Selon plusieurs critères, nous devons opter pour l’une ou l’autre méthode : données manquantes, durée des données, perdus de vues, standardisation des variables, …

Lister les critères et expliquer quelles méthodes dans quels cas de données (ex : si trop peu d’amélioration d’ESS, ne pas faire un modèle joint mais plutôt un modèle mixte)

Faire un tableau de synthèse :

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Méthode** | **Description** | **Avantages** | **Limites** | **Pour quel type de données ?** | **Quand l’utiliser ?** | **Package R (?)** |
|  |  |  |  |  |  |  |

# III/ Exemple sur nos données (?)

Expliquer la base de données, quelle(s) méthode(s) on a choisi et pourquoi + quelques résultats (?)

1. E-Meuse
2. Agir A Dom’, MARS

# Recherche PUBMED/InsermBiblio

Donner les mots-clefs utilisés + faire un tableau récapitulatif de la biblio (titre, auteur, date, méthode utilisée, type de données, objectif)

# Références